Kalle Remm, Jaanus Remm, Ants Kaasik

Ruumiliste loodusandmete statistiline analüüs

Õpik-käsiraamat

Tartu Ülikooli Ökoloogia ja Maateaduste Instituut

Tartu 2012

Ruumiliste loodusandmete statistiline analüüs. Õpik-käsiraamat.
ISBN: 978-9985-4-0712-7 (pdf)

Publitseeritud eestikeelsete digitaalsete õpikute hoidlas
http://site.ebrary.com/lib/tartu
aadressil
http://hdl.handle.net/10062/26456

Tehnilised toimetajad:
Joonas Remm
Allan Rajavee
Egle Rüütli
Annika Murov

Tiitellehe kujundus:
Jane Remm, Kalle Remm ja Tiiu Kelviste

Veebileht: http://kalleremm.ee/RASA/

Saateks

Kättesaadava teabe hulk on suurem, kui üksikisik seda vastu võtta suudab. Maailma mõistmisel, sobiva eeskuju, töövahendi ja meetodi leidmisel on abiks andmebaasid, infosüsteemid ja tegutsemisjuhised. Loodetavasti ei ole vajalikkust minetanud ka ühe teadusvaldkonna teavet süstematiseerivad, üldistavad ning näiteid esitavad emakeelsed õpikud, käsiraamatud ja monograafiad. See tekst on mõeldud nii abimaterjaliks andmetöötluse ja ruumilise analüüsiga seotud õppeainete õppimisel kui ka teaduri käsiraamatuks, millest võib abi olla uurimistöö planeerimisel, andmete töötlemisel ja tulemuste kirjapanekul. Eeldatud on, et õppevahendina kasutamisel lisandub iga õppeaine puhul õppeaine rõhuasetustele, tudengite huvidele ja varasemate teadmiste tasemele vastav tööjuhis või veebilahendus.

Raamatu aluseks oli dotsent Jüri Roosaare õppeaine *Ruumiandmete analüüs* ajel tekkinud huvi ja sellest huvist kantuna esimese autori poolt aastatel 1998 kuni 2008 koostatud käsikiri, mida aastatel 2011 ja 2012 nii uuemate kui ka vanemate publikatsioonide alusel täiendati. Jaanus Remm kohandas esimese peatüki toorteksti selle raamatu jaoks sobivaks, kirjutas peatüki 4.2.2 ning lisas loomaökoloogilisi näiteid ja ülesandeid. Ants Kaasik kontrollis valemeid ja matemaatika terminoloogia kasutust ning täiendas teksti vastavalt vajadusele. Kõik kolm tiitellehel märgitud autorit osalesid kogu teksti viimistlemisel. Peatüki 6.2.6 kirjutas Tiiu Kelviste. Oleme igati püüdnud vigu vältida, aga inimene jääb ekslikuks – raamatu vigade parandused ja täiendused on selle raamatu veebilehel kalleremm.ee/RASA. Viidatud kirjanduse andmebaasi ja viidatud artiklite koopiaid võib küsida esimeselt autorilt.

See teos on valminud paljude inimeste koostöö tulemusel. Autorid avaldavad tänu kõigile, kes on selle kirjatöö valmimisele kriitiliste märkuste, hea sõna ja rõõmsa meelega kaasa aidanud. Esile tõstmist väärivad tehnilised toimetajad Joonas Remm ja Egle Rüütli, keeletoimetajad Allan Rajavee ja Annika Murov, kujunduse ideede pakkuja Jane Remm ning jooniste vormistamise nõustaja Liina Remm. Lisaks valdavale osale tehnilise toimetamise tööst juhtis Joonas Remm tähelepanu paljudele segastele kohtadele algses käsikirjas ja tegi sisukaid parandusettepanekuid, mis on lähedane autori loomingulisele tööle. Täname kolleege, kes nõuannetega abistasid ja suhtusid mõistvalt autorite pühendumusse selle raamatu kirjutamise perioodil. Täname kõiki retsensente, eriti tänulikud oleme retsensentidele Jonne Kottale, Jaan Liirale, Toomas Tammarule ja Tõnu Mõlsile, kes saatsid arvukalt asjatundlikke parandusettepanekuid.

Peatükk 5.6 on pühendatud Hans Remmi mälestusele, kes õpetas 1960ndatel ja 1970ndatel aastatel zoogeograafiat nii zooloogidele kui ka geograafidele.

Raamatu koostamist toetati Tartu Ülikooli kirjastamisnõukogu vahenditest ja Eesti Vabariigi Teadus- ja Haridusministeeriumi sihtfinantseeritavast uurimisteemast SF0180049s09 ja SF0180122s08 ning teaduse tippkeskuse FIBIR poolt.

Sisukord

Eessõna

Kalle Remm jt. „Ruumiliste loodusandmete statistiline analüüs“ pakub väärtuslikku täiendust geoinformaatika ja ökoloogia suuna tudengite ja teadlaskonna käsiraamatute hulka. Õpikust leiavad aga huvitavat lugemist ka teiste uurimissuundade esindajad, kellel tuleb oma töös rohkem või vähem kokku puutuda elusloodu, elusloodust kujundavate protsesside ning nende protsesside tagajärjel tekkinud mustritega.

Soovitan julgesti süveneda Kalle Remmi ja kaasautorite poolt avaldatud käsiraamatusse ja seda mitmel heal põhjusel.

(1) Pole ju teadmata, et ajal, mil inglisekeelne maailm tungib kõikidesse eluvaldkondadesse ja eestikeelne terminoloogia ei jõua tohutult kiire arenguga sammu pidada, pakub käesolev käsiraamat pidepunkte statistika aga ka näiteks intellektitehnika alase emakeelsete mõistete koha pealt.

(2) Raamat on kirjutatud lihtsas ja ladusas keeles ning arusaamist ei piira ka napid statistikaalased teadmised. Nende tarbeks, kes vajavad väikest sissejuhatust statistiliste analüüside maailma, on autorid esimeses kolmes peatükis üldistavalt kokku võtnud olulisema, mida läheb tarvis raamatu teises pooles esitatud ruumiliste meetodite mõistmiseks. Spetsiaalselt ruumilise analüüsi meetodeid käsitlevad peatükid on neljas, viies ja kuues, mis annavad lisaks meetodite põhialuste kirjeldamisele ka ülevaate iga teema erialasest kirjandusest ja autoritepoolseid seisukohti.

(3) Aastasadu on inimesed imetlenud meid ümbritsevat elustikku, püüdnud mõista elustikku mustreid kujundavaid protsesse ning (tihti praktilistest kaalutlustest lähtuvalt) üritanud ennustada liikide ja elupaikade paiknemist ruumis ning ajas. Kõigi nende analüüside puhul on ruumiliste meetodite osatähtsus pidevalt kasvanud. Ühelt poolt peegeldab see kaugseire andmekogumite paremat kättesaadavust ja arvutustehnika kiiret arengut, aga vähemtähtsam pole ka arusaam, et paiknemine ruumis on midagi sellist, mida ei ole võimalik pelgalt klassikaliste keskkonnatunnuste abil ennustada. Ruumi- ja ajamustrite uurimine iseenesest on üks võti ökoloogilistest protsessidest arusaamiseks.

(4) Praktilise poole pealt on tervitatav, et autorid on lisaks teoreetilistele ülevaadetele toonud näiteid erinevate meetodite kasutusest. Abiks on ka viited mitmetele tarkvaralahendustele (sh. autorite poolt loodud Constud programmile), mis võimaldab lugejal iseseisvalt kätt harjutada.

(5) Loomulikult ei võimalda raamatu formaat kõiki meetodeid süvitsi käsitleda ja valdkonna mahukus tingib osade peatükkide konspektiivsuse. Kiirel ajal on aga selline stiil tervitatav ning kellel tekib huvi mingi spetsiifilise teema vastu, leiab abi autorite poolt pakutud mahukast kirjandusülevaatest.

(6) Raamatust ei puudu ka autorite isiklik suhe valdkonna suhtes. Mitmest peatükist saab lugeda põnevatest ja põhjalikest uurimustest, milles autorid on rakendanud erinevaid ruumianalüüse, et paremini tundma õppida Eestimaa loodust.

Arvan, et käesolev käsiraamat on suundanäitava ja märgilise tähendusega suurendades eestlaste hulgas valdkonna populaarsust ning kasvatades teadmiste pagasit ruumilistest meetoditest. Minu jaoks oli raamat ütlemata inspireeriv ning siit ammutatud ideed leiavad lähiaastatel kasutust mereteaduse arendamisel. Minu lugupidamine autorite suhtes!

Jonne Kotta
merebioloog

Sissejuhatus

Enamik andmeid, mille põhjal ökoloogilistes ja sotsiaalteaduslikes uurimustes järeldusi tehakse, on ruumis positsioneeritavad ehk igal andmetabeli kirjel on olemas geograafiline asukoht. Sellest johtuvalt on õigustatud tõstatada mitmesuguseid küsimusi – alates asukoha arvestamise vajalikkusest ja olulisusest ühe või teise analüüsi puhul kuni olukordadeni, kus uuritavate objektide paiknemine ruumis ongi uurimisküsimuseks. Näiteks: kas puud paiknevad looduslikus metsas juhuslikult või on sama liiki puudel kalduvus grupeeruda; või kas kased kasvavad sagedamini kuuskede või mändide läheduses. Sageli pakuvad huvi uurimisobjektide esinemise või omaduste seosed teiste objektide paiknemisega või heterogeense keskkonna struktuurielementide konfiguratsiooniga. Seda küsimuste puhul: millisel määral on metsa koosseisu laigulisus või kartulimardika levik põllul või talude koondumine küladesse seletatav mullastiku ja teiste keskkonnatingimuste muutlikkusega ruumis, millisel määral koha ajalooga, millisel määral teadliku inimtegevusega, millisel määral puhtalt juhusega; millised on need tegurid, mis põhjustavad ruumilist muutlikkust ja kas neid teades õnnestuks paiknemismustri muutusi ette ennustada; või kui palju mõjutaks Ida-Virumaa põtrade paiknemist ühe uue kaevandusraudtee rajamine. Üheks levinumaks komplikatsioonide allikaks ruumiliselt paiknevate objektide uurimisel on ruumiline autokorrelatsioon – lähestikku paiknevad kohad kipuvad olema sarnased ainuüksi selle pärast, et on lähestikku. Näiteks kui naabrite juures on palju hiiri, siis kipuvad hiired teie korterisse ka juhul, kui seal neid midagi head ees ei oota. Siit kerkivad küsimused, mida on asjakohane uurimuste planeerimisel arvestada, nagu näiteks: kuidas arvestada põtrade arvukuse prognoosis põtrade oodatavat arvukust vaadeldava koha ümbruses; kui kauget ümbruskonda tuleks ühel või teisel juhul arvestada; või kuidas metsa struktuuri või maastikumustrit arvutimängude jaoks tõetruult modelleerida.

Ruumilisi loodusandmeid kasutavad uurimused on näiteks sellised, mis kirjeldavad looduse ruumilist struktuuri ja modelleerivad seda; käsitlevad elupaikade, liikide või populatsioonide ja erinevate geeni alleelide levikut ruumis, koosluste ruumilist struktuuri, planktoni paiknemist veekogudes, liikide ruumilist koosesinemist, puistu struktuuri, taimekahjurite paiknemist põllul, metsatulekahjude riski kaardistamist. Ruumikäsitlus võib seejuures olla nii ühe-, kahe-, kolme- või enamamõõtmeline.

Ökoloogiliste probleemide ruumilisele käsitluseni on jõudnud nii populatsiooniökoloogid, sünökoloogid, maastikuökoloogid, geneetikud, evolutsiooni uurijad kui ka liikide ja elupaikade kaardistajad. Kasutusse on tulnud terminid ruumiline ökoloogia (*spatial ecology*) ja ruumiline geneetika (*spatial genetics*). Näiteks on taimkatte ruumiline varieeruvus pikka aega intrigeerinud taimeökolooge ja geograafe. Paarkümmend aastat kestnud ja nüüd taanduv maastikumustrite analüüsi buum on seotud muuhulgas konkurentsiteooriaga, mis seletab koosluste struktuuri eelkõige liikidevahelise konkurentsiga, see on kooslusesiseste parameetritega. Liikide arvukuse ja leviku põhjuseid otsiti ka kooslusevälistest maastikuomaduste konfiguratsioonist. Konkurents ei pruugi olla looduslike koosluste kujundamisel määrav faktor, kuid mingid seosed liikide keskkonna ruumilise struktuuri ja liikidevaheliste suhete vahel siiski on.

Viimasel aastakümnel on looduse ruumilise struktuuri uurimist tugevasti stimuleerinud arvutusvõimaluste ja geoinformaatika kiire areng. Esmane etapp ruumiliste nähtuste analüüsis on, nagu looduse uurimisel ikka, kirjeldav. Ruumilise struktuuri ehk mustri otsimine ja kirjeldamine on tänapäeval saanud juba nii loomulikuks uurimisalaks, et seda võib pidada kirjeldava andmeanalüüsi üheks osaks. Ruumimustrite analüüsi kaugem eesmärk on ruumimustreid tekitavate protsesside parem mõistmine ja nende uute teadmiste kasutamine nii ajaliste kui ka ruumiliste prognooside andmisel.

Ruumiliste protsesside uurimise tasemel on väga oluline hüpoteeside õnnestunud püstitamine, uurimuse planeerimine ja protsesside jäljendamine. Mustri statistilise analüüsi puhul, mis on selle raamatu põhiteemaks, eeldatakse, et mustrit saab käsitleda kui vähemalt osaliselt juhusliku protsessi tulemust.

See raamat püüab olla abivahendiks leidmaks meetodeid, mille abil kirjeldada looduses esinevaid ruumilisi protsesse ja seaduspärasusi ning arvestada looduse uuritava osa ruumilisust ja objektide paiknemisest johtuvaid mõjusid. Raamatu kuuest peatükist kolm esimest ruumilisi ja ökoloogilisi andmeid otseselt ei käsitle, vaid on eelkõige mõeldud andmetöötluse põhitõdede meeldetuletamiseks. Alates neljandast peatükist, kus algavad ruumilise andmeanalüüsi meetodite teemad, on kolmanda taseme alapeatükkide lõpus lühidalt refereeritud teemakohaseid üksikuuringuid. Iga peatüki lõpus on kordamisküsimused, mille vastuseid saab kontrollida selle teose veebilehel http://kalleremm.ee/RASA/ pärast oma vastuse saatmist. Näiteid olulisematest publikatsioonidest iga teema kohta on lühidalt refereeritud kolmanda taseme alapeatükkide lõpus.

Põhjalikumalt on teoses käsitletud meetodeid, millega on autorid oma uurimistöös tegelenud. Need on sarnasuskordajad (ptk 2.2.1), sarnasusele tuginev järeldamine (ptk 2.3.3), punktmustrite kirjeldamine (ptk 4.1.2), suundade analüüs (ptk 4.2.2), üleminekuala eristamine (ptk 4.3.4), ruumilise autokorrelatsiooni modelleerimine (ptk 5.1), ümbruse mõjude kaasamine prognoosimudelisse (ptk 5.5), elupaigasobivuse modelleerimine ja kaardistamine (ptk 5.6.2), liikide esinemise puudumise kohtade sarnasusele tuginev hinnanguline kaardistamine ja selleks loodud infosüsteem Constud (ptk 3.4.6.1 ja 5.6.8.3) ja taimkatte välikaardistamise subjektiivsus (ptk 5.10.2). Mitmel joonisel on näitena esitatud varem publitseerimata tulemusi, mida on joonise allkirjas selgitatud. Teiste teemade käsitlus on valdavalt referatiivne ja vähem põhjalik – selle teose autorite panus oli peamiselt teavet süstematiseeriv ja eestikeelset esitust arendav.

Olulisemad mõisted on tekstis rasvases kirjas esile toodud kohas, kus mõistet selgitatakse. Terminite registris on viidatud ka mõistete teistele mainimislehekülgedele. Muutujate ja konstantide tähistamiseks kasutatakse sageli kreeka tähti. Kreeka tähestik koos tähtede eestikeelsete nimedega on lisas 1.

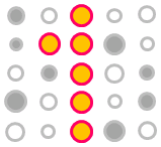
Tekstis olevate viitade klõpsamise järel saab järjehoidja juurest tagasi kohta, kus lugemise järg pooleli jäi, klahvikombinatsiooniga *Alt* ←, Adobe PDF-lehitseja Macintoshi versioonis käsuga ⌘ ← (*Cmd* ←).

Tarkvaralahendusi iga andmetöötlusliku probleemi lahendamiseks on viimasel ajal interneti kaudu saadaval nii palju, et neid ei jõua võrrelda ja soovitada. Pealegi tuleb igast suuremast paketist iga aasta või paari järel uus ning uusi võimalusi pakkuv versioon. Seetõttu ei ole tarkvarapakette pikemalt käsitletud, mainitud on vaid üksikuid. Vabalt allalaaditavaid programme kõikvõimalike statistiliste meetodite kasutamiseks on tarkvarakeskkonnas R (http://www.r-project.org), statistika tarkvara (sealhulgas vabavara) kohta on omaette artikkel Wikipedias (http://en.wikipedia.org/wiki/List_of_statistical_packages), ruumiandmete analüüsi tarkvara leiab ka Arizona ülikooli geo-andmete keskuse veebilehelt http://geodacenter.asu.edu/software ja Minnesota ülikooli ruumiliste andmete ja andmekaevandamise töörühma veebilehelt (http://www.spatial.cs.umn.edu/sdm_software.htm), geostatistika alal töötavate asjatundjate saidist AI_GEOSTATS (https://wiki.52north.org/bin/view/AI_GEOSTATS/WebHome); suur valik numbrilise ökoloogia tarkvara on alla laaditav Pierre Legendre veebilehelt (http://www.bio.umontreal.ca/legendre) ja Jari Oksase veebilehelt (http://cc.oulu.fi/~jarioksa/softhelp). Ruumiliste andmete analüüsi vahendeid on vabavaras SAM (*Spatial Analysis in Macroecology*) (Rangel *et al.* 2006, 2010, http://sam-spatial-analysis-in-macroecology.software.informer.com/), PASSaGE (Rosenberg ja Anderson 2011, http://www.passagesoftware.net) ja Spatstat

(http://www.spatstat.org). Ülevaade ruumilise analüüsi meetoditest koos näidete ja tarkvara soovitustega võib leida raamatust Fischer ja Getis (2010).

Kaardi- ja kaugseireandmeid Eestimaa kohta saab riigi maaameti geoportaalist (http://geoportaal.maaamet.ee), Tartu Ülikooli Ökoloogia ja Maateaduste Instituudi töötajad instituudi digiarhiivist (http://digiarhiiv.ut.ee). Infotehnoloogia ja ruumiandmete analüüsi terminite eestikeelseid määratlusi ja selgitusi on veebis http://www.geo.ut.ee/gis2000/terminid.html, http://vallaste.ee, http://isi.cbs.nl/glossary/blokes78.htm, Constud süsteemi õpiku (Remm ja Kelviste 2011c) lõpus (http://kalleremm.ee/Constud_Tutorial/Constud_est_terminid.pdf) ning mujalgi. Inglisekeelsed infotehnoloogia ja andmetöötluse mõisteid, termineid ja lühendeid selgitavad veebilehed on näiteks http://computer.yourdictionary.com/gis-glossary-h-t, http://isi.cbs.nl/glossary/bloken78.htm, http://www.gartner.com/technology/it-glossary, http://www.statsoft.com/textbook, http://www.itl.nist.gov/div898/handbook/index.htm, Constud süsteemi õpiku (Remm ja Kelviste 2011b) lõpus (http://kalleremm.ee/Constud_Tutorial/Constud_eng_terms.pdf). Loetelu lõpetuseks – ärge unustage Wikipediat!

1. Andmetöötluse alused



Teaduse olemuseks on peetud teadmiste korrastamist ja edasiarendamist. Teadmised omakorda on informatsioon ehk andmed, mis on teadvustatud ning mingil kujul talletatud, näiteks inimese mälus, raamatutes, arvuti kõvakettal vms. Järelikult on andmetöötlus teaduse üks keskseid osasid. Kitsamas mõttes andmetöötlus ehk andmevektoritesse ja -maatriksitesse koondatud arvudega (või muul viisil väljendatud väärtustega) tehtavad tehted ja toimingud on keskse tähtsusega enamuses teaduslikes uuringutes andes põhjenduse tehtavatele järeldustele. Uuringuid viiakse läbi selleks, et kirjeldada huvi all olevate objektide hulka ehk **üldkogumit**, hinnata mitte teada olevaid tunnuste väärtusi või ennustada üldkogumis toimuvaid protsesse.

Kogu huvi all olevat objektide hulka kirjeldada on lihtne, kui objekte ei ole palju. Paraku on ökoloogias ja maateadustes uurimisobjektide vähene hulk pigem erand. Enamasti ei suudeta kõiki loodusobjekte vaadelda ja kirjeldada, sest suure hulga uurimisobjektide puhul on kõigi uurimis-objektide kohta tunnuste registreerimine enamasti liiga kallis ja aeganõudev. Statistilised meetodid võimaldavad teha järeldusi kogu üldkogumi kohta väiksema hulga objektide ehk **valimi** põhjal. Statistilises terminoloogias on uurimisobjektiks üldkogum ja tunnuseid mõõdetakse üldkogumi osadel ehk elementaarobjektidel. Objektid ise ja ühe objekti kohta ühe vaatlusega saadud tunnuste väärtuste kogum võib olla mitmesugune, üldises tähenduses nimetatakse seda lihtsalt **vaatluseks**.

Andmetöötluse põhilised vahendid pärinevad matemaatilisest statistikast, kuid lisaks kombinee-ritakse ka teisi meetodeid, nagu näiteks graafilist analüüsi, analoogide otsimist, andmete tähenduse selgitamist. Andmetöötluse oskused hõlmavad lisaks analüüsi meetoditele ka andmete säilitamise ja organiseerimise viiside ning analüüsi tehniliste vahendite tundmist.

1.1. Andmed ja valikumeetodid

Andmeanalüüsis kasutatavate muutujate väärtused võivad olla kas otseselt reaalsest maailmast pärit ehk **empiirilised** või üldistuse tulemusena saadud ehk **teoreetilised**. Empiirilised andmed võivad olla rohkem või vähem üldistatud. Kuna üldistamine lähtub alati teatud reeglitest, siis on üldistatud empiirilised andmed osaliselt ka teoreetilised.

Empiirilised andmed on saadud katsest, vaatlusest või muust kogemusest. Empiirilisteks võib lugeda näiteks püütud hiirte kehakaalu, toitainete lisamise järel tõusnud taimede biomassi ja vihmaste ilmade sageduse. Ka mälestused võiks lugeda empiiriliste andmete hulka, mida säilitatakse kellegi teadvuses. Ulmes esinevad objektid on küll pigem teooria kui empiiria, aga kujutelmad või ulme võivad olla ka empiirilise uurimise objektiks.

Teoreetilised andmed tuletatakse teatud teoreetilistest printsiipidest või mudelist ning kehtivad enamasti vaid kindlate eelduste korral. Näiteks sademete tõenäosus homme või liigi väljasuremise tõenäosus saja aasta jooksul on olemas vaid teoreetiliselt. Teoreetilised on ka prognoositud väärtused ja ennustused. Näiteks ilmaennustus on teoreetiline ilm, mille prognoosimiseks kasutatakse vaatlusandmeid.

1.1.1. Muutujate ja tunnuste tüübid

1.1.1.1. Arvuline ehk kvantitatiivne muutuja

Muutuja (*variable*) on abstraktne objekt, millele on antud nimi ja mille väärtused ei ole ette teada. Arvuline muutuja võib olla kas pidev või diskreetne, kuid mõlemal juhul on väärtused järjestatavad ja väärtuste sisulised erinevused on võrdelised arvude erinevusega. **Pidev muutuja** võib mingis muutumisvahemikus omada kõiki reaalarvulisi väärtusi.

Diskreetne muutuja saab omada vaid teatud väärtusi. Enamasti on diskreetseks muutujaks mingi loendatav tunnus ja diskreetse muutuja väärtuseks täisarv. Pidevad muutujad on näiteks pikkus, vanus, aeg ja asukoht. Diskreetsed muutujad on näiteks järglaste arv ja täringuviske tulemus. Loomade arv vaatlusalal on diskreetne ja täisarvuline muutuja, loomade keskmine tihedus pindalaühiku või uurimisala kohta on aga pidev muutuja.

Pideva ja diskreetse muutuja erinevus ei ole absoluutne. Pideva tunnuse mõõtmine toimub tavaliselt mingi mõõtmistäpsusega ja seetõttu vahepealseid väärtusi mõõtmistulemustes ei esine. Samuti saab pidevat tunnust diskreetseks klassifitseerida. Kui diskreetsel muutujal esineb palju erinevaid väärtusi, saab seda käsitleda pideva muutujana. Näiteks vanust aastates ja lõpetatud klasside või kursuste arvu võib vaadelda diskreetsete tunnustena, kuid üldjuhul on need pigem pideva tunnuste väiksema täisarvuni ümardatud väärtused. Väiksema täisarvu poole ümardatud tunnustest keskväärtuse arvutamise puhul on asjakohane liita tulemusele 0,5.

1.1.1.2. Mittearvuline ehk kvalitatiivne muutuja

Erinevalt kvantitatiivsetest muutujast ei ole kvalitatiivse muutuja väärtused arvude tähendusega, kuigi võivad olla arvudena kodeeritud. Järjestatavad võivad kvalitatiivse muutuja väärtused olla mingis konkreetse uurimuse jaoks olulises aspektis, kuid mitte universaalselt arvudena. Järjestatavat kvalitatatiivset muutujat nimetatakse **järjestusmuutujaks** (*ordinal variable*) või järjestustunnuseks. Järjestustunnused on näiteks mitmesugused hinnangud, sealhulgas ka hindeskaala A, B, C, D, E, F.

Kui kvalitatiivset muutujat nimetatakse **nominaalseks** (*nominal variable*) ehk nimeliseks, siis

soovitakse rõhutada, et tunnuse väärtused ei ole sisuliselt järjestatavad (näiteks rahvus, õpitav eriala). Formaalselt järjestatavad on ka nominaalsete tunnuste väärtused, kui mitte muul moel, siis kasutatud nimetuste või koodide tähestikjärjekorras. Sellise järjestuse puhul ei tulene vaadeldavate objektide järjekord sisulistest omadustest. Muutuja tüübi puhul eelistatakse terminit nominaalne, jaotuste ja mudelite puhul on tavaks keskmine n täht ära jätta (binomiaalne ja multinomiaalne), kuigi sõna tüve päritolu on sama – kreeka keeles *nomos*, ladina keeles *nomen*. Nominaalset muutujat nimetatakse ka **kategooriliseks muutujaks.**

Kui kvalitatiivset muutujat nimetatakse **binaarseks** ehk dihhotoomseks ehk kaheväärtuseliseks, siis juhitakse tähelepanu selle tunnuse vaid kahele võimalikule väärtusele (näiteks ei/jah vastusevariantidega küsimused, sugu). Binaarse muutuja variantide järjestamine ei oma tavaliselt sisulist tähendust ja seetõttu võib ka binaarseid tunnuseid nominaaltunnusteks nimetada.

Arvulise ja mittearvulise tunnuse erinevus ei ole absoluutne. Esiteks võib kvalitatiivne tunnus olla andmestikus arvuliselt kodeeritud, mis muidugi ei tähenda, et tunnus muutuks sisuliselt arvuliseks. Teiseks, tunnus võib olla põhimõtteliselt küll arvuliselt mõõdetav, kuid kuna mõõtmine on tülikas, on selle asemel kasutatud mittearvulisi hinnanguid ehk klassidesse jagamist. Näiteks juuste värvust, nagu igasugust värvitooni, saab arvuliselt mõõta tumeduse ja põhitoonide osakaalu kaudu, aga võib klassifitseerida ka klassidesse: blond, linalakk, punapea, punkar, brünett, hall, kiilas. Kodeerimine on tunnuse väärtuste asendamine väärtusklassi koodidega. Kuigi kodeerimisel kasutatakse reeglina positiivseid täisarve, kodeeritakse reeglina kvalitatiivseid tunnuseid.

Kui kvalitatiivne tunnus on järjestatav, kuid väärtuste erinevused ei ole selgelt määratletud ega omavahel võrreldavad, siis ei ole aritmeetilised tehted sellise tunnusega sisukad

1.1.1.3. Tunnuste funktsionaalne liigitus

Tunnustevaheliste seoste uurimisel kasutatakse üldjuhul üht tunnust **funktsioontunnusena** ehk tunnusena, mille väärtus sõltub teistest (*response*, *outcome*, *dependent variable*) ja teist või teisi **argumenttunnustena** ehk tunnustena, mis määravad funktsioontunnuse väärtuse (*predictor*, *independent variable*, *explanatory variable*). Terminid sõltuv ja sõltumatu tunnus ei ole õnnestunud, sest jätavad määramata, millest sõltuv või sõltumatu. Pealegi kui sõltumatu tunnus ei seostu (sõltu) argumenttunnustega, siis ei ole ka statistilisi seoseid.

Mitmetunnuselise analüüsi puhul on mitu argumenttunnust ja mitmemõõtmelise analüüsi puhul on funktsioontunnuseid mitu. **Funktsioon** on eeskiri, mille alusel arvutatakse funktsioontunnuse väärtused argumenttunnuste väärtuste järgi. Graafikutel kujutatakse funktsioontunnust enamasti püstteljel (y) ja argumenttunnust rõhtteljel (x). Olukorras, kus tunnuste vaheline põhjuslik seos on vastastikune või ei ole põhjuslikkuse suund teada, sobib seose uurimisel ja graafilisel kujutamisel argumenttunnuseks paremini see tunnus, mida on lihtsam määrata või mille mõõtmisviga on väiksem.

Kuna valim ei haara tervet üldkogumit ja tunnuste mõõtmine toimub mingi **mõõtmisveaga**, siis sisaldavad statistilised funktsioonid lisaks argumenttunnustele ka juhuslikku komponenti. Statistilised seosed ei ole ühesed funktsioonid, mille puhul igale argumendi väärtusele vastab kindel funktsiooni väärtus. **Statistiline funktsioon** omab teatud väärtusi teatud tõenäosusega ning alati on asjakohane kasutada saadud tulemust koos vea hinnanguga (ptk 1.4).

Statistilised funktsioonid ei võimalda funktsioontunnuse väärtusi absoluutse täpsusega määrata

1.1.2. Valikumeetodid

Statistika üks põhieesmärke on teha valimi põhjal järeldusi ja üldistusi suure ja raskestiuuritava üldkogumi kohta. Valimi uurimine on lihtsam, kiirem ja odavam kui üldkogumi uurimine. Mõnikord on üldkogumi uurimine lihtsalt ebareaalne. Kujutlegem näiteks ülesannet võrrelda liivaterade keskmist suurust Pirita ja Pärnu rannas. Et valimi põhjal tehtud järeldused üldkogumi kohta oleksid põhjendatud, peab valim olema esinduslik ehk representatiivne. **Esinduslik valim** on piisavalt suur ja valitud nii, et kajastaks üldkogumis olevaid seaduspärasusi tõeselt ja üldkogumi objektidel on võrdne võimalus valimisse sattuda.

Kui üldkogumis on teadaolev seesmine struktuur, siis peaks ka kõigil struktuuriüksustel olema võrdne võimalus olla valimis esindatud. Esinduslikkus on suhteline ning esinduslikkuse tõstmine üldkogumi sisestruktuuri mingis aspektis loob enamasti ebavõrdset esindatust mingis teises aspektis. Lisaks sellele tuleb uuringute kavandamisel arvestada projekti maksumust. Kokkuvõttes võib esindusliku valimi kavandamise lugeda optimeerimisülesannete hulka kuuluvaks.

Tõenäosusteoorias määratletakse katse kindla protseduuriga korratava toiminguna, millel on ette teadmata ehk juhuslik tulemus. Katsed, mille kordamisel on katsetingimused ühesugused, moodustavad katsete jada ehk **katseseeria**, mille tulemustest moodustuv valim kirjeldab üldkogumit, kuid üldjuhul mitte absoluutse täpsusega. Sõltuvalt katse tingimustest ja valimistoimingust eristatakse järgmisi valimeid.

- **Sõltuv valim**, milles samu tunnuseid on mõõdetud korduvalt samadel katsetingimustel ja vaatlused/katsed moodustavad sestap omavahel seotud paare või kõrgema astme kogumeid. Näiteks mõõtmistulemused enne ja pärast uurimisobjekti töötlemist.
- **Sõltumatus valimis** ei ole vaatlused/katsed valimite vahel seotud. Iga objekt on kaasatud vaid ühel korral või vähemalt ei moodusta objektid uurija poolt planeeritud või ette teada olevaid paare.
- **Tagasipanekuga valiku** puhul võib sama objekt korduvalt valimisse sattuda.
- **Tagasipanekuta valiku** puhul ei saa sama objekti korduvalt valida.

Esinduslik valim võimaldab teha üldkogumi kohta tõeseid järeldusi

1.1.2.1. Juhuslikud valikumeetodid

Juhuvalik on kõige lihtsam esindusliku valimi saamise meetod. Siiski tasub meeles pidada, et juhuslikustamine ei ole eesmärk, vaid üks võimalik viis esindusliku valimi moodustamiseks. Juhuslikud valikumeetodid on kõige objektiivsemad, kuna valimi koostamine on uurija suvast sõltumatu. Valikumeetodi valimisel on oluline, et valim peab olema küll representatiivne, aga seejuures ka mõistlike kulutuste ja ajakuluga saavutatav ehk piisavalt odav. Väliuuringute puhul on täiesti juhuslik valik sageli suhteliselt kallis meetod, sest võrreldes järgnevate meetoditega kipuvad kulutused ühe vaatlusobjekti kohta olema kõrgemad ning sama esinduslikkuse tagab üldjuhul mõnevõrra suurem uuritud objektide hulk.

Juhuvalik tagab sõltumatuse uurija suvast, kuid ei pruugi olla kõige odavam meetod esindusliku valimi ja usaldusväärsete järelduste saavutamiseks

Juhusliku valimi koostamiseks tuleb juhuslikkust kuidagi tekitada. Käepärane viis on valida juhuslike järjekorranumbritega või koordinaatidega üldkogumi objektide. Juhuarvude generaator on mehhanism või algoritm, millega on võimalik tekitada pideval skaalal näiliselt juhusliku jaotusega

väärtusi. Arvutitarkvarades olevad juhuslikke arve moodustavad funktsioonid kasutavad keerukaid algoritme, mille abil saab juhuslikuna näivaid ehk **pseudojuhuslikke arve**. Kui juhuarvude generaator kasutab lisaks ajahetke arvulist väärtust, siis saadakse iga kord erinev juhuarvude jada. Kui aga kasutatakse vaid ühte ja alati sama pseudojuhuslikkust loovat reeglit, saadakse igal samade eeltingimustega genereerimisel sama jada.

Juhuslikkust genereerib ka täringuvise ja kaarditõmbamine, aga need meetodid annavad vaid diskreetseid väärtusi. Tavaliselt tekitavad juhuarvude generaatorid ühtlase jaotusega arve, aga võimalik on moodustada ka etteantud jaotustüübiga juhuslike arvude kogumeid.

1.1.2.2. Planeeritud valikumeetodid

Kvootide meetodi kasutamisel määratakse valimi struktuur enne valimist. Näiteks otsustatakse enne valimist, mitu alla 100 aastast ja mitu üle 100 aastast leht-, sega- ja okaspuistut peab valim sisaldama. Kui kvootide täitmine ei toimu juhuslikult, siis ei ole valimi alusel tehtud otsustuste täpsust võimalik statistiliste meetoditega määrata.

Ekspertvalik on subjektiivne valimine, valitakse tüüpilisi objekte. Valim sõltub eksperdi kogemustest ja sellest, mida ekspert hetkel peab tüüpiliseks. Ekspertotsustuste paikapidavust on raske hinnata, kuid see on sageli kõige kiirem ja odavam viis saada suhteliselt täpne hinnang tunnuse kohta, mille otseselt mõõtmine ei ole jõukohane. Suhtelisus tähendab siinkohal võrdlust väiksema vilumusega valijatega.

1.1.2.3. Kombineeritud meetodid

Klastervalik (*cluster sampling*) on valikumeetod, mille puhul jagatakse üldkogum gruppideks, valitakse juhuslikud grupid ja analüüsitakse kõiki valitud grupi liikmeid.

Kihiline valik (*stratified sampling*) on valikumeetod, mille puhul üldkogum klassifitseeritakse enne valimi võtmist mingi olulise tunnuse või tunnuste järgi kihtidesse. Kihtide piires kasutatakse enamasti juhuslikku valikut. Kihiline valik on üks eelistatumaid valikumeetodeid, sest see tagab ühelt poolt kõigi uurimuse jaoks oluliste kihtide esindatuse valimis, säilitades samas kõigi elementaarobjektide võimaluse valimisse sattuda. Kihiline valik eeldab mingeid eelteadmisi uuritava üldkogumi kohta. Tunnus, mille alusel üldkogum kihtidesse jagatakse võib olla ka objektide paiknemine ruumis. See tähendab, et valik võib olla kihiline nii temaatiliselt kui ruumiliselt. Temaatiline kihilisus tähendab vaatluste etteantud vahekorda valimis mingite kategooriate osas, ruumiline kihilisus ruumiliste eraldiste või piirkondade osas.

1.1.2.4. Süstemaatiline valik

Süstemaatilise ehk regulaarse ehk korrapärase valiku puhul toimub vaatluste valimisse võtmine fikseeritud sammu tagant. Kui esimene objekt valitakse juhuslikult, on tagatud kõigi üldkogumi objektide võrdne tõenäosus valimisse sattuda ja meetodit võiks tinglikult lugeda juhuslikuks. Regulaarse valimivõtu peamiseks puuduseks on valiku lähtumine uurija määratud sammust, millest johtuvalt võib saada ebaesindusliku valimi, kui üldkogumis on valimi sammule vastav seesmine struktuur. Lisaks sageduste ja keskmiste hinnangu nihkele on tagajärjeks alahinnatud varieeruvus, mille tagajärjeks omakorda on tehtavate järelduste usaldusväärsuse ülehindamine. Enamasti ei ole üldkogumi sisemine struktuur eelnevalt teada ning ebasoodsal juhul, kui valimi samm juhtub sellega samas taktis olema, ei ole tsüklilisuse olemasolu võimalik valimi põhjal kontrollida. Seetõttu soovitatakse täielikult regulaarset valimivõttu vältida.

Süstemaatiline valik tagab ühtlase esindatuse, kuid kätkeb ohtu ülehinnata tulemuste usaldusväärsust

Kui üldkogumi seesmine struktuur on kõigi tunnuste osas juhuslik, annab süstemaatiline valik samaväärse tulemuse kui juhuslik valik. Õnnestunult valitud regulaarsusega valim võib üldkogumit igakülgsemalt esindada kui juhuslik valim, kuna viimases võivad üldkogumi olulised aspektid juhuslikult esindamata olla. See tähendab, et esinduslikkuse saavutamiseks võib regulaarselt võetud valim olla isegi väiksem kui juhuslikult valitud valim. Valik võib olla regulaarne nii ruumiliselt kui ka mingi tunnuse suhtes (temaatiliselt). Ruumilise korrapära puhul paiknevad vaatluskohad ühesuguste vahedega või mingil muul viisil korrapäraselt. **Temaatiline korrapära** tähendab korrapärast valikut mingi tunnuse või tunnuste järgi järjestatud vaatlustest. **Ruumiliselt korrapärane** valik tagab ruumiosade ühtlasema esindatuse ja ruumilise keskmise täpsema hinnangu.

1.1.2.5. Kõikne valik

Kõikse valiku puhul tehakse vaatlused kõigi üldkogumi objektide kohta. Üldkogum = valim = uurimisobjekt. Suurte üldkogumite kõikne uurimine on aeganõudev ja kallis. Kõikse valiku näide on rahvaloendus küsitlusankeedi selles osas, millele peavad kõik küsitlusalused vastama. Kõikne valik annab üldkogumi kohta täpsed andmed, mis tähendab, et järelduste usaldusväärsuse hindamiseks ei ole vaja kasutada tõenäosusteooriale tuginevaid statistilisi meetodeid. Otsuste langetamiseks ei ole absoluutne täpsus enamasti vajalik. Piisab ligikaudsest hinnangust või otsusest, mis kehtib piisavalt suure tõenäosusega/täpsusega.

1.1.2.6. Käepärane valik

Käepärase valiku puhul lähtutakse eelkõige töömahu piiramise vajadusest ja kaastakse valimisse eelkõige objektid, mida on lihtsam ja odavam uurida. Käepärane valik ei pruugi olla täiesti subjektiivne, enamasti tähendab käepärane valik kergemini ligipääsetavate või kättesaadavate objektide või uurimispiirkonna eelistamist või suuremat tööd nõudvate objektide ja piirkondade eemaldamist valimist. Valimi põhiosa võib seejuures olla siiski rangelt planeeritud või juhuslik. Looduse välivaatlused liigituvad käepäraseks valimiks näiteks siis, kui vaatlusintensiivsus on vaatleja kodu lähedal keskmisest suurem. Samuti juhul, kui elurikkuse mõõtmisel määratakse lihtsa vaevaga äratuntavad organismid liigi tasemeni, raskestimääratavaid arvestatakse aga kõrgema taksoni tasemel.

1.2. Tõenäosusteooria

Statistilises andmetöötluses analüüsitakse suurt hulka tehtud katsete või vaatluste tulemusi või muid andmeid. Tõenäosusteooria seevastu tegeleb toimumata katsete tulemuste ehk üldkogumist vaatlemata jäänud objektide või mitte teada olevate sündmuste väärtuste võimalike väärtuste hindamisega. Tõenäosusteooria põhimõisted on järgmised.

- **Katse** (*experiment*) on toiming, millega saadakse tulemus (sündmus) ühe objekti kohta üldkogumit moodustavate objektide hulgast. Näiteks: kaardi tõmbamine kaardipakist, täringuvise, mündi viskamine. Loodusuuringutes on selle termini tähendus mõnevõrra laiem.
- **Sündmus** (*event*) on katse tulemus üldises mõttes. Juhtum, mille tõenäosust saab arvutada sarnaste objektidega tehtud katseseeria tulemuste või muude eelteadmiste põhjal. Näiteks: kaardipakist saadi poti kuningas, kaardipakist saadakse must äss, poisi või tüdruku sünd.
- **Elementaarsündmus** (*elementary event*) on katse tulemuse detailseim võimalik väärtus eeldusel, et võimalik on kindel lõplik arv sündmuste väärtusi. Ühegi elementaarsündmuse toimumine mõne teise elementaarsündmuse suhtes ei ole eelistatud. Näiteks: kaardipakist saadi poti kuningas on elementaarsündmus aga kaardipakist saadakse must äss ei ole elementaarsündmus vaid sündmus, mis koosneb kahest elementaarsündmusest.
- **Tõenäosus** (*probability*) on soodsate elementaarsündmuste ja kõigi elementaarsündmuste suhe. Sündmuse A tõenäosus $P(A)$ on alati vahemikus $0 \leq P(A) \leq 1$ (ei saa olla negatiivne ega üle ühe). Tõenäosust saab hinnata, kui katse tingimused on fikseeritud ja on tarvis hinnata katse võimalike tulemuste oodatavaid sagedusi. Tõenäosus on iseenesest ühikuta suurus, mida saab tavatekstis väljendada protsentides, promillides või muudes suhtarvu ühikutes, kui see parandab teksti mõistetavust lugeja jaoks. Parsimooniareegli kohaselt tuleks siiski vältida mittevajaliku ühiku lisamist ühikuta suurusele.
- **Tõepära** (*likelihood*) kasutatakse kindla teadaoleva tulemuse suhtes – tulemus on fikseeritud, aga seda kujundanud tingimused ei ole kindlalt teda. Tõepära väljendab katse tingimuste tõenäosust. **Suurim tõepära** näitab kõige tõenäolisemat eeltingimuste kombinatsiooni, mis võis viia antud tulemuseni.
- **Sõltumatud sündmused** (*independent event*) on selline sündmuste paar, mille korral ühe sündmuse toimumine ei muuda teise sündmuse toimumise tõenäosust. Näiteks: kaardipakist saadi must kaart ja kaardipakist saadi äss on sõltumatud, sest mustade kaartide hulgast ässa saamine on sama tõenäone kui kogu kaardipakist ässa saamine. Samas kaardipakist saadi must kaart ja kaardipakist saadi ristiäss ei ole sõltumatud.
- **Kindel sündmus** (*definite event, certain event*) on sündmus, mis sisaldab kõiki elementaarsündmusi. Kindla sündmuse tõenäosus on 1. Näiteks vähem kui 7 silma saamine täringuviskel tavalise täringuga või ühe silma saamine täringuga, mille kõikidel tahkudel on üks silm.
- **Võimatu sündmus** (*impossible event*) on sündmus, mis ei sisalda ühtegi elementaarsündmust. Võimatu sündmuse tõenäosus on 0. Näiteks üheksa silma saamine täringuviskel tavalise täringuga.
- **Sündmuse järeldusseos** (*inferencial relationship*) tähendab, et kui sündmuse A toimudes toimub kindlasti ka sündmus B, siis sündmus B järeldub sündmusest A. Järeldusseos ei ole reeglina vastastikune. Vastastikuse järeldusseose puhul on sündmused identsed. Kui sündmus B järeldub sündmusest A, siis ei ole sündmuse A tõenäosus suurem kui sündmuse B tõenäosus. Näiteks täringuviske tulemus paaritu arv silmi järeldub tulemusest viis silma.

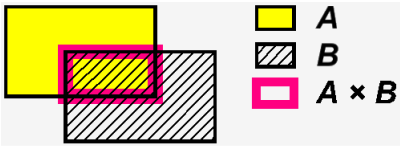
- Sündmuse *A* **vastandsündmus** (*complementary event*) on sündmus, mis toimub siis (ja ainult siis) kui sündmus *A* ei toimu. Sündmuse ja tema vastandsündmuse tõenäosuste summa võrdub ühega. Kindel sündmus ja võimatu sündmus on vastandsündmused. Näiteks paaritu arvu ja paaris arvu silmade saamine täringuviskel on vastandsündmused.
- **Teineteist välistavad sündmused** (*mutually exclusive events*) on sündmused, mis ei saa sama katse tulemusena üheaegselt esineda. Näiteks punase masti (ruutu või ärtu) kaardi ja potiässa saamine samast kaardivalikust eeldusel, et võetakse vaid üks kaart.
- **Tinglik tõenäosus** (*conditional probability*) antud sündmuse tõenäosus mingi teise sündmuse toimumise korral.

Tõenäosuste teadmine võimaldab ennustada sündmuse toimumist, mittetoimumist või oodatavat esinemissagedust. Üldiseid eeskirju sündmuse tõenäosuse määramiseks ei ole. Sageli kasutatakse tõenäosuse hinnanguna sündmuse **suhtelist sagedust**. Seda võimaldab **suurte arvude seadus**, mille kohaselt sündmuse suhteline sagedus läheneb pika katseseeria puhul sündmuse tõenäosusele. Sellisel juhul kasutatakse terminit **statistiline tõenäosus**, mis on juhuslik suurus, mille väärtus sõltub konkreetsest katseseeriast.

1.2.1. Tehted tõenäosustega

Sündmuste korrutis – sündmus, mis toimub siis, kui toimuvad mõlemad sündmustest *A* ja *B* ehk sündmus *AB*. Näiteks ärtu piltkaardi saamine 52-lehelisest kaardipakist on ärtukaardi saamise tõenäosuse (1/4) ja pildiga kaardi saamise tõenäosuse (12/52) korrutis (ässasid ei ole siin piltideks loetud).

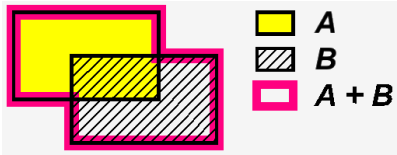
Sündmuste korrutis on elementaarsündmuste hulkade ühisosa (joonis 1-1), mida tähistatakse: $A \cap B$, seega sõltumatuse korral: $P(A \cap B) = P(A) \cdot P(B)$.



Joonis 1-1. Tõenäosuste korrutise graafiline vaste.

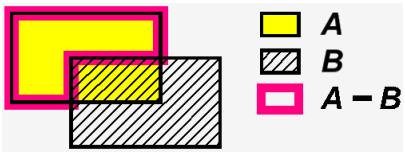
Sündmuste summa – sündmus, mis toimub siis, kui toimub vähemalt üks sündmustest *A* ja *B* (joonis 1-2). Näiteks täringuviskel alla kolme silma saamise tõenäosus võrdub ühe silma saamise ja kahe silma saamise tõenäosuste summaga, mis on elementaarsündmuste hulkade ühend.

Üldjuhul $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Teineteist välistavate sündmuste puhul võrdub sündmuste summa tõenäosus liidetavate sündmuste tõenäosuste summaga: $P(A \cup B) = P(A) + P(B)$.



Joonis 1-2. Tõenäosuste liitmise graafiline vaste.

Sündmuste vahe – sündmus, mis toimub siis, kui sündmus *A* toimub, aga sündmus *B* ei toimu (joonis 1-3). Näiteks täringuviskel kolme silma saamise tõenäosus võrdub alla nelja silma saamise tõenäosuse ja üle kahe silma saamise tõenäosuse vahega. Tähistatakse: $A \setminus B$. $A \setminus B = P(A) - P(A \cap B)$. Kui *A* ja *B* on teineteist välistavad, siis $A \setminus B = A$ ja $B \setminus A = B$, kui *B* järeldub *A*-st, siis $P(B \setminus A) = P(B) - P(A)$, aga $A \setminus B$ on võimatu sündmus.



Joonis 1-3. Tõenäosuste vahe graafiline vaste.

Tinglik tõenäosus – sündmuse *A* tõenäosus eeldusel, et sündmus *B* toimub. Tähistatakse: $P(A|B) = P(A \cap B)/P(B)$. Kui sündmus *A* järeldub sündmusest *B* $[B \Rightarrow A]$, siis $P(A|B) = 1$. Kui *A* ja *B* on teineteist välistavad, siis $P(A|B) = P(B|A) = 0$. Tingliku tõenäosuse alternatiiv on **tingimatu tõenäosus**. Tõenäosusteooria haru, mis tegeleb tinglike tõenäosusi sisaldavate teooriate ja arvutustega nimetatakse **Bayesi tõenäosusteooriaks**. See on klassikalise tõenäosusteooria edasiarendus juhtudele, kus lisaks katsega saadud uuele teabele saab kasutada ka enne katset teada olevaid tõenäosusi. Kui tõenäosusteooria klassikaline haru tegeleb eelkõige sündmuste oodatavate sagedustega, siis Bayesi tõenäosusteooria kvantifitseerib otsuse ebakindluse sõltuvust tõenditest.

Bayesi tõenäosusteooria põhisisu väljendatakse Bayesi teoreemina ehk Bayesi lausena.

Bayesi lause – sündmuse *H* variandi H_i toimumise tõenäosust tingimusel, et on toimunud sündmus *A*, see on $P(H_i|A)$, saab arvutada enne katset teada olevate (aprioorsete) tõenäosuste: sündmuse H_i üldine ehk sündmuse *A* eelne tõenäosus $P(H_i)$, sündmuse *A* üldine tõenäosus $P(A)$ ja sündmuse *A* tõenäosus sündmuse H_i toimumise korral $P(A|H_i)$ järgi vastavalt Bayesi valemile

$$P(H_i \mid A) = \frac{P(A \mid H_i) \cdot P(H_i)}{\sum_i P(H_i) \cdot P(A \mid H_i)} = \frac{P(A \mid H_i) \cdot P(H_i)}{P(A)}. \qquad [1\text{-}1]$$

Näiteks olgu olukord, kus tihasepesa rüüstamise tõenäosus rähni poolt on üldjuhul 20%, rähniõõntes on 15% pesadest ja rüüstatud pesadest 50% on rähniõõntes. Seega tihasepesa rüüstamise tõenäosus juhul, kui meil on täiendavalt teada, et pesa asub vanas rähni rajatud puuõõnes (aposterioorne tõenäosus), on 50 × 20 / 15 = 67%.

1.2.2. Dempster-Shaferi teooria

Dempster-Shaferi teooria lähtub eeldusest, et sündmuse tõenäosus ja vastandsündmuse tõenäosus ei pruugi kokku olla 100%. Lisaks nendele võib katse võimalike tulemuste hulgas olla teatud hulk määramatust või tundmatuid tulemusevariante. Peale selle võivad andmed viidata teatud sündmusekombinatsioonide tõenäosusele. Näiteks hele toon must-valgel kaugseire fotol võib viidata nii viljapõllu kui roostiku esinemisele. Dempster-Shaferi teooria põhimõisted on järgmised.

- **Omistatud põhitõenäosus** (*basic probability assignement*) väljendab ühte sündmusevarianti või ühte konkreetset sündmustekombinatsiooni toetavate tõendite massi. Tähistatakse *m(A)*. Mingi hüpoteesi *A* põhitõenäosus *m(A)* võib olla määratud ekspertotsusega või empiirilistest andmetest.
- **Teadmatus** (*ignorance*) väljendab suutmatust (määramatust) katsetulemuste tõenäosuste üle otsustada.
- **Uskumus** (*belief*) koosneb variandi põhitõenäosusest ja variandi komponentide tõenäosustest. Üksikvariandi puhul võrdub kogutõenäosus põhitõenäosusega. Kogutõenäosus väljendab kõigi hüpoteesi toetavate tõendite kogumassi.
- **Uskumatus** (*disbelief*) ei ole kogutõenäosuse täiend, vaid kõigi nende hüpoteeside

tõenäosuste, mis mingil määral ei sisalda antud variandi kogutõenäosust, summa. Võib tõlgendada kui katsetulemuse ebausutavust.

- **Usutavus** (*plausibility*) väljendab tõenäosust, millisel määral võiks hüpotees parimal juhul kehtida või katsetulemust oodata. Uskumus väljendab katsetulemuse võimaliku tulemuse alumist piiri ja usutavus ülemist piiri, nende vahele jääb **uskumisvahemik** (*belief interval*).

1.3. Jaotused

1.3.1. Juhusliku muutuja jaotus

Juhusliku muutuja jaotus on eeskiri, mis seab juhusliku muutuja iga väärtusega vastavusse selle väärtuse tõenäosuse. Kumulatiivse jaotuse esitust nimetatakse jaotusfunktsiooniks. Mittekumulatiivset jaotust nimetatakse diskreetse muutuja puhul tõenäosusjaotuseks (joonis 1-4). Kuna pideva muutuja iga üksiku väärtuse esinemise tõenäosus läheneb nullile, siis kasutatakse tõenäosusfunktsiooni asemel tihedusfunktsiooni (joonis 1-5).

Pideva tunnuse puhul on iga konkreetse väärtuse tõenäosus null

Juhusliku suuruse **jaotusfunktsioon** kohal a on tõenäosus, et juhusliku suuruse X väärtus on väiksem või võrdne kui a. Tähistatakse $F_X(a) = P(X \leq a)$ ja diskreetse muutuja puhul võib kirjutada ka kujul

$$F_x(a) = \sum_{X_i \leq a} p_i, \qquad [1\text{-}2]$$

kus $p_i = P(X = x_i)$.

Jaotusfunktsiooni väljendab seega kumulatiivset tõenäosust. Pideva juhusliku muutuja iga üksikväärtuse esinemise tõenäosus on null ja seetõttu iseloomustab tihedusfunktsioon muutuja väärtuste teatud väärtusvahemikku kuulumise tõenäosust. Pideva tunnuse jaotuse tihedusfunktsioon on jaotusfunktsiooni tuletis. Graafiliselt on tuletis võrdeline graafiku tõusunurga tangensiga. Seega näitab tihedusfunktsioon tõenäosusfunktsiooni juurdekasvu kiirust antud kohas. Jaotusfunktsiooni omadused on järgmised:

- jaotusfunktsioon on mittekahanev;
- jaotusfunktsiooni piirväärtused on 0 ja 1;
- jaotusfunktsioon on pidev.

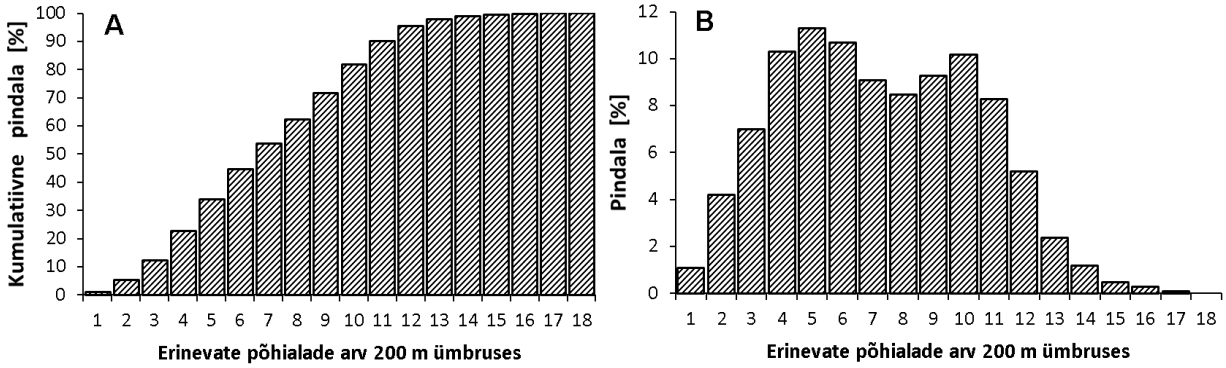
Tõenäosusfunktsioon – eeskiri, mis seab iga juhusliku suuruse väärtuse vastavusse selle tõenäosusega. On kasutatav diskreetsete tunnuste puhul. Tõenäosusfunktsiooni abil on võimalik arvutada iga üksiksündmuse tõenäosust. Seejuures kõigi üksiktõenäosuste summa võrdub ühega.

Tihedusfunktsioon – juhusliku suuruse tõenäosuse tihedus, mis avaldub jaotusfunktsiooni tuletisena. Pideva tunnuse jaotuse jaotusfunktsioon on selle tunnuse väärtuste tihedusfunktsiooni integraal. Graafiliselt vastab integraalile graafikualuse ala pindala integreeritavas vahemikus, mis tähendab, et sündmuse tõenäosus kuuluda etteantud väärtusvahemikku on võrdne väärtusvahemikku jääva tihedusfunktsiooni aluse pindalaga (kogu piirväärtuste vaheline pindala = 1). Tihedusfunktsioon iseloomustab muutuja X tõenäosust kuuluda vahemikku $[x, x + dx]$.

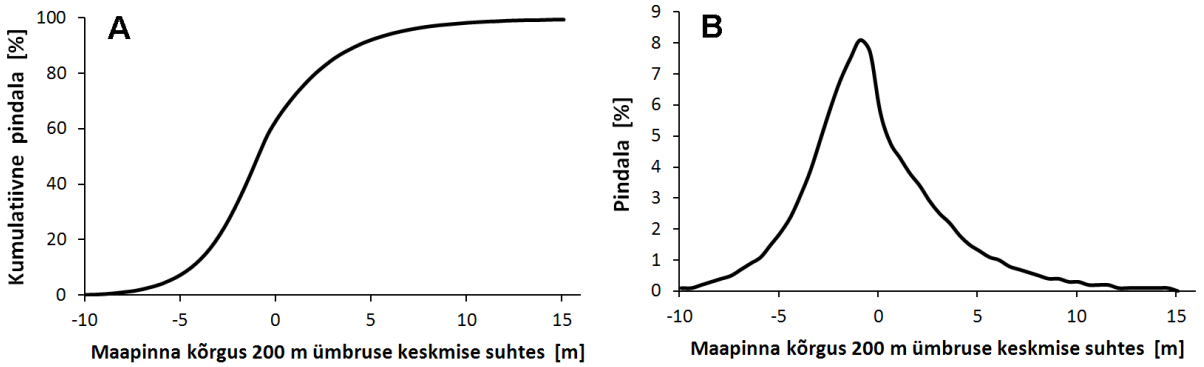
$$f(x) = \frac{d}{dx} F(x) \qquad [1\text{-}3]$$

Tihedusfunktsiooni omadused on järgmised:

- tihedusfunktsioon on alati mittenegatiivne;
- tihedusfunktsiooni integraal üle juhusliku suuruse väärtuste piirkonna ehk kogu graafiku alune pindala võrdub ühega;
- tõenäosus, et juhusliku suuruse väärtus satub antud lõiku, võrdub tihedusfunktsiooni integraaliga üle selle piirkonna;
- tihedusfunktsioon on defineeritud vaid pidevate jaotuste korral.



Joonis 1-4. Diskreetse järjestatava muutuja (Eesti põhikaardi erinevate põhialaklasside arv 200 meetri raadiuses kaardilehel 5434 arvutatuna 10 m vahega võrgustiku punktide ümber) kumulatiivne jaotus (A) ja sagedusjaotus (B). Kaardileht 5434 paikneb Otepää kõrgustikul, kus maastik on mitmekesine – kumulatiivne jaotus näitab, et umbes poolel kaardilehe pinnal on 200 m ümbruses üle kuue erineva põhiala. Kõige sagedamini on 200 m ümbruses põhikaardil esindatud viis erinevat põhiala. Eesti põhikaardi põhialade loetelu on lisas 2.



Joonis 1-5. Pideva muutuja (maapinna suhteline kõrgus 200 meetri raadiuses oleva ümbruse keskmise kõrguse suhtes Eesti põhikaardi lehel 5434) kumulatiivne jaotus (A) ja tihedusjaotus (B). Horisontaaltelje nullväärtusele vastavad kohad, kus maapinna kõrgus on võrdne maapinna keskmise kõrgusega ümbruses. Pane tähele, et Otepää kõrgustiku keskosas oleval kaardilehel 5434 on kõige suurema kogupindalaga ümbruse keskmisest veidi madalamad kohad ehk negatiivsed pinnavormid (küngaste jalamid ja soostunud nõod).

1.3.2. Jaotusparameetrid

Konkreetsest katseseeriast leitud statistiliste tõenäosuste jaotust nimetatakse juhusliku suuruse **empiiriliseks jaotuseks**. Empiiriline jaotus sõltub konkreetsest katseseeriast ja selle täpne tüüp on väikeste valimite puhul raskesti kirjeldatav. Katseseeria pikenemisel läheneb juhusliku suuruse empiiriline jaotus selle juhusliku suuruse **teoreetilisele jaotusele**. Kui teoreetiline jaotus ei ole teada ega arvutatav, siis kasutatakse selle hinnanguna empiirilist jaotust ning jaotusparameetrite väärtused hinnatakse empiiriliste andmete põhjal.

Selles alapeatükis on juhusliku suuruse jaotuse põhilised karakteristikud defineeritud tõenäosuste kaudu, mis on kasutatavad vaid teoreetiliste jaotuste ja kõikse analüüsi puhul. Jaotuskarakteristikute arvutamist valimi andmete põhjal vaadeldakse peatükis 1.4. Teoreetiliste jaotuste iseloomustamiseks piisab jaotusfunktsioonist ja selle parameetritest.

Keskväärtus (*mean value*) – juhusliku suuruse juhusest sõltumatu arvkarakteristik, mis väljendab tõenäosuse masskeset ja mis diskreetse juhusliku suuruse korral võrdub juhusliku suuruse väärtuste ja iga väärtuse tõenäosuse korrutiste summaga. Seda nimetatakse ka üldkogumist tehtud juhusliku valimi väärtuste **statistiliseks ootuseks** (*expectation*) ehk juhusliku muutuja esimest järku momendiks. Tähistatakse *EX* või μ. Diskreetse muutuja keskväärtus arvutatakse väärtuste esinemistõenäosuste p_i kaudu:

$$\mu = \sum_i p_i x_i , \qquad [1\text{-}4]$$

kus x_i on muutuja väärtus väärtuste järjekorraindeksiga i ning p_i on väärtuse x_i esinemise tõenäosus.

Juhusliku suuruse korrutamisel konstandiga muutub ka keskväärtus sama konstandi kordsena. Mõõtühikute muutmisel muutuvad samavõrd nii algandmete numbrilised väärtused kui ka keskväärtus. Juhuslike suuruste summa keskväärtus võrdub nende keskväärtuste summaga. Diskreetse täisarvulise muutuja keskväärtus ei pruugi olla täisarv.

Dispersioon (*variance*) ehk keskmine ruuthälve on juhusliku suuruse hajuvust iseloomustav arvkarakteristik, mis võrdub selle suuruse ruuthälvete keskväärtusega. Seega on juhusliku suuruse dispersiooni mõõtühikuks selle juhusliku suuruse ruutühik. Üldkogumi dispersiooni tähistatakse *DX* või σ^2, valimist arvutatuna s^2.

$$\sigma^2 = \sum_i p_i \left(x_i - \mu\right)^2 \qquad [1\text{-}5]$$

Dispersioon on seda suurem, mida suuremad on juhusliku suuruse hälbed keskväärtuse ümber ja mida sagedamad on suured hälbed. Dispersioon ei ole kunagi negatiivne. Konstandi liitmine juhuslikule suurusele ei muuda selle dispersiooni. Juhusliku suuruse korrutamisel konstandiga suureneb selle dispersioon võrdeliselt konstandi ruuduga. Varieeruvus ja hajuvus on tavakeelsed üldterminid, mis ei täpsusta varieeruvuse mõõtmise viisi.

Standardhälve (*standard deviation*) – ruutjuur dispersioonist ehk ruutkeskmine hälve. Tähistatakse *S(X)* või σ, valimist arvutatuna *s*. On mugav kasutada, kuna mõõtühik on sama, mis kirjeldataval juhuslikul muutujalgi. Standardhälve on alati positiivne. Konstandi standardhälve on null. Standardhälve on invariantne nihke suhtes, see tähendab, et kõigile muutuja väärtustele sama konstandi liitmine või lahutamine ei muuda muutuja standardhälvet.

Juhusliku suuruse standardiseerimisel lahutatakse kõigist väärtustest keskväärtus ning jagatakse väärtused standardhälbega, mis on standardiseeritud väärtuste mõõtühikuks

Parema võrreldavuse tagamiseks muutujaid ja jaotusi tsentreeritakse, normeeritakse ja standardiseeritakse **Tsentreerimine** on juhusliku muutuja lineaarteisendus, kus selle väärtustest lahutatakse muutuja keskväärtus. Tsentreeritud juhusliku muutuja keskväärtus on null. Tsentreerimine ei muuda juhusliku muutuja hajuvust, vaid muudab juhusliku suuruse väärtusi ja nende paiknemist arvsirgel. **Normeerimine** on juhusliku muutuja lineaarteisendus, mille puhul juhusliku muutuja väärtused jagatakse muutuja jaotuse standardhälbega. Normeeritud juhusliku muutuja dispersioon ja standardhälve võrduvad ühega. Normeerimine muudab juhusliku muutuja skaalat. Normeerimiseks

nimetatakse ka juhusliku muutuja väärtuste läbijagamist skaalasid ühtlustavate arvudega (näiteks keskväärtusega), mille abil saab erinevad juhuslikud suurused paremini võrreldavaks muuta. Sisuliselt tähendab selline teisendus mõõtühikute muutmist. **Standardiseerimine** on normeerimine ja tsentreerimine kokku. Standardiseeritud muutuja keskväärtus võrdub nulliga ja standardhälve ühega.

1.3.3. Parameetrilised jaotused

Parameetrilised jaotused on täielikult kirjeldatavad jaotusfunktsiooni koostisesse kuuluvate konstantide – **jaotusparameetrite** abil. **Mitteparameetrilist jaotust** võib ette kujutada empiirilise jaotuse näitel – seda ei ole võimalik jaotusparameetrite abil mõistlikult kirjeldada (vajalik jaotusparameetrite arv on ühe võrra väiksem kui esinevate väärtuste arv).

1.3.3.1. Ühtlane jaotus

Ühtlase jaotuse tõenäosus on konstantne ja jaotusfunktsioon lineaarne. Ühtlane jaotus esineb lõigul (*a*,*b*). Diskreetsel ühtlasel jaotusel on lõplik arv (*k*) võimalikke väärtusi ja kõik väärtused on võrdvõimalikud, näiteks täringuviske ja mündiviske tulemused.

Ühtlase jaotuse keskväärtus

$$\mu = \frac{a+b}{2}. \qquad [1\text{-}6]$$

Ühtlase jaotuse dispersioon

$$\sigma^2 = \frac{(b-a)^2}{12}. \qquad [1\text{-}7]$$

Diskreetse ühtlase jaotuse puhul

$$P(X=i) = \frac{1}{k}, \qquad [1\text{-}8]$$

$$\mu = \frac{k+1}{2} \qquad [1\text{-}9]$$

(eeldusel, et miinimumväärtus = 1),

$$\sigma^2 = \frac{(k^2-1)}{12}. \qquad [1\text{-}10]$$

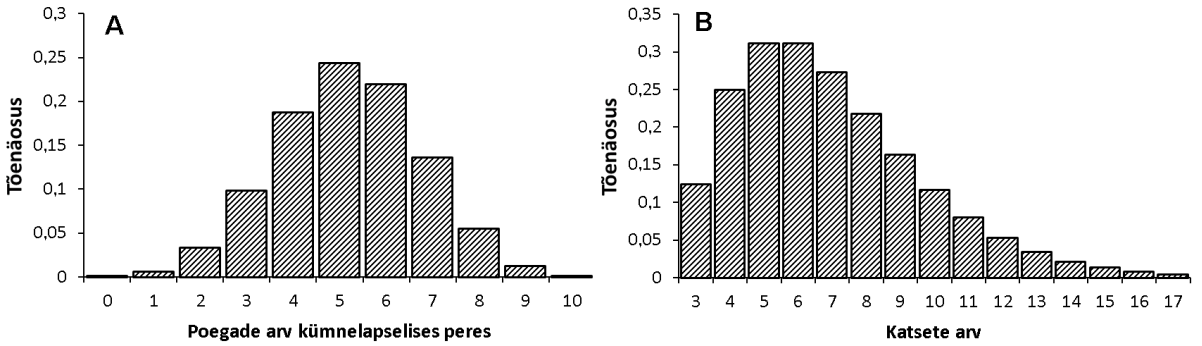
1.3.3.2. Bernoulli jaotus

Bernoulli jaotus on sündmuse toimumise ja mittetoimumise tõenäosuse kaheväärtuseline jaotus. Kui sündmus toimub, siis sündmuse indikaator on *X = 1*, kui ei toimu, siis *X = 0*. Näiteks poja või tütre sündimise tõenäosus. Bernoulli jaotuse parameeter on sündmuse *X* toimumise tõenäosus *P(X)*, mida tähistatakse *p*. Sündmuse mittetoimumise tõenäosus on *1 – p*. Bernoulli jaotus on binoomjaotuse (vt järgnev) erijuht tingimusel, et katsete arv on 1. Bernoulli jaotuse dispersioon avaldub kujul

$$DX = p(1-p). \qquad [1\text{-}11]$$

1.3.3.3. Binoomjaotus

Binoomjaotus (nimetatakse ka binomiaaljaotuseks) on sündmuse *A* esinemiskordade tõenäose jaotus *n* sõltumatust katsest koosnevas katseseerias (joonis 1-6).



Joonis 1-6. Poegade arvu tõenäosuse jaotus kümnelapselises peres juhul, kui poja sündimise tõenäosus on 0,52 (A). Märka asümmeetriat. Kolme eduka katsetulemuse tõenäosuse sõltuvus katseseeria pikkusest, kui ühe katse puhul on õnnestumise tõenäosus 0,5 (B). Kolm edukat tulemust on kõige oodatavamad viie ja kuue katse puhul, suurema arvu katsete puhul suureneb võimalus saada üle kolme eduka tulemuse.

Binoomjaotus eeldab, et katseseeria pikkus on lõplik loendatav suurus ja katsed on üksteistest sõltumatud. Binoomjaotuse parameetriteks on p (sündmuse X tõenäosus) ja n (katseseeria pikkus). Tähistatakse $B(n, p)$. Binoomjaotusega on näiteks pikas pisiimetajate püükide seerias tühjaks jäänud püüniste osakaal kui iga kord on üles seatud kindel hulk püüniseid. Samuti on binoomjaotusega kahealleelse geeni alleelikombinatsioonide esinemise sagedus populatsioonis. Katseseeria pikkuseks on diploidsete organismide puhul 2, tõenäosuseks on vaatlusaluse alleeli sagedus. Binoomjaotusega juhusliku suuruse väärtuste esinemise tõenäosus avaldub valemiga

$$P(X = k) = n!/(k!(n-k)!) \cdot p^k \cdot (1-p)^{n-k}. \qquad [1\text{-}12]$$

Kui katseseeria pikkuseks on kaks, siis esinevad tulemuste variandid sagedusega: p^2, $2pq$, q^2, kus $q = 1 - p$.

Binoomjaotusega on edukate katsete arvu jaotus kindla pikkusega katseseerias, kui üksikkatsed on juhuslikud ja omavahel sõltumatud

Binoomjaotusega suuruse keskväärtus ja dispersioon avalduvad järgnevalt:

$$\mu = np, \qquad [1\text{-}13]$$

$$\sigma^2 = np(1-p). \qquad [1\text{-}14]$$

Suure katsete arvu ja mõlema alternatiivse sündmuse võrdse tõenäosuse korral läheneb binoomjaotus normaaljaotusele

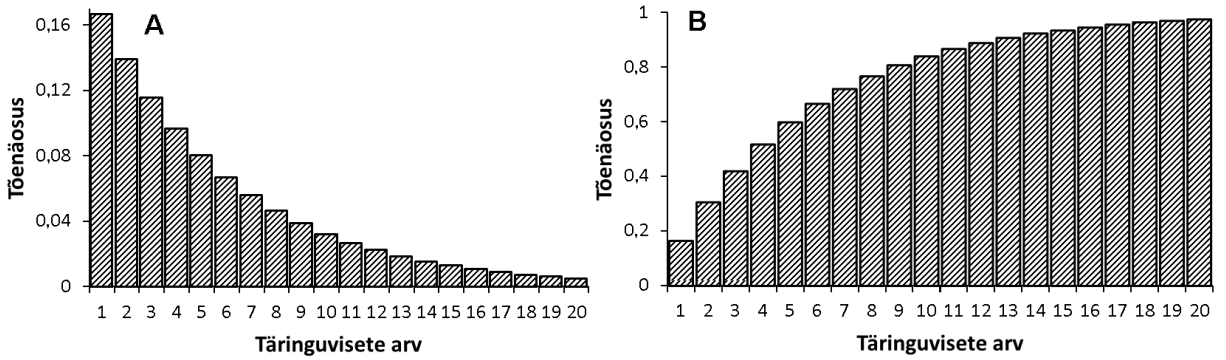
Suure katsete arvu ja mõlema alternatiivse sündmuse võrdse tõenäosuse korral läheneb binoomjaotus normaaljaotusele, tänu millele on kasutatud rusikareeglit, et kui katseseeria pikkuse ja harvemini esineva sündmuse tõenäosuse korrutis on suurem kui 5, siis võib binoomjaotuse asemel kasutada normaaljaotust ja normaaljaotust eeldavaid meetodeid. Väikese soodsa sündmuse tõenäosuse ja pika katseseeria puhul saab binoomjaotuse asemel kasutada ka Poissoni jaotust (ptk 1.3.3.5).

1.3.3.4. Geomeetriline jaotus

Geomeetrilise jaotusega juhusliku suuruse väärtuseks on selle katse number, mille tulemusena oodatud sündmus esmakordselt esines (joonis 1-7). Jaotus eeldab katsete sõltumatust ja avaldub kujul

$$P(X = k) = p \cdot (1 - p)^{k-1}, \qquad [1\text{-}15]$$

kus *k* on 1, 2, ... katse järjekorranumber ja *p* on sündmuse tõenäosus.



Joonis 1-7. Kuue täringusilmaga alustatava lauamängu alustamise tõenäosus viskekordadel (tõenäosusfunktsioon) (A). Iga järgnev viskekord saab võimalikuks vaid juhul, kui eelnevatel kordadel ei õnnestunud mängu alustada. Tõenäosus, et mängu alustamiseks piisab näidatud täringuvisete arvust (kumulatiivne tõenäosus) (B). Esimeses viskevoorus saab mängu alustada tõenäosusega 0,167, aga on tõenäosus 0,026, et kuue silma saamiseks tuleb täringut visata üle 20 korra.

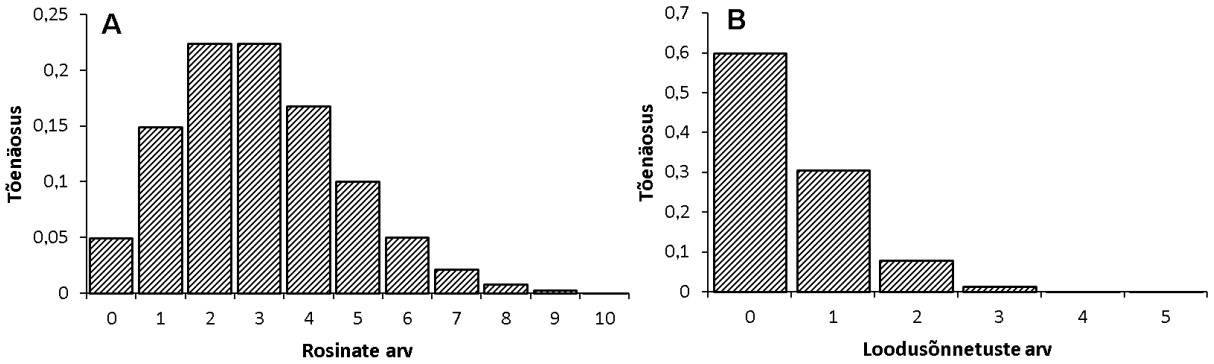
1.3.3.5. Poissoni jaotus

Poissoni jaotus on binoomjaotuse erijuht, kui pikendada katseseeriat piiramatult ja vähendada nähtuse esinemise tõenäosust üksikkatsel. Tähistatakse *P(λ)*. Poissoni jaotust kasutatakse väikesearvuliselt või harva esinevate täisarvuliste ehk loenduvate muutujate modelleerimiseks. Poissoni jaotuse tihedusfunktsiooni väljendatakse valemiga

$$P(X = k) = e^{-\lambda} \lambda^k / k!, k = 0,1,..., \qquad [1\text{-}16]$$

kus *k* on otsitav kordade arv, λ on keskmine kordade arv ja *e* on naturaallogaritmide alus ≈ 2,718.

Poissoni jaotuse parameetriks on keskväärtus λ, mis on võrdne jaotuse dispersiooniga. Parameeter λ võib olla mistahes positiivne arv. Poissoni jaotus sobib juhusliku harvaesineva sündmuse sageduse modelleerimiseks, nagu mutatsioonide sagedus ajaühikus (eeldades, et mutatsioonide keskmine sagedus ei muutu), järglaste arv pesakonnas (eeldades, et keskmine järglaste arv on teada ja järglase saamine või mittesaamine on juhuse asi). Poissoni jaotuse kuju sõltub parameetri λ väärtusest (joonis 1-8).

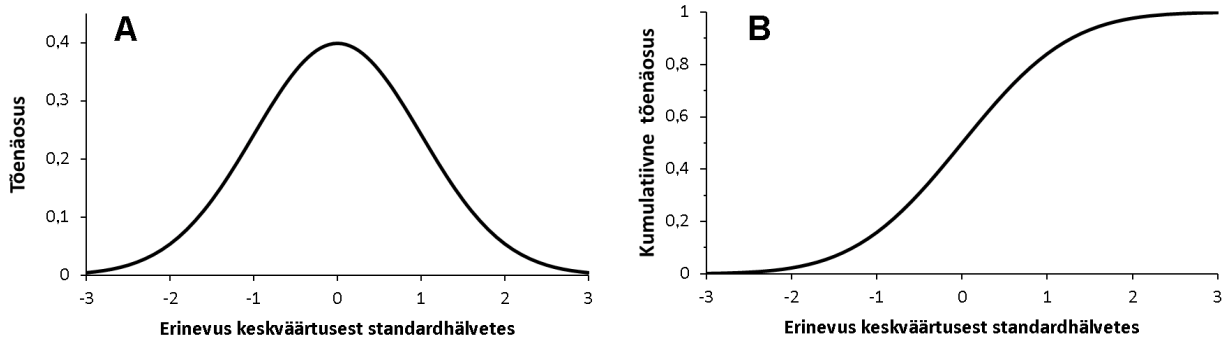


Joonis 1-8. Ühes rosinakuklis olevate rosinate arvu tõenäosused, kui keskmiselt on ühes kuklis kolm rosinat (λ = 3) ja rosinate paiknemine on juhuslik (A). Pane tähele, et oodatavalt umbes 5% rosinakuklitest on juhuslikult ilma ühegi rosinata. Loodusõnnetuse kordade arvu tõenäosus juhul, kui ühe õnnetuse tõenäosus sel perioodil on 0,4 (λ = 0,511) (B). Nende tingimuste juures on siiski tõenäosus 0,0133, et õnnetus kordub sama perioodi jooksul koguni kolm korda ning tõenäosus 0,0152, et kolm või enam korda. Eeldatud on, et õnnetus on muutumatu tõenäosusega juhuslik sündmus.

Vastavalt **Poissoni piirteoreemile** läheneb binoomjaotusega juhuslike suuruste jada katsete arvu kasvamisel Poissoni jaotusega juhuslikule suurusele. Binoomjaotus käsitleb sündmuse tõenäosust kindla pikkusega ja mitte väga pikas katseseerias, Poissoni jaotus sündmuse tõenäosust eeldusel, et katseseeria on lõpmata pikk või vähemalt küllaltki pikk. Binoomjaotusega on näiteks isaste järglaste oodatav arv kindla suurusega kurnas (nt metskurvitsa täiskurnas alati 4 muna). Katseks on siin tibu koorumine (õigemini viljastamine), katseseeria pikkuseks on kurna suurus. Poissoni jaotusega on juhuslikult paiknevate linnupesade arv kindla suurusega uurimisaladel.

1.3.3.6. Normaaljaotus

Normaaljaotus on pidev sümmeetriline jaotus, mille väärtuste hulgaks on kõigi reaalarvude hulk (joonis 1-9). Normaaljaotus on määratletud keskväärtuse (μ) ja standardhälbe kaudu (σ). Standardhälbe asemel võib kasutada dispersiooni (σ^2), sest need on omavahel lihtsasti teisendatavad. Tähistatakse *N(μ, σ)*. Normaaljaotuse keskväärtus võib olla mistahes arv, standardhälve ei saa olla negatiivne arv. Standardiseeritud normaaljaotuse $\mu = 0$ ja $\sigma = 1$.



Joonis 1-9. Standardiseeritud normaaljaotuse (keskmine = 0, standardhälve = 1) tihedusfunktsioon (A) ja jaotusfunktsioon (B). Standardiseeritud jaotuse puhul on mõõdetud suuruse keskväärtus null ja väärtuste mõõtühikuks on standardhälve.

Normaaljaotuse tihedusfunktsiooni graafik on sümmeetriline, punktid $\mu + \sigma$ ja $\mu - \sigma$ on selle kõvera käänupunktideks, kus kumerus läheb üle nõgususeks. Normaaljaotuse tihedusfunktsiooni graafikut nimetatakse **Gaussi kõveraks**. Normaaljaotuse tihedusfunktsiooni esitatakse valemiga

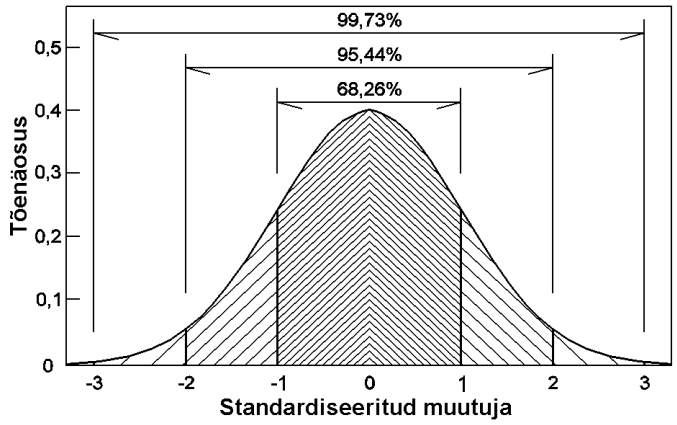
$$Z(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma}}. \qquad [1\text{-}17]$$

Muutuja Z määrab Gaussi kõvera kõrguse, mis näitab väärtuste esinemise suhtelist sagedust. Parameeter μ on sümmeetrilise tihedusjaotuse keskväärtus, parameeter σ määrab kõvera teravuse/lameduse. Kõveraalune pindala kahe argumenttunnuse väärtuse vahel näitab vahemikku jäävate väärtuste esinemissagedust.

Teoreetiliselt võib normaaljaotusega muutuja omada väärtusi vahemikus $+\infty \ldots -\infty$, seejuures:

- keskväärtusest ±1 standardhälbe ulatuses on oodatavalt 68,26% väärtustest,
- ±2 standardhälbe ulatuses 95,44% väärtustest,
- ±3 standardhälbe ulatuses 99,73% väärtustest (joonis 1-10).

Kui normaaljaotusega tunnuse keskväärtus ja standardhälve on teada, siis saab tänu sellele leida tunnuse mingisse väärtusvahemikku kuulumise tõenäosust. Normaaljaotuse väärtused, mis erinevad keskväärtusest rohkem kui kolm standardhälvet, on haruldased ja neid nimetatakse **erinditeks**. Väikeses valimis olevate erindite puhul võib kergesti tegemist olla mõõtmisveaga. Vigaselt mõõdetud andmete puhul on põhjendatud nende välja jätmine analüüsitavast andmestikust. Erindeid aitab leida sagedushistogramm ja normaaltõenäosuse graafik (*normal probability plot*).



Joonis 1-10. Standardiseeritud normaaljaotusega muutuja oodatav vaatluste hulk vastavalt vahemiku laiusele standardhälvetes.

Klassikaline piirteoreem – sõltumatute, ühesuguse või mõõdukalt erineva jaotusega juhuslike suuruste normeeritud summade jada koondub liidetavate arvu lõpmatul suurendamisel standardiseeritud normaaljaotuseks. Teisisõnu on normaaljaotusega juhuslikud suurused, mille väärtus sõltub paljudest faktoritest. Kusjuures kõik faktorid avaldavad vaid nõrka mõju, mis võib olla nii juhuslikku suurust vähendav kui seda suurendav ning faktorite mõjud liituvad. Pidevate juhuslike suuruste puhul, mille väärtused on kujunenud suure hulga suhteliselt sõltumatute ja küllaltki ühetaolise mõjurite tulemusel, on jaotus modelleeritav normaaljaotusena. Piirteoreemi kasutatakse üldkogumi prognoositava keskväärtuse usalduspiiride leidmiseks regressioon- ja dispersioonanalüüsis.

Paljudest suhteliselt väikese mõjuga faktoritest sõltuvad juhuslikud suurused on oodatavalt normaaljaotusega

Mitmete statistikameetodite puhul on eelduseks, et uuritavad tunnused ja/või mudeli jäägid on normaaljaotusega. Kui tunnused ei ole normaaljaotusega, võivad normaaljaotust eeldavad meetodid – näiteks t test (ptk 1.4.3.1), korrelatsioonikordaja (ptk 1.5.2.2), lihtne lineaarne regressioon (ptk 1.5.3), dispersioonanalüüs (ptk 1.5.4) – hälbinud tulemusi anda. Nende asemel tuleks kasutada mitteparameetrilisi meetodeid.

Normaaljaotust või muud teoreetilist jaotust eeldavate meetodite kasutamine on teoreetiliselt õigustatud, kui eeldatav jaotus ei ole põhimõtteliselt välistatud ja kui erinevus normaaljaotusest ei ole statistiliselt oluline

Ühest küljest me teame, et normaaljaotus on teoreetiline üldistus, mida puhtal kujul looduses ei esine. R.C. Geary (1947) järgi tuleks kõigisse uutesse (statistika) õpikutesse kirjutada: "*Normality is a myth; there never was, and never will be, a normal distribution*". Eesti keeles oleks see: "(Jaotuse) normaalsus on müüt; normaaljaotust ei ole kunagi olnud ega saa kunagi olema".

Teisalt on normaaljaotus praktikas kasulik vahend, millele tuginevate meetodite kasutamine on õigustatud, kui uuritav jaotus võiks eeldatavasti olla normaaljaotuse lähedane või erinevus normaaljaotusest ei ole statistiliselt olulisel tasemel tõestatud. Kui erinevus normaaljaotusest on tõestatud, kuid väike, ei ole normaaljaotust eeldatavate meetodite kasutamine teoreetiliselt õige, kuid võib siiski olla põhjendatud, kuna sellest tulenev eeldatav viga on aktsepteeritavates piirides. Suurte valimite (mahuga üle paarisaja) puhul võib arvestada, et usaldusväärset erinevust normaaljaotusest õnnestub tõestada enamasti ka nii väikese kõrvalekalde puhul, mis analüüsi tulemusi märkimisväärselt ei mõjuta.

Teoreetilist jaotust eeldavate meetodite kasutamine teoreetilise jaotuse mittekehtimisel on praktikas põhjendatud kui sellega kaasneva eksimuse suurus on lubatud piires

1.3.4. Empiirilised jaotused

Empiirilise ja teoreetilise jaotuse (ptk 1.3.2) võrdlemine võimaldab kontrollida teoreetilise jaotuse eelduste paikapidavust empiiriliste andmete puhul. Seejuures tuleb silmas pidada, et statistika võimaldab tõestada/kontrollida vaid jaotuste erinevust, mitte sarnasust. Erinevuse puudumist tuleks eeldada seni, kui erinevus ei ole tõestatud, kuigi teoreetiliselt on erinevus alati olemas. Erinevuse olemasolu on sisukas järeldus ja kergekäelisi piisavalt tõendamata sisukaid järeldusi tuleks vältida. Empiirilise andmestiku erinevust mingist teoreetilisest jaotusest on praktikas reaalne kontrollida vaid juhul, kui andmestiku dimensionaalsus on madal, või kui uuritakse igat tunnust eraldi. Empiirilise jaotuse kirjeldamiseks on mitmeid arvkarakteristikuid, mis jagunevad laias laastus keskväärtuse hinnanguteks ja varieeruvuse hinnanguteks. Sõltumata sellest, milliseid näitajaid kasutatakse, tuleb töö aruandes kasutatud karakteristik selgelt ja korrektselt ära mainida.

Empiirilise jaotuse arvkarakteristikud eeldavad muutuja tüübi teadmist – näiteks nominaalseid klasse tähistavate koodide keskmine ei oma sisulist tähendust.

1.3.4.1. Keskmised

Aritmeetiline keskmine – üldkogumi puhul kasutatakse terminit keskväärtus. Valimil saab arvutada aritmeetilist keskmist, mis tähistab ühel kindlal viisil arvutatud üldkogumi keskväärtuse hinnangut. Tähistatakse ja arvutatakse järgmiselt.

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad [1\text{-}18]$$

Aritmeetilise keskmise erijuht on **kronoloogiline keskmine**, mida arvutatakse ajateljel olevatest väärtustest.

Ruutkeskmine võimendab suurte väärtuste osatähtsust ja võimaldab näiteks diameetri või raadiuse andmete abil hinnata keskmist pindala. Seda kasutatakse variatsioonistatistikutes ning näiteks metsatakseerimisel ja metsamaterjalide keskmise mahu hindamisel palkide diameetri järgi.

$$\overline{x}_{ruut} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}} \qquad [1\text{-}19]$$

Harmooniline keskmine sobib kasutamiseks positiivsete suhtarvuliste muutujate puhul, kus on eelkõige oluline murru nimetajas olev suurus, näiteks keskmise kiiruse arvutamisel.

$$\overline{x}_{harm} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \qquad [1\text{-}20]$$

Geomeetriline keskmine sobib positiivsete muutujate puhul, mille kordsed muutused omavad samasuurt tähendust. Saab kasutada näiteks keskmise juurdekasvuprotsendi ja suhtarvude keskmise arvutamisel.

$$\overline{x}_{geom} = \exp\left(\frac{\sum_{i}^{n} \ln(x_i)}{n}\right) = \sqrt[n]{\prod_{i}^{n} x_i} \qquad [1\text{-}21]$$

Geomeetrilist keskmist on nimetatud ka logaritmkeskmiseks, kuna see on logaritmskaalas kasutatav aritmeetiline keskmine.

$$\ln \overline{x}_{geom} = \frac{\sum_{i}^{n} \ln x_i}{n} \qquad [1\text{-}22]$$

Kaalutud keskmine on kasutusel juhul, kui muutuja väärtustele soovitakse näiteks lähteandmete erineva usaldatavuse tõttu omistada erinev osakaal (*w*) keskmise arvutamisel. Suurema kaaluga andmed mõjutavad keskmist rohkem kui väiksema kaaluga andmed.

$$\overline{x}_w = \frac{\sum_{i}^{n} w_i x_i}{\sum_{i}^{n} w_i} \qquad [1\text{-}23]$$

Tinglik keskmine on juhusliku suuruse selliste väärtuste keskmine, mis vastavad seatud tingimusele. Näiteks ainult teatud vahemikus olevate väärtuste keskmine.

Mood on juhusliku suuruse kõige tõenäolisem väärtus. Kui muutuja tõenäosusfunktsioon on ühetipuline, siis nimetatakse seda jaotust unimodaalseks, kui kahetipuline, siis bimodaalseks. Moodile vastavat väärtusklassi nimetatakse **modaalklassiks** ehk **moodklassiks**. Moodi saab kasutada igasuguste tunnuste, ka kvalitatiivsete tunnuste, korral. Kui valimis ei ole korduvaid mõõtmistulemusi, siis saab moodi kasutada vaid tunnuse väärtuste klassifitseerimise järel.

Mediaan on 50% kvantiil ehk suurus, millest juhuslik suurus on võrdse tõenäosusega suurem või väiksem. Sümmeetrilise jaotuse mediaan ja keskväärtus langevad ühte. Kui kogumis on paarisarv vaatlusi, siis arvutatakse mediaan kahe keskmise vaatluse aritmeetilise keskmisena. Uuritava tunnuse väärtuste järgi järjestatud vaatluste ehk variatsioonirea puhul saab mediaaniks oleva vaatluse järjekorranumbri arvutada valemi järgi

$$Me = \frac{n+1}{2}. \qquad [1\text{-}24]$$

Mediaani tuleks aritmeetilisele keskmisele eelistada asümmeetrilise jaotusega muutujate kirjeldamisel.

1.3.4.2. Variatsiooninäitajad

Miinimum ja maksimum on vastavalt suurim ja väikseim mõõtmistulemus.

Variatsiooniulatus ehk haare (*range*) on maksimumi ja miinimumi vahe.

Keskmine lineaarhälve (*average deviation*) on hälvete absoluutväärtuste keskmine.

$$\bar{d} = \frac{\sum_{i}^{n} \left| x_i - \bar{x} \right|}{n}. \qquad [1\text{-}25]$$

Dispersioon (*variance*) on keskmine ruuthälve, mis üldkogumi puhul avaldub:

$$\sigma^2 = \frac{\sum_{i}^{n} \left(x_i - \bar{x} \right)^2}{N}. \qquad [1\text{-}26]$$

Valimi järgi hinnates (*sample variance*) avaldub dispersioon järgmiselt.

$$s^2 = \frac{\sum_{i}^{n} \left(x_i - \bar{x} \right)^2}{n-1} \qquad [1\text{-}27]$$

Kaheväärtuselise tunnuse puhul saab valimi dispersiooni arvutada edukate katsete tõenäosuse hinnangu (sageduse) kaudu.

$$s^2 = \hat{P}\left(1 - \hat{P}\right) \qquad [1\text{-}28]$$

Standardhälve (*standard deviation*) ehk ruutkeskmine hälve avaldub üldkogumis kujul

$$\sigma = \sqrt{\sigma^2}; \qquad [1\text{-}29]$$

ja valimi järgi hinnates

$$s = \sqrt{s^2}; \qquad [1\text{-}30]$$

kaheväärtuselise tunnuse puhul avaldub valimi standardhälve variantide tõenäosuste (P ja $1-P$) kaudu

$$s = \sqrt{\hat{P}\left(1 - \hat{P}\right)}. \qquad [1\text{-}31]$$

Standardhälbe lisamine mõõtmistulemuste keskmisele näitab andmete hajuvust mõõtmistulemuste normaaljaotuse eeldusel (joonis 1-10).

Variatsioonikoefitsient (*coefficient of variation – CV*) võimaldab võrrelda eri skaalades olevate muutujate varieeruvust. Enamasti kasutatakse aritmeetilise keskmisega normeeritud standardhälvet.

$$V = \frac{s}{\bar{x}} \qquad [1\text{-}32]$$

Kui variatsioonikoefitsiendis kasutatakse keskmist lineaarhälvet, tuleks seda kirjutise tekstis mainida. Variatsioonikoefitsienti võib avaldada ka protsentides. Samuti võimaldab variatsioonikoefitsient võrrelda erinevate tunnuste varieeruvust, näiteks mis varieerub rohkem, kas mulla toitaineterikkus või taimkatte biomass.

Standardviga (*standard error – SE*) on valimi järgi arvutatud hinnangu standardhälve, mis kirjeldab üldkogumi keskväärtuse või muu parameetri valimist saadud hinnangu täpsust. Samast üldkogumist korduvalt võetud valimite keskväärtuste jaotus läheneb normaaljaotusele, mille standardhälve võrdub hinnangu standardveaga. Seda standardhälvet kasutatakse hindamaks kindla suurusega *n* valimite keskväärtuste usaldusväärsust üldkogumi keskväärtuse hindamisel; samal viisil, nagu üksikvalimi standardhälve iseloomustab selle valimi võimet hinnata üldkogumi keskväärtust. Standardviga kasutatakse ka üldkeskmise hinnangu usalduspiiride leidmiseks. Kuna üldkogumi standardhälve ei ole teada, siis tuleb selle asemel kasutada valimi standardhälvet *s*. Seega on valimi järgi arvutatud standardvea valem järgmine:

$$SE = \frac{s}{\sqrt{n}}. \qquad [1\text{-}33]$$

Standardvigade võrdlemine eeldab vigade jaotuse ühtlust prognoositavate väärtuste erinevates suurusvahemikes. Kui väiksemate prognoositud väärtuste korral on ka standardviga keskmiselt väiksem, siis on erineva keskväärtusega hinnangute vea võrdlemiseks õigem kasutada suhtelist standardviga: *SE/hinnang*.

Asümmeetriakordaja (*skewness*) näitab jaotuse poolte erinevust. Kui juhusliku suuruse tõenäosusfunktsioon on sümmeetriline, siis tema asümmeetriakordaja võrdub nulliga. Kui asümmeetriakordaja on positiivne, siis on jaotusel raske saba suuremate väärtuste pool (vasakkaldeline muutuja), kui negatiivne, siis väiksemate väärtuste pool (paremkaldeline muutuja). Raske saba tähendab ebaproportsionaalset hulka keskväärtusest kaugel olevaid ehk väljaulatuvaid väärtusi.

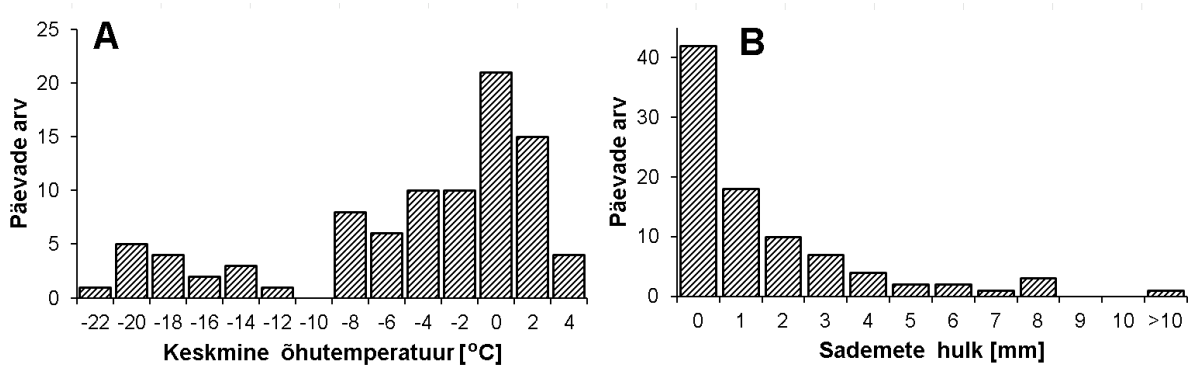
$$A = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^3 \qquad [1\text{-}34]$$

Selliselt arvutatud asümmeetriakordaja väärtused sõltuvad mõõtühikutest ja seetõttu tuleb erinevate muutujate asümmeetrilisuse võrdlemiseks asümmeetriakordaja selle standardhälbega läbi jagada. Asümmeetriakordaja standardhälbe saab suurte valimimahtude korral arvutada järgneva valemi järgi.

$$s_A = \sqrt{\frac{6}{n}} \qquad [1\text{-}35]$$

Matemaatilises statistikas defineeritakse asümmeetriakordajat (a_X) ka kui standardiseeritud juhusliku suuruse kolmanda tsentraalse momendi ja standardhälbe kuubi suhet.

Asümmeetriakordaja järgi on näiteks talvised ööpäeva keskmised temperatuurid negatiivse asümmetriaga (joonis 1-11A). Ööpäevase sademete hulga jaotus on positiivse asümmeetriaga (joonis 1-11B).



Joonis 1-11. A – ööpäeva keskmised temperatuurid Tartus 2011. aasta jaanuaris, veebruaris ja märtsis. Asümmeetria on negatiivne (A = −1,14345) ja ekstsess (e_x = 0,28252) suhteliselt väike, kuid siiski positiivne, kuna nullilähedase temperatuuriga päevi oli 2011. aasta talvel suhteliselt palju. B – ööpäeva sademete hulga jaotus Tartus samal ajaperioodil. Asümmeetria ja ekstsessi väärtused on suured (A = 5,6177; e_x = 2,3153), kuna palju sajab harva ja sademete hulk ei saa olla negatiivne.

Ekstsess (*kurtosis*) on jaotusfunktsiooni karakteristik, mis on terava tipu ja raskete sabadega jaotuste puhul positiivne, lamedatipulise või nõgusa jaotuse puhul negatiivne (kuid mitte väiksem kui –2). Kolme lahutamine tagab ektsessi nullväärtuse standardiseeritud normaaljaotuse korral.

$$e_x = \frac{1}{n}\sum_{i}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^4 - 3 \qquad [1\text{-}36]$$

Erinevate muutujate ekstsesside võrreldavaks muutmiseks tuleks ekstsessid nende standardhälvetega läbi jagada. Suure valimimahu eeldusel saab ekstsessi standardhälvet arvutada järgmiselt.

$$s_E = \sqrt{\frac{n}{24}} \qquad [1\text{-}37]$$

Negatiivse ektsessiga on näiteks ööpäevase sademete hulga jaotus suvel, sest üksikutel päevadel sajab keskmisest oluliselt rohkem, teistel päevadel ei saja üldse. Talveperioodil võib ööpäevase sademete hulga ekstsess olla ka positiivne, sest vähese sajuga päevi on palju (joonis 1-11B).

Kvantiil (*quantile*) ehk jaotuse α–kvantiil on arv, millest väiksemate väärtuste sagedus on α·100%. Empiirilise jaotuse kvantiil jagab järjestatud variatsioonirea suhtes α / (1–α). Tähistus q_α. Kvantiilid sobivad kasutamiseks kõigi järjestustunnuste korral. Mediaan on 0,5 kvantiil.

Kvartiilid (*quartiles*) **–** 0,25 ja 0,75 kvantiili nimetatakse vastavalt alumiseks ja ülemiseks kvartiiliks.

Detsiilid (*deciles*) **–** kümnendik-kvantiilid.

Tsentiilid (*percentiles*) **–** sajandik-kvantiilid.

Täiendkvantiil (*1– α quantile*) – kvantiil kohal *1– α*. Tähistatakse $q_{1-\alpha}$.

Kvantiilihaare (*quantile range*) – mediaani suhtes sümmeetriliste kvantiilide vahel paiknev väärtusvahemik. Näiteks 0,05 ja 0,95 kvantiilile vastavad väärtused. Enamasti kasutatakse kvartiilihaardena, mis on vahemik alumise ja ülemise kvartiili vahel.

1.4. Üldkogumi parameetrite hindamine valimi alusel

1.4.1. Punkt- ja vahemikhinnangud

Üldkogumi parameetrite tegelikud väärtused ei ole tavaliselt teada. Neid on vaja valimi abil võimalikult täpselt või nõutava täpsusega hinnata. Hinnang võib koosneda ühest väärtusest (**punkthinnang**) või vahemikust, millesse tegelik väärtus peaks mingi tõenäosusega kuuluma (**vahemikhinnang**). Parameetrite hinnangute puhul lisatakse parameetri tähisele "katus", näiteks $\hat{x}$.

Juhuslikust valimist arvutatud hinnang on juhuslik suurus; üldkogumi parameetri tegelik väärtus ei ole juhuslik

Ühest ja samast üldkogumist võib erinevate valimite järgi saada samade parameetrite erinevaid hinnanguid. Parameetrite kaudsel hindamisel on eesmärgiks saada **nihketa** hinnang, mille puhul langeb hinnangute keskväärtus kokku hinnatava parameetri tegeliku väärtusega. Hinnangute varieeruvus sõltub hinnatava parameetri hajuvusest, kasutatavast valiku viisist, mõõtmismetoodikast ja **valimi mahust** (*sample size*) ehk üldkogumist uuringuks valitud üksuste arvust. Täpsema hinnangu saamiseks on kõige lihtsam valimi mahtu suurendada. Enamikel juhtudest on n korda täpsema hinnangu saamiseks vaja valimi mahtu suurendada n^2 korda. Hinnang on **efektiivne**, kui see on nihketa ja minimaalse hajuvusega.

Üldkogumi parameetrite punkthinnangutena kasutatakse üldjuhul valimi põhjal arvutatud karakteristikuid. Üldkogumi keskväärtuse, mediaani, dispersiooni ja standardhälbe hinnanguks sobib kõige paremini valimi vastav karakteristik. Näiteks sündmuse tõenäosuse hinnanguks binoomjaotusega muutuja puhul tuleb võtta sündmuse suhteline sagedus.

Kuna valimi järgi ei saa üldkogumi parameetreid absoluutse täpsusega hinnata, on asjakohane kasutada vahemikhinnanguid. Valimi põhjal on võimalik määrata **usaldusintervalli** ehk usalduspiirkonda, see on väärtusvahemikku, mis sisaldab etteantud tõenäosusega parameetri tegelikku väärtust. Usaldusintervalli otsapunkte nimetatakse **usalduspiirideks**. Kui esineb vaid üks usalduspiir, siis on tegemist ühelt poolt tõkestamata usalduspiirkonnaga. Kui usalduspiiride asümmeetrilisus ei ole piisavalt tõendatud või on selle arvutamine liiga tülikas, siis kasutatakse sümmeetrilisi usalduspiire (näiteks $x = 3 \pm 1$). Kas valimi järgi hinnatav üldkogumi parameeter on tegelikult usaldusintervalli piires või mitte, ei ole kindlalt teada. Etteantud tõenäosus määrab vaid üldkogumi parameetri katmise tõenäosuse korduvate sõltumatute katsete (juhusliku valimi võtmine ja usalduspiiride arvutamine) käigus. Igal konkreetsel korral usalduspiirid kas katavad üldkogumi parameetrit või siis mitte.

Keskväärtuste määramise puhul tehakse vahet usalduspiiridel (*confidence limits*) ja **prognoosipiiridel** (*prediction limits*). Neist esimene näitab, millises vahemikus on etteantud tõenäosuse korral hinnangute keskväärtus ja teine näitab, millises vahemikus on üksikvaatluse prognoositav väärtus.

1.4.1.1. Keskväärtuse usalduspiirid

Eeldatavalt normaaljaotusega juhusliku muutuja keskväärtuse usalduspiiride leidmisel kasutatakse suure valimi korral normaaljaotust ning väikese korral t jaotust. Viimati mainitu sarnaneb normaaljaotusega, kuid väikeste valimite puhul eeldab keskväärtusest kaugele hajuvate vaatluste suuremat osakaalu. Kahepoolse usalduspiiri leidmisel tuleb olulisuse nivoo α jagada kahega, sest viga on nii alumisest usalduspiirist väiksem väärtus kui ka ülemisest usalduspiirist suurem väärtus. Vaid

ühepoolse usalduspiiriga piirdumine on põhjendatud, kui hinnangu hälve vastassuunas on välistatud. Vabadusastmete arvuks *t* jaotuse puhul on *n–1*. Vabadusastmete arv on sõltumatute juhuslike tunnuste arv miinus neid tunnuseid siduvate tingimuste arv.

Seega, alumine ja ülemine usalduspiir leitakse vastavalt järgmistele valemitele

$$x_{\_} = x - t_{\frac{\alpha}{2};n-1}\frac{s}{\sqrt{n}}, \qquad [1\text{-}38]$$

$$x^{-} = x + t_{\frac{\alpha}{2};n-1}\frac{s}{\sqrt{n}}, \qquad [1\text{-}39]$$

kus *t* tähistab parameetriga (*n*–1) *t* jaotuse *α/2* täiendkvantiili, *n* on vaatluste arv, *s* on valimi standardhälve ja *α* tähistab olulisuse nivood. *t* testi puhul tuleb teada, kas tegemist on sõltuvate või sõltumatute valimitega.

Usaldusintervalli laiuse reguleerimine

Usaldusintervalli laius sõltub uuritavate väärtuste hajuvusest (standardhälbest) ja valimi suurusest. Kuna üldkogumi dispersiooni muuta ei saa, siis tuleb täpsema keskväärtuse hinnangu saamiseks valimi mahtu suurendada. Kui vaatlused on kallid või töömahukad, siis on otstarbekas eeluuringute abil enne lõpliku uuringu tegemist hinnata nõutava täpsusega tulemuse saamiseks vajaliku valimi suurus. Vajaliku valimi ligikaudset suurust saab leida valemiga

$$n \geq \left(t_{\frac{\alpha}{2};n_0-1}\frac{s}{d} \right)^2, \qquad [1\text{-}40]$$

kus ±*d* on soovitav täpsus, *α* olulise nivoo, *s* proovivalimi standardhälve ja n_0 proovivalimi maht.

Kuna üldkogumi hajuvust muuta ei saa, on hinnangu usaldusväärsust võimalik tõsta vaid suurendades valimi mahtu ja tagades valimi esinduslikkust

1.4.2. Statistilised hüpoteesid

Praktikas võib huvi pakkuda mitte ainult üldkogumi karakteristikute võimalikult täpne hinnang, vaid ka nende vastavus või mittevastavus etteantud tingimusele. Näiteks, kas ühes kasvukohatüübis on puude kasvukiirus keskmiselt suurem kui teises, või kas majandustegevus looduskaitseala naabruses mõjutab kaitseala asustavate liikide arvukust rohkem kui kümme protsenti või mitte? Selliste küsimuste vastus ei ole arvuline, vaid on ei/jah otsustus. See on otsustus, kas vaatlusandmed on mingi oletusega kooskõlas või ei ole. Kui otsustus üldkogumi kohta tuleb langetada valimi põhjal, siis ei ole võimalik täiesti kindlaid järeldusi teha – ei või jah otsustus on tõenäosusliku iseloomuga.

Valimi põhjal ei ole võimalik teha absoluutse kindlusega järeldusi üldkogumi kohta

Statistiline hüpotees on üldkogumi kohta tehtud oletus. Enamasti koosneb see kahest alternatiivsest oletusest: oletus kehtib (sisukas hüpotees) ja oletus ei kehti (nullhüpotees). Sisukaid hüpoteese võib olla rohkem kui üks. **Nullhüpotees** on statistilise hüpoteesi see osa, mille kohaselt püstitatud oletus ei kehti (otsitavat seost või erinevust ei ole, uuritav faktor ei mõjuta tunnuse jaotust

jne). Enamasti on nullhüpotees see, mille puhul eeldatakse vaid juhuslikkuse mõju. Ilma sisuka hüpoteesi kehtivuse tõestuseta eeldatakse üldjuhul nullhüpoteesi kehtivust. Nullhüpoteesi tähistatakse H_0. Hüpoteeside kontrollimise loogika alustab nullhüpoteesi kehtimise eeldusest, see tähendab, et esmalt uurime, kui tõenäone on antud andmete saamine nullhüpoteesi kehtimise korral.

Sisukas hüpotees on statistilise hüpoteesi see osa, mille kohaselt püstitatud hüpotees kehtib (otsitav seos on olemas, uuritav faktor mõjutab tunnuse jaotust, üldkogumi karakteristik erineb etteantud väärtusest, valimid on pärit erinevatest üldkogumitest). Sisukat hüpoteesi nimetatakse ka alternatiivseks hüpoteesiks, sest see väljendab alternatiivi nullhüpoteesile. Sisukat hüpoteesi tähistatakse H_1, mitme sisuka hüpoteesi korral H_2, H_3 jne.

Kuna valim on vaid väljavõte üldkogumist, siis ei saa valimi põhjal täie kindlusega otsust ühe või teise hüpoteesi kasuks langetada. Valimi põhjal otsuste langetamisel on alati olemas väära otsuse langetamise risk. Saab vaid väita, et üht või teist hüpoteesi võib vastu võtta eksimise teatud tõenäosusega ehk teatud **usaldusnivool**. Usaldusnivoo sõltub mõõdetud tunnuse väärtuste hajuvusest ja erinevuse suurusest uuritava karakteristiku kriitilise ja valimist arvutatud väärtuse vahel.

Valimi alusel otsuseid langetades võib eksida kahel viisil, mida nimetatakse esimest liiki veaks ja teist liiki veaks. **Esimest liiki viga** on mittekehtiva sisuka hüpoteesi vastuvõtmine ning **teist liiki viga** on mittekehtiva nullhüpoteesi juurde jäämine. Sisukale hüpoteesile piisava kinnituse puudumisel nullhüpoteesi juurde jäämine ei tähenda siiski nullhüpoteesi tõestamist. Nullhüpoteesi juurde tuleb jääda, kui sisuka hüpoteesi poolt ei ole piisavalt tõendeid. Statistiliste hüpoteeside puhul on vaikimisi eelduseks juhuslikkus, mida deterministlikus maailmas tuleb mõista analüüsi mitte kaasatud faktorite mõjude summana. Esimest liiki vea tagajärjed on enamasti halvemad kui teist liiki vea puhul, seetõttu püütakse eelkõige vältida esimest liiki viga.

Uurija poolt seatud piiri esimest liiki vea tõenäosusele nimetatakse **olulisuse nivooks** ehk riskitasemeks (tähistatakse α). Sagedamini kasutatakse olulisuse nivood 0,05; kuid ka 0,1 ja 0,01. Esimest liiki vea tõenäosus on statistilise olulisuse tõenäosus ehk **olulisustõenäosus** (p). Sisuka hüpoteesi võib lugeda tõestatuks, kui $p \leq \alpha$. Kokkuleppeliselt loetakse loodusteadustes vaikimisi nõutavaks olulisuse nivooks $p < 0{,}05$. Eksimise tõenäosuse ühe või teise kindla taseme fikseerimine on seejuures siiski vaid kokkulepe teatud kultuuriruumis.

Olulisustõenäosus on eksimise tõenäosus sisuka oletuse õigsuse uskumisel, kasutatud meetodi eelduste kehtimisel ja valimite esinduslikkuse korral

D.H. Johnson (1999) juhib tähelepanu, et **olulisustõenäosus** (p) näitab vaid, kuivõrd oodatavad on vaadeldud või veelgi äärmuslikumad tulemused nullhüpoteesi kehtimisel, eeldades kasutatud mudeli õigsust ja valimi esinduslikkust. Kuna kõik mudelid on üldistused ja looduses kogutud vaatlusandmed ei ole kunagi igas mõttes esinduslikud, siis ei saa päriselt õigeks lugeda ühtegi järgnevast kolmest väitest:

- statistiline olulisus (p) on tõenäosus, et tulemus on juhuslik;
- $1-p$ näitab tulemuse usaldatavust;
- p on tõenäosus, et nullhüpotees on õige.

Statistilisi teste ja olulisuse nivoosid ei peaks ületähtsustama, sest erinevus tulemuses $p = 0{,}049$ ja $p = 0{,}051$ on illusoorne, kuna p arvutamisel kasutatud meetodite eeldused on täidetud vaid ligikaudselt. Arvuna esitatud olulisustõenäosus on seejuures siiski informatiivsem kui tulemuse vastavus/mittevastavus mingile olulisusnivoole. Olulisustõenäosuse tõlgendamisel tuleks arvestada ja teistele

selgitada riske ühe või teise otsustuse puhul. Näiteks kestev ülepüük võib kalavarude hävingule kaasa aidata ka juhul, kui kalastamise mõju ei ole suudetud tõestada. Mudelite puhul ei tohiks eeldada, et esialgselt parim mudel on ainuõige.

Olulisusnivoo valikul ja olulisustõenäosuse tõlgendamisel tuleks arvestada esimest ja teist tüüpi veaga kaasnevaid riske

Piirväärtust teist liiki veale nimetatakse **võimsuse nivooks** (tähistatakse β). Erinevate otsuse langetamise meetodite, reeglite, kriteeriumite ja statistiliste testide puhul on tõenäosus sama olulisuse nivoo juures sisukat hüpoteesi ära tunda (kriteeriumi võimsus) erinev. **Kriteeriumi võimsus** on tema võime vältida teist tüüpi viga ehk võime kehtivat sisukat hüpoteesi tõestada. Kui meil on kaks statistilist testi, mille korral on esimest liiki vea tegemise tõenäosus täpselt võrdne, siis on eelistatumaks suurema võimsusega test. Sageli ei ole praktikas olukord sedavõrd lihtne ja võimsuse ning esimest liiki vea tegemise ja teist liiki vea tegemise tõenäosuste vahel on lõivsuhe – võimsama testi korral suureneb ka esimest liiki vea tegemise tõenäosus. Kriteeriumi võimsus sõltub enamasti tema sobivusest uuritavate andmete jaotuse suhtes. Kuna erinevad testid võivad samale küsimusele anda ka samade andmete põhjal erineva vastuse, siis tuleb uurimuses alati kirja panna, millist testi kasutati.

Tavaliselt langetatakse otsus hüpoteesi kehtivuse kohta selleks otstarbeks arvutatud väärtuse ehk **statistiku** põhjal. Statistilise olulisuse hindamise testid lähtuvad igaüks mingitest teoreetilistest eeldustest, mistõttu on arvutatud statistiline olulisustõenäosus täpne vaid nende eelduste kehtimise korral. Olulisustõenäosust saab arvutada vaid nullmudeli suhtes, sest see on lihtne teoreetiline mudel, mis vastab etteantud tingimustele. Sisukale oletusele vastavad seosed on keerukamad ja reeglina me ei tunne neid nii hästi, et sisuka hüpoteesi kohta teoreetilist mudelit koostada. Olulisustõenäosos pole mitte nullhüpoteesi kehtimise tõenäosus, vaid tõenäosus saada selline või veelgi äärmuslikum valim tingimusel, et nullhüpotees kehtib ja tehtavad eeldused (nt sõltumatus, dispersioonide võrdsus vms) on täidetud. Sellest tulenevalt võib väike olulisustõenäosus olla märk sellest, et nullhüpotees ilmselt ei kehti, aga ka hoopis sellest, et tehtud eeldused ei ole meie andmete korral täidetud.

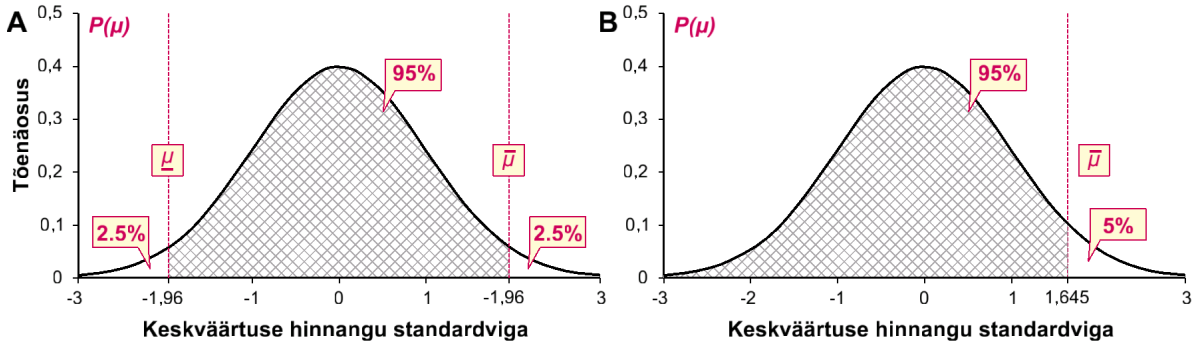
Olulisustõenäosus ei ole nullhüpoteesi kehtimise tõenäosus, vaid tõenäosus saada selline või veelgi äärmuslikum valim nullhüpoteesi kehtimise korral

1.4.2.1. Hüpoteesid üldkogumi keskväärtuse kohta

Keskväärtuse, aga ka teiste arvuliste parameetrite valimi järgi määramise puhul huvitab meid tavaliselt üldkogumi vastava näitaja vastavus mingile etteantud väärtusele (μ_0). Hüpoteesi tingimusega, et üldkogumi parameeter erineb etteantud väärtusest ($\mu \neq \mu_0$), nimetatakse **kahepoolseks hüpoteesiks** (joonis 1-12A). Kahepoolse hüpoteesi korral uuritakse, kas etteantud väärtus on alumise ja ülemise usalduspiiri vahel. Kui kriitiliste olulisuste väärtuste tabel või arvutiprogramm annab vaid ühepoolse hüpoteesi usalduspiirile vastava väärtuse, siis kahepoolse hüpoteesi usalduspiirile vastava väärtuse saamiseks tuleb reeglina kasutada kaks korda väiksemat olulisuse nivood. Eeldatakse, et pooltel juhtudel võib esimest liiki viga anda tulemuse allapoole alumist usalduspiiri ja pooltel juhtudel ülespoole ülemist usalduspiiri.

Hüpoteesi $\mu < \mu_0$ ja hüpoteesi $\mu > \mu_0$ nimetatakse **ühepoolseteks hüpoteesideks** (tabel 1, joonis 1-12B). Ühepoolse hüpoteesi korral uuritakse vaid ühtepidi erinevust, erinevus teises suunas on kas välistatud või ei paku huvi. Kui sisukaks hüpoteesiks on erinevuse olemasolu, siis võib sisukat hüpoteesi lugeda kehtivaks, kui etteantud keskväärtus ei ole valimi järgi hinnatud keskväärtuse usal-

dusintervallis. Siinjuures tuleks tähelepanu juhtida, et üldkogumi keskväärtusi ja teisi parameetreid hinnatakse küll valimi järgi, aga hüpoteesid käivad üldkogumi keskväärtuste kohta, mitte valimist arvutatud keskväärtuste kohta.



Joonis 1-12. Statistilise ootuse sümmeetriliselt paiknevate usalduspiiridega tihedusfunktsioon kahepoolse hüpoteesi korral (A) ja statistilise ootuse ühelt poolt tõkestatud usalduspiiriga tihedusfunktsioon ühepoolse hüpoteesi korral (B). Ristviirutusega ala tähistab kõvera alust pinda piirväärtuste vahel; selle ala pindala väljendab piirväärtuste vahele jäävate väärtuste oodatavat sagedust.

Tabel 1. Keskväärtuse kohta käivad hüpoteesid (μ – uuritava üldkogumi tegelik keskväärtus, μ_0 – ette-antud suurus, millega keskväärtust võrreldakse).

Hüpotees	Tingimus		
	kahepoolne	ühepoolne	
H_0	$\mu = \mu_0$	$\mu \leq \mu_0$	$\mu \geq \mu_0$
H_1	$\mu \neq \mu_0$	$\mu > \mu_0$	$\mu < \mu_0$

1.4.3. Kahe üldkogumi võrdlemine

Sageli on vaja otsustada, kas kahte andmehulka võib käsitleda ühe kogumina või on nende vahel oluline erinevus. Enamasti uuritakse valimi põhjal, kas mingil olulisuse nivool võib järeldada kahe jaotuse keskmiste või jaotuse erinevust. Üldkogumite võrdlemiseks kasutatavad valimid võivad olla sõltumatud või sõltuvad. Sõltumatud valimid ei sisalda sama objekti kohta mitut vaatlust. Sõltumatud on näiteks mikroelupaikade rohkus kahes eri tüüpi metsas.

Sõltuvates (ehk korduvmõõtmistega) valimites kasutatakse vaatluseid ettemääratud paaridena või rohkemaarvuliste kombinatsioonidena – sama objekti vaadeldakse kord ühtedes, kord teistes tingimustes või kasutatakse kombineeritult sarnaseid objekte. Näiteks kui mikroelupaikade rohkust on mõõdetud samades metsades enne ja pärast metsamajanduslikku tegevust. Ka mikroelupaikade rohkust eri kaitsealades ja nendega piirnevates majandusmetsades võib käsitleda sõltuvat valimit moodustavate paaridena. Sõltuvates valimites moodustuvad vaatluste paarid ja järeldusi tehakse ühe uue valimi alusel, mille moodustavad vaatlustulemuste vahed. Tüüpiliselt kontrollitakse kas paariliste vaatluste keskmine vahe erineb nullist. Valimi tüüpide kohta vaata ka peatükk 1.1.2.

1.4.3.1. Keskväärtuste võrdlemine

Keskmiste võrdlemisel kasutatakse statistilist hüpoteesi: H_0 – üldkogumite keskväärtused on võrdsed, H_1 – üldkogumite keskmised on erinevad. Kui üldkogumite erinevuse suund on ette teada, siis saab sisuka hüpoteesi esitada ühepoolse hüpoteesina kujul: H_1 – esimese üldkogumi keskväärtus on suurem või H_1 – teise üldkogumi keskväärtus on suurem. Tuleb siiski meeles pidada, et

looduslikest objektidest koostatud valimite puhul ei ole võimalik kahte absoluutselt võrdse keskväärtusega valimit koostada. Seda isegi mitte kõikse valimi puhul, välja arvatud juhul, kui valimid just samadest objektidest ei koosne. Seega käivad statistilised hüpoteesid pigem küsimuse kohta, kas valimid on pärit ühest objektide homogeensest üldkogumist või eri kogumitest.

Normaaljaotust eeldavad võrdlusmeetodid Z statistik ja t test

Suurte valimite korral on üldkogumite keskväärtuste vahe hindamine piisavalt usaldusväärne ka siis, kui uuritava tunnuse jaotus mõnevõrra erineb normaaljaotusest. Kahe sõltumatu suure mahuga valimi järgi hinnatud keskväärtuste ($\bar{x}_1$ ja $\bar{x}_2$) erinevuse standardhälvet (s_*) saab üksikvalimite standardhälvetest (s_1 ja s_2) ning valimi mahtudest (n_1 ja n_2) arvutada valemiga

$$s_* = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad [1\text{-}41]$$

Keskväärtuste erinevuse standardhälbega normeeritakse valimite põhjal hinnatud keskväärtuste vahe ja arvutatakse statistik *Z*.

$$Z = \frac{|\bar{x}_1 - \bar{x}_2|}{s_*} \qquad [1\text{-}42]$$

Kui nullhüpotees kehtib, on *Z* standardiseeritud normaaljaotusega. Statistiku *Z* väärtus kaks viitab näiteks, et valimite keskmised erinevad üksteisest kahe standardhälbe võrra. Samast üldkogumist võetud valimite vahel tekib nullhüpoteesi kehtimisel selline või suurem erinevus vaid tõenäosusega 0,0456. Seega on üsna tõenäoline, et valimid on pärit erinevatest üldkogumitest.

Sõltuvate valimite korral arvutatakse *Z* statistik vastavalt valemile

$$Z = \sqrt{n}\frac{\bar{V}}{s_V}, \qquad [1\text{-}43]$$

kus *n* on vaatluspaaride arv, $\bar{V}$ on vaatluste keskmine erinevus, S_V on vaatluste erinevuse standardhälve.

Keskväärtuste vahe usaldusväärsuse kontrollimiseks tuleb Z statistiku väärtust võrrelda standardiseeritud normaaljaotusega. Selle tulemusena saadakse hinnang sisuka hüpoteesi usaldatavuse kohta. Kui valimid on väikesed ja/või tegemist ei ole normaaljaotusega, annab *Z* test põhjendamatult kõrge usaldusnivoo. Kui valimid on väikesed ja võrreldavad tunnused on normaaljaotusega, siis saab keskmiste võrdlemiseks kasutada *t* jaotust ja *Z* testi asemel *t* testi. Kui tunnuste hajuvus on ühesugune, siis on nende keskmiste vahe standardhälve arvutatav valemi järgi

$$s_* = s\sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}} = \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \qquad [1\text{-}44]$$

Üldkeskmiste mitteerinevuse korral, see tähendab nullhüpoteesi kehtimise korral on *T* statistik *t* jaotusega

$$T = \frac{|\bar{x}_1 - \bar{x}_2|}{s_*}, \qquad [1\text{-}45]$$

kus $\bar{x}_1$ ja $\bar{x}_2$ on vastavalt esimese ja teise valimi keskmine, s_* on keskmiste vahe standardhälbe hinnang vastavalt eeltoodud valemile ning *t* jaotust kasutatakse vabadusastmete arvuga ($n_1 + n_2 - 2$). *t* test sobib ka normaaljaotusega väikeste sõltuvate valimite abil võrreldavate keskväärtuste vahe

võrdlemiseks. Sel juhul on vaatluste arv mõlemas valimis sama ja *t* statistik arvutatakse samuti, nagu eeltoodud *Z* statistik. Hüpoteesi kontrollimisel võetakse *t* jaotuse parameetriks *n–1*.

Mitteparameetrilised võrdlusmeetodid

Mitteparameetrilised meetodid ei eelda andmete parameetrilist jaotust. Klassikalised mitte-parameetrilised meetodid ei kasuta otseselt vaatlustulemusi, vaid nende erinevuse suunda (märgitest) või **astakuid** ehk suuruse järgi järjestatud vaatluste järjekorranumbreid (*U* test, *W* test). Kui kaks valimit oluliselt ei erine, siis ühiselt järjestatud vaatluste reas on mõlema valimi komponendid hästi segunenud. Kui valimid erinevad, siis tekivad ühise rea algusesse või lõppu ühte või teise valimisse kuuluvate vaatluste kogumid. Tunnuse jaotustüüp ei ole seejuures üldse oluline, tunnus ei pea tingimata olema pideval skaalal, kuid väärtused peavad olema järjestatavad. Mitteparameetrilisi teste võib kasutada ka normaaljaotusega tunnuste ja suurte valimite puhul, kuid enamikel juhtudel on need parameetrilistest meetoditest vähem võimsad. See tähendab, et osadel juhtudel õnnestuks parameetrilise meetodiga nullhüpotees ümber lükata, mitteparameetrilisega see aga ei õnnestu.

Märgitest

Märgitest (*sign test*) on sõltuvate valimite *t* testi mitteparameetriline analoog. See test on kasutatav muutumistendentsi olemasolu kontrollimiseks sõltuvates valimites. Märgitestiga saab kontrollida valimite erinevust igasuguste järjestatavate tunnuste puhul. Testiks on tarvis fikseerida vaid muutuse suund igas vaatluspaaris. Kui erinevused *n* vaatlusest koosnevates valimites on juhuslikud, siis on positiivsed ja negatiivsed muutused võrdvõimalikud ning positiivsete ja samuti negatiivsete muutuste arv on binoomjaotusega *B(n;0,5)*. Võrdse väärtusega vaatluspaarid võib kas testi arvutamisest välja jätta või omistada neile väärtus 0,5. Viimane variant on eelistatum juhul, kui erinevuse puudumine vaatluspaaris on sisulise tähendusega.

Wilcoxoni test

Wilcoxoni test on sõltuvate valimite *t* testi teine mitteparameetriline analoog. Selle testi kasutamisel asendatakse arvtunnuse väärtused järjekorranumbrite ehk astakutega. Mõlemas valimis olevad vaatlused järjestatakse ühisesse variatsiooniritta. Võrdsed vaatlustulemused saavad ühise astaku, mis on nende järjekorranumbrite keskmine. Valimite segunemise määra kirjeldamiseks arvutatakse Wilcoxoni statistikud W_1 ja W_2, mis on vastavalt esimese ja teise valimi astakute summad. *W* statistikute õigsust saab kontrollida valemi järgi

$$W_1 + W_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2} \qquad [1\text{-}46]$$

Sisuka hüpoteesi kehtivuse kontrollimiseks võrreldakse *W* väärtust väiksemate kriitiliste väärtustega. Test sobib paremini selliste valimite võrdlemiseks, kus ei ole korduvaid väärtusi. Väikese mahuga valimite jaoks on olemas tabelid *W* statistiku kriitiliste väärtustega (näiteks Tiit 1972). Samuti saab *W* statistiku kriitilisi väärtusi kontrollida statistikatarkvaraga. Peale selle on *W* statistikut lihtne ümber arvutada *U* statistikuks.

$$U_- = W_1 - \frac{n_1(n_1 + 1)}{2} \qquad [1\text{-}47]$$

$$U_+ = W_2 - \frac{n_2(n_2 + 1)}{2} \qquad [1\text{-}48]$$

Sõltuvate vaatluste puhul saab kasutada ka Wilcoxoni **astakmärgitesti**. Selle puhul tuleb moodustada paariliste vaatluste vahede absoluutväärtuste variatsioonirida. Nulliga võrduvad vahed võib vaatlusreast välja jätta. Iga vahe jaoks tuleb leida astak ja siis positiivsete astakute ja negatiivsete astakute summa *W*– ja *W*+. Kui väärtuste erinevused paarilistes vaatlustes on juhuslikud, on negatiivsed ja positiivsed astakud astakute absoluutväärtuste reas ühtlaselt segunenud. Sisuka hüpoteesi kontrollimiseks võetakse väiksem statistikutest *W*– ja *W*+ ning võrreldakse seda kriitilise väärtusega.

Mann-Whitney U test

U test on sõltumatute valimite *t* testi mitteparameetriline analoog. *U* testi kasutamiseks peavad uuritava tunnuse väärtused olema järjestatavad – tunnus peab olema pidev või omama palju väärtusi. Testi kasutamiseks tuleb loendada esimese valimi iga vaatluse jaoks temast väiksemate ja temaga võrdsete teise valimi vaatluste arv. Kui esimese valimi vaatlus on väiksem, siis liidetakse loendisse 1, võrdsete vaatlustulemuste korral liidetakse loendisse 0,5. Väiksemate vaatluste loend annab vahestatistiku U_-. Suuremate väärtuste loendi vahestatistiku U_+ arvutamiseks liidetakse vaatlusest suuremate või võrdsete vaatluste arv teises valimis. Võrdsete mõõtmistulemuste korral kasutatakse taas koefitsienti 0,5. *U* statistikuna kasutatakse nendest kahest arvust väiksemat. Ei oma tähtsust, kumba valimit nimetada esimeseks või teiseks. Arvutuste kontrollimiseks võib arvestada, et $U_- + U_+ = n_1 \cdot n_2$. *U* statistiku väärtuse võrdlemiseks kriitiliste väärtustega võib kasutada spetsiaalseid tabeleid (näiteks Tiit 1972). Sisuka hüpoteesi saab vastu võtta juhul, kui leitud *U* statistik on kriitilisest väärtusest väiksem. *U* testi võib kasutada ka sõltuvate valimite korral, kuid sellisel juhul on see palju väiksema võimsusega kui märgitest.

Wald-Wolfowitzi test

Wald-Wolfowitzi seeriatest (*runs test*) on sõltumatute valimite *t* testi teine mitteparameetriline analoog, mis sobib kasutamiseks igasuguste kaheväärtuseliste vaatluste puhul, mida saab teise tunnuse alusel jadasse järjestada. Test võimaldab kontrollida, kas väärtused paiknevad jadas (mingil viisil järjestatud reas) juhuslikult või mitte. Juhusliku paiknemise korral on paljude samasuguste väärtuste kõrvuti paiknemine vähetõenäoline, samuti on vähetõenäolised väärtuste korrapärased paiknemismustrid. Mitmeväärtuselise tunnuse saab alati klassifitseerida kaheväärtuseliseks ja kontrollida sellise jaotuse statistilist seost jada järjestuse suhtes. Kaheväärtuselise tunnusena võib käsitleda ka vaatluse kuuluvust ühte või teise valimisse.

Kui jada pikkus on *n* elementi (n_1 ühe väärtusega ja n_2 teise väärtusega), siis seeriate arv (*N*) selles võib muutuda vahemikus 2 kuni *n*. Pika vaatlusrea ning seeriate juhusliku arvu korral peaks seeriate arv olema normaaljaotusega, mis võimaldab kasutada *Z* statistikut. Seeriate arvu keskmine m_N arvutatakse valemist

$$m_N = \frac{2n_1n_2}{n_1+n_2} + 1; \qquad [1\text{-}49]$$

ja seeriate arvu standardhälve

$$s_N = \sqrt{\frac{2n_1n_2\left(2n_1n_2 - n_1 - n_2\right)}{\left(n_1+n_2\right)^2\left(n_1+n_2-1\right)}} \qquad [1\text{-}50]$$

ning *Z* statistiku saab leida

$$Z_N = \frac{N - m_N}{s_N}. \qquad [1\text{-}51]$$

Z statistiku kriitilised väärtused leitakse standardiseeritud normaaljaotuse tabelitest või selle tõenäosusfunktsiooni arvutava tarkvara abil.

1.4.3.2. Protsentide võrdlemine

Protsent on suhtarv, seega kaheväärtuselise tunnuse karakteristik. Korduvalt mõõdetud kaheväärtuselisi tunnuseid kirjeldatakse binoomjaotusega. Kui mõõtmiste arv on piisavalt suur (>60) ja protsent ei ole väga väike (<10%) ega väga suur (>90%), siis võib eeldada, et selle protsendi viga on normaaljaotusega ja usalduspiiride arvutamisel võib kasutada ka eeltoodud vahendeid. Seejuures Bernoulli muutuja dispersioon arvutatakse valemist *DX = p(1–p)* ja standardhälve on ruutjuur dispersioonist.

Binoomjaotusega muutujate (protsentide) erinevuse kontrollimisel ei saa enam lähtuda Bernoulli jaotusest, mille puhul vaatluste arv (*n*) võrdub ühega. Suurte valimite põhjal arvutatakse protsentide erinevuse standardhälve *s** järgmise valemi järgi. Edasine erinevuse kontroll toimub normaaljaotuse abil.

$$s^* = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \qquad [1\text{-}52]$$

Väikese valimi puhul ei ole normaaljaotuse kasutamine keskmiste erinevuse kontrollimisel üldiselt põhjendatud. Selle kasutamine võib vääralt viidata kõrgele usaldusnivoole. On olemas teisendus (Fisheri φ-teisendus), mille kasutamisel läheneb jaotus valimi kasvades normaaljaotusele hoopis kiiremini kui teisendamata protsentide puhul. Kui normaaljaotuse kasutamiseks peaks binoomjaotusega tunnuse valimi maht olema vähemalt 60, siis Fisheri φ-teisenduse kasutamisel piisab keskmiste võrdlemiseks normaaljaotuse abil umbes kümnest vaatlusest. φ-teisendus arvutatakse valemi järgi

$$\varphi = 2\arcsin\sqrt{p}, \qquad [1\text{-}53]$$

kus *p* tähistab katsetulemuse ühe variandi (eduka tulemuse) tõenäosust. Olulisuse hindamiseks kasutatav statistik *Z* arvutatakse valemist

$$Z = \frac{\varphi_1 - \varphi_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}. \qquad [1\text{-}54]$$

Edasine erinevuse kontroll toimub *Z* testi abil.

1.4.3.3. Jaotuste võrdlemine

Eelnevalt käsitleti kahe üldkogumi keskväärtuste erinevuse testimist. Erinevus üldkogumite vahel ei pruugi aga väljenduda keskväärtuses. Kahe tunnuse keskmised võivad küll samad olla, aga jaotuste kuju on erinev. Jaotuste võrdlemise meetodid aitavad leida ja tõestada faktorite mõjusid ning kontrollida andmete jaotustüübi vastavust eeldatavale, näiteks normaaljaotusele.

Šansside suhe

Šansside suhet (*odds ratio – OR*) kasutatakse üldjuhul kahe binaarse tunnuse vahelise seose kirjeldamiseks. Šanssi määratletakse kui võimalike variantide (nähtuse esinemise ja puudumise) vahekorda; näiteks emasloomade arv isaste arvu suhtes. Šansside suhte kasutamisest klassifikatsioonitäpsuse hindamisel on juttu peatükis 2.3.7.4.

Kaheväärtuselise tunnuse puhul saab šansside suhte arvutada 2 × 2 sagedustabelist selle ühte pidi diagonaalil olevate vaatluste arvu või nende osakaalude (p_1, q_2) korrutise ja teistpidi diagonaali olevate vaatluste arvu või osakaalude (p_2, q_1) korrutise suhtena.

$$OR = \frac{p_1 / q_1}{p_2 / q_2} = \frac{p_1 \cdot q_2}{p_2 \cdot q_1}. \qquad [1\text{-}55]$$

Kuna šansside suhte logaritm ehk logit on ligikaudu normaaljaotusega, siis saab šansside suhte usalduspiirid leida selle standardvea abil: $1{,}96 \cdot SE(\ln(OR))$. Risttabelist lahtritega *a, b, c, d* saab šansside suhte standardviga (*SE*) hinnata valemiga

$$SE(\ln(OR)) = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}. \qquad [1\text{-}56]$$

χ² test

Empiirilist ja teoreetilist jaotust võrdlevaid teste nimetatakse **sobivustestideks** (*goodness of fit test*). Täisarvuliste muutujate sobivustestina kasutatakse kõige enam χ^2 testi (hii-ruut test). χ^2 jaotusega on standardse normaaljaotusega muutuja ruut. χ^2 test võrdleb mingite nähtuste loendamise tulemusel saadud sagedusjaotusi jaotustega, mis oleksid oodatavad tunnustevaheliste seoste puudumisel (H_0). Kuna pidevaid tunnuseid saab jagada väärtusklassidesse, on test rakendatav ka arvuliste ja järjestatud tunnuste korral, kuid klassipiiride määramise suvalisuse tõttu seda ei soovitata.

χ^2 testi kasutamisel tuleks jälgida, et ühtegi oodatava jaotuse klassi ei langeks liiga vähe (<5) vaatlusi. Vastasel juhul ei ole testi tulemused usaldatavad, sest nullsageduste korral tekib χ^2 statistiku arvutamisel nulliga jagamine ehk määramatus. Samuti eeldab χ^2 test, et kõik üksikvaatlused on üksteisest sõltumatud.

Hii-ruut testi kasutamiseks peab valimi suurus olema piisav, et nullhüpoteesi korral oodatavas jaotuses jääks igasse kategooriasse vähemalt 5 vaatlust

χ^2 testi saab kasutada nii kahe empiirilise jaotuse kui ka empiirilise jaotuse ja teoreetilise jaotuse erinevuse olulisuse kontrollimiseks. Sageduste summa teoreetilises ehk oodatavas jaotuses võrdub nihketa hinnangu puhul sama summaga empiirilises jaotuses. Kui χ^2 testiga soovitakse kontrollida vaid kahe empiirilise jaotuse kuju erinevust, tuleb võrreldavate valimite mahud võrdseteks normeerida. Esimese valimi mahu normeerimiseks teise valimi mahule vastavaks tuleb sagedused esimeses jaotuses korrutada teise ja esimese jaotuse summade suhtega.

χ^2 testi kasutamiseks määratakse iga klassi kohta vaadeldud väärtuste esinemiskordade arv (empiiriline sagedus – F_{emp}) ja teoreetiline sagedus, mis on oodatav sagedus eeldades nullhüpoteesi, et mõlemad valimid on juhuslik väljavõte ühest ja samast üldkogumist (oodatav sagedus – F_{exp}). Teisisõnu, teoreetiline sagedus on oodatav sagedus juhul, kui vaadeldavad tunnused teineteisest ei sõltu. Oodatava sageduse arvutamine lähtub tõenäosuste korrutamise reeglist: sõltumatute juhuslike sündmuste koosesinemise tõenäosus võrdub üksiksündmuste tõenäosuste korrutisega. Sagedustabeli puhul tähendab see, et oodatav sagedus tabeli igas lahtris võrdub lahtrile vastav reasumma korrutatud

vastava veerusummaga ja jagatud vaatluste üldarvuga. Oodatava ja tegeliku jaotuse vahelise erinevuse olulisuse hindamisel arvutatakse χ^2 statistik valemist

$$\chi^2 = \sum_j^k \frac{\left(F_{emp} - F_{\exp}\right)^2}{F_{\exp}}, \qquad [1\text{-}57]$$

mida kahe empiirilise jaotuse vahelise erinevuse võrdlemisel võib väljendada ka järgmiselt:

$$\chi^2 = \sum_j^k \frac{\left(F_{emp1} - F_{\exp}\right)^2}{F_{\exp}} + \sum_j^k \frac{\left(F_{emp2} - F_{\exp}\right)^2}{F_{\exp}}, \qquad [1\text{-}58]$$

kus k on väärtusklasside arv, j on väärtusklassi indeks, F_{exp} valemis [1-58] tähistab oodatavate sageduste jaotust eeldusel, et empiirilised jaotused on pärit samast üldkogumist.

Kui sagedusjaotus vastab täpselt nullhüpoteesi korral oodatavale, siis $\chi^2 = 0$. Kui χ^2 statistik on suurem kui *df* vabadusastmega χ^2 jaotuse täiendkvantiil, tuleb vastu võtta sisukas hüpotees, et jaotused on erinevad. Kui χ^2 on väiksem kriitilisest väärtusest, siis tuleb jääda nullhüpoteesi juurde. Vabadusastmete arv (*df*) avaldub sagedustabeli ridade arvu k ja veergude arvu m kaudu ja ei ütle otseselt midagi seose tugevuse kohta, statistilise seose tugevuse mõõtmiseks on mitmesugused seosekordajad.

$$df = (k-1)\cdot(m-1) \qquad [1\text{-}59]$$

Eeltoodu oli Pearsoni pakutud klassikaline χ^2 test, mis lähtub ruuthälbe minimeerimise nõudest. χ^2 statistikut saab arvutada ka lähtudes **tõepärasuhte** statistikust (*likelihood ratio*). Testi nimetatakse **G testiks** ja see sisaldab nullmudelist saadud ja katsest saadud suhteliste sageduste (P_{exp} ja P_{emp}) suhet.

$$G = 2\sum_j^k P_{emp} \ln \frac{P_{emp}}{P_{\exp}} \qquad [1\text{-}60]$$

Statistik G on ligikaudu χ^2 jaotusega, mille vabadusastmete arvuks on sõltumatute vaatluste arv k.

Kolmogorov-Smirnovi test

Kolmogorov-Smirnovi testi saab kasutada pidevate tunnuste võrdluseks. Valimid võivad seejuures olla väiksemad kui χ^2 testil. Test võrdleb jaotusfunktsioonide maksimaalset erinevust. Tuletame meelde, et jaotusfunktsiooni väärtuseks kohal x on tõenäosus, et tunnus omab sellest argumendist väiksemat või võrdset väärtust. Kui eelnevalt ei ole teada, kumb jaotusfunktsioon kasvab kiiremini, tuleb otsuse langetamiseks leida empiiriliste jaotusfunktsioonide maksimaalse erinevuse absoluutväärtus ja võrrelda seda kriitilise väärtusega. Kui leitud erinevus D on väiksem kui kriitiline väärtus, tuleb jääda nullhüpoteesi juurde ja tõdeda, et ei ole piisavat alust arvata, et jaotused oleksid erinevad. Kui on põhjust arvata, et ühes kindlas üldkogumis on suuremad väärtused sagedamad kui teises, võib kasutada ühepoolset sisukat hüpoteesi ja jaotusfunktsioonide maksimaalse erinevuse absoluutväärtuse asemel jaotusfunktsioonide maksimaalset vahet.

Dispersioonide võrdlemine

Suurte valimite puhul kasutatakse dispersioonide erinevuse usaldusnivoo määramiseks **F testi**. *F* statistik arvutatakse järgmise valemi järgi

$$F = \frac{s_1^2}{s_2^2}, \qquad [1\text{-}61]$$

kus s_1 on ühe valimi standardhälve, s_2 on teise valimi standardhälve, kusjuures $s_1^2 \geq s_2^2$.

F statistiku väärtust võrreldakse kriitilise väärtusega, vabadusastmete arvuks võetakse (n_1+n_2-2). Dispersioonide võrdlemisele on rajatud dispersioonanalüüs (ANOVA), mida käsitletakse omaette peatükis (ptk 1.5.4).

Tarkvara

Mitmeid statistilisi teste, sealhulgas mitteparameetrilisi, saab teha Vassar College veebilehel (http://vassarstats.net) ja California ülikooli Java tarkvara kasutaval veebilehel *Statistics Online Computational Resource* – SOCR (http://www.socr.ucla.edu). Microsoft Excelil on statistilise analüüsi lisapakett Data Analysis Toolbox. Kui seda menüü tööriistade osas ei ole, tuleb see lisamoodulitest valida.

Märgitestiks sobib ka Exceli binoomjaotuse funktsioon. Selleks tuleks lisaks vaatluspaaride arvule *n* leida harvemini esinenud muutuste arv *m* ja kasutada neid funktsiooniga BINOMDIST(m; n; 0,5; TRUE), mis annab domineeriva muutuse olulisuse tõenäosuse ühepoolse hüpoteesi korral (teistpidi muutuse domineerimine oli välistatud). Positiivsete muutuste arvu saab leida Exceli funktsiooniga IF(tingimus, 1, 0), mille tulemus on 1, kui tingimus on täidetud. Tingimuseks võib olla, et ühe veeru lahtris on väiksem arv kui teise veeru vastavas lahtris. Võrdsete väärtustega vaatluspaarid peaksid arvesse minema kaaluga 0,5.

1.5. Andmeanalüüs

1.5.1. Juhuslik vektor

Mitmest juhuslikust tunnusest koosnevat komplekti nimetatakse juhuslikuks vektoriks. Iga komponendi jaotust eraldi nimetatakse **marginaaljaotuseks** ehk äärejaotuseks, kuna sagedustabelis paikneb see äärmises veerus. Vektori enda jaotust nimetatakse aga komponentide **ühisjaotuseks**. Ühisjaotus sisaldab eneses lisaks kõigi komponentide marginaaljaotustele veel teavet komponentide omavaheliste sõltuvuste kohta. Andmed diskreetse kahemõõtmelise juhusliku vektori kohta esitatakse tavaliselt **risttabeli** kujul (tabel 2). Risttabel võib sisaldada tõenäosusi, summasid, sagedusi või muid ühisjaotuse parameetreid. **Sagedustabel** on risttabeli erijuht, kus esitatakse ühisjaotuse väärtuste esinemissagedusi. **Jaotustabelis** esitatakse ühisjaotuse väärtuste tõenäosusi.

Tabel 2. Kahemõõtmelise vektori kujutamine jaotustabelis. $x_1 \ldots x_k$ ja $y_1 \ldots y_h$ – väärtused teljel X ja Y, $p_{11} \ldots p_{kh}$ – väärtuskombinatsioonide tõenäosused.

$X \setminus Y$	y_1	y_2	...	y_h	P_X
x_1	p_{11}	p_{12}	...	p_{1h}	$p_{1.}$
x_2	p_{21}	p_{22}	...	p_{2h}	$p_{2.}$
...	...	...	...	...	...
x_k	p_{k1}	p_{k2}	...	p_{kh}	$p_{k.}$
P_Y	$p_{.1}$	$p_{.2}$	...	$p_{.h}$	1

Tabelil on päis, mis selgitab tabeli sisu. Veeru i ja rea j ristumiskohas paikneb tõenäosus (p_{ij}), et juhuslik vektor (X,Y) on väärtusega (x_i , y_j). Kusjuures $i = 1, ..., k$ ja $j = 1, ..., h$. Tabeli parempoolses veerus on komponendi X ja alumises reas komponendi Y jaotus. Ühisjaotuse suhtes on need äärejaotused.

Kui ühe jaotuse komponendi jaotus on mingil moel fikseeritud (on määratud mingi tingimus), siis teise komponendi jaotus on **tinglik jaotus**. Tõenäosust, et $X = x_i$ tingimusel, et $Y = y_j$ tähistatakse $P(X = x_i \mid Y = y_j)$ ja see võrdub jaotustabelis $p_{ij} / p_{.j}$. See tähendab, ühisjaotuse komponendi tõenäosuse ja marginaaljaotuse vastava väärtuse suhtega.

Juhusliku vektori uurimise eesmärgiks on enamasti selle komponentide vahelise sõltuvuse kirjeldamine või seose olemasolu tõestamine. Juhuslik suurus X ei sõltu juhuslikust suurusest Y, kui X kõik tinglikud jaotused on omavahel võrdsed. Juhuslikud suurused X ja Y on sõltumatud parajasti siis, kui juhusliku vektori (X,Y) tõenäosusfunktsioon avaldub marginaalsete tõenäosusfunktsioonide korrutisena. Jaotuste kaudu avaldades: juhuslikud suurused on sõltumatud, kui juhusliku vektori (X,Y) jaotus P_{XY} võrdub marginaaljaotuste korrutisega.

$$P_{xy} = P_X \cdot P_Y \qquad [1\text{-}62]$$

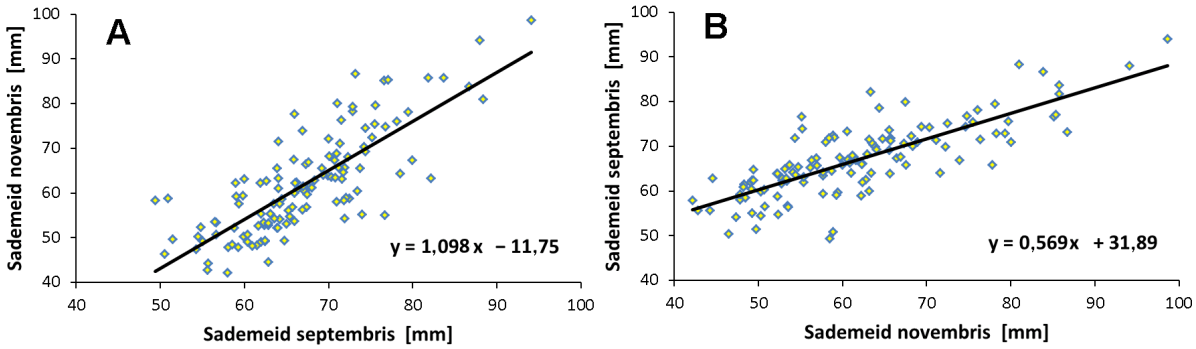
Juhuslike suuruste sõltumatus on vastastikune. Kui juhusliku vektori komponendid on sõltumatud, siis määravad marginaaljaotused ühisjaotuse täielikult. Juhuslike suuruste vaheline sõltuvus on seda tugevam**,** mida täpsemalt võimaldab ühe suuruste väärtuste teadmine määrata teise juhusliku suuruse väärtusi. Sõltumatute suuruste puhul ei lisa ühe muutuja väärtuste teadmine mingit teavet teise muutuja väärtuste kohta.

1.5.2. Seosekordajad

Juhuslike suuruste omavahelise sõltuvuse iseloomustamiseks kasutatakse mitmesuguseid **seosekordajaid**. **Sõltumatute tunnuste** puhul ei ole võimalik ennustada ühe tunnuse väärtuste järgi teise tunnuse tõenäolisi väärtuseid. **Sõltuvate tunnuste** puhul ei ole ühe tunnuse väärtused teise tunnuse väärtuste suhtes täiesti juhuslikud ehk esineb seos. Seega, teades ühe tunnuse väärtuseid on võimalik teha ennustus teise tunnuse väärtuste kohta. Kahe arvtunnuse vahelist statistilist seost ehk **korrelatsiooni** mõõdetakse ja kirjeldatakse mitmesuguste seosekordajatega. Tulemuste interpreteerimisel on seejuures oluline meeles pidada, et korrelatsioon ei pretendeeri põhjuslikkuse kirjeldamisele. Võimalik on ka, et tunnused on sama-aegselt sõltuvad aga ei ole korreleeritud.

Kahe tunnuse ühisjaotuse graafilist esitust nimetatakse **korrelatsioonivälja.** Korrelatsioonivälja võib kujutada punktdiagrammi kujul hajuvusdiagrammina, regressioonielipsina, isoliinide ehk samajoontega või erinevate värvitoonidega. Ühe pideva tunnuse (funktsioontunnuse) oodatava väärtuse sõltuvust teise pideva tunnuse (argumenttunnuse) väärtusest nimetatakse **regressiooniks**. Regressioon on ühepoolne seos (joonis 1-13). Regressiooniseost saab kirjeldada mudeliga, mida võib omakorda esitada võrrandi kujul, seost iseloomustava joonena, mida nimetatakse **regressioonijooneks** või **regressioonipinnana**. Regressioonimudel sobitatakse vaatlusandmetele nii, et funktsioontunnuse väärtused hälbiksid mudelist võimalikult vähe ja hälvete keskväärtus oleks null.

Korrelatsioon on kahepoolne seos, regressioon ühepoolne; kumbki ei pruugi väljenda põhjuslikku seost



Joonis 1-13. Seosed septembri ja novembri paljuaastase keskmise sademete hulga vahel Eesti ilmavaatlusjaamades korrelatsioonivälja ja lineaarse seose regressioonijoon. A – novembrikuu sademete hulga sõltuvus septembri sademete hulgast, B – septembri sademete hulga sõltuvus novembri sademete hulgast. Andmed Jaagus *et al*. (2010). Kuna lähteandmed on paljuaastased keskmised, siis ei saa nende seoste alusel teha järeldusi üksikaasta kohta, vaid ilmavaatlusjaamade kohta. Kohad, mis on sademeterikkad novembris, on suhteliselt sademeterikkad ka septembris. Regressioonijoone tõus ja mudeli parameetrid on erinevad, sest novembri keskmine sademete hulk varieerub vaatlusjaamade vahel enam kui septembri keskmine.

1.5.2.1. Kovariatsioon

Kovariatsioon on kahe muutuja keskmisest hälbimise korrutise ootus. Kui juhuslike suuruste X ja Y kovariatsioon erineb nullist, siis öeldakse, et need juhuslikud suurused on korreleeritud. Kovariatsioon ja korrelatsioon võivad olla nii positiivsed kui negatiivsed. Kovariatsiooni suurus sõltub nii kummagi üksikmuutuja suuruste hajuvusest keskväärtuse ümber kui ka hajuvuste samasuunalisusest. Kui X suurenedes Y keskmiselt suureneb, siis on X ja Y kovariatsioon positiivne. Kovariatsiooni muutujate x ja y vahel arvutatakse järgmise valemi järgi

$$\mathrm{cov}(x, y) = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}. \qquad [1\text{-}63]$$

1.5.2.2. Korrelatsioonikordaja

Korrelatsioonikordaja ehk standardhälvetega normeeritud kovariatsioon on juhusliku vektori komponentide kovariatsiooni suhe nende komponentide standardhälvete korrutisse. Tänu normeerimisele on korrelatsioonikordaja alati vahemikus –1…+1 ja erinevatest andmetest arvutatud korrelatsioonikoefitsiendid on omavahel võrreldavad. Korrelatsioonikordaja tähistus on *r* või *R*.

$$R = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2 \sum_i^n (y_i - \bar{y})^2}} = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y} \qquad [1\text{-}64]$$

Pearsoni korrelatsioonikordaja (*Pearson's product moment correlation*) mõõdab normaaljaotusega juhuslike muutujate vastastikust lineaarset sõltuvust. Kui suurused muutuvad üldiselt samasuunaliselt, on tegemist positiivse, ning kui vastassuunaliselt, siis negatiivse korrelatiivse sõltuvusega. Korrelatsioonikordaja absoluutväärtus näitab lineaarse seose tugevust. Mida väiksem on punktide hajuvus regressioonisirge ümber, seda lähedasem on korrelatsioonikordaja absoluutväärtus ühele. Seose tugevuse verbaalsel iseloomustamisel nimetatakse seost nõrgaks juhul, kui $|R| \leq 0{,}3$; keskmiseks, kui $0{,}3 < |R| < 0{,}7$ ja tugevaks, kui $|R| \geq 0{,}7$.

Pearsoni korrelatsioonikordaja näitab lineaarse seose tugevust ja suunda kahe normaaljaotusega tunnuse vahel

Lineaarne korrelatsioonikordaja on kergesti mõjutatav üksikute tugevasti kõrvalekalduvate mõõtmistulemuste (erindite) poolt. Korrelatsioonikordaja usalduspiiride ja seose olulisuse hindamiseks korrelatsioonikordaja alusel peaksid uuritavad tunnused olema normaaljaotusega. Kui on põhjust arvata, et tunnused ei ole normaaljaotusega, tuleks seose tugevuse hindamiseks kasutada astakkorrelatsiooni kordajat või mõnda teist mitteparameetrilist meetodit (ptk 1.5.2.3).

Korrelatsioonikordaja olulisus ja usalduspiirid

Korrelatsioonikordaja olulisuse ja usalduspiiride arvutamine eeldab tunnuste normaaljaotust. Mittenormaaljaotusega tunnuse korral võib korrelatsioonikordajat kasutada seose tugevuse iseloomustamiseks, kuid mitte hüpoteeside tõestamiseks. Iteratiivsete meetoditega on siiski näidatud, et valimi piisava suuruse korral (>100 vaatluse) andmete normaaljaotuse puudumine enamasti ei mõjuta tulemust märkimisväärselt. Korrelatsioonikordaja usutavust võib olulisemalt mõjutada valimi heterogeensus, kolmandate (varjatud) faktorite mõju ja seose mittelineaarsus. Siinkohal tuleks veelkord rõhutada, et korrelatsioon ei pruugi üldsegi tähendada tunnustevahelise põhjusliku seose olemasolu. Mõlema tunnuse omavahel korreleeruvad väärtused võivad olla tingitud mingist ühisest, aga uuringusse kaasamata põhjusest.

Nullhüpoteesi, mille kohaselt tunnused ei ole omavahel korreleeritud, on võimalik kõrvale lükata ja sisukat hüpoteesi, mille kohaselt tunnused on omavahel korreleeritud, saab korrelatsioonikordaja alusel kontrollida juhul, kui tunnuste ühisjaotus on kahemõõtmeline normaaljaotus. Selleks tuleb arvutada kas *F* statistik või *T* statistik.

$$F = \frac{(n-2)R^2}{1-R^2} \quad [1\text{-}69]$$

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \quad [1\text{-}70]$$

R tähistab siin korrelatsioonikordajat ja n vaatluste arvu. Kui $|T| \geq t_{\alpha/2,n-2}$ (t jaotuse α/2 täiendkvantiili vabadusastmete arvuga n–2) või $|F| \geq F_{\alpha/2,n-2}$ (F jaotuse α/2 täiendkvantiili vabadusastmete arvuga 1 ja n–2). Samuti võib nullhüpoteesi lugeda ümberlükatuks, kui valimi järgi leitud korrelatsioonikordaja usalduspiiride sisse ei jää null.

Korrelatsioonikordaja statistiline olulisus sõltub vaatluste arvust ja korrelatsioonikordaja väärtusest. Seos on statistiliselt olulisem korrelatsiooni kordaja suurema absoluutväärtuse ja vaatluste suurema arvu korral.

Korrelatsioonikordaja statistiline olulisus sõltub vaatluste arvust ja korrelatsioonikordaja väärtusest; muutujate normaaljaotus on eeldus

Valimi põhjal arvutatud korrelatsioonikordaja jaotuse lähendamiseks normaaljaotusele arvutatakse Fisheri *z*-teisendus

$$z = \frac{1}{2}\ln\frac{1+R}{1-R}. \quad [1\text{-}65]$$

Korrelatsioonikordaja usalduspiiride arvutamiseks leitakse valimist arvutatud korrelatsioonikordajale vastav ülaltoodud Fisheri z statistiku väärtus. Seejärel leitakse z statistiku usalduspiirid alltoodud valemite abil.

$$z_{\_} = z - Z_{\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}} \quad [1\text{-}66]$$

$$z^{-} = z + Z_{\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}} \quad [1\text{-}67]$$

$Z_{\alpha/2}$ tähistab siin normaaljaotuse α/2 täiendkvantiili ja α tähistab usaldusnivood. Saadud usalduspiirid teisendatakse Fisheri pöördteisenduse abil üldkogumi korrelatsioonikordaja usalduspiirideks. Usaldusnivoo 0,05 juures on $Z_{\alpha/2} = 1{,}96$. Fisheri teisenduse pöördteisendus on

$$R = \frac{e^{2z}-1}{e^{2z}+1}. \quad [1\text{-}68]$$

Determinatsioonikordaja

Lineaarse korrelatsioonikordaja ruutu (r^2 või R^2) nimetatakse **determinatsioonikordajaks**. See näitab, kui suure osa tunnuste koguvarieeruvusest moodustab nende ühine varieeruvus. Ühise varieeruvuse osa võib olla statistilise seose tugevuse mõõduks.

Determinatsioonikordaja näitab, kui suure osa ühe tunnuse varieeruvusest kirjeldab teise tunnuse varieeruvus

1.5.2.3. Mittelineaarse seose tugevuse mõõtmine

Pearsoni *R* kaudu saab hinnata vaid normaaljaotusega lineaarse või lineaarseks teisendatud seose tugevust ja olemasolu. Mittenormaaljaotusega monotoonse (ühtlaselt kasvava või kahaneva) seose puhul võib kasutada mitteparameetrilisi korrelatsioonikordajaid (Spearmani ρ, Kendalli τ).

Mittemonotoonse seose tugevuse hindamiseks tuleks leida seost kõige paremini kirjeldav mudel ning seejärel mõõta mudeli ja tegelikkuse vastavust mõne sobivustestiga. Kvalitatiivsete tunnuste vahelist seost saab mõõta ja võrrelda Tšuprovi ja Guttmanni kordajatega, kuid viimased ei hinda seose olulisust. Kvalitatiivsete tunnuste vahelise seose olulisuse hindamiseks on vaatluste klassidesse jaotumisel põhinevad sobivustestid, näiteks χ^2 test.

Spearmani astakkorrelatsioonikordaja

Spearmani astakkorrelatsioonikordaja ρ (roo) kuulub, nagu järgnevadki, mitteparameetriliste seose-kordajate hulka ja ei sea eeldusi tunnuste jaotusele. Nii nagu teised mitteparameetrilised meetodid, vähendab ka astakkorrelatsioonikordaja erindite mõju ning võimaldab mõõta ka mittelineaarsete monotoonse tendentsiga seoste tugevust. Kordaja arvutamiseks tuleb mõlema tunnuse väärtused eraldi järjestada ja määrata iga väärtuse järjekorranumber ehk astak. Astakute väärtustest leitakse korrelatsioonikordaja. Juhul, kui korduvaid mõõtmistulemusi ei ole, saab kasutada valemit

$$\rho = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}, \qquad [1\text{-}71]$$

kus d on ühes paaris olevate väärtuste järjekorranumbrite vahe, n on väärtuspaaride arv.

Astakkorrelatsioonikordaja mõõdab seose monotoonsust – kas ühe tunnuse suurenedes suureneb ka teise tunnuse väärtus.

Kendalli hälbimissuundade korrelatsioonikordaja

Kendalli τ (tau) näitab kahe tunnuse võrdluses sama- ja vastassuunaliste vaatluspaaride suhtelise sageduse erinevust. Paarid moodustatakse vaatluste kõigi omavaheliste kombinatsioonide vahel. Kendalli τ on, nagu ka ρ, kasutatav nii arvtunnuste kui ka järjestatavate nominaaltunnuste korral. Kui nii ühel kui ka teisel tunnusel ei ole kokkulangevaid väärtusi, siis

$$\tau = \frac{n_s - n_v}{N}. \qquad [1\text{-}72]$$

Kui kokkulangevaid väärtusi esineb, siis

$$\tau = \frac{n_s - n_v}{\sqrt{\left(n_s + n_v - n_y\right)\left(n_s + n_v - n_x\right)}}, \qquad [1\text{-}73]$$

kus n_s on samasuunaliste paaride arv, n_v on vastassuunaliste paaride arv, n_y on ühe tunnuse kokkulangevate väärtustega paaride arv, n_x on teise tunnuse kokkulangevate väärtustega paaride arv, N on paaride koguarv.

Kendalli tau väärtused on vahemikus −1 kuni +1, juhuslikkuse korral oodatav väärtus on 0. Kui Spearmani ρ on astakutest arvutatud korrelatsioonikordaja, mis mõõdab monotoonse seose tugevust, siis Kendalli tau väljendab eelkõige tõenäosust – tõenäosuse, et kahe võrreldava tunnuse väärtused on

samamoodi järjestatud ja tõenäosuse, et nad on erinevalt järjestatud, vahet. Kendalli korrelatsioonikordaja olulisuse kontrollimiseks saab kasutada statistikut *Z*, mis on nullhüpoteesi kehtides standardiseeritud normaaljaotusega ja mida saab teades valimi mahtu *n* arvutada järgmiselt.

$$Z = \frac{\tau}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}} \qquad [1\text{-}74]$$

Tšuprovi kordaja

Nagu eelpool märgitud, ei sobi χ^2 test kahe tunnuse vahelise seose tugevuse kirjeldamiseks. χ^2 statistik sõltub väärtuskombinatsioonide hulgast (tabeli suurusest) ja vaatluste arvust ning seetõttu ei ole erinevatest tabelitest leitud χ^2 statistikud võrreldavad. **Tšuprovi kordaja** kasutab χ^2 statistikut vaatluste arvuga (*n*) ning χ^2 arvutamise tabeli ridade (*m*) ja veergude arvuga (*k*) normeeritud kujul

$$T = \sqrt{\frac{\chi^2}{n\sqrt{(m-1)(k-1)}}}. \qquad [1\text{-}75]$$

Tšuprovi kordaja sobib seose tugevuse mõõtmiseks kõigil juhtudel, kui χ^2 testi kasutamine on põhjendatud. Nominaalse seose mõõtmiseks kasutatakse ka χ^2 statistiku ja vabadusastmete arvu suhet.

Goodman-Kruskali λ

Goodman-Kruskali seosetugevuse kordaja ehk Goodmani ja Kruskali λ (lambda) sobib kasutamiseks nii väikese väärtustehulgaga arvtunnuste, järjestustunnuste kui ka nominaaltunnuste korral. Pidevaid arvtunnuseid saab kasutada vaid klassifitseeritud kujul. Lähtub suurima tõepära põhimõttest – funktsioontunnuse väärtuseks prognoositakse argumenttunnuse iga väärtuse korral see väärtus, mille esinemise sagedus on kõige suurem (lokaalne mood). Kui risttabeli ridades on argumenttunnuse klassid ja veergudes funktsioontunnuse klassid, siis lokaalne mood on tabeli iga rea see lahter, milles vaatluste arv on suurim.

$$\lambda = \frac{\sum_i^m n_{LM} - n_M}{n - n_M}, \qquad [1\text{-}76]$$

kus *n* on tunnuspaaride arv, n_M on funktsioontunnuse moodi esinemissagedus, n_{LM} on funktsioontunnuse lokaalse moodi esinemissagedus, *i* on argumenttunnuse klassi indeks ja *m* on argumenttunnuse klasside arv.

Lamda mõõdab prognoosivigade osakaalu vähenemist tänu argumenttunnuse kasutamisele ja on seega asümmeetriline seosekordaja. Juhuslikkuse korral, kui argumenttunnuse kasutamine ei vähenda prognoosivigu, on λ väärtuseks null. Kui argumenttunnus võimaldab funktsioontunnuse üheselt määrata, on λ väärtuseks 1. Kui funktsioontunnusel puudub mood, siis ei saa λ arvutada. λ ei ole kunagi negatiivne.

1.5.2.4. Korrelatsioonimaatriks

Rohkem kui kahe tunnuse omavaheliste seoste tugevuse või olulisuse võrdlemiseks võib seosekordajad koondada **korrelatsioonimaatriksisse**. Korrelatsioonimaatriks on turniiritabelit meenutav tabel, mille nii veergude kui ridade päises on võrreldavad tunnused. Maatriksi lahtrites on korrelatsiooni tugevust või olulisuse tõenäosust iseloomustavad arvud.

1.5.3. Regressioonanalüüs

Regressioonanalüüs käsitleb ühe muutuja (funktsioontunnuse) väärtuste prognoosimist teise tunnuse (argumenttunnuse) väärtuste järgi. Funktsioontunnuse väärtuste arvutamise eeskirja nimetatakse funktsiooniks. Tähistatakse $y = f(x)$ ja loetakse: tunnus y on tunnuse x funktsioon. Prognoosimiseks kasutatava valemi ehk seose kuju ehk **regressioonimudeli** valiku langetab enamasti uurija oma kogemuste põhjal subjektiivselt Lähtuda võib seejuures nii sisulistest kui arvutustehnilistest kaalutlustest. Mudeli andmetele sobitamiseks ehk mudeli parameetrite leidmiseks on arvutusmeetodid. Lihtsa lineaarse regressiooni korral on funktsioontunnus pidev ja seos lineaarne. Regressioonimudel võib aga sisaldada ka mitut funktsioon- ja argumenttunnust, samuti võib prognoositavaks tunnuseks olla kaheväärtuselise muutuja ühe või teise variandi esinemise tõenäosus. Klassikalise regressioonanalüüsi eeldused on:

- iga argumenttunnuse väärtuse korral peab funktsioontunnus olema normaaljaotusega ehk kõik tinglikud jaotused on normaaljaotused;
- funktsioontunnuse tinglike jaotuste keskväärtused peavad olema argumenttunnuse funktsioonid $E(Y|x)=f(x)$;
- funktsioontunnuse dispersioon peab olema konstantne;
- kui funktsioontunnuse tinglikud väärtused sõltuvad tundmatutest parameetritest, peab funktsioon olema nende parameetrite suhtes lineaarne.

Regressioonanalüüsile võib järgneda regressioonijääkide analüüs, mis võimaldab selgitada, miks jäi osa varieeruvusest kirjeldamata – kas ei ole regressioonimudeli kuju sobilik või mõjuvad mingid mudelist välja jäänud faktorid. Kui regressioonanalüüsi eeldused on täidetud, peaks prognoosijääkide keskväärtus kõigi argumenttunnuse väärtuste korral olema 0. Prognoosijääke võib standardiseerida, see võimaldab hinnata tõenäosust, et sellise suurusega prognoosijääk võis tekkida juhuslikult.

Kui eesmärgiks on vaid tunnustevahelise seose kirjeldamine, siis ei ole prognoosijääkide normaaljaotuse nõude jälgimine tingimata vajalik. Ka normaaljaotuse puudumisel annab regressioonanalüüs prognoosi, kuid seejuures ei saa olla kindel, et leitud mudel annab sõltuva tunnuse parima prognoosi. Peale selle ei ole normaaljaotusele tuginevad olulisuse hinnangud ja nendel põhinev hüpoteeside kontroll põhjendatud.

1.5.3.1. Regressioonivõrrand

Lihtsa lineaarse funktsiooni graafik on sirge, mille võrrandi saab kirjutada järgmiselt:

$$y = b_0 + b_1 x + \varepsilon. \qquad [1\text{-}77]$$

Seejuures on regressiooni puhul funktsioontunnus y argumenttunnuse keskväärtus kohal x. Selles funktsioonis on kaks parameetrit: regressioonikordaja (*slope*; b_1) ja vabaliige (*y-intercept;* b_0). Regressioonikordaja näitab funktsioontunnuse ootuse muutumist argumenttunnuse ühikulise muutuse

korral. Vabaliige näitab funktsioontunnuse väärtust juhul, kui argumenttunnuse väärtus on null. Kui mudel peaks esindama ka prognoositava tunnuse muutlikkust, siis lisatakse valemisse juhuslikku viga kirjeldav parameeter ε, mida nimetatakse prognoosijäägiks. Saadakse järgmine alltoodud regressioonivalem. Kui juhuslik viga on mudelisse kaasatud, siis prognoosib mudel oodatavaid väärtusi koos nende hajuvusega, kui ei ole, siis oodatavat keskväärtust.

$$y = b_0 + b_1 x + \varepsilon \qquad [1\text{-}78]$$

Regressioonimudeli kasutamiseks tuleb leida regressioonikordajatele kõige sobivamad väärtused. Regressioonanalüüsis eeldatakse enamasti, et juhusliku vea väärtused on normaaljaotusega ja seetõttu sobib b_0 ja b_1 väärtuste leidmiseks **vähimruutude printsiip**. Selle kohaselt tuleb regressioonikordajatele omistada väärtused, mille puhul funktsioontunnuse väärtuste erinevuste ruutude summa vaatlusandmetes ja regressioonivalemist arvutatult erinevad kõige vähem. Seega püütakse minimiseerida prognoosi summaarset või keskmist ruutviga. Graafiliselt tähendaks see regressioonisirge tõmbamist läbi ristteljestikus oleva punktide parve nii, et punktide *y*-koordinaatide summaarne või keskmine kõrvalekalle regressioonisirgest oleks võimalikult väike. Seejuures tähistab võrrandi parameeter b_1 regressioonisirge tõusunurga tangensit ja parameeter b_0 funktsioontunnuse väärtust regressioonisirge *y*-teljega lõikumise kohas (st $x = 0$).

Regressioonisirge sobitatakse läbi korrelatsioonivälja minimeerides funktsioontunnuse vigu, kuid mitte argumenttunnuse vigu; argumenttunnuse puhul eeldatakse, et see on mõõdetud veatult

Lineaarse regressiooni, nagu regressioonanalüüsi puhul üldse, tuleb silmas pidada, et regressioonifunktsiooni kasutamine väljaspool selle sobitamiseks kasutatud empiiriliste andmete muutumispiirkonda ei ole põhjendatud. Keerukamaid regressioonimudeleid käsitletakse peatükis 3.4.1.

Mitmetunnuselise ehk mitmese regressioonanalüüsi puhul käsitletakse funktsioontunnust sõltuvuses mitmest argumenttunnusest. Regressioonimudeli parameetrite leidmine toimub võrrandsüsteemi lahendamisena. Mitmene regressioonanalüüs eeldab kas argumenttunnuste omavahelist sõltumatust (mõjude **liituvust** ehk aditiivsust) või tunnuste koosmõjude lisamist regressioonimudelisse. Mitmemõõtmelist regressiooni, mis sisaldab mitut funktsioontunnust, käsitletakse statistilise modelleerimise peatükis (ptk 3).

1.5.3.2. Regressiooni vastavus andmetele

Iga uuritava objekti empiirilise mõõtmisega kaasnevad juhuslikud hälbed, mida võib käsitleda kui vaatluse all mitteolevate faktorite mõju. Enamasti ei ole funktsioontunnuse väärtused määratud vaid ühe argumenttunnuse poolt. Juhuslike vigade kohta eeldatakse, et need on erinevate vaatluste puhul sõltumatud, nende keskväärtus on null ja nende standardhälve on konstantne – see tähendab oodatav viga on sama suur argumenttunnuse iga väärtuse puhul. Juhuslike hälvete dispersiooni hinnangut nimetatakse keskmiseks ruutveaks ehk **jääkdispersiooniks** – tähistus s^2.

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} \qquad [1\text{-}79]$$

kus n on vaatluste arv, y_i on prognoositava tunnuse y mõõdetud väärtus i-ndal vaatlusel, $\hat{y}_i$ on prognoositava tunnuse y mudeli abil hinnatud väärtus i-ndal vaatlusel.

Ruutjuurt jääkdispersioonist nimetatakse regressioonimudeli standardveaks ehk prognoosiveaks ehk **jääkstandardhälbeks** (*standard error of residuals*). Jääkstandardhälve iseloomustab funktsioontunnuse hajuvust regressioonijoone ümber.

Lihtsa lineaarse regressiooniseose olulisuse kontrollimiseks kasutatakse *F* statistikut vabadusastmete arvuga 1 ja *n* – 2. *F* statistikuks võetakse regressiooni kirjeldatud muutlikkuse ja jääkdispersiooni ehk kirjeldamata muutlikkuse suhe. Selle saab arvutada ka determinatsioonikordaja alusel, mis näitab millisel määral argumenttunnuse teadmine võimaldab funktsioontunnuse väärtust täpsemalt hinnata ehk millise osa kogudispersioonist kasutatud mudel ära kirjeldab.

$$F = (n-2)\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \qquad [1\text{-}80]$$

Kui statistik on suurem kui *F* statistiku kriitiline väärtus etteantud olulisuse nivoo ja vabadusastmete arvu juures, saab vastu võtta sisuka hüpoteesi.

Regressiooni kui prognoosi usalduspiiride arvutamisel tuleb vahet teha prognoosi usalduspiiridel ja üksikväärtuste usalduspiiridel. Prognoosi usalduspiirid arvestavad vaid regressioonivõrrandi parameetrite hindamise täpsust, üksikväärtuse usalduspiirid arvestavad ka üksikväärtuste juhuslikke hälbeid. Regressioonisirge ja üksikväärtuste usalduspiiride arvutamine eeldab regressioonivigade normaaljaotust.

Regressioonimudeli seletusvõime kontrollimiseks kasutatakse ka jääkide analüüsi. **Regressioonijäägiks** (*residual*) nimetatakse tegeliku mõõtmistulemuse ja argumenttunnuse sama väärtuse puhul oleva prognoosi vahet. Kõige kujukam on koostada jääkide graafik hajuvusdiagrammina, mille ühel teljel on jäägi väärtus ja teisel kas argumenttunnuse väärtus, funktsioontunnuse prognoos, vaatluse järjekorranumber või mingi regressioonimudelis kajastamata tunnus. Kui regressioonanalüüsi eeldused on täidetud, peaksid kõik jääkide graafikud olema ühtlased punktiparved, milles positiivsete ja negatiivsete jääkide arv on ligikaudu võrdne ja milles jääkide varieeruvus ei sõltu graafiku argumenttunnusest. Kui jäägid on oma standardhälvetega standardiseeritud, peaks jääkide jaotus olema standardiseeritud normaaljaotusega.

1.5.4. Dispersioonanalüüs

Dispersioonanalüüs (*analysis of variance* – ANOVA) on meetodite kompleks kahe või rohkema üldkogumi keskväärtuste erinevuse kontrollimiseks.

Dispersioonanalüüsi käigus hinnatakse, kui suurt osa kogu valimi ühisest varieeruvusest on võimalik kirjeldada faktorite tasemete keskväärtustega

Enamasti huvitavad uurijat ka keskväärtuste erinevuse võimalikud põhjused. Põhjuste selgitamiseks mõõdetakse või muudetakse katse ajal uurijale huvi pakkuvaid ja eeldatavasti mõju avaldavaid tingimusi – **faktoreid**. Sõltuvalt kasutatud faktorite arvust eristatakse **ühefaktorilist** ja **mitmefaktorilist** ehk faktoriaaldispersioonanalüüsi. Mitmefaktoriliste analüüside puhul on uuritavaid kogumeid gruppideks jagavaid faktoreid mitu. Mitme funktsioontunnuse korral nimetatakse analüüsi **mitmemõõtmeliseks** ehk kanooniliseks analüüsiks.

Faktori väärtusi nimetatakse faktori **tasemeteks** (*levels*). Tasemed on analüüsis kasutatavad väärtusklassid või kategooriad. Sõltuvalt faktorite tasemete kombineerimise viisist eristatakse ristmudelit ja hierarhilist mudelit. **Ristmudeli** korral on vaatlused tehtud kõikvõimalikel faktorite tasemete kombinatsioonidel. **Hierarhilise mudeli** korral on esimese faktori iga tase seotud vaid teise faktori teatud tasemetega. Osa teise faktori tasemeid on esindatud kombinatsioonis esimese faktori mõne teise tasemega. Kui igal faktortunnusel on igal tasemel tehtud ühepalju katseid, nimetatakse mudelit **tasakaalustatuks**.

Kui järeldusi kavatsetakse teha vaid faktorite katseks valitud tasemetel, nimetatakse mudelit **fikseeritud mõjudega** (*fixed effects*) ehk tasemetega mudeliks. Fikseeritud tasemetega mudel ja sellest tuletatud järeldused kehtivad vaid etteantud tasemete korral. Kui faktoril on palju erinevaid väärtusi ja katses kasutatud tasemeid vaadeldakse juhusliku valimina faktori võimalike tasemete hulgast, nimetatakse mudelit **juhuslike mõjudega** (*random effects*) mudeliks. Selle mudeli korral tehakse järeldusi faktori kõigi tasemete kohta, ka nende kohta, mis on katses esindamata.

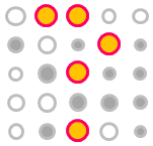
Dispersioonanalüüsi eeldused on:

- juhuslikud vead on üksteisest sõltumatud;
- juhuslike vigade keskväärtus on null ja dispersioon on faktori kõikidel tasemetel ühesugune (dispersioon on homogeenne);
- kõigi faktorite juhuslikud vead on normaaljaotusega;
- kõik faktorite mõjud on aditiivsed.

Probleeme võivad tekitada mittetasakaalulised valimid, puuduvad väärtused, erindid ja faktorite multikollineaarsus. Kui mõni eeldustest on täitmata, on sageli õigem kasutada mõnda mitteparameetrilist dispersioonanalüüsi meetodit, näiteks Kruskal-Wallise testi.

1.5.5. Kruskal-Wallise test

Kruskal-Wallise test sobib kasutamiseks, kui klassikalise dispersioonanalüüsi eeldused ei ole täidetud või kui uurija ei ole eelduste täitmise osas kindel. Meeldetuletuseks, dispersioonanalüüs eeldab juhuslike vigade normaaljaotust, üksteisest sõltumatust ja ühesugust dispersiooni faktori kõikidel tasemetel. Kruskal-Wallise testi puhul on ainsateks nõueteks sõltuva tunnuse pidevus, või järjestatavus ning paljude väärtuste omamine. Eeldatakse, et kasutatavad valimid on juhuslikud ja sõltumatud. Valimite mahud võivad olla erinevad. Test kasutab mõõdetud suuruste asemel nende astakuid ja aritmeetilise keskmise asemel mediaani. Kruskal-Wallise test kahe väärtusklassiga on sama mis Mann-Whitney test.



Küsimused

1. Millisel juhul võrdub muutuja dispersioon nulliga?

2. Kas üksiksündmuse tõenäosust kajastab jaotusfunktsioon või tõenäosusfunktsioon?

3. Milleks standardiseeritakse tunnuseid?

4. Mida saab järeldada, kui empiiriline jaotus erineb oluliselt oodatavast jaotusest?

5. Kui suur on tõenäosus, et viie pojaga kährikkoera pesakonnas on kaks emast ja kolm isast poega, kui eeldada, et isase järglase sündimise tõenäosus on 0,53?

6. Oletame, et õpilaste teadmiste kontrolliks soovitakse kasutada testi, mis koosneb kümnest küsimusest, kusjuures igal küsimusel on ette antud kolm vastusevarianti. Oletame, et test loetakse sooritatuks, kui õigesti on vastatud vähemalt kuus küsimust. Kui suur on tõenäosus test edukalt läbida, kui vastaja teab vastuseid pooltele küsimustele ja teiste küsimuste vastused märgib juhuslikult? Millised on võimalused testi rangemaks muuta?

7. Olgu taignasse, millest valmistatakse 100 rosinakuklit, pandud 300 rosinat. Mitu kuklit on oodatavalt ilma rosinateta eeldades, et taigna segamine tagab rosinate juhusliku paiknemise?

8. Kui suur on tõenäosus, et kahvatõmmetega taimedelt kogutud 50 puugi hulgas on vähemalt üks entsefaliidiviiruse kandja, kui viiruse kandjate osakaal kogu populatsioonist on 1%?

9. Üksikisiku jaoks ei ole lapse sünd juhuslik, kuid rahva puhul võib laste sündi vaadelda juhusliku protsessina. Oletame, et ühe Kukimukimaa täiskasvanud naise kohta on sellel maal keskmiselt 2,3 last. Kui palju peaks 1000 selle maa 50-aastase naise hulgas olema viie lapsega emasid, kui lapsed jaguneksid naiste vahel juhuslikult?

10. Kui suur on tõenäosus, et aasal liblikaid püüdes tabatakse kirju liblikas neljandana, kui lisaks kirjudele püütakse vaid valgeid? Eeldame, et kirjude ja valgete liblikate arvukused, märgatavused ja püütavused on võrdsed.

11. Keskmiselt on viiesajaleheküljelise käsikirja leheküljel 4,5 kirjaviga. Mitmel leheküljel peaks olema üle kümne vea, kui vea tekkimine on juhuslik protsess ja vea tekkimise tõenäosus on käsikirja kõikides osades sama?

12. Kahesajaleheküljelises käsikirjas on kokku 700 kirjaviga. Mitu lehekülge peaks olema vigadeta, kui vea tekkimine on juhuslik protsess ja vea tekkimise tõenäosus on käsikirja kõikides osades sama? Millele viitab suurem vigadeta lehekülgede arv?

13. Millal on Poissoni jaotuse tõenäosusfunktsioon ligikaudu sümmeetriline?

14. Ida-Viru maakonnas on põdra elupaiku (metsamaad) 2280,07 km². 1999.a. oli seal hinnanguliselt 823 põtra. Mitmel ruutkilomeetril on ootuspärane 4 või enama põdra esinemine, kui põtrade

paiknemine oleks juhuslik? Mida võib arvata, kui vaatlusandmete järgi nähti nelja- või enamapealisi põdraseltsinguid suuremas arvus kohtades?

15. Ida-Viru maakonnas on põdra elupaiku (metsamaad) 2280,07 km². 1999.a. oli seal hinnanguliselt 823 põtra. Mitu põtra asuks oodatavalt üksikult (ainsana ruutkilomeetris), kui põtrade paiknemine oleks juhuslik?

16. Mis on nullhüpotees?

17. Mida uuritakse ühepoolse, mida kahepoolse hüpoteesi abil?

18. Mis on süstemaatilise valikumeetodi eelis juhusliku valikumeetodi ees?

19. Kas juhuslike puude vanuse määramine Tartu Toomemäel esindab Eesti pargipuude keskmist vanust? Põhjenda vastust.

20. Mis vahe on keskmisel hälbel ja standardhälbel?

21. Kuidas tuleks vastata küsimusele: "Kui palju tuleks teha mõõtmisi, et saaks etteantud täpsusega ja etteantud kindlusega määrata uuritava üldkogumi keskmist?"

22. Kas keskväärtus võib olla suurem kui ülemine kvartiil?

23. Mis on dispersiooni mõõtühik?

24. Milline oleks sageduste tulpdiagramm (histogramm), kui y-teljel oleks vaatluste arv ja iga kvartiilivahemiku kohta oleks üks tulp? Mitu tulpa oleks, kui kõrged oleksid tulbad?

25. Kas juhusliku muutuja dispersiooni arvuline väärtus võib olla väiksem kui sama muutuja standardhälbe arvuline väärtus?

26. Kas standardiseeritud muutuja keskväärtus võrdub ühega?

27. Loodusmaastiku metsaservadest ning põllumajanduspiirkonna heinamaadelt kütiti kokku 22 juhuslikku rebast ning toitumisanalüüsi jaoks uuriti nende mao sisusid. Uruhiirte alalõualuude arv loodusmaastiku rebaste maosisus olid: 12, 10, 13, 14, 18, 10, 11, 12, 13, 12, 12, 13; ja põllumajandusmaastiku rebastel: 14, 17, 18, 16, 16, 15, 12, 11, 12, 14. Kui kindlalt lubab *U* test nende andmete põhjal väita üldiselt, mitte ainult uuritud isendite kohta, et kahes võrreldud elupaigatüübis toituvad rebased uruhiirtest erineval määral? Eeldame valimite esinduslikkust. Miks on selles ülesandes räägitud valimist ja üldkogumist? Miks eeldatakse valimite esinduslikkust?

28. Vaadeldi 1090 rasvatihase reaktsiooni raudkulli topisele pesapaiga läheduses. 450st vaadeldud isalinnust reageeris agressiivselt 320, emaslindudest näitas agressiivsust välja 420. Kui kindlalt lubab χ^2 test väita, et eri sugu rasvatihaste agressiivsus potensiaalse vaenlase suhtes on erinev? Kas see kindlustunne on teaduslikuks järelduseks piisav?

29. Kas korrelatsioonikordaja mõõdab statistilise seose kuju, suunda, tugevust või seose olemasolu usaldatavust?

30. Mis vahe on argumenttunnusel ja funktsioontunnusel? Kumb kummast sõltub?

31. Mis on kovariatsiooni mõõtühik?

32. Kas korrelatsioonikordaja väärtus sõltub muutujate mõõtühikutest?

33. Mida näitab regressioonikordaja?

34. Zooplankteri kehapikkuse (x) ja biomassi (y) vahel leiti regressioon $y = 3{,}3x - 35$. Kas seda võrrandit on õigem kasutada biomassi hindamiseks kehapikkuse järgi või kehapikkuse hindamiseks biomassi järgi või ükskõik kumba pidi?

35. Mida peaks õppejõud arvama, kui eksami vastustes on tudeng korrelatsioonikordaja väärtuseks märkinud 2,1?

36. Leia täringuviske tulemuse keskväärtus kuuetahulise võrdkülgse ühtlase kaaluga täringu puhul. Mis on selle keskväärtuse mõõtühik? Miks on täringu kuju ja raskuskeskme paiknemine oluline?

37. Kas χ^2 testi järgi võib 95% kindlusega väita, et roomav öövilge (*Goodyera repens*) eelistab mullakaardil oleva lõimise järgi liivmuldi, kui vaatluskohad jagunesid järgmiselt: liivmullal esines 35 korda ja puudus 22 korda, muudel muldadel esines 14 korda ja puudus 27 korda?

38. Mitmeks väärtusklassiks peavad tunnuste väärtused olema jagatud, et saaks kasutada hii-ruut testi?

39. Kas χ^2 test võrdleb keskväärtusi?

40. Kas χ^2 testi sobib protsentide võrdlemiseks?

41. Kas χ^2 testi eeldab tunnuste väärtuste järjestatavust?

42. Milliste tunnuste vahelise seose olemasolu kontrollimiseks on χ^2 test kasutatav, Spearmani ρ aga mitte?

43. Mis on mediaani mõõtühik?

44. Oletame, et meil on normaaljaotusega muutuja, mille keskväärtus on 1 ja standardhälve 2. Mitu protsenti selle muutuja väärtustest on oodatavalt suuremad kui 3?

45. Kahel võrdse asustustihedusega uurimisalal võrreldakse valimite põhjal rohukonnade arvukust hektari kohta kuuel eri päeval. Kui suur on tõenäosus, et ükskõik kummal alal on valimi põhjal saadud hinnang teisest suurem kõigil kuuel päeval?

46. Jääpangal triivivatele polaaruurijatele kavatsetakse visata varustust kahelt erinevalt lennukilt.

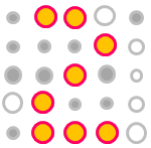
Tõenäosus ühelt lennukilt visatud varustuse sattumiseks jääpangale on 0,5 ja teiselt lennukilt 0,8. Milline on tõenäosus, et nad jäävad varustusest täiesti ilma? Milline on tõenäosus, et nad saavad kätte vähemalt osa varustusest?

47. Kui suur on kahe teineteist välistava sündmuse koosesinemise tõenäosus, kui kummagi sündmuse esinemise tõenäosus on 0,1?

48. Oletagem, et poisi või tüdruku sündimise tõenäosus on võrdselt 0,5. Oletagem veel, et 100 000 elanikuga maal, kus mehi ja naisi on võrdselt ning meeste ja naiste eluiga on võrdne, otsustavad kõik pered saada lapsi seni, kuni sünnib poeg ja seejärel nad rohkem lapsi ei saa. Milline on meeste ja naiste vahekord sellel maal 100 põlvkonna järel?

49. Üks ärimees oli mures, et lennukis, millega ta reisib, võib olla pomm. Ta uuris välja, et pommi olemasolu tõenäosus lennukis on väike, kuid mitte piisavalt väike tema jaoks. Nüüd võtab ta alati oma pommi endaga kaasa, sest kahe pommi samas lennukis oleku tõenäosus oleks ju lõpmata väike. Kas ta käitus õigesti? Kui ei, siis kuidas talle selgitada, et ta eksis?

2. Kirjeldav andmeanalüüs



Kirjeldav andmeanalüüs (*exploratory data analysis – EDA*) otsib seaduspärasusi andmetes ilma eeldusi või hüpoteese kasutamata, vastandudes sellega statistiliste hüpoteeside testimisele ja parameetrite väärtuste hindamisele ehk **kinnitavale analüüsile** (*confirmatory analysis*). Tagasi kirjeldava analüüsi juurde jõutakse sageli ka juhul, kui planeeritud eksperimendist selguvad ootamatud, esialgsetele hüpoteesidele mitte vastavad tulemused. Kirjeldav andmeanalüüs kuulub induktiivsete meetodite, kinnitav analüüs deduktiivsete meetodite hulka. Kirjeldava andmeanalüüsi puhul uuritakse tavaliselt läbi palju faktoreid ja kasutatakse erinevaid meetodeid.

Kirjeldav andmeanalüüs hõlmab traditsioonilist kirjeldavat statistikat (keskväärtuste arvutamine, varieeruvuse kirjeldamine, jaotuste ja mitme muutuja ühisjaotuse kirjeldamine), ordineerimis- ja klassifitseerimismeetodeid ning statistilise modelleerimise vahendeid, kui neid kasutatakse minimaalsete eelnevate hüpoteesidega. Kirjeldava andmeanalüüsi juures omavad tähtsat rolli visualiseerimismeetodid, nagu funktsioonide visuaalne sobitamine, andmepunktide eemaldamise, lisamise või nihutamise mõju vaatamine, erinevat tüüpi graafikute koostamine ja omavahel kombineerimine. Piir kirjeldava statistika ja modelleerimise vahel on hägune, kuna ka lihtsaid üldistusi võib mudeliteks nimetada.

Kirjeldav analüüs on vajalik andmetest ülevaate saamiseks, vajalike teisenduste leidmiseks ja edasise uuringu kavandamiseks

Kirjeldava andmeanalüüsi meetodeid on kasutatud biomeetrias ja geobotaanikas juba alates kahekümnenda sajandi algusest. Matemaatilisse statistikasse tõi selle termini John W. Tukey 1970ndatel aastatel (Tukey 1977). Tukey võrdleb kirjeldavat andmeanalüüsi detektiivi tööga ja kinnitavat analüüsi kohtumenetlusega. Kirjeldav andmeanalüüs on uurimisprotsessi esimene etapp, osa selles etapis leitud tõendeid võivad olla juhuslikud ja uuritava juhtumiga sisuliselt mitte seotud. Sellegipoolest, kirjeldavast andmeanalüüsist loobumine tähendaks piirdumist vaid etteplaneeritud hüpoteeside kontrollimisega ja mitmed ootamatud tulemused jääksid saamata.

Kirjeldava andmeanalüüsi meetodeid kasutatakse lisaks loodusteadustele laialdaselt ka sotsiaalteadustes, poliitikas ja ärijuhtimises. Majandustegevuses ja poliitikas tuleb otsuseid langetada ka olukorras, kus põhjalikuks uurimiseks ei ole aega, kuid kiirest ebatäpsest hinnangust oleks siiski abi. Üsna ligikaudne on muuhulgas õpilaste ja tudengite hindamine, sest täpne hinne ei ole eesmärk, millele tasuks väga palju aega kulutada. Kujutlege kui kaua võtaks eksam aega, kui eksamineerija peaks vastajaid hindama sajandikhinde täpsusega ja eksida ei tohiks. Enamasti on ka ligikaudne või ebakindel hinnang parem kui hinnangu puudumine.

Kirjeldava analüüsi taastähtsustamine on osaliselt seotud statistilise olulisuse liigse ja sisulist tähtsust mitteomava kasutamisega, mis on tekitanud vastuseisu statistiliste testide kasutamisele tervikuna (Johnson 1999). Statistiliste hüpoteeside kasutamine on õigustatum teadusharudes, mis tuginevad matemaatiliselt rangetele teooriatele. Teooria loetakse Karl Popperi järgi kasutuskõlblikuks seni, kuni vastupidist pole õnnestunud tõestada. Ökoloogia ja maateaduste teooriad ei ole nii konkreetsed ja üldkehtivad, et neid saaks mõne vaatlusega ümber lükata. Suurem osa loodusvaatlustest kirjeldab loodust ja kogub tõendeid ühe või teise küllaltki üldsõnalise arusaama kinnituseks.

2.1. Mitmekesisus

Mitmekesisust võib pidada mitmemõõtmeliseks nähtuseks, sest see arvutatakse mitme tunnuse alusel. Arvuna väljenduva koondhinnangu saamiseks kasutatakse indekseid. Klassikuuluvuse mitmekesisuse indeksitega saab määrata nii klassifitseeritud ruumiosade kui ka asukohast sõltumatute objektide mitmekesisust. Lähteandmeteks võib olla nii valim kui ka kogu uuritav andmestik. Kui mitmekesisust on määratud korduvatest ja sõltumatutest ühetaolistest valimitest, saab mitmekesisuse keskmise erinevuse olulisust statistiliste testidega kontrollida. Valimit kasutava uuringu puhul ei pruugi kogu mitmekesisus valimis kajastuda.

Mitmekesisuse indekseid kasutatakse muuhulgas taime- ja loomakoosluste ning maastikustruktuuri iseloomustamiseks. Elurikkust ehk elustiku varieeruvust jagatakse α, β ja γ taseme mitmekesisusteks (Whittaker 1972). α mitmekesisus ehk diversiteet on analüüsiruudu või vaatlusala sisene varieeruvus, β mitmekesisus on vaatluskohtade vaheline ning γ mitmekesisus on piirkondlik. α-mitmekesisuse moodustavad liikide arv ja isendite jagunemine liikide vahel uurimiskoha või elupaiga piires; β-mitmekesisus on mitme sellise koha või elupaiga liigifondide erinevus; γ-mitmekesisus hõlmab nii kohtade sisest (α) kui kohtade vahelist (β) varieeruvust suuremal ala, moodustudes kogu uuritava piirkonna liikide koguhulgast ja isendite jagunemisest liikide vahel. Näiteks tabades kümnelt uurimisalalt igast kümme ühte liiki isendit, mis on aga igal uurimisalal eri liiki, siis α-mitmekesisus on madal, kuid β-mitmekesisus kõrge – kohad on siseselt ühetaolised, kui omavahel erinevad. γ-mitmekesisus on sõltuvalt arvutuskäigust keskpärane või isegi suhteliselt madal. Analoogiliselt oleks α-mitmekesisus kõrge ja β-mitmekesisus madal, kui kõik 10 liiki oleksid esindatud iga koha kümne registreeritud isendi hulgas.

Algselt on γ diversiteeti määratletud α ja β diversiteedi korrutisena (Whittaker 1972), teine ja levinum määratus on, et γ diversiteet on α ja β diversiteedi summa (Lande 1996). Ka β taseme mitmekesisuse tähenduse ja mõõtmise meetodite üle on diskuteerinud Veech *et al.* (2002), Jost (2007) ja Jurasinski *et al.* (2009).

Mitmekesisuse indeks ei ole samatähenduslik, kui see on arvutatud erinevalt määratletud klassifikatsiooniüksuste alusel

Kuna liikide levik varieerub nii geograafilise ruumi kui ka keskkonnatingimuste gradientide suhtes, peaks mitmekesisusel olema kaks komponenti: elupaikade vaheldumisele vastav osa ja igale liigile omase ruumis paiknemise mustriga seotud osa. Teisele komponendile võib lisada ka paiknemise juhuslikkusest tingitud varieeruvuse. Indeksite väärtuste võrdlemisel on lisaks kasutatud valemile oluline tähele panna ka kasutatavaid klassifikatsiooniüksusi (perekond, liik, eluvorm).

Enamkasutatud mitmekesisuse näitajad on taksonite või muude objektitüüpide (klasside) arv valimis, Simpsoni dominantsiindeks ja selle teisendused, Shannoni mitmekesisuseindeks ja selle variandid ning Lorenzi kõverast arvutatav Gini indeks. Vähemkasutatud indeksitest mainiksime Margalefi rohkuse (*richness*) indeksit *SR* (Margalef 1957).

$$SR = \frac{(s-1)}{\ln(N)} \qquad [2\text{-}1]$$

Järgnevalt, kuni peatüki 2.1 lõpuni on kasutatud tähistust: N on ühikute (isendite) arv, s on klasside (taksonite) koguarv, i ja j on klasside (taksonite) indeksid, d_{ij} on klasside i ja j erinevus, p_i on klassi (taksoni) i osa, seega $p_i = n_i / N$.

Hurlberti liikide kohtumise tõenäosus (*probability of interspecific encounter*) näitab teise liigi isendiga kohtumise tõenäosusest juhuslikult liikuvate mitut liiki isendite hulgas (Hurlbert 1971).

$$\mathrm{P}(IE) = \sum_{i=1}^{s} \frac{n_i}{N} \cdot \frac{N - n_i}{N - 1} \qquad [2\text{-}2]$$

Ruutentroopia indeks (*quadratic entropy index*) arvestab klasside omavahelist erinevust (Rao 1982). Klasside erinevuse kaale saab lisada kõigisse mitmekesisuse ja ühetaolisuse indeksitesse, seades kaalu võrdeliseks liigi summaarse erinevusega teistest liikidest (Ricotta 2002).

$$Q = \sum (d_{ij}\, p_i\, p_j) \qquad [2\text{-}3]$$

Mitmekesisuse indeksites on pikemalt juttu publikatsioonides: Jost (2006), Salas *et al.* (2006), Pinto *et al.* (2009).

2.1.1. Dominantsiindeks

Herfindahli või Simpsoni või Hirschmani **dominantsiindeks** (*D*) (Herfindahl 1950, Hirschman 1964, Simpson 1949) ehk ühetaolisuse indeks muutub vahemikus 0...1, kus lõpmata suurt mitmekesisust tähistab null ja täielikku ühetaolisust väärtus üks. Majandus- ja õigusteaduses tuntakse seda eelkõige turu kontsentreerumist ja monopolide moodustumist mõõtva Herfindahli indeksina, ökoloogias Simpsoni indeksina.

$$D = \sum p_i^{\,2} \qquad [2\text{-}4]$$

Sellisel kujul indeksi nullväärtus on teoreetiline abstraktsioon, miinimumväärtus reaalsetes andmetes sõltub üksuste arvust. Saamaks dominantsiindeksi oodatavat väärtust eeldusel, kui üksused on võrdselt esindatud ja sõltumata üksuste arvust (*s*), saab kasutada teisendust

$$E(D) = \left(D - \frac{1}{s}\right) \Big/ \left(1 - \frac{1}{s}\right). \qquad [2\text{-}5]$$

Kuna ökolooge huvitab eelkõige mitmekesisus, mitte ühetaolisus, siis on Simpsoni indeksit teisendatud vastand- ja pöördväärtuseks:

$$1 - \sum p_i^{\,2} \qquad [2\text{-}6]$$

$$\frac{1}{\sum p_i^{\,2}} \qquad [2\text{-}7]$$

Dominantsi tihedus *C* (*concentration of dominance*) väljendub

$$C = \frac{1}{-\ln \sum p_i^{\,2}}. \qquad [2\text{-}8]$$

Siin on kasutatud naturaallogaritme, aga võib kasutada ka kümnend- või kahendlogaritme, see on vaid mõõtühiku valiku küsimus. Indeksite võrdlemisel peaksid mõõtühikud muidugi samad olema.

Dominantsiindeksi ruumiline versioon kasutab liikide (*k*) osakaale (p_k) vaatluspaarides. Mitmekesisus ühes vaatluspaaris vastab tõenäosusele, et kahes vaatlusruudus *i* ja *j*, mis asuvad vahemaaga *h* on erinevad liigid (Bennie *et al.* 2011).

$$H_{ij} = 1 - \sum_{k=1}^{s} p_{ik}\, p_{jk} \qquad [2\text{-}9]$$

Kõigi teatud vahemaaga paiknevate vaatluspaaride keskmine mitmekesisus (*paired-sample diversity*) on sel juhul

$$H(h) = 1 - \frac{1}{N(h)} \sum_{(i,j)\in N(h)} \sum_{k=1}^{s} p_{ik}\, p_{jk}, \qquad [2\text{-}10]$$

kus *N(h)* on vahemaaga *h* vaatluspaaride arv, *i* ja *j* on vaatluste indeksid, *k* on liigi indeks ja *s* on liikide koguarv.

Bennie *et al.* (2011) kujutasid graafikul *H(h)* sõltuvust vaatlustevahelisest vahemaast *h* ja erinevusest piki keskkonnagradienti analoogiliselt semivariogrammiga (ptk 5.3.3 ja 5.3.4). *H(h)* teisendamiseks muutumisvahemiku poolest sarnaseks valemi [2-7] järgi arvutatud mitmekesisusele kasutasid Bennie *et al.* (2011) teisendust

$$D(h) = \frac{1}{1 - H(h)}. \qquad [2\text{-}11]$$

2.1.2. Shannoni mitmekesisus

Shannoni mitmekesisuse (*diversity*) indeks (Shannon 1948) sõltub nii üksuste arvust kui ka üksuste mahu ühetaolisusest. Indeksi väärtused algavad nullist, ülemist piiri ei ole. Shannoni mitmekesisuse tähises on tavaks kasutada ülakoma *H'*, kui peetakse silmas mitmekesisust suures üldkogumis, mida on uuritud valikuliselt; mitmekesisust piiratud mahuga koosluses, kus kõik komponendid on ära loetud või mõõdetud, tähistatakse ilma ülakomata – *H* (Pielou 1977).

$$H = -\sum p_i \left(\log_2 p_i\right) \qquad [2\text{-}12]$$

Claude E. Shannoni publikatsioon *A mathematical theory of communication* (1948) pani aluse informatsiooniteooriale. Shannon võttis kasutusele mõisted bit ja entroopia ning näitas, et teateid saab saata nullide ja ühtede jada abil. C.E. Shannon on ka intellektitehnika-alaste uurimuste autor ning males ja teistes kahe mängija arvutimängudes kasutatava algoritmi üks leiutajatest. 1949. aastal ilmus raamat *A Mathematical Theory of Communication*, mille C.E. Shannon kirjutas koos Warren Weaveriga. Norbert Wiener on aga alusepanija teadusharule küberneetika, mida ta iseloomustas kui juhtimist ja sidet loomas ja masinas (*CYBERNETICS or Control and Communication in the Animal and the Machine*).

2.1.3. Lloydi ühetaolisus

Lloydi ühetaolisus (Lloyd ja Ghelardi 1964) on ühetaolisuse (*evenness, equitability*) klassikaline mõõdik.

$$J = \frac{H}{H'_{\max}}, \qquad [2\text{-}13]$$

kus $H'_{max} = ln(s)$, *s* on taksonite arv, *H* on mitmekesisus Shannoni võrrandi järgi (ptk 2.1.2).

Ühetaolisust saab arvutada ka dominantsiindeksi baasil, kus ühetaolisuse maksimumväärtuseks on dominants antud arvu üksuste võrdse osa korral.

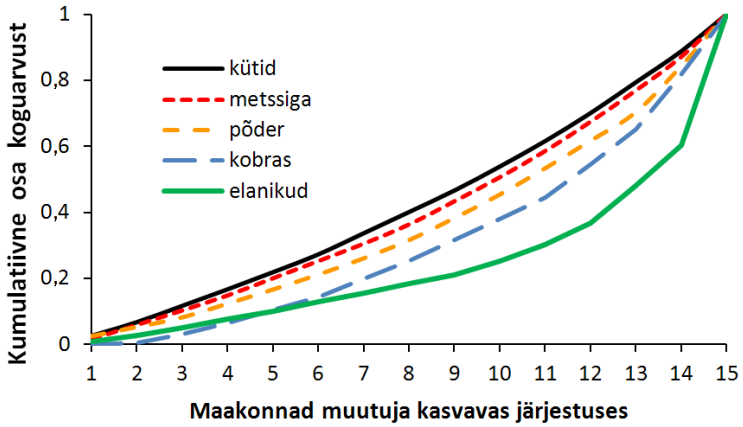
$$E = \frac{-\ln \sum p_i^{\,2}}{H'_{\max}} \qquad [2\text{-}14]$$

2.1.4. Lorenzi kõver ja Gini indeks

Lorenzi kõver (Lorenz 1905) on graafik, mis kujutab suuruse alusel järjestatud arvuliste väärtuste ühetaolisust/ebavõrdsust kumulatiivsel kujul (joonis 2-1). Algselt ja senini kõige sagedamini kasutatakse Lorenzi kõverat sissetulekute ebavõrdsuse iseloomustamiseks. Lorenzi kõver on aga samahästi rakendatav igasuguste arvukogumite ühetaolisuse/erinevuse iseloomustamiseks.

Lorenzi kõver on kas üks-üheselt tõusev sirge või nõgus joon. Kumer kuju ei ole võimalik, sest diagonaalne sirge näitab täielikku võrdsust ja sellest suuremat võrdsust ei saa olla. Ühetaolisuse graafiline esitus on kõnekas, aga võrdlustes eelistatakse sageli arvulisel kujul mõõdikuid. Lorenzi kõverast saab tuletada ebaühtluse ehk Gini indeksi, mis on andmetes oleva ebavõrdsuse suhe maksimaalse ebavõrdsuse suhtes. Seega Gini indeksi miinimumväärtus 0 ja maksimumväärtus 1. Gini indeksi arvutamisreegli puhul on variandid. Kõvera ja diagonaali maksimaalne vahe kõvera lõpu kõrguse (väärtuste summa) suhtes ei nõua erilisi arvutusi, kuid maksimaalset erinevust saab mõõta graafikul nii ühe kui teise telje sihis või siis risti diagonaaliga ning tulemus sõltub väärtuste jaotumise eripärast. Jagades Lorenzi kõvera aluse pindala diagonaaljoone aluse pindalaga, arvestame kõiki kõverat moodustavaid väärtusi.

Lorenzi kõver kujutab suuruse järgi järjestatud muutuja kumulatiivset osa



Joonis 2-1. Lorenzi kõverad mõnede ulukite kütitud arvust aastal 2010 (andmed Männil *et al*. 2011), küttide arvust aastal 2009 (andmed Keskkonnateabe Keskus) ja elanike arvust aastal 2012 (andmed Statistikaamet). Märka, et kütid paiknevad Eesti maakondade vahel palju ühtlasemalt kui elanikud.

Corrado Gini (1884–1965) oli mitmekülgne Itaalia teadlane, kes huvitus sotsioloogiast, tõenäosusteooriast, majandusteooriast, bioloogiast ja statistikast (Giorgi 2005) ning kes on avaldanud üle 800 publikatsiooni.

Lorenzi kõvera konstrueerimiseks tuleb võrdluse aluseks olevad arvud esmalt järjestada suuruse järgi kasvavas järjekorras. Seejärel tuleks arvutada iga arvu kohas selle arvu ja kõigi sellest väiksemate arvude summa ehk kumulatiivne jaotus ning viimaks jagada kumulatiivse jaotuse iga väärtus maksimumväärtusega. Saadakse Lorenzi graafikul kujutamiseks vajalikud kumulatiivsed osad.

2.2. Sarnasus ja erinevus

Sarnasuskordajate abil võrreldakse valimeid ja paljutunnuselisi objekte, nagu näiteks taime- või loomakooslusi või väljavõtteid neist. Tunnusteks on seejuures taksonoomiliste üksuste esinemine või puudumine või mingi ohtruse näitaja. Arvutada saab nii üksuste- kui ka tunnustevahelisi sarnasusi.

Sarnasus on üldmõiste, mille mõõtmiseks on palju erinevaid mooduseid

Sarnasus on üldmõiste, mille tähendust on igal üksikjuhul tarvis täpsustada. Niisamuti, nagu suurus on üldmõiste ning üksivõrdlustes kasutatakse suuruse erinevaid näitajaid – pikkus, laius, kõrgus, paksus, ümbermõõt, mass ning nende üksiknäitajate erinevaid kombinatsioone. Erinevad sarnasuskordajad mõõdavad sarnasust eri aspektidest ja seetõttu annavad erinevaid tulemusi, kusjuures ei saa väita, et mõni neist oleks ekslik. Küsimus võib olla ühe või teise mõõdiku sobivuses mingi probleemi ja andmestiku puhul. Sarnasuskordaja valimisel võib lähtuda järgmistest kriteeriumitest:

- arvutamise lihtsus,
- kasutatavus teiste autorite poolt,
- tõlgenduse loogilisus,
- tundlikkus andmestikus esinevate sarnasuse tasemete juures,
- teisendatavus teisteks sarnasuse või kauguse mõõdikuteks,
- tundlikkus haruldaste liikide suure osakaalu suhtes,
- tundlikkus liikide ohtruse erinevate vahekordade suhtes.

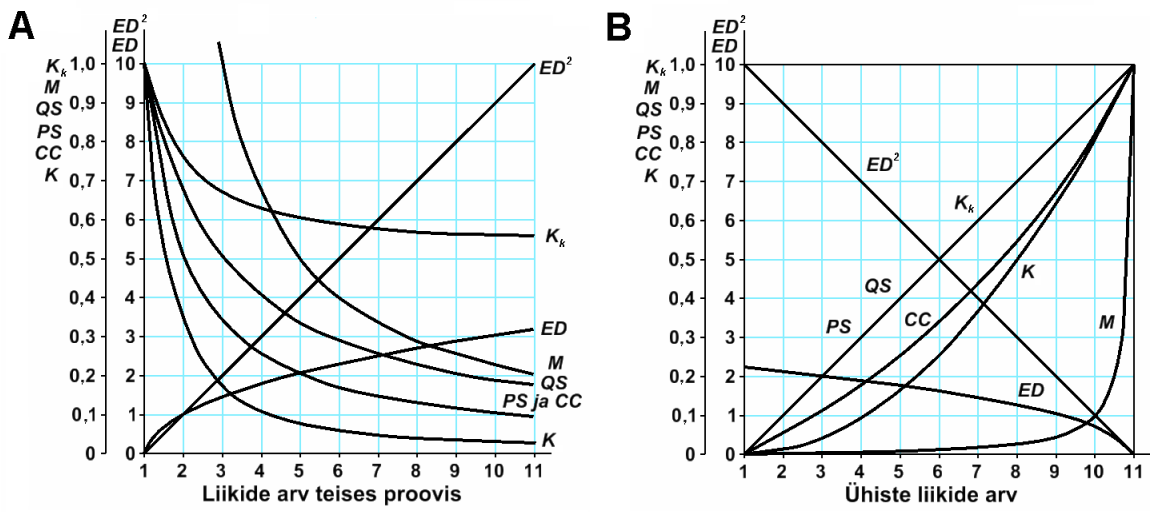
2.2.1. Sarnasuskordajad

Esimesena kasutas sarnasuse kvantitatiivset mõõtmist ja sarnasusmäära võrdlemist Poola antropoloog Jan Czekanowski uurides rahvaste antropomeetrilist sarnasust ja keelte sarnasust, muuhulgas ka balti keelte sarnasust slaavi ja germaani keeltega (Sołtysiak ja Jaskulski 1999). Czekanowski kirjutas põhjaliku ülevaate statistilistest meetoditest ja võttis kasutusele sarnasusmaatriksite graafilise esituse (joonis 2-2), kuid tema poolt algselt kasutatud sarnasusemõõdik oli lihtne tunnuste keskmine erinevus (Czekanowski 1913).

Sarnasuskordajate hulka võib arvata ka segregatsiooniindeksid. Kuna segregatsioon on ruumiline eraldatus ja segregatsiooniindeksid mõõdavad, mis määral on mingi tunnuse järgi erinevatesse gruppidesse kuuluvad objektid (näiteks isikud) ruumis grupeerunud, siis segregatsiooniindekseid käsitletakse punktmustrite paiknemissuhete peatükis (ptk 4.1.3). Paiknemissuhete indeksid näitavad seega territoriaalsete koosluste olemasolu.

	Malaxis monophyllos	*Ophrys insectifera*	*Dactylorhiza russowii*	*Dactylorhiza baltica*	*Dactylorhiza incarnata*	*Epipactis palustris*	*Listera ovata*	*Dactylorhiza fuchsii*	*Epipactis helleborine*	*Goodyera repens*	*Neottia nidus-avis*	*Platanthera chlorantha*
Malaxis monophyllos		0,05	0,04	0,07	0,08	0,1	0,06	0,06	0,03	0	0	0
Ophrys insectifera			0,26	0,08	0,05	0,1	0,03	0	0	0	0	0
Dactylorhiza russowii				0,15	0,09	0,18	0,06	0,07	0,04	0	0,05	0,06
Dactylorhiza baltica					0,2	0,29	0,15	0,11	0,08	0,01	0,03	0,02
Dactylorhiza incarnata						0,4	0,18	0,09	0,02	0	0,03	0,01
Epipactis palustris							0,34	0,26	0,12	0,01	0,05	0,01
Listera ovata								0,45	0,12	0,04	0,09	0,01
Dactylorhiza fuchsii									0,21	0,05	0,13	0,01
Epipactis helleborine										0,06	0,08	0
Goodyera repens											0,06	0
Neottia nidus-avis												0,03
Platanthera chlorantha												

Joonis 2-2. Kaheteistkümne käpaliseliigi koosesinemise suhteline sagedus Elva ja Otepää vahel paikneva kaardilehe 5444 hektarisuurustes ruutudes Czekanowski tabelina; tumedamad ruudud näitavad sagedamat esinemist samal hektaril. Välivaatlused Kalle Remm, Liina Remm, Anneli Palo, Madli Linder 2002-2008. Kõige sagedamini esinevad koos *D. fuchsii* ja *L. ovata*. Koosesinemise suhteline sagedus on mõõdetud Dice-Sørenseni sarnasuskordaja järgi.



Joonis 2-3. Mõnede sarnasuse ja erinevuse mõõdikute väärtused kahe proovi võrdlemisel, kui esimeses proovis on üks liik ja liikide arv teises proovis muutub (A) ning kui kummaski võrreldavas proovis on 10 liiki ja ühiste liikide arv muutub (B). Kõik liigid on esindatud ühe isendiga. Mõõdikud koos valemi numbriga: ED^2 – eukleidilise kauguse ruut [2-26], K_k – Kulczyński teine kordaja [2-17], *ED* – eukleidiline kaugus [2-26], *M* – Mountfordi indeks [2-20], *QS* – Dice-Sørenseni sarnasuskordaja [2-14], *CC* – Jaccardi ühisosa [2-15], *PS* – Renkose sarnasus [2-22], *K* – Vainšteini biotsönoloogiline sarnasus [2-23] (Remm 1976).

Järgnevalt esitatakse mõned koosluste ja vaatluskohtade võrdlemisel kasutatud sarnasuskordajad. Liikide esinemise/puudumise andmeid kasutavate kordajate valemites kasutatakse järgnevalt ühtset tähistust: *A* – liikide arv esimeses (suuremas) valimis, *B* – liikide arv teises (väiksemas) valimis, *C* – ühiste liikide arv. Liikide arvu asemel võivad olla mingid liikide ohtruse näitajad või muud uuritavate klasside esindatust iseloomustavad suurused. Liikidele võib valemites omistada nende arvukusele,

biomassile või muule ohtruse näitajale vastavad kaalud. Sarnasuskordajate pikemaid loetelusid koos valemitega ning nende omaduste võrdlust võib leida teistest kirjandusallikatest (Cheetham ja Hazel 1969, Bloom 1981, Janson ja Vagelius 1981, Wolda 1981, Boyle *et al*. 1990, Podani ja Miklós 2002). Üks põhjalikumaid sarnasuskordajate võrdlusi on Hubálek (1982) uuringus, kus käsitletakse 43 erinevat sarnasuskordajat. Sarnasuse ja erinevuse mõõdikute omadusi kahe proovi võrdlemisel sõltuvalt proovide mahust ja ühiste liikide hulgast näitab joonis 2-3.

Dice-Sørenseni sarnasuskordaja on enim kasutatud lihtne sarnasuse mõõdik näitab kahe kogumi kattumise määra kvantitatiivsete (esinemise/puudumise) andmete korral. Ökoloogias on eelistatud nimetust Sørenseni kordaja (*quotient of similarity*) (Sørensen 1948). Täppisteadustes on enam levinud nimetus Dice'i koefitsient (Dice 1945), mida tuleks varasema publitseerimise tõttu eelistada ka ökoloogias. Ekslikult kasutatakse ka nimetusi Czekanowski kordaja ja Czekanowski-Sørenseni kordaja – kuigi Czekanowski rajas sarnasusanalüsi ja võttis kasutusele sarnasusmaatriksite graafilise esituse, ei kasutanud ta siinesitatud sarnasuskordajat, vaid keskmist erinevust (Czekanowski 1913).

$$QS = \frac{2C}{A+B} \qquad [2\text{-}15]$$

Dice-Sørenseni sarnasuskordajat on nimetatud ka **Bray-Curtise sarnasuseks**, kuigi Bray ja Curtis (1957) algne käsitlus erineb andmete normeerimise poolest. Bray ja Curtis jagasid liikide ohtruse andmed iga liigi maksimaalse ohtrusega uurimisalal saades iga liigi ohtruse lokaalse ökoloogilise optimumi suhtes. Ka vaatluskohtades määratud summaarsed ohtrused tuleb Bray-Curtise järgi maksimumi suhtes normeerida. Yoshioka (2008) juhib tähelepanu, et paljudes tarkvarapakettides olev Bray-Curtise indeksiks nimetatav funktsioon lähteandmeid ei normeeri ja seega arvutatakse vaid lihtne Dice-Sørenseni sarnasus, mille puhul mõjutavad arvukad liigid tulemust ebaproportsionaalselt suurel määral.

Jaccardi ühisosa (*quotient of community*) näitab ühiste liikide osa kõigist liikidest, mis esinevad kahes võrreldavas kogumis. Hulgateooria terminites on see ühisosa ja ühendi suhe.

$$CC = \frac{C}{A+B-C} \qquad [2\text{-}16]$$

Kulczyński esimene kordaja näitab ühiste liikide ja erinevate liikide hulga suhet.

$$K_1 = \frac{C}{A+B-2C} \qquad [2\text{-}17]$$

Kulczyński teine kordaja näitab ühiste liikide keskmist osa võrreldavates kogumites.

$$K_k = \frac{C}{2}\left(\frac{1}{A}+\frac{1}{B}\right) = \frac{C(A+B)}{2AB} \qquad [2\text{-}18]$$

Goweri kaalutud sarnasus ühendab erijuhtudena teisi sarnasuskordajaid (Gower 1971a). Sarnasus arvutatakse valemi järgi

$$S = \frac{\sum_k s_k \delta_k}{\sum_k \delta_k}, \qquad [2\text{-}19]$$

kus k tähistab tunnust, δ_k on tunnuse k kaal, s_k on sarnasus tunnuse k osas.

Kui tunnust k saab selle võrdluse puhul kasutada, siis lihtsamal juhul on kaal 1, kui ei saa, siis δ_k = 0. Kaheväärtuseliste kaalude asemel võib kasutada ka reaalarvulisi tunnuse mõju määravaid kaale. Kvalitatiivse tunnuse sama kategooria puhul on s_k 1, erineva kategooria puhul s_k = 0. Kvantitatiivse

tunnuse puhul arvutatakse s_k järgmiselt:

$$s_k = 1 - \frac{|x_i - x_j|}{R_k}, \qquad [2\text{-}20]$$

kus x_i ja x_j on tunnuse k väärtus vaatlusel i ja vaatlusel j, R_k on tunnuse k väärtuste haare. Kui $x_i = x_j$, siis $s_k = 1$, kui x_i ja x_j on tunnuse k äärmuslikud väärtused, siis $s_k = 0$.

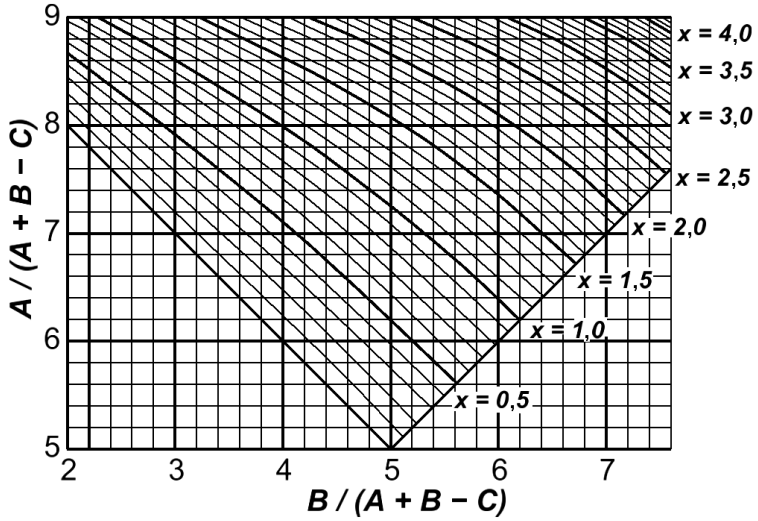
Mountfordi indeks M (Mountford 1962) tugineb seaduspärasusele, et juhuslikult esinevate tunnuste arvu suurendamiseks tuleb valimi mahtu suurendada eksponentsiaalselt. Võrreldavate kohtade tunnustena võib käsitleda neis esinevaid liike, elupaigatüüpe või muid binaarselt mõõdetavaid omadusi. Kui eksponentsiaalne seos kehtib, peaks Mountfordi indeks näitama sama suurt sarnasust ka juhul, kui kahte kogumit võrreldakse erineva suurusega valimite alusel. Teiste sarnasusmõõdikute abil saadud tulemus sõltub valimi mahust. Mountfordi indeksit defineeriv seos on

$$e^{AM} + e^{BM} = 1 + e^{(A+B-C)M} \qquad [2\text{-}21]$$

Mountfordi indeks on kirjanduses sageli kujul

$$M = \frac{2C}{2AB - (A+B)C}, \qquad [2\text{-}22]$$

mida Mountford (1962) soovitas vaid esimese lähendina ja lähendusiteratsioonide algväärtusena. Mountfordi indeksi väärtust saab leida ka nomogrammilt (joonis 2-4). Kui liikide sagedused koosluses on logaritmilise jaotusega, siis vastab Mountfordi indeksi väärtus tõenäosusele, et kummastki võrreldavast kooslusest võetud juhuslikud isendid kuuluvad samasse liiki.



Joonis 2-4. Mountfordi indeksi leidmise nomogramm, kus $x = M \cdot (A + B - C)$ (Mountford 1962 muudetult). Nomogrammi kasutamiseks tuleb valemi 2-22 abil arvutada Mountfordi indeks M, ning parameetrid $A / (A + B - C)$ ja $B / (A + B - C)$. A on liikide arv liigirikkamas valimis, B on liikide arv teises valimis, C on ühiste liikide arv.

Renkose sarnasusprotsenti (Renkonen 1938) kasutatakse sageli entomoloogias, see näitab isendite liikide vahel jaotumise sarnasust.

$$PS = \sum_{i=1}^{s} \min(x_{i1}, x_{i2}), \qquad [2\text{-}23]$$

kus s on liikide arv mõlemas võrreldavas proovis kokku, x_{i1} on i-nda liigi isendite protsent esimeses proovis, x_{i2} on i-nda liigi isendite protsent teises proovis. Isendite protsent arvutatakse kummaski proovis eraldi, selle proovi isendite koguhulga suhtes.

Vainšteini biotsönoloogilise sarnasuse kordaja (K_σ) võtab kokku liikide arvukuste sarnasuse (K_α) ja liigilise koosseisu sarnasuse (K_β). Liikide arvukuste sarnasuse võib arvutada näiteks Renkose järgi [valem 2-23]. Kasutada võiks ka ruutjuurt sellest koefitsiendist, et koondnäitaja mõõtühik ja suurusjärk oleksid samad, mis on komponentidel.

$$K_\sigma = K_\alpha \cdot K_\beta \qquad [2\text{-}24]$$

Ružička sarnasusindeks (Ružička 1958) võrdleb liigi arvukuse ja dominantsi kombinatsiooni tegelikku väärtust selle maksimaalse võimaliku väärtusega. Ružička indeksit on tõlgendatud ka kahes võrreldavas vaatluskohas (1 ja 2) leitud liikide ($i = 1...s$) suuremate ja väiksemate arvukuste x summa suhtena (Pielou 1984).

$$RI = \frac{\sum_{i=1}^{s} \min(x_{i1}, x_{i2})}{\sum_{i=1}^{s} \max(x_{i1}, x_{i2})} \qquad [2\text{-}25]$$

Korrelatsioonikordajad sarnasuse mõõdikutena või nende vastandväärtused ($1-R$) erinevuse mõõdikutena on kasutamiskõlbulikud vaid numbriliste tunnuste korral. Kui ruuthälvete minimeerimine ei ole põhjendatud (tunnuste jaotus ei lähene normaaljaotusele), tuleks eelistada astakkorrelatsiooni kordajaid (ptk 1.5.2.3).

2.2.2. Statistiline kaugus

Rahvusvahelise sõna *distants* asemel eelistame selles raamatus eestikeelseid vasteid vahemaa ja kaugus, kasutades neid sünonüümidena. Vahemaa on liitsõna, kuid neutraalsem, tähistades nii kaugust kui ka lähedust.

Statistiline kaugus (vahemaa) iseloomustab kahe võrreldava objekti erinevust mistahes tunnuste kogumi poolest. Tunnuste arv ei ole piiratud ning nende hulka võib kuuluda ka paiknemine ruumis. Statistiline kaugus jaguneb blokk-kauguseks (*city block distance*), eukleidiliseks kauguseks (*Euclidean distance*) ja Mahalanobise kauguseks (*Mahalanobis distance*).

Blokk-kaugus (*BD*) on tunnuste *i...n* erinevuste absoluutväärtuste summa

$$BD = \sum_{j=1}^{n} \left| x_i - x'_i \right|, \qquad [2\text{-}26]$$

kus i on tunnuse indeks, x_i on i-nda tunnuse väärtus esimesel objektil, x'_i on i-nda tunnuse väärtus teisel objektil.

Võrreldavuse parandamiseks erineva tunnuste arvu korral kasutatakse ka keskmist erinevuse absoluutväärtust, seega *BD/n*.

Eukleidiline kaugus (*ED*) arvutatakse valemi järgi

$$ED = \sqrt{\sum_i \left(x_i - x'_i\right)^2} \qquad [2\text{-}27]$$

ED näitab kahe objekti vahelist kaugust paljumõõtmelises ruumis, kus iga tunnust vaadeldakse kui ruumikoordinaati. Tunnusruum on nii mitme-mõõtmeline, kui palju on kasutatavaid tunnuseid. Eukleidilise kauguse kasutamisel eeldatakse, et tunnused on sõltumatud ja tunnuste ühikulised muutused on samaväärsed. *ED* eelis on, et see rahuldab eukleidilise meetrika kolmnurga reeglit. See tähendab, et kolme suvalise objekti vahelised kaugused võimaldavad alati konstrueerida kolmnurga, mille pikim külg on väiksem või võrdne kahe lühema külje summaga. Kasutatakse ka ruutkaugust ED^2.

Algkujul sõltub *ED* väärtus kasutatavate tunnuste arvust, mille tõttu on tunnuste erineva arvu korral põhjust kasutada tunnuste keskmist eukleidilist kaugust. Kui määratlemata väärtusega tunnused jätta eemaldamata, siis statistiline kaugus loeb valimid sarnaseks ka nende tunnuste poolest, mis on mõlemas valimis mõõtmata või esinevad teabe puudumist tähistava väärtusega.

Ruumiliste andmete analüüsi puhul on võimalik kasutada asukoha koordinaate koos tunnusruumi koordinaatidega, millega loetakse ruumiline lähedus ka üheks sarnasuse komponendiks (Oliver ja Webster 1989).

Mahalanobise kaugus on tunnuste kovariatsioonidega standardiseeritud statistiline kaugus. Punkti ja jaotuskeskme vahemaa korral on see logaritm jaotusfunktsiooni tõenäosuse tihedusest pluss konstant. Kuna kovariatsioonimaatriksid kirjeldavad tunnuste vahelisi seoseid, siis ei eelda Mahalanobise kauguste kasutamine tunnuste sõltumatust. Kauguse isoliinid tunnusruumi teljestikus on eukleidilise kauguse korral sfäärilised ja Mahalanobise kauguse puhul ellipsid.

Prasanta Chandra Mahalanobis (1893–1972) oli India õpetlane ja riigitegelane ning koos Jawaharlal Nehru ja teistega üks India esimeste viisaastakuplaanide väljatöötajatest. P.C. Mahalanobis tegeles uuenduslike rakendusuuringutega antropomeetria, suuremahuliste valikuuringute ja loodus-nähtuste (üleujutuste) statistilise modelleerimise vallas ja oli vabameelne riskialdis teadusorganisaator. Majandusplaanis toetas rasketööstuse arendamist koos käsitööstuse subsideerimisega.

2.3. Klassifitseerimine

Klassidesse jaotada saab objekte, tunnuseid ja asukohti. Ruumiandmete klassifitseerimisest tuleb lähemalt juttu paiknemismustrite kirjeldamise peatükis (ptk 4) ja liikide leviku modelleerimise meetodite juures (ptk 5.6.6 ja 5.6.10.1). Klassifitseerimisskeemid ehk klassifikatsioonipuud ehk otsuste puud võivad baseeruda kindlatel reeglitel (*rule based classification*) või tõenäosusjaotustel. Viimase variandi hulka kuuluvad suurima tõepära klassifikatsioonid (*maximum likelihood classification – MLC*) ja hägused klassifikatsioonid (*fuzzy classification*), mille puhul ei ole objekti klassikuuluvus kindel, vaid tõenäosuslik. Klassid võivad olla klassifitseerimise eelselt ehk *a priori* ette antud või moodustatakse klassid alles klassifitseerimise käigus (*a posteriori*). Esimest varianti nimetatakse kaugseirekujutise klassifitseerimise meetodite liigitamisel ka valvega (*supervised*) ja teist valveta klassifitseerimiseks (*unsupervised*).

Kogemuse-eelsed klassid on ette antud, kogemusejärgsed klassid moodustuvad klassifitseerimise tulemusel

Looduse süsteemi saab käsitleda nii kogemuse-eelse kui ka kogemusejärgse klassifitseerimise näitena. Ükski organism ei sünni liigi, perekonna, sugukonna, seltsi ja kõrgema taksoni sildiga. Taksonid on moodustatud ilma õpetusandmeteta, organismide omadusi omavahel võrreldes ja organisme rühmitades. Sellest lähtuvalt on looduse süsteem kogemusejärgne. Teisalt, kui uurija määrab kogutud näiteid, siis on tema ülesandeks liigitada vaatlused või objektid eelnevalt olemasolevatesse kategooriatesse. Üksikisiku jaoks on kategooriad *a priori* ette antud.

Klasside eristamiseks saab kasutada nii kõigi tunnuste kõiki väärtusi kui ka tunnuste kombinatsioone. Klassifikatsioonimeetodite paindlikkus on nii nende eeliseks kui ka puuduseks, sest suur paindlikkus tähendab parima klassifikatsiooni otsimise variantide paljusust. Kõikide võimalike klassifikatsioonimeetodite läbiproovimine on paraku ebareaalne ja tuleb mingil määral asendada eelteadmise või eelarvamusega.

Lisaks järgnevalt kirjeldatud meetoditele kuuluvad sagelikasutatavate klassifitseerimismeetodite hulka otsuste puud (ptk 3.4.2) ja tolerantsipiiride kombineerimised (ptk 5.6.3). Mõlema meetodi puhul toimub klassifitseerimine kriteeriumite alusel.

2.3.1. Klasteranalüüs

Klasteranalüüsiks (*cluster analysis*) nimetatakse objektide hulga mitme tunnuse järgi liigendamist alamhulkadeks ehk kobarateks ehk klastriteks ehk rühmadeks, millesse kuuluvad mingis mõttes lähedased elemendid. Klasteranalüüsi puhul võib rühmade arv olla ette antud (*k-means clustering*) või ühendatakse objekte klastripuuga järgimööda (*tree clustering, bottom up clustering, agglomerative clustering*), või siis jagatakse andmestikku järgimööda, kuni igasse rühma jääb vaid üks vaatlus (*divisive method, top down clustering*). Klasteranalüüsile peaks eelnema rühmitatavate objektide sarnasuse või vahemaa mõõtmine (ptk 2.2.1 ja 2.2.2).

Klasteranalüüsi kasutatakse eelkõige vaatluste klassikuuluvuse määramiseks ja seoste otsimiseks tunnuste vahel, jälgides vaatluste liigitumist samadesse või erinevatesse klastritesse. Klastritesse rühmitada saab nii objekte kui ka tunnuseid.

Klasteranalüüs rühmitab, aga ei tõesta rühmade olemasolu

Klasteranalüüs kuulub kirjeldava andmeanalüüsi meetodite hulka, mis ei nõua *a priori* hüpoteesi klastrite olemasolu kohta. Klasteranalüüs püüab leida olemasolevate objektide parimat klassifikatsiooni aga ei üritagi tõestada klastrite olemasolu, kuid klastrid on kasulikud mitmekesisuse üldistamiseks ja üksikobjekti omaduste prognoosimiseks klastri üldiste omaduste järgi. Klasteranalüüsi saab kasutada ka *a priori* eeldatavate klastrite olemasolu tõendamiseks.

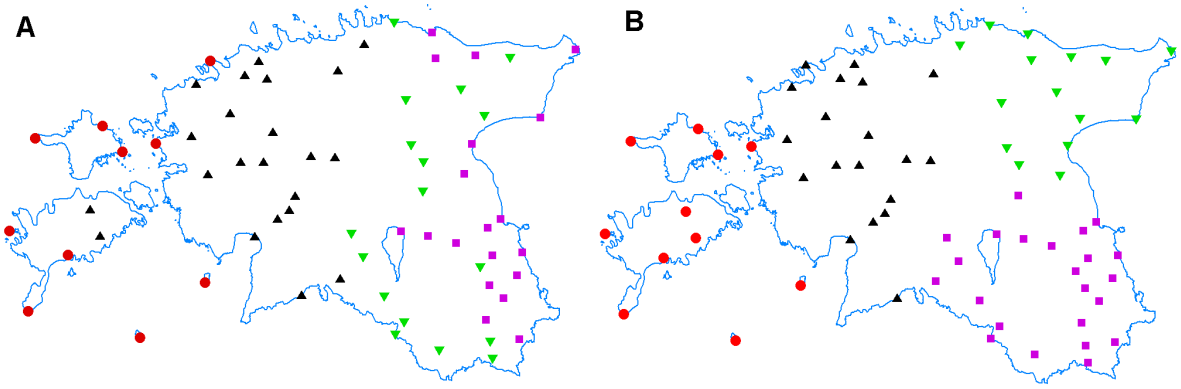
Klasteranalüüs on objektiivne selles mõttes, et parimate klastrite otsimine toimub kindlate reeglite järgi. Klastrite otsimise reeglid tuleb aga enne analüüsi uurijal endal valida. Eri meetoditel saadud klassifikatsioonid võivad olla väga erinevad. Enamgi veel, klasteranalüüs ei tee vahet, kas klastrite olemasolu on üldse statistiliselt oluline. Ka juhuslikke arve saab takistusteta ja täielikult rühmitada.

Klasteranalüüsi kasutas esimesena Robert Tryon (1939), kes töötas California ülikooli psühholoogia osakonnas. Oma doktoritöö lõpetas R. Tryon 1928 olles teinud kõik arvutused käsitsi; hiljem kurtis ta arvutite-eelsel ajal arvutamistele raisatud nooruse üle (Tryon ja Bailey 1970). Klasteranalüüsile on lähedased klassifikatsioonipuude ja regressioonipuude meetodid, mis prognoosivad vastavalt kas nomiaalse muutuja klassikuuluvust või pideva muutuja väärtust jagades andmestiku argumenttunnuste järgi osadeks (ptk 3.4.2).

Klasteranalüüsil on palju variante, mis erinevad lisaks klastrite arvu etteantusele objektidevahelise kauguse arvutamise meetodite ja kobarate moodustamise reeglite poolest (Pielou 1984, Statsoft 2011).

- Lähima naabri otsimine (*nearest neighbour, single linkage*) (kaldub moodustama pikki objektiahelaid).
- Kobarate läbimõõdu minimeerimine (*complete linkage, farthest neighbour*) (töötab hästi ühesuuruste kobarate korral).
- Kobarapaari liikmete omavaheliste kauguste keskmise maksimeerimine (*unweighted pair-group average*). Klastrite vahemaad ei mõõdeta mitte klastrikeskmete, vaid üksikliikmete vahel. Ühendatakse need objektid või klastrid, mille puhul kaalutud kaugus on minimaalne.
- Kahe kobara liikmete omavaheliste kauguste kaalutud keskmise maksimeerimine (*weighted pair-group average*). Kaaludena kasutatakse objektide arvu moodustatud kobarates. Sobib eelmisest meetodist paremini, kui kobarad on erineva suurusega.
- Kauguse maksimeerimine kobarate raskuskeskmete vahel (*unweighted pair-group centroid*).
- Kaalutud kauguse maksimeerimine kobarate raskuskeskmete vahel (*weighted pair-group centroid*). Kauguse kaaluna kasutatakse kobara suurust. Ühendatakse need objektid või klastrid, mille puhul kaalutud kaugus on minimaalne. Eelis eelmise meetodi ees on kobarate suuruse arvestamine.
- Kobarasisese summaarse ruuthälbe minimeerimine (*Ward's method*) (Ward 1963). Wardi meetod annab suhteliselt kompaktsed ja mitmesuguse suurusega grupid. Selle meetodi puhul liidetakse vaatlus grupiga, mille puhul objektidevaheliste kauguste ruutude summa on minimaalne. Üksikute tunnuste domineeriva mõju vältimiseks on soovitav muutujad standardiseerida.

Klasteranalüüsi saab kasutada muuhulgas regioonide eristamiseks kindlates punktides mõõdetud arvandmete järgi. Näiteks saab Eesti ilmavaatlusjaamad kolmekümne kahe aasta (1966–1998) kuude keskmiste sademete hulkade järgi klasteranalüüsi abil nelja rühma jaotada (joonis 2-5A). Kui kuude sademete hulkadele lisada ka vaatlusjaamade koordinaadid, siis saame ruumiliselt pidevamad klimaatilised piirkonnad (joonis 2-5B).



Joonis 2-5. Eesti ilmavaatlusjaamade klastrid aastate 1966–1998 kuude keskmiste sademete hulga alusel (A) ning kuude keskmiste sademete hulga ja lisaks ka asukoha koordinaatide alusel (B). Kasutatud tunnuste poolest sarnased vaatlusjaamad liigituvad samasse rühma ja on kujutatud sama sümboliga. Iga kuu keskmise sademete hulga alusel koonduvad samasse rühma klimatoloogiliselt sarnased jaamad. Asukohakoordinaatide lisamisel üksikkuude keskmistele saadakse klastrid, mis on geograafilises ruumis kompaktsemad ja sobivad paremini klimaatiliste regioonide eristamiseks ja piiritlemiseks.

2.3.2. Bayesi klassifikaatorid

Määratavate klasside esinemise tõenäosuse kogemuse-eelse hinnangu olemasolu korral saab seda ühendada vaatlusandmetest leitud tõenäosusjaotustega. Kogemuse-eelne teadmine võib pärineda varasematest uuringutest, eelarvamustest või etteantud tingimustest (näiteks vajatakse liikide paiknemise prognoosi eeldusel, et liikide arvukuse vahekord on ette antud). Tõenäosusjaotuste kasutamine võimaldab määrata mitte ainult iga koha kõige tõenäolisemat väärtust (klassifikatsiooniüksust), vaid saada ka teiste väärtuste tõenäosuse. Tinglikke tõenäosusi kasutavaid mudeleid nimetatakse **Bayesi mudeliteks** (*Bayesian models*).

Teoreetilises statistikas on **Bayesi järeldamine** (*Bayesian inference*) statistilise järeldamise meetod, milles tõendite mõju uskumuse tasemele määratakse Bayesi valemi abil [1-1]. Tõendite eelset uskumuse taset nimetatakse aprioorseks, tõendite järgset aposterioorseks tõenäosuseks. Bayesi meetodiga arvesse võetav eelnev kogemus (eeljaotus) ja saadav tulemus (järeljaotus) võimaldab käsitleda erinevaid võimalikke stsenaariume ja annab tänu mitmekesisemale sisendile ka mitmekesisema väljundi.

Lihtne ehk **naiivne Bayesi klassifikaator** (*naïve Bayes classifier*) liigitab vaatluse sinna klassi, mille puhul on iga argumenttunnuse järgi eraldi hinnatud tõenäosuste korrutis kõige suurem. Kui otsitav klass pole mõne argumenttunnuse mõne väärtuse korral õpetusandmetes kordagi esindatud, siis on selle klassi tõenäosus argumenttunnuse selle väärtuse järgi null ja selle klassi tõenäosus on null sõltumata teiste tunnuste väärtustest, sest nulliga korrutamise tulemus võrdub nulliga.

$$y = \arg\max_{c_j \in C} P(c_j) \cdot \prod_i P(h_i \mid c_j), \qquad [2\text{-}28]$$

kus y on klassifitseeritavale objektile omistatav üksus, C on kõigi võimalike üksuste kogum, H on hüpoteeside (argumenttunnuste kõigi väärtuste) kogum, c_j on üksik üksus ja h_i on üksik hüpotees.

Naiivse Bayesi klassifikaatori naiivsus seisneb argumenttunnuste koosmõjude ja omavahelise korreleerumise eiramises. Tunnused empiirilistes andmestikes on enamasti aga vähemalt kuigivõrd omavahel seotud. Arvutusliku lihtsuse tõttu on naiivne klassifikaator valdav Bayesi järeldamise meetod. Põhimõtteliselt saaks tunnuste koosmõjusid kaasata kõigi tunnusekombinatsioonide teisen-

damisega omaette tunnusteks, aga kõigi kombinatsioonide esindatus õpetusandmetes on reaalne vaid variantide väga vähese hulga juures.

Naiivne Bayesi klassifikaator selgitab suurima tingliku tõenäosusega klassi, kuid ei arvesta seletavate tunnuste koosmõju

Optimaalne Bayesi klassifikaator (*Bayes optimal classifier*) liigitab vaatluse sinna klassi, mille puhul on suurim summa iga argumenttunnuse väärtuse sagedusest õpetusandmetes korrutatud selle klassi tõenäosus argumenttunnuse sellise väärtuse korral.

$$y = \arg\max_{c_j \in C} \sum_i P(c_j \mid h_i) \cdot P(h_i) \qquad [2\text{-}29]$$

Optimaalne Bayesi klassifikaator on suurema arvutusmahu tõttu rakendatav vaid võimalike klasside ja argumenttunnuste väärtuste vähese arvu korral.

Tinglike tõenäosusjaotuste järgi on tehtud suurem osa satelliidipildipõhisest maakattekaardistusest. Arvutuste kiirendamiseks ei kasuta tarkvaralahendused aga õpetusalade pikslite väärtuste sagedusjaotusi, vaid neid sagedusjaotusi normaaljaotuse eeldusel modelleerivaid keskväärtusi ja hajuvusmaatrikseid.

2.3.3. Näidistega võrdlemine

Näidistega võrdlemise (*case-based reasoning, similarity-based reasoning, analogue matching*) puhul otsitakse prognoostavale vaatlusele, objektile või kohale olemasolevate andmete hulgast kõige sarnasemaid näidiseid. Näidiste abil saab prognoosida nii pideva muutuja väärtust kui ka nomiaalse muutuja klassikuuluvust (klassifitseerimine). Näidiste kasutamise ainus eeldus on, et objektide sarnasust peab olema võimalik arvuliselt mõõta ja võrrelda. Näidiste abil prognoosivat süsteemi käsitletakse lähemalt statistilise modelleerimise peatükis (ptk 3.4.6), siin tuleb juttu eelkõige näidiste abil klassifitseerimisest.

Kujutise töötlemisel kasutatakse etalonidena otsitavate objektide kirjeldusi, mida nimetatakse **signatuurideks**. Signatuur on õpetusalade andmetel koostatud statistiline üldistus, mille põhikomponendid on tunnuste keskväärtused ja varieeruvuse näitajad. Näidistega klassifitseerimise põhimõtteline erinevus signatuuridega klassifitseerimisest on klassifitseeritavate objektide algkujul esindajate, mitte nende statistiliste üldistuste kasutus (joonis 2-6). Kui võrrelda arvutile programmeeritud klassifitseerimise vahendeid ja automatiseerimata meetodeid, siis on tehisnärvivõrk analoogiline eksperdi käest küsimisele, otsuste puud määramistabeli kasutamisele, signatuuride kasutamine taksonite kirjeldustega võrdlemisele ning näidistele tuginev süsteem muuseumis hoitavate eksemplaridega võrdlemisele.

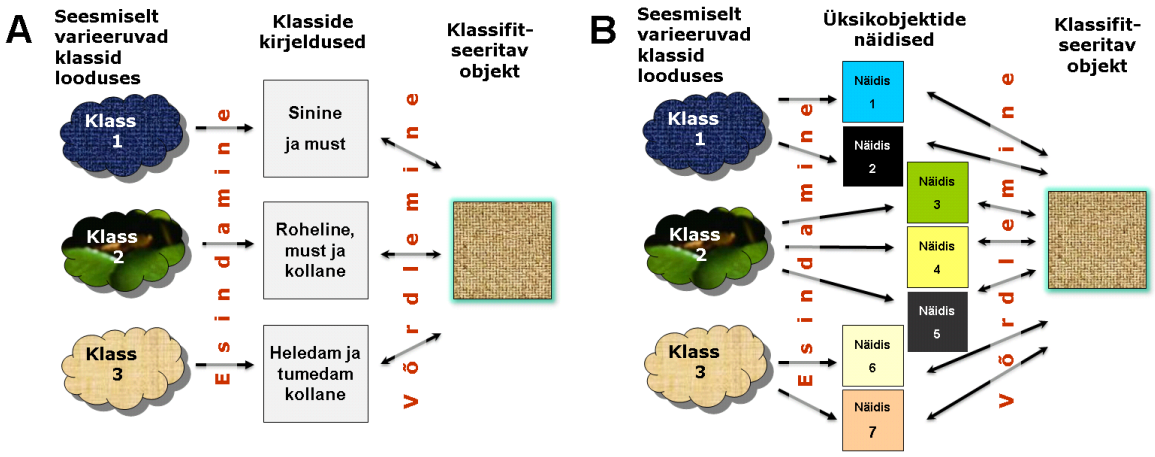
Näidistega võrdlemise teel klassifitseerimine on reeglina arvutusmahukam kui keskväärtuste kasutamine, kuna heterogeenseid objekte peab esindama mitu näidist. Mida rohkem on aga näidiseid, seda rohkem tuleb arvutada sarnasusi ja neid omavahel võrrelda, et kõige sarnasemad näidised näidiste baasist üles leida. Heterogeensete objektide adekvaatsem esindatus on küll näidistega klassifitseerimise üks peamisi eeliseid üldistuste ees, kuid viimane on arvutusmahukas.

Näidistega võrdlemise teiseks puuduseks on variantide rohkus sarnasuse määramisel kasutatavate tunnuste olulisuse ja optimaalse panuse hindamisel ning sobivaima tunnustekomplekti leidmisel. Ligikaudu samavõrd kehtivaid ja samas erinevaid hinnanguid võib saada paljude erinevate komplektidega tunnustest ja näidistest. Klassifitseerimise tulemus sõltub päris palju sellest, kuidas näidised ole-

masolevate andmete hulgast valitakse, kas näidiste hulgas leidub igale vaatlusele sarnane eeskuju, kuidas sarnasusi mõõdetakse, kui mitut sarnasemat näidist ja/või kui sarnaseid näidiseid klassifitseerimisel või muude hinnangute arvutamisel arvesse võetakse. Sarnasuse järgi klassifitseerimine on enamasti ka müra suhtes tundlikum kui klassifikatsioonipuud (Breiman *et al*. 1984)

Kolmandaks puuduseks on asjaolu, et sarnasuse järgi hindamine ei pruugi ära tunda limiteerivaid faktoreid. Mingi koht võib paljude ebaoluliste tunnuste poolest olla sarnane liigi poolt kasutatavate biotoopidega, aga kui üks liigi jaoks limiteeriv faktor on sellele liigile talumatu, siis liiki sellest kohast otsida ei tasu.

Näidistega võrdlemine on mitmevariandiline, arvutusmahukas ja ei erista limiteerivaid tunnuseid



Joonis 2-6. Tundmatu objekti klassikuuluvuse määramine kirjelduste järgi (A) ja näidiste järgi (B).

2.3.5. Jenksi algoritm

George F. Jenks (1977) esitab viis soovitust horopleetkaartide kujundamiseks.

- Standardiseeri andmed. Näiteks indiviidide arvu asemel kujuta kaardil indiviidide tihedust.
- Selle asemel, et kujutada pideva muutuja väärtusi värvitooni või heleduse sujuva ülemineku abil, jaga väärtused piiratud arvuks klassideks. Ei ole mõtet kaardil kujutada suuremat arvu pooltoone, kui vaataja suudab eristada. Katsed on näidanud, et inimene suudab eristada seitset halltooni astet ja kuni 11 värvitooni.
- Kasuta järjestatud väärtuste esitamiseks värvide sujuvat üleminekut.
- Püüa lihtsuse poole. Kaardilt tuleks eemaldada kõik liigne.
- Vali hoolega klassifitseerimismeetodit.

Klassipiiride leidmiseks soovitab Jenks (1967, 1977) Fisheri algoritmi (Fisher 1958), millel on kaks varianti. Esimese kohaselt summeeritakse absoluuthälbed mediaanist, teise puhul ruuthälbed keskmisest. Fisheri algoritmi rakendust pideva muutuja kartograafilisel kujutamisel on hakatud nimetama loomulike klasspiiride (*natural breaks*) otsimiseks ja seda tuntakse **Jenksi algoritmina**. Jenksi algoritm toimib järgnevalt.

- Arvuta kogu andmestiku keskväärtus ja ruuthälvete summa [SDAM].
- Jaga andmestik klassideks, arvuta iga klassi keskväärtus ja ruutude summa klasside keskväärtustest. Summeeri klasside ruuthälvete summad [SDCM].
- Arvuta dispersioonide vastavustase (*goodness of variance fit – GVF*).

$$GVF = \frac{SDAM - SDCM}{SDAM} \qquad [2\text{-}30]$$

- Fikseeri selle korduse GVF väärtus.
- Korda protseduuri alates sammust 2 uue klassideks jagamise variandiga.
- Parim klassifikatsioon on see, mille puhul GVF on suurim.

Kui ruuthälbed ei sobi muutuja varieeruvust kirjeldama, võib nende asemel kasutada absoluuthälbeid mediaanist. Jenks lähtus printsiibist, et optimaalse üldistamise tulemusel peaksid suhtelised või absoluuthälbed üldistatud ja üldistamata andmekihi vahel paiknema kogu pinnal võimalikult ühtlaselt.

Programmid MapInfo, ArcView ja ArcGIS kasutavad Jenksi algoritmi teemakaardi kujundamisel pideva muutuja loomulike klassipiiride järgi.

2.3.6. Diskriminantanalüüs

Diskriminantanalüüsiga ehk eristusanalüüsiga (*discriminant function analysis*) otsitakse tunnuseid ja funktsioone, mis eristavad kõige paremini kahte või enamat vaatluste rühma. Diskriminantanalüüsil moodustatakse iga eristatava klassi jaoks regressioonivõrrand, mille abil saab arvutada vaatluse klassikuuluvuse funktsiooni. Kasutaja saab klassikuuluvuse hinnangule omalt poolt lisada *a priori* tõenäosusi. Meetodit saab kasutada nii hüpoteeside testimisel kui ka kirjeldaval andmeanalüüsil vaatluste klassifitseerimiseks. Meetod on arvutuslikult lähedane dispersioonanalüüsile (ptk 1.5.4) ja seetõttu kehtivad ka dispersioonanalüüsi eeldused:

- tunnused peavad olema normaaljaotusega,
- varieeruvus peab olema homogeenne (samasugune kõigis gruppides ka siis, kui gruppide keskväärtused on erinevad),
- gruppe eristavad faktorid ei tohi olla üksteist dubleerivad (tugevasti korreleerunud), st peavad olema vähemalt teatud määral unikaalsed.

2.3.7. Klassifikatsioonitäpsuse hindamine

Klassifikatsioonimudelite võrdlevaks sobivuse hindamiseks ja klassifikatsiooni optimaalse detailsuse leidmiseks tuleks kontrollida klassifikatsioonireeglite sobivust sõltumatute andmetega. Klassifikatsioonitäpsuse numbrilised mõõdikud sobivad ka kaheväärtuselise tunnuse (näiteks esinemise või puudumise) mudeli tundlikkuse seadistamiseks. Kui mudeli otsene väljund on esinemistõenäosus, siis tuleb valida piirväärtus, millest väiksemad tõenäosused klassifitseeritakse puudumisteks ja suuremad esinemisteks. Erinevate piirväärtuste ehk mudeli tundlikkuse juures on hinnangute täpsus erinev. Kõige täpsemaid hinnanguid pakkuva mudeli leidmiseks tuleb leida sobiv piirväärtus, mille väljaselgitamiseks on tarvis mõõta mudelist saadud hinnangute vastavust kontrollandmetega.

2.3.7.1. Vigade maatriks

Klassifikatsioonitäpsuse üksikasjaliku kirjeldamise vahendid on **vigade maatriks** (*confusion matrix, error matrix, misclassification matrix*) ja hinnangutäpsuse kaardid. Vigade maatriks on $k \times k$ maatriks, kus k on klassifikatsiooniüksuste arv. Nii maatriksi ridades kui ka veergudes on klassifikatsiooniüksused, maatriksi lahtrites on vaatluskohtade arv klassikombinatsioonides. Tavapäraselt on veergudes prognoositud klassifikatsiooniüksused ja ridades otseselt mõõdetud või muul põhjusel tõeseks loetavad klassid. Õiges vastavuses olevad juhud on loendatud maatriksi diagonaalil ja kõigi erinevatesse klassidesse liigitatud juhtude sagedused teistes lahtrites. Kaheväärtuselise tunnuse vigade maatriks on 2 × 2 risttabel (tabel 3).

Vigade maatriksist saab arvutada klassifitseerimistäpsuse mõõtmise ja võrdlemise statistikuid. Põhilised neist on:

- kapa kordaja, mis normeerib klasside leitud kokkulangevust klasside juhuslikul paiknemisel oodatava kokkulangevusega (vt järgmine alapeatükk 2.3.7.2);
- õigesti klassifitseeritud vaatluste osa kõigist vaatlustest (klassifitseerimise edukus – *prediction success*);
- üksiku kategooria õigesti klassifitseeritud juhtude osa selle kategooria vaadeldud esinemissageduse suhtes (tundlikkus selle kategooria suhtes – *prediction sensitivity*, kaheväärtuselise muutuja negatiivsete juhtude puhul – *prediction specificity*);
- õigesti klassifitseeritud üksiku kategooria osa selle kategooria sagedusest hinnangutes (võimsus selle kategooria suhtes – *predictive power*, kaheväärtuselise muutuja positiivsete juhtude suhtes on võimsus positiivne – *positive predictive power*, negatiivsete juhtude puhul negatiivne võimsus – *negative predictive power*);
- keskmine klassifitseerija täpsus (*producer's accuracy*), mis on õigesti klassifitseeritud üksuste keskmine osa tõeseks loetavate esinemissageduste suhtes;
- keskmine klassifikatsioonitäpsus kasutaja jaoks (*user's accuracy*), mis on õigesti klassifitseeritud üksuste keskmine osa klassifitseerimisel saadud esinemissageduste suhtes;
- Hanssen-Kuiperi skoor (*true skill statistic*) , mis on õigete positiivsete osa pluss õigete negatiivsete osa miinus üks (ptk 2.3.7.3);
- šansside suhe, mis on õigete ja väärate variantide sageduse suhe (ptk 1.4.3.3 ja 2.3.7.4).

Tabel 3. Kaheväärtuselise tunnuse vigade maatriksi lahtrid ja koondnäitajad.

		Vaadeldud sagedused		
		Esineb	Puudub	Kokku
Prognoositud sagedused	Esineb	õiged positiivsed	väärpositiivsed	prognoositud esinemised
	Puudub	väärnegatiivsed	õiged negatiivsed	prognoositud puudumised
	Kokku	vaadeldud esinemised	vaadeldud puudumised	vaatluste koguarv

Uurimuse eesmärgist lähtudes ei pruugi kõigi üksikvigade tähendus olla sama suur. Vigade tähenduse arvestamiseks ei pruugi kõiki kõrvalekaldeid vääraks lugeda (näiteks lehtmetsa liigitamine segametsaks ei ole, aga okasmetsaks on viga). Samuti saab vigade maatriksi väljadele omistada erinevad kaalud.

2.3.7.2. Kapa kordaja

Kaugseires tuntud nominaalsete tunnuste vastavust näitav **kapa kordaja** (*Cohen's kappa index of agreement – KIA* ehk häälduse järgi *KHAT*) ehk Coheni vastavuskordaja ehk lihtsalt kapa leiutaja on enamiku õpikute järgi psühholoog Jacob Cohen (1960), kuigi sama valemit kaks korda kaks vigade maatriksist koondnäitaja arvutamiseks kasutas juba palju varem klimatoloog P. Heidke (1926) – sellest kapa kordaja teine nimetus Heidke skoor (*Heidke skill score*). Vigade maatriksis olevate sageduste võrdlemist juhuslikkuse korral oodatava sagedusega kasutas ka juba F. Galton (1892). Cohen (1960) on muutunud viitamise standardiks, kuna kirjeldab kapa kordaja arvutamist ja selgitab selle omadusi kõige põhjalikumalt. Inglise keeles kirjutatakse selle kordaja nimi kahe p-ga, kuid kuna eesti keeles on tavaks kreeka keele κ tähe nime ühe p-ga kirjutada, siis tuleks eestikeelses tekstis järgida eesti-pärasemat tava.

Kapa kordaja mõõdab samade andmete kahe klassifikatsiooni suhtelist vastavust. Kapa eeldab, et kategooriad on üksteisest sõltumatud ja üksteist välistavad ning võrreldavad hinnangud on omavahel sõltumatud. Hinnanguteks võivad olla nii erinevate ekspertide klassifitseerivad otsused mistahes ainevallas kui ka erinevate klassifitseerimisalgoritmide abil saadud tulemused.

Kapa arvutatakse mõlemas võrreldavas klassifikatsioonis samadesse üksustesse klassifitseeritud (P_C) ja juhuslike otsuste korral oodatava (P_0) kokkulangeva klassifitseerimistulemuse osakaalu kaudu.

$$K = \frac{P_C - P_0}{1 - P_0} \qquad [2\text{-}31]$$

Kui m on klassifitseeritavate objektide üldarv, k on klasside arv, $n(i,j)$ on loend vigade maatriksi i-ndale reale ja j-le veerule vastavas lahtris, $n(i,+)$ on rea i summa ja $n(+,j)$ on veeru j summa, siis väljendub P_C ja P_0 järgmiselt:

$$P_C = \frac{\sum_{i-1}^{k} n(i,i)}{m} \qquad [2\text{-}32]$$

$$P_0 = \frac{\sum_{i-1}^{k} \left(n(+,i)n(i,+)\right)}{m^2} \qquad [2\text{-}33]$$

P_C on vigade maatriksi diagonaalil olevate vaatluste (pikslite) osa. P_0 on diagonaalil olevate pikslite osa ootus klassifikatsioonidevahelise seose puudumise (nullhüpoteesi kehtimise) ja klassi-sageduste püsivuse korral. P_C ei sobi klassifikatsioonide kooskõla hindamiseks, sest kui klasse on palju, on diagonaali ruute suhteliselt vähem ja seetõttu ka diagonaalile oodatavate vaatluste osa väiksem. Kaalutud kapa puhul korrutatakse klasside suhtelised sagedused klasside mõju tugevust määravate kaaludega.

Kapa on vastavuse osakaal, mida on korrigeeritud juhuslikkuse korral oodatavaga

Kapa eelis lihtsa kokkulangevuse ees on lineaarse korrelatsioonikordaja taoline universaalne muutumisvahemik miinus ühest pluss üheni. Kapa kordaja väärtus 0 näitab, et klassifitseeritud andmete vaheline kooskõla vastab juhusliku paiknemise korral oodatavale. Väärtus +1 näitab täielikku vastavust, mille puhul nullist suuremad loendid on vaid maatriksi diagonaalil. Negatiivne väärtus näitab süstemaatilist tendentsi objekte (rasterkujul andmekihtide puhul piksleid) valesti

klassifitseerida. Kapa minimaalne väärtus on −1, kuid see realiseerub vaid 2 × 2 sagedusmaatriksi puhul. Üldjuhul on seos klasside arvu *k* ja minimaalse võimaliku kapa väärtuse vahel

$$\kappa_{\min} = -\frac{1}{k-1}. \qquad [2\text{-}34]$$

Cohen (1960) juhib tähelepanu, et kapa maksimum on +1 vaid eeldusel, et sagedusmaatriksi äärejaotused on samad, mis tähendab, et klassifitseerijad on kategooriad käsitlenud samas mahus. Vastavuskordaja maksimaalne võimalik väärtus etteantud äärejaotuste korral võrdub

$$\kappa_{\max} = \frac{P_{0M} - P_C}{1 - P_C}, \qquad [2\text{-}35]$$

kus P_{0M} on minimaalsete väärtuste summa iga klassi osakaalust ühes ja teises klassifikatsioonis. $1 - \kappa_{max}$ näitab äärejaotuste erinevuse tõttu saavutamatut vastavust.

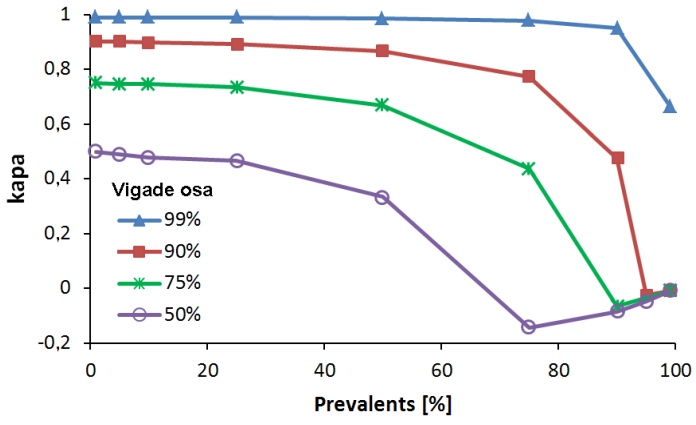
Kapa kordajate erinevuse olulisust saab võrrelda normaaljaotuse *Z* statistiku abil. Erinevus kapa kordajate vahel on statistiliselt oluline, kui |Z| on kriitilisest väärtusest suurem (95% testi puhul >1,96). *Z* statistik arvutatakse valemist

$$Z = \frac{K_1 - K_2}{\sqrt{\mathrm{var}(K_1) + \mathrm{var}(K_2)}}. \qquad [2\text{-}36]$$

Z statistiku määramiseks vajalike kapa dispersioonide [*var(K)*] arvutamiseks leidub kirjanduses keerukaid valemeid (Congalton ja Green 1999, lk. 50).

Kapa on kasutatav nii ruumiliste kui ka asukohaga mitte seotud andmete puhul. Klassikaline kapa ehk tavakapa ei arvesta mittekokkulangevate tulemuste paiknemise vea suurust ning võrdleb kokkulangevust samade ja sama sagedusega klasside juhuslikul paiknemisel oodatava kokkulangevusega. Kapa väärtus sõltub klasside sageduse vahekorrast ehk ühe variandi domineerimistasemest ehk prevalentsist (joonis 2-7). Nende piirangute ületamiseks esitas Pontius (2000) kapa neli varianti:

- tavakapa (*κ standard*) arvutatakse klassikalisel viisil,
- juhusliku asendi ja juhusliku sagedusega klasside puhul oodatava kokkulangevusega võrdlev kapa (*κ no*),
- asukoha kapa (*κ location*), mis arvestab asukoha nihkeid,
- sagedustekapa (*κ quantity*, *κ histogram*), mis mõõdab klassisageduste vastavust.



Joonis 2-7. Klassifikatsioonitäpsuse kordaja (kapa) sõltuvus positiivsete juhtude prevalentsist ja vigade osakaalust kaheväärtuselise tunnuse puhul juhul, kui vigades domineerivad väärpositiivsed. Väärnegtiivsete juhtude domineerimisel langeb kapa väärtus väikeste prevalentside puhul samamoodi, nagu selles näites suurte prevalentside puhul.

Tavakapa arvutamisel on P_0 seose puudumisel oodatavate väärtuste osa. Kusjuures oodatav väärtus osakaalude vastavustabeli igas lahtris võrdub sellele lahtrile vastav reasumma korda veerusumma. P_C on samamoodi klassifitseeritud vaatluste osakaal ning kapa arvutamise valem [2-31] on alapeatüki alguses.

Klassikaline kapa kordaja võrdleb klasside sagedusi ette antud sagedustega

Täielikku juhuslikkust eeldava kapa puhul arvestatakse klasside oodatav sagedus võrdseks, seega iga klassi oodatav osa võrdub üks jagatud klasside arv. Seega oodatav kokkulangevus igas klassis, kui klasside sagedused ei ole ette teada, võrdub klassi vaadeldud sagedus jagatud klasside arvuga.

Sageduskapa arvutamisel on P_C asemel klassisageduste ühisosa P_M ehk iga klassi osakaalu ühes või teises klassifikatsioonis olevate väiksemate väärtuste summa, ehk maksimaalselt võimalik kooskõla ühes ja teises klassifikatsioonis olevate klassisageduste korral. Sageduskapa ei saa olla negatiivne, sest sageduste ühisosa ei saa olla suurem kui seose puudumisel oodatav kokkulangevuse sagedus ($P_M \geq P_C$).

$$K_{Histogram} = \frac{(P_M - P)_0}{(1 - P_0)} \quad [2\text{-}37]$$

Asukohakapa väljendub järgmiselt (Pontius 2000, van Vliet *et al.* 2011).

$$K_{Location} = \frac{(P_C - P_0)}{(P_M - P_0)} \quad [2\text{-}38]$$

Hägune kapa kordaja (*fuzzy kappa)* arvestab nii eristatavate klasside eristamise ebakindlust ja klasside omavahelist sarnasust kui ka asukoha hägusust, hinnates mitte ainult samas kohas olevat üksust, vaid ka naabruses olevaid. Hägune kapa näitab keskmist sarnasust andmestikus juhuslikkuse korral oodatava keskmise sarnasuse suhtes (Hagen-Zanker *et al*. 2005, Hagen-Zanker 2006, 2009). Klasside sarnasuse arvestamiseks tuleb määratleda sarnasus igas klassikombinatsioonis. Asukoha hägususe arvestamiseks tuleb määratleda kaugusest sõltuvad kaalud. Iga koha jaoks arvutatakse interpreteerimisvektor, mis väljendab selles kohas oleva üksuse sarnasust samas kohas ja selle ümbruses olevate üksustega teisel kaardil, arvestades seejuures üksuste omavahelist sarnasust ja kauguskaale. Kui kõrge sarnasusega üksus asub koha lähedal, on kohad sarnasemad kui sama üksuse kaugema asendi korral.

Iga koha sarnasusearvutamiseks leitakse kummaski andmekihis suurima sarnasusega kategooria ja siis nende üksuste sarnasustest väiksem väärtus. Selliselt arvutatud kohtade sarnasust saab kujutada sarnasuskaardil. Üldkeskmine sarnasus on üksikkohtade sarnasuste aritmeetiline keskmine. Saamaks vastavusindeksit, mille oodatav väärtus juhuslikkuse korral võrdub nulliga, tuleb arvutada kahe võrreldava kihi kokkulangevuse ootus kategooriate juhusliku paiknemise korral ühes võrreldavas kihis. Seda saab teha kategooriate sagedusest lähtudes (valemid Hagen-Zanker *et al*. 2005) või ühe andmekihi asendit korduvalt juhuslikult nihutades.

Tavakapa ei sobi ka modelleeritud maastikumuutuste tõesuse kvantifitseerimiseks, sest maastiku-simulaatorid jätavad enamikus kohtades maakasutusüksuse samaks, mis aga ei pruugi alati olla õige tulemus. Tulemuse õigsuse hindamiseks tuleks võrrelda kokkulangevust mitte maakattekategooriate juhusliku paiknemise, vaid juhusliku muutumise tulemusega (van Vliet *et al.* 2011).

Kategooriliste pindade vastavust saab analüüsida ka mitmesuguste sarnasuskordajate abil, mis võrdlevad ühel või teisel moel kokkulangevate ja mitte kokkulangevate vaatluste osa.

2.3.7.3. Hanssen-Kuiperi skoor

Hanssen-Kuiperi skoor (*Hanssen-Kuiper Skill Score – KSS* ehk *True Skill Statistic – TSS*) on kaheväärtuselise muutuja hinnangu täpsuse mõõtmisvahend, mis leitakse vastavalt valemile

$$TSS = TP + TN - 1, \qquad [2\text{-}39]$$

kus *TP* on tõeste positiivsete hinnangute osa ja *TN* tõeste negatiivsete hinnangute osa. Osakaalud leitakse vastavalt positiivsete ja negatiivsete hinnangute arvust eraldi, mitte otsuste koguarvust.

TSS muutub vahemikus +1…–1. Erinevalt kapa kordajast on absoluutse mittevastavuse korral TSS alati –1 ning TSS ei sõltu klasside sageduse vahekorrast (McPherson *et al.* 2004, Allouche *et al.* 2006).

2.3.7.4. Šansside suhe

Kaheväärtuseliste jaotuste vastavust saab mõõta lisaks eelmainitutele ka šansside suhtega, mis võrdleb ühte klassi ja teise klassi kuulumise tõenäosuse suhteid – klassifikatsioonitäpsuse puhul õigete tulemuste ja väärate tulemuste sageduse suhet (vt ka ptk 1.4.3.3).

Kui vaatluse ühte gruppi kuulumise tõenäosus on p_1 ja teise gruppi kuulumise tõenäosus on p_2, siis šansside suhe on

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}. \qquad [2\text{-}40]$$

Kui kaks varianti on teineteist välistavad, siis

$$OR = \frac{p_1/(1-p_1)}{(1-p_1)/p_1}. \qquad [2\text{-}41]$$

Kui kaheväärtuselise tunnuse variantide sagedus kahes jaotuses on esitatud 2 × 2 tabelis, kus p_{11} ja p_{22} tähistavad vastavalt ühe väärtuse ja teise väärtuse kokkulangevust kahes jaotuses (õiged positiivsed ja õiged negatiivsed juhud) ning p_{12} ja p_{21} tähistab väärtuste erinevuse sagedust (väärpositiivsed ja väärnegatiivsed juhud), siis

$$OR = \frac{p_{11}p_{22}}{p_{12}p_{21}}. \qquad [2\text{-}42]$$

Uurimused

Erinevaid liikide esinemise tõenäosuse piirväärtuse mõõdikuid võrdlesid Liu *et al*. (2005). Autorid leidsid, et kõige sagedamini kasutatud mõõdikud – kapa kordaja ja 50% juurde fikseeritud piirväärtus on tundlikud prevalentsi nihkumise suhtes 50% tasemest eemale.

2.4. Ordineerimine

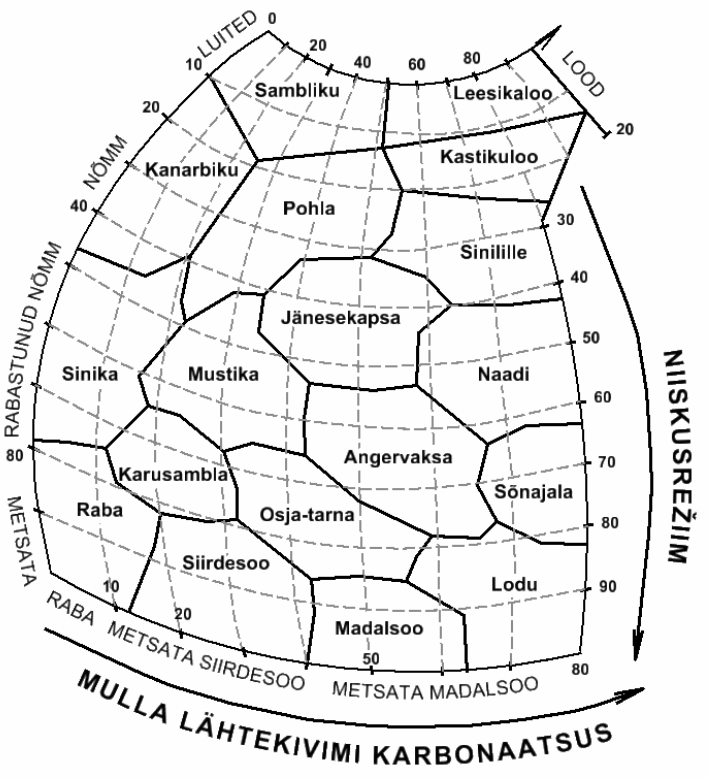
Ordineerimine ehk korrastamine on vaatluste järjestamine nende sarnasuse alusel. Mõiste võttis taimkatteanalüüside järjestamise tähenduses kasutusele Goodall (1954). Ordineerimismeetoditega arvutatakse mingite nähtuste paiknemist mingis teljestikus, eelkõige tunnusruumis. Mitmemõõtmeline ordinatsioon püüab analoogiliselt faktoranalüüsiga kirjeldada vaatluste või tunnuste paiknemist tunnusruumis võimalikult väheste telgede abil. Erinevalt faktor- ja peakomponentanalüüsist ei nõua mitmemõõtmeline ordinatsioon tunnuste normaaljaotust. Ordineerida saab igasuguseid tunnuseid ja vaatlusi, mille puhul on võimalik arvutada nende omavahelist sarnasust. Ordinatsiooni lähteandmed tuleb esitada sarnasuste või siis erinevuste (erinevus = kaugus tunnusruumis) maatriksina. Sarnasuste maatriksina saab kasutada nii parameetrilise kui ka mitteparameetrilise korrelatsiooni maatriksit.

Vaatlusi saab ordineerida nii sarnasuse kui ka paiknemise alusel

Ordineerimismeetodid kuuluvad **gradientanalüüsi** hulka, mis uurib nähtuste paiknemist keskkonnatingimuste suhtes. Kaudse gradientanalüüsi puhul ei ole gradiente (keskkonnafaktoreid) otseselt mõõdetud. Vaatlused (näiteks taimkatteanalüüsid) järjestatakse vaid nende omavahelise sarnasuse järgi. Otsese gradientanalüüsi puhul on keskkonnagradiendid otseselt mõõdetud ja uuritakse liigi esinemissageduse või ohtruse seost keskkonnafaktoriga.

Gradient võib olla kompleksne, näiteks kõrgus merepinnast on otseselt mõõdetav, aga mõjutab organisme temperatuuri, sademete, pinnase jt faktorite kaudu. Gradient võib olla ka uurijapoolne üldistus, näiteks inimmõju intensiivsus või koha sobivus. Gradientanalüüsi võib vaadelda ka ühe tunnusruumi kirjeldamise viisina. Väga paljudes uurimustes kujutatakse liikide, vaatlusalade või elupaigatüüpide paiknemist mingite gradientide teljestikus. Näiteks E. Lõhmuse metsakasvukohatüüpide klassifikatsioon (joonis 2-8). Ordinatsioonide puhul tuleb silmas pidada, et ordinatsioonitelje ei pea tingimata olema omavahel risti (see tähendab üksteisest sõltumatud), nagu on ristkoordinaadid eukleidilises ruumis (Cormack 1979).

Ülevaate ordineerimismeetoditest võib leida Kenti ja Cokeri (Kent ja Coker 1992) raamatust ning veebilehelt http://ordination.okstate.edu.



Joonis 2-8. Eesti metsa-kasvukohatüüpide ordinatsioon (Lõhmus 1984, lihtsustatult). Telgede väärtused on protsentides maksimaalse võimaliku suhtes.

2.4.1. Faktoranalüüs ja peakomponentanalüüs

Faktoranalüüsi ja peakomponentanalüüsi eesmärk on eemaldada niinimetatud paljumõõtmelisuse needus asendades suure hulga omavahel korreleerunud argumenttunnuseid väiksema arvu abstraktsete ja omavahel sõltumatute tunnustega (faktoritega) nii, et faktorid kirjeldaksid võimalikult suure hulga lähtetunnuste hajuvusest. **Paljumõõtmelisuse needus** (*curse of dimensionality*) tähistab mitmesuguseid probleeme, mis kaasnevad tunnuste rohkusega ehk tunnusruumi paljumõõtmelisusega. Põhiprobleemiks on, et tunnusruumi mõõtmete arvu suurenedes täidab olemasolev andmestik tunnusruumi järjest hõredamalt. Tunnuste väärtuste kombinatsioonide esindatuse tagamiseks on tunnuste arvu suurenedes järjest rohkem vaatlusi tarvis.

Faktoranalüüs ja peakomponentanalüüs annavad enamasti lähedasi tulemusi. Erinevused on arvutuslikud ja püstitatud eesmärgis: peakomponentanalüüs püüab kirjeldada varieeruvust vähema hulga ortogonaalsete telgedega ja on eelkõige ordineerimismeetod; faktoranalüüsi peamine eesmärk on leida ja kirjeldada otseselt mittemõõdetavaid üldistatud faktoreid.

Faktoranalüüsi ja peakomponentanalüüsi eeldused on:

- kõik tunnused peavad olema arvulised,
- kõik tunnused peavad olema kõigis katsetes mõõdetud,
- faktoranalüüsid eeldavad tunnuste normaaljaotust juhul, kui lähteandmetena kasutatakse Pearsoni korrelatsioonikordajate maatriksit. Juhul, kui faktorite või peakomponentide leidmisel kasutatakse vähimruutude meetodit, peaks keskmine ruutviga olema tunnuste hajuvuse adekvaatne kirjeldaja.

Faktoranalüüsi (*factor analysis*) puhul eeldatakse, et tegelikult mõjub väike hulk faktoreid, mida on mõõdetud mitme üksiknäitaja abil. Eesmärgiks on need üldistatud faktorid leida ja seostada mõõdetud tunnustega. Teatud osa iga tunnuse varieeruvusest jääb faktoranalüüsil kirjeldamata, seda

nimetatakse **tunnuse omapäraks.** Tunnuse seda osa, mille faktorid ära kirjeldavad, nimetatakse **kommunaliteediks** *(communality*). Faktorkaalud ehk **faktorlaadungid** (*factor loadings*) on faktorite ja lähtetunnuste vahelised korrelatsioonikordajad. Need näitavad, millisel määral üks või teine faktor mingi algtunnusega seondub. Faktoranalüüsi tulemus on faktormaatriks. Mõnikord kasutatakse faktormaatriksi pööramist, et saada paremini tõlgendatavad faktorid.

Faktoranalüüs üldistab mõõdetud tunnused üldistatud faktoriteks

Faktoranalüüsile lähedane on **peakomponentanalüüs** (*principal component analysis – PCA*), mis on andmete ortogonaalne projektsioon vähemõõtmelisse tunnusruumi, säilitades maksimaalse võimaliku dispersiooni. Peakomponentanalüüsil moodustatakse algsesse tunnusruumi uued tunnusteljed nii, et esimene telg on punktiparve maksimaalse hajuvuse sihis, teine esimesega risti ja punktiparve maksimaalse jääkhajuvuse sihis jne. Uued teljed moodustatakse olemasolevate kaalutud lineaarkombinatsioonidena ja neid nimetatakse **peakomponentideks**. Peakomponentide sihid on määratud vastavate **omavektorite** (*eigenvectors*) abil – omavektori elemendid kannavad siin laadungi (*component loading*) nime. Peakomponentide **omaväärtused** (*eigenvalues*) on peakomponentide dispersioonid ehk hajuvuse osa, mida üks või teine peakomponent kirjeldab. Originaalvaatlused projitseeritakse peakomponentidele ja saadakse vaatluste koordinaadid ehk skoorid peakomponentide suhtes (*component scores*).

Peakomponentanalüüs pöörab ja nihutab tunnusruumi vaatlusparve nii, et telgedele moodustatud koondtunnused omavahel enam ei korreleeru

Peakomponentanalüüsi kasutamise korral tekib sageli raskusi esimesele järgnevate telgede interpreteerimisega. Peakomponentanalüüs moodustab ortogonaalsed teljed, mis eeldab koosmõjude puudumist peakomponentide vahel olles tegelikkuse lihtsustus, sest looduses on keskkonnategurid omavahel seotud.

Peakomponentanalüüsi eel on soovitav andmed standardiseerida, nii et kõigi tunnuste keskmine oleks null ja standardhälve üks. Kui standardiseerimist ei kasutata, hakkavad tulemused sõltuma tunnuste mõõtühikutest ja suuremate väärtustega tunnused omandavad suurema osakaalu. Topeltstandardiseerimise puhul standardiseeritakse nii seletavad kui ka funktsioontunnused, näiteks kasutatakse liigi absoluutarvudes ohtruse asemel suhtelist ohtrust. Peakomponentanalüüsi erivorm on SVD (*singular value decomposition*), mille puhul ei lahutata vaatlustest keskväärtust.

Peakomponentanalüüsil on kaks varianti: normaalne ehk R-analüüs ja pöörd- ehk Q-analüüs. Esimene on vaatluste analüüs tunnusruumis, teine tunnuste analüüs vaatluste järgi konstrueeritud telgede suhtes. Termineid Q-analüüs ja R-analüüs kasutatakse ka väljaspool peakomponentanalüüsi konteksti. Meteoroloogias kasutatakse aga termineid T- ja S-peakomponentanalüüs, sõltuvalt sellest, kuidas andmemaatriks sisestatakse. T-analüüsi tunnusteks on mingi ajaühiku väärtused paljudes vaatluskohtades, S-analüüsil on aga tunnusteks mõõtejaamad ja juhtudeks ajaühikud, näiteks erinevad aastad.

R tüüpi analüüsis võrreldakse tunnuseid, Q tüüpi analüüsis objekte

Peakomponentanalüüsi põhimõtteid kirjeldas Karl Pearson (1901). Meetod sai laiemalt tuntuks alles arvutite kasutuselevõtu järel. Geobotaanikas sai meetod populaarseks alates 1966ndast aastast.

2.4.2.1. Empiirilised ristfunktsioonid

Empiirilised ristfunktsioonid (*empirical orthogonal functions – EOF*) on pideva muutuja väärtuspinna üldistatud kirjeldamise vahend. Empiiriliste ristfunktsioonide analüüsil otsitakse ortogonaalseid baasfunktsioone ehk peakomponente, mis kõige enam kirjeldavad muutuja hajuvust.

Empiirilised ristfunktsioonid on meteoroloogias kasutusel 1940ndatest aastatest. Meteoroloogiliste väljade empiirilisteks ortogonaalseteks funktsioonideks lahutamise terminoloogia pärineb E.N. Lorenzilt (1956). Lorenzi käsitluses saab pidevat muutujat kirjeldada summaga ortogonaalsetest kohatunnustest (*orthogonal functions of space*), mille kordajad on aja funktsioonid. Mõned autorid loevad empiirilisi ristfunktsioone peakomponentide sünonüümiks (Björnsson ja Venegas 1997). Teiste allikate järgi on empiiriliste ristfunktsioonide analüüs peakomponentanalüüsi laiendus, mis võimaldab kasutada mittelineaarseid baasfunktsioone ja ajast või kohast sõltuvaid peakomponente. Atmosfäärifüüsikas tähistab empiiriliste ristfunktsioonide analüüs aeg-ruumilise muutuja peakoordinaatanalüüsi, milles leitud peakomponente saab projitseerida nii ajaskaalale kui ka geograafilisele kaardile (Hannachi *et al*. 2007, Straus ja Krishnamurthy 2007).

2.4.2. Mitmemõõtmeline skaleerimine

Vaatlusvektorite projitseerimist peakomponentide poolt määratud tasandile ja vaatluste sellel paiknemise analüüsi nimetatakse **peakoordinaatanalüüsiks** (*principal coordinates analysis – PCO või PCoA*). Peakoordinaate saab kasutada näiteks regressioonanalüüsil, seostamaks vaatlusi tunnustega, mida peakomponentanalüüsil ei kasutatud. Sellist analüüsi nimetatakse ka **peakoordinaatregressiooniks**.

Peakoordinaatanalüüs ehk meetriline mitmemõõtmeline skaleerimine (*metric multidimensional scaling*) on üks **mitmemõõtmelise skaleerimise** (*multidimensional scaling*) meetod. Mitmemõõtmeline skaleerimine üldiselt on sarnasus- või korrelatsioonimaatriksis oleva teabe visualiseerimise meetodite kogum. Mitmemõõtmelise skaleerimise lähteandmeteks on vaatlustevaheliste sarnasuste tabel või korrelatsioonimaatriks. Mitmemõõtmelise skaleerimise abil saab vaatluskohtade - vaheliste vahemaade maatriksist taasluua kohtade paiknemise kaarti, mis säilitab algses maatriksis olevad vaatlustevahelised vahemaad nii palju kui võimalik. Ilmakaarte paiknemine kohtade paiknemismustri suhtes tuleb hiljem endal otsustada.

Tarkvara

Tarkvarapaketi Statistica mooduli *Multidimensional scaling* lähteandmeteks vajaliku korrelatsioonimaatriksi saab luua, kui salvestada parameetrilise korrelatsioonanalüüsi tulemus. Mitteparameetriliste korrelatsioonide arvutamisel saadud tabel *(scrollsheet*) tuleb esmalt salvestada andmeformaati (*file → save as data*), salvestatud tabel tuleb avada ja lisada selle lõppu neli uut rida. Esimese lisatud rea nimeks tuleb kirjutada *Means*, teise lisatud rea nimeks *Std.Dev.*, kolmanda lisatud rea nimeks *No.cases* ja neljanda lisatud rea nimeks *Matrix*. Rea *No.cases* esimesse lahtrisse tuleb sisestada tabeli arvutamisel kasutatud vaatluste arv, ja rea *Matrix* esimesse lahtrisse maatriksi tüübi kood (korrelatsioonid – 1, sarnasused – 2, erinevused või kaugused – 3, kovariatsioonid – 4). Keskmiste ja standardhälvete rea võib jätta arvudest tühjaks. Tunnusruumi dimensioonide arv tuleb endal otsustada. Telgede leidmine toimub iteratiivsete meetoditega ja on võrdlemisi arvutusmahukas.

2.4.3. Sagedustabelite log-lineaarne analüüs

Igasuguseid varieeruvaid mitmetunnuselisi andmestikke saab esitada sagedustabelina, kus on vaatluste sagedused kahe või enama tunnuse klassides (igal tunnusel kaks või enam klassi). Sagedustabelite analüüsi abil püütakse leida ja tõestada statistilisi seoseid võrreldes servasummade järgi arvutatud oodatavaid sagedusi tabelis tegelikult esinevatega. Olulised kõrvalekalded oodatavate ja tegelike sageduste vahel viitavad vaatluste mittejuhuslikule jaotumisele klassikombinatsioonides ja seega seose olemasolule klassifitseerivate tunnuste vahel. Kui tunnuste vahel leitakse seos, tuleb loobuda nullhüpoteesist ja otsida seost kirjeldavat mudelit.

Log-lineaarsed meetodid analüüsivad sagedustabeleid logaritmilise seosefunktsiooniga üldistatud lineaarse mudeli abil. Mudeli sobitamise protsess on iteratiivne ning see põhineb suurimal tõepäral (ptk 1.2). Log-link tagab, et kõik prognoositud väärtused on positiivsed. Juhul, kui ühte sagedustabelis olevat muutujat käsitletakse funktsioon- ja teist argumenttunnusena ning funktsioontunnus on kaheväärtuseline, siis on see logistiline regressioon logit lingiga. Lisaks faktorite individuaalsetele mõjudele võivad esineda ka faktorite koosmõjud, mis ei pruugi olla liituvad ehk aditiivsed, vaid võivad kombineeruda muul viisil ehk olla interaktiivsed. Oodatavate sageduste arvutamiseks mitmemõõtmelistest sagedustabelitest kasutatakse iteratiivset sobitamist.

Sagedustabelite log-lineaarse analüüsi *(log-linear models for contingency tables)* probleemiseade on sarnane dispersioonanalüüsile (ptk 1.5.4). Erinevus on lähteandmetes, milleks log-lineaarse analüüsi puhul on sagedustabelid. Termin log-lineaarne tuleneb sagedustele iseloomulike kordsete muutuste normaliseerimisest logaritmteisendusel. Tänu logaritmilisele teisendusele saab mitmemõõtmeliste sagedustabelite analüüsi ülesannet käsitleda analoogilisena dispersioonanalüüsi ülesandele. Nii sagedustabeli kui ka dispersioonanalüüsi puhul eeldatakse faktorite klassifitseeritud mõjusid, mis koos koosmõjudega tingivad sageduste väärtused sagedustabeli lahtrites.

Log-lineaarse mudeli olulisust saab kontrollida hii-ruut testiga või Monte Carlo meetodil (ptk 3.6.6). Mudeli sobivust saab analüüsida ja graafiliselt esitada jääkide analüüsi abil. Jääksagedused on empiiriliste ja oodatavate sageduste vahed või suhted. Kui mudel on andmetele hästi sobitatud, peaks jäägid paiknema ühtlaselt ja juhuslikult mõlemal pool mudelit.

Tarkvara

Statisticas saab kasutada mudeli automaatset sobitamist, mis toimub järgmise algoritmi abil. Esmalt sobitatakse mudel eeldades, et tunnused on sõltumatud. Kui mudel ei sobi (hii-ruut statistik näitab olulist mudeli ja tegelike andmete erinevust), proovitakse mudelit sobitada kahekaupa koosmõjudega. Kui mudel ei sobi ka nüüd, kasutatakse kolmepoolseid koosmõjusid jne. Kui programm leiab mingil koosmõjude tasemel mudeli rahuldava sobivuse, siis hakkab see mudelist eemaldama koosmõjusid, mille eemaldamisel ei teki olulist erinevust mudeli ja empiiriliste andmete vahel.

2.4.4. Vastavusanalüüs

Vastavusanalüüs (*correspondence analysis – CA*) ehk vastastikune keskmistamine ehk ristkeskmistamine (*reciprocal averaging – RA*) on nominaalsete tunnuste vastavust kirjeldav meetod, mis hindab vastavust sagedustabeli ridade ja veergude vahel. Meetodiga püütakse leida loendite varieeruvust kirjeldavaid peamisi faktoreid analoogselt faktor- ja peakomponentanalüüsile. Vastavusanalüüsi olulisim erinevus on, et varieeruvust kirjeldavaid telgi arvutatakse korraga nii tunnustele kui ka vaatlustele – sagedustabeli ridu ja veerge käsitletakse võrdselt. Saadakse nii tunnuste kui ka

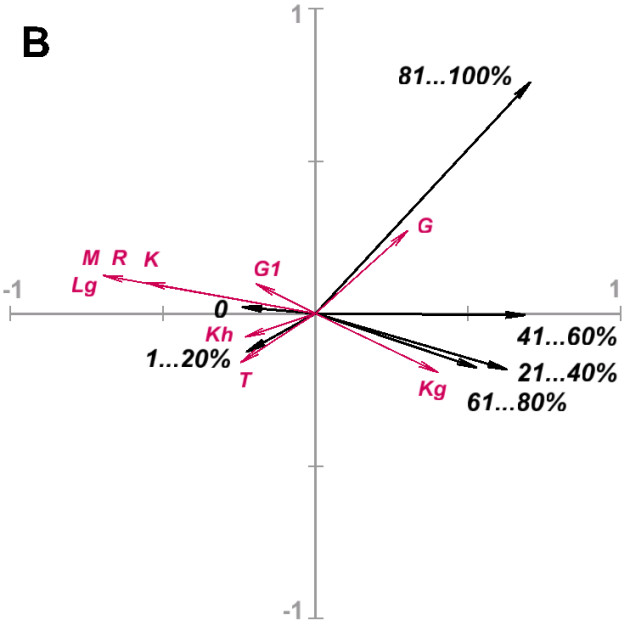
vaatluste ühine ordinatsioon, milles vaatluste ja tunnuste leviku raskuskese on samas kohas ning iga tunnus hälbib keskmest seda tunnust kandvate vaatlustega samas suunas (joonis 2-9). Erinevalt mitmemõõtmelisest skaleerimisest on vastavusanalüüsi lähteandmete allikaks vaatluste sagedustabel, mitte korrelatsioonimaatriks. Erinevalt peakomponentanalüüsist on seostatavad tunnused nominaalsed. Vastavusanalüüsi mitmemõõtmelist varianti nimetatakse kanooniliseks ehk mitmemõõtmeliseks vastavusanalüüsiks (ptk 2.4.6).

Vastavusanalüüs ei tõesta hüpoteese, vaid püüab leida andmetele kõige paremini sobivat mudelit. Ökoloogias on vastavusanalüüsi nimetatud ka kaudseks gradientanalüüsiks ja kaudseks ordinatsiooniks, kuna see otseselt ei mõõda seost liikide ja keskkonnafaktorite vahel, vaid paigutab telgi tunnusruumi liikide koosesinemise sageduse alusel.

A

Muld	Põõsasmarana katvus [%] 0	1...20	21...40	41...60	61...80	81...100	Kokku
T	2	1	0	0	0	0	3
Kh	260	67	16	6	5	3	357
K	92	5	5	0	2	1	100
Kg	98	50	32	13	14	8	215
Lg	2	0	0	0	0	0	2
G	108	39	19	10	7	23	206
G1	17	3	1	0	1	1	23
M	3	0	0	0	0	0	3
R	3	0	0	0	0	0	3
Kokku	585	165	68	29	29	36	912

B



Joonis 2-9. Põõsasmarana katvusklasside ja mullatüübi kombinatsioonide sagedused Keila ja Suurupi piirkonnas (A) ning mullaliikide ja põõsasmarana katvusklasside paiknemine vastavusanalüüsi esimesel ja teisel teljel (B). Horisontaaltelje omaväärtus = 0,345; vertikaalteljel 0,172. Üle 80% katvusega kohad esinevad valdavalt gleimuldadel, keskmised katvused (21...80%) esinevad eelkõige gleistunud karbonaatsel mullal, liigi puudumine või vähene katvus esinevad mitmesugustel muldadel ja seostuvad mullatüübiga nõrgalt. *T* – tehnogeenne ala, *Kh* – paepealne muld, *K* – rähkmuld, *Kg* – gleistunud karbonaatne muld, *Lg* – gleistunud leetunud muld, *G* – gleimuld, *G1* – turvastunud muld, *M* – madalsoomuld, *R* – rabamuld. Vaatlusandmed Kalle Remm 2008–2011.

Vastavusanalüüs arendati välja 1930ndatel aastatel ja esmalt kasutati seda sotsiaalteadustes (Hirschfeld 1935). Ülevaate vastavusanalüüsi arengust ja arvutuslikest üksikasjadest võib leida teostest Beh (2004) ja Greenacre (2007). Ökoloogias hakati vastavusanalüüsi kasutama 1970ndatel aastatel koos kaalutud keskmise meetodiga, sellest ka nimevariant vastastikune keskmistamine. **Kaalutud keskmistamise** meetod lähtub eeldusest, et kui liigi esinemise tõenäosus on keskkonnafaktori suhtes sümmeetriline, on liigi esinemissagedustega kaalutud esinemistingimuste keskmine liigi optimumi nihketa hinnanguks. Kaalutud keskmistamist saab kasutada eelkõige kahel eesmärgil:

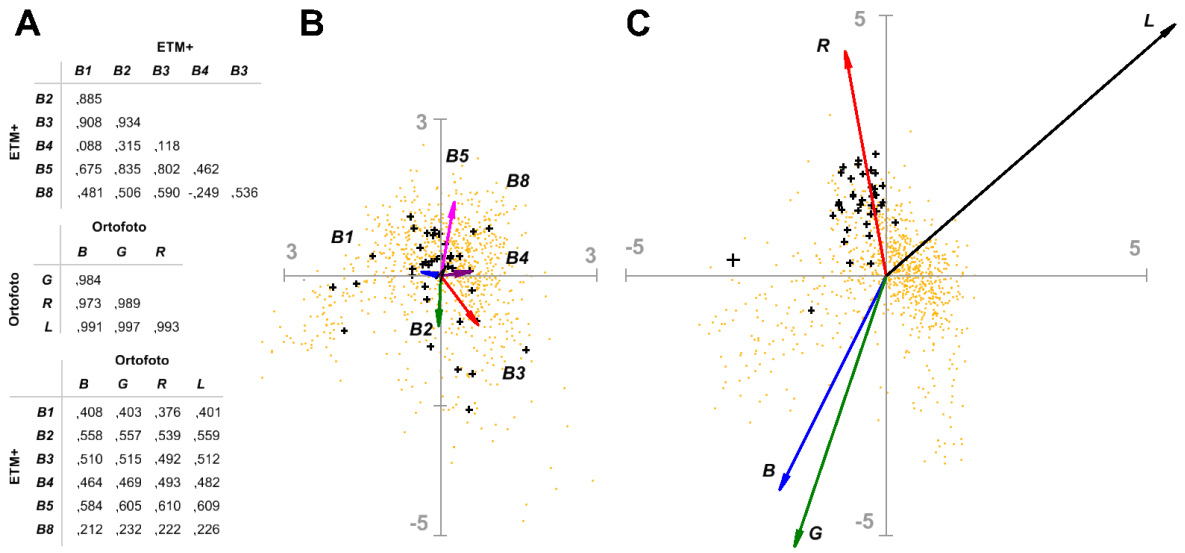
- liikide optimumide leidmiseks keskkonnafaktorite suhtes,
- keskkonnafaktorite kaudseks hindamiseks liikide andmete alusel (bioindikatsioon, kalibreerimine).

Vastavusanalüüsi puhul on ilmnenud kaks olulist probleemi. Esiteks, vastavusanalüüsil leitud teine telg on sageli esimese ruutfunktsioon, seega korrelatsiooni mõttes esimesega tunnusruumis risti ja näiliselt sõltumatu, kuid tegelikul esimese funktsioon (küüru-efekt – *arch effect*). Teiseks, kuna telgede ühikuks on liigilise koosseisu muutumiskiirus, siis surutakse telgede otsad analüüsil kokku. 1980ndatel aastatel taimkatte ordineerimisel entusiastlikult kasutatud programmis DECORANA (Hill ja Gauch 1980) on telgedega seotud probleeme püütud ületada jagades esimesele teljele järgnevad teljed lõikudeks ja kontrollitakse iga lõigu ortogonaalsust eraldi. Kasutatavate lõikude pikkus on paraku tulemusi mõjutav subjektiivne faktor. DECORANAs kasutatud algoritmi nimetatakse trendivabaks vastavusanalüüsiks (*detrended correspondence analysis – DCA*).

Ter Braak (1988a,b) soovitas küüruefekti kõrvaldamiseks faktorlaadungite asemel kasutada polünomiaalregressiooni jääke olemasolevate telgede suhtes.

2.4.5. Kanooniline korrelatsioonanalüüs

Kanooniline korrelatsioonanalüüs, mida sageli nimetatakse lihtsalt kanooniliseks analüüsiks, on kahe tunnusterühma vahelise seose uurimismeetod. Kanooniline korrelatsioonanalüüs on faktoranalüüsi analoog kahe tunnusterühma seostamiseks, mis sobib näiteks seose uurimiseks riskifaktorite ja sümptoome kirjeldavate tunnuste vahel või siis taimekoosluse ja kasvukoha omaduste seostamiseks. Kanoonilises analüüsis võrreldakse kanoonilisi faktoreid, mitte originaaltunnuseid. Kanoonilise korrelatsioonanalüüsiga saab leida vaatluste paiknemist kanooniliste faktorite kui tunnusruumi telgede suhtes ja tunnuste panuseid kanoonilistesse faktoritesse (joonis 2-10). Kanoonilist korrelatsiooni saab väljendada ka korrelatsioonidena: 1) tunnuste vahel ühes tunnuste rühmas, 2) teises tunnuste rühmas ja 3) eri rühma kuuluvate tunnuste vahel.



Joonis 2-10. Kanooniline korrelatsioonanalüüs põõsasmarana vaatluskohtades mõõdetud peegeldunud kiirguse intensiivsuste vahel Landsat ETM+ kujutise kanalites *B1*, *B2*, *B3*, *B4*, *B5* ja *B8* 29. juulist 2001 ning 2009.a. ortofoto kanalites *R*, *G*, *B*, *L*. Kanooniline korrelatatsioon tunnuserühmade vahel = 0,696. A – tunnustevahelise korrelatsiooni maatriksid. B – ETM+ tunnuste kanoonilised kaalud, vaatluskohad (ruuged täpid) ja kohad põõsasmarana katvusega >75% (mustad ristid) kahe esimese kanooniliste faktori teljestikus. C – ortofoto tunnuste kanoonilised kaalud, vaatluskohad ja kohad põõsasmarana katvusega >75% kahe esimese kanooniliste faktori teljestikus. Põõsasmarana ohtra esinemise kohad seostuvad eelkõige ortofoto punase värvitooniga. Enamik siinkasutatud tunnustest on omavahel tugevasti korreleerunud, eriti ortofoto värvitoonid, mis tähendab, et ortofotodel on heleduse-tumeduse varieeruvus palju suurem kui värvitooni varieeruvus.

Kanoonilise analüüsi puhul on soovitatav: 1) et vaatluste arv ületaks faktorite arvu vähemalt 20 korda, 2) et korrelatsioonimaatriks ei sisaldaks seosekordaja väärtusi üks või selle lähedasi väärtusi. Kanooniliste seoste statistilise olulisuse arvutus eeldab muutujate normaaljaotust. Kanooniline korrelatsioon modelleerib vaid lineaarseid seoseid.

Kanoonilist faktorit (*canonical root*) võib ette kujutada kui otseselt mitte mõõdetavat tunnust. Kanoonilised faktorid moodustatakse nii, et nad oleksid üksteisest sõltumatud ning neid võib ette kujutada tunnusruumis omavahel risti paiknevate telgedena.

Kanoonilised kaalud näitavad mõõdetud tunnuste mõju suunda ja osakaalu kanooniliste faktorite moodustamisel. Olulise mõjuga kanooniliste muutujate arv määratakse hii-ruut testiga. Esmalt määratakse kõigi kanooniliste muutujate mõju, seejärel kanooniliste muutujate mõju esimese kanoonilise muutuja eemaldamisel jne.

Kanooniliste muutujate omaväärtused näitavad, kui suur osa dispersioonist on seotud ühe või teise kanoonilise faktoriga või on seletatav ühe või teise kanoonilise korrelatsiooniga. Ruutjuurt omaväärtusest nimetatakse **kanooniliseks korrelatsioonikordajaks**.

Liiasusanalüüsi (*redundancy analysis – RDA*) on mitme funktsioontunnusega ja mitme argumenttunnusega regressioonanalüüs, millele lisandub prognoositud väärtuste peakomponentanalüüs.

Tarkvara

Kanoonilist korrelatsiooni saab arvutada Statistica moodulitega *Canonical Analysis*. Spetsiaalselt kanoonilise korrelatsiooni jaoks on loodud programm CANOCO (http://www.pri.wur.nl/uk/products/canoco), mis arvutab lisaks koosluste sarnasuse ja kasvukohtade sarnasuse hinnangule ka koosluste paiknemise hinnangu kasvukohtade sarnasuse teljestikus.

Programmis Statistica nimetatakse kanoonilise analüüsi tunnuste rühmi listideks. Faktorite jaotuste normaalsuse visuaalseks kontrolliks on mooduli *Canonical Analysis* allmoodulis *Define model* vahend *Means and standard deviations*, mille juurde käib muutujate jaotuste histogrammide koostamine.

2.4.6. Kanooniline vastavusanalüüs

Kanoonilise vastavusanalüüsiga (*canonical correspondence analysis – CCA*) otsitakse parimat vastavust suurt hulka keskkonnatunnuseid üldistatud kujul kirjeldavate faktorite ja liikide nõudluste vahel järgmistel eeldustel:

- liikide tolerantsipiirid on sama laiad,
- iga liigi maksimaalne hulk on sama,
- liikide optimumid paiknevad gradiendil ühtlaselt või juhuslikult,
- liikide nõudlused on keskkonnafaktorite suhtes normaaljaotusega.

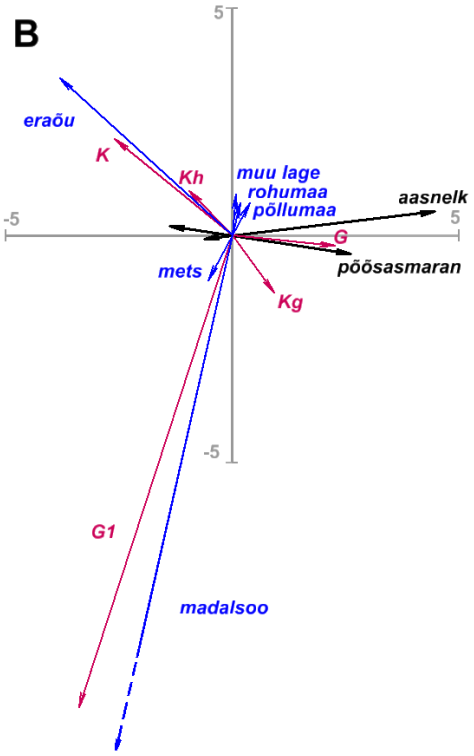
Kanooniline vastavusanalüüs on vähetundlik mõningase kõrvalekaldumise suhtes nendest eeldustest (ter Braak *et al.* 1993, ter Braak ja Juggins 1993). Meetod sobib andmestikele, kus domineerivad liigi puudumise juhtumid. Erinevalt kanoonilisest korrelatsioonist ei eelda kanooniline vastavusanalüüs lineaarset seost tunnuste vahel. Meetod sobib juhul, kui puuduvad andmed liikide ökoloogiliste nõudluste kohta ja kasutada on vaid vaatluskohtades leitud liikide nimestikud. Kanoonilist vastavusanalüüsi on palju kasutatud taimekoosluste ja kasvukohatingimuste omavaheliste sarnasuste kujutamiseks samas teljestikus. Näiteks R. Ferris *et al*. (2000) uuris kasvukoha tunnuste, puistu ja alustaimestiku vahelisi seoseid.

Joonis 2-11 kujutab samasse teljestikku ordineeritud seoseid aasnelgi esinemise ja puudumise, põõsasmarana esinemise ja puudumise, põhikaardi põhialade ja mullakaardile märgitud mullaliigi vahel Keila ja Suurupi vahelises piirkonnas paiknevates vaatluskohtades.

A

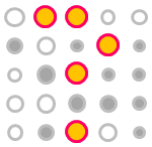
	aasnelk puudub	aasnelk on	põõsasmaran puudub	põõsasmaran on	madalsoo	mets	eraõu	muu lage	rohumaa	põllumaa	Muld: Kh	Muld: K	Muld: Kg	Muld: G	Muld: G1
aasnelk puudub	709	0	499	210	11	178	8	89	145	278	269	100	176	145	19
aasnelk on	0	98	30	68	0	6	1	16	29	46	27	2	28	40	1
põõsasmaran puudub	499	30	529	0	9	123	8	72	124	193	224	91	108	89	17
põõsasmaran on	210	68	0	278	2	61	1	33	50	131	72	11	96	96	3
madalsoo	11	0	9	2	11	0	0	0	0	0	1	0	3	3	4
mets	178	6	123	61	0	184	0	0	0	0	58	12	69	38	7
eraõu	8	1	8	1	0	0	9	0	0	0	4	3	2	0	0
muu lage	89	16	72	33	0	0	0	105	0	0	41	14	24	25	1
rohumaa	145	29	124	50	0	0	0	0	174	0	65	22	34	50	3
põllumaa	278	46	193	131	0	0	0	0	0	324	127	51	72	69	5
Muld: Kh	269	27	224	72	1	58	4	41	65	127	296	0	0	0	0
Muld: K	100	2	91	11	0	12	3	14	22	51	0	102	0	0	0
Muld: Kg	176	28	108	96	3	69	2	24	34	72	0	0	204	0	0
Muld: G	145	40	89	96	3	38	0	25	50	69	0	0	0	185	0
Muld: G1	19	1	17	3	4	7	0	1	3	5	0	0	0	0	20

B



Joonis 2-11. Põõsasmarana ja aasnelgi leiu- ja puudumiskohtade ning sagedasemate põhikaardi põhialade ja mullakaardi mullatüüpide kombinatsioonide sagedused (A) ning samade kohatunnuste paiknemine kanoonilise vastavusanalüüsi esimesel ja teisel teljel (B). Horisontaalteljе omaväärtus = 0,385; vertikaalteljel 0,322. Aasnelgi ja põõsasmarana leiukohad seostuvad eelkõige gleimuldadega.
Kh – paepealne muld, *K* – rähkmuld, *Kg* – gleistunud karbonaatne muld, *G* – gleimuld, *G1* – turvastunud muld.

Küsimused



1. Kumba klassi, kas A või B, määraks vaatluse (tunnus 1 = x, tunnus 2 = R) naiivne Bayesi klassifikaator ja millisesse optimaalne Bayesi klassifikaator, kui õpetusandmestik koosneb järgmistest vaatlustest? Kumb klass on tõenäolisem, kui tunnuste väärtusi ei ole teada?

Klass	Tunnus 1	Tunnus 2
A	x	R
A	x	S
A	y	T
B	z	R
B	x	S
B	x	T
B	y	R
B	y	S

2. Kas kujutise pikslite klassifitseerimisel suurima tõepära meetodil on tarvis teada pikslite omavahelist paiknemist sama andmekihi piires?

3. Mille poolest erineb Mahalanobise vahemaa eukleidilisest vahemaast tunnusruumis?

4. Mille poolest erinevad klassifikatsioonipuu ja regressioonipuu?

5. Milline on kokkulangev osa ja klassifikatsioonitäpsuse kordaja kapa väärtus, kui klasside sagedus vigade maatriksis on järgmine?

0	10	10
10	0	10
10	10	100

6. Milline on kapa väärtus, kui klasside sagedus vigade maatriksi kõigis lahtrites on võrdne?

7. Kas klassifikatsioonitäpsuse kordaja kapa võib olla kokkulangevuse osakaaluga võrdne? Kui jah, siis millisel juhul?

8. Kas klassifikatsioonitäpsuse kordaja kapa saab olla suurem kui kokkulangevuse osakaal?

9. Kas klassifikatsioonitäpsuse kordaja kapa saab olla negatiivne? Kui jah, siis millal?

10. Kas lihtsam on enamik vaatlusi pimesi kahte klassifikatsiooniüksusesse jagada siis, kui klasside sagedus on võrdne või siis kui on ebavõrdne?

11. Kas klasside juhuslikul määramisel oodatavalt õigesti klassifitseeritud vaatluste arv sõltub kasutatud klasside arvust?

12. Mille suhtes arvutatakse vigade maatriksist kasutaja täpsus (*user's accuracy*) ja mille suhtes klassifitseerija täpsus (*producer's accuracy*)?

13. Kas kolmeväärtuselise tunnuse vigade maatriksi kõigis üheksas lahtris võivad olla samad arvud (sagedused)?

14. Kas vigade maatriks on oma diagonaali suhtes sümmeetriline?

15. Kas tundmatu objekti klassikuuluvuse määramisel klassifikatsioonipuu abil saab seletavaid tunnuseid kasutada suvalises järjekorras?

16. Kas klassifikatsioonipuu meetod on rakendatav pideva muutuja klassifitseerimisel? Kui jah, siis millal?

17. Kas objektide klassifitseerimisel kasutatavad tunnused võivad olla nominaalsed?

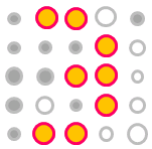
18. Kas klastrite moodustamise reeglid tekivad klasteranalüüsi käigus?

19. Kas klasteranalüüs eeldab klastrite olemasolu andmetes?

20. Arvuta Dice-Sørenseni sarnasuskordaja kahe puistu vahel, milles puude katvused on järgnevad: 1. puistu: lepp 32%, kask 5%, kuusk 17%, mänd 33%; 2. puistu: mänd 32%, kuusk 22%, lepp 32%.

21. Mis on TSS statistiku muutumisvahemik?

3. Statistiline modelleerimine



Mudel on empiiriliste andmete üldistus ja/või korrastatud raamistik, mis võimaldab prognoosida muutujate mitteteadaolevaid väärtusi või jäljendab mingit nähtust või objekti lihtsustatud ja üldistatud kujul. Mudeli koostamisel lähtutakse alati mingitest olemasolevatest arusaamadest ja teoreetilistest eeldustest. Seejuures tuleb valida, kas eelistada mudeli täpsust, üldisust või lihtsust. Keerukate süsteemide puhul, nagu loodus seda on, ei ole võimalik korraga saavutada täpsust, üldisust ja lihtsust. Tuleb valida kas üks või kaks suunda ja teiste osas järeleandmisi teha. Äärmiselt lihtsustatud teoreetilise mudeli tunnetuslik tähtsus võib olla suurem kui rakenduslik tähtsus. Mehhanistlikud protsessimudelid püüavad olla üldised ja realistlikud, aga looduslike protsesside pikemaajaline ette ennustamine on enamasti ebatäpne. Empiirilised statistilised mudelid ei ole reeglina rakendatavad väljaspool lähteandmete väärtuspiirkonda. Viimaste põhiliseks tähtsuseks on nimetatud empiiriliste andmete kondenseerimist (Wissel 1992). Empiirilisi mudeleid saab muuta universaalsemateks teoreetiliste eelduste arvestamisega ja mudelisse lisamisega. Kui statistilisel modelleerimisel eeldatakse juhuslikkuse olemasolu ja mõju (enamasti seda tehakse) nimetatakse modelleerimist **stohhastiliseks** ehk tõenäosuslikuks.

Mudeleid on jagatud veel **kirjeldavateks** ja **prognoosivateks** vastavalt sellele, kas eesmärgiks on eelkõige selliste komplektide otsimine tunnustest, mis tagavad täpseimaid prognoose või on usaldusväärsed hinnangud ainueesmärk. Parima kirjeldava mudeli leidmiseks on kasutatud kahte põhimõtteliselt erinevat teed:

- mudeli statistilise olulisuse hindamine (hüpoteeside kontrolli meetod),
- prognoosivea minimeerimine.

Mudeleid võib klassifitseerida ka vastavalt sellele, kas need lähtuvad staatilistest vaatlusandmetest või siis protsessist. Andmetele orienteeritud mudelid toetuvad kas eksperthinnangutele või statistilistele seaduspärasustele sisend- ja väljundnähtuste vahel. Protsessile orienteeritud deterministlikud mudelid eeldavad, et põhjuslikud seosed protsessis on teada ja võrranditega kirjeldatavad. Statistilised mudelid kasutavad reeglina suurt hulka empiirilisi andmeid, aga võimalik on ka lähtumine vaid genereeritud andmetest või teoreetilistest konstruktsioonidest. Staatilise statistilise modelleerimise eripära on suhtelise stabiilsuse nõue, mis seab teatud piirid hinnangute usaldatavusele. Näiteks liikide leviku staatilised mudelid eeldavad tasakaalu liigi ja keskkonna vahel. Staatilisest mudelist saadud levila ei pruugi tegelikkusele vastata, sest:

- liik ei ole veel jõudnud levida kõigisse talle sobivatesse kohtadesse,
- organismid peavad mõnda aega vastu ka ebasobivates tingimustes ja rände käigus satub osa isendeid neile eluks ebasobivatesse kohtadesse,
- vabalt liikuva liigi ükski isend ei juhtu parasjagu olema vaatluse all olevas kohas,
- mudeli koostamisel kasutatud vaatlusandmete hulgas on registreeritud liigi juhuslikke esinemisi kohtades, kus liik püsivalt elada ei suuda,
- liik on välivaatlusel märkamata jäänud, näiteks pole taimel parajasti märgatavaid maapealseid osi olnud.

Dünaamiline modelleerimine nõuab paljude dünaamikat kirjeldavate parameetrite hindamist, mis on suhteliselt keerukas. Statistilise mudeli puhul piisab etteantud seaduspärasuste konstantsuse eeldusest ja paarist indikaatortunnusest. Määramatust sisaldava dünaamilise mudeli suurem puudus on võimalike arenguvariantide hargnevus ja sellest tulenev vigade võimenduv edasikandumine järgnevate

seisundite hinnangutesse. Väike erinevus modelleeritavate ajaliste muutuste alguses võib viia täiesti erineva lõpptulemuseni. Kui igal ajahetkel on valida *n* arenguvariandi vahel, siis *m* ajahetke pärast võib võimalike variantide arv olla kuni n^m. Ruumiliste protsesside puhul muutuvad dünaamilised mudelid veelgi keerukamaks, sest lisaks protsessi käsitlusele igas lookuses (pikslis, lahtris, kohas) tuleb ruumiliste nähtuste puhul arvestada ka naabruses toimuvate protsesside mõju. Mudeli prognoosiv võime ei pruugi kasvada samaväärselt mudeli keerukuse tõusuga. Seega annavad suhtelises tasakaaluseisundis olevate pindandmete puhul statistilised tasakaaluseisundit eeldavad mudelid suhteliselt kergema vaeva, vähema aja ning oskuste kuluga paremaid tulemusi kui dünaamika mudelid. Siiski on ka uurimusi, kus dünaamilisi mudeleid on õnnestunud rakendada suuremal territooriumil ja ruumiliselt ilmutatud viisil (Urban *et al.* 1991, Moore ja Noble 1993, Roberts 1996, He *et al.* 1999, He ja Mladenoff 1999).

Dünaamika mudelid muudab keerukaks arenguvariantide hargnevus ja vigade võimenduv edasikandumine

Lisaks prognoosile võivad statistilised mudelid anda võrdlusetalone selle kohta, milline oleks liigi levik, koosluse struktuur või muu prognoositav parameeter etteantud tasakaaluliste tingimuste korral. Teoreetilise võrdlusetaloni kõrvutamine empiiriliste andmetega annab võimaluse hinnata prognoosi koostamisel kasutatud eelduste paikapidavust.

Statistiliste mudelite juures eeldatakse seletavate tunnuste ehk argumenttunnuste olemasolu. Olgu nende arv *k*, see tähendab, et argumenttunnused moodustavad k-mõõtmelise tunnusruumi. Teiseks eelduseks on funktsioontunnuse olemasolu. Kui funktsioontunnuseid on rohkem kui üks, on tegemist **mitmemõõtmelise** (*multivariate*) mudeliga. Näiteks tasapinnal paiknemise ja arvukuse dünaamika mudelid on kahemõõtmelised. Dünaamika mudelites saab ühe tunnusruumi dimensioonina käsitleda aega. Tunnusruumi uurimisel on tihti probleemiks suur tunnuste arv (dimensionaalsuse needus) ja tunnuste omavahelised seotused – originaalmõõtmistega määratud tunnusruumi teljed on multikollineaarsed ehk **kollineaarsed**, see tähendab ei ole omavahel risti. Näiteks kaugseirekujutised koosnevad paljudest kiirgusvahemikest, mida ei saa käsitleda omaette tunnustena, sest peegeldunud kiirguse suurema hulgaga ühes kiirgusvahemikus ehk kanalis kaasneb enamasti tugevam kiirgus ka mitmes teises kanalis. Ka multispektraalsed kaugseirekujutised, mis koosnevad rohkem kui sajast kanalist, taanduvad peakomponentanalüüsil vaid kolmeks kuni viieks omavahel sõltumatuks tunnuseks (Curran *et al*. 1998).

Mudeli konstrueerimise eelduseks on veel teatud (enamasti empiirilise) andmestiku olemasolu. Teoreetiliselt tuleks esialgselt ja ilma kindla plaanita kogutud vaatlusandmeid, nagu neid loodusteadustes sageli kohtab, kasutada vaid kirjeldava analüüsi jaoks ja hüpoteeside püstitamiseks. Hüpoteeside kontrollimiseks tuleks andmed koguda uuritavast küsimusest lähtuva täpse plaani kohaselt, mis minimeerib ülejäänud keskkonna kontrollimatu mõju või tuleks korraldada eksperiment. Praktikas kasutatakse aga hüpoteeside tõestamiseks ka kindla plaanita kogutud andmeid lootuses, et need võimaldavad midagi vähemalt 95% kindlusega väita. Andmestik on tavaliselt esitatud andmemaatriksina. Andmestiku üksikmõõtmisi saab kujutada punktidena *k*-mõõtmelises tunnusruumis. Kui tunnusruumi võib käsitleda lõputuna, siis ühe või teise uurimuse andmeid on olemas vaid argumenttunnustega määratud ruumi teatud piirkonnas. Kui seaduspärasuste ekstrapoleerimine väljapoole andmete piirkonda ei ole õigustatud, nimetatakse seda **lubatud piirkonnaks**. Väljaspool lubatud piirkonda andmeid kas puuduvad või ei paku piirkond selle mudeli puhul huvi.

Statistilist modelleerimist võib pidada omaette statistika haruks või ka kirjeldava statistika hulka kuuluvaks. Statistilise modelleerimise eesmärk on leida ja matemaatilises keeles kirjeldada seadus-

pärasused, mis sobivad empiiriliste andmetega kõige paremini, on võimalikult suure üldistusjõuga ja ekstrapoleeritavad ka väljapoole mudeli parameetrite määramiseks kasutatud treeningandmeid ehk **õpetusandmeid**. Mudelisse lülitatavate faktorite mõju statistilise olulisuse kontroll ei ole siiski liiast. Väheoluliste faktorite lisamine mudelisse võib küll prognoosi parandada, kuid muudab mudeli keerukamaks ja väljaspool õpetusandmeid vähemusaldatavaks – toimub mudeli liigsobitumine ehk **ülesobitumine** (*overfitting*).

Ökoloogiliste statistiliste mudelite kolme komponendina on nimetatud ökoloogilist mudelit (reaalsusmudel), andmemudelit ja statistilist mudelit (Austin 2002). **Reaalsusmudel** määrab käsitletavad objektid ja reaalsuses eeldatavasti eksisteerivad seosed ja omadused, **andmemudel** sisaldab andmete kogumise metoodikat ja mõõdetud tunnuseid, statistiline mudel määrab modelleerimise ja mudeli verifitseerimise meetodi. Reaalsusmudelit ja andmemudelit kokku on nimetatud ka **kontseptuaalseks mudeliks**.

Mudeli väärtust hinnatakse eelkõige prognoosi usaldatavusega. On väidetud, et mudeli sobivuse kriteerium on ökoloogilistes uurimustes astumas statistiliste hüpoteeside kontrollimise ja nullhüpoteesi ümberlükkamise asemele (De'ath 2002). See tähendab, et mingi faktori mõju tõestamiseks ei pruugi selle faktori mõju eraldi uurida ning mõju puudumise nullhüpoteesi ümber lükata, piisaks ka faktori mõju olulisuse näitamisest piisavalt täpseid prognoose andvas mudelis. Samasugune eelistus saavutada pigem tõhus mudel kui teoreetiline mõistmine, on ka andmekaevandamise (ptk 3.1) üheks aluspõhimõtteks.

3.1. Andmekaevandamine

Kõige üldisemalt on **andmekaevandamine** (*data mining – DM*) meetodite kogum seaduspärade leidmiseks suures hulgas andmetes. Andmekaevandamise kontseptsioon tekkis koos mahukate andmebaasidega ja seda võib käsitleda ka kui andmebaasidest teabe hankimist (*knowledge discovery in databases – KDD*) (Tan *et al.* 2006). Andmekaevandamisele kitsamas, analüüsimeetodi tähenduses eelneb andmete eeltöötlus ja järgneb järeltöötlus. Automatiseeritud andmekaevandamise viisid kuuluvad ka intellektitehnika ja **tehisõppe** (*machine learning*) valdkonda, statistilised meetodid kirjeldava statistika hulka, osaliselt kattub andmekaevandamine **mustrituvastusega** (*pattern recognition*). Andmekaevandamise meetodid ei ürita määrata seaduspärasuste statistilist olulisust, vaid prognoosivad ainult uuritava tunnuse väärtust või otsivad seaduspärasusi andmetes, sest

- statistiline olulisus kehtib niikuinii vaid kasutatud mudeli õigsuse ja valimite esinduslikkuse eeldusel;
- rakenduslike otsuste aluseks võivad olla ka statistiliselt mitte-olulised tulemused, kuni usaldusväärsemaid mudeleid ei ole.

Andmekaevandamisel otsitakse seoseid suurest andmehulgast, seoste olulisuse tõestamine ei ole eesmärgiks

Andmekaevandamise meetodite erinevus klassikalise statistika vahenditest on eelkõige kontseptuaalne. Kui klassikalises statistikas domineerib algoritmiline lähenemine: täpselt defineeritud töövõtted viivad kindla tulemuseni, siis andmekaevandamise meetodid üritavad suurest andmehulgast seoseid ja reeglipärasusi leida heuristiliselt, nii et analüüsi meetodid ja protseduur täpsustuvad analüüsi käigus. Kui ühe meetodiga head seost ei leita, siis proovitakse järgmist. Kui seose tüübi kohta on vähe ette teada, siis ei saa piirduda teoreetiliselt kõige õigema mudeliga.

Andmekaevandaja on pragmaatik

Modelleerimismeetodina võib andmekaevandamise jagada kolme staadiumi:

- seaduspärasuste otsimine,
- seaduspärasuste modelleerimine,
- mudeli verifikatsioon uute andmetega. Protsessi korratakse, kuni on leitud enam-vähem sobiv viis andmetest leitud teabe kirjeldamiseks.

Lihtsaim andmekaevandamise meetod on sagedustabelite koostamine. Andmekaevandamise sissejuhatavas õpikus (Tan *et al.* 2006) on nimetatud järgmised põhilised meetodid: koondstatistikute arvutamine, andmete visualiseerimine ja tabelite koostamine, klassifikatsioonipuud, lähima naabri klassifikaatorid, Bayesi (tõenäosjaotusi kasutavad) klassifikaatorid, tehisnärvivõrgud, tugivektormasinad, paralleelseid mudeleid kasutavad meetodid ehk ansamblimeetodid, seoste analüüs, klasteranalüüs, anomaaliate tuvastus.

Wu *et al.* (2008) loetlevad järgmised kümme 2006. aasta lõpu seisuga teadusringkondades kõige mõjukamat andmekaevandamise algoritmi: *C4.5*, *k-means*, *SVM*, *Apriori*, *EM*, *PageRank*, *AdaBoost*, *kNN*, *Naive Bayes* ja *CART*. *K-means* algoritm võrdleb klassifitseeritavaid objekte klassikeskmetega (*k* tähistab klastrikesmete arvu); *k-means* meetodist on juttu ka klasteranalüüsi peatükis (ptk 2.3.1), tugivektormasinatest (*SVM*) peatükis 3.4.5.4, *kNN* meetodit on mainitud sarnasusele tugineva järeldamise peatükis (ptk 3.4.6), naiivne Bayesi klassifikaator on klassifitseerimismeetodite peatükis

(ptk 2.3.2), *CART* liikide ja elupaikade leviku modelleerimise osas (ptk 5.6.6.1) ja *AdaBoost* ansamblimeetodite alapeatükis (ptk 5.6.10.1). *C4.5* on otsuste puu algoritm, mis lubab hargnemisi korraga rohkem kui kahte allüksusesse, mis võimaldab mitmetunnuselisi otsustamiskriteeriume ja mis kasutab klassifikatsiooni tõhususe kriteeriumina infoteooria mõõdikuid. Algoritmi uuem versioon on vabalt saadaval aadressil http://rulequest.com/GPL/C50.tgz. *Apriori* algoritm otsib sagedasemeid väärtuskombinatsioone. Kombinatsiooni sageduse määramise arvutusmaht kasvab astmeliselt võimalike kombinatsioonide arvu suurenedes. EM (*Expectation–Maximization*) algoritm sobitab segajaotusi. *PageRank* on veebiotsingu tulemuste järjestamise algoritm, mille mõtlesid välja Sergey Brin ja Larry Page (Brin ja Page 1998) ja millele on rajatud Google otsingumootor. *PageRank* moodustab igale veebilehele väärtushinnangu, mida määrab sellele veebilehele suunatud viitade (häälte) arv veebis. Seejuures arvestatakse kaaluna iga häält andva veebilehe enda väärtust (temale suunatud viitade arvu).

Tarkvara

Andmekaevandamise rakendustarkvara täielikumad komplektid on lisamoodulitena suuremates statistilise andmetöötluse pakettides, nagu *Statsoft Statistica Data Miner*, *SAS Enterprise Miner*. Vabavarast tuntakse eelkõige paketti *RapidMiner*. Andmekaevandamise programmikoode on saadaval ka programmeerimiskeskkonnas R.

3.2. Säästvusreegel

Säästvusreegel ehk parsimooniareegel ehk Occami (Ockhami) habemenuga (*parsimony rule, Occam's razor*) on keskaegsele filosoofile Ockhami Williamile omistatav loogikaprintsiip: *Entia non sint multiplicanda praeter necessitatem* – keerukust ei tohiks eeldada ilma tungiva vajaduseta ehk kui mingit nähtust on võimalik seletada lihtsalt, siis ei ole põhjust kasutada keerukamat seletust.

Säästvusreegli ehk parsimooniareegli järgi tuleb eelistada lihtsamat seletust

Ockhami William sündis 1285 Ockhami külas Inglismaal, liitus varases nooruses frantsiskaanlaste orduga ja õppis teoloogiat Oxfordis. Sattus oma vaadete pärast opositsiooni teaduskonna professoritega. Vastuoludest sai teada ka paavst Johannes XXII, kes kutsus ta 1324. aastal paavsti kohtusse Prantsusmaale Avignoni. Kohtus oli Ockhami Williami oponendiks Jogh Lutterell, endine Oxfordi ülikooli kantsler. Ockhami Williamit ei mõistetud küll kunagi süüdi, kuid tema jätkuvad dispuudid ja vastuolu paavstiga sundisid teda 1328. aastal Avignonist põgenema. Ockhami William leidis kaitset keiser Louis IV juures, kellega ta reisis Münchenisse. Seal kirjutas Ockhami William oma kõige teravamad paavsti võimu vastased traktaadid.

Säästvusreeglit on sõnastatud ka: keerukust ei tuleks eeldada ilma tõsise vajaduseta. Ockhami habemenuga kasutatakse selleks, et nähtuste seletustest maha raseerida kõik kontseptsioonid, muutujad ja konstandid, mis ei ole nähtuse äraseletamiseks tingimata vajalikud. Sellega muutuvad mudelid lihtsamaks ja väheneb ka võimalus eksiteele sattuda tulenevalt andmete ebajärjekindlusest, ebatäielikkusest või dubleerimisest. Säästvus ei välista keerukaid mudeleid, vaid nõuab keerukama mudeli eeliste põhjendamist. Säästvus on seega mudeli omadus. **Parsimoonne** ehk säästlik mudel on selline, mis vastab empiirilistele andmetele ning mille korral on mudeli parameetrite arv minimaalne ja vabadusastmete arv maksimaalne. Parsimoonne on näiteks kladistilises analüüsis rakendatav oletus, et evolutsioon on kulgenud rada mööda, mille puhul on evolutsiooniliste muutuste hulk minimaalne. Eeldatakse, et evolutsioonis ei ole olnud keerdkäike ja tagasipöördumisi. Paraku on selliseid lühimaid võimalikke teid enamasti rohkem kui üks.

Kuigi säästvusprintsiip tundub triviaalne, on see ühest küljest kogu teoreetilise teaduse ja modelleerimise alus. Teisest küljest jääb aga kasutatud keerukustaseme põhjendus tihti vaid (subjektiivse) arvamuse tasemele – kas põhjusel, et mudeli keerukust ei ole alati lihtne mõõta ja võrrelda või põhjusel, et mudelite headust ei saa üheselt ja täpselt hinnata. Milline on ühel või teisel juhul vajalik mudeli täpsus ja milline on sobiv keerukusaste, on tihti raske otsustada.

Iga valimi puhul eksisteerib lõputu hulk võimalikke mudeleid, mis valimis olevaid väärtusi põhjendada võivad, kuna mudel püüab esindada üldist seaduspära lõpmata suures üldkogumis, millest olemasolevad andmed on vaid üks lõpliku suurusega valim. Esindamata vaatlusi püüavad ära arvata mudelis olevad reeglid ja parameetrid. Näiteks läbi korrelatsioonivälja võib tõmmata mitmeid kõveraid, mis kirjeldavad sama hästi kahe nähtuse vahelist statistilist seost. Säästvusreegli kohaselt tuleks seose kirjeldamiseks olemasolevate andmete piirkonnas eelistada lineaarset mudelit ka teades, et lineaarne seos on põhimõtteliselt sobimatu väljaspool seda andmestikku.

Lihtsamat seost tuleks eelistada seni, kuni teatud keerukam seos ei ole andmete järgi ilmne või siis teoreetilistest teadmiste kohaselt eelistatud. Lihtsamate mudelite eelistamist põhjendatakse ka võimalike lihtsamate mudelite väiksema arvuga ja sellest tuleneva väiksema võimalusega, et mudel sobib vaatlusandmetega vaid tänu juhuslikule kokkusattumisele. Säästvusreegli järgi ei tuleks mudelitesse

lülitada mittevajalikke muutujaid. Liialt suur arv seletavaid tunnuseid suurendab vabadusastmete arvu ja ohtu saada näiliselt tugevaid seoseid, mis väljaspool mudeli sobitamise andmeid ei kehti. On näidatud, et argumenttunnuste arvu suurendamine üle optimaalse lausa halvendab prognoose, kuna kõrvalised tunnused toovad endaga kaasa vaid müra (Harell *et al*. 1996, Remm 2004).

Liigkeerukas mudel kaldub ülesobituma

3.3. Modelleerimise etapid

Modelleerimise neli suurt tööetappi on:

- õpetusandmete hankimine ja mudeli formuleerimine,
- mudeli kalibreerimine ehk sobitamine õpetusandmetele,
- mudeli hindamine (kontrollimine sõltumatute kontrollandmetega, prognoosijääkide ja mudeli tundlikkuse analüüs, mudeli täiustamine ja uuesti hindamine),
- mudeli kasutamine.

3.3.1. Andmete kogumine

Mudeli tüübi valikul ja mudeli moodustamisel saab lähtuda olemasolevatest teoreetilistest teadmistest ja kogemusest. Mudeli parameetrite sobitamiseks on vajalik mingi õpetusandmestik, mis reeglina on väljavõte uuritavast üldkogumist – õpetusvalim ehk õpetuskogum. Ka uurimuse mõõtkava ja määratavad faktorid sõltuvad kontseptsioonist ja hüpoteesidest. Mõnikord ka vastupidi. Enamasti on uurijal siiski enne andmete kogumist mingid hüpoteesid selle kohta, millised faktorid võivad olla olulisemad.

Andmete kogumist saab stratifitseerida vastavalt olulisematele faktoritele või väärtuskombinatsioonide esinemise aladele ehk **kombineeritud eraldiste** ehk *gradsect* meetodil (Gillison ja Brewer 1985, Austin ja Heyligers 1989, Neldner *et al.* 1995, Wessels *et al.* 1998). Kombineeritud eraldiste meetodi järgi jagatakse teemakihid osadeks vastavalt faktorite väärtustele. Iga eraldis esindab faktorite väärtuste ühte kombinatsiooni ja iga kombinatsiooni kohta võib olla mitu eraldist. Vaatlusalal valitakse juhuslikud eraldised igast väärtuskombinatsioonist ning igasse eraldisse luuakse juhuslikult paiknevad vaatluskohad. Eraldiste ja vaatluspunktide valikul võib olla täiendavaid eraldise suuruse ja omavahelise paiknemisega seotud piiranguid. Kui faktoreid ja liike on palju ja eri liikide jaoks on olulised erinevad faktorid, siis võib suhteliselt homogeensete eraldiste moodustamine olla keerukas. Wessels *et al.* (1998) on erinevaid valikumeetodeid võrrelnud ja leidnud, et faktorite järgi valik ja ruumiliselt kihistatud valik annavad enam-vähem samavõrd esindusliku väljavõtte, aga süstemaatiline ja lihtne juhuvalik on reeglina vähemesinduslik.

Kui soovitakse kasutada fikseeritud suurusega osavalimeid igast faktorkihist, aga andmestikus ei ole osa faktorväärtuste kombinatsioone piisavalt esindatud, siis tuleks kaaluda täiendavate vaatluste lisamist. Osade faktorkombinatsioonide puudumine andmestikus mõjutab prognoosi usaldusväärsust tõenäoliselt rohkem kui valimi moodustamise põhimõtted. Lünklike andmestike kasutamisel tuleb kas puuduvate väärtustega vaatlused analüüsidest välja jätta või siis täita tühjad lahtrid teiste tunnuste järgi prognoositud või juhuslike väärtustega. Prognoositud väärtused leitakse enamasti mingi esialgse regressioonimudeli järgi. Kui muud mudelit ei ole, siis on juhusliku muutuja väärtuse parimaks prognoosiks tema keskväärtus. Puuduvate väärtuste asendamisel tunnuse keskmisega väheneb tunnuse hajuvus. Kuid kui tunnuse varieeruvust ei uurita ja ka mudel on loodud eelkõige muutuja oodatava väärtuse hindamiseks, mitte muutuja varieeruvuse hindamiseks, siis võib puuduvaid vaatlustulemusi asendada vastava muutuja keskväärtusega. Kasutatakse ka puuduvate väärtuste asendamist prognoositud väärtuse ja juhusliku arvu summaga (Haining 2003) ning asendamist juhuslikult valituga teada olevate väärtuste hulgast.

Puuduvate väärtuste asendamisel tunnuse keskmisega väheneb tunnuse hajuvus

Andmete kogumise eeltööna tuleks uurida nähtuse ruumilist autokorrelatsiooni (ptk 5.1). Kui osa vaatluspunkte on üksteisele lähemal kui ruumilise autokorrelatsiooni ulatus, tuleb need lugeda pseudoreplikatsioonideks ehk **näivkordusteks** (vt ka ptk 5.1). See tähendab, et vaatlused osaliselt kordavad üksteist ehk naabervaatluste väärtuseid põhjustavad samad protsessid (sündmused) või need ei ole üksteisest sõltumatud ja vabadusastmete arv ei vasta vaatluste arvule. Kui planeeritud **replikatsioonid** ehk kordused on statistiliste andmete kogumisel vajalikud andmete varieerumise ja mõõtmisvea hindamiseks, siis pseudoreplikatsioonid on teadmata ulatusega ja mõjutavad tulemusi ning selle mõju suurust on keerukas hinnata.

3.3.2. Mudeli formuleerimine

Mudeli formuleerimine koosneb mudeli tüübi ehk modelleerimise meetodi ning parameetrite ja nende määramise algoritmi valikust. Enamik statistilisi mudeleid eeldavad kindlat tüüpi muutujaid ja nende muutujate jaotustüüpi. Järelikult tuleb kõigepealt kas teoreetiliselt teada või siis olemasolevate andmete põhjal selgitada muutujate jaotustüübid. Näiteks liikide ohtruse vaatlusandmed ja mis olulisem – prognoositavate väärtuste hälbed, on harva normaaljaotusega. Seetõttu ei ole õige kasutada arvukuse prognoosimisel normaaljaotust eeldavaid meetodeid. Mitteparameetrilised meetodid ei sea küll eeldusi jaotustüübile, kuid nõuavad siiski nominaalsete ja numbriliste muutujate eristamist.

Mudeli valikul võib alustada kas täisparameetrilisest mudelist ja siis järk-järgult parameetreid vähendada, kuni saavutatakse sobivalt parsimoonne variant, või siis alustada kõige lihtsamast mudelist ja kontrollida, kui palju uute parameetrite lisamine mudelit parandab.

Paljude tunnuste puhul tekib küsimus, milliseid teisendusi kasutada ehk millises teljestikus väärtusi esitada. Kas nõlva kaldenurka tuleks mõõta kraadides või protsentides, kas faktori väärtuste puhul on olulised võrdsed lineaarsed väärtusvahemikud või kordsed vahemikud, milliseid tunnuseid oleks õigem käsitleda logaritmskaalas või muul viisil teisendatult? Kuidas mõõta reljeefi keerukust ja suhtelist kõrgust, millise näitaja järgi mõõta koosluse keerukust, kas nähtuse osakaalu on õigem hinnata protsentides või nähtuse esinemise ja puudumise vahekorra logaritmi ehk logiti järgi? Kas inimestele olulisi vahemaid tuleks mõõta pikkuse ühikutes või vahemaa läbimisele kuluvas ajas? Nii nagu vahemaa on suhteline geograafilises ruumis, on suhtelised ja mitmeti mõõdetavad ka sarnasused ja erinevused ehk vahemaad tunnusruumis. Valdavalt on siiski üheselt mõistetav väärtuste järjestatavus. Keegi ei kahtle selles, et 20° kaldenurgaga nõlv on etteantud koordinaatsüsteemis järsem kui 10° nõlv, vaieldav on aga, kas 11° nõlv erineb 2° kaldenurgaga nõlvast rohkem kui 20° nõlvast.

3.3.2.1. Ökoniši mudelid

Liikide leviku modelleerimisel tehakse vahet ökoloogilise **põhiniši** (teoreetilise ehk fundamentaalse niši) ja **realiseerunud niši** vahel. Põhinišš sõltub liigi füsioloogilistest nõudlustest ja looduses esinevate tingimuste kompleksidest. Realiseerunud nišš sisaldab lisaks põhiniši tingimustele ka organismidevahelisi suhteid (biootilisi interaktsioone). Põhiniši ja realiseerunud niši vahel vahetegemine modelleerimisel näitab, kas liigi prognoositud levik on saadud lähtudes liigi füsioloogilistest nõudlustest või liigi esinemise ja ohtruse välivaatlustest.

Enamikus ökoloogiliste statistiliste prognoosimudelite puhul on eeldatud, et ökonišši väljendava seose kuju on sümmeetriline ja vastab normaaljaotuse tihedusfunktsioonile. On näidatud, et mitmete liikide esinemissagedus ei ole keskkonnateguri teljel sümmeetriline, kuigi sümmeetrilised seosed võivad olla sagedasemad (Austin 1987, Oksanen ja Minchin 2002). Sobimatu seosetüübi eeldamine võib viia ebaõnnestunud mudelini. Ökonišile vastava levila kaardistamisest on juttu peatükis 5.6.

Universaalse valemi ökonišši kirjeldava kõvera modelleerimiseks on pakkunud Huisman *et al*. (1993).

$$E = M \frac{1}{\left(1 + e^{(a+bx)}\right)\left(1 + e^{(c-dx)}\right)}, \qquad [3\text{-}1]$$

kus E on oodatav väärtus, mis sõltub keskkonnafaktori gradiendi x väärtusest, faktori maksimaalsest võimalikust väärtusest M ja neljast kasutaja poolt määratud väärtusega parameetrist (a, b, c, d). Parameetrite fikseerimise või nulliga võrdsustamise abil võib saada nii asümmeetrilisi, sümmeetrilisi, konstantseid, ühtlaselt lineaarselt muutuvaid kui ka platookujulisi seoseid (Oksanen ja Minchin 2002).

Uurimused

Libisevat keskmist liigi arvukuse ja keskkonnatingimuste vahelise seose kirjeldamiseks on kasutanud Remm (1987), Huntley *et al.* (1989), Prentice *et al.* (1991), Bartlein ja Whilock (1993). Normaaljaotust on liikide vastavusfunktsiooni modelleerimiseks kasutanud ter Braak ja van Dam (1989), Birks *et al.* (1990). Ter Braak *et al.* (1993) arendasid välja multinomiaalse logitmudeli, kus regressioonikordajad on seotud liikide *n*-mõõtmelise normaaljaotusega unimodaalse vastavuspinna optimumidega. Mudeli unimodaalsus tähendab, et igal liigil on vastavuspinnal üks (teistest liikidest erinev) optimum.

3.4. Mudelite tüübid

Loodusuuringutes kasutatava statistilise modelleerimise põhilised meetodid on: regressioonanalüüs, mitmesugused muud aditiivsed mudelid (üldistatud, mitteparameetrilised, mittelineaarsed), dispersioonanalüüs, klassifikatsioonipuud, ordinatsioon, sagedustabelite analüüs, tehisnärvivõrgud ja teised intellektitehnika meetodid, sarnasusele tuginev järeldamine ja stohhastiline modelleerimine. Aegridade ja ruumiandmete puhul lisandub aja ja ruumi arvestamine. Põhiliselt tuleb arvestada trende ning autokorrelatsiooni ajas ja ruumis. Statistilise modelleerimise puhul tuleb arvestada ka tunnusruumi pidevust – autokorrelatsioon esineb ka tunnusruumis.

3.4.1. Regressioonimudelid

Regressioonimudel on statistilise seose lihtsustatud ja formaliseeritud esitus. Regressioonimudelit üldiselt käsitleb peatükk 1.5.3.

3.4.1.1. Lihtsad lineaarsed mudelid

Lihtsad lineaarsed mudelid sisaldavad vaid lineaarseid seoseid ja omavahel liituvaid mõjusid, neid kasutatakse regressioon-, dispersioon- ja kovariatsioonanalüüsis. Lisaks liituvatele mõjudele eeldavad lineaarsed mudelid iga vaatluse sõltumatust ja ühekordset esindatust, kõigi olulise mõjuga argumenttunnuste kaasamist mudelisse, prognoosi ja vaatluste hälvete normaaljaotust ja **homoskedastilisust** (konstantne dispersioon argumenttunnuste kõigi väärtuste korral). Mainitud kitsendustest vabanemiseks on välja töötatud üldised ja üldistatud lineaarsed mudelid, mida iseloomustatakse alljärgnevalt eraldi peatükkides. Regressioonanalüüsi ja dispersioonanalüüsi iseloomustus on esitatud andmetöötluse põhitõdede peatükis (ptk 1.5.3 ja 1.5.4). Siin peatükis püütakse neid meetodeid iseloomustada eelkõige statistilise modelleerimise aspektist.

Lineaarne regressioonanalüüs

Lihtsa lineaarse regressiooni mudel avaldub valemina, mille viimane liige on normaaljaotusega juhuslik viga keskväärtusega 0 ja standardhälbega σ

$$Y = b_0 + b_1 X + viga\left\{\sim N\left(0, \sigma^2\right)\right\}. \qquad [3\text{-}2]$$

Lineaarset regressioonimudelit kasutatakse:

- muutuja *Y* prognoosimiseks muutuja *X* kaudu,
- muutujate *X* ja *Y* vahelise põhjusliku seose hüpoteesi tõendamiseks (mitte tõestamiseks),
- muutuja *Y* varieeruvuse osaliseks seletamiseks muutuja *X* abil.

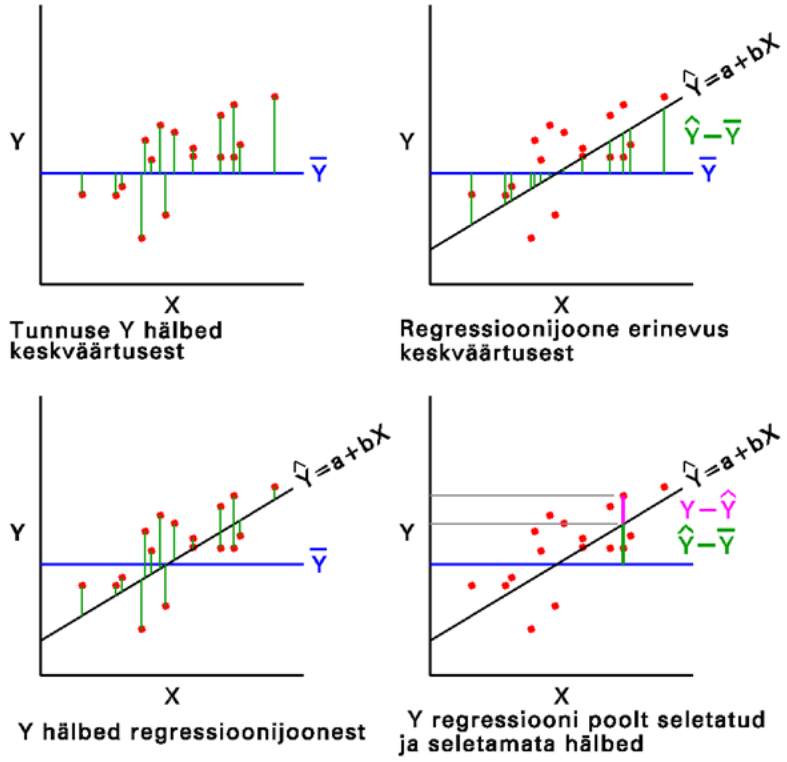
Lihtsa lineaarse mudeli graafiline kuju on **regressioonisirge** (*regression line*), mis sobitatakse läbi kahemõõtmelisel pinnal olevate punktide **vähimruutude meetodil** (*ordinary least squares – OLS*). Vähimruutude meetod on vahend leidmaks selline lahend, mille puhul regressioonijääkide ruutude summa on minimaalne (determinatsioonikordaja on maksimaalne). Vähimruutude meetod annab suurima tõepära (ptk 1.2 ja 3.6.1) hinnangu juhul, kui lihtsa lineaarse mudeli eeldused on täidetud. Vähimruutude meetodil leitud joon läbib muutujate keskväärtustele vastava punkti juhul kui vigade normaaljaotuse eeldus on täidetud. Vaatluste hälbed oma keskväärtusest saab jagada regressiooniga

äraseletatud ja seletamata hälveteks (joonis 3-1).

Mitmetunnuselise ehk mitmese regressiooni korral on argumenttunnuseid mitu. Kui ühe pideva argumenttunnuse korral on regressiooni graafiline kujutis sirgjoon, siis kahe ja enama pideva argumenttunnuse korral moodustub **vastavuspind** (*response surface*).

Kuna **kovariatsioonanalüüsi** (*analysis of covariance – ANCOVA*) puhul on üks seletav tunnus pidev ja teine diskreetne, siis selle tulemuste graafiline kujutis on eraldi regressioonijoon kategoorilise muutuja iga väärtusklassi kohta (vt ka ptk 1.5.2.1). Kategoorilise muutuja mõju avaldub nende regressioonijoonte erinevuses. Kui kõik argumenttunnused on kategoorilised, kasutatakse dispersioonanalüüsi (ptk 1.5.4). Dispersioonanalüüs on meetod pideva tunnuse varieeruvuse jaotamiseks etteantud gruppide vaheliseks ja gruppide siseseks varieeruvuseks.

Argumenttunnuse teisendatud väärtusi kasutavad mudelid, nagu näiteks polünomiaalgressioon, loetakse lineaarsete mudelite hulka, sest teisendatud argumenttunnust võib käsitleda omaette tunnusena. Teisendatud funktsioontunnuse puhul võib lineaarseks pidada mudelit, mille oodatavad vead on aditiivsed. See tähendab, et juhuslik viga liitub faktorite mõjudele ja mitte muul moel, näiteks ei astenda mõjusid. Polünomiaalgressioon kasutatakse ökoloogiliste nähtuste puhul keskkonnafaktori optimumi tuvastamiseks uuritavas väärtusvahemikus ning teadaolevalt mittelineaarse, kuid täpsemalt tundmata seose üldistatud kujul kirjeldamiseks. Näiteks võib liigi või muu nähtuse logaritmitud arvukust kirjeldada ruut-teisendatud keskkonnagradiendi järgi.



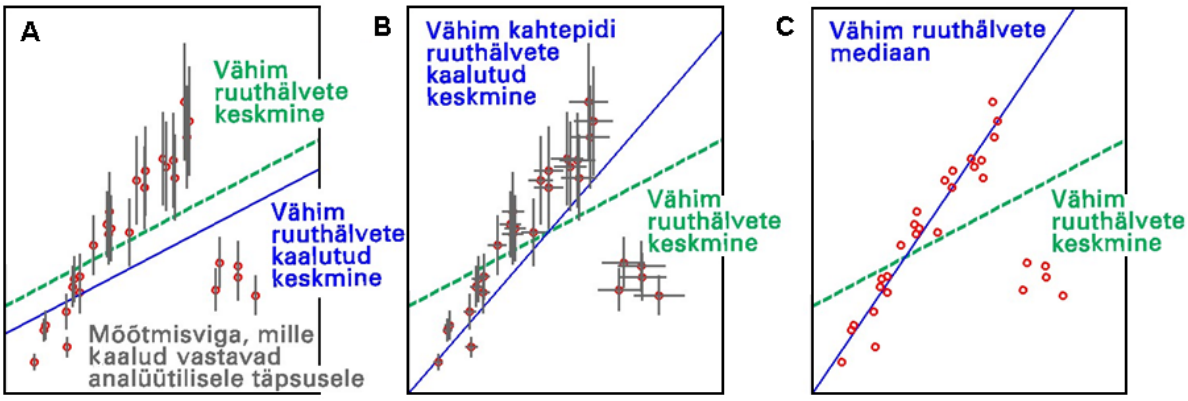
Joonis 3-1. Funktsioontunnuse hälbed keskväärtusest ja regressioonimudelist. Hälbeid mõõdetakse reeglina piki Y-telge, aga on ka teisi variante. Mudeli abil ära seletatud ($\hat{Y}-\overline{Y}$) ja seletamata ($Y-\hat{Y}$) hajuvuse vahekord näitab mudeli kasulikkust.

Regressioonimudelite puhul tehakse veel vahet, kas argumenttunnuse *X* puhul eeldatakse mõõtmisvigade olemasolu või ei. Enamasti eeldatakse, et muutuja *X* ei sisalda mõõtmisvigu (fikseeritud mõjudega mudel ehk esimest tüüpi mudel) (ptk 1.5.4). Esimest tüüpi mudeli puhul mõõdetakse

regressioonijääke vaid y-telje sihis (funktsioontunnuse skaalas) ja mudel on kasutatav vaid funktsioontunnuse prognoosimiseks argumenttunnuse järgi. Teist tüüpi mudeli puhul eeldatakse ka argumenttunnuse juhuslike mõõtmisvigade olemasolu (joonis 3-2). Sellele lisaks tuleb eeldada, et tunnuste varieeruvuse ainus põhjus on nende mõõtmisvead. Vastasel juhul jõutakse määramatusse. Teist tüüpi mudeli sirge paikneb tunnuse *Y* tunnuse *X* järgi prognoosimise ja tunnuse *X* tunnuse *Y* järgi prognoosimise regressioonijoone vahel. Ülevaade vähimruutude regressiooni kasutamisega seonduvatest probleemidest annab tabel 4.

Tabel 4. Vähimruutude regressiooni kasutamisel esinevad probleemid Haining (1990) järgi.

Probleem	**Tagajärg**	**Kontrollimisviis**	**Abinõu**
Jäägid ei ole normaaljaotusega	*F* statistikul baseeruvad testid ei anna õigeid tulemusi	Jääkide graafik, Shapiro *W* test.	Funktsioontunnuse teisendamine (Box-Cox teisendus). Üldistatud lineaarsed mudelid.
Jääkide hajuvus ebaühtlane	Regressioonivea hinnang hälbega	Jääkide graafik, Anscombe test.	Funktsioontunnuse teisendamine. Kaalude omistamine vaatlustele. Robustsed regressioonid.
Jäägid ei ole sõltumatud	Ülehinnatud R^2	Jääkide graafik, autokorrelogramm.	Iteratiivsed üldistatud vähimruutude meetodid.
Seose mittelineaarsus	Mudeli vähene kirjeldav võime, sisutud tulemused.	Korrelatsiooniväli graafikuna.	Muutujate teisendamine. Robustsed regressioonid.
Argumenttunnuste multikollineaarsus	Ebastabiilne mudel	Korrelatsioonimaatriks	Argumenttunnuste arvu vähendamine, argumenttunnuste teisendamine, kantregressioon.
Võimalike argumenttunnuste suur hulk	Optimaalse tunnustekomplekti valik raske	Tunnuste ühisjaotuste graafikud.	Samm-sammuline regressioon, muutujate teisendamine, mudeli lihtsustamine.
Erindid	Mudeli parameetrid sõltuvad liigselt erinditest.	Kirjeldav andmeanalüüs.	Erindite eemaldamine. Robustsed regressioonimeetodid.
Vead andmetes	Sisutud tulemused	Vigade otsimine kirjeldavate meetoditega.	Vigaste andmete eemaldamine või asendamine.
Lünklikud andmed	Olemasolevate andmete ebaefektiivne kasutus. Ebatäpsed tulemused.		Lünklike vaatluste eemaldamine või andmete asendamine prognoositud väärtustega.
Nominaalsed tunnused	Normaaljaotust eeldav regressioon ei sobi		Üldistatud lineaarne mudel.



Joonis 3-2. Robustsed regressioonid Szava-Kovats (2001) järgi. Eeldatakse, et suuremate väärtuste puhul on suurem mõõtmisviga. A – Vertikaaltelje suhtes ruuthälvete keskmist ja ruuthälvete kaalutud keskmist minimeerivad mudelid, B – Mõlema telje suhtes ruuthälvete keskmist ja ruuthälvete kaalutud keskmist minimeerivad mudelid, C – Vähim ruuthälvete mediaan kirjeldab tõhusalt enamikku vaatlustest.

Lineaarne segamudel

Lineaarne segamudel hõlmab fikseeritud mõjusid, mudeli parameetreid ja juhuslikke hälbeid:

$$Y = Xa + Zb + viga\left\{\sim N\left(0,\sigma^2\right)\right\}, \qquad [3\text{-}3]$$

kus X on argumenttunnuse fikseeritud mõjude maatriks, Z on argumenttunnuse juhuslike mõjude maatriks, a ja b on mudeli parameetrid, $N(0,\sigma^2)$ on normaaljaotusega juhuslik viga keskväärtusega 0 ja standardhälbega σ.

Segamudeli sobitamine toimub suurima tõepära (ptk 1.2 ja 3.6.1) kriteeriumi järgi.

Box-Cox teisendus

Mittelineaarseid seoseid saab sageli esitada lineaarse mudelina pärast seost lineariseerivat teisendust. Universaalne valem andmete teisendamiseks on Box-Cox teisendus (Box ja Cox 1964) kujul

$$T\left(Y\right) = \frac{Y^{\lambda} - 1}{\lambda}, \qquad [3\text{-}4]$$

kus λ on vahemikus –1…+1 olev astendaja. Kui λ = 1, siis on teisendus lineaarne, kui λ = 0, siis kasutatakse logaritm-teisendust.

Box-Cox teisendust on kasutanud näiteks Puttock *et al.* (1996) põdra arvukuse modelleerimisel.

3.4.1.2. Üldised lineaarsed mudelid

Üldised lineaarsed mudelid (*general linear models – GLM*) on meetodid pidevate muutujate prognoosimiseks ja seostamiseks argumenttunnuste komplektiga. Üldiste lineaarsete mudelite hulka kuuluvad muu hulgas eelpool mainitud dispersioonanalüüs, kovariatsioonanalüüs ja regressioonanalüüs. Erinevalt lihtsast mitmetunnuselisest regressioonanalüüsist võib üldises lineaarses mudelis olla mitu funktsioontunnust – funktsioontunnuste plaanimaatriks. Regressioonikordajate vektori asemel on

regressioonikordajate maatriks, mis võib sisaldada rohkem kui ühte funktsioontunnuse veergu.

Üldine lineaarne mudel lähtub samadest eeldustest kui dispersioonanalüüski (ptk 1.5.4), aga lubab omavahel korreleeruvaid mõjusid ning pidevate ja diskreetsete argumenttunnuste kombineerimist. Seosed argumenttunnuste vahel ei pruugi olla lineaarsed ja pööratavad. Argumenttunnused võivad olla omavahel korreleerunud juhuslikult (valim on liiga väike), ekslikult (sisuliselt sama muutuja on kaks korda argumenttunnuste hulka sattunud) või planeeritult (mees/naine eraldi tunnustena).

Kuna funktsioontunnused võivad üldises lineaarses mudelis olla omavahel korreleerunud, siis ei pruugi ühetunnuselised seose olulisuse testid õigeid tulemusi anda ja üldiste lineaarsete mudelite olulisuse testimiseks kasutatakse mitmemõõtmelisi ehk multivariantseid teste. Mitmemõõtmelised testid võimaldavad hinnata ka funktsioontunnuste omavahelise seotuse olulisuse tõenäosust. Lisaks mitmemõõtmelisele funktsioontunnusele võimaldavad üldised lineaarsed mudelid analüüsida korduvalt mõõdetud faktoreid.

Kolmas üldiste lineaarsete mudelite erinevus mitmesest regressioonist on võimalus kasutada nominaalseid argumenttunnuseid analoogiliselt dispersioon- ja kovariatsioonanalüüsiga. Nominaalsete muutujate kasutamisel teisendatakse need kaheväärtuselisteks tunnusteks. Üldised lineaarsed mudelid eeldavad, erinevalt üldistatud lineaarsetest mudelitest, funktsioontunnuste normaaljaotust. Peale selle, erinevalt üldistatud lineaarsetest mudelitest, ei teisendata üldistes lineaarsetes mudelites funktsioontunnuse ootust. Kui teisendus on vajalik, siis tuleb see enne analüüsi ära teha.

Üldises lineaarses mudelis võib olla mitu funktsioontunnust, faktorid võivad olla nominaalsed ja/või omavahel seotud

Kui mudelis on rohkem kui üks funktsioontunnus, siis nimetatakse seda mudelit mitmemõõtmeliseks ehk multivariantseks. Mitmemõõtmelised seosekordajad on keerukamad kui nende ühemõõtmelised analoogid, nagu näiteks korrelatsioonikoefitsient. Mitmemõõtmelise seose mõõtmisel tuleb arvestada mitte ainult argumenttunnuste mõjuga ühele funktsioontunnusele, vaid ka funktsioontunnuste omavahelise mõjuga.

Mitmemõõtmelises mudelis olevate statistilise seose tugevust näitavaid **mitmemõõtmelisi statistikuid** arvutatakse enamasti faktorite (i) omaväärtuste (λ_i) abil. Need on näiteks:

- Wilksi lambda (W) muutumispiirkonnaga 0 … 1 (1 tähendab seose puudumist) on determinatsioonikordaja mitmefaktoriline analoog, mis näitab argumenttunnuste abil ärakirjeldatud varieeruvuse osa,
- Pillai indeks (PI),
- Hotelling-Lawley statistik (HL) ja
- Roy suurim juur (RJ).

$$W = \prod_i \frac{1}{1+\lambda_i} \qquad [3\text{-}5]$$

$$PI = \sum_i \frac{\lambda_i}{1+\lambda_i} \qquad [3\text{-}6]$$

$$HL = \sum_i \lambda_i \qquad [3\text{-}7]$$

$$RJ = \max_i \left(\lambda_i\right) \qquad [3\text{-}8]$$

3.4.1.3. Üldistatud lineaarsed mudelid

Üldistatud lineaarsetes mudelites (*generalized linear models – GLMZ, GLZ*) lisatakse üldiste lineaarsete mudelite omadustele võimalus asendada funktsioontunnuse ootus selle teisendusega. Funktsioontunnuse teisendust nimetatakse **seosefunktsiooniks** (*link function*). Seosefunktsiooni puhul eeldatakse, et sellele vastab üks-üheselt pöördfunktsioon, mida nimetatakse **vastavusfunktsiooniks** (*response function*). Seosefunktsioon võimaldab funktsioontunnuse lineariseerimist ja hoiab prognoositud väärtused võimalike väärtuste vahemikus. Üldistatud lineaarsete mudelite puhul eeldatakse, et funktsioontunnuse keskväärtuse funktsioon $f(\mu_i)$ on valitud nii, et see on plaanifunktsioonide järgi lineaarsete mudelite kaudu kirjeldatav:

$$f(\mu_i) = \sum_j \beta_j f_j(x_i), \qquad [3\text{-}9]$$

kus f_j on argumenttunnuste x_i funktsioon.

Tänu seosefunktsioonidele ei pruugi sõltuva muutuja jaotus olla normaaljaotus, see võib olla ka Poissoni, gamma, pöördnormaalse jaotusega või isegi nominaalne (kaheväärtuseline, mitmeväärtuseline või järjestatav nominaalne). Prognoositava tunnuse jaotustüübi arvestamine on vajalik selleks, et prognoos ei omandaks loogiliselt võimatuid väärtusi. Näiteks üksikobjektide loendi tulemus ei saa olla negatiivne ja nominaalse tunnuse väärtus peab piirduma klassikoodide hulgaga (planeeritud laste arv on positiivne täisarv, puu on kas kask, kuusk, mänd või muust liigist).

Üldistatud lineaarne mudel lubab mitmeid funktsioontunnuse jaotusi ja vaid ühte funktsioontunnust

Sagedamini kasutatavad seosefunktsioonid on vastavalt funktsioontunnuse jaotustüübile järgnevalt välja toodud.

Identity link: **samasusseose funktsioon** ehk teisendamisest loobumine, eeldades vigade normaaljaotust.

$$f(x) = x \qquad [3\text{-}10]$$

Log link: **log-seose funktsioon** kasutatakse Poissoni jaotusega loendusandmete puhul.

$$f(x) = \log(x) \qquad [3\text{-}11]$$

Logit link: **logit-seose funktsioon** kirjeldab ühesuunalist tendentsi binaarsetes andmetes, kus mõlema variandi esinemise tõenäosus on vahemikus 0…1. **Logit** on logaritm esinemise ja puudumise vahekorrast.

$$f(x) = \log\left(\frac{p(x)}{1-p(x)}\right) \qquad [3\text{-}12]$$

Logitist saab tõenäosused logit-teisenduse pöördteisendusega.

$$p(x) = \frac{e^{\text{logit}(p(x))}}{1+e^{\text{logit}(p(x))}} \qquad [3\text{-}13]$$

Probit link: **probit-seose funktsioon** kasutab standardiseeritud kumulatiivse normaaljaotuse pöördväärtust (*invnorm*).

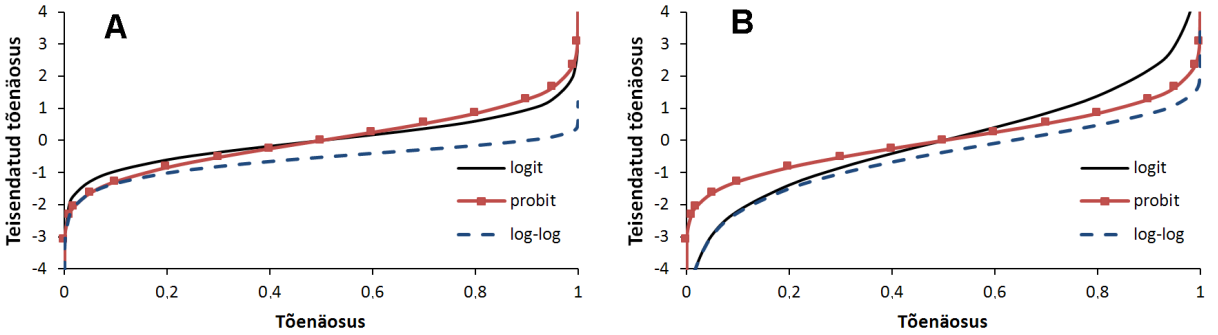
$$f(x) = invnorm(x) \qquad [3\text{-}14]$$

Seega probitregressiooni mudel on

$$y = NP\left(b_0 + \sum_i b_i x_i\right), \qquad [3\text{-}15]$$

kus *NP* on standardiseeritud kumulatiivse normaaljaotuse kõvera alusele pindalale vastav tõenäosus.

Probit-teisendus annab enam-vähem sama tulemuse kui kümnendlogaritmidaga logit-teisendus ja komplementaarne log-log teisendus (joonis 3-3A). Kuna logit-mudel on veidi lihtsam ja kergemini interpreteeritav, siis eelistatakse logit-teisendust.



Joonis 3-3. Tõenäosuse teisendamine logit, probit ja komplementaarse log-log funktsiooniga kasutades logit ja log-log teisenduses kümnendlogaritme (A) ja kasutades naturaallogaritme (B).

Complementary log-log link: **komplementaarse log-log** seose funktsiooni puhul läheneb tõenäosus ühele järsemalt kui nullile (logit ja probit teisendused on nullväärtuse suhtes sümmeetrilised.

$$f(x) = \log(-\log(1-x)) \qquad [3\text{-}16]$$

Väikeste tõenäosuste ja kümnendlogaritmide kasutamise korral annab komplementaarne log-log teisendus enam-vähem sama tulemuse kui logit ja probit teisendus, naturaallogaritmide kasutamisel ja väikeste väärtuste korral on probit-teisenduse tulemus logit- ja log-log teisenduse tulemusest mõnevõrra suurem (joonis 3-3).

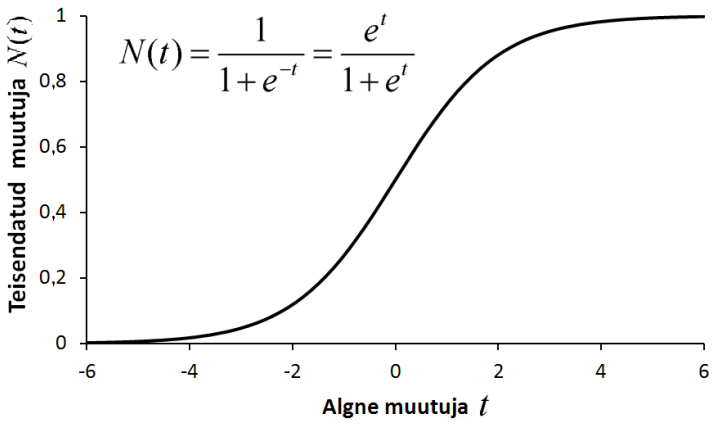
Generalized logit link: **üldistatud logit-seose funktsioon** on logit-teisenduse multinominaalne variant, kus mudelis on $c-1$ kategooriat.

$$f(x_1 | x_2, \ldots, x_c) = \log\left(\frac{x_1}{1 - x_1 - \ldots - x_c}\right) \qquad [3\text{-}17]$$

Logistiline funktsioon kirjeldab piiratud muutumisvahemikuga funktsioontunnuse sõltuvust pidevast argumenttunnusest. Kuna funktsioontunnuse muutumisvahemik on piiratud, siis on seose joon **sigmoidne** ehk S-kujuline kõver (joonis 3-4). Lihtne logistiline funktsioon avaldub võrrandina

$$N(t) = \frac{1}{1+e^{-t}} = \frac{e^t}{1+e^t}, \qquad [3\text{-}18]$$

kus *N(t)* on funktsioontunnuse väärtus argumenttunnuse väärtuse *t* juures, näiteks populatsiooni arvukus ajahetkel *t*. Eeldatakse argumenttunnuse skaala pidevust ja tsentreeritust nullkoha juurde.



Joonis 3-4. Logistiline kõver.

Logistilisele kõverale andis 1845. aastal nime Pierre-François Verhulst (1838), kes avaldas populatsiooni juurdekasvu võrrandi alltoodud kujul.

$$\frac{dp}{dt} = mp - \varphi(p), \qquad [3\text{-}19]$$

kus: p on populatsiooni suurus (isendite arv), dp/dt on populatsiooni juurdekasv ajaühikus dt, m on juurdekasvu kiirust määrav kordaja, $\varphi(p)$ tähistab juurdekasvu kiiruse sõltuvust populatsiooni suurusest.

Verhulst (1838) näitas ka, et populatsioonid ei suurene piiramatult – populatsiooni suurusel on ülempiir. Populatsiooniökoloogias kasutatakse Verhulsti võrrandit enamasti kujul

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right), \qquad [3\text{-}20]$$

kus: N on populatsiooni suurus (isendite arv), dN/dt on populatsiooni juurdekasv ajaühikus dt, K on keskkonna kandevõime, r on juurdekasvu kiirust määrav kordaja. Parameetri r väärtusest sõltub populatsioonidünaamika tüüp – liik on kas r-strateeg või K-strateeg.

Logit on logaritm esinemise ja puudumise vahekorrast; logistiline funktsioon on logit-teisenduse pöördteisendus

Verhulsti võrrand on universaalne mudel negatiivse tagasisidega protsessidele, mille puhul ühe variandi (liigi poolt asustatud ala) suurenemine on võrdeline vastupidise variandi (asustamata ala) suurusega.

Loodusuuringutes on sageli ülesandeks hinnata mingi liigi või muu nähtuse esinemise või puudumise tõenäosust. Kuna tõenäosus ei saa olla väiksem kui 0 ja suurem kui 1, siis lineaarne regressioonimudel selliste ülesannete lahendamiseks ei sobi, küll aga sobib logistiline regressioon. **Logistiline regressioon** ehk logit-regressioon on logit lingiga üldistatud lineaarne mudel, mis kirjeldab kaheväärtuselise muutuja variantide esinemistõenäosust. Logit-regressiooniga modelleeritav kaheväärtuseline muutuja võib olla mingi bioloogilise liigi või muu loodusnähtuse (maalihke, äikese, metsatulekahju) esinemise/puudumise tõenäosus. Otsestes vaatlusandmetes on esinemise ja puudumise juhud. Logistilist regressiooni saab kasutada ka igasuguste osatähtsuste, olgu need esitatud protsentides või mõne muu suhtarvu kujul, modelleerimiseks. Osakaaluna esitatud funktsioontunnus (F) tuleb logistilise mudeli jaoks teisendada kujule $log[F/(1–F)]$.

Logistilist mudelit saab kasutada ka multinominaalse tunnuse kategooriate tõenäosuse hindamiseks. Kui funktsioontunnus koosneb *M* võimalikust üksteist välistavast kategooriast, tuleb moodustada *M–1* logistilist mudelit. Järjestatavate kategooriate tõenäosuste modelleerimisel arvutatakse kumulatiivne tõenäosus, et lahend kuulub ühte või teise kategooriasse. Järjestatava muutuja logitmudeli näide on toiduobjekti liigi (multinominaalne muutuja) tõenäosuse sõltuvus alligaatori suurusest (pidev muutuja) (Agresti 1996).

Logistilist regressiooni kasutatakse eelkõige pidevate argumenttunnustega. Kui argumenttunnused ei ole pidevad ja eesmärgiks on seoste leidmine, mitte prognoosimudeli koostamine, võib esinemise ja puudumise andmete log-lineaarne analüüs sobivam olla.

Üldistatud lineaarsed mudelid pakuvad vahendeid ka loodusuuringutes sageli ette tulevate rohkete nullidega vaatlusandmete modelleerimiseks. Kui nullväärtused on sagedad funktsioontunnuse vaatlustes, siis võib olla mõistlikum analüüsida nähtuse esinemist/puudumist või siis kasutada log-teisendusega mudelit. Rohkete nullidega argumenttunnuse puhul saab kõigile väärtustele liita mingi väikese arvu ja seejärel väärtused logaritmida või siis kasutada tunnust kaheväärtuselisena.

Üldistatud lineaarsete mudelite parameetrite leidmist raskendab mudelivea teadmata jaotus. Parameetrid leitakse iteratiivselt ja hinnangu viga määratakse suurima tõepära (ptk 1.2 ja 3.6.1) meetodil. Selleks on mitmeid iteratiivseid meetodeid, millest efektiivseimad on Newton-Raphsoni meetod ja iteratiivselt ümberkaalutud vähimruutude meetod (*Fisher-scoring*). Statistilise olulisuse testimiseks kasutatakse Waldi statistikut, tõepärasuhet või skoori statistikut. Neist kõige aeganõudvam, kuid asümptootiliselt kõige efektiivsem on tõepärasuhe ja vastupidi, arvutuslikult kõige kiirem, kuid kõige vähem usaldatav on skoori statistik. Mudeli sobivuse võrdlemise vahendeid käsitletakse mudeli hindamise allpeatükis (ptk 3.6). Üldistatud lineaarse mudeli jääkide analüüsi kaks põhitüüpi on Pearsoni jäägid (erinevus vaadeldud ja prognoositud väärtuste vahel) ja hälbejäägid (*deviance residuals*), mis põhinevad vaatluste panusel log-tõepärasse.

3.4.1.4. Üldistatud aditiivsed mudelid

Eelnevalt käsitletud klassikalised vähimruutude regressioonanalüüsi meetodid eeldavad mõõtmisviga vaid sõltuval tunnusel, sõltuva tunnuse mõõtmisvea konstantsust, erindite suuremat tähtsust võrreldes tüüpiliste vaatlustega ja mitmese regressiooni puhul argumenttunnuste omavahelist sõltumatust. Peale selle nõuab regressioonanalüüs kas seose kuju eelnevat otsustamist või algandmete normaliseerimist. **Üldistatud aditiivseid mudeleid** (*generalized additive models – GAM*) loetakse mitteparameetriliste meetodite hulka ehk need ei sea eeldusi seosefunktsioonidele, küll aga nõuavad siiski faktorite mõjude aditiivsust. Lisaks sellele tuleb ka mitteparameetriliste meetodite puhul moodustada mingi mudel. Ülevaate üldistatud aditiivsetest mudelitest võib leida Trevor Hastie ja Robert Tibshirani (1990) raamatust.

Aditiivsete mudelite ainus eeldus on seletavate tunnuste mõjude liidetavus

Aditiivsed mudelid on lai klass paindlikke regressioonimudeleid, millesse kuuluvaid üksikmeetodeid on loetletud ja lühidalt iseloomustatud allpool. Tasub meeles pidada, et piirangute vähenemisega tõuseb tõenäosus kirjeldada uuritava tunnuse juhuslikku varieeruvust kui seletavate tunnuste põhjuslikku mõju ehk võib kasvada mudeli ülesobitumise oht.

Kui ka aditiivsuse piirang ahistab, võib regressioonimudeli asemel kasutada vaatluste klassifitseerimist ja väärtuste hindamist klassikuuluvuse alusel (klassifikatsiooni ja regressioonipuud), kõige sarnasemate näidiste kasutamist või intellektitehnika valda kuuluvaid meetodeid.

Lineaarsed mitteparameetrilised meetodid

Kaalutud vähimruutude regressioonis (*weighted least squares regression*) kaalutakse vaatlused funktsioontunnuse dispersiooni hinnanguga iga vaatluse puhul.

L-hinnangu (*L-estimator*) regressiooni puhul kasutatakse tõesuskriteeriumina funktsioontunnuse lineaarhälvete summat.

M-hinnangu (*M-estimator*) regressiooni puhul sobitub regressioonijoon nii, et mõlemal pool joont on ühe palju vaatlusi. Mitmevariandilisuse vältimiseks lisatakse erinevaid täiendavaid piiranguid. LMS-regressiooni (*least median squares regression*) puhul minimiseeritakse hälbe ruutude mediaani. BI-regressiooni (*bounded influence*) puhul omistatakse erinditele väike mõjukaal.

Kantregressioon (*ridge regression*) on omavahel tugevasti korreleerunud argumenttunnuste jaoks kohandatud meetod, mis aga paraku annab regressioonikordajate nihkega hinnangud. Meetod võimaldab paljude omavahel korreleerunud faktorite hulgast suurema kindlusega leida neid, mis on olulisemad.

Totaalvähimruutude regressioonid (*total least squares ehk errors in variables regressions*) arvestavad hälbeid ka argumenttunnuste väärtustes, mitte ainult funktsioontunnuse hinnangutes.

Peakomponentregressioon (*principal component regression*) minimeerib vaatluste nii x-telje kui ka y-telje suunas olevate hälvete ruutude summat.

Partsiaalvähimruutude regressioon (*partial least squares regression – PLS*) kasutab abstraktseid faktoreid, mis kirjeldaksid kõige paremini seoseid kahe muutujategrupi vahel. PLS sobib näiteks mitmeliigilise koosluse liigilise koosseisu prognoosimiseks paljude keskkonnafaktorite järgi. PLS regressiooni saab kasutada ka sobivate argumenttunnuste valimiseks ja erindite leidmiseks enne klassikalist regressiooni. Meetod on kasulik eelkõige juhul, kui mudelisse soovitakse lülitada suhteliselt palju faktoreid, kuid mitte kõiki algselt mõõdetud tunnuseid.

Kahetine vähimruutude regressioon (*bivariate least squares*) arvestab lisaks dispersioonile mõlema telje suunas ka üksikvaatluse usaldatavuse hinnangut.

Kaalutud jääkide mediaanregressioon (*median sum of weighted residuals regression*) on välja töötatud Tartu Ülikoolis Robert Szava-Kovatsi poolt, selle puhul minimeeritakse vaatluste eukleidilis-te kauguste summat regressioonijooneni. Minimeerimiseks ei kasutata seejuures mitte kõiki vaatlusi, vaid mingit etteantud osa nendest ning kaugusi kaalutakse vaatluste mõõtmistäpsusega (Szava-Kovats 2001).

Mittelineaarsed mudelid

Mittelineaarsed mudelid jagatakse vormilt ehk **kujult mittelineaarseteks** ja **seesmiselt** (*intrinsically*) **mittelineaarseteks**. Kujult mittelineaarsed mudelid on kas argumenttunnuse või funktsioontunnuse või mõlema teisendustega lineaarseteks muudetavad. Sisuliselt mittelineaarseid mudeleid ei õnnestu teisendustega lineariseerida, sest mudeli vead ei allu samale teisendusele kui tunnused.

Seos argumenttunnuse ja funktsioontunnuse vahel ei ole loodusandmete puhul sageli lihtsa parameetrilise funktsiooniga kirjeldatav. Sellisel juhul kasutatakse pideva funktsiooni kirjeldamiseks libiseva keskmise abil saadud seosekõveraid, lokaalselt kaalutud regressiooni või lokaalselt kaalutud tõenäosusjaotusi. Silumisega saadud funktsioone saab kasutada nii prognoosimudelina (Brown 1994) kui ka seose kirjeldamiseks (Austin ja Mayers 1996, Remm 1987). Lokaalseid meetodeid saab kasutada nii ühe kui ka mitme argumenttunnusega.

Ühe tunnuse **splain-regressiooni** ehk lokaalselt sobituva regressiooni puhul kasutatakse lokaalses sobitamises enamasti kuupfunktsiooni, aga ka muud funktsioonid on võimalikud. Splain-mudeli puudus on tundlikkus sõlmekohtade paiknemise suhtes (Wooda ja Augustin 2002). Splain-regressiooni mitmetunnuseline variant **MARS** (*multivariate adaptive regression splines*) (Friedman 1991) sobitab lokaalselt mitmemõõtmelisi seosejooni.

Libiseva keskmise ehk kernelregressiooni puhul tuleb kasutajal otsustada lokaalse akna ehk kerneli ulatus ja akna keskkoha kaugusest sõltuv kaalufunktsioon.

LOWESS ehk lokaalse kaalutud regressiooniga silumine (Cleveland 1979, Cleveland ja McGill 1985, Cleveland ja Devlin 1988, Trexler ja Travis 1993). Lokaalne regressioon tähendab, et funktsioontunnuse prognoosimiseks mingis kohas x kasutatakse vaid selle koha lähedal olevaid vaatlusi, mis muudab meetodi väga paindlikuks. Otsustada tuleb vaid mitu lähimat vaatlust moodustavad naabruse ning leida prognoositavas kohas vahemaa neist kaugeimani (d). Järgnevalt omistatakse igale naabervaatlusele (x_k) akna keskpunkti (x_i) kaugusest sõltuv kaal

$$w_i(x_k) = \left(1 - \left|\frac{x_i - x_k}{d}\right|^3\right)^3. \qquad [3\text{-}21]$$

Kaugemal kui d omistatakse kaalule väärtus null. Kuupfunktsiooni kasutatakse lihtsalt seetõttu, et see annab parajalt laia maksimumi ja kenasti sujuva servaga kaalufunktsiooni.

Seejärel sobitatakse aknas olevatele vaatlustele kaalutud vähimruutude meetodil lineaarne regressioon. Nii korratakse iga argumenttunnuse x väärtuse korral. Esmane regressioonimudel on seega käes. See mudel ei ole robustne, vaid on erindite suhtes tundlik. Robustse hinnangu saamiseks arvutatakse esmase hinnangu regressioonijäägid (r). Regressioonijääkidest ja nende absoluutväärtuse mediaanist arvutatakse robustsuskaalud (rw)

$$rw_i = \left(1 - \left(\frac{r_i}{6m}\right)^2\right)^2, \qquad [3\text{-}22]$$

kaugemal kui d omistatakse kaalule väärtus null.

Kaalud rw_i annavad suurema tähtsuse vaatlustele, mis hälbivad regressioonijoonest vähem ja erindite mõju vähendatakse. Regressioonijoonest rohkem kui kuue mediaani võrra hälbivate vaatluste kaal muutub nullilähedaseks. Viimase sammuna arvutatakse lokaalne regressioon uuesti, nüüd juba robustsuskaaludega.

ACE (*Alternating Conditional Expectations*) (Breiman ja Friedman 1985) on iteratiivne optimaalse transformatsiooni otsimise meetod, mis maksimeerib korrelatsiooni teisendatud sõltuva muutuja ja teisendatud sõltuvate muutujate summa vahel. ACE arvutamisel silutakse korduvalt liikuva aknaga muutujat Y piki muutuja X väärtusi ja saadakse Y väärtuse hinnang $f(X)$, seejärel silutakse hinnangut piki muutujat Y ning asendatakse esialgsed väärtused silutud prognoosiga. Protsessi korratakse, kuni prognoosi ja vaatluste ruuthälvete keskväärtus stabiliseerub. ACE eeldab, et tunnused on aditiivsed ja omavahel sõltumatud.

Kombineeritud meetodid

Liialdatud nullidega andmetes (*zero inflated data*) esineb loendustulemuste nullväärtusi oluliselt rohkem kui Poissoni jaotuse kohaselt peaks olema. **Liialdatud nullide regressiooni** (*zero inflated regression*) puhul modelleeritakse nähtuse puudumine logistilise regressiooniga, nähtuse hulk tema esinemise korral aga kas kärbitud Poissoni jaotusega või negatiivse binoomjaotusega. Nähtuse hulka (liigi arvukust) käsitletakse seejuures tinglikuna, see tähendab liigi esinemise eeldusel. Tulemuseks on kaks eraldi mudelit – mudel liigi esinemise ja puudumise kohta ning mudel liigi hulga kohta liigi esinemise korral (Lambert 1992). **Kärbitud jaotuse** (*truncated distribution*) võimalike väärtuste hulk on piiratud. Piirang võib ette anda lubatud vahemiku (näiteks ema vanus järglase sünni hetkel) või välistada mingid väärtused (näiteks nullväärtused Poissoni jaotusest).

Struktuurse statistilise modelleerimise (*structural equation modelling – SEM*) puhul üritatakse statistilist seost formuleerida võrrandite võrgustiku abil. SEM-mudel sisaldab tavaliselt latentseid vahe-muutujaid, mida otseselt mõõta ei saa ja mis ei kuulu ka tulemuste hulka.

3.4.2. Otsuste puu

Otsuste puu (*decision tree*) on argumenttunnustega seotud kriteeriumite hierarhiline kogu funktsioontunnuse prognoosimiseks. Otsuse puu otsuseid tuleb langetada kriteeriumite alusel, mis jagunevad faktoriteks ja piiranguteks. **Piirang** (*constraint*) välistab mingi valikuvariandi, faktor muudab alternatiivide tõenäosusi. Otsuseid saab langetada numbrilise argumenttunnuse mistahes "suurem kui" või "väiksem kui" tingimuse järgi. Nominaalse tunnuse korral sobivad jagamise kriteeriumiks suvalised klassikombinatsioonid. Otsuste puu koostatakse nii, et puu harude homogeensus oleks võimalikult suur. Homogeensuse mõõtudena saab nominaalsete tunnuste puhul kasutada näiteks Shannoni entroopia indeksit (ptk 2.1.2) või Gini indeksit (ptk 2.1.4), pideva tunnuse puhul muutuja grupisisest ja gruppidevahelist hajuvust ning klassifitseerimisega ärakirjeldatud hajuvuse osa.

Otsuste puu koosneb hierarhiliselt kasutatavatest piirangutest ja faktoritest

Otsuste puu moodustamise algoritmid kuuluvad osaliselt intellektitehnika valdkonda. Masinõppe meetodid võimaldavad lisaks parimale struktuurile otsida ka puu sobivaimat üldistustaset. Kui sama struktuuriga puu suudab juhuslikest andmetest sama tugevaid seoseid leida sagedamini kui uurija poolt aktsepteeritud riskimäär, siis on puu üle sobitatud. Ülesobitunud mudelid annavad õpetusandmetes täpsemaid prognoose kui sõltumatutes kontrollandmetes. Ülesobitumine võib tuleneda puu liigsest detailsusest ning tuleneda (juhuslikest) vigadest ja ebaolulistest tunnustest õpetusandmetes. Otsuste puud ei pruugi anda parimat klassifikatsiooniskeemi ka teises mõttes: nad kalduvad eelistama neid klassifikaatoreid, mida õnnestub hierarhilises skeemis paremini kasutada.

Otsuste puud on laialdaselt kasutusel suuremahuliste kaugseireandmete töötluses, kuna nende kasutus on arvutuslikult kiire. Otsuste puud ei sea eeldusi andmete tüübi ja statistilise jaotuse osas, kuid võimaldavad samaaegselt kasutada erinevas mõõtkavas lähteandmeid ning tarkvara on piisavalt saadaval.

Otsuste puud jagunevad vastavalt funktsioontunnuse tüübile klassifikatsiooni- ja regressioonipuudeks. Klassifikatsioonipuud kasutatakse siis, kui klassifitseeritakse nominaalset tunnust ja regressioonipuud pideva muutuja väärtuste prognoosimiseks.

Klassifikatsioonipuu abil otsustatakse klassikuuluvust, regressioonipuu abil reaalarvulist väärtust

3.4.2.1. Klassifikatsioonipuu

Klassifikatsioonipuu (*classification tree*) on meetod diskreetse olemi klassikuuluvuse määramiseks tema kvalitatiivsete ja kvantitatiivsete omaduste järgi. Statistika teoorias on see suhteliselt uus (Breiman *et al.* 1984, Quinlan 1986), mitmetel muudel uurimisaladel aga ammu kasutatusel olev meetod (bioloogiline süsteem ja määramistabelid K. Linnést alates, otsustuste teooria, meditsiiniline diagnostika). Puuks sobib nimetada eelkõige järk-järgulist klassifitseerimiseeskirja, mis jagab vaatlusi korduvalt osadeks. Klassifikatsioonipuu nõuab kriteeriumite kasutamist kindlas järjekorras. Klassifikatsioonipuude metoodika on paindlik, kuna otsustamise kriteeriumid võivad olla väga erinevad. Otsustuste juures saab kasutada tunnuste kombinatsioone ja tagasipöörduvaid suunamisi. Kuna sama faktor võib otsuste langetamisel osaleda mitmel tasemel ja kombineeritult teiste faktoritega, siis on võimalik arvesse võtta ka keerukaid seoseid. Kui otsuse aluseks on vaid üks faktor, siis on otsustuspiir vastava teljega risti olev sirge.

Klassifikatsioonipuu nõuab kriteeriumite kasutamist kindlas järjekorras

Klassifikatsioonipuude kasutamisel võib olla raskusi sobivaima üldistustaseme otsustamisega. Nagu kõigi mudelite puhul, võib ka klassifikatsioonipuude puhul väga detailne variant vaid näida nähtuste hea seletajana. Detailsete mudelite vastavus väljaspool lähteandmeid on enamasti halvem kui paraja üldistustasemega mudelitel. Sobivat üldistuse taset aitab leida ristkontroll ja nullhüpoteesile vastavate randomisatsioonidega võrdlemine. Üldistustaseme suurendamiseks saab klassifikatsioonipuud sobiva üldistatuse tasemini maha **pügada** (*pruning*).

Klassifikatsioonipuud erinevad diskriminantanalüüsist otsustuste hierarhiliste järjestatuse poolest ja võimaluse poolest kasutada igasugust tüüpi tunnuseid.

3.4.2.2. Regressioonipuu

Regressioonipuu (*regression tree*) eeldab, et prognoositav muutuja on pidev, argumenttunnused võivad olla mistahes tüüpi. Kuna enamasti on regressioonipuu hierarhiliselt jagunev, siis kasutatakse selle tähistamiseks ka terminit **hargneva puu regressioon** (*hierarchical tree-based regression – HTBR*). Regressioonipuud moodustavate kriteeriumite otsimisel kasutatakse argumenttunnuste klasteranalüüsi või muul viisil klassifitseerimist. Erinevalt klasteranalüüsist on regressioonipuu mudel, mida saab kasutada funktsioontunnuse väärtuse prognoosimiseks uutel vaatlustel. Erinevalt klassifikatsioonipuust ja sarnaselt klasteranalüüsile annab regressioonipuu lisaks objektide klassikuuluvusele ka iga objekti klassi sees paiknemist iseloomustavaid ja klasside homogeensuse parameetreid. Näiteks, kas objekt on klassikeskme lähedal (tüüpiline klassi esindaja) või pigem mingite klasside vahepealne vorm ning keskmine ruuthälve grupi keskväärtusest või keskmine absoluuthälve grupi mediaanist klassi seesmise ühetaolisuse mõõduna.

Regressioonipuu eelised regressioonanalüüsi ja vastavusanalüüsi ees on:

- robustsus, see tähendab, tunnuste tüüpidele ja väärtuste jaotusele ei seata piiranguid,
- regressioonipuu sõltumatus funktsioontunnuse monotoonsetest teisendustest,
- regressioonipuu võimaldab paindlikumalt arvestada argumenttunnuste iseloomulikke kombinatsioone,
- regressioonipuu hargnemiskriteeriumiks ei pruugi olla seletavate tunnuste kriitilised väärtused, vaid ka väärtuste kombinatsioonid või isegi lokaalsed regressioonimudelid.

Regressioonipuud võivad olla ühemõõtmelised (üks funktsioontunnus) või mitmemõõtmelised (mitu funktsioontunnust). **Ühemõõtmelise regressioonipuu** moodustamisel jagatakse vaatlusi argumenttunnuste järgi gruppidesse nii, et grupid oleksid seesmiselt võimalikult ühetaolised ja eristuksid omavahel võimalikult hästi. Vaatlusi jagatakse senikaua, kuni gruppide arv või mingi gruppide homogeensuskriteerium vastab oodatavale. Parimat prognoosi andev puu (*the best predictive tree*) on see, mis annab argumenttunnuste järgi kõige täpsema funktsioontunnuse prognoosi. Sobiva klassifikatsioonitaseme leidmiseks kasutatakse ristkontrolli ehk ristvalideerimist.

Mitmemõõtmelises regressioonipuus on mitu prognoositavat muutujat ja klastri homogeensust saab mõõta keskmiste ruutvahemaade kaudu tunnusruumis või mingi sarnasuskordajaga. Lisaks funktsioontunnuse hinnangule saab mitmemõõtmelisest regressioonipuust teha järeldusi funktsioontunnuste (näiteks elupaigatingimuste järgi prognoositavate liikide) eripära kohta. De'ath (2002) nimetab liikide kui funktsioontunnuste näitel kolme tüüpi järeldusi:

- saab eristada, milline liik mõjutab klassideks jagunemist rohkem, milline vähem,
- saab leida gruppidele iseloomulikke liike,
- saab liikide elupaiganõudlusi võrrelda.

3.4.3. Ordinatsioonid

Reeglina loetakse ordinatsioonimeetodeid kirjeldava statistika hulka kuuluvaks (ptk 2.4), kuid abstraktsioonidena võib ordinatsioonimeetodeid pidada mudeliteks. Ordinatsiooni võib käsitleda mudelina ka lähtuvalt sellest, et tunnuste ja vaatluste koordinaate ordinatsioonitelgedel saab kasutada prognoosimudelites nii seletavate kui ka prognoositavate tunnustena.

3.4.4. Markovi ahel

Markovi ahel (*Markov chain*) on vahend objekti järgneva seisundi tuletamiseks eelnevast. Protsessi modelleeritakse juhuslikuna ja ühesuunalisena. Eeldatakse katseseeriat, kus iga katse tulemusel muutub objekti **seisund** (*state*). Võimalike seisundite (sündmuste) hulk on lõplik ning iga seisundi tõenäosus sõltub talle eelnenud seisundist (aga mitte varasematest) ja **üleminekutõenäosustest** (*transition probabilities*) Kui üleminekutõenäosused on teada, saab arvutada kõigi võimalike seisundite tõenäosusi kogu seeria ulatuses. Omavahel üleminekutõenäosustega seotud seisundite jada ongi Markovi ahel. Markovi ahela ülemineku-tõenäosusi esitatakse maatriksi kujul, mille iga rida on tõenäosusvektoriks ja selle komponentide summa on 1. Klassikalisel juhul käsitletakse Markovi mudelis aega diskreetsena ja üleminekutõenäosusi ajas konstantsetena.

Varjatud Markovi mudeli (*hidden Markov model – HMM*) ehk varjatud Markovi ahela puhul sisaldab süsteemi iga seisund sümbolite hulka. Seisundid ise on varjatud, vaadelda saab sümbolite esinemissagedust. Varjatud Markovi mudeli topoloogia on defineeritud varjatud seisundite arvu, sümbolite arvu ja nulliga võrdsustatud üleminekutõenäosuste ja sümboli väljastamistõenäosuste arvuga. Seisundeid ühendavad üleminekutõenäosused leitakse vaatlusandmetest. Varjatud Markovi mudeleid on kasutatud kõnetuvastuses, kujutise analüüsis (Aas *et al*. 1999) ja bioinformaatikas.

3.4.5. Intellektitehnika

Intellektitehnika (*Artificial Intelligence – AI*) eesmärgiks on luua inimese intellekti jäljendavaid tehisintellektisüsteeme. Intellektitehnika hulka kuuluvad nii füüsilised süsteemid kui ka arvutusmeetodid, milles kasutatakse automatiseeritud iteratiivset optimeerimist ehk tehisõpet. Tehisintellektisüsteem on inglise matemaatiku A.M. Turingi järgi masin, mis vastab küsimustele nii, et pole võimalik vahet teha, kas vastajaks on masin või inimene. Intelligentsuse oluline tunnusjoon on õppimise ja muutuvates oludes kohanemise võime.

Intellekt on õppimisvõime, intellektitehnika meetodid kasutavad tehisõpet

Tehisintellekti loomise suure optimismi periood oli 1960ndatel, mil USA kaitseministeerium selle alaseid uuringuid lahkelt finantseeris. Tähelepanuväärne on, et inglisekeelses maailmas nimetatakse valitsustele teavet koguvaid ja analüüsivaid ametkondi *Intelligence Agency* ja *Intelligence Service*. 1970ndate teisel poolel kärbiti AI projektide finantseerimist (*AI winter*). 1980ndatel jõudsid ekspertsüsteemid ärilise eduni. 1990ndatest aastatest alates on AI meetodite rakendused olnud edukad mitmetel eraldiseisvatel tehnika aladel ja andmetöötluses.

Ühtse AI teooriani ja AI üldprintsiipideni ei ole siiski senini jõutud ning andmeanalüüsi meetodite puhul on AI meetodeid muudest iteratiivsetest meetoditest raske eristada. Valdavat osa statistilise andmeanalüüsi AI meetodeid loetakse samal ajal kuuluvaks ka andmekaevandamise meetodite hulka (ptk 3.1). Intellektitehnika hulka loetakse klassifikatsiooni- ja regressioonipuud, tugivektormasinad ja sarnasusele tuginevad meetodid juhul, kui nendes kasutatakse kaalude või muude parameetrite iteratiivset sobitamist. Alljärgnevalt on need metoodikavaldkonnad esitatud omaette alapeatükkides.

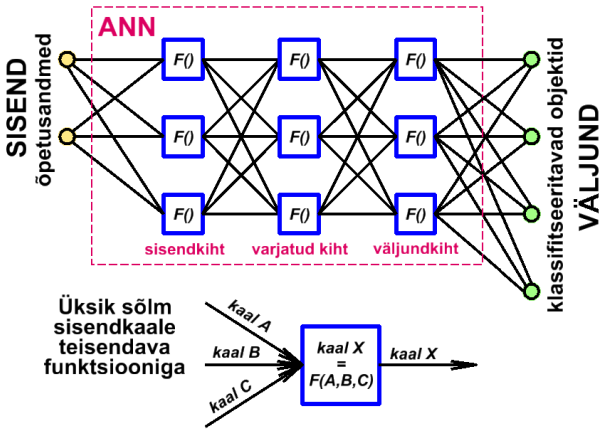
Eesti keeles tehakse vahet tehisintellektil ja intellektitehnikal, mille hulka kuuluvad igasugused tehisintellekti loomise vahendid, seega ka tehisõpet kasutavad arvutusmeetodid.

3.4.5.1. Tehisnärvivõrgud

Tehisnärvivõrkude (*artificial neural networks – ANN*) idee pärineb kujutlusest, kuidas töötab bioloogiliste olendite aju. Aju kõige tähelepanuväärsem omadus on õppimisvõime ja võime vähese lähteinfo järgi üldistusi teha. Esimesed katsed järele aimata bioloogiliste närvivõrkude talitlust pärinevad II Maailmasõja perioodist. Ühe esimese publikatsioonina viidatakse F. Rosenblatti (1958) kirjutisele. Laialdasemalt hakati tehisnärvivõrkudega tegelema 1980ndatel aastatel, osaliselt seoses arvutamisvõimaluste hoogsa kasvuga. ANN-lähenemise olulisim erinevus traditsioonilistest statistilistest meetoditest on heuristilisus – ANN ei lähtu mingist olemasolevast mudelist ega jaotustüübist, ANN nõuab vähem statistika alaseid eelteadmisi ja samas võimaldab kirjeldada keerukaid statistilisi seoseid ja arvestada kõiki argumenttunnuste koosmõjusid. ANN metoodikat saab kasutada nii objekti grupikuuluvuse kui ka numbrilise muutuja väärtuse prognoosimiseks. ANN metoodika ei pruugi anda ühest vastust oodatava kategoorilise väljundi kohta, selle asemel on kuuluvusfunktsioon, mis näitab väljundi variantide tõenäosust.

ANN koosneb omavahel seotud **sõlmedest** (*nodes*), mis on aju neuronite analoogid ja mis moodustavad ANN kihte (joonis 3-5). Osa kihte võivad olla varjatud, s.t neil puudub otsene side sisendiga. ANN moodustamise järel hakatakse seda võrgustikku õpetama. Tehissüsteemi õpetamine ehk treenimine on regressioonimudeli parameetrite sobitamise analoog ja toimub sisendandmete ja neile vastavate väljundandmete ehk õpetuskogumi järgi. Treenimise käigus püütakse minimiseerida väljundi viga. Algselt omistatakse ANN sõlmi ühendavatele kaaludele (regressioonikordajate

analoogid) juhuslikud väärtused. Võrgustiku treenimine toimub kaalude iteratiivse täpsustamise teel. Uute andmete laekumisel saab ANNi aina edasi õpetada. Seega võib tehisnärvivõrk olla aina uuenev teadmiste baas. See teadmiste baas on aga niinimetatud must kast, mis annab küll prognoose, aga ei esita ei statistilisi seaduspärasusi ega tüüpilisi näidiseid.



Joonis 3-5. Ühe varjatud kihiga ja vaid edasisidega tehisnärvivõrgu ülesehitus.

ANN annab prognoose, aga ei selgita statistilisi seaduspärasusi

Tehisnärvivõrkude puudusena mainitaksegi eelkõige nende sarnasust musta kastiga. ANN-mudel võib anda hea prognoosi, aga reeglina ei anna arusaama sellest, millised seosed uuritud andmetes esinevad ja milline on faktorite mõju vahekord. ANNi struktuuri moodustamine sõltub täiesti uurijast ja tema kogemusest. ANN oskab opereerida vaid õpetusandmete muutumispiirkonnas. Ka siis, kui ANN kasutab diskreetseid funktsioone ja püüab seoseid väga detailselt kirjeldada, on oht, et õpetuskogumis mitteesinenud lähteandmete puhul annab ANN ootamatuid ja vähe usaldusväärseid tulemusi. ANNis kasutatavate vea minimeerimise algoritmide jaoks on probleemiks prognoosivigade lokaalsed miinimumid, mille tõttu piisavalt hea lahend ei pruugi olla parim lahend. Keerukate ANNide puhul võib probleemiks olla ka üleõppimine, mille tõttu hakkab ANN seoseid leidma ka mürast. ANNi ja teiste tehisintellektisüsteemide ületreenimine on analoogne statistilise modelleerimise säästvusreegli rikkumisega – keerukam mudel võib küll paremini sobida õpetusandmetele, aga sugugi mitte kontrollandmetele.

Ülevaateartikli tehisnärvivõrkudest ja nende kasutamisest ökoloogilises modelleerimises on kirjutanud S. Lek ja J.F. Guégan (1999). ANN eelised ja puudused on ühes teises publikatsioonis kokku võetud järgmiselt (Jarvis ja Stewart 1996).

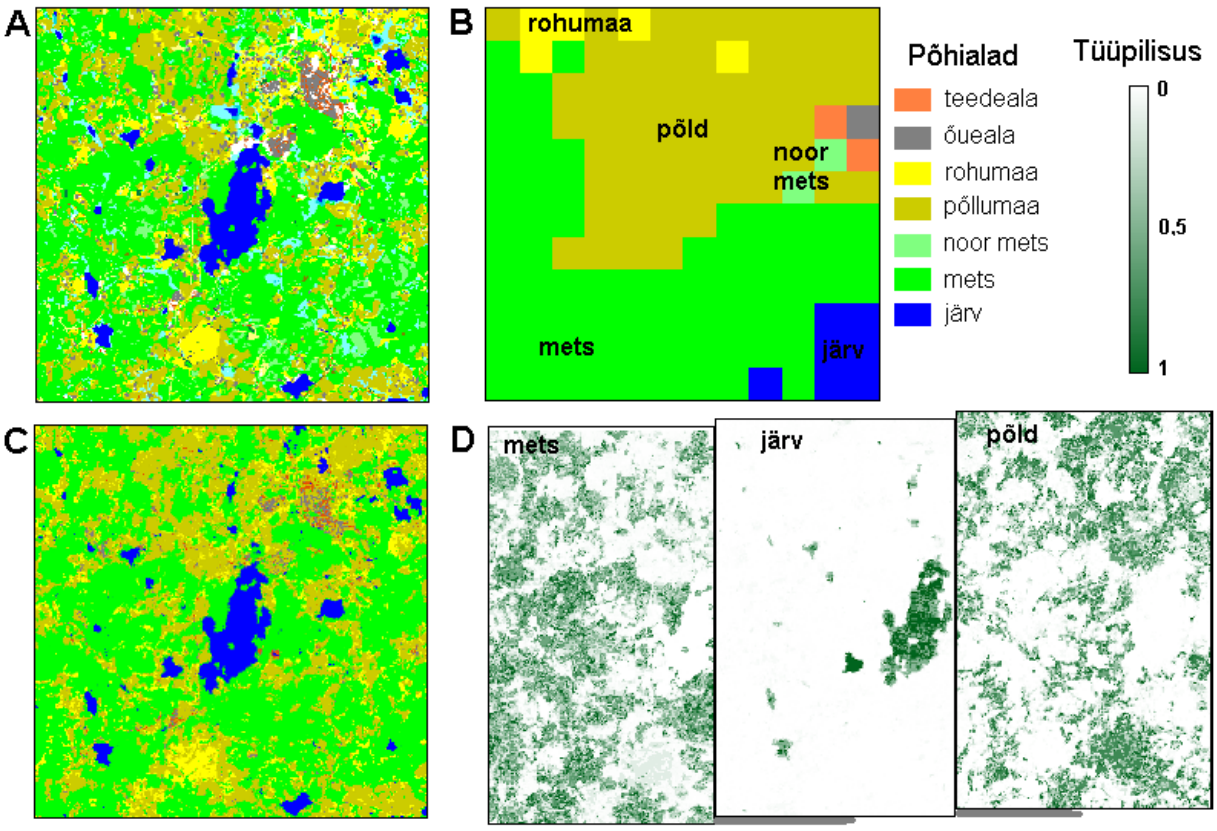
Eelised: mittelineaarsete seoste kirjeldamine, puuduvate andmete ja müra taluvus, nominaalsete ja numbriliste tunnuste kooskasutamise võimalus, piirangute puudumine andmete statistilise jaotuse osas, arvutuslikult efektiivne lahendite arvutamine pärast treenimist, võime õppida väikestest andmestikest, võime kohanduda muutuvate õpetusandmetega, võimaldab kasutada konteksti andmeid, võimaldab väljastada hägusaid tulemusi.

Tehisnärvivõrkude puuduseks on peetud aeglast õpet suurte andmestike puhul, raskusi õpetusandmetes puuduvate juhtumite klassifitseerimisel, suurt mälunõudlikkust keerukamate võrgustikutüüpide puhul, võrgustikutüübi valiku on subjektiivsust, takerdumist lokaalselt parimate lahendite juurde, tulemuste saamise läbipaistmatust, vähest kasutust laialtlevinud tarkvarapakettides ja vähest võimekust sisuliselt oluliste ja ebaoluliste tunnuste eristamisel. Enamik nendest puudustest kehtib kõigi keerukate mudelite puhul.

3.4.5.2. Kohoneni iseorganiseeruv tunnuskaart

Kohoneni iseorganiseeruv tunnuskaart ehk **Kohoneni kaart** (Kohonen 1982, 1984) (*Self-organizing feature map – SOFM, Self-Organizing Map – SOM, Kohonen map*) on ANN-tehnoloogia erijuhtum, mis klassifitseerib andmeid sarnasuse järgi ja aitab visualiseerida andmetes olevat korrapära. SOM on loodud eelkõige ordineerimiseks ja visualiseerimiseks, on lausa väidetud, et Kohoneni kaart astub edukalt traditsiooniliste ordinatsioonimeetodite asemele (Recknagel 2001), kuid kõige enam on seda kasutatud siiski klassifitseerimiseks. Erinevalt teistest ANN-meetoditest ei ürita SOM funktsioontunnuse väärtusi prognoosida, kuid SOM kaardi abil eristatud klasse saab rakendada väljaspool õpetusandmeid.

Klassikaline Kohoneni kaart on kahemõõtmeline maatriks, milles sarnased objektid paiknevad lähestikku ning seda võib liigitada järelvalveta (ehk õpetusandmeteta) klassifikatsiooni meetodiks või ordinatsioonimeetodiks, mis ei kasuta ette antud ordinatsioonitelgi (joonis 3-6). Nagu teisedki ordinatsioonimeetodid, vähendab SOM andmete dimensionaalsust – paljude tunnustega kirjeldatud vaatlusvektorid asendatakse väiksema arvu ordinatsioonitelgedega. Selles mõttes on SOM peakomponentanalüüsi mitteparameetriline variant.



Joonis 3-6. Kaardilehel 5434 olevad põhikaardi põhialad (A), põhialade ordinatsioon Landsat 5 TM kujutise kanalite kiirgusväärtuste ja NDVI indeksi järgi Kohoneni tunnuskaardina (B), sama kaardilehe põhialade kaart arvutatuna Kohoneni kaardi sõlmede kaalude ja samade kaugseireandmete põhjal (C) ning iga koha tüüpilisus metsa, järvede ja põllu suhtes (D). Märka, et Kohoneni tunnuskaart on abstraktses ordinatsiooniteljestikus, mitte ruumikoordinaatides. Kaardilehe keskel on Pühajärv.

Kohoneni kaart erineb tavalisest edasisidega tehisnärvivõrgust varjatud kihi puudumise poolest. Sisendkihi sõlmed on kaalude kaudu seotud otse väljundkihi sõlmedega. Lisaks tavapärasele ANN struktuurile on Kohoneni kaardi puhul ka väljundkihi sõlmed omavahel kaaludega seotud. Kohoneni kaardi iseõppimise käigus leitakse sellised seosekaalud, mille puhul lähestikku olevad väljundkihi sõlmed on omavahel võimalikult sarnased. Väljundkihi sõlmede topoloogilist asendit ei muudeta, muudetakse vaid nende sarnasust ehk kaugust tunnusruumis. Ka iga vaatlusvektori seotust iga väljundsõlmega näitavad kaalud. Väljundkihis saab vaadelda iga tunnuse ja vaatluse paiknemist telgede suhtes ja omavahelisi seoseid – lähestikku paiknevad üksused on sarnased.

SOM loob väljundkihi, milles lähestikku paiknevad sõlmed on sarnased

Kohoneni kaart ei sea mingeid eeldusi lähteandmetele, kuid väljundkihi ridade ja veergude arv ning dimensionaalsus tuleb kasutajal endal otsustada. Kui eesmärgiks on ordinatsioon, siis kasutatakse reeglina kahemõõtmelist ristkülikmaatriksit. Iseorganiseeruva kaardi puhul peab kasutaja otsustama parameetrid, sarnasuse mõõtmise viisi, võrgustiku tüübi ja sõlmede arvu. Väljundkaardil olev klaster ei pruugi olla vastavuses mingi *a priori* klastriga, sest SOM lähtub sisend- ja väljundkihi sõlmede sidumisel õpetusandmetest, mitte etteantud klassidest.

SOM on tehisõppe abil sobitatav sarnasusele tuginev klassifitseerimine ja ordineerimine, mille puhul klassid ja ordinatsiooniteljed ei ole ette antud

Kohoneni kaardi õppe eel omistatakse väljundvõrgustiku sõlmedele juhuslikud kaalud. Sisendvektorid standardiseeritakse samasse muutumisvahemikku nagu kaaludki. Õppe käigus leitakse ükshaaval igale sisendvektorile kõige sarnasem väljundsõlm ja seejärel kohandatakse selle väljundsõlme kaale ja vähemal määral väljundsõlme naabrite kaale sisendile sarnasemaks. Iga õppevooru järel muudetakse kohandamismäära väiksemaks.

Klassifitseerimisülesande puhul võib SOM väljundsõlmede arv olla vastavuses soovitud klasside arvuga, aga võib olla ka suurem – väljundi tõlgendaja võib ise lähedased väljundsõlmed ühte klassi koondada. Kaugseire kujutise klassifitseerimisel on sisendvektoriteks kujutise pikslid ja tunnusteks pikslivääртused eri kanalites või andmekihtides. Kui soovitakse SOM sõlmi siduda etteantud klassidega, siis kasutatakse sisendvektoritena vaid kontrollitud klassikuuluvusega piksleid õpetus-valimis. SOM klassifikatsiooni rakendamisel väljaspool õpetusandmeid hinnatakse klassifitseeritava vektori sarnasust SOM väljundkihi sõlmedega, mis erinevalt *k* lähima naabri meetodi klassifitseerimise tulemustest on üldistused, mitte üksikud väljavalitud vaatlusvektorid. Tundmatu klassiga vektorile omistatakse kõige sarnasema SOM sõlmega seotud klass. Saab kujutada ka iga klassifitseeritava koha sarnasust iga üksiku klassiga või SOM sõlmega ning sarnasustaset otsuse langetamisel.

3.4.5.3. Evolutsioonilised ja geneetilised algoritmid

Evolutsiooniliste algoritmide hulka loetakse mitmesugused stohhastilised meetodid, mis püüavad leida parimat mudelit jäljendades looduslikku ehk loomulikku valikut. Tähtsamad evolutsioonilised algoritmid on geneetilised algoritmid ja geneetiline programmeerimine. **Geneetiliste algoritmidega** otsitakse parimat hüpoteesi või mudelit vaatlusandmetes olevate seoste kirjeldamiseks, kusjuures evolutsioneeruvateks objektideks on mudelit iseloomustavad parameetrid. **Geneetilise programmeerimise** puhul on arenevateks objektideks arvutuseeskirjad.

Evolutsioonilised algoritmid jäljendavad loomulikku valikut; geneetilised algoritmid modelleerivad seoseid kasutades evolutsioonilisi algoritme

Evolutsiooniline algoritm loob iseorganiseeruva tehissüsteemi, mis võib otsida varemvalitud mudelile sobivamaid parameetreid või otsida keerukale probleemile head lahendit paljude võimalike hulgast, nagu loodus otsib kõige sobivamat eluvormi. Evolutsioonilised algoritmid on rakendatavad igasuguste statistiliste mudelite ja andmete jaotustüüpide puhul. Geneetilised algoritmid sobivad parimate lahendite otsimiseks paljude omavahel seotud tunnuste puhul, kus iga üksiku faktori eraldi mõju on raske modelleerida. Evolutsiooniliste algoritmide eelis intellektitehnikas tavapäraste sammhaaval parema lahendi poole liikuvate algoritmide ees on lõtv sõltuvus sobivuspinna gradiendist. Tänu sellele ületavad evolutsioonilised algoritmid sobivuspinna orge kergemini. Meetodi puudused on äärmiselt suur arvutusmahukus ja võimalus, et vaatamata mutatsioonidele ei leita keeruka seose parimat lahendit. Lisaks sellele sõltub meetodi tõhusus tõesuskriteeriumist, mille valik ei pruugi alati õnnestuda.

Geneetilise algoritmi töö algab suure hulga kandidaatlahendite loomisega. Kandidaatlahendid annavad järglasi ja alluvad valikule, mis tähendab sobivamate lahendite suuremat arvu järgmisel iteratsioonil. Teatud hulk lahendeid muudetakse geneetiliste operaatoritega, mis on geneetilise muutlikkuse analoog. Geneetilist muutlikkust imiteeritakse nii mutatsioonide kui ka kombinatiivse muutlikkuse kujul. Mutatsioonid on seejuures vajalikud, et protsess lokaalsete optimumide juurest välja viia ja jätkata globaalse optimumi otsimist. Lahendite populatsiooni keskmine sobivus suureneb iteratsioonide käigus ja paljude iteratsioonide järel jääb reeglina alles andmetele kõige paremini sobiv lahend. Kui looduslik evolutsioon toimub pidevalt muutuvas keskkonnas, siis andmetöötlusliku probleemi puhul on tingimused, sealhulgas tõesuskriteerium, fikseeritud. Tänu sellele on tehisevolutsiooni puhul suurem lootus jõuda etteantud tingimustele kõige paremini vastavate lahenditeni kui looduses, kus liikide kohastumine muutuva keskkonnaga on igikestev protsess. Sellegipoolest tuleb arvestada asjaoluga, et ka muutumatute tingimuste korral võib tingimustega hästi sobivaid eluvorme olla mitmeid. Ka geneetilised algoritmid võivad anda erinevatel katsetel samadest õpetusandmetest erinevaid tulemusi.

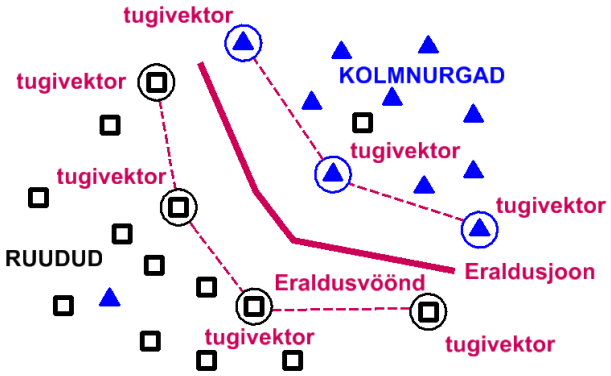
Evolutsiooniliste algoritmide meetodi alusepanijaks on J.H. Holland, ülevaateid võib leida töödest Holland (1992a, b), Chatterjee *et al*. (1996), Reggiani *et al*. (2001).

3.4.5.4. Tugivektormasinad

Tugivektormasinad (*support vector machines – SVM*) on õpetusandmeid kasutav klassifitseerimismeetod, mis tugineb Vladimir Vapniku välja töötatud statistilise õppe teooriale (Vapnik 1995, 1998). Tugivektormasinad kuuluvad suurima eraldusvööga klassifikaatorite (*maximum margin classifiers*) hulka. SVM-meetod otsib tunnusruumi jagavat pinda, millele lähimate andmepunktide (tugivektorite) omavaheline kaugus on kõige suurem (joonis 3-7). Klasse eraldava pinna asend sõltub vaid tugivektoritest.

Tugivektorid on klasside eralduspinnale kõige lähemal olevad andmepunktid

Tugivektormasinad on algselt leiutatud nominaalse funktsioontunnuse ja pidevate seletavate tunnuste jaoks. Nominaalsete argumenttunnuste puhul moodustatakse igast väärtusklassist omaette binaarne tunnus. Tugivektormasinad, nagu ka teised klassifitseerimismeetodid, on rakendatavad ka pideva muutuja väärtuste prognoosimiseks ehk regressiooniülesandeks.



Joonis 3-7. Tugivektorid (ringiga ümbritsetud sümbolid) klassipiiride määrajana ruutude ja kolmnurkade klassi vahel.

Uurimused

Tehisnärvivõrkude kasutamine on olnud suhteliselt edukas aegridade modelleerimisel, näiteks on koostatud USAs jahu hinna prognoosimise mudel ja saadud korrektseid hinnaprognoose (Chakraborty *et al*. 1992). Palju on tehisnärvivõrke kasutatud ka kujutise, eriti satelliidipildi töötlemisel (Dekruger ja Hunt 1994, Carpenter *et al*. 1999, Jensen *et al*. 1999, Kimes *et al*. 1999). Üksikuid katseid on tehtud ka taimkatte ja elupaikade leviku modelleerimiseks (Lek *et al*. 1996, Özesmi ja Özesmi 1999). Mineviku taimkatte leviku modelleerimiseks kasutasid ANNi Hilbert ja Ostendorf (2001). Metsa suktsessiooni on ANNga modelleerinud Gullison ja Bourque (2001). Tappeiner *et al*. (2001) leidsid, et ANN annab lumikatte kestvuse modelleerimisel Alpides tunduvalt täpsema prognoosi kui regressioonanalüüs. Schleiter *et al*. (1999) näitasid, et tehisnärvivõrgud sobivad vee kvaliteedi hindamiseks nii reostuse sissevoolu ja ilmastiku järgi kui ka põhjaloomastiku järgi. Walley ja Fontama (1998) kasutasid ANNi jõgede elustiku rikkuse prognoosimiseks jõe omaduste järgi. Mas *et al*. (2004) koostasid ANNi abil metsade kadumise prognoosikaarte kasutades satelliidipilte ja asukoha omadusi. Kimes *et al*. (1996) kasutasid tehisnärvivõrke metsa vanuse määramiseks Landsat TM kujutisest ja kõrgusmudeli andmetest.

SOM tehnoloogia abil ja kaugseireandmeid kasutades on kaardistatud taimkatet (Foody 1999), maakatteüksusi (Ji 2000, Bagan *et al*. 2005, Jianwen ja Bagan 2005) ja puiduvaru (Stümer *et al*. 2010). Grebby *et al*. (2010) kasutasid geoloogilisel kaardistamisel maapinna kõrgusandmetest moodustatud topograafilisi andmekihte. R. Céréghino *et al*. (2005) kujutasid sisserännanud kalaliikide elupaiga-eelistuste levila sarnasust Kohoneni kaardil.

Geneetilist algoritmi kasutas Whigham (2000) Austraalia metsades elava kukkurlooma *Petauroides volans* asustustiheduse modelleerimiseks. Seppelt ja Voinov (2002) modelleerisid geneetilise algoritmiga optimaalset maakasutust. D'heygere *et al*. (2003) kasutasid geneetilist algoritmi bentiliste suurselgrootute esinemise modelleerimiseks parima muutujate komplekti valimisel. Esinemise/puudumise mudel koostati otsuste puuna. Ortega-Huerta ja Peterson (2004) modelleerisid Mehhiko lindude ja imetajate ökoloogilisi nišše ja potentsiaalset levikut kasutades geneetiliste algoritmide tarkvara Desktop GARP (ptk 5.6.6.2).

Tugivektormasinaid on kaugseirepõhises maakattekaardistuses kasutanud näiteks Foody ja Mathur (2004) ning Keramitsoglou *et al*. (2006). Remm *et al*. (2011) võrdlesid tugivektormasinaid teiste meetoditega paljuaastase keskmise sademete hulga modelleerimisel Baltimaade ilmavaatlusjaamade andmete järgi. SVM tulemused ei olnud parimate hulgas, arvatavasti seetõttu, et SVM on loodud klassifitseerimisülesannete lahendamiseks, mitte pideva muutuja (milleks on sademete hulk) prognoosimiseks.

3.4.6. Sarnasusele tuginev järeldamine

Sarnasusele tuginevate meetodite ehk juhtudele ehk **näidistele tuginevate järelduste** (*similarity-based reasoning – SBR, case-based reasoning – CBR, instance-based reasoning – IBR*) puhul otsitakse olemasolevate andmete hulgast prognoositavale vaatlusele, objektile või kohale kõige sarnasemaid näidiseid. Näidistega sarnasuse järgi saab prognoosida nii pideva muutuja väärtust (prognoosida) kui ka nominaalse muutuja klassikuuluvust (klassifitseerida). Näidiste kasutamise ainus eeldus on, et objektide sarnasust peab olema võimalik arvuliselt mõõta – mõõdetud sarnasusi peab saama võrrelda ja järjestada. Kuna sarnasust käsitletakse kaugusena tunnusruumis, siis nimetatakse sarnasusele tuginevaid meetodeid ka lähima naabri meetoditeks.

Näidistele tuginev järelduste tegemine on psühholoogias tunnetusprotsessi mudel, intellektitehnikaga tegelejatele arvutusalgoritmi tüüp ja ekspertsüsteemide loojatele ekspertsüsteemi ülesehituse tüüp (Aha 1998). Suurem osa inimeste mõtlemisest ja tegutsemisest arvatakse tuginevat kogemusele – teadlikult või alateadlikult meelde jäänud juhtude lahendustele. Laps õpib kõndimagi proovides ja kukkudes, mitte etteantud reeglite järgi. Samamoodi püüab tudeng eksamiks valmistuda ja sportlane võistluseks valmistuda viisil, mis on varem häid tulemusi andnud. Isegi siis, kui sellisele käitumismallile on teoreetilisi vastuväiteid. Uutmoodi toimides soovitakse kogemustepagasit täiendada.

Näidistele tuginevat järeldamist käsitletakse erinevas mahus. Kitsamas tähenduses koosneb see järgmistest etappidest:

- leia kõige sarnasemad näidised,
- eralda näidistest probleemi varasem lahendus,
- kohanda lahendust uuele olukorrale,
- säilita uus juhtum näidisena järgmiste juhtude jaoks.

Inglise keeles võetakse need etapid kokku nelja r tähega algava sõnaga: *retrieve, reuse, revise, retain*.

Näidistele tugineva järeldamise etapid on: sarnaste näidiste leidmine, eeskuju võtmine, eeskuju kohandamine ja juhtumi lisamine näidiste baasi

Näidistele tuginev järeldamine on erinevalt teistest sarnasusele tuginevatest meetoditest suunatud eelkõige juhtumite lahendamisele ja erineb eelkõige näidistelt saadud eeskuju kohandamise poolest.

Näidiste abil klassikuuluvuse prognoosimisest oli juba juttu klassifitseerimismeetodite peatükis (ptk 2.3.3), näidiste abil saab aga hinnata igasuguseid muutujaid ja selliste hinnangute automatiseeritud andmiseks luuakse näidistega sarnasust kasutavaid ekspertsüsteeme. Sarnasusele tugineva prognoosisüsteemi teadmised on talletatud näidiste ja nende omaduste komplekti ehk **näidiste baasi** (*case base*). **Näidis** (*exemplar*) on õpetusandmete hulgast valitud juhtum või vaatlustulemus, mis kannab **elementaartunnuseid** (*features*), mis aitavad tuvastada prognoositava muutuja väärtust.

Lisaks näidistele on sarnasusele tuginevaks järeldamiseks tarvis valida sarnasuse mõõtmise reeglid. Mahukast näidistebaasist kõige sarnasemate näidiste leidmisel on oluline andmete otsingu tehnoloogia. Muud teooriat ja mudelit sarnasusele tuginev järeldamine ei vaja.

Sarnasuse järgi otsustamise kesksed küsimused on: kuidas valida ja kaaluda tunnuseid ja näidiseid ning kuidas kohandada hinnang uuele juhtumile

Näidistele tuginevad tehisõppesüsteemid erinevad mustrituvastuse vallas juba aastakümneid tuntud *k* lähima naabri (*kNN*) meetodist tehisõppe kasutuse poolest. *kNN* meetodi puhul kasutatakse kõiki tunnuseid, seevastu otsivad tehisintellekti süsteemid hinnangu andmiseks olulisi tunnuseid ja

kasutavad vaid neid. Suure hulga kirjeldavate tunnuste puhul piisab tavaliselt 6...10 tunnusest, liigsete tunnuste lisamine vähendab hinnangute täpsust. See on nähtus, mida nimetatakse **dimensionaalsuse needuseks** (*curse of dimensionality*). Faktor- ja peakomponentanalüüs ühendavad tunnuseid üldistatud faktoriteks, näidistele tugineva süsteemi tehisõppes toimub põhiliselt tunnuste valimine ja kaalumine ning näidiste kiiremat leidmist toetavate indeksite loomine näidiste baasi.

Dimensionaalsust saab vähendada tunnuseid ühendades ja tunnuseid valides

Eri valdkondades kasutatavaid näidistele tuginevaid intellektitehnika vorme ühendab **laisk probleemikäsitlus** (*lazy problem solving*). Laisa käsitluse puhul arvutatud sarnasusi ja töödeldud näidiseid ei säilitata, vaid käivitatakse iga kord uus päring näidistebaasi.

Laisale käsitlusele on iseloomulik:

- lähteandmete töötlus vaid otsese vajaduse korral,
- lahenduse otsimine olemasolevate töötlemata andmete kombinatsioonidest,
- vaheproduktide hülgamine.

Laisa käsitluse vastand on **innukas probleemikäsitlus** (*eager problem solving*). Innukad käsitlused asendavad lähteandmed esimesel võimalusel üldistustega (mudelitega, üksikute arvnäitajatega, kirjeldustega, teooriaga, reeglitega, abstraktsioonidega). Innukate algoritmide eelis on suure lähteandmete hulga asendamine väiksema hulga üldistustega.

Laisk käsitlus väldib üldistusi, innukas käsitlus asendab toorandmed üldistusega esimesel võimalusel

Laisa käsitluse eelistena loetletakse järgmist:

- pole vaja reegleid ja mudeleid, pole vaja ka abstraktsioonide loomisele vahendeid kulutada,
- toorandmed saab kasutada mitme eri tüüpi ülesande lahendamiseks, mudelid koostatakse või vähemalt sobitatakse ühe kindla probleemi lahendamiseks,
- laisad käsitlused annavad paremaid tulemusi keerukate nähtuste ja seoste korral, kus lihtsad abstraktsioonid ei ole tõepärased,
- tulemuste interpretatsioonil saab tugineda konkreetsetele pretsedentidele, sarnaseid näidiseid saab üldistada seletusteks,
- näidiste omaduste hulgas on lihtne säilitada andmeid näidise kujunemisloost,
- näidiste abil probleemide lahendamine on intuitiivselt hästi mõistetav (Aha 1998).

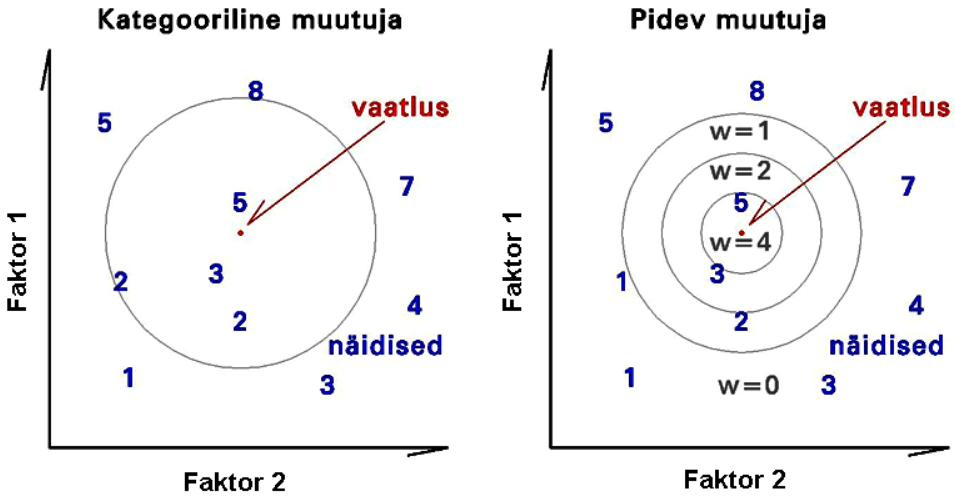
Kui statistilised mudelid tuleb vaatlusandmetele sobitada, siis tehisintellektisüsteemide puhul nimetatakse süsteemi õpetusandmete järgi optimeerimist tehisõppeks. Laisale probleemikäsitlusele vastab **laiskõpe** (*lazy learning, memory-based learning*), mille põhivahendid on näidiste ja tunnuste valik, nende kaalumine ning prognoosiva süsteemi tööd kiirendavate eelklassifikaatorite valik. Parimaks prognoosiks ei ole enamasti vaja kõiki näidiseid ja näidistel mõõdetud tunnuseid. Optimaalne tunnuste komplekt võib olla erinev, kuid enamasti on komplekssete muutujate prognoosimiseks rohkem tunnuseid tarvis. Tunnuste kaalumine ei ole kasulik mitte ainult parimate prognooside saamiseks, vaid see võib anda ka sisulisemat teavet faktorite mõju vahekordade kohta. Teisest küljest annab faktorite olulisuse hindamine teavet selle kohta, milliseid tunnuseid on kindlast vaja mõõta ja milliste mõõtmine ei ole eriti oluline. Samuti saab näidiste kaalude järgi leida need kirjeldused, mis otsitavaid klasse

(näiteks taimekooslusi) kõige paremini iseloomustavad ja teistest klassidest eristavad. Ebatüüpilised või tõenäoliselt ebatäpselt määratud või teisi dubleerivad näidised võib edaspidi kasutusest kõrvale heita. Kui kasutatakse vaid ühte kõige sarnasemat näidist ja see ei vasta täpselt uuele juhtumile, tuleb näidis kohandada ehk adopteerida.

Näidistega sarnasuse abil prognoosiva süsteemi võib optimeerida kas:

- ühe ja parima näidise leidmisele ning seejärel näidise adopteerimisele,
- teatud arvu parimate näidiste leidmisele (*k* **lähima naabri meetod** – *kNN* ehk *k-NN*),
- sarnasusega kaalutud *k* lähima naabri meetod (joonis 3-8),
- kõigi teatud minimaalselt nõutava sarnasusega näidiste kasutamisele (*d* lähima naabri meetod – *dNN* ehk *d-NN*),
- minimaalselt vajaliku või maksimaalselt kasutatava summaarse sarnasuse kasutamisele,
- eeltoodud kriteeriumite kombinatsioonidele.

Siinmainitud *dNN* algoritmi kasutavaid meetodeid nimetatakse ka **kernelmeetoditeks,** tuumik-meetoditeks või tuumameetoditeks (*kernel methods*), kernelit ennast aga tuumikuks või tuumaks.



Joonis 3-8. A – kategoorilise muutuja klassifitseerimine kaalumata *k* lähima naabri meetodil. Ühe kõige sarnasema näidise järgi tuleks vaatlus lugeda klassi koodiga 5, nelja näidise kasutamisel klassi 2. B – pideva muutuja prognoos kaalutud *k* lähima naabri meetodil. Joonisel näidatud kaugusest tunnusruumis sõltuvate kaalude (w) järgi tuleks vaatlusele omistada väärtus 4.

Vaid ühe näidise kasutamise korral on prognoos tundlik müra, see on ebatüüpiliste näidiste suhtes. Tunnusruumi fikseeritud osa kasutav *dNN* meetod on arvutuslikult kiirem kui *kNN*, kuid jätab prognoosi andmata siis, kui piisavalt sarnast näidist ei leidu. Fikseeritud ruumiosa kriteerium kipub ebaühtlase andmestiku korral andma ebaühtlase usutavusega prognoosi, kuna fikseeritud ruumiosa sisse jääb erinev arv näidiseid. *kNN* meetod on arvutatav igasuguse andmete tiheduse korral, see tähendab aga, et meetod annab prognoosi ka olukorras, kus ühtegi sarnast näidist tegelikult ei ole.

Prognoos *k* lähima naabri järgi saadakse ka siis, kui usaldusväärse sarnasusega näidiseid selle arvutamiseks ei ole.

Sobiv *k* suurus määratakse kas intuitiivselt või võetakse see omaette uurimisprobleemiks. Parim variant sobiva *k* otsustamisel võib olla sarnasusele vastavate kaalude kasutamine. *kNN* meetodi kasutamisel vähese arvu nominaalsete tunnuste korral, kui sarnasus iga tunnuse suhtes on kaheväärtuseline (vaatlus kas kuulub samasse klassi või ei kuulu), võib sageli leiduda palju

samaväärse sarnasusega näidiseid, millest parima valimine on problemaatiline. Kuna suure arvu tunnuste korral enamasti ei leidu klassifitseeritavale juhtumile täpselt vastavat näidist, siis kasutatakse pidevate andmete puhul prognoosina sarnasemate näidiste nende sarnasusega kaalutud keskmist. Kategooriliste andmete puhul loetakse prognoosiks kas kõige sarnasem näidis või samasarnaste näidiste hulgas kõige sagedamini esinev variant. Prognoosi saab anda ka tõenäosusjaotusena.

Näidiste abil liikide leviku modelleerimise ülesande puhul prognoositakse liigi võimalikeks esinemiskohtadeks kohad, mis on tingimustelt sarnased nende kohtadega, kus liiki on senini leitud. Tingimuste kompleksi sarnasust mõõdetakse sarnasusindeksitega või kaugusega tunnusruumis (vt ka ptk 5.6.8).

Näidistega samastamine on küll ühest küljest väga paindlik ja universaalne klassifitseerimis- ja prognoosimeetod, kuid teisest küljest jätab suur paindlikkus ruumi ka subjektiivsusele. Näidiste komplekt või selle koostamise reeglid väljendavad uurija arusaama klassifikatsiooniüksustest. Oma arusaamade vormistamine näidiste komplektiks võib aga olla nii kunst kui teadus. Näidisteks võivad olla juhuslikud objektid empiirilistest vaatlustest, sihilikult valitud tüüpiliseks peetavad eksemplarid või tüüpiliste näidiste konstruktsioonid. Konstrueeritud näidiseid on kasutatud näiteks puistu struktuuri äratundmiseks aerofotode järgi (St-Onge ja Cavayas 1997, Silbernagel ja Moeur 2001). Teiseks näidistele tugineva järeldamise puuduseks on vaid lokaalse sarnasuse arvestamine, seetõttu on näidistele tuginevatel tehisõppemeetoditel kalduvus sobivuspinna lokaalsete maksimumide juurde kinni jääda.

Näidistega sarnasusele tugineva hindamise tugevus on keerukate seoste ja erandlike juhtude äratundmise võime, üldised seaduspärasused jäävad kergesti märkamata

Näidistele tuginevate prognoosisüsteemide eelistena mainitakse järgmist:

- järkjärgulise õppimise võimalus, mille tõttu ei ole tarvidust luua ennatlikke abstraktsioone,
- parem prognoosivõime keerukate andmestike puhul,
- sobivus kasutamiseks sündmuseridade puhul (malekäigud),
- järelduste lokaalne (päringuspetsiifiline) iseloom analoogiliselt lokaalselt sobitatud regressioonimudelitega,
- näidistele tuginev prognoosisüsteem on kasulik juhtudel, kui andmete salvestamise ja töötluse aega tasub kokku hoida ja kui parimate näidiste ülesleidmine on efektiivselt optimeeritud,
- kui osade tunnuste väärtused puuduvad osade näidiste puhul (puuduvad väärtused ei sega oluliselt sarnasuse arvutust),
- kui parima hinnangu saamiseks on vajalikud nii tüüpilised näidised kui ka näidiste andmebaasis säilitatavad erandlikud vaatlused (Aha 1998),
- sarnasusele tuginevad klassifikaatorid suudavad moodustada suvalise kujuga otsustuspiire (Tan *et al*. 2006),
- näidistele tuginevat prognoosisüsteemi on muutunud lähtetingimuste korral lihtne kohandada näidiste väljavahetamise teel.

Sarnasusele tuginevat hinnangute metoodikat on kõige enam kritiseeritud tugeva teoreetilise põhjenduse puudumise pärast. Lihtsamad ülesanded võivad tõesti valmis reeglitele taanduda. Sarnasusele tuginevad hinnangud on eelkõige keerukate olukordade ja andmestike jaoks, kus efektiivseid reegleid ei ole leitud. Näiteid ja näidiseid esitada ja mingil viisil sarnasust määrata saab nii lihtsatel kui ka keerukatel juhtudel. Sarnased näited sobivad muidugi ka neile, kes tõhusaid reegleid ei tunne.

Näidiste järgi hindamise metoodika edasiarendused hõlmavad mitmesuguseid laiskade ja innukate algoritmide kombinatsioone, nagu näiteks üldistatud näidiste ehk **prototüüpide** loomine asendamaks omavahel sarnaseid näidiseid (Rajamoney ja Lee 1991) ning **laisad otsuste puud** (*lazy decision tree)*, mille puhul kasutatakse näidiseid otsuste puu teatud tasemetel, kas lõppharudes või raskesti defineeritavate hargnemiste kohtades (Indurkhya ja Weiss 1995, Friedman *et al.* 1996) või siis tunnuste lokaalset kaalumist ja näidiste intelligentsemat valikut (Zhang *et al*. 1997). Praktikas kasutatakse deduktiivse ja induktiivse lähenemise kombinatsioone. Teoreetilist teadmist saab koos empiiriliste andmetega kasutada näiteks mudeli tüübi valikul ja võimalike algparameetrite valikul, mudeli sobitamise protsessi suunamisel ja tõesuskriteeriumi valikul.

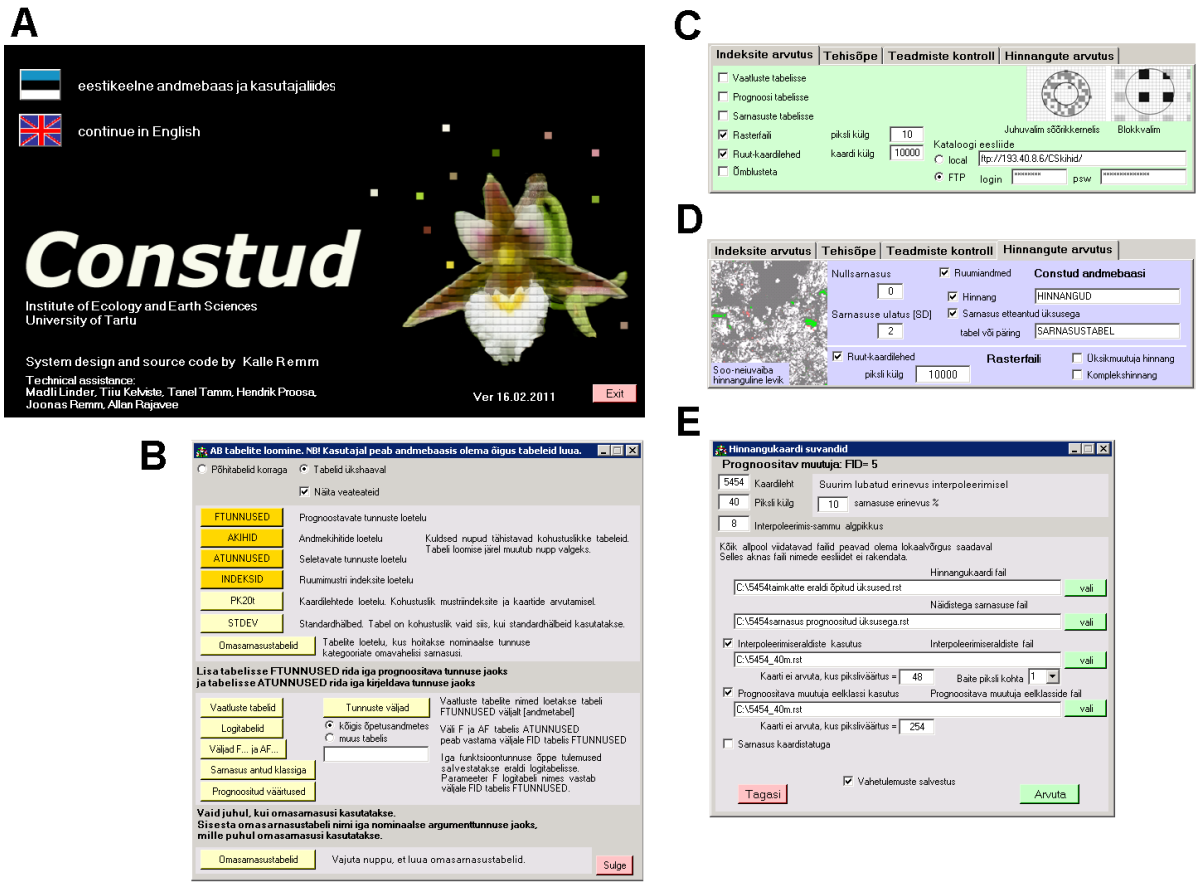
Näidistele tugineva järeldamise lähedane meetod on **analoogia otsimine** (*analogue matching*), mille puhul otsitakse sarnasust nähtuste struktuuris, mitte sisulisemat tähtsust omavates tunnustes. Analüütilise järeldamise (*analytical reasoning, explanation based learning*) puhul ei otsita parimat seletavat mudelit või hüpoteesi mitte vaatlusandmetest, vaid eelnevast teadmisest ja deduktiivsest seletusest. Vaatlusandmeid ei kasutata mitte mudeli koostamiseks, vaid deduktiivselt koostatud mudeli kontrollimiseks. Analüütilise järeldamise eelised ilmnevad eelkõige juhul, kui empiiriline andmestik on vähene ja ei võimalda piisavalt usaldusväärset mudelit koostada. Varasemad üldistused võimaldavad koostada mudelit nii väljapoole vaatlusandmeid kui ka varakult loobuda ebareaalsetest mudeli variantidest ja eemaldada ebaolulisi muutujaid. Teoreetiliste teadmiste kasutamise edukus sõltub nii teooria paikapidavusest lahendatava ülesande puhul kui ka sellest, millisel määral kasutatav teoreetiline printsiip vaatlusandmete varieeruvust ära seletada suudab.

3.4.6.1. Tarkvarasüsteem Constud

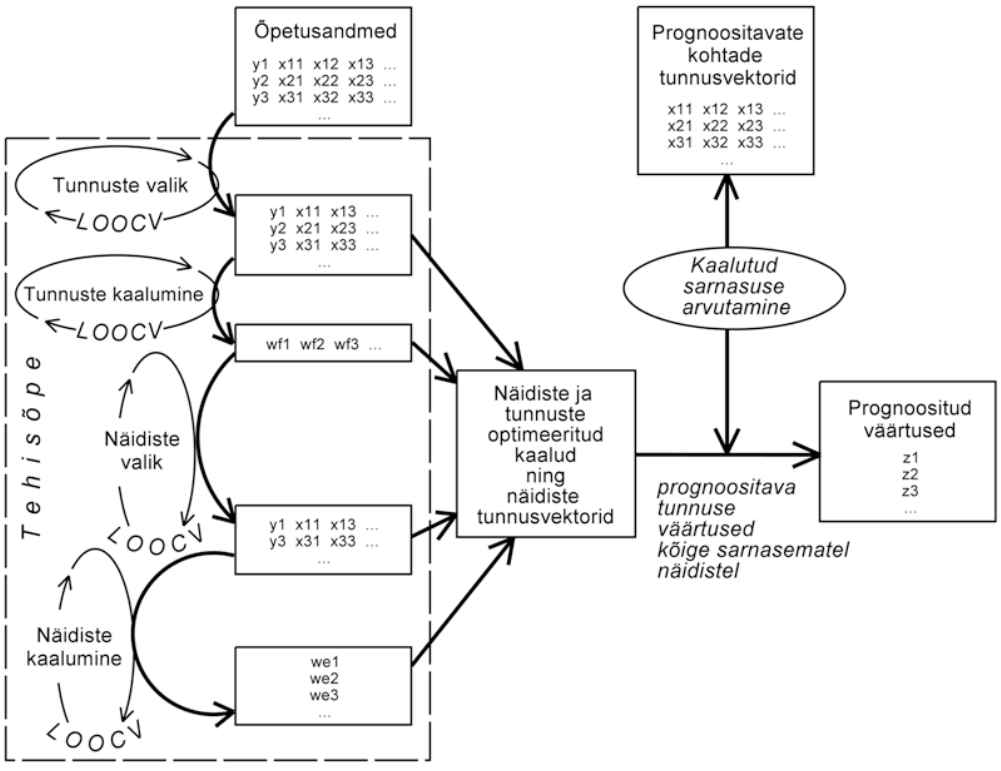
Constud on eelkõige sarnasusele tugineva järeldamise ja sarnasusele tuginevate hinnangukaartide loomise süsteem, mis koosneb:

- seletavate tunnuste, eelklassifikaatorite ja interpoleerimiseraldiste rasterkujul andmekihtidest (kui kasutatakse ruumiandmeid);
- vaatluste, tehisõppe parameetrite, tehisõppe tulemuste ja tabelisse prognoositud väärtuste andmebaasist;
- Constud tarkvarast (joonis 3-9).

Constud tarkvara keskne osa on sarnasusele tuginev tehisõpe, mille tulemusel valitakse funktsioontunnuse prognoosimiseks vajalikud tunnused ja näidised ning leitakse neile optimaalsed kaalud (joonis 3-10).



Joonis 3-9. Mõned Constud tarkvara dialoogiaknad (Remm ja Kelviste 2011c). A – tervitusaken. B – andmetabelite loomine. C – ruumiliste indeksite arvutuse suvandid. D – hinnangute arvutuse suvandid. E – hinnangukaardi suvandid.



Joonis 3-10. Constud tarkvara tehnoloogiline skeem (Remm ja Kelviste 2011c).

Constud süsteemi ülesehituse kavandas ja tarkvara lähtekoodi kirjutas Kalle Remm aastatel 2002–2011. Constud nimi tuleneb ingliskeelsetest sõnadest *Continuous Study*. Süsteemi kolm põhilist kasutusvaldkonda on järgmised:

- ruumimustrit kirjeldavate indeksite arvutamine rasterkujul andmekihtidest andmebaasi tabelitesse või rasterfaili;
- tehisõpe ehk tunnuste ja vaatluste komplekti iteratiivne valik ning kaalude sobitamine õpetusandmetele;
- nominaalsete ja arvuliste muutujate sarnasusele tuginev prognoosimine andmebaasi tabelitesse või rasterfaili.

Constud on vabavara, mille installeerimisviit on süsteemi veebilehel www.geo.ut.ee/CONSTUD. Põhjalikuma ülevaate Constud süsteemist saab selle kasutamisõpetusest (Remm ja Kelviste 2011a, b, c) ja Constud veebilehelt. Constud süsteemi kasutamisest liikide leviku kaardistamisel on juttu peatükis 5.6.8.3.

Uurimused

Blum ja Langley (1997) võrdlevad tunnuste ja näidiste valimise algoritme. Wilson ja Martinez (2000) on kirjutanud ülevaate näidiste valimise algoritmidest. Tunnuste kaalumisest on ülevaateartikli kirjutanud Wettschereck ja Aha (1995) ja Wettschereck *et al.* (1997), uuemate ülevaadete autor on C.J.C. Burges (2009, 2010).

Sarnaseid analooge on kasutanud karude elupaiga eelistuste modelleerimiseks Clark *et al.* (1993), taimkatte modelleerimiseks Wilds *et al.* (2000), rekultiveeritavatele kaevandusaladele kujuneda võiva taimkatte prognoosimiseks Osborne ja Brearley (2000), kasvukohatüüpide kaardi koostamiseks ja puuliikide katvuse hindamiseks Otepää looduspargis Remm (2002). Analoogmeetodit kasutatakse ka paleoökoloogilistes rekonstruktsioonides (Birks 1993, Flower *et al*. 1997).

Kompleksne, neljale erinevale teadmiste baasile tuginev ekspertsüsteem on moodustatud Kaspia mere kalavarude hindamiseks ja püügikvootide jagamiseks (Sazanova *et al*. 1999).

Džeroski ja Drumm (2003) võrdlesid lineaarse regressioonimudeli, regressioonipuu ja *kNN* meetodi täpsust merikurkide (*Holothuria leucospilota*) arvukuse prognoosil ja leidsid, et *kNN* meetod andis korrelatsioonikordaja ja keskmise lineaarhälbe järgi kõige täpsema tulemuse.

Whigham (2005) prognoosis klorofüll *a* dünaamikat sarnaste järgnevuste alusel.

Näidisalade ja kaugseireandmete järgi puistu takseertunnuste määramise koolkond on Soome metsakorralduses. Lisaks Soomele kasutatakse *kNN* meetodit ka USA (Franco-Lopez *et al*. 2001, McRoberts *et al*. 2006), Rootsi (Reese *et al*. 2003), Kanada (Labrecque *et al*. 2006) ja Uus-Meremaa (Tomppo *et al*. 1999, Trotter *et al*. 1997) metsanduses. Ülevaade metsa takseerandmete hinnangulise kaardistamise tulemustest on peatükis 5.8.

Constud tarkvara on rakendatud lasteaialaste enterobiaasiriski hindamisel (Remm ja Remm 2008), taimkatte kaardistamisel (Linder *et al*. 2009), okas-, leht- ja segametsa eristamisel (Oviir *et al.* 2008), käpaliste leviku hinnangulisel kaardistamisel (Remm ja Remm 2009) ja sademete hulga kaardistamisel (Remm *et al.* 2011).

3.5. Aegridade modelleerimine

3.5.1. Aja võimalikud rollid mudelis

Ajal on mudelitest kolm võimalikku rolli. Funktsioontunnusena käsitletakse aega (elu)kestusanalüüsis (*life expectancy analysis*), kus uuritavaks suuruseks on aeg, mis kulub teatava sündmuseni (surmani, haigestumiseni, detaili riknemiseni jmt). Aeg on müratunnuseks siis, kui katse on kavandatud nii, et vaatlused tuleks teha ühel ajamomendil, kuid tehnilistel põhjustel on vaatlusaeg veninud pikemaks. Selle vältimiseks tuleks aeg võtta argumenttunnuste hulka ja arvestada temast sõltuvust.

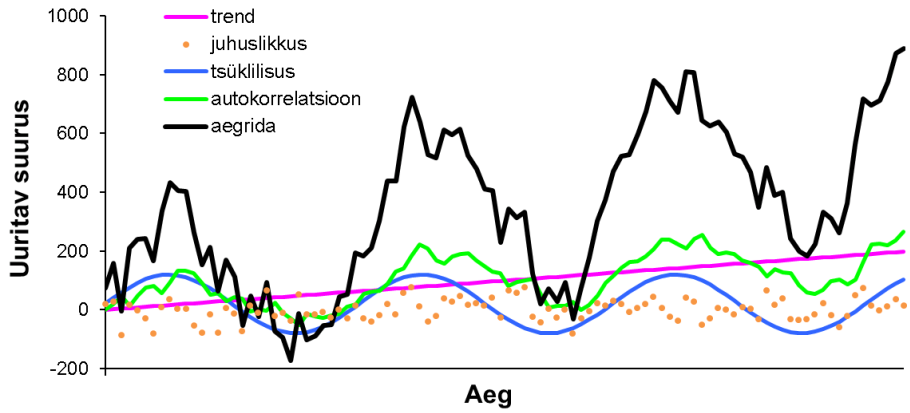
Ajas muutuva tunnuse või tunnuste väärtusi nimetatakse **aegreaks** (*time series*). Aegridade mudelites on aeg põhiline ja tihti ainus argumenttunnus. Aegrea andmestik erineb klassikalisest valimist, sest tüüpiliselt on igal ajahetkel olemas vaid üksainus mõõtmistulemus. Samas võib mõõtmistulemuste hulk (mõõtmishetkede arv) olla küllaltki suur. Aegrida on üldiselt defineeritud diskreetsena ja enamasti eeldatakse, et mõõtmised toimuvad ajas korrapäraselt, fikseeritud sammuga. Kui üheaegselt on mõõdetud mitut tunnust, siis on aegrida mitmemõõtmeline.

Aeg võib mudelites olla funktsioontunnuse, müratunnuse või argumenttunnuse rollis

Aegrea analüüsi eesmärgiks on reeglina aegrea jaotamine komponentideks:

- kindla suunaga muutuseks ehk trendiks,
- ühetaolisena korduvaks mustriks ehk tsüklilisesks komponendiks,
- ajalise autokorrelatsiooni mõjuks ja
- juhuslikeks hälveteks (joonis 3-11).

Aeg on üks argumenttunnus ka **kordusmõõtmiste** (*repeated measures*) mudelites. Kordusmõõtmiste puhul on tegemist kolmemõõtmelise andmestikuga, milles ühe dimensiooni moodustab uuritavate tunnuste loetelu, teise mõõdetavate objektide loetelu ja kolmanda mõõtmishetkede loetelu. Ideaaljuhul on mõõtmised tehtud kõigi objektide ja kõigi tunnuste puhul samadel võrdse vahega ajahetkedel ja saadakse mitmemõõtmeline aegrida iga üksikobjekti kohta.



Joonis 3-11. Konstrueeritud aegrea neli komponenti ja nende komponentide liitumisel moodustav aegrida. Juhuslikud mõjud on normaaljaotusega (keskväärtus = 0, standardhälve = 40), autokorrelatsiooni mõju on arvestatud kolme ülejäänud faktori summaarse väärtuse eelmise kolme ajaühiku kaalutud summana (kaalud vastavalt ajavahele: 0,1; 0,3 ja 0,5).

3.5.2. Autokorrelatsioon ajas

Kuna muutused toimuvad mingi lõpliku kiirusega ja on lisaks sellele tihti ajalise viivitusega, siis on ajateljel lähestikku paiknevad vaatlused sarnased. Lähedaste vaatluste sarnasuse ehk autokorrelatsiooni tugevus sõltub arvesseminevast ulatusest ehk **laagist**. Aegridade puhul tähendab laag ajavahemikku, ruumiliste nähtuste korral teadaolevate väärtuste otsinguraadiust (ptk 5.1.1). Nagu enamiku teiste analüüside puhul, nii eeldatakse ka aegridade analüüsil, et andmed sisaldavad süstemaatilist osa (lineaarset trendi ja tsüklilisi muutusi) ning juhuslikke hälbeid (müra). Enamik aegridade analüüsi meetodeid sisaldab mingeid vahendeid müra eemaldamiseks.

Enamik aegridu saab kirjeldada trendi ja sesoonsuse kaudu. Trendi automaatseks leidmiseks aegreast ei ole ühte ja ainuõiget meetodit, ent üldistel eeldustel on trendi olemasolu võimalik kontrollida näiteks **Mann-Kendalli testi** abil. Esimene samm trendi määramisel on tavaliselt silumine. Silumine sisaldab alati mingil kujul lokaalset ehk libisevat keskmistamist. Seejuures eeldatakse, et juhuslikud hälbed kompenseerivad üksteist. Lokaalne silumine toimub mingi **silumisulatuse** piires, mis võib haarata ka kogu andmerida. Sel juhul kaalutakse andmed vastavalt nende kaugusele ajateljel. Asümmeetriliste hälvete ja erindite esinemise korral tuleks lokaalsele keskmisele eelistada lokaalset mediaani. Splainide kasutamine on mõistlik korrapäraselt paiknevate punktide puhul.

Sesoonne aegrea komponent on formaalselt defineeritud kui korrelatsioon aegrea *i*-nda elemendi ja *i*·*k*-nda elemendi vahel. Sesoonsus on autokorrelatsiooni tsükliline vorm, konstant *k* on selle ulatus ehk laag. Sesoonseid mustreid aegridades saab kindlaks teha tsüklilist mudelit aegreale sobitades ja autokorrelogrammi abil. Autokorrelogramm näitab graafiliselt autokorrelatsiooni sõltuvust laagist.

Autokorrelogrammi puhul tuleb silmas pidada, et autokorrelatsioonid järjestikuste laagide puhul ei ole sõltumatud. Seetõttu võib korrelogramm esimest järku autokorrelatsioonide eemaldamise järel oluliselt muutuda. Erinevate laagide autokorrelatsioonide omavahelist sõltuvust on võimalik eemaldada ja arvutada osalised autokorrelatsioonid (*partial autocorrelations*), mis ei sisalda sõltuvust laagi piiresse jäävate elementide hulgast ja paiknemisest. Kui laag võrdub ühega, siis osa-autokorrelatsioon võrdub autokorrelatsiooniga. Sesoonsust saab eemaldada ka aegrea diferentsimisega, teisendades iga *i*-nda elemendi erinevuseks *i* miinus *k*-ndast elemendist. See võimaldab avastada varjatud sesoonsust.

3.5.3. Autoregressiivne libisev keskmine

Enamik aegridade modelleerimise meetodid on mitmefaktorilise **autoregressiivse libiseva keskmise** (*autoregressive moving average – ARMA)* mingid variatsioonid. ARMA metoodika on paindlik, kuid komplitseeritud viis varjatud korrapära leidmiseks ja prognooside koostamiseks aegrea andmetel. ARMA mudel koosneb autoregressiivsest (AR) komponenedist ja libisevast keskmisest protsessist (MA). ARMA meetod eeldab, et aegrida on statsionaarne ja vähemalt 50 vaatlust pikk.

Statsionaarsus – aegrida on statsionaarne vaid siis, kui autoregressiooni võrrandi parameetrite väärtused on teatud piires. Kui varasemad efektid akumuleeruvad, siis ei ole aegrida statsionaarne, vaid liigub kas lõpmatusse või nulli. Statsionaarse aegrea keskväärtus, dispersioon ja autokorrelatsioon on ajas konstantsed. Mingist statsionaarsuse eeldusest lähtuvad kõik statistilised mudelid. Statsionaarsus on seejuures teoreetiline lähtekoht ja modelleerija otsustus, mitte uuritava nähtuse püsiv omadus. Näiteks ökoloogiliste protsesside statsionaarsust muutuva kliima ja inimtegevuse tingimustes on naiivne oodata.

Autoregressiivne komponent on kirjeldatav regressioonivõrrandiga

$$x_t = \mu + \varepsilon_t + b_1 x_{(t-1)} + b_2 x_{(t-2)} + b_3 x_{(t-3)} \ldots = \mu + \varepsilon_t + \sum_i^T \left(b_i x_{(t-i)} \right), \quad [3\text{-}23]$$

kus μ on konstant, b_1, b_2, b_3 on autoregressioonimudeli parameetrid, t tähistab ajahetke, T ajahetkede arvu (sõltuvuse ulatust) ja ε_0 on juhuslik hälve. See tähendab, et iga vaatlus on määratud eelnevate vaatluste lineaarkombinatsiooni ja juhusliku hälbega. Kui vaatlus sõltub vaid ühest eelnevast vaatlusest, on see esimest järku autoregressiivne mudel, kui kahest eelnevast, siis teist järku mudel jne.

Libisev keskmine protsess – autoregressiivsest protsessist sõltumatult võib iga element olla mõjutatud eelnenud juhuslikest hälvetest.

$$x_t = \mu + \varepsilon_t + q_1 \varepsilon_{(t-1)} + q_2 \varepsilon_{(t-2)} + q_3 \varepsilon_{(t-3)} \ldots = \mu + \varepsilon_t + \sum_i^T \left(q_i \cdot \varepsilon_{(t-i)} \right), \quad [3\text{-}24]$$

kus μ on x ootusele vastav konstant (üldjuhul null) ja q_i on autoregressioonimudeli parameetrid, ε_i on juhuslik hälve.

ARMA mudeli koondkuju ühendab muutuja x ootusele vastava konstandi α (üldjuhul $\alpha = 0$) ning valemid [3-22] ja [3-23].

$$x_t = \alpha + \varepsilon_t + \sum_i^{T_1} \left(b_i x_{(t-i)} \right) + \sum_i^{T_2} \left(q_i \varepsilon_{(t-i)} \right) \quad [3\text{-}25]$$

ARMA modeli üldistus on ARIMA (*autoregressive integrated moving average*), mis sisaldab vahendeid protsessi mittestatsionaarsuse käsitlemiseks. ARIMA määratlemise faasis diferentseeritakse aegrida senikaua, kuni saadakse statsionaarne protsess. Diferentsimise astme määramiseks tuleks jälgida andmete graafikut ja autokorrelogrammi. Järsud üles-alla muutused nõuavad esimest järku mittesesoonset diferentsimist kasutades ühikulist ajavahet (laag = 1). Tugevad kaldenurga muutused nõuavad teist järku mittesesoonset diferentsimist. Sesoonne regulaarsus nõuab sesoonset diferentsimist. Kui autokorrelatsioon väheneb suuremate laagide puhul aeglaselt, on vajalik esimest järku diferentsimine.

ARIMA mudeli määratlemise faasis tuleb otsustada, kui palju autoregressiivseid ja kui palju libiseva keskmise parameetreid on vaja efektiivse mudeli jaoks. Praktikas on harva vaja enam kui kahte parameetrit kummastki klassist. Mudeli parameetrite arvu määramiseks saab kasutada ka andmete graafikut ja autokorrelatsioonide (AKK) ja osaautokorrelatsioonide korrelogrammi (OAKK). Parima mudeli leidmiseks on vaja erinevaid variante katsetada. Enamiku empiiriliste aegridade puhul sobib mõni järgnevatest variantidest piisavalt hästi:

- üks autoregressiivne parameeter: AKK langeb eksponentsiaalselt, OAKK tipuga esimese laagi kohal, teiste laagide puhul olulist korrelatsiooni ei ole;
- kaks autoregressiivset parameetrit: AKK lainetab, OAKK tipuga laagide 1 ja 2 juures, teiste laagide puhul olulist korrelatsiooni ei ole;
- üks libiseva keskmise parameeter: AKK tipuga laagi 1 juures, teiste laagide puhul olulist korrelatsiooni ei ole. OAKK langeb eksponentsiaalselt;
- kaks libiseva keskmise parameetrit: AKK tipuga laagide 1 ja 2 juures, teiste laagide puhul olulist korrelatsiooni ei ole, OAKK lainetab;
- üks autoregressiivne parameeter ja üks libiseva keskmise parameeter: AKK ja OAKK langevad eksponentsiaalselt.

ARIMA mudeli parameetrite väärtused määratakse vähimruutude meetodil. Määratud parameetreid kasutatakse prognoosi koostamiseks väljapoole lähteandmete piirkonda. Parameetrid määratakse diferentseeritud andmetest. Enne prognoosi koostamist tuleb need tagasi teisendada (integreerida) originaalandmete kujule. ARIMA mudeli konstandi tähendus sõltub mudelist. Kui mudelis ei ole autoregressiivseid parameetreid, siis väljendab konstant aegrea keskväärtust. Kui mudelis on autoregressiivsed parameetrid, siis väljendab konstant mudelijärgset väärtust ajahetkel 0.

Ajas korduva sesoonsuse modelleerimiseks tuleb lisaks eeltoodud parameetritele määrata sesoonsuse parameetrid: sesoonse autoregressiivsuse (*ps*), sesoonse diferentseerumise (*ds*) ja sesoonse libiseva keskmise parameeter (*qs*). Näiteks mudeli tähistus ARIMA(0, 1, 2)(0, 1, 1) tähendab, et mudel ei sisalda autoregressiivseid parameetreid, sisaldab kahte tavalist libiseva keskmise ja ühte sesoonse libiseva keskmise parameetrit ning parameetrid arvutati pärast ühekordset diferentsimist laagiga 1 ja ühekordset sesoonset diferentsimist. Sesoonse diferentsimise laag tuleb määrata juba identifitseerimise faasis. Sesoonse mudeli parameetrite arvu määramiseks soovitatakse samalaadseid korrelogrammidele tuginevaid kriteeriume kui lihtsa mudeli puhulgi. Põhiline erinevus on, et sesoonse mudeli puhul tuleb jälgida väärtusi korrelogrammil sesoonse laagi kaugusel.

Kui ARIMA mudeli hindamisel parameetrite standardhälvete järgi arvutatud *t* statistiku väärtused ei ole olulised, võib vastava parameetri mudelist välja jätta ilma, et mudeli sobivus oluliselt halveneks. Peale selle võib mudeli sobivust hinnata testandmete abil. Hea mudeli jääkide autokorrelogrammis ei tohiks olla lähipunktide sõltuvust. Prognoosijääkide jaotus peaks vastama normaaljaotusele.

Aegridade uurimise sage probleem on diskreetse sündmuse mõju hindamine aegreale. Aegrea katkestus võib olla järsk ja püsiv, sujuv ja püsiv või järsk ja ajutine. Sujuv ja ajutine mõju ei ole aegrea katkestus. Järsk püsiv katkestus muudab aegrea keskväärtust. Sujuv katkestus ilmneb alles teatud aja möödumisel. Sujuvat püsivat ja järsku ajutist katkestust iseloomustab mõju kestvus ja tugevus.

3.5.4. Eksponentsiaalne silumine

Eksponentsiaalne silumine on populaarne eelkõige oma lihtsuse ja heade prognoosiomaduste tõttu. Ajas kaugematele vaatlustele omistatakse eksponentsiaalselt väiksemad kaalud. Eksponentsiaalse silumise mudelisse saab lisada trendi. Trend võib olla lineaarne, eksponentsiaalne, silutud või prognoosi käigus korrigeeritav. Eksponentsiaalne silumise korral arvutatakse iga järgmine väärtus eelmiste väärtuste kaalutud keskmisena, kusjuures kaalud vähenevad kaugusega eksponentsiaalselt sõltuvalt parameetrist α. Kui α = 1, siis eelmisi vaatlusi ignoreeritakse täielikult, kui α = 0, siis ignoreeritakse antud vaatlust ja kõik vaatlused võrduvad esimese vaatlusega. Sesoonsus võib eksponentsiaalse silumise mudelis olla aditiivsena või multiplikatiivsena

$$S_t = aX_t + (1-a)S_{(t-1)} \qquad [3\text{-}26]$$

kus S_t on aegrea silutud väärtus ajahetkel *t* ning X_t on aegrea mõõdetud väärtus ajahetkel *t*.

3.5.5. Sesoonne jaotamine

Aegrida võib kujutleda koosnevana neljast komponendist: sesoonsus (S_t), trend (T_t), tsüklilisus (C_t) ja juhuslik komponent (I_t). Tsükliline komponent erineb sesoonsusest ebaühtlase pikkuse poolest. Sesoonse dekompositsiooni ehk sesoonse jaotamise eesmärgiks on eristada need komponendid. Eeldatakse, et komponendid on kas aditiivsed või multiplikatiivsed.

Aditiivse mudeli järgi avaldub vaadeldud väärtus ajahetkel *t* kujul

$$X_t = T_t + C_t + S_t + I_t; \quad [3\text{-}27]$$

multiplikatiivse mudeli järgi

$$X_t = T_t C_t S_t I_t. \quad [3\text{-}28]$$

Ameerika Ühendriikide rahvaloendusbüroos 1920ndatel aastatel välja töötatud Census I metoodika kohaselt ühendatakse sesoonsus ja tsüklilisus kas aditiivselt või multiplikatiivselt. Esmalt arvutatakse liikuv keskmine silumisulatusega üks sesoon. See eeldab sesooni pikkuse eelnevat teadmist. Aditiivse mudeli puhul lahutatakse silutud andmed originaalandmetest, multiplikatiivse mudeli puhul jagatakse originaalandmed silutud andmetega. Sellega eemaldatakse sesoonne ja juhuslik varieeruvus.

Census II meetod ei eelda *a priori* mudeli olemasolu. Selle asemel kasutatakse mitmeid *ad hoc* täpsustusi, mis on praktikas kasulikuks osutunud. Nende hulka kuulub näiteks tööpäevade arv kuus. Teiseks, Census II meetod võimaldab sesoonse elemendi määramisel erindeid välja jätta. Lisaks aegrea põhikomponentide leidmisele võib arvutada mitmesuguseid koondstatistikuid. Näiteks dispersioonanalüüsi tabeleid ja kuude protsentuaalseid erinevusi aegrea komponentide kaupa. Kui ajavahemik võrreldavate perioodide vahel suureneb, suureneb tõenäoliselt ka sesoonse trendi erinevus, kuid juhusliku komponendi erinevus ei tohiks muutuda. Keskmist intervalli, mille jooksul sesoonne komponent võrdsustub juhusliku komponendiga, nimetatakse **tsüklilise dominantsi perioodiks** (kuuks, kvartaliks) (*the month (quarter) for cyclical dominance – MCD*). MCD = 2 tähendab, et keskmiselt alates kahekuulisest vahest hakkavad sesoonsed muutused ületama juhuslikke fluktuatsioone.

Aegrea mudeli sobitamine Census II meetodil toimub seitsmeetapiliselt. Etappe tähistatakse traditsiooniliselt tähtedega A…G:

A. Eelsobitus (*prior adjustment*). Mitmesugused kasutajapoolsed sobitused ja kaalud, mis kas liidetakse või lahutatakse vaatlusandmetest või millega vaatlusandmeid kas korrutatakse või jagatakse.

B. Tööpäevade arvu varieeruvuse ja erindite mõju vähendamine.

C. Tööpäevade arvu ja ebaregulaarsete kaalude lõplik hindamine.

D. Sesoonse faktori, tsükli, juhusliku komponendi ja sesoonselt korrigeeritud aegrea lõplik määramine.

E. Originaalrida, sesoonsuse ja ebakorrapärase faktori mõju täpsustamine vähendades erindite mõju.

F. Koondnäitajate, tsüklilise dominantsi perioodi ja libiseva keskmise arvutamine.

G. Graafikute esitamine.

3.5.6. Jaotunud laagide analüüs

Jaotunud laagide analüüs (*distributed lags analysis*) on metoodika nihkega muutujate analüüsiks. Saadud tellimuste hulk ja täidetud tellimuste hulk on omavahel korreleeritud ajalise nihkega. Ka seaduste muutmise mõju majandusele ilmneb teatud nihkega, nagu ka investeeringute hulk ja tootlikkus.

Metoodika lähtub üldiste lineaarsete mudelite võrrandist, kus suurust *x* on mõõdetud erinevate laagidega. Kui regressioonikordaja b_i on mingi laagi puhul statistiliselt oluline, siis võib järeldada, et muutuja *y* on muutuja *x* alusel vastava laagiga seletatav ja prognoositav vastavalt valemile 3-29.

$$y_t = a + \sum_{i=0}^{T} b_i x_{t-i} \quad [3\text{-}29]$$

Üldine probleem eeltoodud regressioonimudeli korral on, et argumenttunnuse ajas lähestikku paiknevad väärtused on omavahel tugevasti korreleeritud, mudelisse tuleks lisada funktsioontunnuse sõltuvus iseendast ja seetõttu on regressioonikordajate määramine ebastabiilne. Küsimus on näiteks, milline alltoodud mudelitest on parim.

$$y_t = a + by_{t-1} + x_t \quad [3\text{-}30]$$

$$y_t = a + by_{t-1} + x_{t-1} \quad [3\text{-}31]$$

$$y_t = a + by_{t-1} + x_{t-2} \quad [3\text{-}32]$$

$$y_t = a + by_{t-1} + x_t + x_{t-1} + x_{t-2} \quad [3\text{-}33]$$

Jõumeetodil küsimust lahendades tuleks kõik võimalikud mudeli variandid läbi proovida. Almon (1965) pakkus välja metoodika, mis vähendab laagidega mudeli regressioonikordajate määramise probleemi. Ta näitas, et kui väljendada regressioonikordajaid valemiga

$$b_1 = a_0 + a_1 i + \ldots + a_q i^q; \quad [3\text{-}34]$$

on α väärtusi lihtsam määrata kui b väärtust. Hinnang sõltub paraku polünoomi astmest, mis ei ole *a priori* teada ja mis tuleb järelikult endal otsustada.

3.5.7. Spektraalanalüüs

Spektraalanalüüs ehk Fourier' analüüs tegeleb tsükliliste struktuuridega andmetes, selle eesmärgiks on keeruka struktuuriga aegrea jaotamine eri lainepikkusega sinusoidaalseteks (siinus või koosiinus) funktsioonideks. Spektraalanalüüsi võib võrrelda valge valguse suunamisega läbi klaasprisma. Analoogiliselt prismaga püüab spektraalanalüüs aegrea eri lainepikkusega protsessideks eraldada. Analüüsi kutsutakse Fourier' analüüsiks idee autori Jean Baptiste Fourier (1768–1830) järgi. **Harmooniline analüüs** on spektraalanalüüsi matemaatiline üldistus. Aegrea aluseks olevaid laineid nimetatakse harmoonikuteks (*harmonics*).

Spektraalanalüüs jagab aegrea sinusoidaalseteks struktuurideks

ARMA (ptk 3.5.3) ja eksponentsiaalse silumise (ptk 3.5.4) korral peab tsüklilise komponendi lainepikkus kas teada olema või tuleb see *a priori* otsustada, spektraalanalüüsi puhul ei ole vaja lainepikkust *a priori* teada. Ka lainelise tsükli alguskoht ei ole oluline. Andmed peavad aga olema pidevad, vähemalt ilma suuremate lünkadeta ja igas vaatluskohas peab olema vaid üks vaatlus. Andmete jaotumist ajateljel saab ühtlustada interpoleerimise teel. Spektraalanalüüsis eeldatakse vigade normaaljaotust, aga väikesed kõrvalekalded normaaljaotusest väidetavalt suuri probleeme ei põhjusta (Platt ja Denman 1975). Peale selle eeldab spektraalanalüüs, nagu enamik teisigi aegridade analüüsi meetodeid, et trendid kas puuduvad või on enne analüüsi eemaldatud.

Spektraalanalüüs võib olla **ühe spektri analüüs** (*single spectrum analysis*) või **mitme spektri analüüs** (*cross-spectrum analysis*). Ühe spektri analüüsi eesmärk on aegreast leida ja kirjeldada lainelised struktuurid.

Mitme spektri analüüsi eesmärgiks on leida korrelatsioone erinevate aegridade vahel, näiteks päikese aktiivsuse ja ilmastiku vahel. Kuna mõlema muutuja rida võib sisaldada mitme erineva

sagedusega tsükleid, on korrelatsioonid eri sagedustel erinevad. Mõne sageduse juures tsüklid korreleeruvad, mõne juures mitte. Mitme spektri analüüsi variandiks võib pidada tasapinnalist ehk kahemõõtmelist spektraalanalüüsi, mille lihtsaima variandi korral analüüsitakse mingi tunnuse muutlikkust mitte ajateljel, vaid tasapinna *x* ja *y* teljel.

Aegreas oleva tsüklilise struktuuri lainepikkuse mõõduna kasutatakse tavaliselt sagedust. **Periood** on täistsükliks kuluv aeg. **Faasinihe** (*phase shift*) on mõõt, mis näitab, kui palju üks sageduse komponent teist edestab.

Aegrea spektraaltihedus ehk võimsus ehk intensiivsus arvutatakse siinusfunktsiooni ruut pluss koosinusfunktsiooni ruut korda pool aegrea pikkusest iga sageduse *k* juures. Spektraaltiheduse sõltuvust sagedusest kujutab **periodogramm**. Periodogramm varieerub ka ajas mingi perioodiga, kord on suurem, kord väiksem. Kahe muutuja puhul saab esitada ristperiodogrammi (*cross periodogram*).

Rist-amplituud on kahe muutuja sageduskomponentide vahelise kovariatsiooni mõõt. Rist-amplituude saab standardiseerida neid ruutu võttes ja jagades kummagi üksikspektri tiheduste korrutisega. Tulemust nimetatakse **ristkoherentsiks** ja seda võib interpreteerida analoogiliselt korrelatsioonkordaja ruuduga. Ettevaatust – mõlema seeria väikeste tiheduse väärtuste korral võib saada kahtlaselt suuri ristkoherentse. **Tulu** (*gain*) on rist-amplituudi ja ühe spektri tiheduse jagatis. Tulu on analoogiline regressioonikordajaga ja on erinev mõlema spektri jaoks.

Praktikas esineb sageli olukord, kus üks sagedus lekitab osa varieeruvust naabersagedusele. Selle vastu aitab polsterdamine (*padding*), piiramine (*tapering*) ja silumine (*smoothing*). **Polsterdamine** on ajaskaala ühiku väiksemaks muutmine, mis võimaldab võnkumiste detailsemat kujutamist. **Piiramine** on aegrea alguse ja lõpu vaatluste kaalu muutmine.

Silumine libiseva keskmisega võib toimuda mitmesuguse aknaga laiusega, mis on järgnevalt tähistatud *m*, aken haarab seega kummaltki poolt $p = (m–1)/2$ naabervaatlust. Vaatluse *j* kaalud on tähistatud w_j. Kaalud akna mõlemas pooles on võrdsed, seega kui $j < 0$, siis $w_j = w_{-j}$. Alljärgnevalt on toodud mõned silumisvalemid.

Danielli silumine toimub tasase aknaga. Akna sees olevate väärtuste kaalud on võrdsed.

Tukey akna puhul arvutatakse kaalud iga sageduse jaoks järgmiselt.

$$w_j = 0{,}5 + 0{,}5\cos\left(\frac{\pi \cdot j}{p}\right) \qquad [3\text{-}35]$$

Hammingi akna puhul arvutatakse kaalud iga sageduse jaoks järgmiselt.

$$w_j = 0{,}54 + 0{,}46\cos\left(\frac{\pi \cdot j}{p}\right) \qquad [3\text{-}36]$$

Bartletti akna puhul arvutatakse kaalud iga sageduse jaoks järgmiselt.

$$w_j = 1 - \frac{j}{p} \qquad [3\text{-}37]$$

Aegrea analüüsiks ettevalmistamisel on soovitav kõigepealt aegrida tsentreerida keskmise lahutamisega väärtustest ja eemaldada trend. Kui aegreas trendi ja tsükleid ei ole, siis peaksid aegrea väärtused olema juhuslikud (normaaljaotusega) ja sellist aegrida nimetatakse **valge müra** reaks. Aegrea erinevust valgest mürast saab vaadelda väärtuste histogrammil ja Kolmogorov-Smirnovi *d* statistiku abil.

3.6. Mudeli kalibreerimine ja mudeli hindamine

Mudeli kalibreerimine ehk sobitamine ehk lähendamine andmetele on mudeli parameetritele selliste väärtuste leidmine, mis tagavad mudelist saadud prognooside kõige parema vastavuse õpetusandmetele.

Klassifikatsioonipuu tasemete või regressioonimudeli parameetrite arvu suurendamisel võib nii klassifikatsiooni kui ka regressioonimeetodite puhul saavutada mudeli üks-ühese vastavuse andmetele. Sellisel juhul ei ole mudel tegelikkuse üldistus ja on suhteliselt kasutu. Sobiv klassifikatsioonitasemete või parameetrite arv tuleb otsustada kas mudeli koostajal, või siis kasutada samm-sammulist protseduuri, kus on mudeli minimaalselt vajalik sobivus või üksikfaktori minimaalne olulisus ette antud.

Mudeli empiirilistele andmetele vastavuse kontrollimise kohta kasutatakse termineid hindamine (*evaluation*), tulemuste kontrollimine (*validation*) ja täpsuse hindamine (*accuracy assessment*).

Mudeli hindamisel kasutatakse mitmeid üksikstatistikuid (ptk 3.6.1) ja mudeli usaldusväärsuse hindamisviise:

- ristkontroll ehk tulemuste hinnang sõltumatute kontrollandmestiku järgi, kontrollandmeteks võib olla mudeli sobitamisest kõrvale jäetud lähteandmete osa (ptk 3.6.2),
- prognoosi hajuvuse hinnang üksikvaatluste kordamööda kõrvalejätmisega (jäta-üks-välja meetod, ptk 3.6.4),
- bootstrap (prognoosi hajuvuse hinnang vaatluste korduva tagasipanekuga juhusliku valimise abil, ptk 3.6.5).
- prognoosi juhuslikkuse hindamine juhuslike prognooside jaotuse järgi (nullhüpoteesi ümberlükkamine Monte Carlo meetodiga, ptk 3.6.6).

Eristatakse täpsust õpetusandmetes ehk **õpetustäpsust** (*training accuracy*) ja täpsust sõltumatutes kontrollandmetes ehk **kontrolltäpsust** (*test accuracy*) vahel. Esimesel juhul saadakse mudeli vastavuse hinnang samadest andmetest, mille järgi mudel koostati, teisel juhul on kontrollandmestik õpetusandmestikust sõltumatu.

Kontrolltäpsus arvutatakse andmetest, mida mudeli kalibreerimisel ei kasutatud

Vähimruutude regressiooni puhul nimetatakse prognoosiveaks uue prognoosi oodatavat ruutviga, klassifikatsioonimudeli puhul väärklassifikatsiooni tõenäosust. Prognoosivea hindamise parameetrilised mudelid kipuvad mudeli keerukuse suurenedes prognoosiviga alahindama. Selle vastu aitab kas mudeli keerukusastme arvestamine (näiteks Akaike informatsiooni kriteeriumi abil) või kontroll sõltumatute vaatlustega. Ka nullhüpoteesi meetodit on kritiseeritud, kuna nullhüpoteesi tingimused on enamasti ebareaalsed ja parameetrilised meetodid kalduvad ülehindama suure hulga seletavate tunnustega regressioonimudelite statistilist olulisust (Jones ja Matloff 1986, Berger ja Sellke 1987, Draper 1995, Stewart-Oaten 1996, Fernandez-Duque 1997).

Parima kirjeldava mudeli leidmiseks on kasutatud kahte põhimõtteliselt erinevat teed:

- mudeli statistilise olulisuse hindamine (nullhüpoteesi meetod),
- parima prognoosi meetod (prognoosivea minimeerimine).

Kui mudeli lähteandmed sisaldavad ebamäärasust või juhuslikke vigu, siis kanduvad need vead läbi mudeli prognoositud väärtustesse ebaühtlaselt. Vigade ja hägususe edasikandumine ja võimendumine andmeteisenduste käigus on omaette uurimisvaldkond informaatikas. Nii nagu prognoosegi, saab ka prognoositäpsust esitada graafiliselt – jääkide graafikuna, prognoosi ja vaatlusandmete korrelatsiooni-väljana või vigade paiknemise kaardina. Ruumiliste prognooside korral aitab vigade kartograafiline esitus vigade põhjustele jälile saada ja mudelit parandada.

Mudeli omadusi uuritakse lähemalt **määramatuse** ehk ebakindluse **analüüsiga** (*uncertainty analysis – UA)* ja **tundlikkuse analüüsiga** (*sensitivity analysis – SA*). Ebakindluse analüüs hindab lähteandmete (eba)usaldatavuse edasikandumist mudeli tulemustele. Tundlikkuse analüüs hindab mudeli väljundi sõltuvust lähteandmete varieeruvusest ja mõõtmisvigadest. Tundlikkuse analüüs võimaldab ühelt poolt leida olukordi ja tunnuste kombinatsioone ning väärtuspiirkondi, mis lõpptulemust kõige enam mõjutavad. Teisalt annab tundlikkuse analüüs vihjeid selle kohta, milliseid lähteandmeid oleks vaja kas rohkem või täpsemalt mõõta.

Lisaks tavapärastele vigade allikatele (mõõtmisvead, ebasobiv klassifikatsioon, ebapiisav andmehulk) võivad ruumiandmete vead ja ebamäärasused olla tingitud ruumiandmetele eripärasematest põhjustest: ebasobivast mõõtkavast, asukoha määramise vigadest, nähtuste ruumilisest heterogeensusest ja ruumilisest struktuurist. Ruumiandmete vead jagunevad põhiliselt objektiveaks, atribuudiveaks ja asukohaveaks. Ruumiandmestike usaldatavust mingi uurimuse kontekstis tuleb enamasti eraldi uurida (ptk 5.10).

3.6.1. Mudeli hindamise statistikud

Nominaalse muutuja klassikuuluvuse prognoosi täpsuse kirjeldamise tavapärased vahendid on vigade maatriks ja kapa koefitsient. Neid käsitleti klassifikatsioonimeetodite osas (ptk 2.3.7). Peale kapa saab vigade maatriksist arvutada ka teisi sagedusjaotusi võrdlevaid statistikuid, näiteks hii-ruutu. Pidevate muutujate korral kasutatakse prognoositäpsuse hindamiseks mitmesuguseid prognoosivigade keskmise suuruse näitajaid ning mitmesuguseid vaatluste ja prognooside korreleeruvuse näitajaid.

Vähimruutude meetodil sobitatud mudeli abil äraseletatud hajuvust mõõdetakse determinatsioonikordaja abil. **Suurima tõepära** (*maximum likelihood*) meetod annab suurima tõepära hinnangud, see on parameetrite väärtused, mis vastavad tõepärafunktsiooni maksimumile ehk väärtused, milliste korral selliste andmete saamine on kõige tõenäolisem (vt ka ptk 1.2). Vigade normaaljaotuse korral kuulub ka vähimruutude meetod suurima tõepära meetodite hulka. Vigade teistsuguse jaotuse korral annab vähimruutude meetod nihkega hinnangu ja tuleks eelistada suurima tõepära meetodit, mis on küll arvutusmahukam.

Suurima tõepära meetodil kalibreeritava pideva funktsioontunnusega mudeli seletava võime mõõduna kasutatakse **hälbe vähenemist** (*deviance reduction − D^2*), mis on determinatsioonikordaja (R^2) analoog:

$$D^2 = \frac{ND - RD}{ND}, \qquad [3\text{-}38]$$

kus *ND* on vaid vabaliikmega mudeli hälve ja *RD* on täismudeli jääkhälve.

Hälbe vähenemist saab korrigeerida vaatluste arvuga *n* ja parameetrite arvuga *p* järgmiselt (Weisberg 1980, Guisan ja Zimmermann 2000).

$$D^2_{korrigeeritud} = 1 - \frac{n-1}{n-p} \cdot \left(1 - D^2\right) \qquad [3\text{-}39]$$

Tõepärasuhte statistik (*likelihood ratio statistic – Λ*) on nullhüpoteesi ja sisuka hüpoteesi kehtivuse tõepära suhe, mida kasutatakse logaritmitud kujul. Tõepärasuhet saab väljendada ka logaritmitud tõepärade ehk **log-tõepärade** (*log-likelihood*) vahena

$$\Lambda = -2\ln(\frac{L_0}{L_1}) = -2(\ln L_0 - \ln L_1), \quad [3\text{-}40]$$

kus L_0 on nullmudeli korral oodatav tõepära ja L_1 on alternatiivse mudeli tõepära.

Kui log-tõepärasuhte arvutamise valem algab miinusmärgiga, siis muutub saadud statistik vahemikus 0...∞. Mida suurem on log-tõepära, seda paremini kirjeldab sisukas mudel andmete jaotumist. Tõepärasuhe võimaldab otsustada, kas parameetri(te) lisamine parandas mudelit. Tõepärasuhte olulisust saab kontrollida hii-ruut jaotuse abil. Hii-ruut jaotuse parameetrite arvuks tuleks võtta mudelite parameetrite arvu vahe.

Tõepäraskoor (*likelihood score*) on log-tõepära meetodil mudeli sobivuse hindamise vahend, mis kasutab tõepärafunktsiooni maksimumi otsimisel selle tõusunurka.

Waldi statistik (*Wald statistic*) on suurima tõepära meetodil saadud mudeli parameetri hinnangu statistilise olulisuse kontrollimise vahend. Waldi statistik arvutatakse parameetri hinnangu ruudu ja parameetri hinnangute asümptootilise dispersiooni suhtena ning selle olulisust kontrollitakse hii-ruut testiga.

Akaike informatsioonikriteerium on eelmisele lähedane mudelite tõhususe võrdlemise vahend, mis arvestab mudeli keerukust parameetrite arvu (q) järgi kahel viisil.

$$AIC = 2[\ln(L) + q] \quad [3\text{-}41]$$

$$AIC = -2[\ln(L) - q] \quad [3\text{-}42]$$

Wilksi λ on determinatsioonikordaja mitmefaktoriline analoog, mis näitab argumenttunnuste abil ärakirjeldatud varieeruvuse osa; muutumispiirkond 0…1 (1 tähendab seose puudumist).

Hajuvuse jaotamine (*variance partitioning – VPA*) selgitab, milline osa funktsioontunnuse väärtuste dispersioonist on selgitatav iga üksiku argumenttunnuse abil, kui suur osa erinevate argumenttunnuste poolt seletatud hajuvuse osast on omavahel kattuv, kui suurt osa hajuvusest seletab argumenttunnuste kooskasutus ning kui suur osa hajuvusest jääb seletamata. Hierarhiline hajuvuse jaotumise analüüs (*hierarchical partitioning*) (Chevan ja Sutherland 1991) võrdleb hajuvuse jagamisel argumenttunnuste kõiki võimalikke kombinatsioone, mitte vaid ühte nagu lihtne hajuvuse jaotumise analüüs.

3.6.2. Ristkontroll

Ristkontrolliks (*cross-validation*) ehk ristvalideerimiseks nimetatakse mudeli testimist kontrollandmetega, mida mudeli koostamisel ei kasutatud. Kontrollimiseks kasutatav andmete osa võib olla sõltumatu valim või mudeli sobitamiseks kasutatud valimist välja jäetud andmete osa. Mudel on kõige paremini sobitatud siis, kui see vastab nii kontrollandmetele kui ka õpetusandmetele.

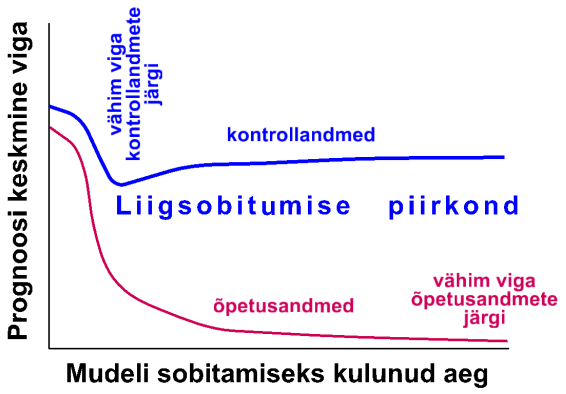
Ristkontroll võib olla:

- ühekordne (*holdout validation*) ehk ühe kontrollvalimiga, milles olevaid vaatlusi õpetusandmetena ei kasutata;
- kahekordne (*two-fold cross-validation*), mille puhul mudel kalibreeritakse ja kontrollitakse kaks korda: kontrollvalim ja õpetusvalim vahetavad oma rolle;
- mitmekordne ristkontroll (*V-fold cross-validation*, *k-fold cross-validation*) on rohkem kui kahe üksteist välistava valimi kordamööda kõrvalejätmine,
- vaatluste ükshaaval kõrvalejätmine (*leave-one-out cross-validation – LOOC*) võimaldab suurima mahuga õpetusandmestikku ja kontrollandmestikku. Ristkontrollile lähedased on tagasipanekuga juhuvalimeid moodustavad meetodid – bootstrap (ptk 3.6.5) ja jack-knife (ptk 3.6.4).

Kui kasutatud vaatlused on esinduslikud ja sõltumatud, siis võib põhjendatult loota, et mudeli prognoosid on kogu uuritaval alal enam-vähem sama hea paikapidavusega. Ristkontroll on eriti vajalik nende meetodite kasutamisel, mis kalduvad andma näiliselt häid prognoose liigsobitatud mudelist, näiteks intellektitehnika ja andmekaevandamise meetodid. **Ülesobitumine** ehk liigsobitumine (*overfitting*) on olukord, milleni jõutakse liiga keerulise mudeli sobitamisel andmetele. Ülesobitatud mudel annab õpetusvalimis tunduvalt täpsema prognoosi kui väljaspool õpetusandmeid. Tehisõppe kontekstis sõltub mudeli keerukus sobiva mudeli otsimisele kulutatud ajast (joonis 3-12).

Termineid **kontrollvalim** ja **kontrollandmed** võiks eelistada terminitele testvalim ja testandmed, sest sõna "test" tähendus on hägusem kui sõna "kontroll" tähendus. Testandmeteks on nimetatud igasuguseid katseandmeid ja näidisandmeid.

Sõltumatu kontrollandmestik aitab vältida mudeli ülesobitumist



Joonis 3-12. Mudeli vastavus õpetusandmetele ja kontrollandmetele sõltuvalt mudeli õpetusandmetele sobitamisele kulutatud ajast. Liigsobitumine tekib mudeli järjest keskendumisel etteantud õpetusandmetele, mille tõttu väheneb mudeli üldistusvõime ja prognoositäpsus väljaspool õpetusandmeid.

3.6.3. Tulemuslikkuse kõverad

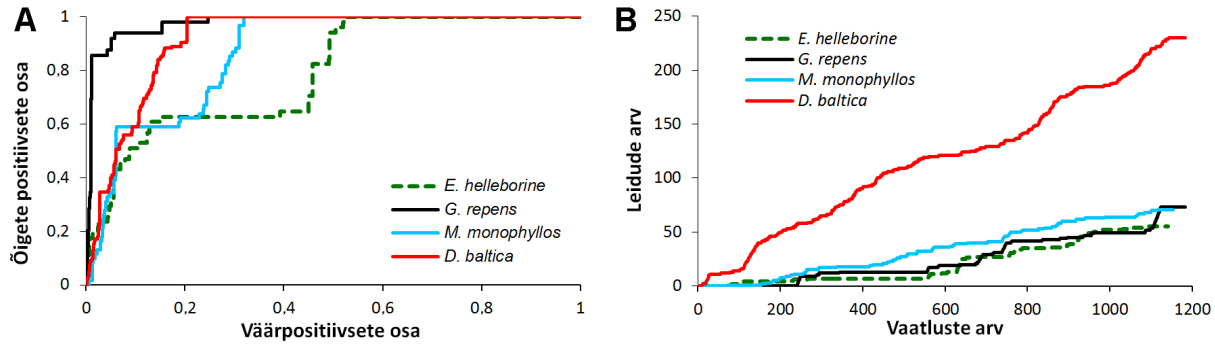
3.6.3.1 Toimimiskõver

Toimimiskõver ehk ROC-kõver ehk ROC-graafik [*receiver* (*relative*) *operating characteristic* – *ROC*] ja toimimiskõvera alune pind võimaldavad hinnata kaheväärtuselisi tulemusi andva mudeli prognoosivõimet erinevate kriitiliste sageduste korral (Hanley ja McNeil 1982, 1983, Zweig ja Campbell 1993, Murtaugh 1996, Bradley 1997, Peterson *et al.* 2008). Toimimiskõvera vertikaalteljel

kujutatakse mudeli **tundlikkust** ehk õigesti prognoositud esinemisjuhtude (*true positive*) osakaalu tegelike esinemisjuhtude hulgast, horisontaalteljel on **ekslikkus** ehk vääralt esinemiseks klassifitseeritud juhtude (*false positive* = 1 – *true negative*) osakaal puudumiskohtade hulgas. Kui andmetes on võrdse tõenäosusega üksuseid, millest osa on esinemisjuhud ja osa puudumisjuhud, eriti kui selliseid üksusi on palju, siis sõltub toimimiskõvera kuju sama väärtusega vaatluste järjestamisest. ROC-graafiku telgedel on osakaalud kõigist vaatlustest, mis eeldab kõigi vaatluste teadmist.

Näiteks Elva ja Otepää vahel paiknevatelt kaardilehtedelt 5434, 5444 ja 5454 kogutud andmete järgi on Russowi sõrmkäpa (*D. russowii*) ja roomava öövilke (*G. repens*) esinemise/puudumise hinnang on kõige tõepärasem (ROC-kõvera alune pind on suurim), laialehise neiuvaiba (*E. helleborine*), soovalgu (*M. monophyllos*) ja hariliku käoraamatu (*G. conopsea*) esinemise/puudumise hinnangud on uuritud liikidest kõige vähem usaldusväärsed. Otsustustaseme nihutamine väiksema tõenäosuse poole suurendaks järsult liikide *D. fuchsii*, *G. conopsea* ja *E. helleborine* puhul vääralt esinemiskohtadeks klassifitseeritud vaatluste osakaalu. Soovalgu puhul oleks paar protsenti rangema esinemiskohaks tunnistamise puhul vastavuse koondnäitaja kõrgem (joonis 3-13A).

Toimimiskõvera vertikaalteljel on õigete positiivsete osa, horisontaalteljel väärpositiivsete osa kõigist vaatlustest



Joonis 3-13. Nelja käpaliseliigi vaadeldud esinemise/puudumise ning koha sarnasuse järgi arvutatud hinnangu vastavus Otepää ja Elva vahel paiknevate kaardilehtede 5434, 5444 ja 5454 andmetest seisuga jaanuar 2008 toimimiskõverana (A) ja tulujoonena (B). Toimimiskõver on koostatud hinnangu otsusekindluse kahanevas järjekorras. Toimimiskõvera alused pinnad on eelmainitud liikidel järgmised: *Dactylorhiz. baltica* – 0,917; *Goodyera repens* – 0,978; *Epipactis helleborine* – 0,790; *Malaxis monophyllos* – 0,864.Tulukõver on koostatud vaatlustulemuste laekumise järjekorras. Kuna tulukõveratel pole langustendentsi märgata, siis tasub välivaatlusi sellel uurimisalal jätkata ka pärast tuhande vaatlustulemuse registreerimist.

Toimimiskõvera alla jääva ala pindala (*area under curve* – *AUC*) ehk **toimimispind** on mudeli diagnoosiva võime näitaja, mis ühendab hinnangute täpsust tundlikkuse kõigi tasemete korral. See hindab tõenäosust, et esinemise ja puudumise andmetest juhuslikult valitud vaatluste paari puhul annab mudel suurema esinemise tõenäosuse sellele vaatlusele, kus nähtus tegelikult esineb. AUC = 1 viitab mudeli absoluutsele täpsusele, AUC = 0,5 vastab prognoosi juhuslikkusele ning väärtused <0,5 viitavad tegelikkusega vastuolus olevate prognooside ülekaalule. ROC-graafiku alune pind ei sõltu positiivsete vaatluste oodatavast sagedusest.

ROC kõvera alune pind on kaheväärtuselise muutuja hinnangute täpsuse koondnäitaja, mis ei sõltu mudeli tundlikkuse üksiktasemest

ROC-kõvera alust pindala saab arvutada ka diagonaali ehk juhusliku prognoosi korral oodatava suhtes. Seda tuleks nimetada suhteliseks kõvera aluseks pindalaks (*relative area under curve – RAUC*) ehk **suhteliseks toimimispinnaks**. RAUC puhul vastaks nullväärtus hinnangu juhuslikkusele ehk tegelike positiivsete ja negatiivsete juhtude ühtlasele jaotumisele kõigi prognoositud tõenäosuste korral ning suurima võimaliku mittevastavuse korral oleks RAUC = –1. AUC ja RAUC on vastastikku teisendatavad:

$$AUC = \frac{1 + RAUC}{2}, \qquad [3\text{-}43]$$

$$RAUC = 2AUC - 1. \qquad [3\text{-}44]$$

ROC-kõvera aluse pindala arvutamine eeltoodud RAUC-kujul ei ole siiski tavaks – ilmselt ei peeta isegi suhtelise pindala puhul negatiivseid väärtusi kohasteks. Klassifikatsioonitäpsuse hindajana ei eelda ROC-kõvera alune pind, erinevalt kapa kordajast ja Hanssen-Kuiperi skoorist, tulemuste jagamist klassidesse – kasutatakse positiivse variandi tõenäosust.

Peterson *et al.* (2008) kirjeldavad liikide leviku hinnangulise kaardistamise mudelitele sobitatud **osalist ROC analüüsi** (*partial ROC analysis*), milles esiteks ei ole ROC graafiku horisontaalteljel väärpositiivsete vaatluste osa, vaid liigi esinemisalaks prognoositud pinna osa. See võib olla põhjendatud vaid esinemiskohti kasutavate algoritmide puhul, mille puhul liigi esinemist modelleeritakse taustaandmete ehk kogu uurimisala suhtes. Teiseks kasutatakse toimimiskõvera aluse pinna vaid seda osa, mille ulatuses algoritm prognoose annab. Kasutada tasub vaid seda toimimiskõvera vahemikku, mis vastab mudeli tundlikkusele ning mille puhul esinemisalaks prognoositud ala osa on suurem kui null ja väiksem kui kogu uurimisala. Lisaks sellele võib kasutaja määrata õigete positiivsete osakaalu nõutava taseme. Kui kuni selle tasemeni mudelit niikuinii ei rakendata, siis tuleks see osa AUC võrdlustest välja jätta.

ROC-kõvera konstrueerimiseks on tarvis järgmist:

- leida positiivsete ja negatiivsete juhtude arv andmestikus,
- järjestada vaatlused uuritava nähtuse esinemistõenäosuse kahanevas järjekorras,
- arvutada iga vaatluse juures selle ja eelnenud vaatluste positiivsete vaatluse arv ja negatiivsete vaatluse arv,
- jagada positiivsete juhtude arv positiivsete juhtude koguarvuga ning negatiivsete juhtude arv negatiivsete juhtude koguarvuga iga vaatluse juures,
- kujundada graafik, mille vertikaalteljel oleks osakaal õigesti positiivseks hinnatud juhtudest ja horisontaalteljel osakaal õigesti negatiivseks hinnatud juhtudest.

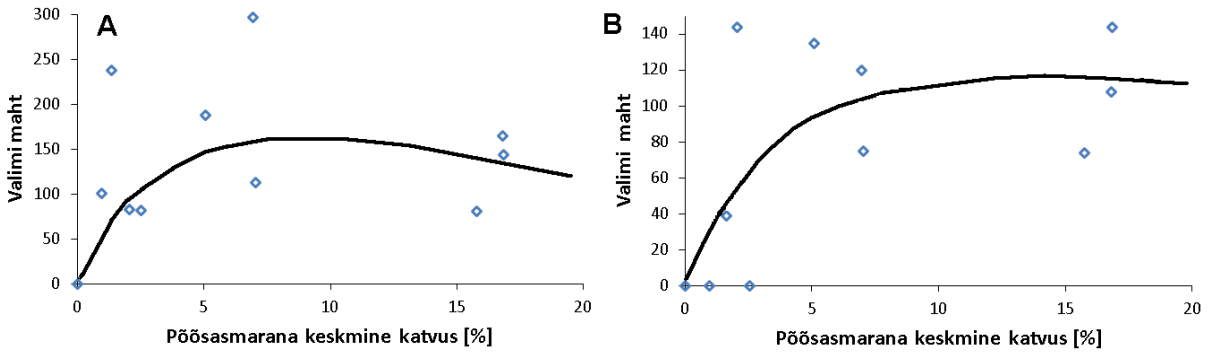
Kõvera aluse suhtelise pinna arvutamiseks kasutatakse kas parameetrilist lähendamist või pinna trapetsoidideks jagamist.

3.6.3.2 Teised tulemuslikkuse kõverad

Toimimiskõverale lähedane meetod on **tulujoon** (*gain chart* ehk *lift chart* ehk *lift curve*), mis näitab positiivsete vastuste osa muutumist vastanute arvu muutudes (joonis 3-13B). Tulujoone graafiku vertikaalteljel võib olla ka positiivsete vastuste hulk või muu reaalarvuline tulemuslikkuse näitaja. Ka horisontaalteljel võib olla kas vaatluste hulk, katsete arv või osakaal vaatluste või katsete koguarvust.

Vajalike vaatluste keskmise arvu kõver (*average sample number curve, average sample function – ASN*) ehk **vaatlusmahukõver** näitab järjestikuse analüüsi (*sequential analysis*) korral, mitu vaatlust

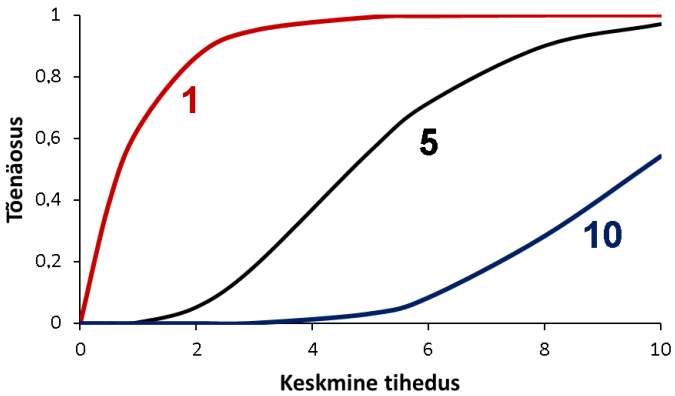
(valimit, analüüsi, mõõtmist) või kui suurt valimit on keskmiselt tarvis, et saaks teha etteantud kindlusega järelduse. Vaatlusmahukõvera graafiku horisontaalteljel on mingi muutuja, millest vajalike vaatluste arv sõltub, reeglina uuritava nähtuse keskmine sagedus. Elusloodusе uuringutes võib selliseks muutujaks olla populatsiooni keskmine tihedus või katvus (joonis 3-14). Siis näitab ASN kõver kindla suurusega prooviruutude oodatavalt vajalikku arvu, määramaks populatsiooni keskmist tihedust kümneprotsendilise suhtelise veaga.



Joonis 3-14. Vaatluste hulk, mis on vajalik põõsasmarana keskmise katvuse määramiseks etteantud täpsusega (A – suhteline viga 20%, B – absoluutviga 5%) ja selle alusel koostatud empiirilised vajaliku vaatlusmahu kõverad. Õigeks loetav keskmine on arvutatud kõigist olemasolevatest vaatlustest samal kaardilehel. 5–20% keskmise katvuse puhul on selle mõõtmiseks vajalik vaatluste arv suur, sest katvuste varieeruvus on suhteliselt suur. Ühtlaselt väikese ja ühtlaselt lausalise katvuse puhul on selle varieeruvus vaatluskohtade vahel väike ja keskmise määramiseks piisab väiksemast vaatluste arvust. Andmed Kalle Remm 2008-2011.

Vaatlusmahukõveraid koostatakse kas eeldades mingit teoreetilist jaotust või tehes korduvaid väljavõtteid empiirilise jaotusega (paiknemismustriga) andmetest või sõltumatute kontrollandmete järgi. ASN kõverad aitavad koostada ja optimeerida katse plaani – seada nõutav hinnangute täpsus vastavusse mõistliku töömahuga. Nõutava täpsuse suurendamiseks vajalik töö maht võib olla ebareaalselt suur. Kuna oodatav keskmine ei ole täpselt ette teada, siis tuleb vaatlusmahu kõverat töö käigus korrigeerida.

Statistilises kvaliteedikontrollis ja järjestikuse analüüsi teoorias esindab **tulemusfunktsiooni** (*operating characteristic function*) termin aktsepteerimistõenäosuse sõltuvust defektide sagedusest. Tulemusfunktsiooni graafiline esitus on **tulemuskõver** (*operating characteristic curve – OC*). Tüüpiline tulemuskõver on mingis osas järsemalt langev. Defektide väikese ja suure sageduse korral on otsus enamasti vastavalt kas positiivne või negatiivne, kõvera langusele vastava defektide hulga juures positiivse otsuse tõenäosus väheneb järsult. Teoreetiline tulemuskõver arvutatakse Poissoni jaotuse abil. Kõvera kuju sõltub valimi mahust, üldkogumi mahust ja lubatud defektide või muude sündmuste hulgast (joonis 3-15).



Joonis 3-15. Vähemalt ühe, viie ja kümne põdra esinemise tõenäosus juhuslikult valitud ruukilomeetril sõltuvalt põtrade keskmisest tihedusest ja eeldades põtrade juhuslikku paiknemist.

Välivaatluste analüüsil saab tulemuskõveraga kujutada näiteks liigi leidmise tõenäosuse ja vaatluste hulga või vaatluse kestuse seost. Vajalike vaatluste arvu kõverat ja tulemuskõverat on kasutatud taimekahjurite jälgimise planeerimisel metsanduses, aianduses ja põllunduses (Binns *et al*. 2000, Koch *et al*. 2006, Nyrop ja Binns 1992).

3.6.4. Liigendnoa-meetod

Jack-knife ehk **liigendnoa-meetodi** kasutamisel arvutatakse statistiku väärtus järjepanu üksikuid vaatlusi andmestikust välja jättes ja iga kord statistikut uuesti arvutades. Mudeli täpsuse hindamisel sobitatakse mudelit korduvalt igat vaatlust ükshaaval eemaldades ja saadakse ühe vaatluse eemaldamise keskmise mõju suuruse hinnang. Ideaaljuhul ei tohiks ühe vaatluse eemaldamine mudeli parameetrite hinnanguid mõjutada.

Meetodi nimi viitab igaks tööks vajalike vahenditega taskunoa universaalsusele ja samas ka asjaolule, et taskunuga on vaid lihtne hädapärane asendaja parema tööriista puudumisel. *Jack of all trades* tähistab meest, kes tuleb toime iga tööga (aga ei ole ühelgi alal tipptegija). Liigendnoa meetodit kasutatakse nii muutuja hajuvust mõõtvate statistikute väärtuste määramisel kui ka mudelist saadud prognooside täpsuse hindamisel.

Enamasti kasutatakse *jack-knife*-protseduuri, mille igas korduses jäetakse arvutustest järjepanu välja vaid üks vaatlus. Mitmekaupa väljajätmisel tuleks kasutada kõiki kombinatsioone väljajäetavatest vaatlustest, mis suurendab arvutusmahtu. Igas korduses vaid ühe vaatluse väljajätmisel nimetatakse *jack-knife*-protseduuri ka jäta-üks-välja (*leave-one-out*) meetodiks.

Liigendnoa meetod jätab vaatlustulemusi ükshaaval arvutusest välja ning hindab mudeli stabiilsust üksikute vaatluste mõju järgi

Liigendnoa-meetodi kasutamisel arvutatakse iga vaatluse eemaldamise järel hinnatava parameetri P pseudoväärtus S

$$S_i = nP' - (n-1)P''_{-i}, \qquad [3\text{-}45]$$

kus n on valimi suurus, P' on parameetri hinnang kõigi vaatluste järgi, P''_{-i} on parameetri hinnang kõigi vaatluste järgi, välja arvatud vaatlus i.

Kõigi vaatluste ükshaaval väljajätmise tulemusel saadakse n pseudoväärtust, mille keskmine on muutuja *jack-knife*-hinnang. Hinnangu dispersioon on

$$Var(P) = \frac{\sum_i^n \left(S_i - P''_{-i}\right)^2}{n(n-1)}. \qquad [3\text{-}46]$$

Hinnangu dispersiooni ja t jaotuse abil saab leida hinnangu usalduspiirid

$$\pm t_{\alpha/2,n-1}\sqrt{Var(P)}. \qquad [3\text{-}47]$$

3.6.5. Bootstrap

Bootstrap hinnangu saamiseks võetakse algsest valimist *k* korda *n* tagasipanekuga vaatlust ja arvutatakse nendest *k* korda parameetri hinnang (*bootstrap replicate*). Hinnatav parameeter võib olla näiteks prognoosi standardhälve või usalduspiirid. Bootstrap on soovitatud meetod, kui hinnatava parameetri jaotus on keerukas või ei ole teada ja kui valimi maht ei ole täpsemate meetodite kasutamiseks piisav.

Bootstrap meetodis saadakse parameetrite hinnangud korduvate tagasipanekuga valimite abil

Bootstrap hinnanguks võetakse üksikhinnangute keskväärtus. Kui uuritava statistiku arvutuskäik on piisavalt lihtne ja arvutusvõimsus võimaldab läbi arvutada väga palju kordusi ($k > 1000$), saab hinnangu usalduspiirid leida kvantiilide meetodil. Näiteks 95% usalduspiiride jaoks leitakse saadud bootstrap hinnangute jaotuse 2,5% ja 97,5% kvantiilid.

Sõna *bootstrap* tähistab inglise keeles saapa taga olevat aasa, mille abil on lihtsam saabast jalga tõmmata. Bootstrap meetod on analoogiline iseenda ülestõstmisega sikutades saapa-aasadest.

3.6.6. Monte Carlo meetod

Monte Carlo meetod on vahend keerukate protsesside modelleerimiseks juhuslike arvude abil. Protsessi võimalike tulemuste valim saadakse korduvate juhuslike jäljenduste abil. Tulemuse variantide sageduse järgi valimis tehakse järeldusi vastavate tulemuste tõenäosuse kohta. Monte Carlo meetod on väga lähedane randomisatsioonitestile (permutatsioonitestile). Kaks põhilist erinevust on:

- Monte Carlo meetod on kasutatav ka pidevate jaotuste korral, randomisatsioonitest eeldab diskreetseid üksusi või vaatlusi mida saab ümber paigutada;
- Monte Carlo meetod korral on võimalik luua piiramatu arv empiirilisele jaotusele vastavaid juhuslikke jäljendusi, mille hulka võib juhuslikult sattuda ka omavahel identseid jäljendusi, randomisatsioonitesti korral on võimalike juhuslike ümberpaigutuste arv piiratud.

Monte Carlo meetodi kasutusalasid tutvustatakse veel punktmustrite statistiliste testide (ptk 4.1.4.2), pindade vastavuse hindamise (ptk 4.3.3.1) ja ruumilise korrelatsiooni teema juures (ptk 5.4.2).

Randomisatsiooni eelis tulemuste usaldusväärsuse hindamisel on tema otsesus. Iteratiivse randomisatsiooni tulemus näitab otseselt, kui sageli jõutakse puht-juhuslikult empiirilistest andmetest saadud tulemuseni või etteantud kriitilise piirini. Kasutatavate juhuslike jäljenduste arvu määrab testi kasutaja.

Monte Carlo meetod mõõdab statistilist olulisust otseselt, teoreetilise jaotuse parameetreid kasutamata

Monte Carlo meetodi on kasutatav olukorras, kus teststatistiku jaotus ei ole teada. Üks või teine statistik läheneb nullhüpoteesi kehtimise korral asümptootiliselt teoreetilisele jaotusele. Erinevad statistikud võivad anda samadest andmetest erinevaid tulemusi. Kui täpne on ühel või teisel juhul olulisuse hinnang, sõltub testi võimsusest ja eelduste paikapidavusest. Üksiku testi tulemuste usaldusväärsuse puhul saab rääkida vaid tõenäosusest. Teststatistikute alusel antavad hinnangud on üsna ebatäpsed, kui andmete hulk on väike.

Monte Carlo meetod on universaalne – see ei sea mingeid eeldusi andmete jaotuse kohta. Monte

Carlo meetodit saab kasutada ka väikeste andmemahtude korral. Näiteks 7 vaatlusest ja 9 vaatlusest koosneva valimi võrdlemiseks võib mõlemad valimid ühendada. Saadakse 16 vaatlusest koosnev koondvalim, mille jagamiseks sama suurteks valimiteks on 11440 varianti. Võttes neist juhuslikke variante, saab kontrollida, kui sageli tekib algvalimites olnud erinevus puhtjuhuslikult. Otsustada on tarvis vaid nullhüpoteesile vastava juhuslikkuse piirid.

Monte Carlo meetodi puhul ei pruugi nullhüpoteesiks olla täielik juhuslikkus

Monte Carlo meetod võimaldab juhuslikkuse genereerimisel arvestada teadaolevat mittejuhuslikkust ja punktprotsessi ebaühtlust. See on oluline näiteks paiknemisseose olulisuse testimisel, kus paiknemisstatistikuid võrreldakse juhusliku paiknemise järgi oodatutega. Nähtustele sisemiselt omane struktuur, näiteks üksikliigi paiknemise laigulisus liikide paiknemise võrdlemisel, tuleks Monte Carlo jäljendustes säilitada. Arvestada saab vaid nähtustele omaste struktuuride teadaolevaid parameetreid. Statistilised testid lähtuvad teatud teoreetilistest eeldustest andmete jaotuse kohta. Üks põhiline eeldus on seejuures üksikvaatluste sõltumatus, mis paraku reaalsete väliuuringute puhul ei ole sageli täidetud. Eelpoolmainitud mittejuhuslikkus võib olla ka üksikvaatluste osaline või täielik seotus.

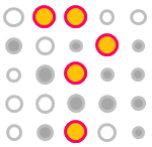
Monte Carlo meetodi puudusena mainitakse eelkõige, et tulemusi ei saada kokkuvõtlikul analüütilisel kujul ning suurt vajalikku arvutusmahtu – varieeruvuse ja usalduspiiride usaldusväärsete hinnangute saamiseks on tarvis kümneid tuhandeid kordusi (Heuvelink 1998). Randomisatsioonides võrreldava teststatistiku valiku subjektiivsus on Monte Carlo meetodi teine nõrk koht – uurijal on võimalik valida endale meelepärasemat tulemust andev statistik paljude alternatiivsete hulgast (Diggle 1983).

3.7. Analüüsimeetodi valik

Loodetavasti aitab sobivat andmete analüüsi või modelleerimise vahendit leida alltoodud tabel (tabel 5).

Tabel 5. Erinevat tüüpi andmete analüüsiks sobivad meetodid.

<table>
<tr><td rowspan="2">ANALÜÜS</td><td colspan="4">ANDMETÜÜP</td></tr>
<tr><td>Nominaalne mittejärjestatav</td><td>Nominaalne järjestatav</td><td>Arvuline mitte-normaaljaotusega</td><td>Arvuline normaaljaotusega</td></tr>
<tr><td>Ühe andmekogumi kirjeldamine</td><td>Sagedustabel</td><td>Mood, mediaan, miinimum, maksimum</td><td>Mediaan, kvantiilid, ekstsess, asümmeetriakordaja, miinimum, maksimum</td><td>Keskväärtus, standardhälve</td></tr>
<tr><td rowspan="2">Kahe andmekogumi erinevuse kontroll</td><td></td><td colspan="2">Kolmogorov-Smirnovi test, märgitest (sõltuvad valimid), iteratsioonitest (sõltumatud valimid).</td><td rowspan="2">Z test (suurte valimite keskväärtuste erinevus), t test (väikeste valimite keskväärtuste erinevus), F test (dispersioonide erinevus)</td></tr>
<tr><td colspan="2">χ^2 test sagedustabelist</td><td>U test
W test</td></tr>
<tr><td>Mitme andmekogumi erinevuse kontroll</td><td colspan="3">Kruskal-Wallise test, Scheffé test, Newman-Keulsi test, Tukey HSD test, Spjotvoll/Stoline test, Bounferroni test</td><td>Dispersioonanalüüs</td></tr>
<tr><td rowspan="2">Tunnuste-vaheliste seoste kirjeldamine (modelleerimine)</td><td colspan="4">Graafik, nomogramm</td></tr>
<tr><td colspan="3">Oodatavate sageduste tabel, üldistatud lineaarsed mudelid</td><td>Regressioonanalüüs, üldised lineaarsed mudelid</td></tr>
<tr><td rowspan="3">Ühe tunnuse prognoos teiste tunnuste järgi</td><td colspan="4">Graafik, nomogramm, tehisnärvivõrgud, vähimruutude osaregressioon, otsuste puu.</td></tr>
<tr><td colspan="2">Log-lineaarsed mudelid, klassifikatsioonipuud</td><td></td><td rowspan="2">Regressioonivõrrand, regressioonijoon, üldised lineaarsed mudelid</td></tr>
<tr><td colspan="3">Üldistatud lineaarsed mudelid</td></tr>
<tr><td colspan="4">Sarnasusele tuginev järeldamine</td></tr>
<tr><td>Tunnuste-vaheliste seoste tugevuse võrdlemine</td><td>Tšuprovi T, Guttmani λ</td><td colspan="2">Spearmani ρ, Kendalli τ, Friedmani ANOVA</td><td>Korrelatsioonikordaja</td></tr>
<tr><td>Vaatluste või tunnuste rühmitamine sarnasuse järgi</td><td colspan="3">Klasteranalüüs, klassifikatsioonipuud</td><td>Diskriminantanalüüs</td></tr>
<tr><td>Üldistatud faktorite või tunnusruumi dimensioonide otsimine ja kirjeldamine</td><td colspan="3">Vastavusanalüüs (lähtub vaatluste ja tunnuste sageduste risttabelist, mis võib sisaldada tunnuste kaale), mitmemõõtmeline skaleerimine (lähtub sarnasuste või kauguste maatriksist)</td><td>Faktoranalüüs, peakomponentanalüüs, kanooniline korrelatsioonanalüüs</td></tr>
<tr><td>Ajaliste muutuste trendid</td><td colspan="4">Aegridade analüüs</td></tr>
</table>



Küsimused

1. Mis on Occami habemenuga?

2. Mis on parsimooniareegel?

3. Mis on mudeli kalibreerimise tulemus?

4. Milliste andmete modelleerimiseks sobib logitmudel?

5. Kas lineaarsetes regressioonimudelites saab kasutada nominaalseid argumenttunnuseid? Kui jah, siis millistes?

6. Mida näitab regressioonikordaja?

7. Mida näitab regressioonivõrrandi vabaliige?

8. Mida peaks tegema lünklike andmetega enne regressioonanalüüsi?

9. Mida peaks ette võtma, kui regressioonijäägid ilmselt ei ole normaaljaotusega?

10. Millised kolm tunnust iseloomustavad laiska probleemikäsitlust?

11. Mille poolest erineb *k-means* meetod k-NN meetodist?

12. Millal sobib mudeli täpsuse hindamiseks ruutkeskmine viga?

13. Milliseid näidiseid kasutatakse k-lähima naabri meetodi puhul?

14. Milliseid näidiseid kasutatakse d-lähima naabri meetodi puhul?

15. Millised on aja võimalikud rollid statistilistes mudelites?

16. Mis on mudeli liigsobitamine?

17. Mis on treeningtäpsus ja mis on kontrolltäpsus?

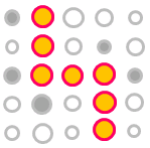
18. Mis on kontrollandmestik ja mis on õpetusandmestik?

19. Mis on mudeli ristkontroll?

20. Millise muutuja mudeli täpsuse hindamiseks sobib kapa kordaja?

21. Mis muutujad on toimimiskõvera ehk ROC-kõvera telgedel?

4. Paiknemise kirjeldamine



Selles peatükis käsitletakse punktide, joonte ja pindade ruumis paiknemise struktuuri ehk **ruumimustri** (*spatial pattern*) kirjeldamist, mida nimetatakse ka kirjeldavaks ruumiandmete analüüsiks (*Exploratory Spatial Data Analysis – ESDA*). Paiknemise struktuur ja ruumilised seosed võivad omavahel kõikvõimalikel viisidel kombineeruda. Ka paljud eluteaduste üldistused eeldavad ruumilise struktuuri olemasolu. Näiteks kogu biogeograafia eeldab paiknemise käsitlust, metapopulatsioonide dünaamika tegeleb aeg-ruumiliste mustrite uurimisega ning looduskaitse planeerimine ei saa hakkama ilma elupaikade kaardita ja selle kaardi analüüsita.

Paiknemist saab iseloomustada nii raster- kui ka vektorkujul andmete järgi – paiknemise näitajaid saab arvutada on nii objektidel kui ka väljadel. Küsimuse, kas tegeletakse ruumi ja selle omadustega või ruumis paiknevate kindlapiiriliste objektidega, on H. Couclelis (1992) kokku võtnud fraasi: inimesed käsitsevad objekte, kuid harivad välju (*people manipulate objects but cultivate fields*).

Konkreetseid paiknemismustreid jagatakse mustri tüüpidesse. Mustri tüüp iseloomustab objektide paiknemist üksteise suhtes, kuid ei sisalda teavet mustri suuna ega mõõtkava kohta. Mustri pööramine ja nihutamine mustri tüüpi ei muuda, suurendamine ja vähendamine üldjuhul ka mitte. Mustri piiride muutmine muudab sageli ka mustri tüüpi (vt näiteks joonis 4-1).

Andmed looduse kohta sisaldavad ühel või teisel viisil ruumidimensiooni. Objekte jagatakse vastavalt nende dimensionaalsusele punktideks (0D), joonteks (1D), pindadeks (2D) ja kehadeks (3D). Mustri substraat on vastavalt dimensionaalsusele kas 1D (joon, telg, tunnuse muutumisskaala), 2D (pind), 3D (ruum), 4D (aegruum). Neljast kõrgem dimensionaalsus on tavaline tunnusruumis. Tunnusruumi telgedeks võivad olla kas mõõdetud tunnused (parameetrid) või abstraktsed faktorid. Viimased kujutavad endast mõõdetud tunnuste mingit kombinatsiooni või teisendust.

Maavarade otsimisel on geoloogiliste objektide paiknemise seaduspärasuste teadmine abiks maardlate leidmisel, meteoroloogias tuleb tunda atmosfääri osade ja seisundite ruumilise struktuuri seaduspärasusi. Ökoloogia peamine eesmärk on seoste ja protsesside tundmaõppimine. Paiknemise kirjeldamine võib olla omaette eesmärgiks, aga ökoloogiliste uurimuste kaugem siht on siiski looduses toimuvate protsesside kirjeldamine ja võimalike arengute ettenägemine. Paiknemismuster on nii protsesside mõjutaja kui ka protsesside tulemus. Paiknemismustri järgi püütakse teha järeldusi, millised protsessid on selle mustri loonud ja millised on mustri komponentide omavahelised mõjud. Ruumimustri kaardistamine on enamasti palju lihtsam kui mustrit mõjutavate protsesside mõõtmine. McIntire ja Fajardo (2009) nimetavad ruumi kaardistamist protsessi kirjeldamise aseaineks (*space as a surrogate*). Ruumimustri kirjeldamise kui aseaine edukus sõltub kolmest komponendist:

- *a priori* hüpoteesi oskuslik seadmine,
- uuritava protsessi teooria või eelneva kogemuse olemasolu,
- sobivate ruumilise analüüsi meetodite kasutamine.

Paiknemismustreid uuritakse nende kujunemist mõjutanud protsesside ja mustri osade omavaheliste mõjude selgitamiseks ning teatud tüüpi struktuuride leidmiseks

Ülevaate ja võrdluse ruumiliste ökoloogiliste andmete üldistavaks kirjeldamiseks, modelleerimiseks ja paiknemismustriga seotud hüpoteeside kontrollimiseks kasutatavatest statistilisel meetoditest on esitanud Perry *et al.* (2002).

4.1. Punktmustrid

Punktmustrit käsitletakse kui objekti asukohta määrava protsessi ehk **punktprotsessi** tulemust või vähemalt seostatakse mustrit tõenäoliselt mõjutavate protsessidega. Termin punktprotsess ei viita protsessi iseloomule või põhjustele, need võivad olla erinevad, vaid objektide tüübile – neid käsitletakse punktidena.

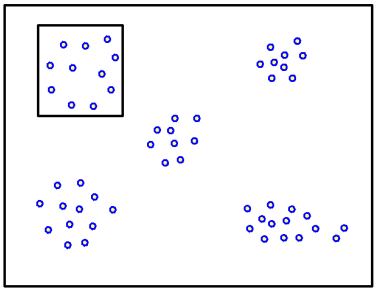
Üksikpunkte nimetatakse ruumistatistikas sündmusteks (*events*) või kohtadeks (*locations*). Sündmused võivad olla kõik võrdväärsed, aga need võivad oma omadustelt ka erineda. Objekti või sündmuse lahutamatuid omadusi nimetatakse **atribuutideks**. Puhas punktmuster on vaid asukoha-koordinaatidest koosnev andmestik. Atribuut-andmetega punktprotsessi nimetatakse **märgistatud punktprotsessiks** (*marked point process*). Punktmustri kujunemist võib mõjutada korraga mitu protsessi, sel juhul kasutatakse termineid liitprotsess ja **liitmuster** (*compound process, compound pattern*). Teisest küljest – mitmed üksikprotsessid võivad anda tulemuseks sama punktmustri, sellised protsessid on **ekvifinaalsed**.

Taimede ja loomade paiknemismustrit mõjutavad järgmised faktorid.

- Levikuviis (piiratud levikuga liigi isendid paiknevad laigulisemalt).
- Olelusvõitlus (karjas on ringkaitse võimalik, üksikisendi puhul mitte).
- Liikidevaheline afiinsus (parasiiti, sümbionti või kommensaali võib leida põhiliselt sealt, kus on peremeesorganism; toitujat sealt, kus on toiduobjekt; teatud keskkonda eelistavat liiki sealt, kus on sobivad tingimused).
- Liigisisene konkurents ja sellest tulenev territooriumivajadus (taimedel näiteks konkurents valguse, vee ja toitainete pärast).
- Liigisisene afiinsus (koos liigikaaslastega luuakse sobivam elukeskkond).
- Keskkonna heterogeensus (metsaloomade ja -taimede paiknemismuster vastab metsade paiknemise struktuurile).
- Keskkonna arenguloo heterogeensus (enamasti avalduvad keskkonna varasemad seisundid ka praegustes mõõdetavates tunnustes, näiteks suurem lämmastikusisaldus kunagise lauda kohas).
- Mustri varasemad staadiumid. Vaadeldavale hetkele eelnenud paiknemismustri teada saamiseks tuleb teha kas pidevaid vaatlusi, või otsida kaudseid tõendeid. Näiteks looma (tegutsemis)jäljed näitavad, et mingil eelneval ajahetkel viibis loom selles kohas. Selle looma praegune paiknemine sõltub tema eelnevast paiknemisest, sest tema liikumiskiirus ei ole lõpmata suur. Mustri varasemate staadiumite mõju võib avalduda ka populatsiooni erinevas vanuses. Pikka aega ühes kohas asunud populatsioonil võib olla enam vähem stabiilne välja kujunenud paiknemismuster. Uue populatsiooni paiknemismuster võib olla laiguline vaid selle tõttu, et ta ei ole veel jõudnud ühtlasemalt laiali levida. R-strateegid on reeglina rohkem grupeerunud kui K-strateegid.

4.1.1. Punktmustrite tüübid

Punktmuster võib paikneda kas joonel, pinnal või ruumis (ka aegruumis). Punktmustri tüüp võib olla ruumis muutuv ja sõltuda vaatlusala piiridest (joonis 4-1). Sündmuste paiknemist joonel uuritakse näiteks aegridade analüüsi puhul ja nukleotiidide järjestuse uurimisel. Sündmuste aegruumis paiknemise kirjeldamine on pigem dünaamilise modelleerimise ülesanne. Maastiku mõõtkavas uuritakse eelkõige objektide paiknemist maapinnal kasutades kaarte ja kaugseire andmeid. Kõrgusdimensiooni lisamine järgnevalt esitatud meetoditele on võimalik, aga vähemalt senini vähe kasutatud.



Joonis 4-1. Punktide paiknemismustri tüübi sõltuvus vaatlusala piiridest. Suure ristküliku sees paiknevad punktid rühmadena, väikese ruudu sees hajusalt ja korrapärasemalt kui juhuslikkuse korral oodatav.

4.1.1.1. Korrapärane

Korrapärases (*regular*) punktmustris paiknevad objektid suhteliselt ühetaoliste vahedega. Korrapärase punktmustri kohta on ökoloogias kasutatud ka termineid hajunud (*dispersed*) ja ülehajunud (*over-dispersed*) rõhutamaks asjaolu, et korrapärane muster täidab ruumi ühtlasemalt kui juhuslik muster. See termin on siiski eksitav, kuna objektide arv ühetaolistel vaatlusaladel varieerub korrapäraselt paiknevate objektide puhul vähem kui juhusliku paiknemismustri korral. Üldiselt arvatakse, et organismide korrapärane paiknemine viitab nende vahel olevale konkurentsile.

4.1.1.2. Koondunud

Koondunud ehk **agregeerunud** (*aggregated, clumped*) muster viitab kas objektidevahelisele afiinsusele või siis keskkonna laigulisusele. Geobotaanikas eristatakse ka **liitelist** ja **sporaadilist** paiknemismustrit. Liitelise paiknemise korral on objektid omavahel kontaktis, sporaadiline paiknemis-tüüp tähistab suurte vahedega gruppide eraldatult paiknemist. Liiteline ja sporaadiline muster geobotaanilises käsitluses võivad kuuluda nii juhuslike, korrapäraste kui ka koondunud mustrite hulka. Kuna nende objektide kontaktist sõltuvate paiknemistüüpide eristamine eeldab objektide käsitlemist pindadena, siis ei kuulu need punktmustrite teoreetilisse käsitlusse, mille kohaselt punktsündmusel pindala ei ole.

4.1.1.3. Juhuslik

Juhusliku punktmustri iga punkti asukoht on juhuslik ja teiste punktide asukohast sõltumatu. **Täielik ruumiline juhuslikkus** (*complete spatial randomness – CSR*) tähendab, et igal sündmusel on võrdne tõenäosus esineda ükskõik millises uuritava ala osas ja sündmused paiknevad üksteisest sõltumatult. Enamik ruumimustri analüüse algab juhuslikkuse nullhüpoteesi kontrollimisega. Kui paiknemismuster ei erine oluliselt juhuslikust, siis ei ole edasistel struktuuri otsingutel suurt mõtet. Kuna tõenäosus (P_n) leida ühikulisel pinnal n juhusliku mustri punkti on ligikaudu Poissoni jaotusega ja kuna punktmustrite analüüsi kasutatakse sageli puude paiknemise seaduspärasuste uurimisel, siis nimetatakse juhuslikku punktmustrit **Poissoni metsaks**.

Täielik juhuslikkus on abstraktsioon, nagu seda on null pikkusega joon, täiesti must keha või normaaljaotus. Kasu on juhuslikust jaotusest ka siis, kui seda reaalsuses ei esine, sest juhuslikku paiknemismustrit kasutatakse etalonina teiste mustrite võrdlemisel ja vaatleja suvast sõltumatute vaatluskohtade valimisel.

Juhuslik muster on nullmudelina kasutatav abstraktsioon ja üks meetod esindusliku valimi saamiseks

Juhuslikuna paistev muster tekib enamasti paljude kindlate põhjuste tõttu. Alati on mingi väike tõenäosus, et korrapärane või koondunud muster on tekkinud juhuslikult ehk paljude väikeste teadmata põhjuste koosmõjul. Väga väikese tõenäosusega nähtusi nimetatakse tavakeeles imeks.

4.1.1.4. Liitmustrid

Kompleks- ehk **liitmustrid** (*compound pattern*) tekivad juhuslike liitprotsesside tulemusena. Liitprotsesse vaadeldakse lähemalt paiknemismustrite moodustamise osas (ptk 6). Liitmustri näiteks on juhusliku suurusega loomakarjad rohtlas või juhuslikult paiknevad metsatukad. Kui loomade paiknemist karjas ja puude paiknemist metsas määravad ühe taseme protsessid, siis karjade ja salude paiknemist maastikul määravad teise taseme protsessid. Üksikindiviidide paiknemist maastikul määrab nende kahe protsessi koondmõju. Empiirilistes andmetes esinevaid liitmustreid püütakse lähendada mõnele teoreetilisele mudelile. Näiteks kui juhuslik muster koosneb Poissoni jaotusega gruppide arvust etteantud suurusega vaatlusaladel ja indiviidide arv grupis on logaritmilise jaotusega, siis indiviidide arv vaatlusaladel on negatiivse binoomjaotusega (Pielou 1977).

Joonis 4-2. 20 juhuslikku kohta igast maakonnast. Peipsi ja Võrtsjärv on välja arvatud. Punktide keskmine tihedus sõltub maakonna suurusest – Hiiumaal on juhupunktid tihedamalt kui Pärnumaal. Maakondade piirid Maa-amet, 2. 05. 2012, juhukohtade koordinaadid Eesti ruutkilomeetrite andmebaasist (Remm 2000).

Kombineeritud punktmustri näiteks võib tuua juhuslikult järjestatud ruutkilomeetrid Eesti ruutkilomeetrite andmebaasis (http://kalleremm.ee/Eesti_KM.aspx), kus igat ruutkilomeetrit käsitletakse kui objekti, mille atribuutide hulka kuulub ka juhuslik järjekorranumber (joonis 4-2). See juhuslik järjekorranumber on genereeritud kasutades paigastikutüüpide, maakattetüüpide ja maakondade üleesindatust vältivaid dünaamilisi kvoote (Remm 2000). Dünaamiliste kvootide kasutamise eelis kihilise valiku ees on asjaolu, et ka haruldase paigastikutüübi ja harva esineva maakattega ruutkilomeetritele jääb võimalus valimisse sattuda. Juhusliku indeksi abil valitud ruutude koordinaate saab kasutada ruumiliselt juhusliku või kihilise ruumiliselt juhusliku valimi koostamisel, kus kihid määratakse mõne ruutkilomeetrite andmebaasis oleva tunnuse alusel.

4.1.2. Punktmustri kirjeldamine

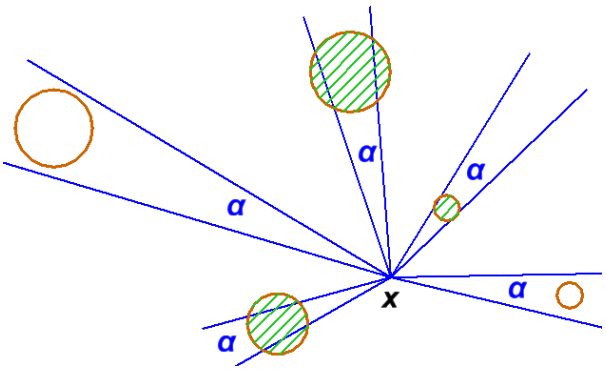
Punktmustreid saab kirjeldada graafiliselt, kaardina, mustri tüüpi iseloomustavate arvuliste näitajate ehk parameetrite kaudu, aga ka mustrit tekitanud protsessi parameetrite abil. Protsessi parameetrite kasutamisel jääb alles **ekvifinaalsuse** probleem – sama paiknemismuster võib olla tekkinud erinevate protsesside tulemusel.

Punktmustrid ebaühtlasel pinnal sõltuvad pinna omadustest. Pinna ebaühtluse problemaatikat käsitletakse pikemalt eraldi alapeatükis (ptk 4.1.5). Punktmustri kirjeldamise meetodeid jagatakse ühelt poolt ruudumeetoditeks, vahemaameetoditeks ja punktide kaardistamist nõudvateks meetoditeks. Punktmustrit kirjeldavaid arv-näitajaid jagatakse esimese järgu (*first order*) statistikuteks ja teise järgu (*second order*) statistikuteks. **Esimese järgu statistikud** kirjeldavad mustri üldkeskmisi näitajaid. **Teise järgu statistikud** kirjeldavad objektide tiheduse varieeruvust ruumis (ja ajas).

Ruudumeetodite (*quadrat sampling*) puhul loendatakse objektide arv mingitel proovialadel. Objektidevahelisi vahemaid mõõta ja asukoha koordinaate määrata ei ole tarvis.

Vahemaameetodid (*distance sampling*) tuginevad vahemaade mõõtmisele. Objektide asukohta ei ole vaja kaardistada. Vahemaad kirjeldavaid statistikuid on mugav kasutada ka siis, kui iga objekti asukoht on *x* ja *y* koordinaadi abil määratud, sest koordinaatide abil saab vahemaad arvutada.

Mingi vaatenurga sisse jäävate ja etteantust suurema nurkmõõduga objektide loendamist ehk **nurkloendamist** (*angle count sampling, Bitterlich sampling*) kasutatakse eelkõige metsanduses puistu tagavara määramisel (joonis 4-3).



Joonis 4-3. Nurkloendamine. Vaatleja asub punktis x; arvesse lähevad objektid (puud), mille nurkmõõt vaatleja asukohast (*x*) vaadates on suurem kui etteantud nurk α.

Kaardistatud objektide puhul on võimalik kasutada igasuguseid meetodeid, sest ülepinnalisele kaardile saab paigutada tinglikke vaatlusruute ja kaardistatud objektide vahemaad. Samuti on lihtne kaardistatud andmetest lihtne tuletada vahemaade jaotumist.

Ülevaated punktmustri kirjeldamise meetoditest on kirjutanud Tomppo (1986), Stoyan ja Penttinen (2000), Wiegand ja Moloney (2004), Perry *et al.* (2006), Law *et al.* (2009). Nagu ikka, ei tasu otsida ühte ja igas olukorras parimat meetodit. Erinevad meetodid täiendavad üksteist, tuues esile punktmustri erinevaid tahke.

4.1.2.1. Tihedus

Täiesti juhuslik punkte tekitav protsess on kirjeldatav ühe parameetri – **intensiivsuse** (*intensity*) abil. Staatilisel punktmustril vastab sellele oodatav tihedus pinnaühiku kohta, mida määratakse mõõtes keskmist tihedust. Nii intensiivsust kui ka tihedust tähistatakse Poissoni jaotuse parameetri λ kaudu. Keskmise tiheduse asemel saab kasutada ka selle pöördväärtust – keskmist pinda punkti kohta.

Tihedus on esimese järgu statistik. Juhuslikust mustrist erinevate mustrite kirjeldamiseks vaid tiheduse ootusest ei piisa, vaja on ka mõnda teise järgu statistikut. Keskmist tihedust saab arvutada mitmesugustest lähteandmetest, nii vahemaastatistikutest kui ka prooviruutude andmetest. Kui keskmine tihedus mustri tüüpi ei iseloomusta, siis tiheduse muutlikkus ruumis teeb seda üsna hästi. Millise ulatusega ümbruses lokaalset tihedust mõõta, tuleb uurijal endal otsustada.

Punktmustri keskmine tihedus λ avaldub n juhuslikust lähtekohast vastavate lähimate objektideni mõõdetud kauguste järgi järgmiselt (Pollard 1971):

$$\lambda = \frac{n}{\pi \sum_{i=1}^{n} r_i^2} \qquad [4\text{-}1]$$

ning mitte esimese, vaid mõne järgmise (k-nda) objekti kasutamisel:

$$\lambda = \frac{\sum_{i}^{n} k_i}{\pi \sum_{i=1}^{n} r_{ik(i)}^2}, \qquad [4\text{-}2]$$

kus n on juhuslike punktide arv, r_i on vahemaa juhuslikust lähtekohast i lähima punktobjektini. Järgmiste naabrite kasutamise eelis on tiheduse hinnangute väiksem varieeruvus.

4.1.2.2. Loendid

Loendid (*counts*) on objektide arv mingis ühikus. Punktobjektide arv juhuslikust punktmustrist võetud ühesuurustel proovialadel on ligikaudu Poissoni jaotusega. Erinevus ühesuurustelt vaatlusaladelt saadud loendite tegeliku ja juhuslikust mustrist oodatava jaotuse vahel iseloomustab mustri tüüpi. Loendite jaotus ruutudes ei sõltu ainult objektide paiknemismustrist, vaid ka ruudu suurusest ja kindla suunaga mustri olemasolul ka loendusalade kujust. Soovitatud on, et loendusruutude pindala oleks umbes ühe- kuni kahekordne uurimisala pindala jagatud punktide koguarvuga. Loendite meetod on väheefektiivne, kui objektid paiknevad hõredalt ja enamikus vaatlusalades ei oleks ühtegi objekti. Sellisel juhul tuleks eelistada kaugusmeetodeid.

Juhusliku punktmustri puhul on objektide oodatav arv samasuurtel vaatlusaladel võrdne. Empiiriliste loendite ebavõrdsuse statistilist olulisust saab kontrollida χ^2 jaotuse abil vabadusastmetega $m - 1$

$$\chi^2 = \sum_{i=1}^{m} \frac{(x_i - \bar{x})^2}{\bar{x}}, \qquad [4\text{-}3]$$

kus m on vaatlusalade arv ja x_i on objektide arv vaatlusalal i (Diggle 1983).

4.1.2.3. Dispersiooniindeksid

Lihtsaim loendite hajuvust kirjeldav näitaja on variatsioonikoefitsient, see on kindla pindalaga vaatlusaladel leitud objektide arvu standardhälbe s ja keskmise $\bar{x}$ suhe (ptk 1.3.4.2). Ökoloogias on tavaks saanud loendite hajuvust kirjeldada pigem dispersiooni s^2 ja keskmise suhtega, mida on nimetatud ka **dispersiooniindeksiks** (Ludwig ja Reynolds 1988) ja **Coxi indeksiks** *CI* (Cox 1971, Neumann ja Starlinger 2001). Coxi indeksi miinimumväärtus on null, punktide juhusliku paiknemise korral on oodatav väärtus 1 ja maksimum ei ole piiratud.

$$CI = \frac{s^2}{\bar{x}} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^2}{\bar{x}(m-1)}, \quad [4\text{-}4]$$

kus m on vaatlusalade arv ja x_i on objektide arv vaatlusalal i.

On kasutatud ka variatsioonikoefitsiendi modifitseeritud variante, näiteks David ja Moore (1954) on esitanud **koondumisindeksi** (*index of cluster size – ICS, index of clumping – IC*) kujul

$$ICS = \frac{s^2}{\bar{x}} - 1, \quad [4\text{-}5]$$

mille juhuslikkuse korral oodatav väärtus on 0. Indeksi positiivseid väärtusi saab tõlgendada kui juhusliku objektiga liituvate teiste objektide keskmist arvu. Indeksi negatiivsed väärtused viitavad hajutatud paiknemisele. Mõlema eelmise indeksi puudus on maksimumväärtuse piiramatus ja sõltuvus vaatluste arvust.

Greeni indeksis (*GI*) jagatakse hajuvuse ja keskmise suhe loendatud objektide koguarvuga, mille tulemusel on maksimaalse koondumise korral oodatav väärtus võrdne ühega (Green 1966).

$$GI = \frac{(s^2 / \bar{x}) - 1}{(\sum x) - 1} \quad [4\text{-}6]$$

Fisheri paralleelvaatluste dispersiooniindeksis q kasutatakse andmestiku jagamist sama mahuga alajaotusteks või samades tingimustes tehtud korduvvaatlusi. Originaalpublikatsioon käsitleb mikroobide loendamist Petri tassidel inkubeeritud mullalahuses (Fisher *et al*. 1922).

$$q = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^2}{\bar{x}} = \frac{(m-1)s^2}{\bar{x}}, \quad [4\text{-}7]$$

kus x_i on objektide arv alajaotuses i, s^2 on loendite dispersioon, $\bar{x}$ on objektide keskmine arv alajaotuse kohta ja m on alajaotuste arv. Objektide juhusliku paiknemise korral on Fisheri dispersiooniindeksi oodatav väärtus $m-1$.

Taimestiku kirjeldamise kontekstis nimetab Diggle (1979a) seda meetodit Fisheri kvadraatide meetodiks ja soovitab vaatlusala jagamist mõlemat pidi viieks moodustades kokku 25 alajaotust.

Johnson ja Zimmer (1985) esitasid **dispersiooniindeksi I**, mille parameeter V_i võrdub kahemõõtmelisel pinnal πR_i^2, kus R_i on vahemaa juhuslikust kohast i lähima punktobjektini ja n on juhuslike lähtekohtade arv.

$$I = \frac{(n+1)\sum_{i=1}^{r} V_i^2}{\left(\sum_{i=1}^{r} V_i\right)^2} \quad [4\text{-}8]$$

Selle dispersiooniindeksi väärtused on $1 + 1/n \leq I \leq \infty$; juhusliku punktmustri korral oodatav väärtus $E(I) = 2$ ja dispersioon $Var(I)$ arvutatakse järgmiselt:

$$Var(I) = \frac{4(n-1)}{(n+2)(n+3)}. \quad [4\text{-}9]$$

4.1.2.4. Lloydi grupeerumisindeks ja laigulisuse indeks

Lloydi grupeerumisindeks (*index of mean crowding – IMC*) näitab juhusliku objektiga samas prooviruudus koosesinevate teiste objektide keskmist arvu (*mean crowding*) ja sõltub prooviruudu suurusest. **Laigulisuse indeks** (*index of patchiness – IP*) näitab objektide ruumilise jaotuse erinevust juhuslikust jaotusest (Lloyd 1967). Juhusliku jaotumise korral on *ICS* = 0, seega *IMC* = $\bar{x}$ ja *IP* = 1.

$$IMC = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{\sum_{j}^{m} x_j\left(x_j - 1\right)}{\sum_{j}^{m} x_j} = \bar{x} + \frac{s^2}{\bar{x}} - 1 = \bar{x} + ICS, \qquad [4\text{-}10]$$

$$IP = \frac{IMC}{\bar{x}} = \frac{\bar{x} + \frac{s^2}{\bar{x}} - 1}{\bar{x}}, \qquad [4\text{-}11]$$

kus *N* on isendite koguarv, ja X_i – iga isendi kaaslaste arv vastavas prooviruudus (valimis).

4.1.2.5. Morisita agregatsiooniindeks

Morisita indeks (*Morisita's index of aggregation*) (1959a) väljendab suhtelist tõenäosust, et kaks juhuslikult valitud objekti paiknevad samas loendis (vaatlusruudus). Lloydi grupeerumisindeksist (*IMC*) varem publitseeritud Morisita indeks on esimesele lähedane (Ludwig ja Reynolds 1988).

$$I_\delta = \frac{m\sum_{i=1}^{m} x_i\left(x_i - 1\right)}{N(N-1)} = \frac{m}{N-1} IMC, \qquad [4\text{-}12]$$

kus *N* on objektide üldarv kõigis valimites kokku, *m* on valimite arv ja x_i on objektide arv valimis *i*. Kui $I_\delta > 1$, siis paiknevad objektid grupeerunult, kui $I_\delta < 1$, siis paiknevad objektid korrapäraselt.

4.1.2.6. Astmefunktsioon

Perioodil 1950ndad kuni 70ndad aastad oli ökoloogias aktuaalne teema prooviruutudes loendatud isendite arvu keskmise ja dispersiooni vaheline seos. Taylor (1961) leidis paljude andmestike ülevaatamisel, et seos keskmise (*m*) ja dispersiooni (s^2) vahel on astmefunktsiooniga.

$$s^2 = am^b \qquad [4\text{-}13]$$

Taylori järgi on *b* liigispetsiifiline parameeter, mille seos ruumimustriga paraku raskesti interpreteeritav. Seda seost on hakatud nimetama **Taylori astmeseaduseks** (*Taylor's power law*).

Astmefunktsiooni on kasutatud ka uuritud ala pindala ja seal esinevate liikide arvu modelleerimiseks võrrandiga

$$S = S_0 A^z, \qquad [4\text{-}14]$$

kus *S* on liikide arv, *A* on ala pindala, S_0 ja *z* on mudeli parameetrid.

S_0 saab interpreteerida kui ühikulisel pindalal oodatavat liikide arvu. Kui S_0 sõltub pindala mõõtmise ühikutest, siis liigirikkuse ja pindala suhte mudeli parameeter *z* peaks teoreetilise ökoloogia seisukohast midagi näitama. On leitud, et mandritel on *z* väärtused enamasti 0,1 ja 0,2 vahel, saartel aga on graafiku tõus järsem (*z* >0,3) (Rosenzweig 1995). Ulrich ja Buszko (2003) järgi on vahemereliste liblikaliikide astmefunktsiooni *z* parameeter palju suurem (*z* = 0,49) kui Põhja- ja Ida-Euroopa maadele iseloomulikel liikidel (*z* = 0,10). See võib viidata elupaikade suuremale varieeruvusele Vahemeremaades.

4.1.2.7. Erisuurused vaatlusalad

Greig-Smith (1952, 1964) soovitas taimede paiknemismustri kirjeldamisel suure arvu juhuslikult paiknevate vaatlusalade asemel kasutada erineva suurusega külgnevaid alasid ja pakkus kahte moodust, kuidas vaatlusalasid suurusega 1, 2, 4, 8, 16, 64 ühikut üksteise kõrvale paigutada (joonis 4-4). Erineva suurusega vaatlusalade abil saab määrata üksikobjektide ruumilise kobara tüüpilist suurust. Hilisemas publikatsioonis tõdeb Greig-Smith, et enamikes väliuuringutes on lihtsam kasutada samas sihis paiknevaid eri suurusega vaatlusalasid või muutuvat vaatluste vahemikku vaatlustransektil (Greig-Smith 1961). Vaatlustransekti puhul oleks võrdluse alus mitte vaatlusala pindala, vaid vaadeldud lõigu pikkus.

Olgu alade suurusklassid tähistatud vastavalt pindalale (*r* = 1, 2, 4, …). Seejärel saab iga suurusklassi puhul arvutada erinevuse (G_r) kahe hajuvusnäitaja vahel – kahekordne loendite hajuvus (T_r) väiksema suuruse juures ja kaks korda suurema vaatlusala (T_{2r}) puhul.

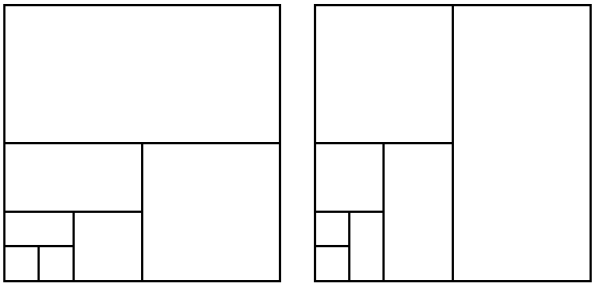
$$G_r = 2T_r - T_{2r} \qquad [4\text{-}15]$$

$$T_r = \frac{\sum_i^n \left(x_{i1} - x_{i2}\right)^2}{n} \qquad [4\text{-}16]$$

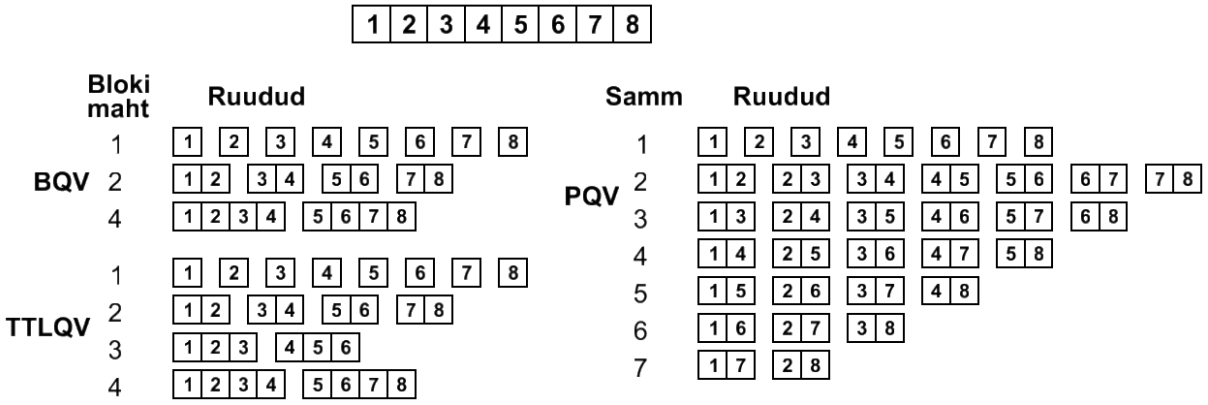
Siin *n* on külgnevate vaatlusalade paaride arv suurusel *r*, x_{i1} ja x_{i2} on objektide arv vaatlusalade paari esimesel vaatlusalal ja teisel vaatlusalal.

Seejärel saab graafikul kujutada G_r ja *r* vahelist seost. Objektide juhusliku paiknemise korral on G_r enam-vähem konstantne, agregeeritud paiknemise korral näitab G_r järsu tõusu koht suurusele *r* vastavat klastri suurust.

Kui vaatlusruudud on ühesuurused, saab neid mitmel viisil küljetsi suuremateks üksusteks ühendada määramaks indiviidide koondumise ja tõukumise tendentsi mõõtkava (joonis 4-5). Greig-Smithi (1952) kahekordse pinnaga vaatlusalade kompaktset ühendamisviisi tuntakse **ühendatud vaatlusruutude hajuvuse** meetodina (*blocked quadrat variance – BQV*), **mitmeti ühendatud vaatlusruutude hajuvuse** (*two term local quadrat variance – TTLQV*) (Hill 1973) meetod kaasab kahekordsetele muutustele ka vahepealse mahuga külgnevate vaatlusalade kombinatsioonid, seega moodustatakse ka liidetud ruudud suurusega 3, 5, 6, 7, 9 jne algset ruutu (vt ka ptk 4.3.1.1). **Paarikaupa vaatlusruutude hajuvuse** meetod (*paired quadrat variance – PQV*) ühendab erineva kaugusega vaatluspaare ja on seega variogrammi analoog. Kui vaatluspaarid valitakse juhuslikult, tähistatakse viimast RPQV. RPQV võimaldab osa vaatluspaare jätta sõltumatuks kontrollandmestikuks. BQV, TTLQV ja RPQV meetod võivad näidata täiesti erinevat seost vahemaa ja varieeruvuse vahel (Turner *et al.* 1991).



Joonis 4-4. Külgnevate vaatlusalade paigutuse kaks varianti Greig-Smithi (1952) proovivõtumetoodika järgi.



Joonis 4-5. Transektil asuvad külgnevad vaatlusruudud ja nende kombineerimise meetodid: BQV, TTLQV, PQV (Ludwig ja Reynolds 1988, muudetud).

Külgnevate proovialade meetodi puudustena märgivad Upton ja Fingleton (1985) ning Diggle (2003), et G_r väärtus sõltub:

- proovialade paigutamise viisist,
- klastrite kuju vastavusest proovialade kujule ning
- *r* väärtustele vahepealse suurusega klastrite olemasolu ei pruugi ilmneda ja
- G_r maksimumide statistilise olulisuse määramiseks ei ole valmis statistilisi teste.

Seetõttu soovitatakse algse BQV meetodi asemel kasutada TTLQV meetodit (Ludwig ja Reynolds 1988). Kriitikale vaatamata võib teha järelduse, et vaatlusalade suuruse varieerimine annab paiknemismustri kohta lisaandmeid ja võimaldab sama vaatlusseeriaga saada andmeid erineva suurusega organismide kohta.

4.1.2.8. Klastrite suuruste jaotus

Kui punktmustri sees on eristatud parved ehk kogumid ehk klastrid, siis saab mustrit iseloomustada kogumitevaheliste vahemaade, kogumite suuruste ja üksikpunktide paiknemise järgi kogumikeskmete suhtes. Sama punktmustrit saab kogumiteks kui kõrgema üldistustasemega objektideks jagada paljudel erinevatel viisidel. Kogumi suurust saab mõõta kogumi pindala või kogumisse kuuluvate objektide arvu abil. Goodall (1974) soovitas klastri suuruse määramiseks mõõta erinevatel kaugustel olevates juhuslikes ruutudes leitud isendite arvu varieeruvust.

Klastrikeskmete koordinaate ja klastrisse kuuluvate objektide arvu määramiseks saab kasutada klasteranalüüsi, mille lähteandmeteks on ruumikoordinaadid.

4.1.2.9. Pielou indeks

Pielou korrapäraindeks (*Pielou's nonrandomness index – PNI*) (Pielou 1959) on punktide tiheduse (*n/A*) ja nende vaheliste vahemaade (*r*) ruutkeskmise korrutis, kus *k* on punktide arv ja λ on keskmine tihedus.

$$PI = \pi \frac{n}{A \cdot k} \sum_{i=1}^{k} r_i^2 = \pi\lambda \frac{\sum_{i=1}^{k} r_i^2}{k} \qquad [4\text{-}17]$$

4.1.2.10. Kaugus lähima objektini

Kaugust lähima punktobjektini saab arvutada nii juhupunktist kui ka korrapäraselt paiknevatest punktidest lähtudes. Vahemaade jaotust juhupunktist lähima objektini nimetatakse ka **tühiku statistikuks** ehk sfäärilise kontakti jaotusfunktsiooniks (*empty space statistic, spherical contact distribution function*) ja tähistatakse *F(x).* Praktikas on sellised tühikud näiteks häilud metsas. *F(x)* väljendab tõenäosust, et juhuslikult paigutatud ringis raadiusega *x* on vähemalt üks punktobjekt.

Kuna juhusliku paiknemise korral on punktobjekti leidmise tõenäosus igas pinna osas ühesugune, siis saab juhusliku paiknemise korral oodatavaid lähima naabri kaugusi arvutada ühtede ja samade valemite järgi; sõltumata sellest, kas lähtepunktideks on olemasolevad objektid, juhuslikud punktid või korrapäraselt paiknevad punktid. Kui punktide paiknemine ei ole juhuslik, siis sõltub lähima naabri kauguste jaotus mõõtmise lähtekohtade paiknemisest. Juhuslikust kohast lähima naabri kauguste keskmise ja olemasolevatest punktidest lähima naabri kauguste keskmise suhet on kasutatud punktmustrit kirjeldava indeksina (Hopkins ja Skellam 1954, Coops ja Culvenor 2000).

Juhuslikus punktmustris on oodatav vahemaa lähima punktobjektini võrdne nii lähtudes punktobjektidest, juhuslikest kohtadest kui ka korrapäraselt paiknevatest kohtadest

Eberhart (1967) on uurinud, kuidas arvutada tõenäosusjaotust, et juhusliku koha ümber olevas etteantud raadiusega ringis ei ole ühtegi punktobjekti ja kuidas see sõltub juhusliku ja regulaarse komponendi vahekorrast mustris. Seejuures sõltub regulaarse komponendi arvutamise viis korrapära tüübist ja ülesanne ei ole analüütiliselt üheselt lahendatav. Somers ja Oderwald (1988) kasutasid keerukat suurima tõepära (ptk 1.2 ja 3.6.1) hinnangut regulaarsete paigutuste osakaalu määramiseks. Nad näitasid, et juhusliku ja regulaarse mustrikomponendi vahekorra üle saab otsustada lähima naabri kauguste varieeruvuse põhjal.

4.1.2.11. Lähima naabri kaugus

Keskmine lähima naabri kaugus (*mean distance to the nearest neighbour – MNND*) võib kirjeldada nii ühte tüüpi objektide kui ka erinevat tüüpi punktobjektide paiknemist. Kui analüüsitakse kahe paiknemismustri assotsieerumist, siis leitakse kaugused ühe klassi kõigist objektidest lähima teise klassi objektini (ptk 4.1.3). Punktide juhuslikul paiknemisel lõpmata suurel pinnal on keskmine lähima naabri kaugus ($MNND_{exp}$) keskmise tiheduse λ kaudu arvutatav (Clark ja Evans 1954)

$$MNND_{\exp} = \frac{1}{2\sqrt{\lambda}} \qquad [4\text{-}18]$$

Tegeliku ($MNND_{obs}$) ja oodatava keskmise lähima naabri kauguse erinevuse statistilist olulisust

saab hinnata normaaljaotuse *Z* statistiku abil.

$$Z = \frac{\left| MNND_{obs} - MNND_{\exp} \right|}{SE_{\exp}} \qquad [4\text{-}19]$$

Eeltoodud valemi nimetajas olev juhupaiknemisest oodatava keskmise lähima naabri kauguse standardviga (SE_{exp}) avaldub kujul

$$SE_{\exp} = \frac{0{,}26136}{\sqrt{n \cdot \lambda}}, \qquad [4\text{-}20]$$

kus *n* on punktide koguarv ja λ on keskmine tihedus.

Lähima naabri kauguse tegeliku ja oodatava väärtuse suhe ehk **Clark-Evansi agregatsiooniindeks** (*Clark-Evans index of aggregation*) näitab, mitu korda on keskmine lähima naabri kaugus suurem juhusliku paiknemise korral oodatavast.

$$CE = \frac{MNND_{obs}}{MNND_{\exp}} = \frac{2\sqrt{\lambda}\sum_{i=1}^{n} r_i}{n}, \qquad [4\text{-}21]$$

kus r_i on kaugus lähima naaberobjektini.

Suhtarvu eelis keskmise lähima naabri kauguse ees on sõltumatus mõõtühikust. Kui kõik objektid paiknevad ühes kohas, on *CE* = 0; juhusliku paiknemise korral *CE* = 1; maksimaalse korrapära korral paikneksid kõik isendid ühesuguste korrapäraste kuusnurksete lahtrite keskkohtades ja *CE* = 2,1491 (Clarke ja Evans 1954).

Lähima naabri kauguste jaotust nimetatakse ka **Diggle G-funktsiooniks** (*Diggle's G function*) (Diggle 1983, 2003, Barot *et al.* 1999). Kumulatiivse jaotusfunktsioonina tähistatakse seda *G(x)*, kus parameeter *x* tähistab lähima naabri kaugust üldjuhul ja funktsiooni väärtuseks on kas oodatav lähima naabri esinemise tõenäosus või empiirilistest andmetest leitud lähimate naabrite suhteline sagedus. Üksikobjekti *i* lähima naabri kaugus on x_i. Täht *G* tähistab vähemalt ühe objekti esinemise tõenäosust ehk tõenäosust, et lähim naaber on kaugusel *x* või lähemal. *G(x)* funktsioon on lähedane radiaal-jaotusele *G(r)* (ptk 4.1.2.18), mis arvestab mitte vaid esimest naabrit, vaid ka järgmisi ning mida enamasti kasutatakse mittekumulatiivsel ja normeeritud kujul. Juhuslikule kohale lähima naabri jaotust tähistatakse *F(x)* või *F(r)* (ptk 4.1.2.10).

Lähima naabri kaugusele tuginevad meetodid arvestavad vaid kõige lähimat naabrit, mis ei iseloomusta kogu mustrit

Empiirilist lähima naabri kauguste jaotust saab võrrelda juhusliku jaotuse järgi oodatava jaotusega. Empiirilist jaotust tähistatakse valemina

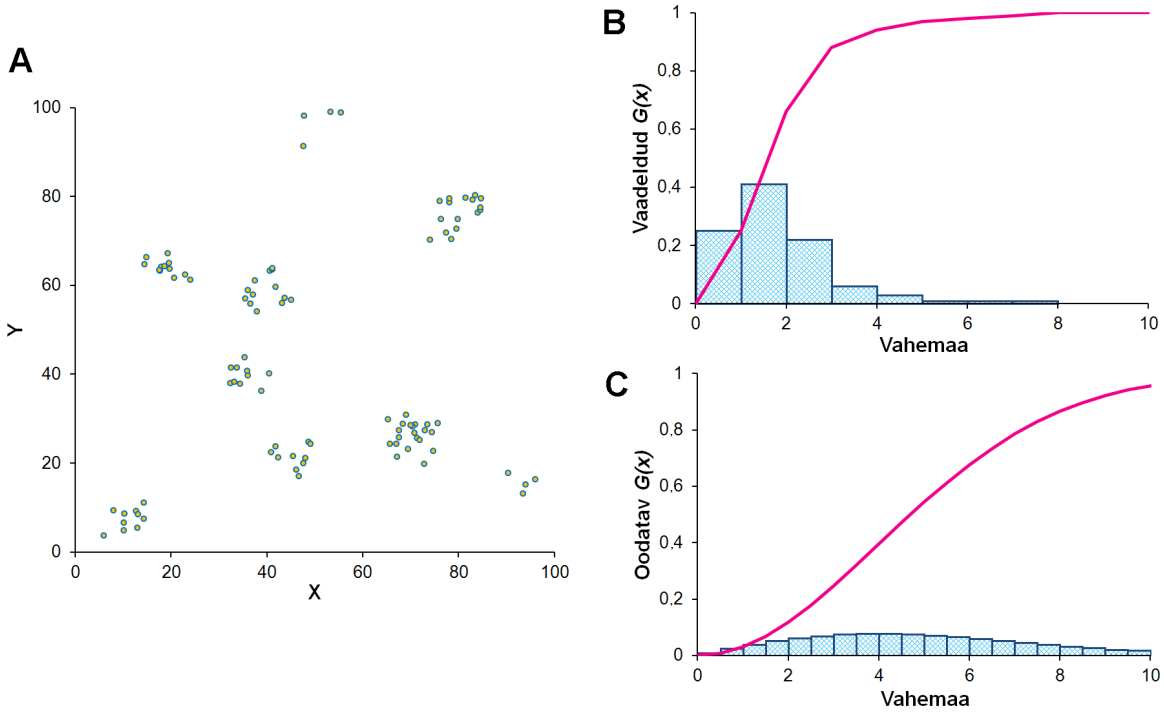
$$\hat{G}(x) = \frac{\varphi(d_i \le x)}{n}; \qquad [4\text{-}22]$$

objektide juhusliku paiknemise korral oodatavat (teoreetilist) jaotust (joonis 4-6)

$$G(x)_{theor} = 1 - e^{-\pi\lambda x^2}, \qquad [4\text{-}23]$$

kus *x* on etteantud kaugus, d_i on objekti *i* lähima naabri kaugus, λ on punktide keskmine tihedus, $\varphi()$ sulgudes olevale tingimusele vastavate objektide arv.

Punktide koondumise või tõukumise mõõduna saab kasutada tegeliku ja oodatava lähima naabri kauguste jaotuse maksimumide vahet. Kui vaatlusandmete kohaselt domineerivad suuremad lähima naabri kaugused kui juhuslikust paiknemisest võiks oodata, siis ilmneb objektide vahel tõukumis-tendents.



Joonis 4-6. 100 punktobjekti agregeerunud paiknemine (A), nende lähima naabri kauguste jaotus ja suhteline sagedus vahemaaklassides (B) ning juhusliku paiknemise korral oodatav lähima naabri kauguste jaotus ja sagedus vahemaaklassides (C).

G(x) jaotuse diferentsimisel saadakse tõenäosusjaotus, mis näitab lähima naabri teatud kaugusel paiknemise tõenäosust. Tõenäosus (*P*), et punktobjekti lähima naabri kaugus (*d*) juhuslikus mustris on vahemikus $d_1 \ldots d_2$ avaldub kujul

$$P(d_1 < d < d_2) = e^{-\lambda d_1^2} - e^{-\lambda d_2^2}. \qquad [4\text{-}24]$$

Lähima naabri kaugused saab järjestada suuruse järgi. Vaadeldud ja juhuslikust jaotusest oodatavate lähima naabri järjestatud jaotusi on mugav graafikul võrrelda. Juhusliku paiknemise puhul oodatavad lähima naabri kaugused on lineaarses seoses järgmise funktsiooniga

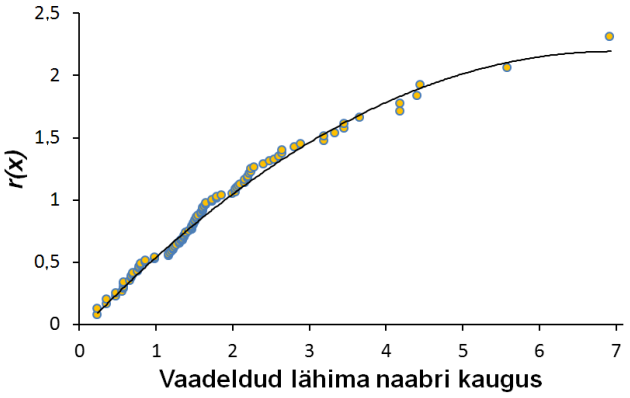
$$r(x) = \sqrt{-\ln\left(1 - \frac{x - 0{,}5}{n}\right)}, \qquad [4\text{-}25]$$

kus *n* on valimi suurus, *x = 1, 2, 3 … n* ja funktsiooni tõusunurga ruut näitab punktmustri tihedust (Campbell 1992, 1995, 1996).

Juhupaiknemisest oodatava ja tegeliku standardiseeritud lähima naabri kauguste jaotuse erinevus on üsna hea punktmustri iseloomustaja (Frelich *et al.* 1993, Gappa *et al*. 1997). Kui võrrelda graafikul lähima naabri järjestatud vahemaid juhuslikus ja empiirilises mustris eeldades, et vaadeldud vahemaad on horisontaalteljel ja *r(x)* funktsioon vertikaalteljel, siis kõvera ülespoole kumerus viitab agregeeritusele, nõgus kõver näitab paiknemismustri korrapärasust (joonis 4-7). Gappa *et al.* (1997)

jagasid tegelikud ja teoreetilised lähima naabri kaugused kaugusklassidesse ja võrdlesid sagedusjaotusi χ^2 testiga.

Lähima naabri kauguste kasutamist kritiseerib B.D. Ripley (1985). Kuna lähima naabri kauguste jaotus on alati tugevasti asümmeetriline, siis soovitab Ripley selle asemel kasutada lähima naabrini ulatuva ringi keskmist pindala.



Joonis 4-7. Järjestatud lähima naabri kaugused (punktid) joonisel 4-6A kujutatud andmetest juhuslikkuse korral oodatavate lähima naabri kauguste *r(x)* suhtes. Trendjoon on väiksemate väärtuste juures enam-vähem lineaarne, keskmiste ja suuremate vahemaade puhul kumer, mis viitab punktide juhuslikule paiknemisele klastrite sees ja agregeeritusele laiemas ümbruses.

4.1.2.12. k lähima naabri kaugus

Thompson (1956) esitas idee kasutada liigi leiukohtade paiknemise heterogeensuse määramisel mitte ainult lähima naabri jaotust, vaid *k* lähima naabri kaugusi (*k nearest neighbours – kNN*). Juhuslikus paiknemismustris on vaatenurgas Θ radiaani olevate lähima kuni *k*-lähima naabri kauguste tihedus oodatavalt järgmine (Holgate 1972).

$$r_k = (\lambda\Theta)^k \prod_{i=1}^{k} x_i e^{-\frac{1}{2}\lambda\Theta x_k^2} \qquad [4\text{-}26]$$

Manly (1997) kasutas tegeliku ja randomiseeritud $k = 1\ldots m$ lähima naabri keskmise kauguse võrdlemist männitaimede paiknemismustri analüüsil. Kui Manly arvutas keskmise kauguse vaid paari esimese astme naabrini, siis Davis *et al.* (2000) arendasid meetodit edasi, arvutades keskmise kauguse kuni *n–1* naabrini. *n* tähistab siin punktide koguarvu ja üks on lahutatud, sest objekt ei ole naabriks iseendale. Saadud jaotusi võrreldi juhukordustel saadud tulemuste usalduspiiridega. Juhuslike vahemaade saamiseks võeti uuritud kohtade komplektist 5000 juhuslikku valimit, mille maht võrdus liigi leiukohtade arvuga. Juhukorduste genereerimisel jälgiti liikide suhtelist sagedust elupaigatüüpides ja uurimisala allosades.

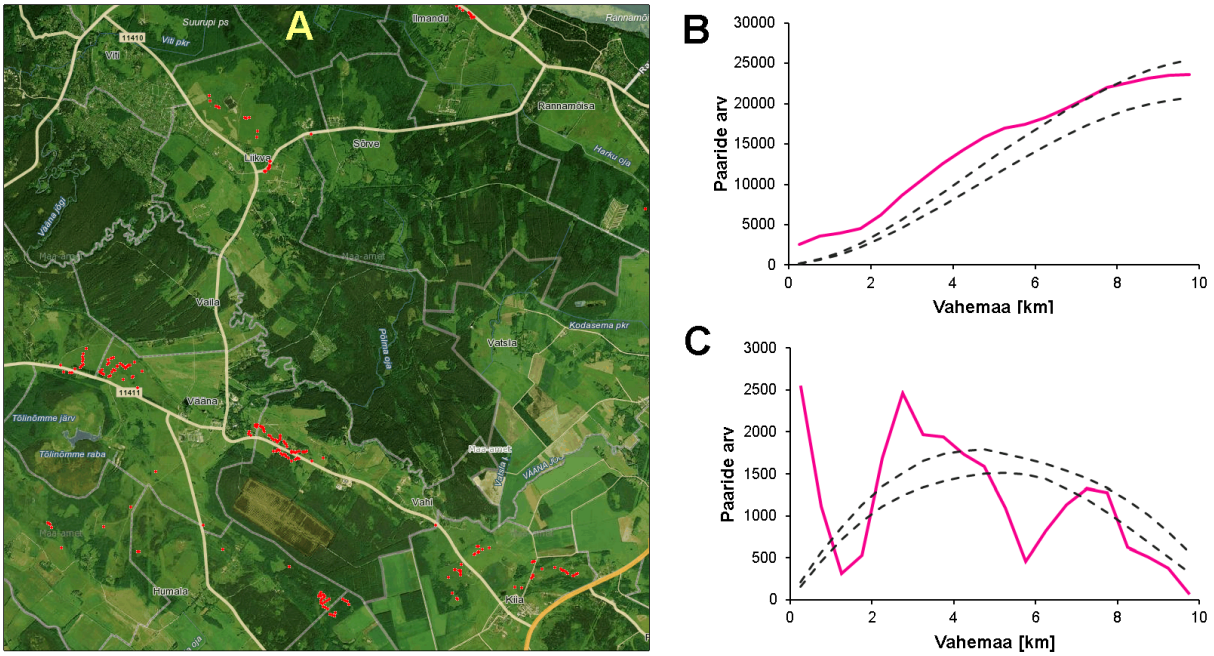
4.1.2.13. Kõigi vahemaade jaotus

Juhusliku paiknemismustri kõigi objektide omavaheliste vahemaade oodatava väärtuse saab leida, kui punktidevaheliste vahemaade kogupikkus jagada juhuslikult *S* lõiguks ja järjestada lõigud suurenevas järjekorras MacArthur (1957). Kusjuures *m*-inda lõigu oodatav pikkus – N_m on

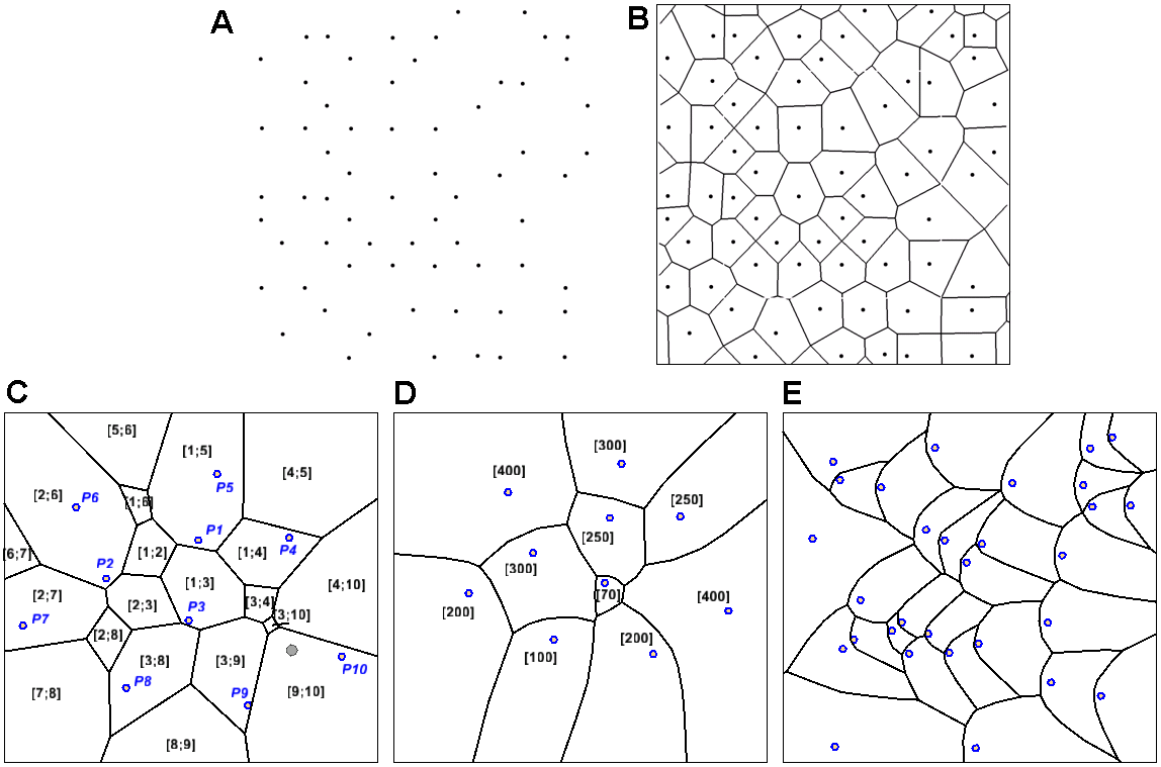
$$N_m = \frac{T}{S}\sum_{i=1}^{m}\frac{1}{S-i+1}, \qquad [4\text{-}27]$$

kus *i* on objekti indeks, *S* on *n* objekti vahemaade arv = *n(n−1)/2*; *T* on vahemaade summa.

Kõrvalekalded vahemaade tegeliku jaotuse ja oodatava jaotuse vahel näitavad paiknemismustri iseloomu (joonis 4-8). Üldiselt on väiksemate vahemaade jaotus informatiivsem.



Joonis 4-8. A – aasnelgi leiukohad (punased täpid) Keilast kirdepool oleval kaardilehel 6382 maa-ameti kaardi ja ortofoto taustal. B – kõigi vahemaade kumulatiivne jaotus (pidevjoon) aasnelgi leiukohtade vahel ja sama arvu kohtade juhuslikul paigutamisel saadud vahemaade 95% usaldusvahemik (katkendjooned), C – sama sagedusjaotustena. Leiukohad on juhuslikust sagedamini vahemaa <1 km ja 2 kuni 3 km vahemaa juures. Kaardilehe serva pikkus = 10 km.



Joonis 4-9. A – punktmuster; B – punktidevahelise ala tesselatsioon; C – eraldised kahe lähima punkti järgi (nurksulgudes eraldisega seotud punktide numbrid); D – eraldised kaalutud mõjuga punktide ümber (nurksulgudes eraldisega seotud punkti kaal); E – eraldised voolavas keskkonnas. Joonise osad C, D ja E on koostatud Okabe *et al.* (1994) järgi.

Bartlett (1964) soovitab objektidevaheliste kauguste jaotumist analüüsida spektraalanalüüsi meetoditega. Spektraalanalüüsist ehk Fourier' analüüsist on juttu aegridade modelleerimise osas (ptk 3.5.7).

4.1.2.14. Tesselatsioonipindade jaotus

Pinda saab mingite reeglite järgi osadeks jagada ehk **tesseleerida**. Objekti ümber piiritletud ruumiosa, mille igale kohale on just see objekt mingi reegli järgi mõõdetult kõige lähem, nimetatakse selle objekti ümbruseks ehk **proksimaalregiooniks** ehk Thiesseni polügooniks (joonis 4-9). Proksimaalregioone saab moodustada üsna mitmete reeglite abil (Okabe *et al.* 1992, 1994). Tesselatsiooniüksuste pindalade ja muude parameetrite jaotus iseloomustab tesselatsioonikeskmete paiknemismustrit.

4.1.2.15. Kaugus korrapärani ja kaugus grupeerumiseni

Kaugus korrapärani ja **kaugus grupeerumiseni** (*distance to regularity, distance to crowding*) on J.N. Perry esitanud mõõdikud kaugusindeksite ruumilises analüüsis (*spatial analysis by distance indices – SADIE*) (Perry ja Hewitt 1991, Perry 1995, 1998). SADIE võrdleb valimi ruumilist struktuuri samadest andmetest moodustatud maksimaalselt grupeerunud, juhusliku ja maksimaalselt korrapärase paigutusega.

SADIE kaugus korrapärani algoritm on järgmine.

- Teisenda vaatluskohtade koordinaate nii, et ristkülikukujulise vaatlusala nurgakoordinaadid oleksid: (0,0), (0,A), (1,A), (1,0), kus $0 < A < 1$.
- Moodusta iga vaatluskoha ümber Thiesseni polügoon.
- Nihuta vaatluskohti nende kõige kaugemete naabrite poole. Iga naabri panus uue koha leidmisel on võrdeline ühise külje pikkusega tesselatsioonis. See tagab punkti liikumise oma lähimate naabrite juurest kõige kaugemete naabrite poole. Naabritena arvestatakse neid punkte, mille tesselatsioonipolügoonil (ümbrusalal) on vaadeldava punkti ümbrusalaga ühine külg. Servaalale lisatakse ajutised punktid, et vähendada servaefekti. Punkte nihutatakse mustriseisude kaupa nii, et vaid eelmise mustriseisu asukohad mõjutavad punktide uusi asukohti.
- Uue mustriseisu moodustumise järel tesseleeritakse see uuesti ja mõõdetakse ümbrusalade pindalade varieeruvust. Kui varieeruvus vähenes rohkem kui etteantud konstant, siis jätkatakse protsessi punktist 2. Vastasel juhul mustri korrastamine lõpetatakse.
- Mõõdetakse kaugus korrapärani, mis on esialgse ja lõpliku asukoha vaheline vahemaa iga punkti puhul. Kogu uuritava kogumi kaugus korrapärani *D* võrdub üksikpunktide summaarse kaugusega korrapärani.
- Tulemuse olulisust võrreldakse Monte Carlo iteratsioonidel (ptk 3.6.6 ja 4.1.4.2) juhumustritest saadud tulemustega.

Perry (1998) esitas kaks koonduvuse indeksit

$$I_a = \frac{D}{E_a} \qquad [4\text{-}28]$$

$$J_a = \frac{F_a}{C}, \qquad [4\text{-}29]$$

kus E_a on genereeritud juhumustrite keskmine kaugus korrapärani, D on empiirilise mustri kaugus korrapärani, C on kaugus koondumiseni, minimaalne summaarne vahemaa punktide liigutamiseks ühte kohta, F_a on genereeritud juhumustrite punktide keskmine summaarne kaugus koondumiseni.

$I_a > 1$ ja $J_a > 1$ viitab enamasti koondunud paiknemismustrile; $I_a = 1$ and $J_a = 1$ on juhumustrist oodatavad väärtused ja $I_a < 1$ ja $J_a < 1$ viitab punktide korrapärasele paiknemisele.

Genereeritud juhumustrite selle osa suurust, mille puhul E_a on võrdne või suurem kui empiiriliste andmete kaugus korrapärani, tähistatakse P_a. Juhumustri nullhüpoteesi võib koonduvuse kasuks kõrvale heita, kui $P_a < 0{,}025$. Kui $P_a > 0{,}975$, siis võib otsustada korrapärase paiknemise hüpoteesi kasuks.

SADIE metoodikal on puudusi.

- Vaatlusala servad mõjutavad tulemusi.
- Korrapära ala sees sõltub korrapära paiknemisest alast väljaspool.
- Kaugust koonduvuseni on mõtet kasutada vaid siis, kui on põhjust oletada vaid ühe kobara esinemist kogu vaatlusalal.
- SADIE indeksi väärtus sõltub mitte ainult loendite suhtelisest paiknemisest, vaid ka koondunud väärtuste absoluutsest asendist uurimisala sees (Xu 2003). Koondumisindeks on suurem, kui suurte väärtuste laik on uurimisala servas, sest summaarne vahemaa koondumiskeskmesse on sealt suurem.

4.1.2.16. Amalgamatsiooniindeks

D.A. Coomes *et al.* (1999) on välja pakkunud SADIE meetodiga sarnase metoodika, milles koonduvuse mõõduna kasutatakse reduktsioonide arvu objektide ruumilisel rühmitamisel. Klastrikeskmete leidmiseks kasutatakse klasteranalüüsi, mille lähteandmeteks on ruumikoordinaadid. Agregeeritud paiknemisega objekte klasterdatakse senikaua, kuni saadud klastrikeskmete muster ei erine oluliselt juhuslikust. Klastritesse liidetud objektide arvu saab kasutada agregeerituse mõõtmisel. Reduktsioonide keskmine arv ehk **amalgatsiooni indeks** (*amalgamation index – A*) arvutatakse klastritesse liidetud objektide ja klastrite arvu suhte ($\check{N}_{clump}$) järgi

$$A = 1 - \frac{1}{\check{N}_{clump}}. \qquad [4\text{-}30]$$

Kui uuritaval alal pindalaga Q on N_i objekti liigist i ja kui need objektid paiknevad juhuslikult, siis keskmine naaberobjektide arv raadiuses r juhuslikust objektist on

$$E(N_{ii}) = \frac{\pi r^2 (N_i - 1)}{Q}. \qquad [4\text{-}31]$$

Seetõttu saab agregeeritust mõõta tegeliku ja oodatava naabrite arvu suhtega $AI(r)$

$$AI(r) = \sum_{1}^{Ni} N_{ii}(r) / [\pi r^2 N_i (1 - N_i)/Q] \qquad [4\text{-}32]$$

kus: N_{ii} on samast liigist naabrite arv raadiuses r. $AI > 1$ näitab grupeerumist, $AI < 1$ on korrapära.

4.1.2.17. Ripley K funktsioon

Juhusliku punktprotsessi esimese järgu karakteristik on protsessi intensiivsus ehk punktide keskmine tihedus. Keerukamate, mitmes mõõdus struktuure sisaldavate mustrite kirjeldamiseks tuleb kasutada erineval kaugusel paiknevate naaberobjektide hulka või tihedust. Lihtne näitaja on punktide keskmine arv naabruses, aga see sõltub naabruse ulatusest ja punktide keskmisest tihedusest (Ripley 1981).

Ripley (1976, 1977, 1985) võttis punktmustrite kirjeldamiseks kasutusele funktsiooni *K(t)*, mis sisaldab nii objektide vahemaade mõõtmist kui ka objektide loendamist etteantud raadiuses. Sellepärast loetakse seda teise järgu statistikuks. Punktide oodatav arv lähtekohast kauguseni *t* on punktide juhusliku paiknemise ja keskmise tiheduse λ korral võrdne $\lambda\pi t^2$. Kauguse tähistamiseks kasutatakse *t* asemel ka tähti *d* või *r*. Traditsioon tähistada kaugustsooni tähega *t* pärineb aegridade analüüsist, *d* viitab ladina keelest pärit sõnale distants ja *r* rõhutab vahemaa mõõtmist raadiusena. Selles tekstis on eelistatud Ripley funktsiooni originaaltähistust.

Kui empiirilise punktmustri keskmine punktobjektide arv kauguseni *t* jagada punktide keskmise tihedusega, saame tulemuseks ringi pindala, mille raadius on *t*, mis ongi *K(t)* funktsiooni väärtus empiirilise mustri korral.

$$\hat{K}(t) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\delta_{ij}(t)}{N\lambda} = \frac{A}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\delta_{ij}(t) = \frac{\overline{N}_{t>tij}}{\lambda}, \qquad [4\text{-}33]$$

kus λ on punktide keskmine tihedus; $\delta_{ij} = 0$, kui *c* ja $\delta_{ij}=1$, kui $t > t_{ij}$; *A* on uuritava ala pindala ja *N* on punktide koguarv. $\Sigma\Sigma\delta_{ij}(t)$ tähistab kaugusel *t* või lähemal olevate naaberpunktide hulka. Kui see jagada punktide üldhulgaga, siis saame keskmise naabrite hulga kauguseni *t*. *A/N* on punktide keskmise tiheduse pöördväärtus. $\overline{N}_{t>tij}$ tähistab keskmist naaberpunktide arvu raadiuses *t* kõigi lähtepunktide ümber. Kui punktid on klastritesse grupeerunud, on naaberpunktide tihedus väikestel kaugustel suurem kui juhusliku jaotuse järgi oodatav tihedus ja $K(t) > \pi t^2$. Punktide korrapärase paiknemise korral on vastupidi.

Ripley K(t) funktsioon on raadiuses t olevate naaberobjektide keskmine arv jagatud objektide keskmise tihedusega, st normeeritud objektide oodatava arvuga ühikulisel pinnal; λK(t) on oodatav naabrite arv raadiuses t

Punktmustri kirjeldamiseks on mõnikord eelistatud punktobjektide tegeliku ja juhupaiknemise korral oodatava *K(t)* erinevust ehk $K(t) - \pi t^2$ (Diggle 1979a). Kuna lineaarmõõdud on paremini võrreldavad ja stabiilsema dispersiooniga kui pindmõõdud, siis enamasti kasutatakse *K(t)* funktsiooni kas *L(t)* või $L(t) - t$ statistiku kujul. *t* asemel võib samahästi olla ka tähistus *d* või *r*.

$$L(t) = \sqrt{\frac{K(t)}{\pi}} \qquad [4\text{-}34]$$

L(t) statistikut on kasutatud ka kujul $t - L(t)$ (Upton ja Fingleton 1985). *L(t)* statistiku oodatav väärtus ühtlasel piiramatu suurusega pinnal paikneva juhusliku punktprotsessi korral on 0. Usalduspiirid *K(t)* ja *L(t)* statistikutele saadakse Monte Carlo meetodiga (ptk 3.6.6 ja 4.1.4.2).

Ripley *K(t)* on nimetatud punktmustrite analüüsi standardmeetodiks (He ja Duncan 2000). Meetod on teadusmaailmas laialt aktsepteeritud, sest see ei sõltu punktide tihedusest ja seda on võimalik

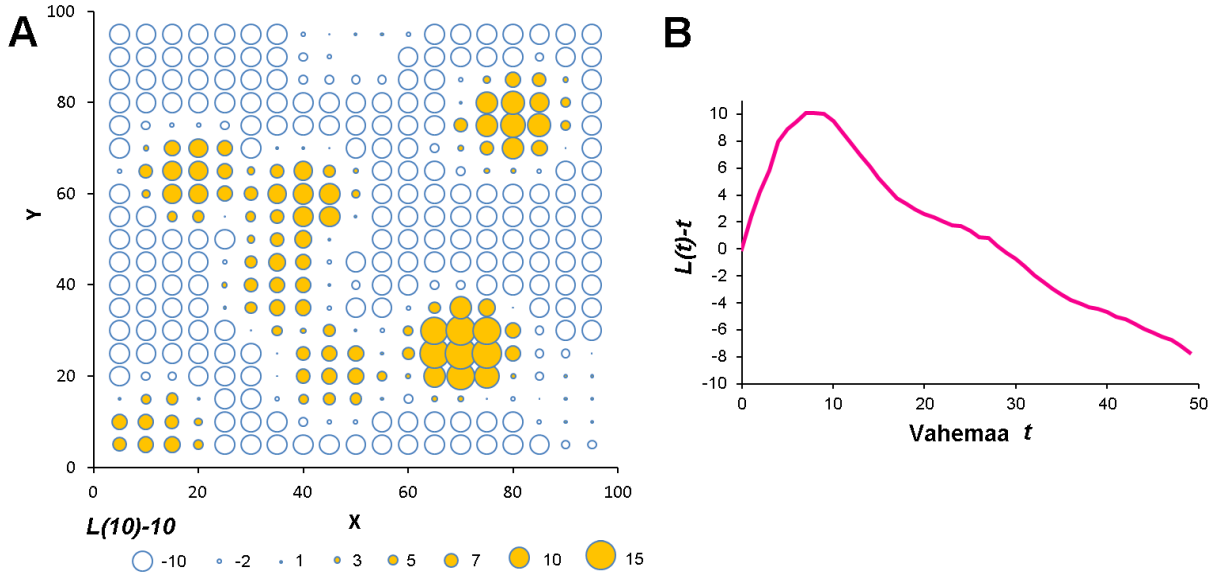
kasutada nii sama tüüpi punktide paiknemise kui ka samal alal esinevate punktmustrite teineteise suhtes paiknemise analüüsiks.

Enamasti on *K(t)* või *L(t)* statistikut kasutatud mingil alal olevate punktide paiknemismustri summaarse näitajana, kuid neid statistikuid on võimalik arvutada ka lokaalse indikaatorina iga üksikobjekti juures (Getis 1984, Getis ja Franklin 1987) või etteantud (korrapärase võrgustiku) punktides, nagu teisigi ruumimustrit kirjeldavaid näitajaid (joonis 4-10). Ripley *K(t)* funktsiooni on arvutatud ka aegruumis (Lynch ja Moorcroft 2008).

Servalähedaste punktide korral tuleks *K(t)* funktsiooni arvutamisel naaberpunkti etteantud raadiuse sees või väljas olemise indikaatori δ_{ij} asemel kasutada servakorrektuuri kaale w_{ij}:

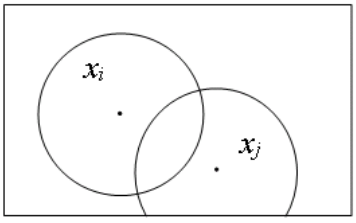
$$\hat{K}(t) = \frac{A}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(t), \qquad [4\text{-}35]$$

kus $w_{ij} = 1$, kui $t \leq b_{ib}$ ja $w_{ij} > 1$, kui $t > b_{ib}$ ning b_{ib} on kaugus punktist *i* uurimisala servani.



Joonis 4-10. Joonisel 4-6A olevast punktmustrist arvutatud *L(t)-t* statistikud arvutatuna lokaalselt 10 kaugusühiku raadiuses (A) ja arvutatuna globaalselt (B).

Kuna servalähedaste lähtepunktide ümber olev uuritud ala pind ja selles oodatav naabrite arv on väiksem kui raadiuse kaudu arvutatava ringi pind ja sellel oodatavalt olev naabrite arv, siis tuleks servalähedaste punktide korral korrigeerida naabrite arvu vaid uuritud ala sisse jääva pindala osaga ringi kogupindalast (joonis 4-11). Servakorrektuuri kaalude arvutamise eeskirju on käsitlenud Barot (1999), Cressie (1993), Diggle (1979a, 1983, 2003), Ripley (1977, 1985), Getis ja Franklin (1987), Goreaud ja Pelissier (1999).



Joonis 4-11. Ripley *K(t)* kaalude arvutamine servalähedaste objektide korral. Kaal punktis $x_i = 1$; kaal punktis $x_j > 1$, kuna selle punkti uuritavast alast välja jäävas ümbruses võib olla naaberobjekte (Diggle 2003, muudetud).

Servakorrektuuri kaal K(t) arvutamisel peaks vastama uurimisala sisse jäävale osale vaadeldavas raadiuses olevast alast

Kui uuritav punktmuster paikneb ebaühtlasel pinnal või soovitakse arvestada objektide paiknemise erinevat tõenäosust pinna eri osades, on põhjendatud iga lähtekoha ümber oleva ümbruse kaalutud pindala kasutamine raadiuse kaudu arvutatud pindala asemel. Raadiuses *t* olevate naaberobjektide arvu saab sel juhul korrigeerida sobivuskaalude abil arvutatud pindala ja raadiuse kaudu arvutatud pindala erinevuse järgi.

4.1.2.18. n-osakese jaotus, paariline korrelatsioon, radiaaljaotus

n-osakese jaotus on tõenäosusjaotus, et *n* osakest, mis kuuluvad teatud punktobjektide hulka *N*, paiknevad teatud etteantud ruumiosas. *n*-osakese jaotust kasutatakse statistilises füüsikas ja molekulaarkeemias ning see tuletatakse Gibbsi jaotusest, mis kirjeldab tõenäosust $P(X = x)$, et süsteem X on olekus x (juhuslik muutuja X omab väärtust x). Gibbsi jaotus sõltub aine struktuurist ja on näiteks kristalsetel ning vedelatel ainetel erinev.

$$P(X = x) = \frac{1}{Z(\beta)} e^{-\beta E(x)}, \qquad [4\text{-}36]$$

kus $E(x)$ on muutuja X võimalike väärtuste funktsioon, β on jaotuse parameeter ja $Z(\beta)$ on normaliseeriv teisendus.

Kui *n*-osakese jaotus on seotud paiknemisega mingi teise osakese suhtes, siis nimetatakse seda **n-osakese korrelatsiooniks** ja tähistatakse $g^{(n)}$. Ühe osakese kohast sõltuva paiknemise tõenäosust kirjeldab ühe osakese jaotus $g^{(1)}$, mis on homogeenses gaasis või vedelikus keskmise tihedusega võrdne konstant. Kristalsete ainete puhul on $g^{(n)}$ laineline. *n*-osakese jaotust kasutatakse ka taimede ja loomade paiknemise analüüsil. Viimasel juhul esindab *n*-osakese jaotus teatud arvu isendite leidumise tõenäosust teatud mahuga naabruses.

Paariline korrelatsioon [*pair correlation* – $g_{ij}(r)$] on *n*-osakese korrelatsiooni erijuht tingimusel, et $n = 2$ ehk vaadeldakse ühe osakese paiknemist teise suhtes. Paariline korrelatsioon näitab tõenäosust leida teist osakest teatul kaugusel esimesest osakesest. Ühetaoliste sfääriliste osakeste homogeenses keskkonnas omavahelise paiknemise tõenäosus on vaid osakestevahelise radiaalkauguse funktsioon ja sellisel juhul nimetatakse paarilist korrelatsiooni **radiaaljaotuseks** (*radial distribution function*). Kumulatiivne radiaaljaotus $G(r)$ näitab tõenäosust, et raadiuses r ühest osakesest leidub mõni teine osake. Radiaaljaotust kasutatakse enamasti mittekumulatiivsena ja keskmise tihedusega (gaaside puhul ideaalse gaasi tihedusega) normeeritud kujul ning tähistatakse sel juhul $g(r)$

$$g(r) = \frac{\rho(r)dr}{\rho^{ID}(r)dr}, \qquad [4\text{-}37]$$

kus $\rho(r)dr$ on osakeste tegelik arv kaugustsoonis dr ning nimetajas on ideaalse gaasi osakeste arv tsoonis dr.

$$\rho^{ID}(r)dr = \frac{N}{V}dr \qquad [4\text{-}38]$$

Radiaaljaotuse maksimumid näitavad vahemaid, kus osakesed üksteise suhtes sagedamini viibivad. Osakeste tüüpiline esinemistõenäosus on null kuni mingi läheduseni, sest kaks osakest ei saa asuda

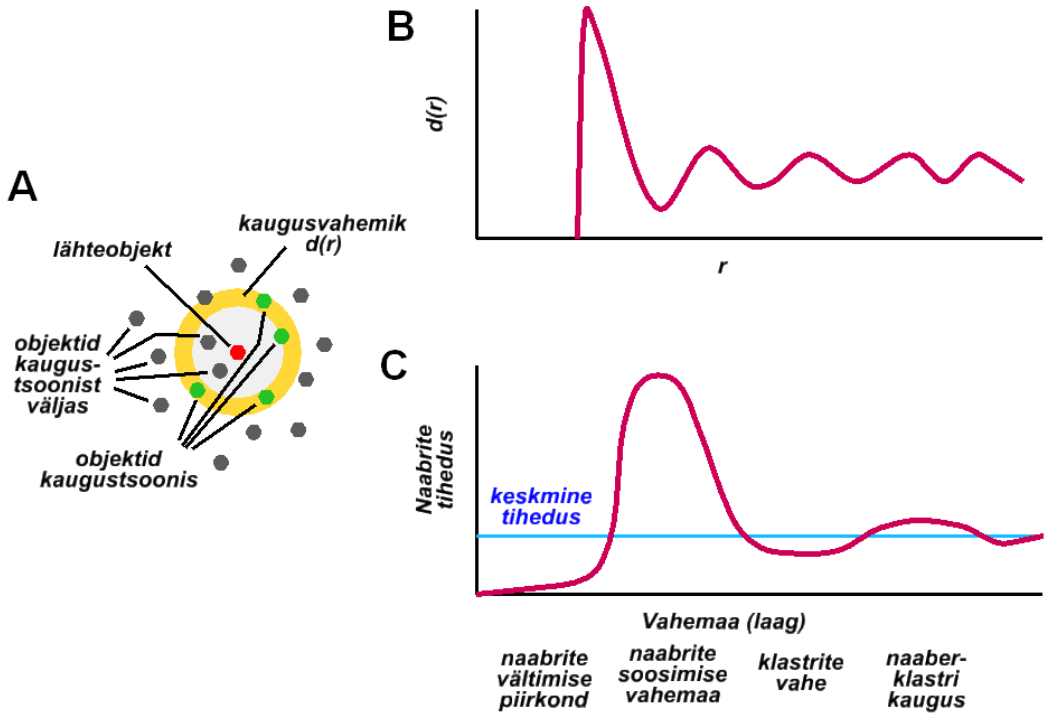
samas kohas ega väga lähedal teineteisele. Naabrite vältimise kaugust nimetatakse keemias **van der Waali diameetriks**. Nullile järgneb maksimum, sest naaberosakesed püüavad difundeeruda antud osakese asukohta, kuid kohtavad takistust. Maksimum põhjustab sellele järgneva miinimumi. Järgnevad osakeste esinemise tõenäosuse sumbuvad võnked suurematel kaugustel (joonis 4-12B). Kuna radiaaljaotus normeeritakse enamasti osakeste keskmise tihedusega, siis vedelike puhul sumbub see tasemele 1. Vedelike osakeste radiaaljaotus on üldiselt sujuvam kui kristalsetel ainetel. Mittehomogeensete ainete (segude) puhul aitab radiaaljaotus kirjeldada aine struktuuri.

Ripley *K(r)* on radiaaljaotuse kumulatiivne variant ja radiaaljaotus on Ripley *K* funktsiooni mittekumulatiivne kuju ehk diferentsiaal (Stoyan 1988, Comas ja Mateu 2007).

$$g_{ij}(r) = \frac{1}{2\pi r}\frac{d\left(K_{ij}(r)\right)}{dr} \qquad [4\text{-}39]$$

$$K_{ij}(r) = \int_0^r g(r)2\pi r dr \qquad [4\text{-}40]$$

Radiaaljaotus on teise objekti leidumise suhteline tõenäosus ühetaolises keskkonnas sõltuvalt ümbruse raadiusest



Joonis 4-12. A – punktobjektide paiknemine; B – tüüpiline radiaaljaotus; C – naabrite tiheduse jaotus agregeerunud punktmustris, kus klastrite piires paiknevad objektid suhteliselt korrapäraselt.

Radiaaljaotust kasutatakse statistilises füüsikas, keemias, astronoomias ja ökoloogias objektide omavahelise paiknemise kirjeldamiseks, molekulaarbioloogias nukleotiidide paiknemise analüüsil ja epidemioloogias haiguste leviku modelleerimisel. Aatomfüüsikas kirjeldatakse radiaaljaotusega elektronide orbitaale. Ökoloogilises kirjanduses käsitletakse paarilist korrelatsiooni ja radiaaljaotust enamasti samas tähenduses.

4.1.2.19. Märgikorrelatsioon

Punktobjektide atribuudi ehk märgi väärtuste vahemaast sõltuvat seost nimetatakse **märgikorrelatsiooniks** (*mark correlation*), kui atribuudi väärtused on arvulised ja **märgiühenduseks** (*mark connection*), kui atribuut näitab klassikuuluvust (Stoyan ja Penttinen 2000). Sisuliselt mõõdavad need näitajad ruumilist autokorrelatsiooni (ptk 5.1) ja segregatsiooni (ptk 4.1.3). Märgikorrelatsiooni käsitleme siimnkohal punktmustrite peatükis, kuna seda käsitletakse paarilise korrelatsiooni edasiarendusena ja punktide atribuutide paiknemise seaduspärasuste näitajana.

Märgikorrelatsioon k_r arvutatakse kaugustsoonide kaupa ja selles kasutatav seosefunktsioon $t(m_i,m_j)$ võib olla erinev. Enamasti on seosefunktsiooniks lihtsalt märgi väärtuste korrutis ühe objektil ja teise objektil, mis jagatakse seosefunktsiooni ootuse (c) ja kaugustsoonis r olevate objektipaaride arvu (n_r) korrutisega (Getzin *et al*. 2011).

$$k_r = \frac{\sum_i^n \sum_j^n t\left(m_i, m_j\right)_r}{cn_r} \qquad [4\text{-}41]$$

Kui seosefunktsioonina kasutatakse väärtuste korrutist, siis seosefunktsiooni ootuseks on märgi üldkeskmise ruut. Märgikorrelatsiooni kasutatakse näiteks puude paiknemise seaduspärasuste kirjeldamisel, kus märgi väärtuseks on igat puud iseloomustav tunnus (diameeter, kõrgus, puuliik vms).

4.1.2.20. Naabrite tiheduse jaotus

Ökoloogilistes uuringutes on naaberobjektide sagedust mõõtvaid arvutuseeskirju nimetatud mitmeti: radiaaljaotuseks ja paariliseks korrelatsiooniks (ptk 4.1.2.15), **O-ring statistikuks** (*O-ring statistic*) (Wiegand *et al.* 1999, Wiegand ja Moloney 2004), **suhteliseks naabrustiheduseks** (*relative neighbourhood density – RND*) (Condit *et al*. 2000) ja **naabrite tiheduse jaotuseks** (*neighbours' density distribution*) (Remm ja Luud 2003), **mittehomogeenseks paariliseks korrelatsiooniks** (Law *et al.* 2009). Põhiline asjaolu kõigi nende puhul on, et naabrite hulka ei loendata kumulatiivselt kogu raadiuse ulatuses alates fokaalkohast, vaid kaugustsoonides.

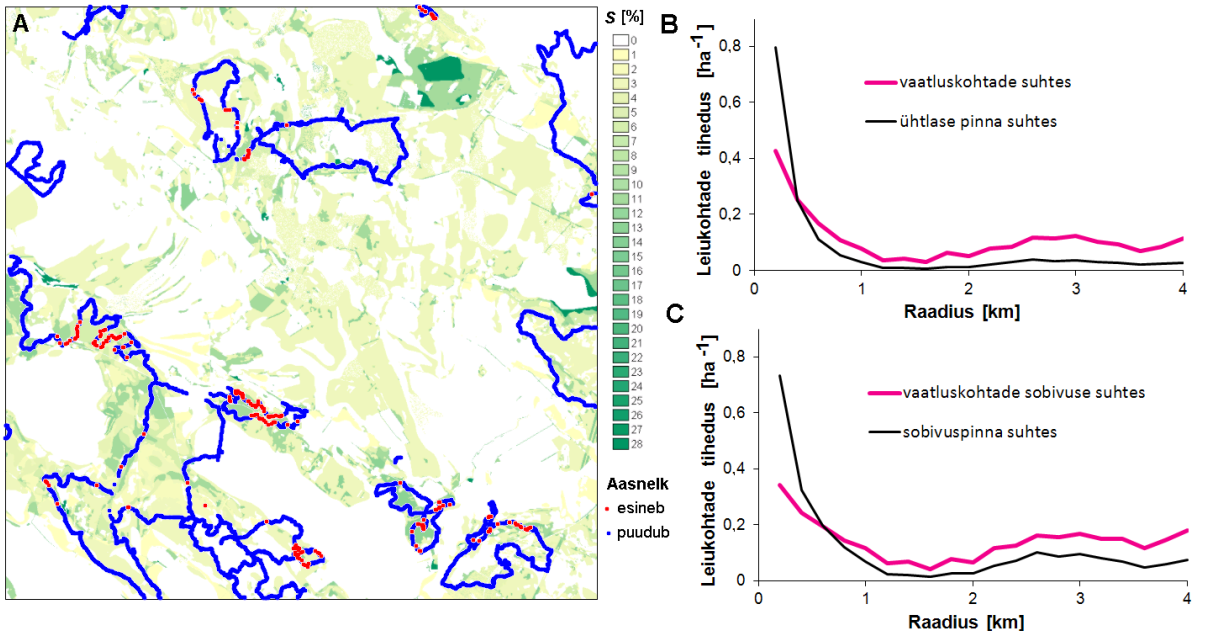
O-ring statistikut ja naabrite tiheduse jaotust ei normeerita keskmise tihedusega, paarilist korrelatsiooni ja radiaaljaotust normeeritakse (joonis 4-12C). Radiaaljaotuse puhul võrreldakse osakeste lokaalset tihedust osakeste keskmise tihedusega ja eeldatakse pinna ühetaolisust. Keskmise tihedusega normeerimine on põhjendatud objektide paiknemise seaduspära võrdlemiseks erineva keskmise tihedusega punktmustrites. Kui peamine eesmärk on punktmustri kirjeldamine, on otstarbekas naaberobjektide tihedust mitte normeerida, vaid kasutada tihedust pinnaühiku suhtes. Law *et al.* (2009) normeerisid radiaaljaotust punktobjektide lokaalse tihedusega, eeldades punktprotsessi piirkondlikult erinevat intensiivsust.

O-ring statistikut ja naabrite tiheduse jaotust ei normeerita keskmise tihedusega, paarilist korrelatsiooni ja radiaaljaotust normeeritakse

Kui statistilises füüsikas kasutatava radiaaljaotuse puhul eeldatakse ruumi homogeensust, siis ökoloogias ei ole selline eeldus enamasti mõistlik. Ökoloogiline pind on ikka ebaühtlane ja seda ebaühtlust on võimalik ka kuidagi kirjeldada. Pinna kvaliteedi varieeruvus põhjustab objektide koondumise, mis ei ole tingitud objektide omavahelistest suhetest, vaid sarnasest asukoha-eelistusest. Sobivamad elupaigad on tihedamini asustatud, sest seal on pinnaühiku kohta rohkem eluks vajalikke

ressursse. Naabrite puudumine naabruses olevates täiesti sobimatutes elupaikades ei kajasta mitte niivõrd uuritava nähtuse üksikobjektidele endile omast koondumise või tõukumise tendentsi, vaid elupaikade paiknemismustrit. Seega saab naabrite tiheduse jaotust kasutada keskkonna mitteteadaoleva laigulisuse tuvastamiseks ja kirjeldamiseks. Kui eesmärk on uurida objektide omavahelisi suhteid, mitte keskkonna laigulisust, tuleks naabrite tihedus arvutada sobivusega kaalutud pinna suhtes, nagu seda tegid Remm ja Oja (2001), Remm ja Luud (2003), Remm *et al*. (2006).

Lisaks pinna sobivuse laigulisusele mõjutab objektide tihedust ka vaatluste ebaühtlane ruumiline intensiivsus ja vaatluskohtade ebaühtlane paiknemine (joonis 4-13). Sobivuspinna kasutamine suurendab küll arvutuste mahtu iga statistiku arvutamisel, aga ruumistatistikas on arvutuste suur maht tavaline nähe ja tänapäeva arvutustehnika juures ei ole naabritiheduste arvutusaeg enam otsustava tähtsusega.



Joonis 4-13. Naabrite tiheduse jaotus sobivuspinna arvestamisel ja mittearvestamisel. A – aasnelgi leiu- ja puudumiskohtade vahekorra järgi igal mullatüübi ja põhikaardi põhiala kombinatsioonil hinnatud elupaigasobivus (*S*) maksimaalse võimaliku suhtes Keilast kirdepool oleval kaardilehel 6382 (kaardilehe külje pikkus on 10 km); B – naaberleiukohtade tiheduse jaotus ühetaolise pinna suhtes ja vaadeldud kohtade suhtes (vaadeldud koha pinnaks on arvestatud vaatluskohta sisaldav 10 × 10 m ruut). C – naaberleiukohtade tiheduse jaotus sobivusega kaalutud pinna suhtes ja sobivusega kaalutud vaatluskohtade suhtes.
Leiukohtade koondumine umbes paarisaja meetri raadiuses on nende andmete alusel poolenisti seletatav ebaühtlase vaatlusintensiivsusega (tihedus esimeses kaugustsoonis kogu pinna suhtes on umbes kaks korda suurem kui vaatluskohtade suhtes), liigiomase ruumilise koondumisega (kõik graafikud näitavad leiukohtade koondumist) ja vaid vähesel määral mullatüübi ja põhikaardi abil kirjeldatud kasvukoha sobivuse laigulisusega (vaatluskohtade sobivuse arvestamine ei muuda tiheduse jaotust).

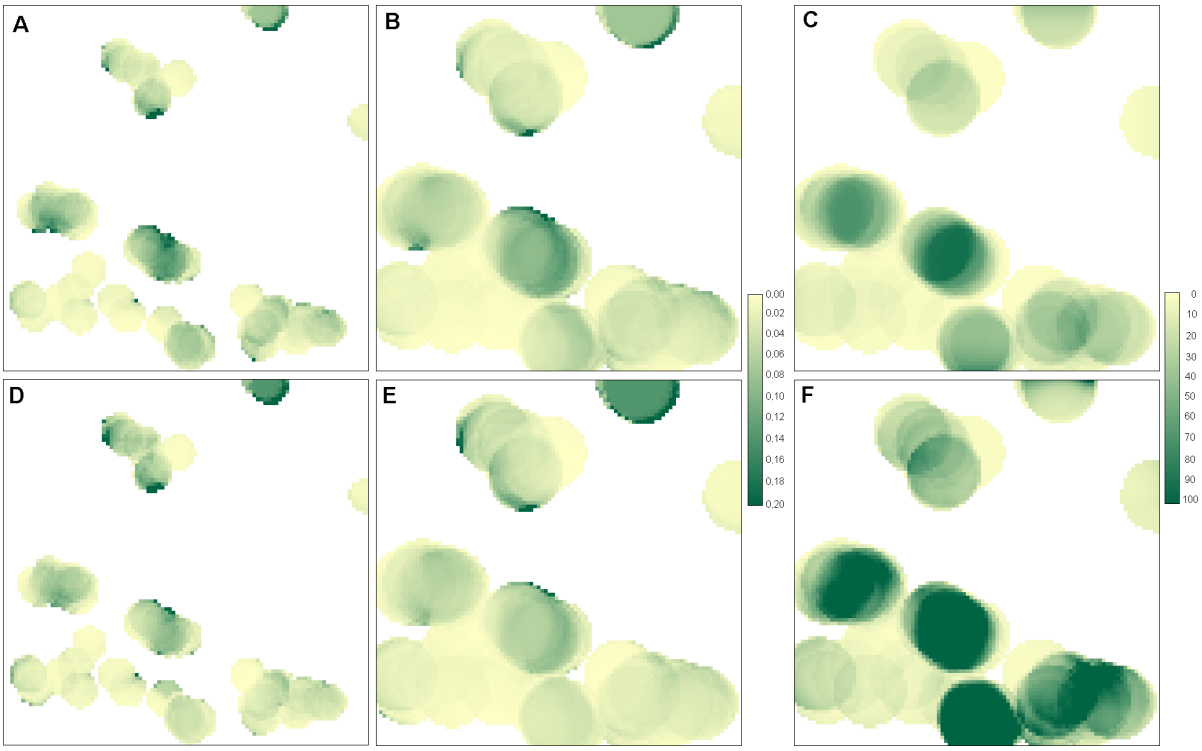
Isendite looduses paiknemise seaduspärasuste kirjeldamisel ja modelleerimisel tuleks arvestada elupaigasobivuse muutlikkust

Ideaalsetest punktidest koosnevate punktmustrite jaoks välja töötatud meetodite rakendamisel loodusandmetele on segavaks faktoriks objektide suurus. Näiteks suured puud ja ka muud suuremad taimed katavad märkimisväärse pinna, muutes teise objekti paiknemise sellel pinnal võimatuks. Wiegand *et al.* (2006) soovitavad selliste andmete korral objekte tähistavad eraldised rasteriseerida ja genereerida eraldise igasse pikslisse selle eraldise klassile vastav punktobjekt. Nuske *et al.* (2009) kohandasid paarilise korrelatsiooni arvutust eraldiste jaoks, kasutades seejuures eraldiste piirjooni

tsentroidide asemel. Vahemaad objektide vahel arvuti nende servast, mitte keskkohast. Meetodit näitlikustati metsa häilude paiknemise analüüsiga.

Naabrite tihedus ja selle analoogid arvutatakse eraldi iga kaugustsooni jaoks. Naabrite tiheduse kujutamiseks kaardil tuleks kas kasutada vaid ühte kaugustsooni, luua iga kaugustsooni jaoks omaette kaardikiht või ühendada statistiku väärtused erinevates kaugustsoonides üheks koondstatistikuks (joonis 4-14).

Ripley K(t) ja naabrite tiheduse jaotus täiendavad teineteist ja soovitav on punktmustri iseloomustamiseks kasutada mõlemat



Joonis 4-14. Aasnelgi leiukohtade tihedus [1/ha] Keilast kirdepool asuval kaardilehel 6382: A – vaatluskohtade suhtes kaugusel kuni 500 m, B – vaatluskohtade suhtes kaugusel kuni 1000 m, C – kogu pinna suhtes kaugusel kuni 1000 m, D – pinna sobivusega kaalutud vaatluskohtade suhtes kaugusel kuni 500 m, E – pinna sobivusega kaalutud vaatluskohtade suhtes kaugusel kuni 1000 m, F – sobivusega kaalutud pinna suhtes kaugusel 1000 m. Leiukohad, puudumiskohad ja pinna sobivus on joonisel 4-13A. Vaatluskohtade suhtes arvutatud tihedused on kõrged mitmes kohas, kus etteantud raadiuse piiresse jääb vähe registreeritud puudumiskohti.

Naabrite tiheduse jaotusel on võrreldes Ripley *K(t)* funktsiooniga nii eeliseid kui ka puudusi. Naabrite tiheduse eelised on järgnevad.

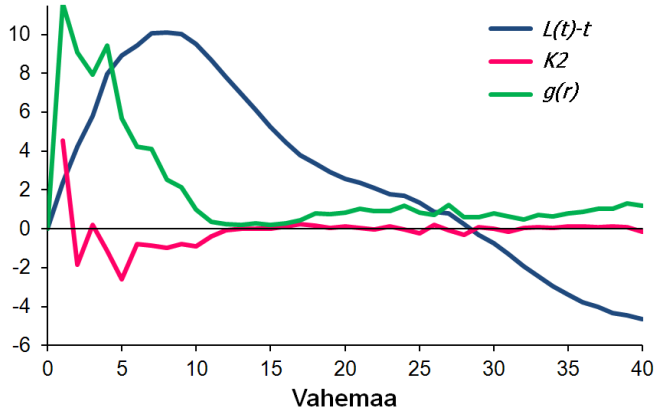
- Lihtsam mõistetavus ja tõlgendatavus. Naabrite tiheduse maksimum on väiksema raadiuse juures kui samadest andmetest arvutatud kumulatiivse kõvera maksimum.
- Kaugusvahemike sõltumatu käsitlus. Ripley *K(t)* summutab seose, kui agregeeritus ühel kaugusel vaheldub regulaarsusega mingil teisel kaugusel. Naabrite tiheduse jaotus on tundlikum, kui tsoonid ei ole liiga laiad
- Naabrite tihedus ei sõltu uurimisala formaalsetest servadest. Ripley *K(t)* on tugevasti sõltuv servaefektist ja seetõttu tuleks uuritava ala servade või laigulise ala laikude servade lähedal kasutada servakorrektuure.

Naabrite tiheduse jaotuse puudustena mainitakse järgmist.

- Liialt kitsaste kaugustsoonide puhul kaotab tihedusfunktsioon üldistusvõime.
- Naabrite tiheduse arvutamisel on arvutuste maht suurem kui *K(t)* puhul, eriti kui tiheduse arvutamisel mõõdetakse pinna suurust ja iga koha sobivuskaalu väga suurest pinnaüksuste hulgast. *K(t)* arvutamiseks ei ole tarvis pinda elementaarüksusteks jagada – lähteandmetena piisab *x* ja *y* koordinaatidest, kuid pinna ebaühtluse arvestamine Ripley *K(t)* puhul viib arvutuste mahu sama suureks kui naabritiheduse puhul.
- Naabrite tiheduse jaotuse kirjeldamine eeldab kauguste jagamist kas diskreetseteks kaugustsoonideks, libiseva akna või muu diferentsimise või interpoleerimise meetodi kasutamist. Andmete silumine või tsoonidesse jagamine on aga originaalandmete üldistus ning naabrite tiheduse jaotus sõltub oluliselt tsooni või silumisel kasutatava akna laiusest. Tänu kumulatiivsusele Ripley *K(t)* funktsiooni kuju tsoonide tihedusest oluliselt ei sõltu.

4.1.2.21. Radiaaljaotuse tuletis

Radiaaljaotuse tuletis ehk *K2* statistik (Schiffers *et al*. 2008) näitab naabrite tiheduse muutumist kaugusvahemiku muutumisel. Erinevalt *L(t)-t* statistikust ei ole *K2* kumulatiivne, mis tähendab, et naabrite sagedus lähemas ümbruses ei mõjuta statistiku suuremas raadiuses arvutatud väärtusi (joonis 4-15). *K2* statistikut ja selle olulisust saab arvutada R keskkonna tarkvarapaketis *spatstat* (Baddeley ja Turner 2005).



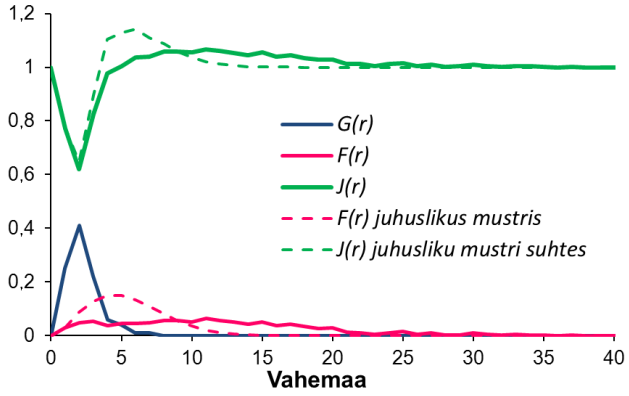
Joonis 4-15. *L(t)-t* statistik, radiaaljaotus *g(r)* ja radiaaljaotuse tuletis (*K2*) sõltuvalt vahemaast arvutatuna joonisel 4-6A esitatud punktmustrist. *K2* näitab naabrite tiheduse muutumist etteantud vahemaa juures.

4.1.2.22. J-funktsioon

J(r) funktsioon (van Lieshout ja Baddeley 1996) on määratletud lähima naabri vahemaade jaotuste kaudu – lähima naabri leidmise tõenäosusena raadiuses *r* mõõdetuna olemasolevatest punktidest [*G(r)*] ja juhuslikest lähtekohtadest [*F(r)*]. Ühest lahutamisel saab tõenäosuse, et etteantud raadiuses ei ole ühtegi objekti.

$$J(r) = \frac{1 - G(r)}{1 - F(r)} \qquad [4\text{-}42]$$

Statsionaarse juhusliku punktmustri korral on *J(r)* oodatav väärtus 1. Suuremad väärtused viitavad korrapärasele mustrile, väiksemad grupeerumisele. Lähimate tegelike kauguste mõõtmine objektide juhupunktidest arvestab olemasoleva paiknemise ebaühtlust, valemi 4-23 kasutamisel võrreldakse olemasolevat mustrit täielikult juhusliku paiknemisega (joonis 4-16). *J(r)* funktsiooni on arendatud võrdlemaks antud punktide või joonte paiknemist juhusliku punktmustri punktide paiknemisega (Foxall ja Baddeley 2002).



Joonis 4-16. *G(r)*, *F(r)* ja *J(r)* funktsioon joonisel 4-6A esitatud punktmustrist. *J(r)* funktsioon toob esile lähima naabri sagedama paiknemise vahemaal 1 kuni 4 ühikut ning mõningase tõukumistendentsi vahemaal 5 kuni 35 ühikut. Kasutatud mustris on objektideta piirkonnad suuremad kui objektide juhusliku paiknemise korral oodatav, seetõttu on suuremad kaugused juhupunktist lähima objektini sagedamad.

4.1.2.23. Punktmustri anisotroopia

Anisotroopia ehk suunalisus tähistab suunast sõltuvate struktuuride või paiknemise seaduspärasuste olemasolu. Punktmustri puhul võib suunast sõltuda eelkõige punktide tiheduse autokorrelatsioon, see tähendab, et mustri tihedamad ja hõredamad piirkonnad on valdavalt kindla suunaga – ühes suunas kitsamad, teises suunas laiemad.

Punktmustri anisotroopia kirjeldamiseks on vähemalt kolm lähenemisviisi. Esiteks võib moodustada punktide lokaalse tiheduse pinna ja kirjeldada seda numbrilise muutuja pinnana näiteks kasutades suunaga (auto)korrelogramme. Teine võimalus on kasutada tavapäraseid punktmustri kirjeldamise meetodeid igas etteantud suunavahemikus ehk sektoris eraldi. Esimesel juhul tuleb eelnevalt otsustada lokaalse tiheduse arvutamise ulatus, teisel juhul kasutatavad suunavahemikud. Kolmas võimalus mustris domineerivate suundade määramiseks on spektraalanalüüs (Mugglestone ja Renshaw 1996) ja lainekeste analüüs (*wavelet analysis*) (vt ka ptk 4.3.5), mis on vahend sinusoidsete struktuuride modelleerimiseks.

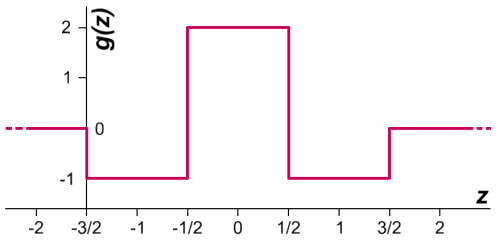
Lainekeste analüüs käsitleb tüüpilisel juhul ajutisi võnkumisi, nagu maavärinad, mis tekivad ja siis hääbuvad. Ökoloogiasse tõid lainekeste analüüsi Bradshaw ja Spies (1992), kirjeldades selle meetodiga puistu struktuuri. Lainekeste analüüsi on kasutatud ka klimatoloogias, mullastiku, metsa alustaimestiku ja maakatte kirjeldamisel. Täpsemad viited leiab kirjandusest (Saunders *et al.* 2005, Schröder ja Seppelt 2006). Lainekeste analüüs modelleerib muutlikkust lähendades andmetele erineva mõõtkavaga ja erineva lähtekohaga funktsioonide lineaarset kombinatsiooni.

Rosenberg (2004) järgi toimub lainekeste analüüsi kasutus punktmustri anisotroopia kirjeldamiseks järgmiselt. Iga punkti ümbrus jagatakse kitsasteks sektoriteks. Igasse sektorisse jäävate naabrite arv loendatakse. Kui vastassuunad on ühetähenduslikud, siis nende loendid ühendatakse. Kui sektorid olid ühe nurgakraadi laiused, siis saadakse 180 loendist koosnev jada iga lähtekoha kohta. Neid jadasid üldistatakse lainekeste funktsiooni abil, mis on kohalik teisendus. Võimalikke kohalikke teisendusi on mitmeid, näiteks niinimetatud kaabufunktsioon (*French top hat*) on kolmeosaline (joonis 4-17). Pane tähele, et kaabufunktsiooni positiivsete väärtuste osa katab ühikulise laiusega osa muutuja *z* teljest.

$$g(z) = -1, \text{ kui } 1/2 < |z| < 3/2, \quad [4\text{-}43]$$

$$g(z) = 2, \text{ kui } |z| < 1/2, \quad [4\text{-}44]$$

$$\text{muul juhul } g(z) = 0. \quad [4\text{-}45]$$



Joonis 4-17. Kaabufunktsioon.

Ruumistruktuurile iseloomuliku mõõtkava selgitamiseks rakendatakse lainekese teisendust erinevates mõõtkavades ehk erineva ruumilise sammuga (b_k) ja igal positsioonil (x_i) etteantud ulatuses olevatel naaberpositsioonidel (x_j) olevatele väärtustele $y(x_j)$ kogu n loendist koosneva transekti ulatuses. Punktmustri kirjeldamisel on etteantud ulatuses olevaks väärtuseks punktobjektide arv antud sektoris.

$$W(b_k, x_i) = \frac{1}{b_k} \sum_{j=1}^{n} y(x_j) g\left(\frac{x_j - x_i}{b_k}\right) \quad [4\text{-}46]$$

Sobivaim mõõtkava on suurima hajuvusega $V(b_k)$.

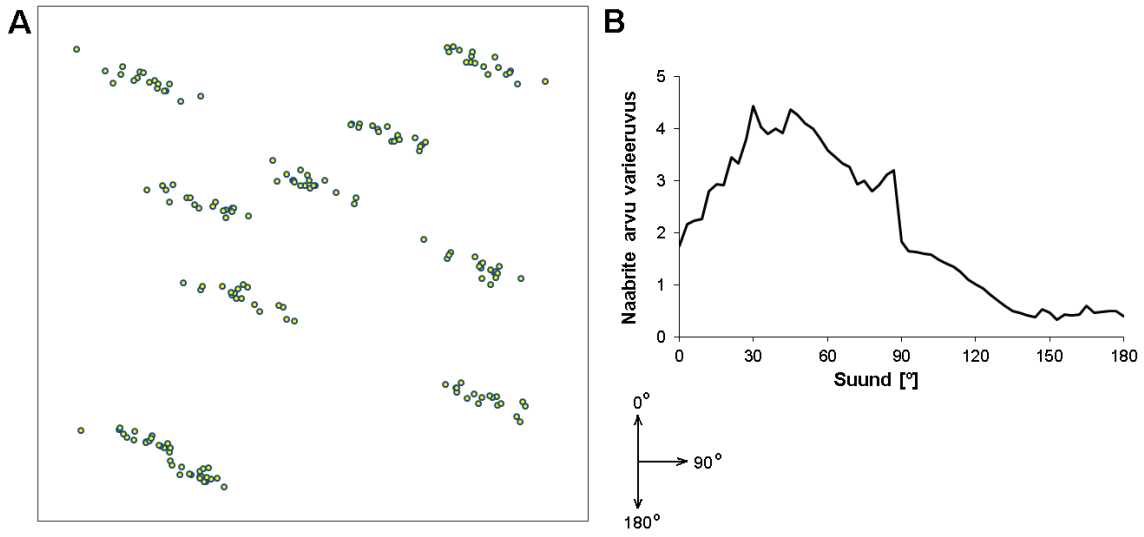
$$V(b_k) = \frac{1}{n} \sum_{i=1}^{n} W^2(b_k, x_i) \quad [4\text{-}47]$$

Iga suunaga suundade reas ehk positsiooniga transektil seotud hajuvus arvutatakse üle erinevate mõõtkavade m.

$$P(x_i) = \frac{1}{m} \sum_{k=1}^{m} W^2(b_k, x_i) \quad [4\text{-}48]$$

Tulemusi visualiseeritakse näidates joongraafikul hajuvuse [$V(b_k)$] sõltuvust mõõtkavast või lokaalse hajuvuse $P(x_i)$ sõltuvust suunast (joonis 4-18) või siis pinnana, mis kujutab $W(b_k,x_i)$ sõltuvust nii suunast kui ka mõõtkavast. Lainekeste analüüsi tulemuste statistilist olulisust saab kontrollida sama arvu punktobjekte korduvalt uuritavale alale juhuslikult paigutades.

Lainekeste analüüsi saab kasutada ka ruumiliste struktuuride mõõtkava määramisel (ptk 4.3.5), transektilt saadud vaatlusandmetes korrapära otsimisel ja objektide paiknemismustri kirjeldamisel transektide abil (Dale ja Mah 1998, Perry *et al*. 2002). Kõige informatiivsem on mõõtkava, mille juures on valemis 4-46 esitatud W statistiku hajuvus suurim.



Joonis 4-18. Punktmustri anisotroopia kirjeldamine lainekeste funktsiooniga. A – punktobjektid suunatud kobaratena. B – naabrite keskmise arvu $g(x)$ funktsiooni varieeruvuse sõltuvus suunast.

Uurimused

Ripley *K(t)* funktsiooni või *L(t)–t* statistikut on kasutatud väga paljudes töödes, eriti puistu struktuuri analüüsis. Sterner *et al*. (1986) kasutasid *L(t)* statistikut puude paiknemise seaduspärasuste uurimiseks troopilises metsas. Szwagrzyk ja Czerwczak (1993) määrasid puude paiknemise seaduspärasusi Ripley *K(t)* funktsiooni abil ning seletasid puuliikide mõningast eraldihoidmist üle 15 m vahemaa puhul Kesk-Euroopa põlismetsas eelkõige kasvukoha varieeruvusega. Grau (2000) uuris Lõuna-Ameerika puu *Cedrela lilloi* regeneratsioonimustrit Ripley *K(t)* abil ja leidis, et noortaimed on küll grupeerunud, kuid ei paikne reeglina vahetult vanade puude lähedal (kaugusel <15 m), nende tihedus on statistiliselt olulisel määral üle keskmise >50 m kaugusel sama liigi suurtest puudest. Kuna teiste liikide suhtes oli *C. lilloi* paiknemine juhuslik, siis järeldas autor, et selle liigi noorte isendite esinemismustrit mõjutab just sama liigi vanade puude paiknemine. Vanade puude raiumisel tekkivad häilud ei pruugi liigi järelkasvu suurendada, kuna nendes kohtades noortaimed puuduvad.

Chen ja Bradshaw (1999) uurisid Ripley *K(t)* abil Hiinas asuva Changbaishani looduskaitseala põlismetsa struktuuri. Nad väidavad, et puude paiknemine põlismetsas ei ole nii juhuslik kui on näidanud paljud varasemad uuringud (Szwagrzyk 1990, Tomppo 1986). Samuti ei vii puudevaheline konkurents ja puude loomulik surm korrapärase paiknemismustrini. Pigem muudab konkurents noorte puude grupeerunud paiknemise juhuslikuks. Näiteks enamik kuuski tärkab kõdunevatel tüvedel, kuid suurte kuuskede paiknemine ei ole agregeerunud. Metsapuude paiknemismustri laigulisuse olulisemate põhjustena tuuakse välja ühelt poolt järglaste piiratud levikut ja suurte põlispuude väljalangemisest tekkinud häile ning teiselt poolt suuremate esimese rinde puude paiknemist. Suurte puude paiknemine võib ise olla suhteliselt juhuslik, kuid määrab seejuures väiksemakasvuliste puude paiknemist.

Gappa *et al.* (1997) uurisid tõruvähi *Balanus amphitrite* isendite paiknemist kaardistades neid 12 korda ligi nelja aasta jooksul samal 0,8 m^2 suurusel vaatlusalal Argentina rannikul. Lähima naabri keskmise kauguse järgi olid isendid enamik aega agregeerunud. Asustustiheduse languse järel muutus ruumiline koondumine statistiliselt ebaoluliseks.

North ja Greenberg (1998) kasutasid Ripley *K(t)* funktsiooni teralise hirvepähkli (*Elaphomyces granulatus*) liigisisese ja puude suhtes paiknemise analüüsiks. Autorid näitasid, et hirvepähklid hoiavad *Tsuga heterophylla* isenditest eemale statistiliselt olulisel määral.

Radiaaljaotust on metsa struktuuri kirjeldamiseks kasutanud Ward *et al*. (1996), Wiegand *et al.* (2007b, 2009) ja Yu *et al.* (2009). Wiegand *et al*. (1999) käsitlesid maastikumustri piksleid punktobjektidena ja arvutasid *O*-ring statistiku abil tõenäosust leida ümbruses sama tüüpi maastikku.

Getzin *et al*. (2006) mõõtsid paarilise korrelatsiooni abil puuliikide liigisisest ja liikidevahelist konkurentsi. Getzin *et al.* (2011) määrasid lisaks puude paiknemisele ka puude rinnaspinna ja juurdekasvu paiknemise seaduspärasid kasutades märgikorrelatsiooni. Hao *et al.* (2007) ja Zhang *et al.* (2010) analüüsisid *O*-ring statistiku abil puistu erinevates rinnetes olevate erinevat ja sama liiki puude vahelisi paiknemissuhteid ja konkurentsi.

Puuliikide liigisiseseid ja liikidevahelisi paiknemise seaduspärasusi Jaapani metsas kirjeldasid Morisita indeksite abil ja sõltuvalt vaatlusalade suurusest Miyadokoro *et al.* (2003).

Li *et al.* (2009) kirjeldasid puude paiknemist Hiina metsas suhtelise naabrustiheduse indeksi abil. Puude paiknemismuster sõltus liigi levimisviisist, liigi ohtrusest, puude suurusest ja kasvukoha omadustest.

Jacquemyn *et al.* (2010) uurisid nurmenuku noortaimede ja täiskasvanud taimede suhteid Flandria kraavikallastel *O*-ring statistiku ja paarilise korrelatsiooni abil. Leiti, et noortaimede ellujäämus ei sõltu täiseas taimede paiknemisest, vaid majandamisviisist. Nurmenuku populatsioonide püsimiseks

tuleks kraavikaldad aeg-ajalt puhtaks rehitseda.

Rayburn *et al.* (2011) iseloomustasid taimede liigisisese paiknemise seaduspärasusi K2 statistiku abil ja liikidevahelist paiknemist *O*-ring statistiku abil USA lääneosa põõsastikus. Koondusid eelkõige sama liigi isendid, liikidevahelist koondumist leiti vaid kahe liigi puhul kahekümnest paarist. Rayburn ja Wiegand (2012) kirjeldasid põõsaste paiknemist paarilise korrelatsiooni abil.

Tarkvara

Punktide lokaalse tiheduse arvutamise vahendeid leiab igast suuremast geoinfo tarkvarapaketist.

Ripley *K(t)* funktsiooni koos nullmudeli usalduspiiridega saab arvutada tarkvaras ArcGis. Servakorrektuuridest on saadaval punktobjektide genereerimine väljapoole uurimisala, lähtekohtadest loobumine uurimisala serva lähedal ning servalähedastele vaatluste kaalu muutmine.

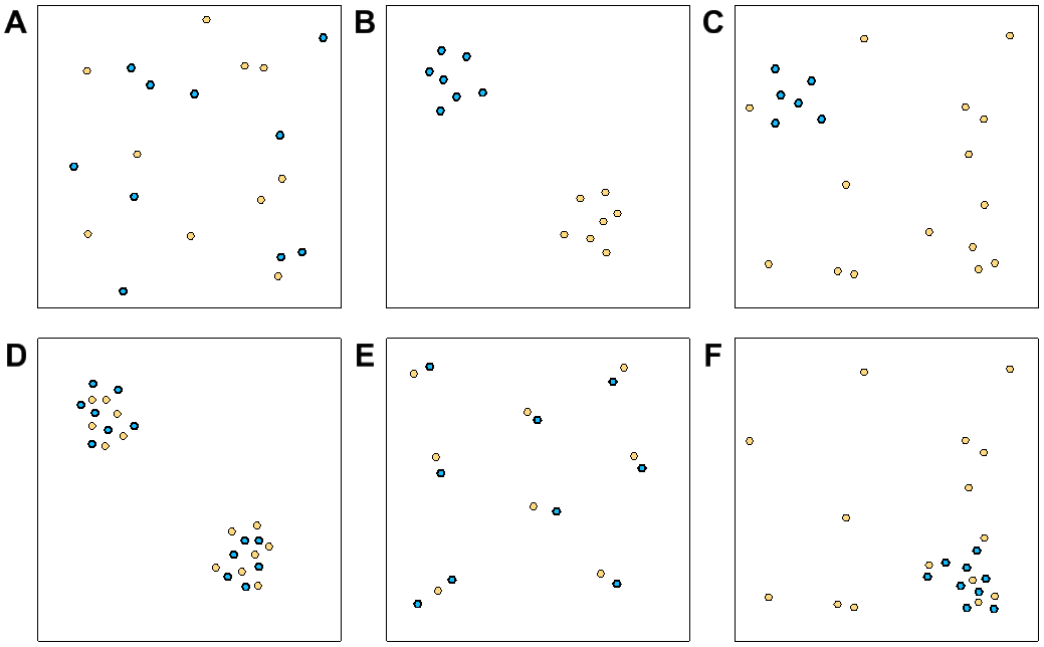
Ripley *K(t)* on suhteliselt lihtsasti arvutatav ka, kui on olemas kõigi punktide asukohakoordinaadid ristkoordinaatidena, millest tuleb moodustada kõigi omavaheliste kauguste maatriks.. Kuna *K(t)* arvutamiseks on vaja otsida teatud raadiusesse jäävate naabrite arvu, siis tuleb loendada teatud vahemaast väiksemale kaugusele jäävad objektidevahelised kaugused. Seda on võimalik teha Exceli funktsiooniga COUNTIF. Teine võimalus on järjestada kaugused kasutades Exceli funktsiooni SMALL(andmeplokk, järjekorranumber) või LARGE(andmeplokk, järjekorranumber). Kuna vahemaade maatriksi diagonaalidel olevad nullid (punktide kaugused iseendaga) ei peaks arvesse minema, siis tuleb väikseimate kauguste arvust maha lahutada maatriksi küljepikkus.

Punktmustrit kirjeldavaid näitajaid saab arvutada ka tarkvarapaketis PASSaGE (Rosenberg ja Anderson 2011, http://www.passagesoftware.net). Samuti programmiga NDENS, mis on saadav selle arvuti veebilehelt ja mis arvutab nii $L(t) - t$ statistikut kui ka naabrite tihedust. Programm võimaldab arvestada pinna sobivust.

4.1.3. Punktmustrite paiknemissuhe

Ühte tüüpi objektide ruumilist koondumist nimetatakse **agregatsiooniks**, mitme tüübi esindajate ruumilise koondumise tendentsi korral räägitakse **assotsieerumisest**. Punktmustrite assotsieerumine mingil kaugusel näitab, et sellel kaugusel on naaberpunkte rohkem kui juhuslikust protsessist oodata võiks. **Segregatsiooni** (*segregation*) ehk eraldatuse puhul on objektitüüpidel tendents üksteist vältida (joonis 4-19). **Assotsieerumise analüüsiks** (*association analysis*) nimetatakse igasugust seoste uurimist, terminit **segregatsiooni analüüs** (*segregation analysis*) kasutatakse eelkõige geneetikas. Ruumimustrite omavahelise paiknemise analüüsi puhul tuleks ära märkida, et silmas peetakse ruumilist assotsieerumist või segregatsiooni. Assotsieerumise, segregatsiooni ja sõltumatu paiknemise üldiseks koondnimetus võiks olla **paiknemissuhe** (*spatian relation*).

Punktmustrite kirjeldamiseks kasutatavaid parameetreid (nagu kaugus lähima naabrini, Ripley *K(t)*, naabrite tihedus) saab kasutada ka paiknemissuhte mõõtmiseks. Kahe mustri paiknemissuhte mõõtmisel tuleb eristada lähteobjektid (*i*) ja sihtobjektid (*j*). Paiknemissuhte arvutamisel määratakse iga lähteobjekti paiknemisstatistik vaid sihtobjektide suhtes. Kui kaks punktmustrit paiknevad teineteise suhtes juhuslikult, siis Ripley *K(t)* kasutamisel: $K_{ij}(t) = K_{ji}(t) = \pi t^2$ ning $L(t)–t$ statistiku positiivsed väärtused näitavad liikide atraktsiooni ja negatiivsed väärtused segregatsiooni. Ripley *K(t)* kasutamisel kahe punktmustri võrdlemisel tuleb valemis olev N^2 asendada ühte liiki punktide arvu ja teist liiki punktide arvu korrutisega. Kui üheväärtuselise punktmustri kirjeldamisel *K(t)* funktsiooniga on probleemiks servakorrektuurid, siis kahe punktmustri võrdlemisel võib servakorrektuuridest loobuda, eeldadades, et servaefekt mõjutab mõlemat mustrit ühetaoliselt.



Joonis 4-19. Kahte liiki punktobjektide paiknemise näited: A – juhuslik; B – liigisiseselt agregeerunud ja liikidevaheliselt segregeerunud; C – segregeerunud ja erineval määral agregeerunud; D – assotseerunud ja agregeerunud; E – assotseerunud ja segregeerunud; F – assotseerunud ja erineval määral agregeerunud.

Punktmustrite paiknemissuhte kirjeldamiseks saab kasutada samu statistikuid, mida kasutatakse ühe mustri kirjeldamisel

Paariviisilise korrelatsiooni oodatav väärtus juhusliku paiknemise korral on 1 ja naabrite tiheduse väärtus on võrdne naabrite keskmise tihedusega. Seejuures võib punktmustrite kovarieerumine sõltuda mõõtkavast. Eri klassi punktid võivad näiteks väikestel kaugustel üksteist vältida, suurematel kaugustel aga koonduda.

Paiknemissuhe ei pruugi olla sümmeetriline ega eri kaugustel sama

Kahe eri klassi objektide omavahelist atraktiivsust saab võrrelda ka naabritevahelisi kaugusi ja lähima naabri kaugust kasutavate statistikute abil. Mõned meetodid on esitatud järgnevalt.

Morisita (1959b) liikidevahelise korreleerumise indeksis kasutatakse sama autori agregatsiooniindeksit.

$$R_{\delta} = \frac{2(q\sum x_i y_i - N_x N_y)}{q(\delta_x + \delta_y)N_x N_y}, \qquad [4\text{-}49]$$

kus x_i on esimest liiki objektide arv loendis i, y_i on teist liiki objektide arv loendis i, δ_x on ühe liigi agregatsiooniindeks, δ_y on teise liigi agregatsiooniindeks, N on objektide üldarv ja q on loendite arv.

Pielou (1961) klassifitseeris iga indiviidi vastavalt liigikuuluvusele ja vastavalt tema lähima naabri liigikuuluvusele. Igasse klassi kuuluvad isendid loendati (tabel 6).

Tabel 6. Kahe võrreldava liigi isendite loendid.

	Liik A	Liik B	Kokku
Lähim naaber A	*a*	*b*	*m*
Lähim naaber B	*c*	*d*	*n*
Kokku	*r*	*s*	*N*

Tabelis olevatest loenditest saab arvutada segregatsiooni tugevust näitavat indeksit *S*, mis võrdub üks miinus liikide A ja B omavahel lähimaks naabriks olemise vaadeldud ja juhuslikkuse korral oodatava sagedusega.

$$S = \frac{\text{segapaaride vaadeldud arv}}{\text{segapaaride arvu ootus}} = 1 - \frac{N(b+c)}{ms+nr} \qquad [4\text{-}50]$$

Kui lähim naaber on alati samast liigist, siis $S = 1$. Kui kõik nende liikide isendid külgnevad teise liigiga ja mõlemat liiki on ühepalju, siis $S = -1$. Liikide ebavõrdse sageduse korral on *S* miinimumväärtus −1 ja 0 vahel. Liikide teineteise suhtes juhusliku paiknemise korral $S = 0$.

Pielou test on lihtne, aga kasutab vaid lähima naabri andmeid. Järgmiste naabrite kaugus enam rolli ei mängi. Paiknemisindeksi *S* mitte-juhuslikkuse statistilist olulisust on määratud hii-ruut testiga, kuigi lähimaks naabriks olemise kordade arvud ei ole sõltumatud loendid, nagu χ^2 test seda eeldab. Seetõttu annab χ^2 test olulisuse hälbinud hinnangu (Maegher ja Burdick 1980).

Dixon (1994) kasutas lähima naabri andmetest segregatsiooni mõõtmisel teist lähenemisviisi. Selle asemel, et oletada nullhüpoteesina punktide juhuslikku paiknemist teise klassi punktide suhtes, lähtutakse klassikuuluvuse juhuslikkusest ja paiknemise juhuslikkust ei eeldata. Isendi ja tema lähima naabri liigikuuluvuse kombinatsioonide oodatavad sagedused avalduvad isendite üldarvu *N*, liigi *A* isendite arvu N_A ja liigi *B* isendite arvu N_B kaudu.

$$EN_{AA} = N_A \frac{N_A - 1}{N - 1} \qquad [4\text{-}51]$$

$$EN_{AB} = N_A \frac{N_B}{N - 1} \qquad [4\text{-}52]$$

$$EN_{BB} = N_B \frac{N_B - 1}{N - 1} \qquad [4\text{-}53]$$

$$EN_{BA} = N_B \frac{N_A}{N - 1} \qquad [4\text{-}54]$$

EN_{AA} on kombinatsiooni: liigi A esindaja on lähimaks naabriks liigi *A* esindajale, oodatav sagedus. Teiste kombinatsioonide tähistus on analoogiline. Üks lahutatakse, kuna endale ei saa naabriks olla. Kui oodatav sagedus jagada isendite koguarvuga, siis saame tõenäosuse, et juhuslikult valitud naabritepaaris on eeltoodud valemile vastav liigikombinatsioon.

Dixon (1994) esitas segregatsiooniindeksit kujul

$$S_{AA} = \ln\left[\frac{N_{AA} / N_{AB}}{(N_A - 1) / N_B}\right], \qquad [4\text{-}55]$$

kus N_A ja N_B on liikidesse *A* ja *B* kuuluvate objektide arv, N_{AA} on liiki *A* kuuluvate objektide arv, mille lähim naaber kuulub samasse liiki, N_{AB} on liiki *A* kuuluvate objektide arv, mille lähim naaber kuulub liiki *B*.

$S_{AA} > 0$ näitab, et liigi *A* naabriteks on enamasti liigikaaslased, seega liikidevahelist segregatsiooni; $S_{AA} < 0$ näitab liikidevahelise assotsieerumise tendentsi. Analoogiliselt saab arvutada segregatsiooni-indeksi liigi *B* jaoks.

Dixon (1994) soovitab ka statistikut (Dixoni *C*) lähima naabri kombinatsioonide sageduste ja ka segregatsiooniindeksi statistilise olulisuse hindamiseks.

Coomes *et al.* (1999) soovitasid indeksit

$$SI_{ij}(r) = \sum_{i=1}^{N_i} N_{ij}(r) \frac{\pi r^2 N_i N_j}{Q}, \qquad [4\text{-}56]$$

kus N_{ij} on liigi *i* liiki *j* kuuluvate naabrite arv raadiuses *r* ja *Q* on ruudu pind. SI_{ij} positiivsed väärtused näitavad assotsieerumist, negatiivsed segregatsiooni.

Shimatani (2001) esitas kaks indeksit *α(r) ja β(r)*, mis on vaatluste erinevatesse kategooriatesse kuulumise tõenäosus vastavalt kaugusest *r* väiksema vahemaaga vaatluspaaris ja vaatluspaaris vahemaaga *r*.

$$\alpha(r) = 1 - \sum_m \frac{\lambda_m^2 K_m(r)}{\lambda^2 K(r)}, \qquad [4\text{-}57]$$

$$\beta(r) = 1 - \sum_m \frac{\lambda_m^2 g_m(r)}{\lambda^2 g(r)}, \qquad [4\text{-}58]$$

kus *K(r)* on Ripley *K* funktsioon, *g(r)* on radiaaljaotuse funktsioon, *m* on liigi indeks.

Uurimused

Harkness ja Isham (1983) uurisid kahte liiki sipelgate pesade omavahelist paiknemist. Nullhüpoteesile vastav pesade paiknemine saadi ühe liigi pesade paiknemise asendamisel sama arvu pesade juhupaigutusega. Lähima naabri meetod näitas teatud lähima teise liigi pesa kauguse sagedamat esinemist kui nullmudeli järgi peaks olema. Ripley *K(t)* test statistiliselt olulist seost liikide paiknemises ei näidanud.

Duncan (1991) näitasid Ripley *K(t)* testi abil kahe puuliigi *Dactrycarpus dacrydioides* ja *Dacrydium cupressinum* liigisisese konkurentsi mõju puude paiknemisele, kuid ei leidnud märke nende liikide omavahelisest konkurentsist. Need liigid on puistu jaganud kummalegi soodsama kasvukoha laikude järgi.

Peterson ja Squiers (1995) näitasid *L(t)* testi abil, kuidas valge männi (*Pinus strobusi*) järelkasvu ruumiline koondumine sõltub vanemate haabade (*Populus grandidentata* ja *P. tremuloides*) paiknemisest ja hukkumisest.

Lynch ja Moorcroft (2008) näitasid Ripley *K(t)* funktsiooni abil, et metsapõlengute ja puistu mähkurikahjustuste vahel on Briti Columbias ruumiliselt 25 km ulatuses ja ajaliselt 6 aasta ulatuses negatiivne seos. Üldistades mähkurikahjustuse sagedust põlengu järel ja vastupidi leiti, et mähkurikahjustus ei soodusta metsapõlenguid ja metsapõlengud ei soodusta mähkurikahjustusi, nagu võiks arvata.

4.1.4. Statistilised testid punktmustritele

Punktmustrite testimised jagunevad mustri tüübi testimiseks ja mustrite vahelise suhte testimiseks. Need kaks testi tüüpi on omavahel sõltumatud. Seega on kõik kombinatsioonid võimalikud, näiteks võivad kaks eraldi vaadates juhuslikuna paistvat mustrit omavahel tugevasti assotsieeruda ning kaks agregeeritud mustrit võivad teineteise suhtes juhuslikult paikneda. Nullhüpoteesiks on mustri testimisel **täiesti juhuslik paiknemine** (*complete spatial randomness*) ja mustritevahelise suhte testimisel liikide omavahelist sõltumatust eeldav **juhuslik märgistamine** (*random labelling*). Juhuslik märgistamine säilitab objektide asukohad, juhuslikult omistatakse vaid klassikuuluvus.

Enamikku eelnevates alapeatükkides mainitud ühe mustri juhuslikkusest erinevuse määramise meetodeid saab kasutada ka mitme mustri assotsieerumise kontrollimiseks. Ülevaateid punktmustrite statistilistest testidest võib leida järgmistelt autoritelt, Diggle (1979a), Cormack (1979), Upton ja Fingleton (1985, lk. 60), Kulldorff (2006).

4.1.4.1. Kaugusmeetod

Ökoloogiliste uuringute osaks on sageli loomade või taimede ohtruse (tiheduse, arvukuse, katvuse) määramine. Traditsiooniliselt toimub olendite tiheduse hindamine proovialadel ja hinnang saadakse hulga ja pindala suhtena. Lisaks proovialadele tuginevale proovivõtu metoodikale on olemas ka kauguste ja suundade mõõtmisele tuginevad populatsiooni tiheduse meetodid. **Kaugusmeetodite** (*distance sampling*) puhul võib kauguse mõõtmise lähtekohaks olla olemasolev objekt, (juhuslik) koht vaatlusalal või vaatlustrass. Kaugusmeetodite kasutamise eripäraks on objekti avastamise tõenäosuse ja hinnangute täpsuse sõltuvus kaugusest. Kaugusmeetodite puhul määratakse ruumiliste punktmustrite parameetreid mõõtes vahemaid (Holgate 1972). Põhjaliku ülevaate kaugusmeetoditest on kirjutanud S.T. Buckland *et al*. (1993).

Kaugusmeetodi statistilisi teste on mitmeid. Mõned neist on kirjeldatud järgnevalt. Ülevaate kauguste abil paiknemise juhuslikkuse kontrolliks kasutatavatest testidest annab Cormack (1979).

Clark ja Evans (1954) testisid lähima naabri kauguse keskväärtusi juhuslikkuse suhtes. Nende test ignoreeris aga vastastikuste naabrite olemasolu (Diggle 1979a).

Hopkins-Moore test kasutab *h* statistikut, mis on nullhüpoteesi kehtivuse korral beetajaotusega *B(m,m)* (Hopkins ja Skellam 1954, Moore 1954)

$$h = \frac{\sum X_i^2}{\sum X_i^2 + \sum Y_i^2} \qquad [4\text{-}59]$$

kus X_i on kaugus juhuslikust punktist *i* lähima sündmuseni, Y_i on kaugus juhuslikult valitud sündmusest *i* lähima teise sündmuseni, *m* on sündmuste ja juhupunktide arv.

Holgate (1965) soovitas statistikut Z_{st}, mis kasutab kaugusi *m* juhupunktist *s*-nda ja *t*-nda lähima sündmuseni ja on nullhüpoteesi kehtimisel beetajaotusega vabadusastmete arvuga *2ms* ja $2m(t-s)$.

$$Z_{st} = \frac{\sum X_s^2}{\sum X_t^2} \qquad [4\text{-}60]$$

Diggle (1983) teisendas Holgate statistiku kujule

$$H_N = \frac{\sum_{i=1}^{m} \frac{X_{1i}^2}{X_{2i}^2}}{m}, \qquad [4\text{-}61]$$

milles on kaugused $i = 1...m$ vaatluskohast esimese ja teise lähima objektini. H_N on objektide juhusliku paiknemise korral ligikaudu normaaljaotusega keskväärtusega 0,5 ja dispersiooniga $(12m)^{-1}$, kui $m \geq 10$.

Besag ja Gleaves (1973) esitasid h_B statistiku, mis kasutab vahemaid Z sündmusest Q (mis on lähim juhuslikule punktile P) lähima naabersündmuseni pinna selles osas, mis on punkti P suhtes teisel pool punkti Q läbivat joont, mis on risti joonega PG.

$$h_B = \frac{\sum X_i^2}{\sum X_i^2 + \sum \frac{Z_i^2}{2}}, \qquad [4\text{-}62]$$

kus X_i – vahemaa punktide P ja Q vahel. h_B on sündmuste juhusliku paiknemise korral beetajaotusega $B(m,m)$. See test on võimsam kui Holgate test, sest X ja Z on omavahel sõltumatud (Diggle 1979a).

4.1.4.2. Monte Carlo test

Punktmustri juhuslikust erinevuse Monte Carlo testi (ptk 3.6.6) saab rakendada korraga kogu punktmustri suhtes kui ka lokaalselt, eraldi uuritava ala igas alajaotuses, iga objekti ümber või korrapärase võrgustiku sõlmekohtades. Jäljendatud paiknemismustrite ning Aafrika poolkõrbes esineva tiigerpõõsa isendite paiknemismustri lokaalset koondumist ja üksteise vältimist ning paiknemismustri erinevuse olulisust juhusliku paiknemise nullmudelist kaardistasid näiteks Couteron *et al.* (2003). Ruumiliselt ilmutatud populatsioonimudelite Monte Carlo testides kasutamise näiteid esitavad Coomes *et al.* (1999). Juhupunktide lokaalne paigutamine võimaldab kaardistada punktmustri või eri liiki punktide omavahelise paiknemise nullmudelist erinevuse olulisust ja erinevuse suunda (joonis 4-20).

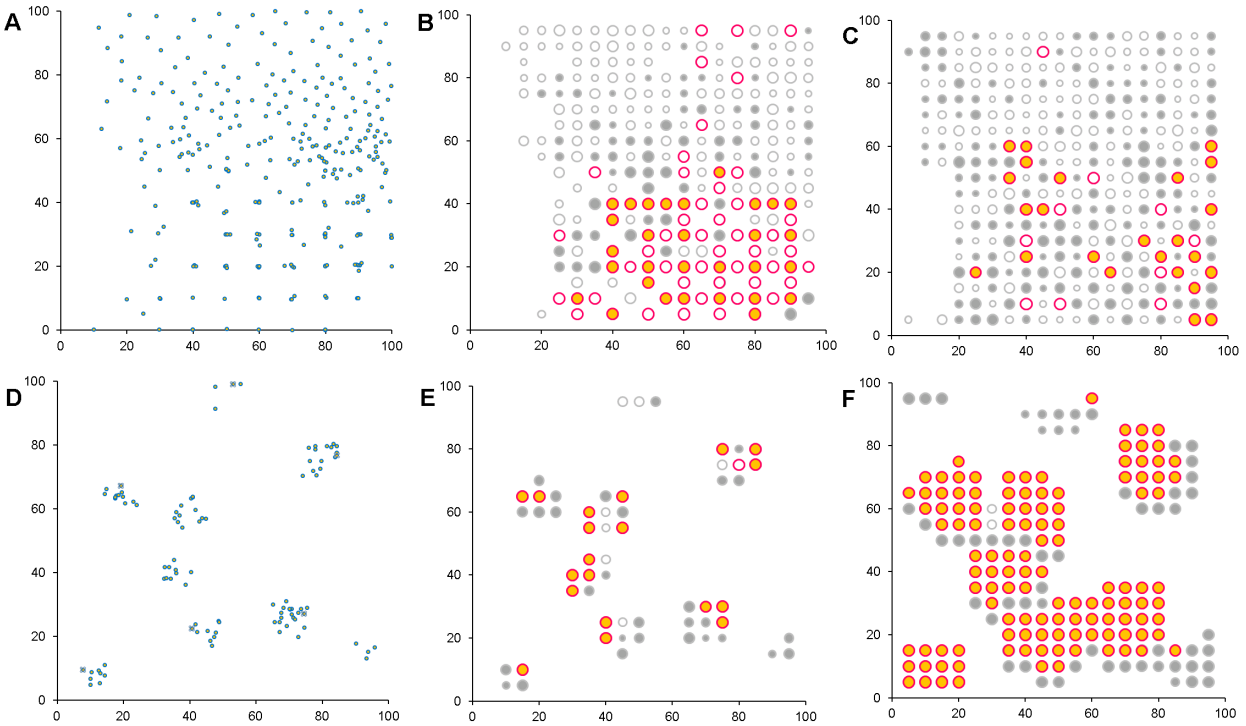
Punktmustrite Monte Carlo testiga võrreldakse teatud parameetreid: lähima naabri kaugust, vahemaad võrgustiku lähimast punktist, punktide keskmist vahemaad, vahemaad korrapärase võrgustiku silmast lähima punktini, teatud vahemaast väiksemal kaugusel olevate naabrite arvu või mõnda keerukamat statistikut. Ka naabrite tiheduse jaotuse usalduspiirid saadakse Monte Carlo meetodiga, see on korduvalt juhuslikke naabreid genereerides.

Kahe nähtuse paiknemissuhte mittejuhuslikkuse testimisel Monte Carlo meetodiga ei asendata mustreid juhuslike punktidega, vaid muudetakse juhuslikult kas seesmiselt fikseeritud mustrite asendit üksteise suhtes või siis punktsündmuste liigikuuluvust.

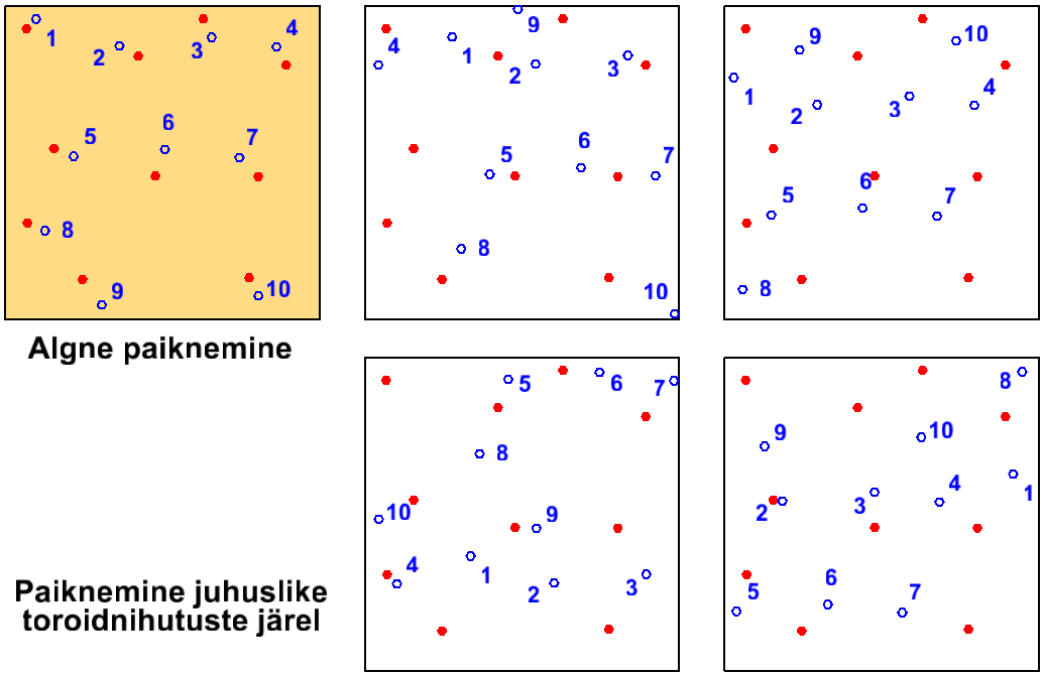
Punktmustrite omavahelise ruumilise sõltumatuse testimiseks on seega kaks põhilist moodust:

- mustri pinna vastasservade ühendamine (pööramine toruks) ja seejärel ühe mustri juhuslik nihutamine – **toroidnihutus** (*toroidal wrapping, toroidal shift*) (joonis 4-21),
- sündmuste juhuslikult ühte ja teise liiki nimetamine ehk juhuslik märgistamine, säilitades seejuures sündmuste algsed kohad (Diggle 1983).

Liikide paiknemissuhte võrdlemisel juhusliku suhte korral oodatavaga nihutatakse juhuslikul määral ühe liigi paiknemismustrit tervikuna või randomiseeritakse punktide liigikuuluvust



Joonis 4-20. Punktmustrite (A ja D) mittejuhuslikkuse lokaalselt arvutatud olulisus piirkonnas olevate objektide omavaheliste vahemaade ruutude keskmise järgi (statistiku suurem väärtus näitab paiknemist suurte vahedega, väiksem väärtus ühes grupis paiknemist) raadiuses (*r*) viis ja kümme ühikut: B – vahemaade ruutude keskmine mustrist A (*r* = 5); C – mustrist A (*r* = 10); E – mustrist B (*r* = 5); F – mustrist B (*r* = 10). Juhuslikust mustrist erinevuse olulisustõenäosus arvutati paigutades 2000 korda iga arvutatava koha ümber etteantud raadiusse tegeliku punktide arvuga võrdse hulga juhuslikult paiknevaid punkte ja arvutades nende juhupunktide omavaheliste vahemaade ruutude keskmist. Täidetud ringid näitavad koondumistendentsi, tühjad ringid eraldumist, hallide märkide kohas on juhuslikkuse tõenäosus > 0,05.



Joonis 4-21. Punaste punktide ja siniste ringide algne paiknemine ja paiknemine siniste ringide juhuslike toroidnihutuste järel. Punaseid punkte ei nihutata. Siniste ja punaste objektide juhuslikult tekkinud nii tugev ruumiline assotseerumine on äärmiselt vähetõenäoline.

4.1.4.3. Ruumiline ellujäämusanalüüs

S. Reader (2000) väidab, et kaheväärtuselise punktmustri punktide omavaheliste vahemaade kirjeldamiseks sobib ka ruumiline **ellujäämusanalüüs** (*spatial survival analysis*), mida üldiselt kasutatakse ajaliste tõenäosuste arvutamisel kliinilistel katsetel ja tehnilise kvaliteedi kontrollis. Ellujäämusanalüüsi moodulite üks eelis paljudes statistikapakettides on valmis testid kumulatiivsete jaotuste võrdlemiseks. Punktmustri analüüsil on sellisteks jaotusteks sündmuste arv lähtekohast kuni teatud kauguseni. Ruumilise ellujäämusanalüüsiga saab arvutada ka näiteks haigestunute hulka epideemiakoldest kaugenedes või haigestunute hulka ühe nakkuse levitaja ümber.

4.1.4.4. Tühimike suurus

Dale ja Powell (2001) näitasid, oodatav juhuslike punktide arv (e_k) kolme üksteisest võimalikult kaugel asuva punkti abil moodustatud ringi sees võrdub

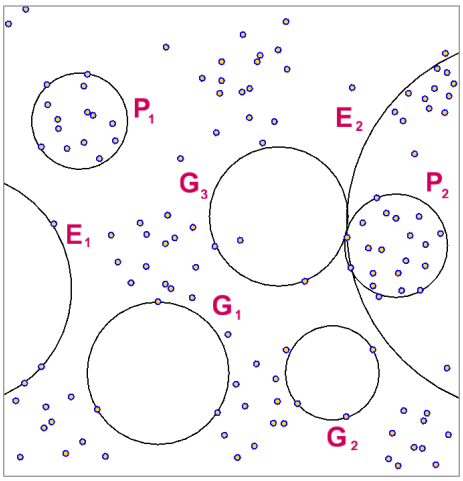
$$e_k = \frac{a_k (n-3)}{A} \qquad [4\text{-}63]$$

ning vaadeldud ja oodatavat punktide arvu saab võrrelda Z statistiku abil

$$z_k = \frac{n_k - e_k}{\sqrt{e_k}}, \qquad [4\text{-}64]$$

kus n on punktide arv uuritaval alal, n_k on punktide arv ringis, a_k on uurimisala piires oleva ringi osa pindala, A on uurimisala pindala.

Kui z_k on väiksem kui −1,96 või on suurem kui 1,96, siis võib punktide tiheduse lokaalse kõrvalekalde lugeda statistiliselt oluliseks (joonis 4-22). Dale ja Powell andsid meetodile nimeks **ümbritsevate ringide meetod** (*circumcircle method*).



Joonis 4-22. Piirkondi ümbritsevad ringid Dale ja Powell (2001) järgi: G – tühikut ümbritsev ring, P – punktide koondumispiirkonda ümbritsev ring, E – vaid osaliselt uurimisalas olev ring. Ringide *z* statistiku väärtused: P_1: 3,97; P_2: 5,47; G_1: −3,09; G_2: 1,90. Kuigi G_2 on tühi, on selle z_k liiga väike, et seda tühikut statistiliselt oluliseks pidada. G_1 ja G_2 on absoluutselt tühjad, G_3 on suhteliselt tühi. Osaliselt uurimisalast väljas olevate ringide puhul vaid uurimisala piires olevat pinda arvestades E_1: −2,59; E_2: 2,72. Kui E_2 puhul arvestada ka väljaspool uurimisala piire olevat pindala ning lugeda see tühjaks, siis oleks selle z = 6,01.

4.1.4.5. Testide hälbimise põhjused

Testid, mis määravad punktmustri erinevust juhuslikust, võivad väärtulemusi anda mitmel põhjusel: juhuslike mõõtmisvigade tõttu, punktsündmuste vastastikku lähimaks naabriks olemise tõttu, kaugusvahemike kasutamisel tulemuste vähese arvu tõttu mõnes vahemikus, heterogeensete andmete ja arvestamata kõrvaliste mõjude tõttu ning uurimisala serva läheduse tõttu. Sageli arvestamata jääv kõrvaline mõju ökoloogilistes uuringutes on punktmustri aluspinna ebaühtlased omadused, mida käsitletakse eraldi (ptk 4.1.5).

Juhuslikud vead võivad punktmustrite korral olla eelkõige asukoha määramise ja vahemaade mõõtmise vead või andmete ümberkirjutamise vead. Mõlemad vead vähendavad punktmustrist leitud tendentside statistilist olulisust, mõõtmisvead kipuvad ka struktuuride mõõtkava muutma (Freeman ja Ford 2002).

Lähima naabri kauguse kasutamisel on osa sündmusi vastastikku teineteise lähimad naabrid ja see vähendab lähima naabri kauguste hajuvust (Campbell 1996). Seetõttu ei saa lähima naabri kaugusi käsitleda ka sõltumatute vaatlustena.

Hii-ruut testi Yates korrektuuri soovitatakse kasutada, kui pideva suuruse väärtused on klassidesse jagatud. Sama korrektuuri soovitatakse ka siis, kui mõne klassi oodatav sagedus on alla 5.

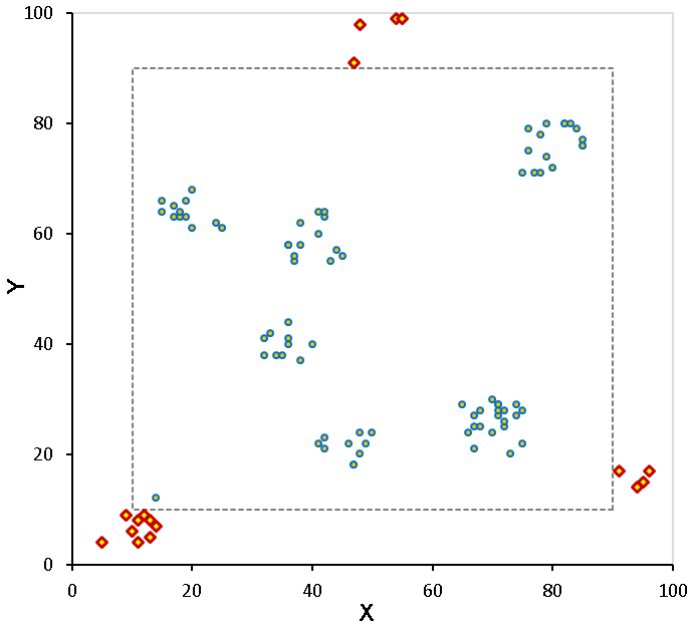
Andmete heterogeensus võib tähendada oluliste faktorite mõju ignoreerimist. Andmete mittevõrreldavust võib põhjustada vaatluste päritolu erinevatest populatsioonidest, erinevatest piirkondadest, eri aastaajast.

Servaefektide vältimise esimene nõue on, et uurimisala peab olema piisavalt suur (kindlasti suurem kui punktide rühmad). Uurimisala peaks olema piisavalt suur ka objektide tiheduse ruumilise trendi ilmnemiseks (Radeloff *et al.* 2000). Uurimisalal on paratamatult siiski piirid ja seetõttu tuleb praktiliselt kõigi punktmustrite analüüsil uurimisala servade lähedal arvutatud või mõõdetud parameetreid kas korrigeerida või kriitiliselt hinnata. Näiteks keskmine naabrite arv etteantud kauguseni saab serva lähedal alahinnatud ja keskmine kaugus lähima naabrini ülehinnatud. Vaid uurimisala piiresse jääval pinnal arvutatud naabrite keskmine tihedus ei hälbi, kuid selle hinnang on servade lähedal vähem täpne. Samuti muutub elupaiga kvaliteet elupaiga serva lähedal. Elupaiga sobivus serva lähedal kas langeb sujuvalt või on siis elupaikade piirialad just liigile kõige sobivamad. Põhjalikumalt käsitleb ruumiliste punktprotsesside servakorrektuure Ripley (1982).

Serva mõju vältimise vahendid on toroidteisendus, puhvertsoonide kasutus ja arvutuslikud servakorrektuurid. Toroidkorrektuuri (*toroidal edge correction*) ehk **toroidteisenduse** puhul ühendatakse mõtteliselt uuritava ala vastasservad – tasapinnast moodustatakse toru (joonis 4-22)

(Diggle 1983, Haase 1995, Couteron ja Kokou 1997, Batista ja Maguire 1998, Mateu *et al.* 1998, Grau 2000). Toroidteisenduse kasutamine on hõlbus vaid siis, kui uurimisala on riskülikukujuline. Muudel juhtudel tuleb tulemusi korrigeerida nihutamise järel kokku langeva uuritud ala osaga.

Kui uuritud ala ja andmestik on piisavalt suur, võib servalähedases puhvertsoonis olevaid punkte kasutada vaid naabrite loendamisel ja loobuda neist lähtepunktidena (joonis 4-23). See on kõige lihtsam lahendus, aga tähendab osadest andmetest loobumist. Puhvertsooni on kasutanud näiteks Clarke ja Evans (1954) ning Campbell (1995).



Joonis 4-23. Punase rombiga tähistatud objektid on vaadeldud ala äärele lähemal kui 10 ühikut. Punktmustrit kirjeldava statistiku arvutamisel raadiuses 10 kaugusühikut tuleks neid kasutada neid naabritena ja mitte lähteobjektidena.

Arvutuslikest servakorrektuuridest on Ripley *K* funktsiooni servakorrektuuri kirjeldatud peatükis 4.1.2.17. Coomes *et al.* (1999) on seostanud servakorrektuuri suuruse punktide arvu ja lähima naabri kaugusega. Ripley *K* servakorrektuuri arvutamiseks on Ohser ja Stoyan (1981) pakkunud valemi

$$\hat{K}(r) = \left(\frac{A}{n}\right)^2 \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} \frac{I(d_{ij} < r)}{w(d_{ij})}, \qquad [4\text{-}65]$$

kus A on vaatlusala pindala, n on objektide arv, d_{ij} on vahemaa objektide i ja j vahel, i ja j on lähte- ja sihtobjekti indeks, $i = 1$ kui $d_{ij} < r$ ja $i = 0$ muudel juhtudel, I on vahemaa d_{ij} raadiuses r paiknemise indikaator, $w(d_{ij})$ on servakorrektuur. Ristkülikukujulise servapikkustega a ja b ($a < b$) ala puhul arvutatakse servakorrektuur järgmiselt:

$$w(d_{ij}) = ab - \frac{d_{ij}(2a + 2b - d_{ij})}{\pi}, \text{ kui } 0 < d_{ij} < a. \qquad [4\text{-}66]$$

Seda korrektuuri on kasutanud Ohser (1983), Penttinen *et al.* (1992) ning Batista ja Maguire (1998).

4.1.5. Punktobjektide ja pinna ebaühtlus

Elustiku ebaühtlase paiknemise põhjuseid on palju. Nende põhjuste mõistmiseks on tarvis kirjeldada nii paiknemise kui ka seda mõjutavate faktorite ebaühtlust. Kusjuures kohaga seotud tunnuste hulka kuuluvad ka selle koha omadused eelnevatel ajahetkedel.

Ruumilise mõju poolest võib ökoloogilisi faktoreid jagada kolmeks:

- tagasiside efektiga faktorid ehk uuritava nähtuse mingis kohas olemasolevast tunnusest (näiteks asustustihedus) sõltuvad faktorid,
- mingi koha omadused, mis uuritavast nähtusest ei sõltu,
- selle koha nii ajalise kui ruumilise ümbruse omadused ehk keskkonna varieeruvus.

Ebaühtlase punktmustri kirjeldamiseks esitasid Baddeley *et al.* (2000) Ripley *K* funktsiooni ja paarilise korrelatsiooni versiooni mittestatsionaarse punktmustri jaoks. Comas *et al.* (2009) kasutasid neid mudeleid mittehomogeense metsa struktuuri uurimisel ja lisasid mittehomogeense Ripley *K* funktsiooni ja paarilise korrelatsiooni hajuvuse ning mittehomogeensuse statistikud. Paiknemismustrite üldistatud kirjeldamiseks on kasutatud ka mittehomogeensete protsesside mudeleid, näiteks puistu struktuuri uurimisel (Stoyan ja Stoyan 1998) ning sademete modelleerimisel (Onof *et al*. 2000).

4.1.5.1. Pinna ja keskkonnategurite ebaühtlus

Lihtsamad punktmustrite kirjeldamise meetodid eeldavad mustrit kandva ruumi omaduste ühetaolisust või pinna ebaühtluse ja trendide kõrvaldamist enne analüüsi. Populatsiooniökoloogias ja eriti mitmeosaliste metapopulatsioonide modelleerimisel on elupaiku käsitletud selgepiiriliste ja ühetaoliste, maatriksit moodustavate laikudena (Hanski 1994, 1999). Füüsilisest ruumist olulisem limiteeriv faktor on enamasti kättesaadavate ressursside hulk. Saadaolevate ressursside hulga hindamise esmane lihtne vahend on indiviidi mõjusfääri suuruse või ümbruskonnas oleva elamiskõlbuliku ala suuruse määramine. Kui elupaiga sobivuse või ressursside hulga määramiseks on täpsemaid võimalusi, siis saab neid kasutada kas ruumiressursi kaaludena või ruumiressursi asemel.

Mitmed autorid on hoiatanud, et kui paiknemismustri analüüsi juures ei arvestata pinna ebaühtlust, võib jõuda väärate järeldusteni (Pielou 1960, Chapin *et al.* 1989, Wilson 1991, Duncan 1995, He ja Duncan 2000). Kaasaegses maastikuökoloogias rõhutatakse laikude omavahelist seotust – laigud on maastiku kompleksi komponendid (Wiens 1997), on pidevas muutumises (Akçakaya 2001, Walters 2001) ja sageli selgete piirideta. Seega, kaasajal tuleks arvestada, et mitte ainult populatsioonid ei ole dünaamilised, vaid ka elupaik kui populatsiooni substraat on pidevas muutumises. Paiknemise laigulisus või korrapära ja samuti liikide koosesinemine võib olla tingitud liikide samalaadsest sõltuvusest keskkonnatingimustest ebaühtlases keskkonnas.

Pinna ebaühtluse saab esitada pinna mõõdetavate omaduste väljana (kaardina) või omaduste mõju summeeriva sobivuspinnana. Viimane on pinna omadustest sõltuva tunnuse esinemise/puudumise, arvulise väärtuse või klassikuuluvuse hinnang. Elupaigasobivuse indeksite arvutamise meetoditest on peatükk 5.6.2. Nii elupaiga kvaliteedi ruumilise varieeruvuse kui ka elupaikade ebaselgete piiride korral soovitatakse modelleerimisel eelistada eraldiste korrapärast ruumilist võrgustikku (Thomas ja Kunin 1999, Luoto *et al*. 2001).

Punktmustri heterogeense pinnana saab käsitleda ka ebaühtlaselt paiknevat punktobjektide populatsiooni, millest teatud hulk punkte on teistest erinevalt märgistatud. Näiteks haiguse levimise pinnana on sisuliselt õigem kasutada potentsiaalsete haigestujate punktmustrit ja mitte maapinda. Punktmustri loomisel saab pinna ebaühtlust arvestada mustrit moodustava protsessi intensiivsuse reguleerimisega,

aga ka reguleerides punktide kadumist punktmustrist ehk **harvendamist** (*thinning*) (ptk 6.1.3). **Punktprotsessi intensiivsuse** reguleerimine tähendab punktobjekti tekkimise tõenäosuse sõltuvusse seadmist pinna omadustest (ptk 6.1.2.2). Näiteks nii, et soodsas kohas on uue taime tärkamise tõenäosus suurem. Pinna ebaühtlust võib ka kirjeldada uuritavate objektide optimaalse tiheduse väljana. Ka servaefekti probleemi võib käsitleda laiemas, pinna ebaühtluse kontekstis.

Esimestes pinna ebaühtluse punktmustri analüüsi kaasamise katsetes eeldati, et uurimisala sisaldab piirkondi, kus objektid ei saa asuda. Ökoloogilises kontekstis tähendab see, et uurimisala käsitleti kaheväärtuselisena – koosnevana elupaiga ja mitteelupaiga laikudest. Kaheväärtuselisel pinnal olevast juhuslikust agregeeritud punktmustrist võetud prooviruutudes loendatud objektide arvu jaotust nimetatakse **Shorti jaotuseks**. Ripley (1985) esitas koha ümbruses oleva sobiva elupaiga pindala silmamõõdulise määramise metoodika, milles kasutatakse läbipaistvat trafaretti.

Isendite tihedam paiknemine suurema mahtavusega alal ei tõenda liigiomast koondumistendentsi, vaid pigem elupaigaeelistust

Mingi koha sobivuskaal peaks sisaldama andmeid ka selle koha ümbruse kohta. Erinevad ei ole reeglina mitte ainult sama koha tingimused, vaid ka tingimused ümbruskonnas ja need mõjutavad koha väärtust elupaigana ja ressursside allikana. Näiteks metsa koosseisu uuenemine sõltub ümbruses oleva metsa koosseisust ja tihedusest (Frelich *et al.* 1993, Batista ja Maguire 1998). Ruumilist heterogeensust on lülitatud ka röövloom-saakloom mudelitesse (McLaughlin ja Rouchgarden 1992). Pinna ebaühtluse kirjeldamiseks võib koostada loendamatu hulga algoritme. Paraku tuleb silmas pidada, et objektide ruumilist paiknemist iseloomustavate näitajate väärtused sõltuvad pinna ebaühtluse arvestamise metoodikast. Erinevate meetodite abil saadud tulemused ei pruugi võrreldavad olla.

Ripley (1985) tõdeb linnupesade paiknemise analüüsi lõpus, et linnupesade paiknemise seaduspärasuste määramisest on vist võimatu lahutada oletusi elupaigamustri mõjudest nendele seaduspärasustele.

4.1.5.2. Objektide ebaühtlus

Enamasti lähtutakse punktmustri kirjeldamisel eeldusest, et kõik indiviidid on ühetaolised, kuigi selline eeldus on ebareaalne (Cormack 1979). Inimeste erinevust tunnetame hästi, aga isegi ritsikate puhul on näidatud, et osa isendeid looduslikus populatsioonis koondub gruppidesse ja teised hoiduvad omaette (Campbell ja Shipp 1979). Selleks, et indiviidide elujõudu ja konkurentsivõimet punktmustri kirjeldamisel arvestada, tuleb seda igal indiviidil kas kuidagi mõõta või omistada indiviidi omadused juhuslikult.

4.1.5.3. Varasema arengu ruumiline ebaühtlus

Ökosüsteemi varasema arengu olulisust kaasaegse objektide paiknemise mustri kujundajana on rõhutatud mitmes uurimuses. Näiteks Nicotra *et al.* (1999) on näidanud, et uute puude tärkamine troopilises vihmametsas võib sõltuda valguslaikude varasemast paiknemisest. Praeguse taimkatte laigulisust võivad põhjustada varasemate taimede jäänused (Bergelson 1990) ning seemnete ebaühtlane levik ja ebaühtlane seemnevaru mullas (Batista ja Maguire 1998, Cantero *et al*. 1999). Mineviku maastiku tõenäolist mõju pärdikute levikule on ära märkinud Lindenmayer *et al*. (1999). On näidatud, et metsastunud alade taimestik sõltub sellest, milline oli maakasutus enne metsa (Wulf 2004).

4.2. Joonte, suundade ja kaugussuhete kirjeldamine

4.2.1. Jooned

Joonmustri põhilised omadused on joonte tihedus ja suundade jaotus. Suundade jaotuse kirjeldamise meetodeid käsitletakse järgmises peatükis.

Joonte tiheduse ruumilist jaotust saab mõõta liikuva aknaga. Liikuva akna suurus ei pruugi seejuures olla sama, mis akna nihutamise samm. Saadakse tiheduste väli, mille analüüsimiseks saab kasutada väärtuspindade analüüsi meetodeid. Joonte tiheduse kirjeldamiseks sobivad ka mitmed punktmustrite kirjeldamise meetodid, sest jooned saab asendada joonel paiknevate punktidega, mille vahemaa on konstantne, juhuslik või mingi muu reegli järgi määratud.

4.2.2. Suunad

Selle alapeatüki kirjutas **Jaanus Remm**.

Suunaandmete puhul on asjakohane kasutada **tsirkulaarstatistika** (*circular statistics*) meetodeid, mis on kohandatavad kõigile andmetele, mis on pärit tsükliliselt sama väärtuseni jõudvatest ehk perioodiliselt korduvatest jaotustest. Tsirkulaarsele jaotustele on iseloomulik tähendusliku alguse ja lõpu puudumine ning väärtuste suurteks ja väikesteks jagamise tähendusetus. Suundade analüüsi läheb tarvis näiteks loomade liikumissuundade, tuulesuundade, puuõõnsuste avade paiknemise, kasvukoha ekspositsiooni ja pinnastruktuuri orienteerituse uurimisel.

Suuna-andmete analüüsi põhiline omapära on **nurkade tsirkulaarsus** ehk perioodilisus – null ja 360° on sama suund; 1° ja 359° vahe on 2° ning nende keskmine ei ole 180°. Seega lihtsa aritmeetilise keskmise arvutamine suunanurkadest ei oma tähendust. Sarnaselt on asjakohane läheneda ka perioodilise ajamõõtmise tulemustele nagu kellaaeg ööpäeva lõikes või kuupäev aastas. Näiteks kui uuritavad sündmused toimuvad südaöö paiku kella 23:30 ja 00:30 vahel, siis oleks ilmselgelt väär käsitelda nende toimumise keskmise kellaajana 12:00 päeval. Erandina, kui uuritavate väärtuste nurkvahemik on suhteliselt kitsas ja ei sisalda nullväärtust (sellega piirneda võib), siis saab tsirkulaarseid andmeid üldistatult käsitleda ka lineaarsetena. Suunanurka saab esitada kahe ristkoordinaadi abil – põhi-lõuna, ida-lääs või Δx ja Δy.

Suunda saab mõõta millegi suhtes, seega on suund alati suhteline ehk nagu laulis John Lennon: „*How can I go forward if I don't know which way I'm facing*?“ (Kuidas saan liikuda edasi, kui ma ei tea, mispidi ma seisan.) Suuna mõõtmise pidepunkt võib olla kõigi vaatluste jaoks ühtne koordinaatsüsteemi null-suund või iga vaatluskoha puhul eraldi rakendatav kohalik suund. Näiteks suund lähima naabri poole, suund lähima eluhooneni, suund kohaliku valgusallika suhtes, suund liikumissuuna suhtes, suund väärtusvälja lokaalse miinimumi või maksimumi suhtes.

Suund on alati suhteline

4.2.2.1. Keskmine suund

Selleks, et leida paljude suunaga objektide keskmine suund, tuleks vaadeldavaid objekte käsitleda vektoritena. Igat vektorit iseloomustavad suund ja pikkus. **Suund** (Θ) võib olla mõõdetud mingi üldise või lokaalse koordinaatsüsteemi suhtes, näiteks vaadeldava objekti nurk mõne teise objekti suhtes. Suunavektori pikkus (R) peaks olema võrdeline iga vaatluse kaaluga (w). Juhul, kui kõik vaatlused

omavad võrdset kaalu, võrduks kõigi vektorite pikkus ühega. Sellisel juhul ei pea keskmise suunavektori leidmisel kaalusid arvestama.

Keskmise suuna leidmiseks suundadest Θ_i ... Θ_n saab kasutada valemit (Upton ja Fingleton 1989)

$$\overline{\Theta} = \arctan \frac{\sum_{i=1}^{n} w_i \sin \Theta_i}{\sum_{i=1}^{n} w_i \cos \Theta_i} \qquad [4\text{-}67]$$

Suunavektorite keskmine pikkus, mis kajastab suundade hajuvust, on keskse tähtsusega hüpoteeside kontrollimisel tsirkulaarsete andmete alusel, leitakse vastavalt valemile

$$\overline{R} = \frac{\sqrt{\left(\sum_{i=1}^{n} w_i \sin \Theta_i\right)^2 \left(\sum_{i=1}^{n} w_i \cos \Theta_i\right)^2}}{n}. \qquad [4\text{-}68]$$

Nendes valemites tähistavad $w_i \sin\Theta_i$ ja $w_i \cos\Theta_i$ vektori lõpp-punkti x ja y koordinaate eeldusel, et vektori algpunkt on kasutatava koordinaadistiku 0-punkt. *Y*-telje suund võetakse koordinaatsüsteemi nullsuunaks ja x-telg on sellega 90°-se nurga all, suundasid arvestatakse päripäeva (nagu kompassi asimuut).

Ristkoordinaatide abil esitatud suunaandmete, nagu näiteks liikumise alg ja lõppunkti puhul saab keskmise suuna leidmiseks kasutada koordinaatide vahesid (Δx ja Δy). Vahed tuleb keskmistada või summeerida ning seejärel teisendada keskmistatud koordinaadid nurga ühikuks (kraad, radiaan, täispööre)

$$\overline{\Theta} = \arctan \frac{\sum_{i=1}^{n} w_i \Delta x_i}{\sum_{i=1}^{n} w_i \Delta y_i}. \qquad [4\text{-}69]$$

Tulemuste interpreteerimisel tuleb tähele panna, et trigonomeetrilised pöördfunktsioonid annavad tulemusi 180° ulatuses (−90°…+90°). Tulemuseks saadakse nurk 0-suuna suhtes, kuid määramata jääb keskmise suunavektori siht. Kumba täisringi poolde tulemuseks saadud vektori siht jääb, tuleb tuvastada algandmete ja koordinaatide keskväärtuste alusel (Jammalamadaka ja SenGupta 2001). Kui $\sum\cos\Theta_i$ või $\sum\Delta y_i$ on positiivne ja arvutuse tulemuseks saadud nurga väärtus on positiivne ($\Theta' \geq 0$), siis $\Theta = \Theta'$, kui $\Theta' < 0$, siis $\Theta = 360 + \Theta'$; kui $\sum\cos\Theta_i$ või $\sum\Delta y_i$ on negatiivne, siis $\Theta = 180 + \Theta'$.

4.2.2.2. Keskmise suuna usalduspiirid

Eeldusel, et suunad jaotuvad keskväärtuse ümber vastavalt normaaljaotusele ning kõigil vaatlustel on võrdne kaal, saab keskmisele suunale leida usalduspiirid valemiga

$$\overline{\Theta} \pm \arcsin\left[u_\alpha \sqrt{n \frac{1 - \frac{1}{n}\left(\sin 2\overline{\Theta} \sum_{i=1}^{n} \sin 2\Theta_i + \cos 2\overline{\Theta} \sum_{i=1}^{n} \cos 2\Theta_i\right)}{2\left(n\overline{R}\right)^2}} \right], \qquad [4\text{-}70]$$

kus u_α tähistab olulisusnivoole α vastavat erinevust keskväärtusest standardiseeritud normaaljaotuse korral (Upton ja Fingleton 1989). Juhul, kui $\alpha = 0{,}05$, siis $u_\alpha = 1{,}96$.

4.2.2.3. Keskmine suund võrreldes juhusliku jaotusega

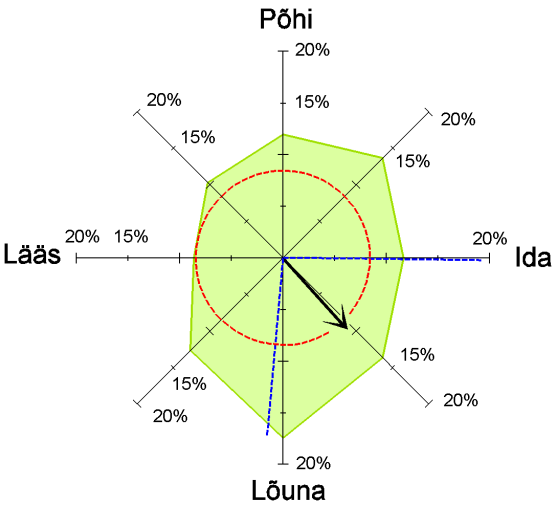
Suundade jaotuse, aga ka keskmise suuna erinevust juhuslikust jaotusest näitab vaadeldavate objektide suundadest leitud keskmise vektori pikkus. Seejuures eeldatakse, et juhusliku jaotuse korral on kõikide suundade esinemise tõenäosus täisringil võrdne. Seega läheneks lõputult suure juhusliku jaotuse korral suundadest arvutatud keskmise vektori pikkus nullile. Eeldusliku juhusliku suundade jaotusele 95% usalduspiiri leidmiseks saab kasutada Monte Carlo meetodit (ptk 4.1.4.2). Genereerides juhuslikke suundasid korduvalt ja leides keskmisi suunavektorite pikkuseid kontrollitava valimiga sama suurtes juhuslikes valimites, on võimalik hinnata, millised suunavektorite keskmised pikkused tekivad juhuslikult väiksema tõenäosusega kui uuringus kasutatav olulisusnivoo. Empiirilisele keskmisele suunavektori pikkusele on võimalik usalduspiirid leida bootstrap protseduuriga, moodustades juhuslikke valimeid uuritavate andmete hulgast (ptk 3.6.5). Suundade jaotust, keskmist suunda ja selle usalduspiire saab graafiliselt kujutada suundade roosina (joonis 4-24).

Rayleigh' test võimaldab kontrollida tsirkulaarse jaotuse ebaühtlust ehk erinevust juhuslikkusest koondades vaatlused ühte suunda. Eeldatakse, et andmed vastavad **von Mises'e jaotusele** ehk tsirkulaarsele normaaljaotusele. Juhul, kui vaatlused on koondunud sümmeetriliselt vastassuundadesse või võrdsete vahedega radiaalselt mitmesse suunda, oleks keskmise vektori pikkus hoolimata korrapärast nullilähedane nagu ka juhusliku paigutuse korral. Teststatistik (Z) leitakse valimi suuruse (n) ja suundade keskmise vektori pikkuse ($\overline{R}$) ruudu korrutisena (Zar 1996).

$$Z = n \cdot \overline{R}^2 \qquad [4\text{-}71]$$

Otsus nullhüpoteesi juurde jäämise või selle kõrvalejätmise kohta tehakse võrreldes empiirilist Z väärtust vastava valimi suuruse puhul eri uusaldusnivoode kriitiliste Z väärtustega (Zar 1996). Umbkaudse usaldusväärsuse hinnangu saab leida ka vastavalt valemile 4-72 (Berens 2009).

$$p = e^{\sqrt{1+4n+4[n^2-(n\cdot\overline{R})^2]}-2n-1} \qquad [4\text{-}72]$$



Joonis 4-24. Puuõõnte ava suundade diagramm Alam-Pedja lammihaavikutes (n = 418, avade keskmine asimuut 138° ± 47°, erinevus juhuslikkusest 9,3%, p = 0,029). Kaheksanurk näitab ava suundade protsentuaalset jagunemist 8 ilmakaare vahel, noole suund keskmist ava suunda ning pikkus erinevust juhuslikkusest, sinised katkendjooned keskmise suuna 95%-usaldusvahemiku ning punane katkendjoonega ring näitab 1000 Monte Carlo iteratsiooni ava suundade juhusliku jaotumise 95%-usalduspiiri (8,4%) (Remm *et al*. 2006 muudetud).

Uuritava valimi keskmise suuna erinevuse olulisust nullhüpoteesile vastavast suunast (Φ) ehk suunast, mille ümber vaatlused juhuslikkuse korral koodnuksid, võimaldab kontrollida V test.

Teststatistik (*V*) arvutatakse järgmiselt:

$$V = \sum_{i=1}^{n}(\Delta y_i \cdot \cos\Phi) + \sum_{i=1}^{n}(\Delta x_i \cdot \sin\Phi) = n \cdot \overline{R} \cdot \cos(\overline{\Theta} - \Phi). \quad [4\text{-}73]$$

Seejärel soovitatakse *V* statistikut vaatluste arvu järgi korrigeerida *u* statistikuks (Zar 1996, Berens 2009).

$$u = V \cdot \sqrt{\frac{2}{n}}. \quad [4\text{-}74]$$

Kasutuse näidis

Korduva kasutamise hõlbustamiseks on valemid 4-67 kuni 4-70 salvestatud Exceli faili *suunad.xls*, mis on zip arhiivina saadaval selle raamatu veebilehelt. Selle faili töölehel *Näidis* veerus *A* on 128 Pedja jõe äärsetest haavikutest leitud rähniõõnsuse ava asimuudid ning veerus *B* on iga õõnsuse ava pindala, mida on kasutatud vaatluse kaaluna, mille bioloogiline tähendus võib olla näiteks õhus lendlevate seeneeoste puutüve sisemuse sattumise sagedus. Kasutaja saab veergudes *A* ja *B* olevad näidisandmed asendada oma vaatlusandmetega. Lahtrisse *K2* arvutab Excel õõnsuste avade keskmise suuna arvestamata vaatluse kaalu ning lahtrisse *J2* kaalutud keskmise suuna. Lahtris *N2* on keskmise kaalumata vektori pikkus. Keskmise vektori pikkus võib varieeruda vahemikus 0…1 (0…100%). Lahtris *M2* on keskmise kaalumata vektori pikkus, mille suurus on samades ühikutes kui vaatluste kaaludki. Vaatluste keskmist vektori pikkust saab võrrelda juhuandmetest genereeritud keskmise vektori pikkustega ja niimoodi selgitada, kuivõrd oluline on keskmise suuna erinevus juhuslikust suunast.

Näidisfaili töölehel *Näidis* lahtrites *K2* ja *K3* on keskmise kaalumata suuna 95%-usalduspiirid. Näidisfaili töölehel *Monte Carlo* on vaatlustega sama arv juhuslike suundasid genereeritud 42 korda (veergudes *C*, *I*, *O* jne) ning lahtrites *B2* kuni *B43* on igale juhuslike suundade komplektile arvutatud keskmise vektori pikkus. Töölehel *Näidis* lahtris *P2* on vektori pikkus, millest 95% eelnimetatud keskmistest vektoritest on lühemad. Samuti on leitud 95%-usalduspiir kaalutud vaatluste juhuslikule jaotusele (töölehel lahtris *O2*). Kui eesmärgiks on kontrollida keskmise suuna, aga mitte kaalude jaotuse erinevust juhuslikust, siis tuleks kaalud (töölehel *Monte Carlo* veergudes *D*, *J*, *P* jne) jätta sellisteks, nagu need loetakse töölehelt *Näidis*. Juhuslikuna genereeritakse vaid suunad ja vaatluste kaale ei muudeta. Kaalumata vaatluste puhul tuleks võrdsustada kaalude töölehel *Näidis* veerus *B* olevad väärtused.

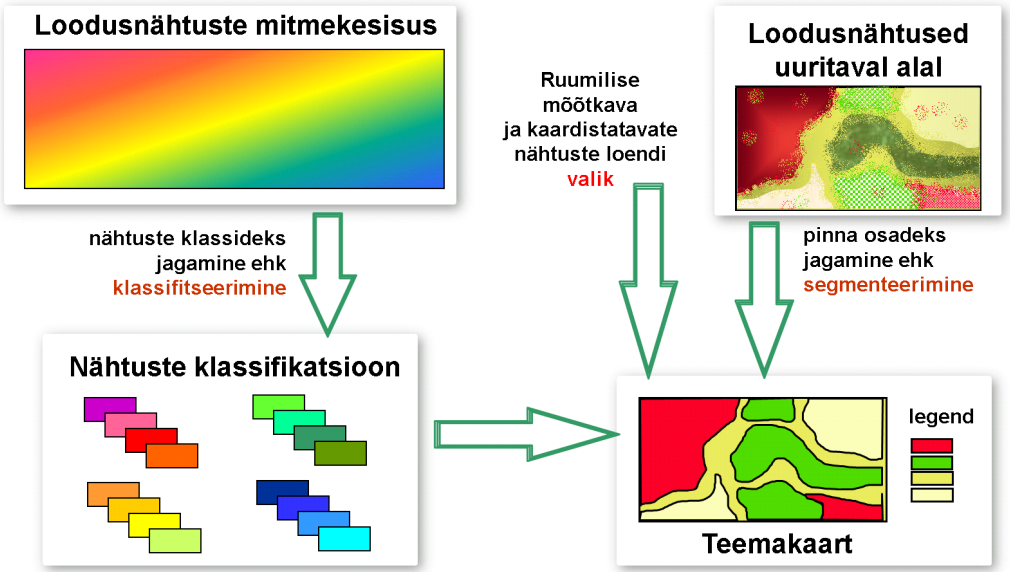
Programmeerimiskeele R keskkkonnas on suunaandmete analüüsimise vahendid koondatud pakettidesse *circular* ja *CircStats* (Jammalamadaka ja SenGupta 2001).

4.3. Väärtuspinna kirjeldamine

Reaalsed pinnad on ruumiliselt heterogeensed, neil on struktuur. Pindade heterogeensust iseloomustab väärtuste varieeruvus ja väärtuste paiknemine. Väärtuste paiknemist iseloomustatakse ruumilise trendi, ruumilise autokorrelatsiooni tugevuse ja ulatuse, anisotroopia ja juhusliku varieeruvuse (müra) tugevuse abil.

4.3.1. Kategooriline pind

Mingi ala mingis mõttes suhteliselt homogeenseteks eraldisteks jagamine ehk pinna segmenteerimine (ptk 4.3.2.2) ja eraldiste kaardina esitamine on looduse kirjeldamise üks levinumaid viise. Eraldisi kasutatakse näiteks elupaikade kaardistamisel, taimestiku kaardistamisel, maakatte kaardistamisel, mulla kaardistamisel ja metsakorralduses. Kõigil neil juhtudel tegeletakse klassifitseerimisega: looduse suhteline pidevus jagatakse nii temaatiliselt kui ka ruumiliselt kindlapiirilistesse klassidesse (joonis 4-25).



Joonis 4-25. Klassikaline teemakaart kui väärtuste klassifitseerimise ja pinna segmenteerimise tulemus.

Kui mingi ala on kaetud selgepiiriliste ja klassifitseeritavate, see tähendab nominaalse ehk kategoorilise muutuja areaalidega, siis nimetatakse seda **kategooriliseks pinnaks.** Kui kategooriline pind on tervenisti kategoorilise muutujaga kaetud, nimetatakse seda **kategooriliseks katteks** (*categorical coverage*). Kategooriline pind on reeglina **laiguline** (*patchy*). Kategoorilise pinna ruumilist heterogeensust võib jagada viieks komponendiks:

- kategooriate arv,
- iga kategooria osakaal,
- laikude paiknemine,
- laikude kuju,
- laikude naabrussuhted (Li ja Reynolds, 1994).

B. Boots (2006) lisab loetelusse klasside paiknemise vaadeldava ala keskkoha suhtes ehk ekstsentrilisuse.

Pindobjektide paiknemisstruktuuri uurimismetoodikaga tegeles juba Greig-Smith (1952), kes soovitas punktmustrite ja laigulisuse uurimiseks kasutada külgnevaid eri suurusega analüüsiruute, mida on kasutatud ka punktmustrite kirjeldamiseks (ptk 4.1.2.7). Laigulisuse kirjeldamiseks saab kasutada teisigi punktmustrite kirjeldamise meetodeid, kui pind asendada korrapärase võrgustikuga ja võrgustiku iga silma käsitleda punktina. Laigulisust saab iseloomustada ka ruumilise autokorrelatsiooni kirjeldamise vahenditega (ühisloend-statistik, korrelogramm – ptk 5.1).

Pinna erijuhtum on läbilõige ehk **transekt**, kui seda kasutatakse kindla laiusega väljavõttena (reeglina pikk ja kitsas riba) suuremast pinnast. Enamasti käsitletakse transekti laiuseta läbilõikena.

4.3.1.1. Maastikumeetrika

Maastikumeetrika on kvantitatiivse maastikuanalüüsi valdkond, mis kirjeldab kvantitatiivseid seaduspärasusi kategoorilise kattena kaardistatud maastikus. Enamasti seostatakse maastikumeetrikat maakatte (*land cover*) kirjeldamisega seostamaks indeksite erinevusi maastiku mingite omadustega ja maastikul toimuvate protsessidega. Paljusid kategoorilise katte indekseid saab arvutada nii iga üksiku kategooria kohta kui ka uuritava maastiku kohta tervikuna, mõningaid indekseid (näiteks kuju indekseid) saab arvutada ka iga laigu puhul eraldi. Maastikumeetrikat võib siiski mõista ka laiemalt, mitte ainult maakatte ruumilist struktuuri kirjeldavana, vaid ka pidevate maastikuparameetrite paiknemisstruktuuri iseloomustajana. Pidevateks maastikuparameetriteks on näiteks maapinna kõrgus, nõlvakalle ja ekspositsiooninurk. Näiteks maakatteklasside mitmekesisuse asemel võib maastikumeetrika indeksite abil kirjeldada ka topodiversiteeti või biodiversiteeti.

Maastikumeetrika indekseid on välja mõeldud sadu. Kuna paljud neist on lähedased ja omavahel tugevasti korreleerunud, siis on leitud, et kuus kuni üheksa indeksit esindavad maastikumustrist valdavat osa (Riitters *et al.* 1995, Botequilha Leitão ja Ahern 2002, Cushman *et al.* 2008). Ülevaate maastikumeetrika rakendustest on kirjutanud Botequilha Leitão ja Ahern (2002) ning Uuemaa *et al.* (2009). Meeles tasub pidada, et ökoloogilist tähendust omav maastikumuster sõltub iga organismi tegutsemisala ulatusest, levimisvõimest, vajalikest ressurssidest ja ressursside kasutuse intensiivsusest, mitte inimeste loodud kartograafilisest kujutisest (Milne 1992).

Maastikumeetrika indeksite arvutamise standardpakett on FRAGSTATS (http://www.umass.edu/landeco/research/fragstats/fragstats.html). Idrisi vormingusse rasteriseeritud pindade lokaalseid tunnuseid saab arvutada ka tarkvaras LSTATS (www.geo.ut.ee/LSTATS) ja Constud (ptk 3.4.6.1 ja 5.6.8.3).

Järgnevalt on esitatud mõned näited maastikumeetrika indeksitest. Täiendust võib leida kirjandusest (näiteks O'Neill *et al*. 1988, Miller *et al*. 1997, Boots 2006, Uuemaa *et al.* 2009) ja Fragstats programmi veebilehelt (http://www.umass.edu/landeco/research/fragstats/documents/Metrics/Metrics%20TOC.htm). Alljärgnevast loetelust on välja jäetud üldised mitmekesisuse ja ühetaolisuse mõõtmise vahendid, millest on juttu peatükis 2.1 ja lihtsad koonstatistikud (keskmine, tihedus ja hajuvuse indeksid), mis kuuluvad üldstatistika osasse (ptk 1).

Agregatsiooni indeksid näitavad sama kategooria pikslite koondumist. Lihtsaim agregatsiooniindeks mõõdab sama klassi külgnevuste osa kõigi pikslikülgnevuste suhtes (Bregt ja Wopereis 1990):

$$AG = \frac{\sum_i \sum_j ab_{ij}}{\sum_i \sum_j a_{ij}}, \qquad [4\text{-}75]$$

kus a_{ij} = 1 kui pikslid *i* ja *j* on naabrid ja muul juhul a_{ij} = 0, ab_{ij} = 1 kui pikslid *i* ja *j* on naabrid ja kuuluvad samasse kategooriasse, muul juhul ab_{ij} = 0.

He *et al*. (2000) agregatsiooni indeks *AI* näitab sama kategooriaga külgnemise osa maksimaalse võimaliku suhtes, kusjuures maksimaalne võimalik külgnevuste arv sõltub kategooria pindalast ja uuritava ala kujust

$$AI = 100\sum_{i}^{n} AI_i \cdot A_i, \qquad [4\text{-}76]$$

kus A_i on pindkategooria *i* osa uuritava ala pindalas, AI_i on pindkategooria *i* sama kategooriaga külgnemiste osa maksimaalsest võimalikust, *n* on pindkategooriate arv.

Kompaktsusindeks sõltub samasse kategooriasse kuuluvate piksliteks jagatud laikude suurusest ja arvust. Bregt ja Wopereis (1990) lugesid minimaalseks kompaktseks regiooniks kolmest külgnevast sama kategooria pikslist koosneva ala. Kui jagada kompaktsusindeks uuritava ala maksimaalse kompaktsusega, saab suhtelise kompaktsuse mõõdu, mis muutub vahemikus 0 … 1.

$$TCP = \sum_{i=3}^{n} W_i E_i, \qquad [4\text{-}77]$$

kus *n* on kompaktse regiooni maksimaalne suurus pikslites, minimaalne suurus on 3, W_i on pikslite arv kompaktses regioonis *i* ehk areaali *i* suurus, E_i on suurusega *i* kompaktsete regioonide arv.

Servakontrasti indeks on kolme omavahel külgneva ja erinevatesse kategooriatesse kuuluvate pikslite arv. Kui jagada servakontrasti indeks uuritava ala maksimaalse servakontrastiga, saab suhtelise servakontrasti mõõdu, mis muutub vahemikus 0 … 1.

Suhteline laigulisus (*relative patchiness – RPI*) arvestab külgnevate laikude erinevuse määra (Romme 1982).

$$RPI = 100\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{E_{ij}D_{ij}}{N_b}, \qquad [4\text{-}78]$$

kus *n* on kategooriate arv, E_{ij} on servade arv kategooria *i* ja *j* vahel, D_{ij} on kategooriate *i* ja *j* erinevuse mõõt, N_b on piksli või eraldise servade arv (ristkülikul 4).

Fragmentatsiooni indeks mõõdab kategooriate arvu (*M*) ja pikslite arvu (*N*) suhet.

$$TI = \frac{M-1}{N-1} \qquad [4\text{-}79]$$

Koonduvus (*contagion*) näitab, millisel määral on laigud agregeerunud. Suhteline koonduvus *RC* arvutatakse maksimaalse võimaliku koonduvuse $C = 2n\cdot\ln(n)$ suhtes.

$$C = 2n \cdot \ln n + \sum_{i}^{n}\sum_{j}^{n} P_{ij} \cdot \ln P_{ij} \qquad [4\text{-}80]$$

$$RC = \sum_{i}^{n}\sum_{j}^{n}\frac{P_{ij} \cdot \ln P_{ij}}{2n \cdot \ln n}, \qquad [4\text{-}81]$$

kus *n* on kategooriate arv, P_{ij} on tõenäosus, et külgnevad pikslid kuuluvad kategooriatesse *i* ja *j*.

Sidusus (*cohesion*) (Schumaker 1996) mõõdab kaheväärtuselise maastiku läbitavust eeldusel, et liikuda saab vaid ühte tüüpi piksleid pidi. Sidususe väljendamisel protsentides tuleb valemisse lisada sajaga korrutamine.

$$COH = \frac{1-\frac{\sum_{j}^{m} p_{ij}}{\sum_{j}^{m} p_{ij}\sqrt{a_{ij}}}}{1-\frac{1}{\sqrt{A}}} = \left[1-\frac{\sum_{j}^{m} p_{ij}}{\sum_{j}^{m} p_{ij}\sqrt{a_{ij}}}\right]\left[1-\frac{1}{\sqrt{A}}\right]^{-1}, \quad [4\text{-}82]$$

kus p_{ij} on laigu ij ümbermõõt pikslikülgede arvuna, a_{ij} on laigu ij pindala pikslite arvuna, A on pikslite arv uuritavas maastikus.

Suuruse ebavõrdsuse indeks (*size disparity*) on Lorenzi kõvera (ptk 2.1.4) ja graafiku diagonaali aluse pindala suhe. Kui kõik eraldised on võrdse suurusega, on Lorenzi kõver diagonaalsuunaline sirge.

Gini indeksit üldjuhul on kirjeldatud peatükis 2.1.4. Les ja Maher (1998) andsid Gini indeksile binaarse jaotuse jaoks järgmise kuju

$$\hat{\Omega}_t = \frac{N}{N-1}\sum_{j\neq 1} p(j\,|\,t)p(i\,|\,t), \quad [4\text{-}83]$$

kus N on klasside arv, $p(i|t)$ on pinnal t oleva juhusliku objekti tõenäosus kuuluda klassi i, $p(j|t)$ on pinnal t oleva juhusliku objekti tõenäosus kuuluda klassi j. Seega, mitmekesisus on suurim juhul, kui kaardil on kategooriaid i ja j võrdsel hulgal.

Fraktaalne dimensioon. Maastiku fraktaalset dimensiooni saab määrata üksiklaikude pindalade ja ümbermõõtude vahelise regressioonisirge tõusunurga järgi (O'Neill *et al*. 1988, Li ja Reynolds 1994). Laigu kuju fraktaalne dimensiooni D leitakse sel juhul valemist

$$D = \frac{2\ln(cP_k)}{\ln(A_k)}, \quad [4\text{-}84]$$

kus D on laigu kuju fraktaalne dimensioon, A_k on laigu k pindala, P_k on laigu k ümbermõõt, c on arvutusviisist sõltuv konstant, ruudukujuliste pikslitega rastri puhul = 0,25.

Teine maastikumustri fraktaalse dimensiooni arvutamise variant lähtub Pielou (1977) hierarhilisest mitmekesisusest, mis omakorda arvutatakse Shannoni mitmekesisusest $H(\varepsilon)$ (ptk 2.1.2), sõltuvalt mõõtkavaühikust ε (Li 2000). $H(0)$ tähistab mitmekesisust maksimaalselt detailses mõõtkavas.

$$D_I = \frac{H(\varepsilon)-H(0)}{\ln(\varepsilon)} \quad [4\text{-}85]$$

TTLQV indeksi (*two-term local quadrate variance – TTLQV* ehk V_b) väärtusi kujutatakse graafikul sõltuvalt analüüsiruudu suurusest b (b = 1, 2, 3, …, n/2) (Hill 1973) (vt ka ptk 4.1.2.7).

$$V_b = \frac{\sum_{j=0}^{n-2b-1}\left[\sum_{i=j+1}^{j+b}(x_i - x_{i+b})\right]^2}{2b(n+1-2b)}, \qquad [4\text{-}86]$$

kus x_i on katvus transekti *i*-ndas ruudus, *j* on ruudu indeks blokis suurusega *b*, *n* on ruutude arv transektis.

Dale ja MacIsaac (1989) näitasid, et TTLQV-indeks sõltub laikude tihedusest. Suuremate laikude suurema tiheduse korral ei suuda meetod väiksemat struktuuri avastada. Mainitud autorid pakkusid ka iteratiivse võimaluse V_b ja tiheduse hindamiseks eri mõõtkavas laikude puhul. Seejärel on võimalik arvutada osadispersioonid ja seega tiheduse erinevuse mõju kõrvaldada.

4.3.1.2. Kategoorilise pinna mitmekesisuse sõltuvus mõõtkavast

Ruumiandmeid saab järk-järgult üldistada nii, et moodustuvad sama maastiku erinevad mõõtkavatasandid – iga tasand omaette andmekihina. Moellering ja Tobler (1972) järgi võiks mitmekesisuse mõõtkavatasandeid mudeliks kirjutada kujul

$$X_{ijk} = \mu + \alpha_i + \beta_{ij} + \gamma_{ijk}, \qquad [4\text{-}87]$$

kus X_{ijk} on mitmekesisus maakonnas *k*, osariigis *j*, mis on piirkonnas *i*. μ on piirkonna omamitmekesisus, α piirkondlik mitmekesisuse osa, β väljendab osariigi panust mitmekesisusse ja γ maakonna panust. Iga maakonna mitmekesisus moodustub nende faktorite summana. Moellering ja Tobleri printsiipi saab kohandada ka tesitsugustele hierarhilistele tasemetele.

Mitmekesisuse mõõdikud reageerivad mõõtkava sama maastiku puhul muutmisele erinevalt. Näiteks Wu *et al*. (2000) leidsid, et laikude arv, laikude tihedus, servajoone pikkus, laigu keskmine suurus, laikude suuruste variatsioonikoefitsient muutuvad mõõtkava muutmisel sujuvalt, samas teised mõõdikud, nagu mitmekesisuse indeks, fraktaalne dimensioon ja kuju indeks võivad teha ette-arvamatuid hüppeid.

4.3.1.3. Elupaikade fragmenteerumine

Tervikliku elupaigalaigu jagunemist mitmeks väiksemaks laiguks nimetatakse elupaiga killustumiseks ehk **fragmenteerumiseks**. Elupaikade fragmenteerumisega kaasneb ühenduse märkimisväärne vähenemine või isegi katkemine elupaigalaike asustavate populatsioonide vahel. Üldiselt kaasneb elupaiga fragmenteerumisega kogu asurkonna elujõulisuse langus, kuid mitmest lokaalpopulatsioonist koosneval metapopulatsioonil võivad olla ühe tervikliku populatsiooni ees mõned eelised, näiteks riskide hajutatus. Ühe suure või mitme väikese populatsiooni eelistamine on looduskaitse planeerijate hulgas juba aastakümneid kestnud debatt (Diamond 1975, Simberloff ja Abele 1976, Araújo ja Williams 2000).

Elupaikade üksteisest eraldumise ehk fragmenteerumise mõju hindamisel on rõhutatud eelkõige liigi mobiilsuse, elupaiganõudlikkuse (spetsialiseerumise) ja populatsiooni tiheduse arvestamise olulisust. Fragmentatsiooni mõju ei sõltu ainult eraldiasuvate eraldiste hulgast ja suurusest, vaid ka nende omavahelisest paiknemisest, kujust ja suuruste jaotusest (joonis 4-26). Ohustatud liigi säilumiseks on otsustava tähtsusega enamasti just piisavalt suurte tuumalade (*core area*) ehk võtmelaikude (*key patch*) olemasolu.

Joonis 4-26. Intensiivselt raiutud metsaala Kirde-Eestis, kus lendorava esinemistõenäosus on elupaiga fragmenteerumise tagajärjel muutunud nullilähedaseks. Sobiva puistu hulk kujutatud alal on lendorava pesitsemiseks piisav, kuid raiutud ala eraldab kaitsealuse (punase piirjoonega) või vana metsa laigud üksteisest.

Uurimused

Mac Nally *et al*. (2000) juhivad tähelepanu fragmenteerumise erinevale mõjule eri liikide puhul; samuti liikidevaheliste suhete muutumisest tulenevatele raskustele fragmenteerumise mõju hindamisel.

Remmel ja Csillag (2003) juhivad tähelepanu, et maastikumeetrika indeksite väärtuste võrdlemisse tuleks suhtuda suure ettevaatusega, sest väikesed muutused mustri koosseisus ja konfiguratsioonis võivad esile kutsuda drastilisi muutusi indeksite väärtustes

Kumar *et al*. (2006) leidsid, et maastiku mitmekesisuse indeksid seletasid võõrliikide hulga varieeruvust suuremal määral kui kohalike liikide mitmekesisust.

Gaucherel (2007) soovitas mitmekesisuse ja mõõtkava seose näitamiseks kasutada heterogeensusprofiili (*heterogeneity profile – HP*) ja erinevates mõõtkavades mitmekesisuskaartide keskmistamist ühtseks mitmekesisuskaardiks (*multiscale heterogeneity map – MHM*).

Bennie *et al*. (2011) mõõtsid mitmekesisuse ruumilist varieeruvust teatud vahemaaga vaatluspaarides arvutatud mitmekesisuste erinevusena.

4.3.2. Pidev väärtuspind

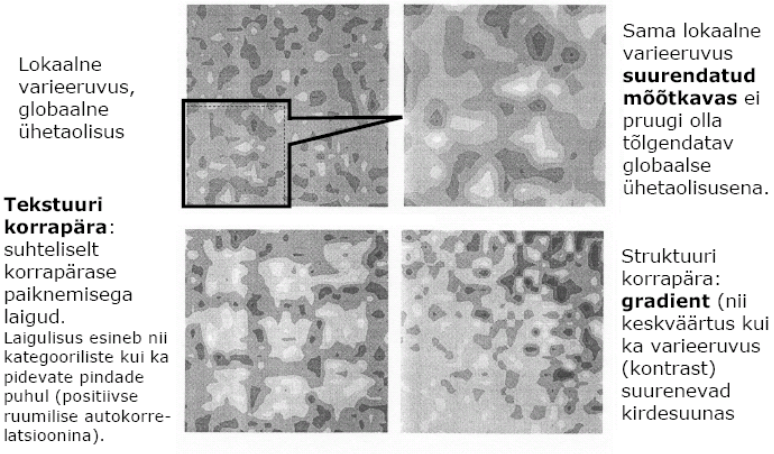
Väärtuspinnaks ökoloogilistes uuringutes on enamasti maastik, kaart (pinna tesselatsioon) või vaatlustransekt, mille punktides (sageli mitte kõigis) on mõõdetud või hinnatud mingi muutuja väärtus. Kujutise töötlemise meetodid eeldavad reeglina, et algandmeteks on üks või mitu rastrikihti, millede pikslite väärtusi (mis enamasti on täisarvud vahemikus 0…255) käsitletakse pideva muutujana. Satelliidipildi töötluses on klassifitseeritavateks väärtuspindadeks kiirgusväärtuste andmekihid. Satelliidipilt haarab suuremat ala ja on vähema geomeetrilise moonutusega, aeropildistamine võimaldab objekte lähemalt seirata ja on vähem tundlik pilvisuse suhtes.

Lennuaparaadile paigutatud sensor registreerib sensori tüübist sõltuvaid kiirgusvahemikke. Kohaga seotud signaali järgi saab tuvastada pikslile vastava koha omadusi. Kaugseireandmete kasutamisel

koha omaduste indikatsioonis on tüüpülesandeks pikslite klassifitseerimine kas varem etteantud klassidesse või siis parima klassifikatsiooni otsimine.

Traditsiooniliselt on taimkatte kaardistamiseks kasutatud Landsat TM sensori andmeid, mille piksli suurus on 25…30 m. Sellise piksli suuruse juures on küllaltki palju liitpiksleid, mis pärinevad osaliselt ühest ja osaliselt teisest maakatteklassist. Uuemate satelliitide ja aeropildistamisel kasutatavad sensorid on alla ühe meetrise lahutusvõimega. Nii detailse info töötlemisel on probleemiks maastiku üksikelementide erinevast peegeldumisest tulenev lisateave, mis ei ole suuremate üksuste eristamisel vajalik. Detailsete kujutiste järgi vähemdetailsete klasside eristamiseks ei sobi pikslikaupa klassifitseerimise meetodid. Nende asemel on asutud välja töötama koha ümbruses (kernelis) olevate teatud tüüpi pikslite sagedust või paiknemise struktuuri arvestavaid meetodeid. Kõrge lahutusega kosmosefotode töötlemine on aktuaalne, kuna selle kasutamine kaardistustöödel pakub kaugseireandmete suuremat ühetaolisust, võib olla odavam ja operatiivsem kui aerofotode kasutamine.

Pideva muutuja väärtuspinna varieeruvus võib avalduda juhuslike hälvetena, kindlasuunalise muutumistendentsina ehk trendina, lainelisusena ja erinevas suuruses laigulisusena või keerukama tekstuurina (ptk 4.3.2.6) (joonis 4-27).

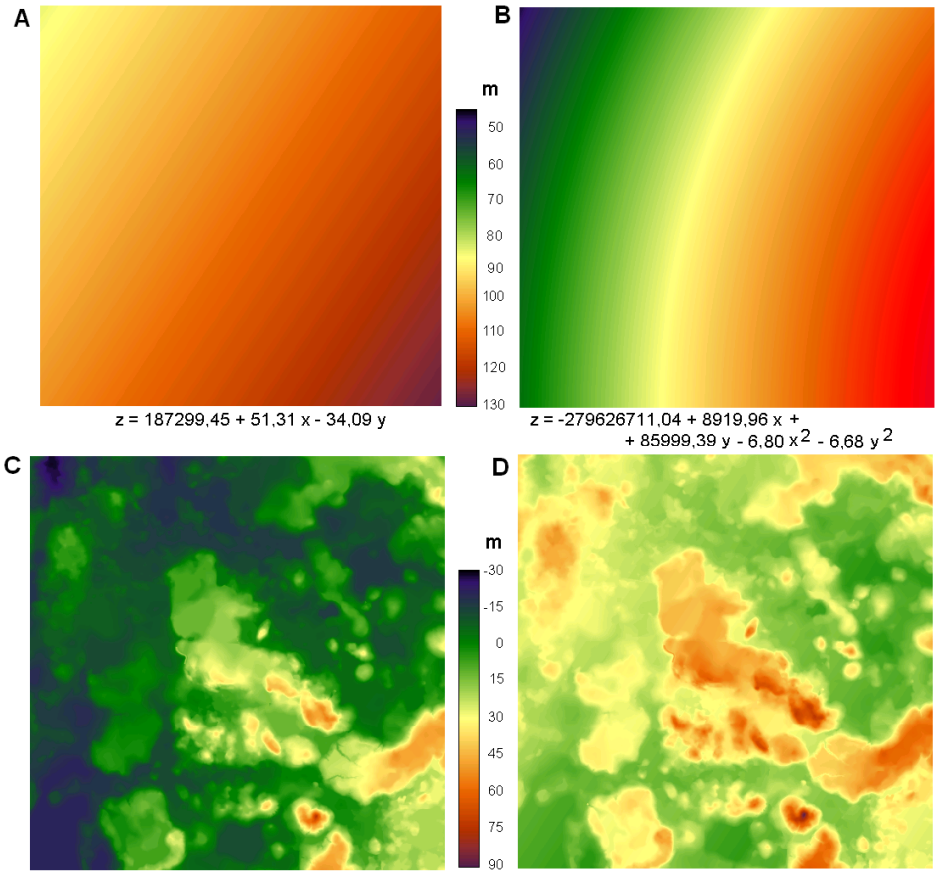


Joonis 4-27. Ruumilised struktuurid pidevates väärtuspindades; halli toonid esindavad pideva muutuja väärtusklasse (Legendre 1993, muudetud).

4.3.2.1. Ruumiline trend

Trend tähistab suundumust ehk muutumise üldistust. **Trendpinna analüüs** on üldjuhul suuremate ruumiliste tendentside otsimine andmetest, mis reeglina toimub regressioonanalüüsi meetoditega. Argumenttunnuseks võivad seejuures olla nii ruumikoordinaadid kui ka nende polünoomfunktsioonid või muud teisendused, funktsioontunnuseks on aga mingi muu uuritav muutuja (joonis 4-28). Ühemõõtmelise (*univariate*) trendpinna analüüsi puhul rakendatakse mitmetunnuselist regressiooni iga uuritava muutuja puhul eraldi. Trendpinna kanoonilise analüüsi puhul analüüsitakse korraga mitut sõltuvat muutujat ehk muutujate kompleksi (Bocquet-Appel ja Sokal 1989).

Trendpinna analüüsil kasutatakse argumenttunnustena vaatluskohtade koordinaate ja nende teisendusi



Joonis 4-28. Maapinna kõrguse (*z*) lineaarne trend kaardilehel 54344 (Otepää) (A), polünoomtrend samadest andmetest (B) ning maapinna suhteline kõrgus samal kaardilehel olevast keskmisest kõrgusest (C) ja lineaarse trendi suhtes (D). Trendpinna valemites on Lambert-Est süsteemi ristkoordinaadid kilomeetrites.

Trendpinna analüüsi eesmärgiks on ruumilise nähtuse asukohast sõltuva suundumuse üldistatud esitamine, lokaalsete nähtuste eristamine üldisest suundumusest ja lokaalsete nähtuste kirjeldamine trendi suhtes. Kui pinnavormide suhteline kõrgus esitatakse mingi kindla kõrgusega tasandi suhtes, siis nõlval olevaid pinnavorme saab kujutada ka nõlva üldistatud kujul kirjeldava kaldpinna suhtes (joonis 4-28D).

Ruumilise trendi lokaalne vaste on lokaalne muutus ehk lokaalne gradient. Lokaalseks gradiendiks on näiteks nõlva kaldenurk, mille suuruse arvutamiseks on tarvis teada kõrvalolevaid väärtusi kas naaberpikslites või lokaalses kernelis. Enamik geoinformaatika tarkvara arvutab nõlva kaldenurka võrreldes külgnevate pikslite väärtusi. Tartu Ülikooli geoinformaatika õppetooli juures loodud tarkvara LSATS (www.geo.ut.ee/LSTATS) oskab keskmise lineaarse gradiendi tugevust leida suvalise suurusega ruudukujulises või ümaras aknas.

Trendpinna kanooniline analüüs (*canonical trend surface analysis – CTS*) otsib mitmest üksiktunnusest koosneva kompleksi üldise muutumistendentsi seost asukoha koordinaatidega. Trendpinna kanooniline analüüs on lähedane kanoonilisele korrelatsioonile (ptk 2.4.5), mis mõõdab kanoonilist seost tunnustekomplektide vahel. CTS analüüsis konstrueeritakse mitmetunnuseline trendpind, mis korreleerub võimalikult tugevasti asukohakoordinaatidega ning teisest küljest ka koordinaatidega määratud kohtades mõõdetud tunnuste kompleksiga.

Kanoonilised trendid näitavad vaid üldtendentse ja on andmetele paremini sobitatud uurimisala keskosas. Originaalandmete hälbeid CTS pinna suhtes saab analüüsida nagu korrelatsioonijääke.

Sealhulgas võib huvi pakkuda korrelatsioonijääkide ruumilise autokorrelatsiooni analüüs.

CTS pinnaga seletatud kovariatsiooni suurust uuritavate muutujate väärtuste ja nende asukoha vahel saab hinnata kanoonilise korrelatsioonikordaja ruudu abil. Taanduvuse koefitsient (*redundancy coefficient of CTS*) näitab CTS pinna vastavust algsetele empiirilistele andmetele (Wartenberg 1985). CTS pind on analoogiline peakomponentpinnale, kui muutujate varieerumise peamine suund geograafilisel pinnal vastab varieeruvuse peakomponendile, mille arvutamisel ei ole asukohta arvestatud (Diniz-Filho ja Malaspina 1995).

4.3.2.2. Väärtuspinna segmenteerimine

Segmenteerimise eesmärk on jagada väärtuspind eraldisteks, mis vastavad reaalse maailma objektidele. Kõige primitiivsem väärtuspinna segmenteerimise viis on jagada kujutis võimalikult ühetaolisteks osadeks. Osadeks jagamise ülesannet saab lahendada kahel viisil: alt üles – külgnevate sarnaste pinna osade (pikslite või eraldiste) ühendamise teel, või ülalt alla – pinna osadeks jagamise teel. Pinna osadeks jagamine võib lähtuda kas kogu andmestikust või vaid lokaalsest varieeruvusest. Mitmetunnuselise pinna või transekti osadeks jagamisel on tegemist klasteranalüüsi tüüpi ülesandega, mille juures saab kasutada vaatluste vahelise sarnasuse/erinevuse mõõte, sealhulgas kaugust tunnusruumis. Erinevate muutumispiirkondadega tunnuste kasutamisel on soovitav erinevused nende standardhälvetega standardiseerida.

Haralick ja Shapiro (1985) järgi peaks segmenteeritud pind vastama järgmistele tingimustele:

- segmendid peaks olema suhteliselt ühetaolised vaadeldavate tunnuste osas,
- külgnevad eraldised peaksid erinema oluliselt,
- eraldised peaksid olema võimalikult lihtsa kujuga ja eelistatult ilma saarteta,
- eraldiste piirjooned peaksid olema võimalikult sirged ja täpse asukohaga.

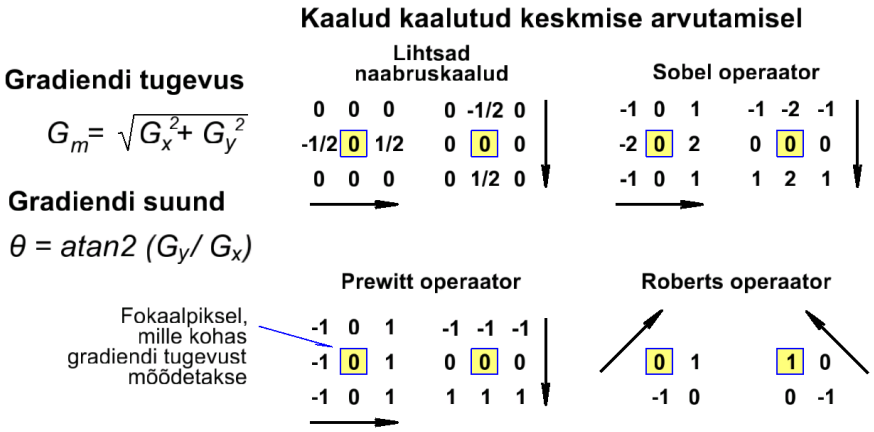
Keerukam, kuid realistlikumaid tulemusi andev segmenteerimine on järkjärguline, igal eraldatud objektil (segmendil) on naabrus, superobjekt, objektisisene varieeruvus ja enamasti võivad olla ka alamobjektid, mis kõik annavad suure hulga võimalikke tunnuseid.

Segmenteerimine võib toimuda mitmete algoritmide järgi. Väärtuspinna segmenteerimise viisid oleksid laias laastus järgmised.

Grupeerimine on olemasolevate ruumiosade (pikslid, eraldised) koondamine külgnevateks ja omavahel maksimaalselt sarnasteks suuremateks ruumiosadeks. Tekstuuri analüüsile tuginevat pikslite grupeerimist on kasutanud Pietikäinen ja Rosenfeld (1981), Chou *et al*. (1994), Spann ja Wilson (1985), Beveridge *et al.* (1989), Woodcock ja Harward (1992), Shandley *et al.* (1996), Ryherd ja Woodcock (1996).

Globaaltsoneerimine (*global zonation*) alustab kogu pinnast ja püüab seda võimalikult homogeenseteks eraldisteks jagada. Näiteks tarkvaras eCognition algab segmenteerimine kõige kaugematest ja kõige erinevamatest pikslitest.

Servatuvastus (*edge detection*) on üks põhilisi pilditöötluse ja eriti objektide automatiseeritud tuvastamise vahendeid, mille käigus püütakse leida kujutise järsema muutumise kohti. Servakontrasti otsimise lokaalsed meetodid kasutavad piiri otsimisel vaid potentsiaalse piiri lähedal olevaid kohti. Lokaalselt kasutatavaid kaalumisfunktsioone nimetakse seejuures operaatoriteks. Operaatorid erinevad naabrite asendist sõltuvate kaalude poolest (joonis 4-29). Objektituvastuse (*feature detection*) algoritmid tegelevad lisaks servatuvastusele veel nurgatuvastusega, ühetaoliste alade tuvastusega (*blob detection*), harja tuvastusega (*ridge detection*) ja mitmesuguste etteantud struktuuritüüpide tuvastamisega.



Joonis 4-29. Lokaalsed filtrid gradiendi tugevuse mõõtmiseks.

Servatuvastuse filtrid toimivad järgmiselt. Aken jagatakse kaheks või enamaks osaks vastavalt sellele, millises suunas või suundades piire otsitakse. Akna osade vahel arvutatakse erinevuse ehk lokaalse gradiendi mõõdik. Akent liigutatakse üle kogu pinna ja fikseeritakse erinevuse mõõdud igas kohas. Suurima erinevuse vööndid on ruumiliste eraldiste piiride paiknemiskohana kõige enam põhjendatud. Üleminekukohtade võimendamiseks saab kasutada Laplace filtrit, mida rakendatakse kujutise intensiivsuse tuletisele.

D. Gill (1970) kasutas piiri parima koha leidmiseks segmendisiseseid (SS_w) ja segmentide vahelisi ruuthälvete summasid (SS_b). Piiri parim koht on seal, kus R on suurim [4-88]. Osadeks jagamist korratakse, kuni R enam ei suurene.

$$R = \frac{SS_b - SS_w}{SS_b} \qquad [4\text{-}88]$$

J. Canny (1986) esitas servatuvastuse teooria, mille kohaselt jaguneb servatuvastuse efektiivsus kolmeks:

- objektide tõene äratundmine,
- objektide tõene piiritlemine,
- vähene tundlikkus kujutises oleva müra suhtes.

Canny algoritmi algab müra vähendamist kujutises normaaljaotuse kõverat järgiva lokaalse filtriga. Sellele järgneb kujutise intensiivsusgradiendi tuvastamine lokaalse operaatoriga eraldi neljas suunas (horisontaalne, vertikaalne ja kaks diagonaalsuunda). Iga leitud piir klassifitseeritakse vastavalt ülemineku suunale. Piiride kujutisel kasutatakse suunaklasside tähistamiseks värve. Siis leitakse gradiendi tugevuse maksimum igas eelnevalt leitud üleminekuvööndis. Seejärel toimub servajoonte pidevuse kontroll ja objekti piiri mitte tähistavate üleminekute eemaldamine. Lõpuks on võimalik objekte tähistavate eraldiste moodustamine vektoriseerimine.

Vombling on gradiendi otsimise lokaalne meetod, mille puhul keskmistatakse ruumiliste muutujate tuletiste absoluutväärtusi. Tuletiste arvutamisel kasutatakse liikuvat akent. Meetod on saanud nime W.H. Womble (1951) järgi (Oden *et al.* 1993, Fortin 1994). Seda on kasutatud geneetikas ja metsanduses. Muutuja Z ruumis muutumise kiirust m saab arvutada tuletisena:

$$m = \sqrt{\left[\partial f(x,y)/\partial x\right]^2 + \left[\partial f(x,y)/\partial y\right]^2}, \quad [4\text{-}89]$$

$$f(x,y) = Z_A\,(1-x)(1-y) + Z_B\,x(1-y) + Z_C\,(1-x)y + Z_D\,xy. \quad [4\text{-}90]$$

Võrevombling nõuab regulaarselt paiknevat andmestikku, triangulaarvomblingu lähteandmed võivad paikneda ebakorrapäraselt. Ebakorrapäraselt paiknevad andmed võib korrapärasesse võrgustikku interpoleerida, kuid interpoleerimine võib olemasolevaid gradiente maskeerida.

Triangulatsioonvombling arvutab lokaalse muutuse kiirust lähima kolme punkti kaalutud erinevuse abil. Ebakorrapäraste kolmnurkade võrgustiku (*triangulated irregular network – TIN*) korral võrdub muutuse kiirus kolmnurga kaldenurgaga. Seega künka ja oru piiritlemise ülesande lahendab vombling kõige järsema nõlva järgi.

Vomblingut saab kasutada ka mitmefaktoriliselt ja kategooriliste andmetega. Esimesel juhul määratakse keskmine muutumise kiirus mitme faktori muutumise kiiruse keskmisena. Teisel juhul omab olulist tähtsust otsus, kas klasside erinevused on sama suured või mitte. Vomblingul leitud piiride kontrastsust saab võrrelda samade andmete korduvate juhuslike ümberpaigutustega.

Struktuursete üksuste eristamise puhul võrreldakse iga koha (piksli) arvulist väärtust ja naabruses olevate väärtustega ja saadakse piiritletud pinnavormid. Näiteks tarkvara Idrisi vahendiga *toposhape* on võimalik eristada 11 topograafilist vormi: tipp (*peak*), hari (*ridge*), sadul (*saddle*), tasandik (*flat*), jäärak (*ravine*), auk (*pit*), kumer nõlv ehk nina (*convex hillside*), saduljas nõlv (*saddle hillside*), ühtlane nõlv (*slope hillside*), nõgus nõlv ehk orvand (*concave hillside*) ja looklev nõlv (*inflection hillside*).

Pinnavormide hulk ja paiknemine sõltub väga olulisel määral üldistusastmest. Näiteks üldistatud mõõtkavas kujutatud mägi võib detailses mõõtkavas koosneda mitmest tipust ja tippudevahelistest sadulatest.

Topograafiliste vormide hierarhilist struktuuri on nimetatud ešelonideks (*echelons*). Ešelonid avalduvad eraldistena kaardil, atribuutide tabelina ja struktuuri ühelt ešelonilt teisele pärandumist esindava puuna (Myers *et al.* 1996, 1997). Ešelonide eristamist on kasutatud bioloogilise mitmekesisuse kaardistamisel (Myers *et al*. 1995).

4.3.2.3. Kujutise objektorienteeritud klassifitseerimine

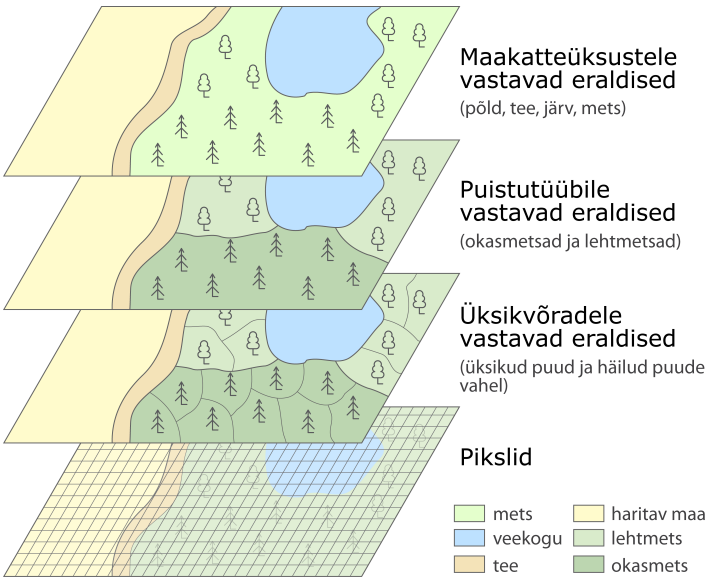
Kui objektid või maakatteklassid, mida kaugseirekujutiselt üritatakse eristada, on suuremad kui kujutise piksel, siis ei esinda üksikpiksel enam eristatavat kategooriat, vaid selle osa. Sama objekti piires võib olla erineva kiirgusväärtusega piksleid ja objekti iseloomustab pikslivääruste muster. Objektorienteeritud kaugseirekujutise töötlust on nimetatud geoinfoteaduse omaette osaks (*geographic object-based image analysis – GEOBIA*) (Hay ja Castilla 2008, Johansen *et al*. 2010), mis tegeleb kaugseirekujutise jagamisega sisukateks objektideks ja seejärel nende objektide klassifitseerimisega. GEOBIA tugineb nii segmenteerimise kui ka mustrituvastuse ideedele ning eristub objektipõhisest pilditöötlusest (*object-based image analysis – OBIA*), mida rakendatakse muudel aladel, nagu biomeditsiinis ja raalnägemises (*computer vision*).

Kaugseirekujutise objektorienteeritud töötluse vajadust põhjendavad Hay *et al.* (2005) järgmiste asjaoludega.

- Detailse kujutise üksikud pikslid ei esinda klassifitseeritavaid objekte, vaid nende osi.
- Inimesed ei taju maailma pikslitena, vaid erinevas suuruses objektidena.
- Objektidest koosneva ruumilise andmestikuga töötamiseks on olemas piisavalt palju meetodeid ja tarkvara ning nende arendamine jätkub.

- Objektideks jagamine suudab paremini kirjeldada ökosüsteemide keerukat mitmemõõtkavalist sisestruktuuri, mille komponendid on omavahel mitmesugusel viisil seotud.
- Objektorienteeritud kaardistamine pakub lahendust muudetava pindüksuse probleemile (ptk 4.3.5).

Kuna looduse kaardistamisel ei ole ühte ja igas olukorras sobivaimat mõõtkava, siis moodustatakse kujutisest mitmel hierarhilisel tasemel objekte esindavad eraldised. Nende eraldiste omadused koosnevad eraldise sees olevate pikslite, eraldist moodustavate detailsemate eraldiste ja eraldist hõlmavate suuremate eraldiste atribuutidest (joonis 4-30).



Joonis 4-30. Mitmemõõtkavaline segmenteerimine. Igal detailsustasemel kasutatakse objektide eristamiseks erinevaid, sealhulgas teiste tasemete tunnuseid. Joonise koostas Tanel Tamm.

Mitmemõõtmelises objektorienteeritud käsitluses moodustatakse mitmel detailsustasemel eraldised

Spetsiaalne kaugseirekujutise segmenteerimise ja eraldiste klassifitseerimise standardtarkvara on eCognition (http://www.ecognition.com/sites/default/files/eCognition_Software.pdf).

4.3.2.4. Üldistatud erinevusanalüüs

Ökoloogilisel regionaliseerimisel liidetakse sarnaste keskkonnaomadustega kohad üheks piirkonnaks, milles tuleks eelistada ühesuguseid majandamisviise ja planeerimispõhimõtteid. Kvantitatiivsel ökoloogilisel regionaliseerimisel on kõige sagedamini kasutatud mitmetunnuselist sarnasusele ja erinevusele tuginevat kirjeldavat rühmitamist, milles rakendatakse vaid keskkonnatunnuseid (Bunce *et al.* 1996, Hargrove ja Hoffman 2005, Leathwick *et al.* 2003, Mackey *et al.* 2008, Metzger *et al.* 2005, Sheail ja Bunce 2003, Trakhtenbrot and Kadmon 2005).

Keskkonnatunnuste järgi rühmitamist on täiustatud elustiku sarnasuse/erinevuse ja keskkonnafaktorite sarnasuse/erinevuse seose kaasamisega. Clarke ja Ainsworth's (1993) esitatud metoodika kohaselt tuleks esmalt arvutada eraldi maatriksid vaatluskohtade biootilise erinevuse kohta ja keskkonnafaktorite väärtuste erinevuse kohta samades vaatluskohtades. Biootilise erinevuse maatrikseid on üks, keskkonna erinevuse maatrikseid mitu – eraldi iga keskkonnatunnuse jaoks ja iga tunnuste kombinatsiooni jaoks. Keskkonnatunnuste kombinatsioonid sisaldavad tunnuseid kahekaupa, kolme-

kaupa jne kuni kõigi keskkonnatunnuste kooskasutuseni. Seejärel tuleb arvutada astakkorrelatsioon biootilise erinevuse maatriksi ja iga üksiku keskkonna erinevuse maatriksi vahel. Suurim korrelatsioonikordaja näitab keskkonnatunnuste kombinatsiooni, mis kõige paremini esindab biootilisi erinevusi uuritaval alal. Meetod võimaldab seoste üksikasjalikku esitust, kuid on arvutusmahukas.

Ferrier ja kaasautorid tutvustasid eelkirjeldatud meetodit andes sellele nime **üldistatud erinevusanalüüs** (*generalised dissimilarity analysis – GDM*) (Ferrier *et al*. 2002b, 2007, Ferrier 2002). GDM metoodika võimaldab määrata keskkonnatunnuste sobivaimad teisendused ja kaalud enne klasteranalüüsi ning visualiseerida koos temaatiliste regioonide piiridega ka nende regioonide omavahelist sarnasust (Snelder *et al.* 2006).

GDM seob elustiku ja keskkonnatunnuste varieeruvuse ning võimaldab valida elustiku erinevustega seonduva keskkonnatunnuste kombinatsiooni

4.3.2.5. Kontekstist sõltuv klassifitseerimine

Pilditöötluses, sealhulgas kaugseirekujutise klassifitseerimisel mõistetakse kontekstist sõltuva klassifikatsiooni all täiendavate andmeallikate kasutamist. Kujutisele lisanduvad täiendavad andmeallikad on eelkõige topograafilised kaardid ja varasemad andmed uuritava tunnuse kohta. Lisaandmeid saab kasutada enne või pärast klassifitseerimist või ka klassifikatsiooni ajal, ruumilise või temaatilise eelklassifikaatorina või järelklassifikaatorina või tulemust mõjutava tunnusena.

Landsat TM pildi eelkorrektuuris kasutatakse näiteks nõlva ja valguse langemise nurka arvestavat korrektuuri Minnaert (1941) konstandi järgi

$$\log(L \cdot \cos(e)) = k \cdot \log(\cos(i) \cdot \cos(e)) + \log(L_n). \qquad [4\text{-}91]$$

Piksli väärtuse korrigeerimiseks kasutatakse Minnaerti korrektuuri (Colby 1991)

$$L_n = \frac{L \cdot \cos(e)}{\cos^k(i) \cdot \cos^k(e)}, \qquad [4\text{-}92]$$

kus L on kiirgus, L_n on normaliseeritud kiirgus, kiirgus kui $e = i = 0$, e on nõlvakalle, i on valguse langemise nurk, k on empiiriliselt leitav konstant.

4.3.2.6. Tekstuuri tuvastamine

Haralick *et al.* (1973) jagasid inimtajule vastava mustrituvastuse kolmeks: **spektraalseks, tekstuurseks** ja **kontekstuaalseks**. Spektraalsed erinevused kirjeldavad erinevate kiirgusvahemike kiirguse vahekorda ja selle varieeruvust, tekstuur kirjeldab kiirguse varieeruvust samas kiirgusvahemikus, konteksti tunnused iseloomustavad koha ümbrust. 1979. aasta töös jagas Haralick tekstuuri omakorda **statistiliseks** ja **struktuurseks** (Haralick 1979). Statistiline tekstuur iseloomustab väärtusklasside vahekorda, struktuurne tekstuur iseloomustab väärtusklasside paiknemist üksteise suhtes.

Tekstuur iseloomustab väärtusklasside vahekorda ja paiknemist üksteise suhtes

Lark (1996) pidas tekstuuri iseloomustamisel vajalikuks nelja omaduse kirjeldamist:

- pikslivääртuste varieeruvus,
- pikslivääртuste varieeruvuse sõltuvus pikslite vahemaast ja mõõtkavast,
- pikslivääртuste varieeruvuse sõltuvus pikslite paiknemise suunast (anisotroopia),
- pikslivääртuste varieeruvuse ruumiline perioodilisus.

Digitaalselt salvestatud kujutised koosnevad reeglina erineva lainepikkusega peegeldunud või emiteeritava kiirguse intensiivsuse arvulistest väärtustest. Iga lainepikkuse vahemikku saab analüüsida omaette halltoonis pikslitest koosneva kujutisena, milles igal pikslil on halltooni intensiivsust kirjeldav reaalarvuline väärtus. Tekstuuri kirjeldamiseks jagatakse halltoon heledusklassidesse. Kuna arvutuste maht ja vajaliku mälu maht suureneb kiiresti heledusklasside arvu suurenedes, piirdutakse tavaliselt mitte enam kui 16 heledusklassiga. Kuna korraga kirjeldatakse kas vaid ühe spektrivahemiku või pankromaatse (halltoonides) kujutise tekstuuri, nimetatakse tekstuuri kirjeldavaid sõltuvusi **halltooni paiknemisseosteks** (*gray-tone spatial dependence*, *gray-level dependences*).

Halltooni klasside külgnevuste sagedust kirjeldatakse maatriksina, mille komponendid $P(i, j, d, \theta)$ tähistavad tõenäosust, et kaks pikslit väärtusklassidest i ja j esinevad teineteisest kaugusel d ja nurga all θ (joonis 4-31). Kui tekstuuri kirjeldamisel kasutada spektri erinevaid osi, lisandub sagedusmaatriksisse veel viies dimensioon. Tekstuuri kirjeldamiseks saab kasutada sagedusmaatriksist tuletatud üksikparameetreid. Kujutise klassifitseerimiseks tuleb tekstuuri iseloomustavaid maatrikseid mingite näitajate alusel võrrelda ja mõõta nende sarnasust etteantud näidistega või nendest arvutatud statistikutega.

Joonis 4-31. Tekstuurimaatriksite moodustamine naabruse neljas suunas.

Enamkasutatud on järgmised sagedusmaatriksite statistikud (Haralick *et al*. 1973): teise järgu nurkmoment (*angular second moment*) [4-93], kontrast [4-94], korrelatsioon [4-95], dispersioon [4-96], homogeensus ehk erinevuse pöördmoment (*inverse difference moment*) [4-97], klassisummade keskväärtus [4-98], klassisummade dispersioon [4-99], klassisummade entroopia (kuna logaritm nullist on määramatus, siis soovitab Haralick sagedustele väikese arvu juurde liita) [4-100] ning entroopia [4-101].

$$f_1 = \sum_{i=1}^{N}\sum_{j=1}^{N}[p(i,j)]^2 \quad [4\text{-}93]$$

$$f_2 = \sum_i \sum_j (i-j)^2 \cdot p(i,j) \quad [4\text{-}94]$$

$$f_3 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} i \cdot j \cdot p(i,j) - \mu_x \cdot \mu_y}{\sigma_x \cdot \sigma_y} \quad [4\text{-}95]$$

$$f_4 = \sum_i \sum_j (i-\mu)^2 \cdot p(i,j) \quad [4\text{-}96]$$

$$f_5 = \frac{\sum_i \sum_j p(i,j)}{1+(i-j)^2} \quad [4\text{-}97]$$

$$f_6 = \sum_k k \cdot p(k) \quad [4\text{-}98]$$

$$f_7 = \sum_k (k - f_6)\, p(k) \quad [4\text{-}99]$$

$$f_8 = -\sum_k p(k) \log[p(k)] \quad [4\text{-}100]$$

$$f_9 = -\sum_i \sum_j p(i,j) \log[p(i,j)] \quad [4\text{-}101]$$

Kasutatud tähistus: *p(i,j)* on klassi *i* ja *j* esinemise suhteline sagedus teatud vahemaaga ja teatud nurga all, külgnevuste sageduse maatriksi element, *N* on halltooni klasside arv, *k* on klassikombinatsiooni indeks, $k = i + j$, μ on klassiväärtuste keskväärtus, μ_x ja σ_x on sagedusmaatriksi ridade summade keskväärtus ja standardhälve, μ_y ja σ_y on sagedusmaatriksi veergude summade keskväärtus ja standardhälve.

Suhteliselt lihtsad tekstuuri parameetrid on liikuva akna sees arvutatud lokaalne standardhälve ja dispersioon. **Lokaaldispersioonide meetodi** (*local variance method*) puhul arvutatakse ümbruses olevate pikslite väärtuste dispersioon või standardhälve iga pildil oleva koha puhul eraldi. Meetodit on kasutatud maakatte ja maapinna kaardistamiseks kaugseire piltide järgi (Logan *et al*. 1979, Woodcock ja Harward 1992, Arai 1993, Chica-Olmo ja Abarca-Hernández 2000, Dirnböck *et al*. 2003). Lisaks lokaaldispersioonile ja lokaalsele standardhälbele on tekstuuri iseloomustamiseks kasutatud ka naaber-pikslite kontrasti (Edwards *et al*. 1988), piksli väärtuste lokaalset haaret ja poolhajuvust (Hudak ja Wessman 1998). Viimatimainitud autorid leidsid selge seose savanni kujutise kahemeetrise küljega pikslite väärtuste poolhajuvuse lävendi ja puude katvuse vahel, suurema katvuse puhul on pikslite lokaalvarieeruvus tugevam. Seejuures sõltub sama maastiku kujutiste tekstuur ka piksli suurusest.

Varieeruvuse partitsiooni meetod (*variance partition*) jagab ökoloogilise nähtuse ruumilise varieeruvuse neljaks osaks: puhtalt keskkonnast tulenevaks, puhtalt ruumiliseks, ruumilistest (ümbruse) ja keskkonnamõjudest tulenevaks ja seletamata varieeruvuseks (Legendre ja Legendre 1998). Kuna meetodi algne variant ei suutnud piisavalt kirjeldada lokaalset varieeruvust, täiendasid

Borcard ja Legendre (2002) meetodit lähestikku olevate kohtade vahemaade peakoordinaatanalüüsiga.

Taimekoosluste äratundmiseks kõrge lahutusega fotodelt on kasutatud ka variogramme (ptk 5.3.3). Erinevale taimkattele vastab erinev variogramm. Variogramme saab võrrelda variogrammi mudeliga parameetrite abil.

4.3.2.7. Spektraalmikstuuri analüüs

Spektraalmikstuuri analüüsis (*spectral mixture analysis – SMA*) käsitletakse piksleid erinevat tüüpi alampikslite seguna. Iga alampiksli tüübile iseloomulik peegeldumisväärtus kalibreeritakse eraldi. Iga piksli puhul analüüsitakse, milliste alampikslite millisest vahekorrast võisid vaadeldud peegeldumisparameetrid tekkida (Sabol *et al*. 2002).

4.3.2.8. Kerneli ümberklassifitseerimine

Kerneli ümberklassifitseerimise algoritmi (*kernel reclassification algorithm*) kohaselt tuleb kõigepealt kõik pikslid klassifitseerida ühekaupa. Klassid ei pruugi olla ette antud, tavaliselt piisab 6 kuni 12 klassist. Seejärel leitakse pikslitele omistatud klasside sagedus ja paiknemine teatud suurusega aknas (tuumikus ehk kernelis). Klasside sagedust vastavalt paiknemisele saab esitada külgnevuste maatriksina (*adjacency-event matrix*) M, kus f_{ij} – piksliklassi *i* ja *j* külgnevuse sagedus.

$$\mathrm{M} = \begin{bmatrix} f_{11} & f_{12} & f_{1n} \\ f_{21} & f_{22} & f_{2n} \\ & f_{ij} & \\ f_{n1} & & f_{nn} \end{bmatrix} \qquad [4\text{-}102]$$

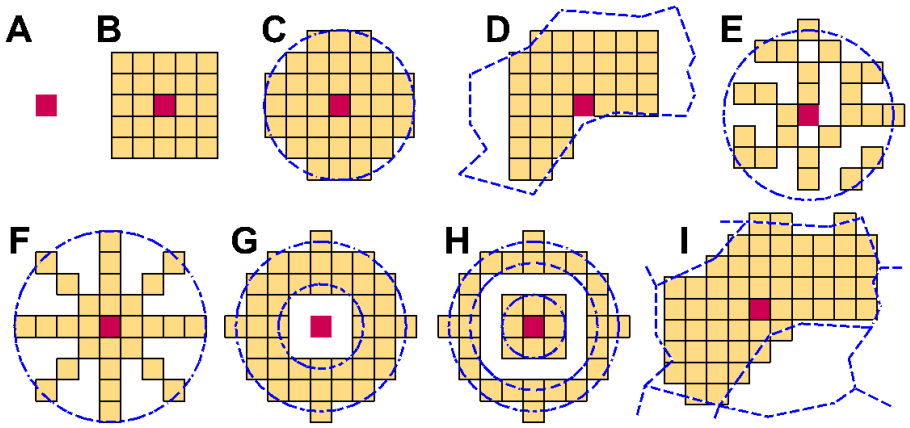
Algoritmi teises etapis võrreldakse iga klassifitseeritava koha ümber olevas aknas arvutatud külgnevuste maatriksi sarnasust näidiskohtades arvutatud külgnevuste maatriksitega. Kohad klassifitseeritakse kõige sarnasemasse üksusesse. Meetod sobib eelkõige juhul, kui otsitavad üksused on heterogeense struktuuriga – koosnevad erineva väljanägemisega allüksustest.

Meetodit on kasutatud linna maakasutusüksuste kaugseires (Barnsely ja Barr 1996) ja mosaiiksete alade maakatteüksuste kaugseirepõhisel kaardistamisel (Keramitsoglou *et al.* 2005, 2006, Kobler *et al.* 2006).

4.3.2.9. Lokaalstatistikud

Lokaalstatistikute mõistet tuntakse eelkõige sotsiaalmajanduslikus tähenduses, kus need märgivad statistilisi näitajaid omavalitsuse või naabruskonna tasandil vastandudes riigi tasandil statistilistele näitajatele ja täiendades neid. Loodusandmete analüüsis käsitletakse lokaalstatistikuid eelkõige lokaalselt arvutatud pilditöötluse või mustrituvastuse vahenditena. Lokaalstatistikud arvutatakse arvesse minevat ala piiravas aknas ehk kernelis või siis kaugusest sõltuvate kaaludega fookuses oleva koha ehk fokaalkoha ümber. Kernel on kas korrapärase kujuga (enamasti ruut, ring, sektor või sõõrik) või kujutise eelneva segmenteerimise abil saadud suvalise kuju ja suurusega eraldis ehk geograafiline kernel (Merchant 1984). Kernelitüüpe saab kombineerida, näiteks võivad eraldised piirata korrapärast või juhusliku valikuga kernelit (joonis 4-32). Sobiva kernelisuuruse valiku küsimust on põhjalikumalt käsitlenud Hodgson (1998), Ricotta *et al*. (2003) ja Linder *et al.* (2008). Idrisi vormingusse rasteriseeritud pindade lokaalseid tunnuseid saab arvutada tarkvaras LSTATS (www.geo.ut.ee/LSTATS).

Lokaalstatistikud iseloomustavad igat kohta ja selle ümbrust ükshaaval, globaalstatistikud iseloomustavad kogu urimisala tervikuna



Joonis 4-32. Erinevatesse kernelitesse kaasatud pikslid (Proosa 2008, Linder *et al.* 2008, täiendatud). A – vaid fokaalpiksel; B – ruudukujuline kernel 5 × 5 pikslist; C – ümar kernel raadiusega 3,5 pikslikülge fookusest; D – eraldisega piiratud ruudukujuline 7 × 7 pikslist; E – pikslite juhuslik 50% valim ümarast kernelist (raadius 4 pikslit); F – kiirtekujuline valim ringikujulises kernelis raadiusega 4,5 pikslit; G – sõõrik-kernel siseraadiusega 2 ja välisraadiusega 4 pikslit; H – ümara sisekerneli (raadius 1,5 pikslit) ja välise sõõri (siseraadius 3, välisraadius 4 pikslit) kombinatsioon; I – eraldise piiridega määratud geograafiline kernel.

Uurimused

Lokaalset maastikuökoloogiliste üksuste piiritlemist kaugseirekujutise alusel on kasutatud näiteks Forman ja Godron (1986), Legendre ja Legendre (1998), Warrick *et al.* (1986), Whittaker (1956, 1960).

Woodcock *et al.* (1988a,b) seostasid metsa struktuuri kujutisest arvutatud variogrammidega. Nad leidsid, et puude kõrgus oli seotud variogrammi lävendiga ja puude suuruse varieeruvus variogrammi kujuga. Kuigi puistu struktuuri tuvastamiseks on üldiselt vaja kõrge resolutsiooniga kujutisi, näitasid Cohen *et al*. (1990), et nõrgad seosed on leitavad isegi Landsat 30 m küljega piksli puhul. Hudak ja Wessman (1998) leidsid selge seose savanni kujutise kahemeetrise küljega pikslite väärtuste poolhajuvuse lävendi ja puude katvuse vahel, suurema katvuse puhul on pikslite lokaalvarieeruvus tugevam.

Järjest suurenevates akendes olevate väärtuste maksimaalset dispersiooni on kasutatud puistu struktuuri kirjeldamiseks fotode ja videokujutise abil (Woodcock ja Strahler 1987, Coops ja Catling 1997, 2001 Coops *et al*. 1998, Coops ja Culvenor 2000). Varieeruvus on maksimaalne kui aken on sama suur kui enamik puuvõrasid. Lokaalvarieeruvus sõltub ka eri liiki puude paiknemismustrist. Kujutise tekstuuri analüüsi on puistu struktuuri kirjeldamiseks edukalt kasutanud ka Atkinson ja Danson (1988), Cohen *et al*. (1990), Nel *et al*. (1994), St-Onge ja Cavayas (1995).

Tokola ja Shrestha (1999) kasutasid vaatlusalal paikneva piksli ja vaadeldava piksli erinevuse hindamiseks järgmist valemit (kõrgus on valemis seetõttu, et uurimus tehti Nepaalis, kus kõrguserinevused on olulisim kasvukohtade erinevusi määrav faktor).

$$d_{ij} = \sqrt{e_{ij}^2 + \sum_{h=1}^{nc} (p_h c_{ijh})^2}, \qquad [4\text{-}103]$$

millest arvutati vaatlusalade kaalud w_{ij} vaadeldava piksli väärtuse hindamiseks järgmiselt:

$$w_{ij} = \left(\frac{1}{d_{ij}} \right)^t, \qquad [4\text{-}104]$$

kus *nc* on kanalite arv, *h* – kanali indeks, p_h on kanali empiiriline konstant, *i* on piksel vaatlusalal, *j* on vaadeldav piksel, e_{ij} on kohtade kõrguste erinevus, c_{ijh} on pikslite spektraalväärtuste erinevus, *t* on kaalufunktsiooni empiiriline konstant.

Muutuja hinnang mingi piksli jaoks saadakse ümbritsevatel vaatlusaladel mõõdetud väärtuste kaalutud keskmisena üle kõigi Landsat TM kanalite.

$$\hat{y}_j = \frac{\sum_{i=1}^{np} w_{ij} x_i}{\sum_{j=1}^{np} w_{ij}} \qquad [4\text{-}105]$$

Hinnangud arvutati ka vaatlusalade pikslite kohta. Satelliidipildi järgi hinnatud ja maa peal mõõdetud väärtuste erinevus annab vea hinnangu. Kui muutuja on enam-vähem normaaljaotusega, siis hinnatakse viga enamasti standardhälbe abil.

Kilpeläinen ja Tokola (1999) kasutasid olemasolevaid metsakorraldusandmeid satelliidipildi järgi puistu tagavara hindamisel. Prognoosis kasutati pikslite kaalumist sarnasuse järgi. Prognoosis arvestati eraldiste piire ja eraldiste piiril olevaid piksleid töödeldi eraldi. Tulemuse täpsus sõltus tugevasti metsakorralduse eraldiste piiride ja satelliidipildi geomeetrilise ühildamise täpsusest. Tokola ja Kilpeläinen (1999) selgitasid puidu tagavara prognoosi täpsust metsakorralduse eraldiste piiril olevate segapikslite kasutamisel ja ilma piiripiksliteta. Eraldise piirialade kohta on prognoos vähetäpne, kuid segapikslite väljajätmine testandmetest põhjustaks nihkega hinnangu. Parim lahendus on kombineerida külgnevate vaatlusalade andmed. Tokola (2000) näitas, et puidu tagavara hindamisel näidisalade ja Landsat TM kujutise järgi peaksid näidisalad paiknema mitte kaugemal kui 20km prognoositavast kohast.

Diniz-Filho ja Malaspina (1995) uurisid kodumesilaste morfomeetriliste tunnuste varieeruvust Brasiilia populatsioonides. Nad leidsid tunnuste geograafilisest varieerumisest kinnitust hüpoteesile, et neotroopilistes tingimustes toimub Euroopa päritolu *Apis mellifera* populatsioonides Aafrika rassi geenide osakaalu suurenemine. Kanoonilise trendpinna korrelatsioonijääkide autokorreleerumist selgitati võrreldes statistiliselt oluliste Morani *I* koefitsientide suhtelist hulka erinevates kaugustsoonides ühelt poolt originaalandmetes ja teiselt poolt trendpinnal.

Segmenteerimisülesande lahendamist eeldab igasugune objektorienteeritud pilditöötlus (Baatz ja Schäpe 1999, 2000, Bian ja Butler 1999), metsa kaugseirekujutise jagamine eraldisteks (Woodcock ja Harward 1992) ja üksikute puuvõrade eraldamine aerofotol (Uuttera *et al*. 1998, Leckie *et al*. 2003). Segmenteerimisega on tegeletud ka liikide leviku modelleerimisel (Schuerholz 1974, Patton 1975), populatsioonidünaamika vallas (Cole 1954), taimkatte gradientide analüüsil (Beals 1969, Wilson ja Mohler 1983) ja muudel teadusaladel.

Andréfouët ja Claereboudt (2000) rakendasid erinevusmaatriksite vahelise korrelatsiooni arvutamist ehk üldistatud erinevusanalüüsi seostamaks kaugseirekujutist ja merepõhja omadusi. Snelder *et al.* (2007) kasutasid tunnustekombinatsioonide otsimisel üldistatud erinevusanalüüsis samm-sammulist protseduuri. Snelder *et al.* (2009) seostasid keskkonnatunnuseid tegeliku maakatte erinevusega, Snelder *et al.* (2010) mudelist prognoositud potentsiaalse maakattega.

4.3.3. Pindade vastavus ja selle statistilised testid

4.3.3.1. Kategooriliste pindade vastavus

Kategooriliste pindade puhul väljendub kovariatsioon kattuvuses olevate kategooriate suhtelise hulgana, mida analüüsitakse ülekatteoperatsioonides. Ülekatet on lihtne mõõta ja väljendada kattuva osa suhtarvuna (protsentides). Suhtarvude puhul tuleb muidugi jälgida, mille suhtes neid on arvutatud – kas kogu analüüsitava pinna või ühe teemakihi pindala suhtes.

Lihtsat pikslikaupa arvutatud ülekatet võivad tugevasti mõjutada väikesed asukoha hälbed ja muud ruumilised moonutused. Kahe sarnase mosaiikse pinna suhteliselt vähene nihutamine vähendab kattuvuses olevat osa kiiresti. Seetõttu võib kattuvuses oleva osa osakaalu määramiseks sobivam olla väikeste asukohahälvete suhtes vähem tundlik kategooriliste pindade erinevuse mõõt (ρ) (Seppelt ja Voinov 2003).

$$\rho_w(H_1, H_2) = \frac{\sum_{z=1}^{R}\sum_{i=1}^{S}\left|g_i\left(H_1 \cap U_w(z)\right) - g_i\left(H_2 \cap U_w(z)\right)\right|}{2 \cdot R \cdot g_i\left(H_1 \cap U_w(z)\right)}, \qquad [4\text{-}106]$$

kus: $\rho_w(H_1, H_1)$ on suhteline erinevus pindade H_1 ja H_2 vahel, z on piksli järjekorranumber R pikslist koosneval võrreldaval pinnal, i on võrreldavatel pindadel esineva klassi indeks S klassi hulgas, w on kerneli suurus (külje pikkus), $g_i\ (H_1 \cap U_w(z))$ on klassi i kuuluvate pikslite arv kernelis suurusega w piksli z ümber.

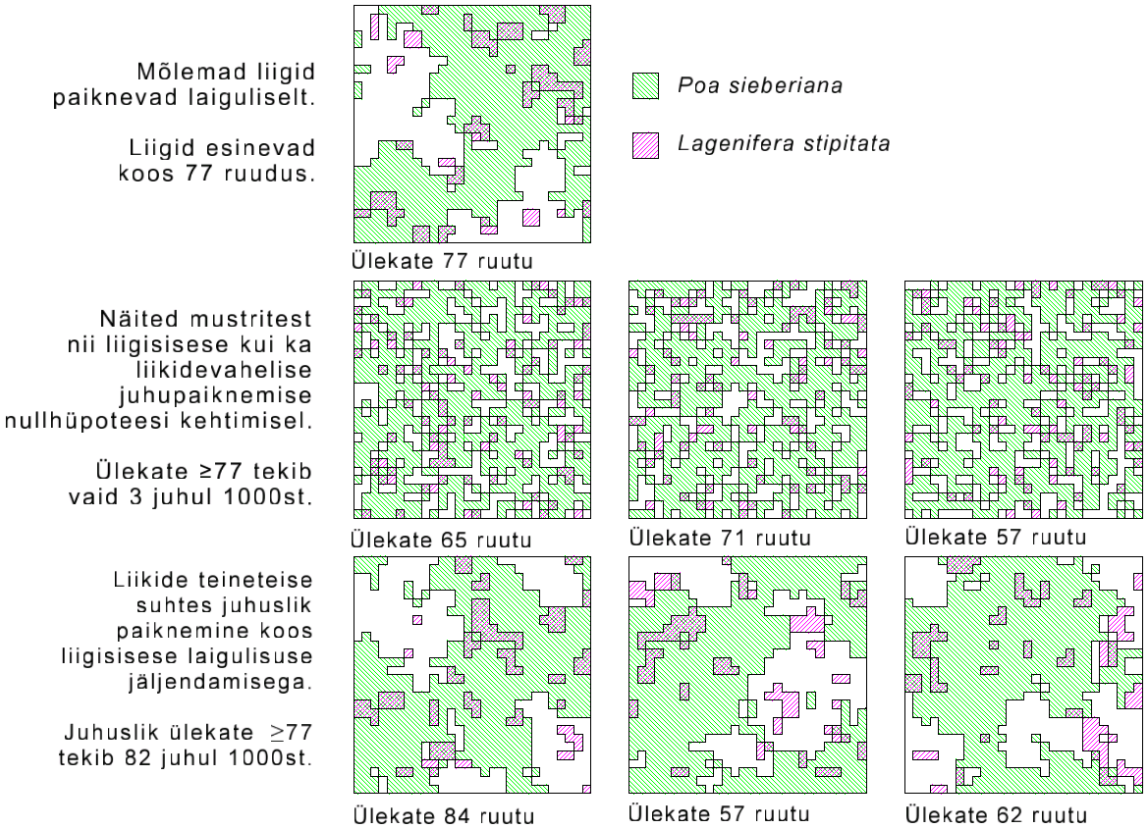
Pindade vastavuse ehk ülekatte mittejuhuslikkuse statistilise olulisuse määramise klassikaline näide on kahe rohttaimeliigi *Lagenifera stipitata* ja *Poa sieberiana* paiknemine vaatlusalal (Roxburgh ja Chesson 1998, joonis 4-33). Vaatlusala igas detailses lahtris loendatud liikide koos ja eraldi esinemise sagedused on tabelis 7. Hii-ruut test selle tabeli andmetest näitab, et liikide teineteise suhtes paiknemine ei ole juhuslik ($\chi^2 = 9{,}634$, $df = 1$, $P(x > \chi^2) = 0{,}003$). Seega võiks teha järelduse, et liigid esinevad sagedamini koos ja ka puuduvad samas vaatluskohas tunduvalt sagedamini kui juhuslike mustrite puhul võiks oodata.

Laikude juhusliku paigutuse tuhatkordne taastekitamine näitab aga, et sellise või suurema koosesinemise tõenäosus on juhumustrites tunduvalt suurem $p = 0{,}082$. Seega, klassikaline hii-ruut test ei arvesta pindade vastavuse hindamisel pindadele omast laigulisust ja võib seetõttu anda täiesti ebaõigeid tulemusi, sest laigulise mustri üksikud pikslid ei ole sõltumatud vaatlused, nagu tavastatistikud eeldavad. Ruumimustrite omavahelise seose statistilise olulisuse otsene määramine toimub korduva juhusliku nihutamise abil, mida kirjeldatakse järgnevalt ja peatükis 4.1.4.2 (Monte Carlo, ptk 3.6.6).

Hii-ruut test eeldab loendatud objektide sõltumatust, samasse laiku kuuluvad vaatluskohad ei ole omavahel sõltumatud

Tabel 7. χ^2 test *Lagenifera stipitata* ja *Poa sieberiana* esinemise vahel (Roxburgh ja Chesson 1998). H_0 = juhusliku paiknemise nullhüpotees.

	Lagenifera stipitata			
	esines		puudus	
Poa sieberiana	tegelik	ootus H_0 järgi	tegelik	ootus H_0 järgi
esines	77	63,25	429	442,75
puudus	21	34,75	257	243,25



Joonis 4-33. Austraalia taimede *Lagenifera stipitata* ja *Poa sieberiana* esinemine vaatlusalal (Roxburgh ja Chesson 1998 muudetult). Kui liikide esinemise laigulisust mitte arvestada, on nende koosesinemine $p \leq 0{,}05$ tasemel statistiliselt oluline, laigulisust arvestadades ei ole.

Põhiline meetod väärtusklasside kokkulangevuse statistilise olulisuse määramiseks arvestades pindade laigulisust või muud paiknemismustrit on pindade juhuslikus ulatuses ja juhuslikus suunas nihutamine teineteise suhtes. Nihutamise asemel saab kasutada ka juhuslikku pööramist või peegeldamist juhuslikult paikneva telje suhtes. Mõnevõrra arvutusmahukam on analoogiliste mustrite korduv jäljendamine (ptk 6.2).

Nihutamisoperatsioonide puudus on, et uuritava ala keskkoha nihke keskväärtus on väiksem kui servapikslitel, mis hüppavad vastasserva sagedamini. See omakorda võib põhjustada süstemaatilist viga ja testi võimsuse langust. Peale selle tekib toroidnihutamisel mustri sisse ühendatud servadele vastav ebanormaalne hüppekoht. Peegeldamisel on selline hüppekoht peegeldustelje kohas. Pööramisel muutub pöördetelje juures olevate kohtade paiknemine vähem kui kaugemal.

Nihutamismeetodeid on kasutatud eelkõige punktmustrite puhul (ptk 4.1.4.2) ja andmetes, kus vähemalt üks nihutatavatest pindadest on kategooriline pind. Iteratiivset nihutamist on aga võimalik kasutada ka pidevate ruumiliste muutujate puhul.

4.3.3.2. Pidevate pindade korrelatsioon

Reaalarvulisi väärtusi kandvate pindade korrelatsiooni saab mõõta tavastatistikas kasutatavate korrelatsioonikordajate abil, mille kasutamist ei ole põhjust nimetada ruumiliseks analüüsiks ja tulemust ruumiliseks korrelatsiooniks, sest väärtuste asukoht tulemust ei mõjuta. Sama tulemuse saaks, kui kohakuti olevate pikslipaaride väärtused suvalises järjekorras tabelisse kanda ja pikslite asukoht unustada. Kui kohakuti olevad pikslid on omavahel seotud, siis muul moel asukoht tulemust ei mõjuta.

Kahe rasterpinna kohakuti olevate pikslite väärtuste vahel arvutatud lineaarne korrelatsioon ei ole ruumiline korrelatsioon

Pindadevahelise korrelatsiooni arvutamisel kasutatakse reeglina lõplikku arvu kokkulangeva asukohaga esinduspunkte uuritavatest pindadest. Tõepärase hinnangu saamiseks peaksid esinduspunktid moodustama esindusliku valimi. Ühest küljest on suurem valim esinduslikum, kuid tiheda võrgustiku puhul paiknevad punktid üksteise lähedal. Kuna loodusele on iseloomulik teatud ruumiline pidevus, siis lähestikku olevad vaatlused ei ole üksteisest sõltumatud. Vaatluste sõltumatuse nõude eiramine mõjutab eelkõige pindadevahelise korrelatsiooni statistilise olulisuse hinnangut. Saadakse vaid näiliselt olulised seosed.

Kuidas leida seoseid mingis kohas oleva väärtuse ja selle ümbruses olevate kas sama tunnuse või teise tunnuse (pinna) väärtuste vahel, käsitletakse ümbruse mõju modelleerimise osas (ptk 5.5).

Tarkvara

Idrisi vormingusse rasteriseeritud pindade vastavuse statistilist olulisust saab juhuslike toroidnihutuste abil hinnata selle raamatu veebilehelt saada oleva programmiga Toroidpinnad.

4.3.4. Üleminekuala eristamine

Servatuvastuse meetoditega tegeletakse eelkõige kujutise automatiseeritud analüüsiga seotud tehnoloogilistel aladel, aga ka ökoloogias, mullateaduses ja geoloogias. Üldiselt seisnevad ruumiliste struktuuride üleminekud kas kvantiteedi muutumises, koosseisu muutumises või struktuuri muutumises. Van der Maarel (1976) jagab taimestiku üleminekud nelja tüüpi:

- piirid looduses olevate koosluste vahel,
- piirid taimestikutüüpide vahel,
- piirid mingis kaardistusprojektis kasutatavate üksuste vahel ning
- koosluse ajalised piirid.

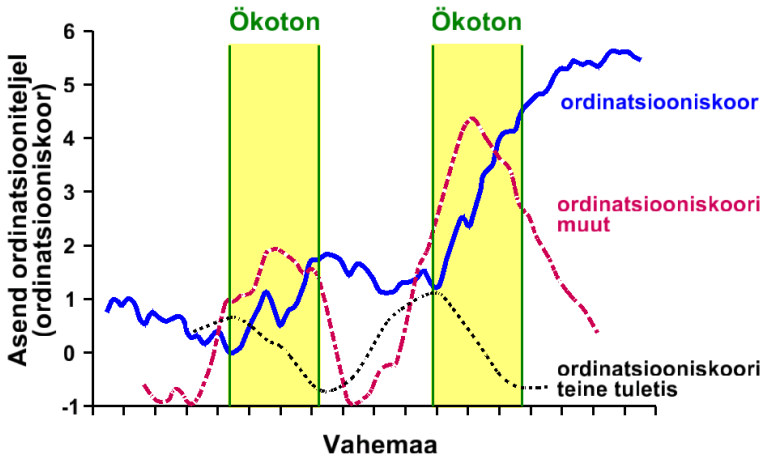
Taimekoosluste muutumise vööndit ehk **ökotoni** on tavapäraselt määratud külgnevate vaatluskohtade kirjelduste sarnasuse järgi. Saadakse sarnasuse või **erinevuse läbilõige** (*differential profile*) (Maarel 1974). Erinevuse läbilõike maksimumid, mis vastavad kohtadele, kus erinevused naabruses olevate vaatlustulemuste vahel on suurimad, näitavad koosluste üleminekukohti. Vaid ühe vaatlustrassil oleva naaberkoha arvestamisel mõjutavad erinevuse läbilõiget kooslusesisene varieeruvus ja juhuslikud erinevused. Ruumiliselt üldistatuma läbilõike saab liikuva akna kasutamisel, mille puhul ühendatakse akna ühes ja teises pooles olevad kirjeldused ja erinevuse mõõdik arvutatakse akna poolte vahel (Whittaker 1960). Ökoton ei pruugi avalduda, kui üleminekuala on kitsam kui akna laius. Soovitav on kasutada erineva suurusega liikuvaid aknaid, mille abil saadakse profiilid erinevas mõõtkavas muutuste jaoks (Webster 1973). Erinevuse mõõdikuks sobivad sarnasuskordajate (ptk 2.2.1) vastandväärtused.

Ökoton on koosluste suhteliselt järsema muutumise vöönd võrreldes ühetaolisema naabrusega

Ülemineku kontrasti võib määrata ka profiilide otstes olevate näidiskohtade suhtes. Walker *et al.* (2003) järjestasid (ordineerisid) vaatlused sarnasuse alusel. Liikuva akna igas peatuskohas sobitati lineaarne regressioonimudel esimesel ordinatsiooniteljel oleva väärtuse ehk ordinatsiooniskoori ja vaatluse asukoha vahel liikuva akna sees. Lokaalne regressioonikordaja näitab regressioonisirge tõusu ja seega muutuse intensiivsust. Üleminekuala piiritlemiseks arvutati ka muutuse teine tuletis, mis näitab muutuse suunda (joonis 4-34).

Üleminekuala eristamiseks on kasutatud ka lainekeste funktsiooni (Bradshaw ja Spies 1992, Camarero *et al*. 2006). Lainekeste funktsiooni abil määratud üleminekuala statistilist olulisust saab hinnata juhuslikult ümberpaigutatud lähteandmetest arvutatud lainekeste funktsioonide suhtes.

Eelmainitud serva tuvastamise meetodid määravad suhtelist kontrasti ja leiavad kõige järsema ülemineku ka juhuslikest arvudest koostatud arvreast. Cornelius ja Reynolds (1991) kirjeldasid kahte moodust, kuidas hinnata üleminekuala statistilist olulisust. Esimese meetodi puhul randomiseeritakse paljukordselt transektil olevate vaatluste asukoht. Randomiseeritud andmetest arvutatakse keskmine erinevus akna poolte vahel ja standardhälve liikuva akna igas peatuskohas. Eeldati, et erinevused on normaaljaotusega ning rohkem kui kahe standardhälbe kaugusel keskmisest peaks olema 5% juhtudest. Suurem erinevus vaadeldud andmetes loeti statistiliselt oluliseks. See meetod on rakendatav iga läbilõike puhul eraldi ja sõltub kasutatud aknasuurusest ning tunnuste varieeruvusest sellel läbilõikel.



Joonis 4-34. Ökotoni eristamine transektil vaatlustulemuste asendi järgi ordinatsiooniteljel (ordinatsiooniskoor) ja ordinatsiooniskoori muutuse järgi (Walker *et al.* 2003, muudetud).

Sõltuvust akna ühest kindlaksmääratud suurusest vähendab erineva suurusega liikuvate akende kasutamine. Väiksema akna puhul on erinevused akna poolte vahel üldiselt suuremad kui suure akna puhul. Erineva suurusega liikuva aknaga saadud erinevusprofiilide võrreldavaks muutmiseks standardiseerisid Cornelius ja Reynolds (1991) erinevusmõõdiku üksikväärtused *Z* skooriks juhukorduste üldkeskmise ja standardhälbe suhtes ja keskmistasid standardiseeritud erinevusprofiilid. Saadud läbilõige üldistab hierarhiliselt erinevas mõõtkavas ilmnevad üleminekud. Seda meetodit on rakendanud Hennenberg *et al.* (2005).

Mõlemad meetodid eeldavad transektil olevate vaatluskohtade erinevusmäärade normaaljaotust. Rangelt võttes ei saa erinevused kunagi normaaljaotusega olla juba ainuüksi seetõttu, et erinevuse miinimumväärtus on null. Vabanemaks normaaljaotuse eeldusest, võrdlesid Pärn *et al*. (2010) vaadeldud taimestikugradiente taimestikuvaatluste juhusliku paiknemise nullmudeliga, kasutamata standardhälbeid. Taimestikuvaatluste asukohad randomiseeriti sama transekti piires tuhat korda ja 5% olulisuse piir määrati 950nda korduse järgi. Kasutati ühepoolset hüpoteesi, sest hinnata sooviti vaid

juhuslikkuse korral oodatava järsema ülemineku tõenäosust. Sisukas hüpotees oli juhuslikust suurem erinevus, mitte ühetaolisus. Saadi hinnang vaadeldud taimkatteerinevuste statistilise olulisuse kohta. Enamikul vaatlustrassidest leiti vaid üks statistiliselt oluline üleminekukoht, aga mõnedel trassidel oli mitu taimkatte hüppekohta olulisustõenäosusega alla 0,05.

Tasub tähele panna gradiendi statistilise olulisuse suhtelisust, sest olulisus ei sõltu ainult muutuse järskusest, vaid ka vaatluste üldisest varieeruvusest trassil. Viimane omakorda võib oluliselt sõltuda vaatlustrassi pikkusest. Mida pikemalt ulatuvad vaatlustrassi otsad kummalgi pool olevale ühetaolisele alale, seda selgemini eristub nende kahe ühetaolise ala üleminekukoht. Kui vaatlused esindavad vaid üleminekuala, siis ei õnnestu seda ala üleminekualana esile tuua.

Üleminekukoha statistiline olulisus sõltub vaatluste hulgast ühetaolisel alal

4.3.5. Mõõtkava valik

Mõõtkava tajumine ja sobivaima detailsustaseme valik on oluline looduse struktuuri, looduse osade omavaheliste suhete ja protsesside mõistmisel. Uurimuseks kõige paremini sobiva mõõtkava ehk ruumilise detailsuse leidmine on ökolooge huvitanud juba aastakümneid (Kershaw 1960, Milne *et al*. 1989, Wiens 1989, Levin 1992). Ökoloogilises kontekstis on eristatud mõõtkava kolme aspekti – vaatluste detailsust, tulemuste üldistatust ja uuritava ala ulatust (Csillag *et al*. 2000, Dungan *et al*. 2002, Borcard *et al*. 2011). Vaatlustulemustes ei ilmne vaatluskohtade tihedusest väiksemad ruumilised struktuurid ja uuritava ala suurusest suuremad struktuurid ning kirjeldatud struktuuridest väiksemaid objekte ei saa analüüsida.

Mõõtkava võib tähendada nii vaatluste detailsust, tulemuste üldistustaset kui ka uuritava ala ulatust

Termineid suur ja väike mõõtkava ei mõisteta ökoloogias samuti, nagu neid mõistetakse kartograafias, kus suuremõõtkavaline kaart on detailsem kui väikesemõõtkavaline. Seetõttu on soovitav eelistada termineid üksikasjalik ehk detailne (*fine scale*) uuring ja laiahaardeline (*broad scale*) uuring, või veel parem – selgesõnaliselt esitada uurimuse ulatus ja detailsuse tase.

Ruumilise detailsuse mingi üks aste võib olla parim ühe kitsapiirilise probleemi ja uurimisobjekti puhul, uuritavad ökoloogilised nähtused on valdavalt keerukamad ja sõltuvad mitmes mõõtkavas teguritest. Seetõttu on mitmed autorid soovitanud töötada korraga mitmes mõõtkavas, näiteks kasutada mitme erineva pikslisuurusega lähteandmeid või fraktaalsust kirjeldavaid näitajaid (Borrough 1983, Milne 1988). Mitmes mõõtkavas käsitluse alternatiiviks võib olla mitmel erineval kaugusel oleva ümbruse omaduste ja võimalike mõjude lisamine koha tunnustele (Remm ja Luud 2003).

Mõõtkava probleemi olulisust on geograafias nimetatud **muudetavate pindüksuste probleemiks** (*modifiable areal unit – MAUP*) (Openshaw ja Taylor 1979, Openshaw 1984, Marceau 1999). Probleem tuleneb asjaolust, et uuritavat ala saab üksteist välistavateks allüksusteks jagada paljudel erinevatel viisidel, kusjuures jaotamise põhimõtted tulenevad enamasti üksikuuringu vajadustest ja võimalustest. Kuna iga sellise uurimuse tulemus sõltub kasutatud üksustest ja kuna üksused on vabalt muudetavad, siis ei ole erinevaid eraldisi kasutanud uuringute tulemused omavahel võrreldavad. Osalist lahendust muudetavate pindüksuste probleemile pakub hierarhiliselt mitmemõõtkavaline segmenteerimine (ptk 4.3.2.2) ning spektraalanalüüs ja selle analoogid.

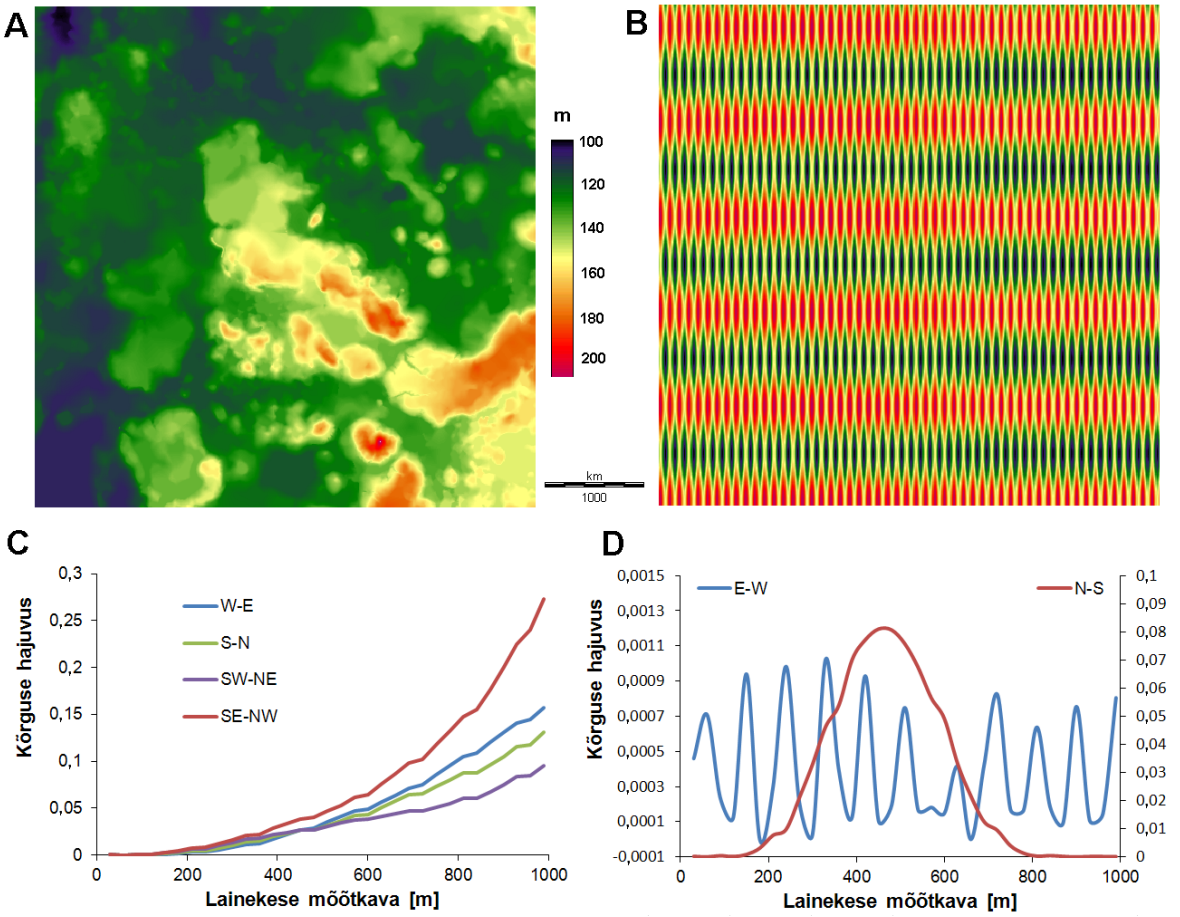
Saunders *et al.* (2005) võrdlesid ruumimustri kõige selgema esinemistaseme määramise kolme meetodit: lakunaarsusindeksit, spektraalanalüüsi ja lainekeste analüüsi (*wavelet analysis*).

Spektraalanalüüs on üldiselt efektiivsem perioodiliselt korduva struktuuriga mustri tuvastamisel. Mainitud uurimuses kasutatud andmestikes suutis lainekeste analüüs ära tunda peenemaid varieeruvuse mustreid kui lakunaarsusindeks. Spektraalanalüüsist on juttu aegridade kirjeldamise peatükis (ptk 3.5). Lakunaarsusindeks on lokaalselt arvutatud variatsioonikordaja analoog, mis arvutatakse järgmise valemi järgi (Plotnick *et al*. 1993).

$$\Lambda(w) = \frac{1+V(w)}{\overline{w}^2}, \qquad [4\text{-}107]$$

kus $V(w)$ on tunnuse w dispersioon, $\overline{w}$ – tunnuse w keskmine.

Ruumilistele struktuuridele iseloomuliku mõõtkava leidmiseks on kasutatud ka lainekeste analüüsi, millest on pikemalt juttu peatükis 4.1.2.23. Lainekeste analüüsi abil saab pinnavorme tuvastada (Kalbermatten *et al*. 2012), kõrgusandmeid üldistada (Bjørke ja Nilsen 2003) ja esile tuua pinnavormide iseloomulikku suurust ja suunda (joonis 4-35).



Joonis 4-35. Maapinna kõrgus [m] kaardilehel 54344 (Otepää) (A) ja sinusoidsete struktuuridega tehispind (B) ning väärtuste kaabufunktsiooniga (joonis 4-17) kaalutud keskmise hajuvuse sõltuvus selle funktsiooni kasutamise mõõtkavast ja suunast arvutatuna sama kaardilehe kõrgustest (C) ja sinusoidsete struktuuridega tehispinnast (D). Kõrgused standardiseeriti arvutuste eel. Graafikute mõningast ebaühtlust põhjustab kõrgusandmete ja lainekesefunktsiooni mõõtkava kasutamine kindla sammuga. Selle kaardilehe reljeefis ei ilmne kuni 1 km suuruste pinnavormide ühte domineerivat suurust. Tehisandmetes domineerivad põhja-lõuna (N-S) suunas umbes 480 m ja ida-lääne (E-W) suunas 48 m raadiusega struktuurid. Selle mõõtkavaga laineke läheb väärtuspinnaga resonantsi ja annab arvutatud kohtades suurima keskmise hajuvuse.

Kui punktmustri kirjeldamisel määratakse punktobjektide arvu hajuvust (valem 4-46) igas sektoris ja etteantud ulatuses ehk punktide lokaalset tihedust, siis väärtuspinna puhul saab määrata väärtuste hajuvust. Rakendades lainekese funktsiooni kui lokaalset fookuskoha ja selle ümbruse erinevust võimendavat filtrit kõrgusmudelile lokaalselt ja muutes funktsiooni skaalat, peaks ilmnema mõõtkava, mille juures läheb funktsioon reljeefivormidega kõige enam sünkrooni.

Lainekeste funktsioonid on kasutusel ka kujutiste ja andmeridade kokkupakkimisel, näiteks kasutatakse lainekesi jpeg failivormingus.

Uurimused

Üksiku liigi maastikueelistuste mõõtkavaga on tegeletud vähestes uuringutes. Schaefer *et al*. (2008) leidsid, et punahirve (*Cervus elaphus*) avatud ala eelistus ja okasmetsa vältimine avaldusid selgemini robustses kodupiirkondade paiknemise skaalas kui GPS vastuvõtjaga fikseeritud asupaigapunktide põhjal kirjeldatud liikumiskohtades.

4.4. Kolmemõõtmelise struktuuri kirjeldamine

Tüüpilised ökoloogide jaoks olulised kolmemõõtmelised struktuurid on atmosfäär, puistu, muld ja geoloogilised vormid. Puistu struktuurne mitmekesisus korreleerub elurikkusega, sest erinevad struktuurielemendid loovad suurema hulga mikroelupaiku. Puistu struktuuri iseloomustavad teatud tüüpi puude tihedus või suhteline hulk puistus, liigirikkus ja erinevate eluvormide hulk, mõõtude varieeruvus ja üksteise suhtes paiknemise eripära. Puistu keerukust on püütud kokku võtta paljude erinevate indeksitega. Osa neist mõõdikutest kirjeldab eelkõige elurikkust ja metsaökosüsteemi keerukust, teised võrade ruumilist struktuuri. Indeksite iga kasutuse korral tuleks selgelt näidata, kuidas kasutatud indeks arvutati ja mida see näitab.

Ülevaate puistu mitmekesisuse indeksitest annavad McElhinny *et al*. (2005). Lihtne metsa alumise vööndi struktuurse keerukuse indeks võrdub rohurinde katvus pluss põõsarinde katvus pluss mahalangenud tüvede katvus pluss varise katvus – kõiki nelja liidetavat komponenti hinnatakse visuaalselt nelja punkti skaalas (0...3) (Barnett *et al.* 1978). Holdridge (1967) puistu keerukuse indeksis korrutatakse puistu kõrgus, rinnaspind, tüvede arv ja liikide arv. McElhinni *et al*. (2006) valisid metsa struktuuri mitmekesisust kirjeldavasse indeksisse 13 komponenti. Estes *et al.* (2010) arvutasid puistu struktuuri üksikkomponentide peakomponendid ja summeerisid puistu struktuuri indeksisse vaid need peakomponendid, mis kirjeldasid vähemalt 15% varieeruvusest. Selline indeks kirjeldas edukalt mikroelupaikade tihedust.

Eelkõige ruumilist struktuuri kirjeldavad indeksid on **vertikaalse ühetaolisuse indeks** (*vertical evenness*) (Neumann ja Starlinger 2001), mis väljendub valemina

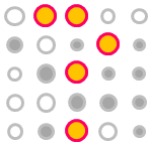
$$VE = \frac{\sum_i p_i\left[-\log\left(p_i\right)\right]}{\log\left(n\right)}, \qquad [4\text{-}108]$$

kus p_i on *i*-nda rinde suhteline katvus ja *n* on puistu rinnete arv.

Struktuurse komplekssuse indeksi (*SCI*) arvutamiseks tuleb esiteks puude asukohtade kahemõõtmeline punktmuster tesseleerida lähimaid naabreid ühendavateks kolmnurkadeks ning teiseks konstrueerida lähimaks naabriks olevate puude ladvapunktidest kolmemõõtmeline tesselatsioon (Zenner ja Hibbs 2000). Struktuurse kompleksussuse indeks võrdub kolmemõõtmeliste kolmnurkade kogupinna ja kahemõõtmeliste kolmnurkade kogupinna suhtega. Vanades metsades on *SCI* suurem ja komplekssuse indeks stabiliseerub alles piisavalt suures (vähemalt 50 × 50 m) proovialas (Zenner 2005).

Metsa struktuurse mitmekesisuse kaardistamisel suuremal alal tuleb kasutada kaugseire andmeid. Coops ja Catling (1997) määrasid metsa struktuuri komponente visuaalselt otse videosalvestiselt. Mitmes uuringus on maastikul mõõdetud struktuuri komponenete seostatud kaugseirekujutiselt arvutatud formaalsete tunnustega (Cosmopoulos ja King 2004, McElhinny *et al*. 2006, Wunderle *et al*. 2007, Pasher ja King 2011).

Viimasel ajal on uusi võimalusi metsa struktuuri ülepinnaliseks kaardistamiske pakkunud LiDAR (*Light Detection And Ranging)* tehnoloogia (Lim *et al.* 2003, Kane *et al*. 2010). LiDAR registreerib välja saadetud ja tagasi peegeldunud valguskiire abil peegeldumispinna kauguse. Metsas peegeldub osa valgust maapinnalt, osa alustaimestikult ja osa võrade eri osadest, mis annab teavet puistu kõrguse, ruumilise struktuuri ja selle varieeruvuse kohta. LiDAR tehnoloogia kasutab valguskiirt mis ei läbi tihedat taimestikku. Seetõttu saadakse tihedate ülemiste rinnete all oleva alustaimestiku ja maapinna kõrguse kohta palju lünklikumad andmed kui hästi nähtavates kohtades.



Küsimused

1. Esita näide, kus punktmustri tüübi tõlgendus sõltub mõõtkavast või kontekstist?

2. Mis näitab naabrite tihedust Ripley K(t) graafikul?

3. Milline punktmustri kirjeldamise meetod väljendab keskmist pindala punktobjekti kohta?

4. Mis on punktmustri kirjeldamise vahemaameetodite puhul kauguse arvutamise lähtepunktiks?

5. Mille poolest erinevad punktmustrite kirjeldamise esimese järgu ja teise järgu meetodid?

6. Mis on lähima naabri keskmise kauguse eelis *k* lähima naabri kauguse ees ja vastupidi, mis on *k* lähima naabri keskmise kauguse eelis ühe lähima naabri kauguse ees?

7. Kas k-NN ja ML meetodi kasutamiseks pilditöötluses on tarvis teada sama andmekihi pikslite omavahelist paiknemist?

8. Mis on pinna segmenteerimise tulemus?

9. Kas kahte liiki punktobjektide paiknemine võib samal ajal olla agregeerunud ja segregeerunud?

10. Milles seisneb lokaaldispersioonide meetod?

11. Mis on Poissoni mets?

12. Miks on vahemaad kasutavad punktmustri kirjeldamise meetodid punktide hõreda paiknemise puhul tõhusamad kui loendimeetodid?

13. Mida näitab tühiku statistik?

14. Millised kaks erinevust on Diggle G(r) jaotusel ja radiaaljaotusel?

15. Nimeta kaks põhjust, miks üldiselt eelistatakse lähima naabri kauguste jaotust kõigi vahemaade jaotusele.

16. Miks on juhusliku punktmustri korral oodatav vahemaa juhuslikust vaatluskohast lähima punktobjektini sama, mis oodatav vahemaa korrapärase võrgustiku vaatluskohast lähima punktobjektini?

17. Kui suur on oodatav naaberobjektide arv raadiuses t, kui Ripley K(t) = 10?

18. Mille tõenäosust näitab radiaaljaotus?

19. Mis erinevus on O-ring statistikul ja radiaaljaotusel?

20. Kas Ripley K(t) funktsiooni eelis naabrite tihedust mõõtvate statistikute ees on tugevam suure või vähese punktobjektide hulga puhul? Miks?

21. Kas Ripley K(t) funktsioon on parameetriline või mitteparameetriline objektide koonduvuse mõõtmise meetod? Põhjenda.

22. Mida näitab segregeerunud paiknemissuhe?

23. Mida kontrollitakse punktmustri statistiliste testidega?

24. Mis on juhusliku toroidnihutuse puudus punktmustrite omavahelise sõltumatuse kontrollimisel võrreldes objektide juhusliku märgistamisega?

25. Mille poolest erineb trendpinna kanooniline analüüs kanoonilisest korrelatsioonanalüüsist.

26. Mis erinevus on väärtuspinna segmenteerimisel ja rasteriseerimisel?

27. Kas toroidnihutust saab kasutada, kui uurimisala on ebakorrapärase kujuga? Kui jah, siis kuidas?

28. Millal võib loota, et hii-ruut test annab kategooriliste pindade omavahelise sõltumatuse määramisel usaldusväärse tulemuse?

29. Millest sõltub ökotoni statistiline olulisus?

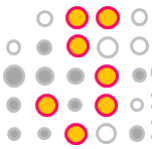
30. Kas Ripley K(t) statistiku ja $L(t) - t$ statistiku väärtused sõltuvad vahemaa mõõtmise ühikutest?

31. Kas naabrite tiheduse arvulised väärtused sõltuvad vahemaa ja pindala mõõtmise ühikutest?

32. Mis on naabrite tiheduse muutumisvahemik?

33. Mis on $L(t) - t$ statistiku võimalike väärtuste vahemik?

5. Ruumilised mudelid



Selles peatükis käsitletakse ruumilistes andmetes olevate statistiliste seoste modelleerimist ja seosemudelite kasutamist hinnangute arvutamisel. Ruumiandmete modelleerimine sisaldab üldjuhul **punktoperatsioone**, mille puhul funktsioontunnus sõltub sama koha argumenttunnuste väärtustest; **naabrusoperatsioone**, mille puhul funktsioontunnus sõltub teatud kohas argumenttunnustest selle koha ümber; ning **globaalseid operatsioone**, mille tulemus sõltub kogu andmestikust või selle suuremast osast.

Ruumikäsitluse aspektist võib prognoose jagada **ruumiliselt ilmutatud** (*spatially explicit*) ja **ruumiliselt ilmutamata** prognoosideks (*spatially implicit*). Ruumiliselt ilmutatud mudelid ja prognoosid käsitlevad nähtuste ruumilist ebaühtlust, ruumiliselt ilmutamata käsitlus ruumilist struktuuri ei ava. Ruumiliselt ilmutamata on näiteks H.T. Odumi ökosüsteemi energeetiline mudel (Odum 1974). Ruumiliselt ilmutamata mudel võib sisaldada kirjeldatava ala tunnuseid globaalsete ehk kogu ala kohta arvutatud karakteristikutena, nagu näiteks elupaiga keskmine kvaliteet, elupaiga osakaal, elupaigalaikude keskmine vahemaa, liigi keskmine asustustihedus, populatsiooni maksimaalne suurus, populatsiooni juurdekasv.

Ruumiliselt ilmutatud käsitlus on olnud domineeriv maastikuökoloogias ja maastikumõõtkavas loodusnähtuste väliuuringutes. Geograafilises mõõtkavas ruumiliselt ilmutatud mudeleid nimetatakse ka **georuumilisteks** (*geospatial models*). Neid võib jagada deduktiivseteks ehk teooriast lähtuvateks ja induktiivseteks ehk andmetest lähtuvateks. Esimesel juhul määrab uurija kasutatavad seosed ja piirangud, teisel juhul leitakse seosed vaatlusandmetest. Liikide leviku hinnangulisel kaardistamisel otsustab uurija esimesel juhul näiteks, et liigile sobivad madalsookaasikud, teisel juhul arvutatakse seos liigi esinemistõenäosuse, mullatüübi ja puistutüübi vahel vaatlusandmetest.

Ruumiline struktuur võib seejuures esile toodud olla:

- punktvaatluste asukohakoordinaatidena või üksteise suhtes paiknemisena,
- pinnalaikude või ka joonstruktuuride suuruse (joonte puhul pikkuse), kuju ja omavahelise paiknemisena,
- korrapärase rasterstruktuurina.

Vormiliselt võib prognoosikaart kujutada:

- nähtuse esinemise või puudumise tõenäosust,
- nähtuse hulka või intensiivsust,
- koha sobivust, mis võib integreerida nii esinemistõenäosust kui ka ohtrust,
- nähtuse kõige tõenäolisemat asukohta,
- kõige tõenäolisemat olemit mingis kohas (dominantliik, maakatteklass, pinnavorm vms.),
- väärtuste (liikide, elupaigatüüpide, asustustiheduse) tõenäosusjaotust kas teatud kohtades või kogu uuritava ala ulatuses,
- sarnasust etteantud olemiga või etteantud näidistega.

Griffith´i ja Layne´i (1999) järgi peaks reaalarvulise muutuja ruumilis-statistiline analüüs sisaldama järgmisi etappe. Protsentide ja loendite puhul tuleks kasutada kas logistilist või Poissoni regressiooni.

- Koosta uuritava tunnuse karp-vurrud diagramm (erindite otsimine ja jaotuse sümmeetria kontroll) ja histogramm (normaaljaotusele vastavuse visuaalne kontroll).
- Arvuta maksimumi ja miinimumi suhe, normaaljaotusele vastavuse Shapiro-Wilki *W* statistik ja mingi dispersiooni homogeensuse kontrollimise statistik (Cochran, Bartlett või Hartley).

- Kui maksimumi ja miinimumi suhe on ≥3, siis proovi, kas astme- või logaritmteisendus parandab normaaljaotusele vastavuse statistikuid. Kui ei, siis kasuta teisendamata andmeid.
- Arvuta Morani *I*, autokorrelatsiooniväli, variogramm või korrelogramm ja hinda ruumilise autokorrelatsiooni mõju.
- Koosta vähimruutude regressioonimudel kasutades erindite eraldamise kriteeriume ja arvuta regressioonimudeli prognoosid ja regressioonijäägid.
- Otsi ruumilise autokorrelatsiooni ilminguid regressioonijääkidest.
- Kui regressioonijääkides on selge ruumiline autokorrelatsioon, siis vali autoregressiooni kaasav mudel (CAR, SAR, krigingu puhul variogrammi mudel) ning määra mudeli parameetrid.
- Arvuta mudeli prognoositud väärtused ja genereeri prognoosikaart.

Ruumiliste andmete statistilise analüüsi ja modelleerimise sagedamini vajaminevad vahendid on vabavaras SAM (Rangel *et al*. 2006, 2010). Enamkasutatud moodulid SAM versioonis 4.0 on järgmised: graafiline andmekirjeldus, Morani *I* ja autokorrelogramm, ruumiline korrelatsioon, regressioon, logistiline regressioon ja partsiaalregressioon, peakomponentanalüüs, autoregressioon (sealhulgas SAR ja CAR), omavektorkaardid, geograafiliselt kaalutud regressioon, Ripley *K*, Mantel test.

5.1. Ruumiline autokorrelatsioon

Autokorrelatsioon kui loodusnähtus on ajas või ruumis lähestikuste vaatluste sarnasus. Autokorrelatsiooni probleemid on oluline osa aegridade analüüsist (ptk 3.5). Kuna eeldame, et aeg on pidev ja muutused toimuvad mingi lõpliku kiirusega ja on lisaks sellele tihti ajalise viivitusega, siis on ajateljel lähestikku paiknevad vaatlused sarnased. Kuna ruumilised nähtused on enamasti samavõrd pidevad kui ajalised, siis on samavõrd põhjust autokorrelatsioonist rääkida ka ruumiliste nähtuste korral. Enamgi veel – aeg on vaid ühemõõtmeline, ruumikäsitlustes on enamasti rohkem dimensioone. Näiteks tavakaart on kahemõõtmeline, maapinna kõrgusmudel kolmemõõtmeline. Ruumilise nähtuse autokorrelatsioon võib olla igas suunas erinev, ei pruugi olla ruumis statsionaarne ega erinevates mõõtkavades ühetaoline.

Autokorrelatsiooni kirjeldamine annab teavet nähtuste ruumilise struktuuri ehk mustri kohta. Ruumilise autokorrelatsiooni olemasolu korral on tunnuse väärtus mingis ruumipunktis prognoositav sama muutuja väärtuste järgi teistes ruumipunktides. Autokorrelatsiooni mõõtmise vahendeid on mitmeid, termin autokorrelatsioon ei ütle, mis viisil on autokorrelatsioon mõõdetud ega anna teada seose positiivset või negatiivset suunda.

Autokorrelatsioonil on ulatus – laag ehk aja või kauguse vahemik, mille piires autokorrelatsiooni kirjeldav indeks on arvutatud. Erinevatel kaugustel on autokorrelatsiooni tugevus erinev. Erinev on tulemus ka sõltuvalt sellest, kas vahet mõõdetakse radiaalulatusena (laagina) või kaugustsoonina (kaugusvahemikuna). Esimesel juhul on tegemist autokorrelatsiooniga vaid ühes, esimeses kaugustsoonis, teisel juhul diferentsiaalse autokorrelatsiooniga. Tunnustevahelise korrelatsiooni selgemaks eristamiseks ühe tunnuse autokorrelatsioonist nimetatakse tunnustevahelist korrelatsiooni ka **ristkorrelatsiooniks** (*cross correlation*) ja vastavat korrelogrammi ristkorrelogrammiks.

Autokorrelatsiooni tugevuse sõltuvust vahemaast saab kirjeldada, nagu igat statistilist seost, vähemalt viiel erineval viisil:

- andmetabelina,
- andmekaardina,
- korrelatsiooniväljana,
- valemina,
- graafikuna.

Ruumilist autokorrelatsiooni võib interpreteerida:

- kaardimustrina,
- nähtuse sisemise, talle omase struktuurina,
- ruumilise laialivalgumisena,
- kaudse teabeallikana otseste mõõtmiste asemel,
- ruumi osadeks jagamise (segmenteerimise) sobivuse mõõduna,
- segava faktorina kuna põhjustab pseudoreplikatiivsust.

Sokal ja Oden (1978b) loetlevad bioloogiliste nähtuste ruumilise autokorrelatsiooni järgmiseid põhjuseid eristades seejuures positiivset ja negatiivset autokorrelatsiooni. Positiivne ruumiline autokorrelatsioon lähikaugustel võib olla põhjustatud keskkonna laigulisusest, migratsioonist, lähestikuste organismide geneetilisest lähedusest, vegetatiivsest paljunemisest. Negatiivne autokorrelatsioon lähikaugustel võib ilmneda väikeste kontrastsete omadustega keskkonnalaikude ja lähedaste omadustega isendite konkurentsi korral. Positiivne autokorrelatsioon suhteliselt suurtel vahemaadel

võib olla seotud keskkonna gradientide või suurte elupaigalaikude korrapärase vaheldumisega. Tugevaim negatiivne autokorrelatsioon suurte vahemaade korral ilmneb, kui kõige erinevamad kohad ja kooslused paiknevad üksteisest eemal. Andmete ruumilise autokorrelatsiooni põhjuste hulgas võivad olla ka vaatlusmetoodikast tingitud hälbed. Näiteks erinevad vaatlejad, erinev vaatlusaeg ja erinev vaatluste intensiivsus vaatlusala eri piirkondades (Dormann 2007a). Vaatlusaladel (vaatlusruutudel) määratud taimkatte katvuse autokorrelatsiooni tugevus sõltub vaatlusruutude suurusest ja kujust (Fortin 1999).

Ruumilise autokorrelatsiooni põhjused on seega:

- nähtuse ruumiline pidevus (nähtuse enda loomus),
- nähtust mõjutavate faktorite ruumiline pidevus,
- nähtuse arenguloo ruumiline pidevus (lähedastel kohtadel on sarnane arengulugu, sest varasemad mõjufaktorid on olnud teatud ruumilise pidevusega) ning
- erinevused andmete kogumisel.

Ökoloogilises kontekstis jagatakse ruumilise autokorrelatsiooni põhjuseid kolme klassi.

- Organismide piiratud levimisvõime.
- Kooslusesisesed suhted konkureerivate liikide, eri troofilistel tasemetel olevate liikide või peremeeste ja parasiitide vahel. Näiteks toiduobjekti ohtruse stabiliseerimine tarbijate (röövloomade) poolt.
- Abiootilise keskkonna ruumiline autokorrelatsioon ja ajalise muutlikkuse ruumiline sünkroonsus. Erinevate keskkonnamõjurite ruumiline, aga ka ajaline autokorreleeritus võivad olla väga erinevad. Näiteks kui Tartumaal on ebaharilik külmalaine, siis on see enamasti vähemalt paari päeva jooksul ka Võrumaal. Taimedele ohtlik rahe on aga palju lokaalsem, kuid vähemalt mõnesaja meetri ulatuses siiski samaaegselt toimuv.

Autokorreleerunud andmete erijuht on **gradiendid** – ruumilised struktuurid, mida saab lihtsa võrrandiga tuletada koha koordinaatidest. Puhas, ilma autokorrelatsioonita gradient peaks olema regressioonivõrrandiga nii kirjeldatav, et regressioonijäägid ruumis ei autokorreleeru. Arvutuslikult võib autokorrelatsioon tekitada näilise gradiendi, kui laag ulatub vaadeldavast alast välja.

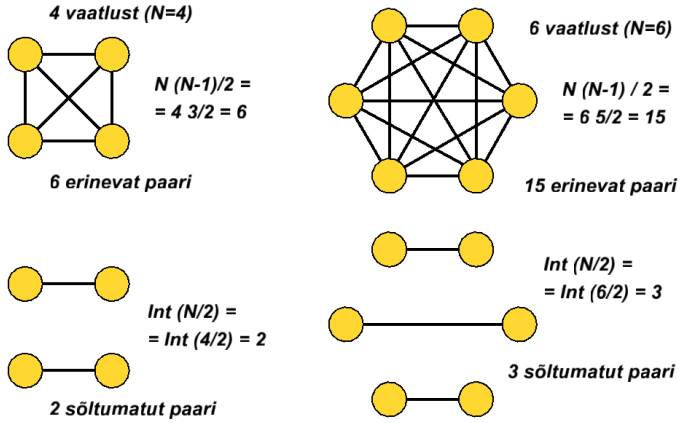
Ruumilise autokorrelatsiooni mõju tähtsustamine on mõjutanud proovide kogumise ja vaatluskohtade valiku metoodikat ökoloogilistes väliuuringutes. Lihtne juhuvalik ei pruugi olla parim lahendus, sest see ei võimalda vältida pseudoreplikatsioone ehk näivkordusi, mille põhjus on nähtuste ruumiline autokorrelatsioon. Ruumilist autokorrelatsiooni tuleb kas spetsiaalselt uurida ja selle mõju prognoosimudelisse lülitada või siis planeerida valik nii, et autokorrelatsiooni ulatusest väiksema vahemaaga paiknevad proovikohad ei satuks samasse valimisse. Proovikoha lähedal paiknevad kohad ei pruugi olla siiski uurimusest elimineeritud, need võivad olla lülitatud kordusvalimitesse. Samuti on hakatud enam tähtsustama ruumiliste nähtuste kaardistamise ja ruumiliselt ilmutatud analüüsimeetodite vajadust.

Positiivset ruumilist autokorrelatsiooni lähivahemaadel on tähistatud ka terminiga **naabrusharmoonia** (*neighbourhood coherence*) (McArdle *et al.* 1997).

5.1.1. Autokorrelatsiooni mõju analüüsi tulemustele

Klassikalised statistilised meetodid eeldavad üksikvaatluste sõltumatust. Ajas või ruumis lähestikku paiknevad vaatlused kipuvad aga olema sarnased ainuüksi seetõttu, et on lähestikku. Lähestikuseid vaatlusi ei tohiks statistilisel analüüsil käsitleda sõltumatutena. Sõltumatute vaatluste korral lisab iga uus vaatlus ühe ühiku teadmisi ja ühe vabadusastme. Sõltuvate vaatluste puhul ei lisandu iga uue vaatlusega üks vabadusaste, vaid mingi osa sellest, mille suurust ei ole võimalik otseselt mõõta.

Autokorrelatsioonikordajate arvutamisel võrreldakse kõigi võimalike vaatluspaaride sarnasust, mida on märkimisväärselt rohkem, kui vaatlusi endid – kokku $N(N-1)/2$ tükki. Juhul, kui vaatlused on omavahel sõltumatud, on maksimaalne üksteisest lahus olevate vaatluspaaride arv vaid $Int(N/2)$ (joonis 5-1). Kui vaatlusi on üks, ei saa vaatluspaare moodustada, suurema vaatluste arvu korral on aga alati $N(N-1)/2 > Int(N/2)$. Seega meetodid, mis kasutavad kõiki võrreldavaid paare üksteisest sõltumatute vaatlustena, ülehindavad vabadusastmete arvu. Peale selle ei pruugi ka üksikvaatlused olla sõltumatud, vaid varjatud kujul kordavad üksteist suuremal või vähemal määral. Lähestikku paiknevad vaatlused võivad iseloomustada sisuliselt sama objekti (populatsiooni, elupaigalaiku) ja neid ei saa seetõttu käsitleda sõltumatute vaatlustena ega planeeritud kordustena – pigem tuleb need näivkordusteks lugeda. Näivkordused on ka ajalise autokorrelatsiooni piiresse jääva ajavahemiku järel tehtud kordusvaatlused.



Joonis 5-1. Paaride arv ja omavahel sidumata paaride arv neljast ja kuuest vaatlusest.

Näivkordused ei ole omavahel sõltumatud vaatlused

Kuidas autokorrelatsiooni erinevatel juhtudel arvestada ja kuidas selle mõju mõõta ehk autokorrelatsiooni modelleerimine on ruumistatistika üks keskseid ülesandeid.

Kui autokorrelatsiooni mitte arvestada, siis hindavad statistilised testid vabadusastmete arvu vääralt kõrgeks ja esimest tüüpi statistilise vea ehk ebaõige sisuka järelduse tõenäosus suureneb (Malanson 1985, Legendre 1993). Mida tugevam on positiivne autokorrelatsioon andmetes, seda suurem on risk ebaolulisi seoseid olulisteks hinnata. Autokorrelatsiooni mittearvestamine mitme faktori mõju võrdlemisel viib tugevamini autokorreleerunud faktorite mõju tõenäolisele ülehindamisele (Lennon 2000). Seda seaduspära on uuritud näiteks kategooriliste pindade ülekatte puhul – laiguliste pindade juhuslikul nihutamisel tekivad suuremad ülekatted sagedamini kui juhusliku mustriga pindade puhul oodatav. On leitud ka, et dispersioonanalüüsi F statistik on andmete ruumilise autokorrelatsiooni korral väiksem kui autokorrelatsiooni puudumisel (Sokal *et al.* 1991, Diniz-Filho ja Malaspina 1995).

Autokorreleerunud andmete kasutamine regressioonimudeli sobitamisel saadakse mudelisse

hälbega parameetrid. Autokorrelatsiooni kaasamine mudelisse muudab mudeli palju keerukamaks. On ka väidetud, et kui eesmärgiks on tulemuste ekstrapoleerimine, siis ei tasu autokorrelatsiooni mudelisse kaasata, sest ei ole selge, kas see kehtib samasugusena väljaspool õpetusandmeid (Guisan ja Thuiller 2005). Kontrollida tasub ka ruumiliste andmete prognoosijääkide (regressioonimudeli korral regressioonijääkide) ruumilist autokorreleerumist kujutades neid kaardil. Kui on tüüpiline, et lähestikku paiknevad jäägid on sarnased (samasuunalised), siis on põhjust arvata, et lähteandmete ruumiline autokorrelatsioon mõjutab hinnanguid. Ruumis autokorreleeruvad jäägid viitavad kas mingi olulise faktori arvestamata jätmisele, milleks võib olla prognoositava nähtuse autoregressiivne loomus ehk sõltuvus iseendast, või mingi ruumiliselt autokorreleerunud faktori mõju ülehindamisele.

Uurimused

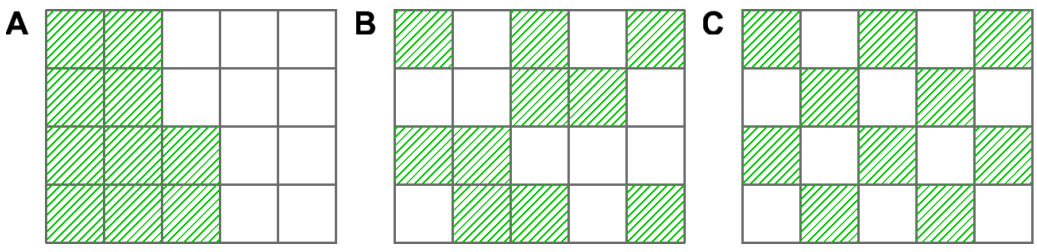
Diniz-Filho *et al*. (2003) väidavad, et ruumiline autokorrelatsioon ei pruugi prognoosihälbeid põhjustada ja suuremaid regressioonikordajaid ruumiliselt autokorreleerunud andmete puhul võib seletada ka nähtuste struktuuri sõltuvusega mõõtkavast või sellega, et mingi laiaulatusliku trendi mõju on tõesti olulisem kui lokaalse mustri mõju.

Muchoney ja Strahler (2002) juhivad tähelepanu, et pikslivääruste autokorrelatsiooni tõttu kalduvad kujutise klassifitseerimise täpsuse hinnangud, mis võrdlevad tulemust ja maapealset tõde pikslikaupa, klassifitseerimistäpsust üle hindama. Betts *et al*. (2006) leidsid, et liikide leviku logistilised mudelid, mis ei arvestanud lindude leviku pidevust, kaldusid elupaiga faktorite mõju suuremaks hindama. Autologistilised mudelid andsid küll parema hinnangute ja vaatluste vastavuse, kuid enamike liikide puhul ei olnud mudelite väljundite erinevus statistiliselt oluline.

5.1.2. Ruumilise autokorrelatsiooni kirjeldamine

Ruumilist autokorrelatsiooni on vaja mõõta selle kui omaette nähtuse kirjeldamiseks, selle lülitamiseks ruumilisi seoseid üldistavatesse mudelitesse ja otsustamaks, kuivõrd tavastatistika testide kasutamine on õigustatud. Ruumiline autokorrelatsioon võib olla negatiivne, positiivne või puududa. Positiivse ruumilise autokorrelatsiooni puhul paiknevad sarnased väärtused ruumis lähestikku (täpsemalt – teatud parajasti käsitletaval kaugusel), negatiivse autokorrelatsiooni korral on etteantud vahemaaga mõõtmistulemuste vaheline erinevus suurem kui andmestikus keskmiselt (joonis 5-2).

Ruumiline autokorrelatsioon on statistiline seos sama tunnuse väärtuste vahel teatud vahemaaga vaatlustel



Joonis 5-2. Ruumiline autokorrelatsioon kaheväärtuselistes andmetes Haining (2003) järgi. A – Positiivne autokorrelatsioon naaberpikslite vahel (samasugused ruudud külgnevad sagedamini kui juhupaiknemisel). B – autokorrelatsiooni puudumine ruutude juhusliku paigutuse korral. C – negatiivne ruumiline autokorrelatsioon (sama tüüpi ruudud on omavahel rohkem eraldatud kui juhusliku paigutuse korral).

Ruumilise autokorrelatsiooni kirjeldamise vahendid võib jagada kolme rühma (Goodchild 1986, Haining 1990):

- globaalsed statistikud,
- lokaalsed statistikud (globaalsed indikaatorid arvutatuna iga koha ümbruse jaoks),
- graafilised kompleksnäitajad (korrelogramm ja variogramm).

ruumilise autokorrelatsiooni kirjeldamiseks leitakse iga mõõtmiskoha (lähtekoha) ja teiste (lähedaste) mõõtmiskohtade (sihtkohtade) vahemaa, mis kuulub mingisse kaugustsooni. Samas kaugustsoonis olevate vaatluspaaride väärtustest arvutatakse mingisugune korrelatsiooni või sarnasuse kordaja ja nii iga kaugustsooni jaoks. Seejuures on iga vaatluspaari üks vaatlus kord lähtekohaks ja kord sihtkohaks. Autokorrelatsiooni tugevuse sõltuvust vahemaast saab graafiliselt esitada korrelogrammil.

Ruumiline autokorrelatsioon on alati seotud vahemaaga

Kõik ruumilise autokorrelatsiooni mõõtmise vahendid eeldavad mingil moel sarnasuse mõõtmist, milleks võib põhimõtteliselt kasutada igasuguseid sarnasuse mõõdupuid. Enamkasutatavad on ühisloend ja Morani *I* (ptk 5.1.2.1 ja 5.1.2.2). Globaalsed statistikud eeldavad ruumilise muutuja teist järku statsionaarsust. See tähendab, et muutuja hajuvus ja oodatav väärtus peaksid olema kogu uuritava ala ulatuses samad. Globaalseteks statistikuteks võivad olla muutuja oodatav keskväärtus ja oodatav dispersioon kogu uuritaval alal. Kõiki globaalseid statistikuid saab arvutada ka lokaalselt, see tähendab ühes kohas või piirkonnas.

Ruumiline autokorrelatsioon on alati seotud vahemaaga vaatluskohtade vahel. Vahemaa on seejuures määratletud piirväärtustega või kasutatakse vahemaast sõltuvaid kaale. Autokorrelatsiooni arvutamiseks tuleb kõigepealt mõõta vaatluskohtade vahemaad, sest iga vahemaaklassi puhul võib autokorrelatsioon olla erineva tugevusega. Mingi etteantud vahemaaga vaatluste autokorrelatsiooni kordaja arvutamiseks võetakse vaid selle vahemaaga vaatluspaarid. Kaalutud keskmine autokorrelatsioon koondab erinevate vahemaade autokorrelatsioone kaugusest (enamasti kauguse pöördväärtusest) sõltuvate kaalude abil. Autokorrelogrammi koostamiseks arvutatakse autokorrelatsioon eraldi kõigis korrelogrammil kujutatavates vahemikes või siis arvestatakse vahemaad muutuva raadiusena.

5.1.2.1. Üldine ristkorrutis-statistik

Ristkorrutis (*cross-product statistic*) on üldine sarnasuste vastavuse mõõtmise viis, mille võtsid kasutusele Knox ja Bartlet (1964) ja mida üldistas N. Mantel (1967). Selle abil hinnatakse, kas ühte tüüpi sarnasusega kaasneb teist tüüpi sarnasus. Ruumilise autokorrelatsiooni hindamise puhul tähistatakse seda

$$r = \sum_i \sum_j W_{ij} Y_{ij}, \qquad [5\text{-}1]$$

kus W_{ij} on kohtade i ja j ruumilise vahemaa mõõt, Y_{ij} on samade kohtade erinevuse või ajalise läheduse mõõt ning $i \neq j$.

Normaaljaotusega muutuja puhul kasutatakse erinevuse mõõduna põhiliselt ruuthälbeid. Vahemaa asemel võib kasutada ka läheduse või ühendatuse mõõtu ning erinevuse asemel sarnasust. Ridade ja veergude arv võrreldavates maatriksites peab olema sama. Enamik tarkvaralahendusi arvutab Mantel korrelatsiooni ehk Mantel statistikut jagades standardiseeritud vahemaade ristkorrutise võrreldud lahtrite (erinevusmaatriksi diagonaali kohal olevate lahtrite) arvuga n – 1. Mantel statistik (*rM*) on vahemikus −1 ja +1.

$$rM = \frac{1}{n-1}\sum_i\sum_j \frac{w_{ij} - \overline{w}}{s_w} \cdot \frac{y_{ij} - \overline{y}}{s_y}, \qquad [5\text{-}2]$$

kus n on võrreldavate vahemaade arv, w_{ij} on väärtus vahemaade maatriksis, y_{ij} on väärtus erinevuste maatriksis, s_w on vahemaade standardhälve ning s_y on erinevuste standardhälve.

Mantel korrelatsiooni arvutamisel vaid ühes kaugustsoonis asendatakse vahemaade maatriksis etteantud kaugustsooni vahemaad ühtedega ja kõik muud vahemaad nullidega või siis piirdutakse mõlemas maatriksis vaid selle kaugustsooni vaatluspaaridega. Nominaalse muutuja puhul on sarnasus Y_{ij} kas 0 või 1 vastavalt sellele, kas vaatlused kuuluvad samasse klassi või teise klassi. Lihtsaima variandi korral arvestatakse nii ruumilist kaugust kui ka sarnasust vaid kaheväärtuselisena – kohad on kas naabrid või ei ning kas kuuluvad samasse klassi või ei. Kaheväärtuseliste muutujate puhul nimetatakse ristkorrutist **ühendusloendiks** (*join-count* või *joint count statistic*). Ühendusloend näitab ühenduste suhtelist sagedust.

K.R. Clarke (1993) võttis kasutusele astakud nii erinevuste kui ka kauguste maatriksis, et muuta Mantel korrelatsioon mitteparameetriliseks.

5.1.2.2. Morani I

Morani *I* (*Moran's I*) (Moran 1950) on kõige sagedamini kasutatav pidevate muutujate ruumilise autokorrelatsiooni mõõt. Morani *I* arvutamisel standardiseeritakse uuritava tunnuse autokovariatsioon sama tunnuse dispersiooniga.

$$I = \frac{N\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i\neq j} w_{ij}\sum_{i=1}^{N}(z_i - \bar{z})^2}, \qquad [5\text{-}3]$$

kus N on etteantud vahemaaga vaatluspaare moodustavate vaatluste arv, i ja j – on koha indeksid, i on lähtekoht ja j on sihtkoht, kusjuures $i \neq j$, w_{ij} on vaatluste omavahelisest paiknemisest sõltuv kaal, z_i on muutuja z väärtus kohas i, z_j on muutuja z väärtus kohas j, $\bar{z}$ on muutuja z keskmine ning s^2 on muutuja z dispersioon. Reeglina arvestatakse $w_{ij} = 1$ kui kohtade i ja j vahemaa on etteantud raadiuses, muidu $w_{ij} = 0$.

Sisuliselt samasugust võrrandit kasutatakse ka Morani *I* väärtuse arvutamiseks kaugustoonis [5-4] ja Morani *I* väärtuse arvutamiseks ruumilise korrelatsiooni mõõduna [5-31]. Viimasel juhul on z_i ja z_j asemel erinevad muutujad (korrelatsioon on kahe muutuja vahel) ning dispersiooni s^2 asemel on ühte tunnuse ja teise tunnuse standardhälbe korrutis.

Morani *I* võrrand erineb Pearsoni lineaarse korrelatsioonikordaja arvutuseeskirjast vahemaast sõltuvate kaalude poolest. Morani *I* on ühikuta suhtarv, mis väljendab vastavas kaugustsoonis olevate vaatluste autokovariatsiooni ja dispersiooni suhet. Kuna vaatluste iseendaga paardumist ei lubata, on

mõõdetud väärtuste juhusliku paiknemise korral Morani *I* väärtused normaaljaotusega ja ootus

$$E(I) = \frac{-1}{n-1}. \qquad [5\text{-}4]$$

Vaatluste suure arvu puhul läheneb *E(I)* nullile. Mitte väga väikese vaatluste hulga puhul näitavad *I* positiivsed väärtused positiivset autokorrelatsiooni, negatiivsed väärtused negatiivset autokorrelatsiooni ja nullväärtus tähendab autokorrelatsiooni puudumist. Positiivne ruumiline autokorrelatsioon tähendab, et vastavas kauguses olevad väärtused kalduvad olema sarnased.

Vaatluste omavahelisest paiknemisest sõltuvate kaalude valiku aspektide üle on arutlenud Jamars *et al.* (1977). Reeglina võrdsustatakse kaal w_{ij} ühega, kui koht *j* on etteantud kaugusvahemikus kohast *i*, muul juhul $w_{ij} = 0$. See etteantud vahemik võib tähendada nii külgnevust kui ka kaugemat vahemikku. Kaalud võib seada proportsionaalseks vahemaa pöördväärtusega või ette anda kasutaja määratud kauguskaalude maatriksina. Külgnevuse määratlemisel on omakorda mitmeid võimalusi.

Kui soovitakse kirjeldada ruumilise korrelatsiooni sõltuvust vahemaast, tuleb korrelatsioonikordaja, mis see ka ei oleks, arvutada iga kaugustsooni jaoks eraldi. Kaugustsoonide kaupa arvutuse äramärkimiseks lisatakse kaugustsooni piires arvutatud statistikute tähistusse sulgudesse tsooni tähis – Morani *I* puhul seega *I(d)* või *I(t)* või *I(r)* või *I(h)*. Ühe kaugustsooni Morani *I* arvutamisel tuleks autokovariatsiooni standardiseerida vaid samas kaugustsoonis olevate vaatluste ruuthälvetega. Sellisel juhul on Morani *I* teatud omavahelise kauguse vahemikku jäävate vaatluste autokovariatsiooni ja samade andmete maksimaalselt võimaliku autokovariatsiooni suhe. Iga kaugustsooni eraldi käsitlemise korral väljaspool kaugustsooni olevad vaatlused Morani *I* väärtust ei mõjuta ja indeksi muutumisvahemik on –1…+1.

$$I(d) = \frac{N(d)\sum_{i=1}^{N}\sum_{j=1}^{N}(z_i - \bar{z})(z_j - \bar{z})}{W(d)\sum_{i=1}^{N}(z_i - \bar{z})^2} \qquad [5\text{-}5]$$

Vahemaaklassi *d* puhul on *N(d)* nende vaatluste hulk, mis moodustavad vaatluspaare vahemaaga *d* ja *W(d)* on nende vaatluspaaride hulk, mille vahemaa kuulub kaugustsooni *d*. Pane tähele, et teatud omavahelise kauguse tsooni jäävate vaatluspaaride hulk ei ole neid paare moodustavate vaatluste arvust tuletatav. Vaatluste arvu ja vaatluspaaride arvu suhe sõltub vaatluste omavahelisest paiknemisest. Näiteks neli vaatlust võib anda null kuni kuus etteantud kaugusvahemikus olevat vaatluspaari.

Kui *I(d)* arvutamisel on nimetajas vaid sama kaugusvahemiku vaatlustest arvutatud ruuthälbed, siis on *I(d)* väärtused vahemikus –1…+1. Kui etteantud vahemaa piiresse jäävat autokovariatsiooni võrrelda kõigist vaatlustest arvutatud dispersiooniga, siis sõltub *I* väärtus kaugemate vaatluste varieeruvusest, seega taustsüsteemist. Kui väärtuste hajuvus on seejuures ruumis ebaühtlane ei ole autokorrelatsiooni muutumispiirid ette teada, teoreetiliselt on need vahemikus $-\infty\ldots+\infty$. Globaalse hajuvuse suhtes arvutatud Morani *I* on <–1 või >1 siis, kui muutuja väärtus lähedastes kohtades varieerub rohkem kui kaugetes kohtades ja kui varieeruvus lähedastes vaatluspaarides on kindla suunaga (vt ka ptk 5.1.2.4 ja joonis 5-4).

5.1.2.3. Geary c

Geary c (Geary 1954) muutumispiirkond on vahemikus 0…+2. Null tähistab kõige tugevamat positiivset ruumilist korrelatsiooni ja 2 näitab maksimaalset negatiivset korrelatsiooni. Väärtus üks viitab ruumilise (auto)korrelatsiooni puudumisele. Erinevalt Morani *I* kordajast kasutab Geary kordaja erinevuse mõõduna mitte kovariatsioone, vaid erinevuste ruute. Seega mõõdab Geary *c* mitte sarnasust, vaid erinevust.

$$c = \frac{(N-1)\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\left(x_i - x_j\right)^2}{2W\sum_{i=1}^{N}(x_i - \bar{x})^2} \qquad [5\text{-}6]$$

Kahega jagamine nimetajas tuleneb vaatluspaaride loomulikust kahekordsest olekust: kord on tinglikul esimesel positsioonil üks paariline, teinekord teine paariline. Geary *c* on vähem kasutatav kui Morani *I* eelkõige selle tõttu, et selle muutumispiirkond ei vasta tavapärasele; aga ka selle tõttu, et see alahindab nende vaatluste osa, millel on vähem naabreid. Vähem naabreid on reeglina näiteks uuritava ala serva lähedal olevatel vaatlustel. Geary *c* on ka tundlikum muutuja jaotuse suhtes kui Morani *I* (Cliff ja Ord 1981).

5.1.2.4. Lokaalne autokorrelatsioon

Eeltoodud autokorrelatsiooni mõõtmise vahendid annavad ruumilise autokorrelatsiooni globaalse (üle kogu andmestiku ulatuva) hinnangu. See tähendab, et üks arv või graafik kirjeldab autokorrelatsiooni erinevate vahemaade korral kogu uuritaval alal. Kui uuritav ala on suur ja heterogeenne, siis ei ole ruumilise statsionaarsuse eeldamine õigustatud. Lisaks muutuja enda ebaühtlasele varieeruvusele võib varieeruda ka selle muutuja ruumiline autokorrelatsioon. Kui punktmustri puhul sündmuste tiheduse ruumiline autokorrelatsioon kirjeldab punktide paiknemismustrit ennast, siis arvuliste ruumiliste muutujate korral ei kirjelda ruumiline autokorrelatsioon mitte otseselt muutuja väärtuste pinda, vaid selle ruumilist statsionaarsust.

Lokaalse ruumilise koondumise mõõtmisel võib olla kolm eesmärki:

- uuritava nähtuse ruumilise statsionaarsuse hindamine,
- ühetaoliste väärtustega laikude esiletoomine,
- ühetaoliste väärtuste laikude suuruse määramine.

Kõiki globaalseid näitajaid saab arvutada ka lokaalselt ehk iga koha ümbruses. Lokaalse autokorrelatsiooni arvutamisel on kaks varianti:

- lokaalseks loetakse vaid lähtepunkt ja iga lähtepunkti puhul kasutatakse vaid selle koha ümber vastavas kaugustsoonis olevaid naabreid,
- lokaalset autokorrelatsiooni arvutatakse liikuva akna põhimõttel.

Viimasel juhul kasutatakse nii lähte- kui ka sihtpunktidena kõiki akna ulatuses olevaid vaatlusi ja mitte kaugemaid. Väärtuste keskmine ja dispersioon, mille suhtes ruumilist varieeruvust mõõdetakse, võib olla arvutatud kas globaalselt või vaid lokaalse akna sees olevatest andmetest – tulemus on erinev.

Võimalik on mõõta iga üksiku vaatluse mõju globaalse statistiku väärtusele ja niimoodi otsida eriti suure mõjuga üksikvaatlusi või ebaolulise mõjuga vaatlusi edasisest analüüsist välja jätta. Regressioonanalüüsil tuntakse seda *hat and leverage* meetodina.

Ruumilise koondumise lokaalsetele indikaatoritele (*local indicators of spatial association – LISA)* (Deichman ja Anselin 1994, Anselin 1995) on statistikud, mis näitavad sarnaste väärtuste ruumilist koondumist ja mille väärtuste summa on proportsionaalne mõne ruumilise koondumise globaalse näitajaga. LISA analoog kategooriliste andmete jaoks on LICD (*local indicators for categorical data)* (Boots 2003, 2006). LICD hõlmab lokaalselt arvutatud maastikumeetrika indekseid.

Lühend LISA tähistab kõiki ruumilise koondumise lokaalseid näitajaid

Lokaalne autokorrelatsioon võib vaatlusala ühes piirkonnas olla positiivne ja mõnes teises osas negatiivne. Sokal *et al.* (1998) nimetavad positiivse autokorrelatsiooni kohti **kuumadeks laikudeks** (*hot spots*) ja negatiivse autokorrelatsiooni alasid **külmadeks laikudeks** (*cold spots*). Liigi ohtruse kaardistamisel näitab kuum laik ühetaolisi elupaigatingimusi ja levimisvõimet, külm laik viitab kas ökotonile või leviku tõkkele.

Lokaalset autokorrelatsiooni on püütud käsitleda ka omaette faktorina, mis lisandub globaalsele autokorrelatsioonile ja ruumilisele trendile (Boots 2003).

Gi ja Gi* statistikud

Spetsiaalselt numbriliste väärtuste lokaalse koondumise näitajad on Getis-Ordi G_i ja G_i* statistikud (Getis ja Ord 1992, 1996). G_i ja G_i* arvutatakse väärtustest koha (*i*) ümbruses ning keskmisest kogu uuritaval alal.

G indeks on koha ümbruses olevate väärtuste summa osa uuritava ala kõigi väärtuste summast

$$G_i(d) = \frac{\sum_j w_{ij}(d) x_j}{\sum_j x_j}, j \neq i, \qquad [5\text{-}7]$$

$$G_i(d)^* = \frac{\sum_j w_{ij}(d) x_j}{\sum_j x_j}, \qquad [5\text{-}8]$$

kus x_j on väärtus kohas *j*, $w_{ij}(d)$ on vaatluse *j* vaatluse *i* naabrusesse *d* kuulumist kirjeldav kaal. Kui koht *j* on etteantud kauguse sees, siis $w_{ij} = 1$, muul juhul $w_{ij} = 0$. Statistiku tärnita ja tärniga variandi erinevus seisneb kohas *i* oleva vaatluse vastavalt kas välja jätmises või arvesse võtmises keskmise arvutamisel.

G statistiku oodatav väärtus *E(G)* on vahemikus 0...1 ja avaldub järgmiselt:

$$E(G_i) = \frac{\sum_j W_{ij}(d)}{n-1}, \qquad [5\text{-}9]$$

$$E(G_i^*) = \frac{\sum_j w_{ij}(d)}{n}. \qquad [5\text{-}10]$$

G_i on loodud lokaalse statistikuna toomaks esile positiivse ruumilise autokorrelatsiooni lokaalseid taskuid, nagu Getis ja Ord (1992) kõrgete väärtustega piirkondi nimetasid. *G* statistiku väärtus on kõrge kohtades, mille ümber on ebaproportsionaalselt palju uuritava muutuja kõrgeid väärtusi.

Autokorrelatsiooni puudumise ja muutuja normaaljaotuse korral võrdub *G** oodatava väärtusega *E(G*)*. Oodatavast erinevuse olulisust saab kontrollida *Z* statistiku järgi normeerides erinevust standardhälbega.

Getis-Ord statistiku originaalkuju eeldab, et uuritava muutuja väärtused on positiivsed ja normaaljaotusega. Vabanemaks sõltuvusest muutuja väärtuste jaotusest, standardiseerisid Ord ja Getis (1995) arvutusvalemid järgnevale kujule:

$$G_i(d) = \frac{\left(\sum_j w_{ij}(d)x_j - W_i\bar{x}(i)\right)\sqrt{n-2}}{s(i)\sqrt{\left((n-1)S_{1i}\right) - W_i^2}}, \text{ kui } j \neq i \qquad [5\text{-}11]$$

ja kõigi vaatluste kasutamisel kujule

$$G_i^*(d) = \frac{\left(\sum_j w_{ij}(d)x_j - \bar{x}W_i^*\right)\sqrt{n-1}}{s\sqrt{nS_{1i}^* - W_i^{*2}}}, \qquad [5\text{-}12]$$

kus w_{ij} on reaalarvulised kauguskaalud, $W_i = \sum_j w_{ij}(d)$ ehk kaalude summa ilma vaatluseta *i*, $W_i^* = W_i + w_{ii}$ ehk kaalude summa koos vaatlusega *i*, $S_{1i} = \sum_j w_{ij}^2$ ehk kaalude ruutude summa ilma vaatluseta *i*, $S_{1i}^* = \sum_j w_{ij}^2$ ehk kaalude ruutude summa koos vaatlusega *i*, *s(i)* on koha *i* ümbruses olevate vaatluste standardhälve, *s* on kõigi vaatluste standardhälve, $\bar{x}$ on valimi keskväärtus.

Täiustatud *G* statistiku jaotus läheneb raadiuse suurenedes normaaljaotusele ka siis, kui muutuja jaotus on asümmeetriline. Seega on *G* statistik sisuliselt *Z* skoor, mis võimaldab hinnata statistiku olulisust normaaljaotuse tõenäosustena.

Leung *et al.* (2003) soovitasid *G* statistikuid lihtsustada jättes ära standardiseerimise ja eelistades ruutu võtmist ruutjuurele. Kui statistik arvutatakse vaid ühes raadiuses, siis ei ole tarvidust märkida valemisse muutujat *d*.

$$\tilde{G}_i = G_i^2 = \left\{\sum_{j\neq i} w_{ij}\left[x_j - \bar{x}(i)\right]\right\}^2 \Big/ \frac{1}{n-1}\sum_{j\neq i}\left[x_j - \bar{x}(i)\right]^2 \qquad [5\text{-}13]$$

$$\tilde{G}_i^* = G_i^{*2} = \left\{\sum_j w_{ij}\left[x_j - \bar{x}\right]\right\}^2 \Big/ \frac{1}{n}\sum_j\left[x_j - \bar{x}\right]^2 \qquad [5\text{-}14]$$

Lokaalne Morani I

Lokaalse Morani *I* arvutamisel on kaks põhimõtteliselt erinevat varianti – kas arvutada see valemi 5-5 järgi kaardistatavale alale paigutatud võrgustiku sõlmedes (joonis 5-3) või arvutada eraldi iga vaatluskoha *i* ümber ja loobuda summeerimisest üle lähtekohtade (valemid 5-15 ja 5-16). Lokaalse Morani *I* väärtus ei pruugi olla vahemikus −1...+1, kui selle arvutamisel kasutatakse globaalset dispersiooni, nagu alltoodud valemites. Muutliku varieeruvuse (heteroskedastilisuse) korral ei ole nende valemite järgi arvutatud lokaalse Morani *I* muutumispiirid ette teada, teoreetiliselt on need vahemikus −∞…+∞. Lokaalne Morani *I* on <–1 või >1 siis, kui muutuja väärtus lähedastes kohtades varieerub rohkem kui kaugetes kohtades ja kui varieeruvus lähedastes vaatluspaarides on kindla suunaga.

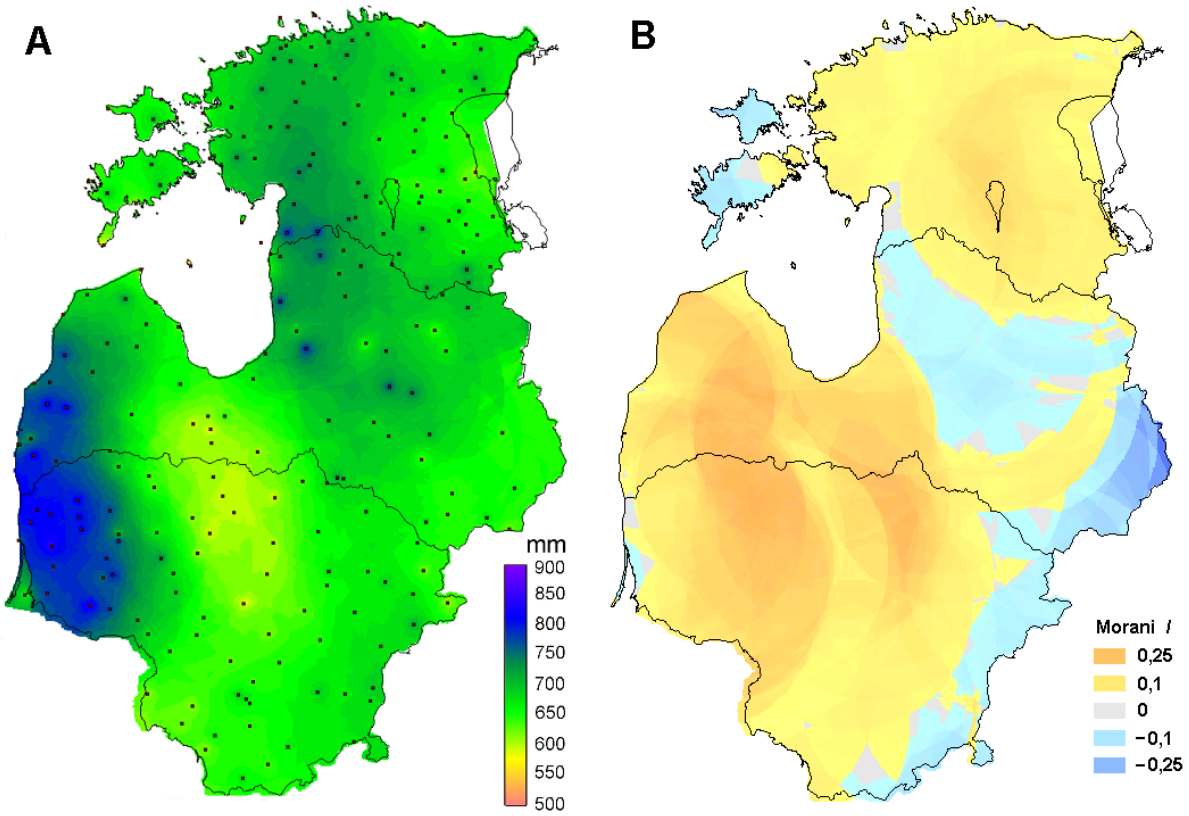
Lokaalse *I* lähedane statistik on lokaalne *c* (valem 5-16) (Anselin 1995, Getis ja Ord 1996).

Lokaalse Morani I väärtused on vahemikus –1…+1 vaid homoskedastilistes andmetes

$$I_i = \frac{(z_i - \bar{z})\sum_{j=1}^{N} w_{ij}(z_j - \bar{z})}{\sigma^2}, \qquad [5\text{-}15]$$

$$c_i = \frac{\sum_{j=1}^{N} w_{ij}(z_i - z_j)^2}{\sigma^2}, \qquad [5\text{-}16]$$

kus z_i ja z_j on väärtused kohas *i* ja *j*, σ^2 on kõigi *z* väärtuste standardhälve, $\bar{z}$ on valimi keskväärtus.

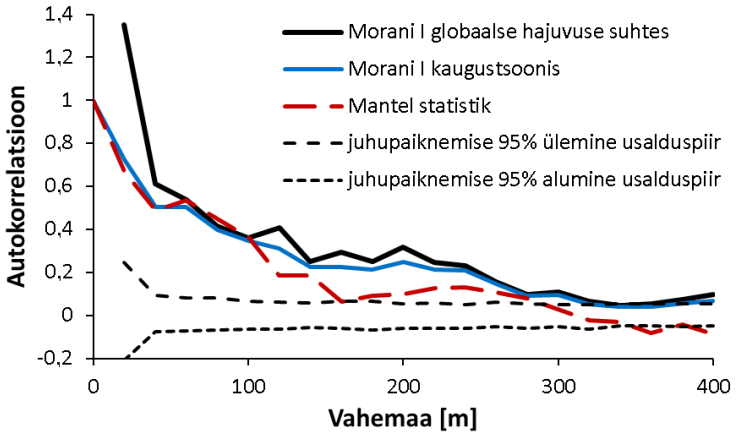


Joonis 5-3. Baltimaade keskmine aastane sademete hulk ilmavaatlusjaamade vahele interpoleerituna (A) ja vaatlusjaamades mõõdetud paljuaastase keskmise sademete hulga vaatlusjaamast kauguse pöördväärtusega kaalutud ruumiline autokorrelatsioon (B). Ühe kilomeetri sammuga kohtades lokaalselt arvutatud autokorrelatsioon on positiivne eelkõige seal, kus ühtlaselt kõrge sademete hulgaga ala (näiteks Lääne-Leedu) lähedal paikneb ühtlaselt kuivem piirkond (näiteks Kesk-Leedu ja Kesk-Läti). Autokorrelatsioon on negatiivne eelkõige ida- ja põhjapoolses Lätis, Leedu idapiiril ning Saare- ja Hiiumaa lääneosas, kus keskmine sademete hulk on ruumiliselt ebastabiilne – see tähendab, et erinevused lähestikku paiknevate vaatlusjaamade andmetes on suhteliselt suured. Andmed Remm *et al.* (2011).

5.1.2.5. Korrelogramm

Korrelogramm on graafik, mis näitab mõõtmistulemuste omavahelise korreleerumise sõltuvust mõõtmiskohtade vahelisest kaugusest (*h*). Kui korrelogramm leitakse sama muutuja väärtustest erinevates kohtades, iseloomustab see ruumilist autokorrelatsiooni ja nimetatakse autokorrelogrammiks (joonis 5-4); kahe muutuja vahelise seose sõltuvust vahemaast näitab ruumiline korre-

latsioon. Ruumilise korrelatsiooni mõõtmiseks saab kasutada lineaarse korrelatsiooni kordajat ning mitmesuguseid teisi kordajaid, autokorrelatsiooni mõõtmiseks kasutatakse eelpool mainitud autokorrelatsiooni kordajaid. Mantel statistiku [5-2] kasutamisel nimetatakse korrelogrammi Mantel korrelogrammiks (Oden 1984, Koenig ja Knops 1998, Bjørnstad *et al*. 1999b).



Joonis 5-4. Autokorrelogramm põõsasmarana katvuse andmetest kaardilehelt 6382 globaalse hajuvusega ja kaugustsoonis olevate vaatluste hajuvusega normeeritud Morani *I* ning Mantel statistiku järgi ja ka Morani *I* oodatava väärtuse 95% usaldusvahemik vaatluste juhuslike ümberpaigutamiste korral. Kuna põõsasmarana katvuse varieeruvus ei ole ruumiliselt ühetaoline, siis on kuni 20 m vahega vaatluste puhul Morani *I* väärtus > 1. Vaatlusandmed K. Remm 2008-2011.

Ruumilise trendi esinemisel on autokorrelatsioon tugevam väiksematel kaugustel ja läheneb nullile suuremate kauguste juures. Ruumilise regulaarsuse korral korrelogramm lainetab. Laigulisuse korral on autokorrelatsioon positiivne laikude keskmisele läbimõõdule vastava kauguseni. Korrelogrammid aitavad ruumilistest andmetest korrapäraseid struktuure leida. Korrelogrammi kaugusklasside arvu ja laiuse valik sõltub andmete hulgast ja kirjeldatava ruumilise korrapära omadustest (nt. laikude suurus ja laikude vahemaa). Liiga laiade kaugusklassidega korrelogramm ei esita iseloomulikke tendentse. Kitsastesse kaugusklassidesse jääb vähe vaatlusi ja nendest arvutatud korrelatsioon ei ole usaldatav. Legendre ja Fortin (1989) ei soovita kasutada nende kaugusklasside andmeid, kus on alla 1% vaatlustest.

Korrelogramm võib olla suunda mitte arvestav ehk **igasuunaline** (*omnidirectional*) korrelogramm. Kui uuritava nähtuse ruumiline struktuur on **isotroopne** (see tähendab ei sõltu suunast), on igasuunalise korrelogrammi kasutamine õigustatud. Mõned nähtused on aga kindlasuunalise struktuuriga ehk **anisotroopsed**. Sel juhul on (auto)korrelatsioon erinevates ruumisuundades erinev ja kasutatakse suunaga korrelogramme. Suunaga korrelogrammi puhul arvutatakse korrelatsioonid eri suunavahemike jaoks eraldi.

Korrelogrammi saab kujutada ka variogrammi kujul, see on $1 - r(h)$ ja h vahelise seose graafikuna. Variogrammi kujul esitatud korrelogramm ei ole sama, mis variogramm. Korrelatsioon muutub vahemikus –1…+1, variogrammil kujutatav poolhajuvus aga vahemikus 0…*lävend* (teoreetiliselt kuni +∞). Variogrammi kujul korrelogramm peaks suuremate laagide korral stabiliseeruma tasemele 1.

Standardiseeritud (õigemini tsentreeritud) korrelogrammi puhul lahutatakse kaugustsoonide korrelatsioonikordajatest üldkeskmine korrelatsioon. Ruumilise autokorrelatsiooni puhul lahutatakse vahemaatsoonides arvutatud autokorrelatsioonidest keskmine autokorrelatsioon kõigis vaatluspaarides, sõltumata vaatluste vahemaast. Kui eeldada, et ka tsentreeritud korrelatsioonid peaksid muutuma vahemikus –1…+1, siis tuleks eelistada järgmist teisendust:

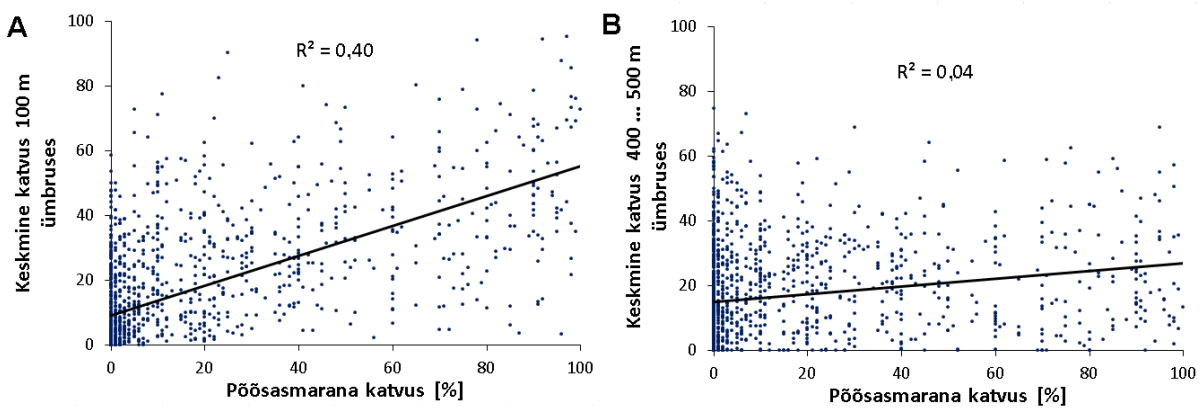
$$R_{sd} = \frac{R - \overline{R}}{1 - \overline{R}}, \text{ kui } R > \overline{R}, \qquad [5\text{-}17]$$

$$R_{sd} = (R - \overline{R}) \times \frac{1}{1 + \overline{R}}, \text{ kui } R < \overline{R}, \qquad [5\text{-}18]$$

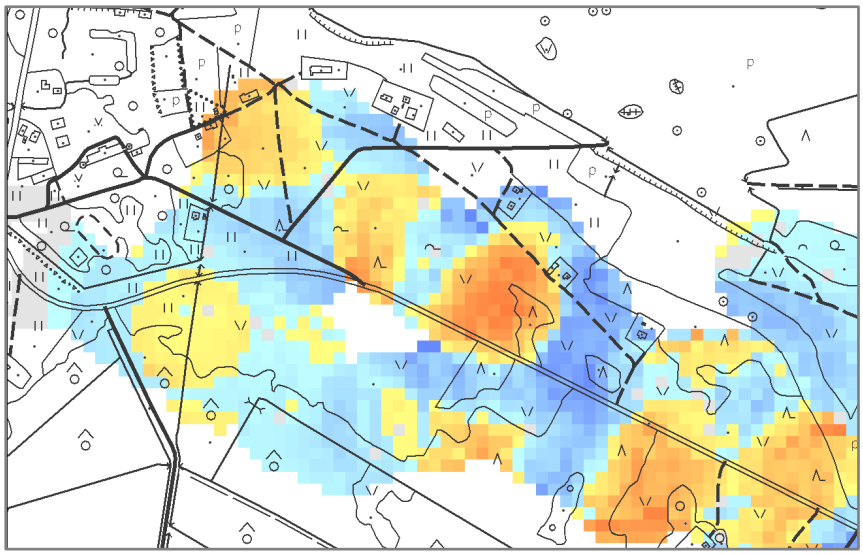
kus R on standardiseerimata autokorrelatsioon kaugustsoonile vastava vahemikuga vaatluspaarides, R_{sd} on standardiseeritud autokorrelatsioon kaugustsoonile vastava vahemikuga vaatluspaarides,$\overline{R}$ on autokorrelatsioon kõigis vaatluspaarides.

5.1.2.6. Autokorrelatsiooni jaotusväli

Autokorrelatsiooni jaotusväli ehk Morani jaotusväli (*Moran scatterplot*) (Anselin 1995) kujutab naabervaatluste sarnasust iga vaatluskoha ümber. Morani jaotusväljal on kaks kuju. Esiteks graafik, mille horisontaalteljel on uuritava muutuja väärtus vaatluskohas ja vertikaalteljel muutuja keskmine väärtus sama vaatluskoha ümbruses etteantud kaugusvahemikus (joonis 5-5A,B). Klassikalises variandis on muutuja väärtused standardiseeritud. Graafiku regressioonijoone kaldenurk näitab autokorrelatsiooni tugevust (joonis 5-5). Teiseks kaart, mis kujutab muutuja suhtelist väärtust ümbruse suhtes. Morani jaotusväli toob esile naabrusest erinevad piirkonnad (joonis 5-6).



Joonis 5-5. Morani jaotusväli põõsasmarana katvuse andmetest kuni 100 m vahemaa korral (A) ja 400...500 m vahemaa korral (B). Väiksema vahemaa juures on seos tugevam.



Joonis 5-6. Põõsasmarana suhteline katvus 100–200 m ümbruse keskmise suhtes Vääna mõisapargist idapool. kollakates ja punakates kohtades on põõsasmarana katvus 0–50 m raadiuses suurem kui 100 m raadiusega ümbruses, sinakates kohtades väiksem. Valgel alal välivaatlused puuduvad. Vektortaust – Eesti põhikaart, Maaamet 2005.

5.1.2.7. Osaautokorrelatsioon

Korrelogrammi saab arvutada nii kaugustsoonide kaupa kui ka korrelatsioonina vaatluste vahel, mis on lähemal kui teatud kaugus (Deichmann ja Anselin 1994). Kaugustsoonide kaupa arvutamisel nimetatakse iga kaugustsooni korrelatsiooni partsiaal- ehk **osaautokorrelatsiooniks** (*partial autocorrelation*).

Kui ruumimuster sisaldab sarnaseid väärtusi või struktuure, mis korduvad vahemaaga *h*, siis väljenduvad need partsiaalkorrelogrammis mitte ainult kaugusel *h*, vaid ka *h* kordsetel kaugustel. Seega võib korrelogramm sisaldada samade struktuuride korduvat kujutamist ja kaugemates tsoonides olevad korrelogrammi tipud ei pruugi viidata iseseisvatele struktuuridele. Aegridade analüüsi jaoks on sarnastel alustel välja töötatud meetodid, mis võimaldavad tuvastada ja eemaldavad mustri perioodilisuse mõju.

5.1.2.8. Omaväärtustele tuginevad meetodid

Peamine ökoloogiliste andmetes oleva ruumilise autokorrelatsiooniga seonduv probleem on, kuidas eristada bioloogiliste muutujate varieeruvuse asukohast sõltuvat ja keskkonnateguritest sõltuvat komponenti. Borcard *et al*. (1992) kasutasid kanoonilist vastavusanalüüsi (ptk 2.4.6) lähtudes eeldusest, et liikide varieeruvust määrab:

- mõõdetud keskkonnatunnuste kompleks,
- asukoht ja selle ümbrus,
- määratlemata tegurid, mille kohta puuduvad andmed ning keskonnatunnuste varieeruvus on teatud osas seotud asukohaga.

Kanooniline vastavusanalüüs tehti neljas variandis samadest andmetest: liikide andmed seotult keskkonnatunnustega ja eemaldatud keskkonnamõjudega ning liikide andmed asukohakoordinaatidega ja asukoha mõju eemaldamise järel. Nelja analüüsi tulemusel saadi liikide keskkonnatunnustega seotud varieeruvuse, asukohaga seotud varieeruvuse, nii keskkonnatunnuste kui ka asukohaga seotud varieeruvuse ja seletamata varieeruvuse osa hinnang.

Naabrusmaatriksi peakoordinaadid

Naabrusmaatriksi peakoordinaadid (*principal coordinates of neighbour matrices – PCNM*) on arendatud alternatiiviks trendpinna analüüsile ja võimaldavad eristada vaatluste paiknemise mõju. PCNM arvutuskäik on järgmine (Borcard ja Legendre 2002).

- Moodusta kõigi vaatluskohtade vaheliste kauguste maatriks D.
- Vali piirkaugus *t* ja omista kõigile sellest suurematele kaugustele väärtus 4*t*.
- Vii läbi kärbitud väärtustega kaugusmaatriksi peakoordinaatanalüüs (PCoA). Kauguste kärpimise tõttu on osa omaväärtusi negatiivsed. Need vastavad negatiivsele ruumilisele autokorrelatsioonile.
- Kasuta positiivse omaväärtusega peakoordinaate regressioonimudelis või kanoonilises analüüsis liigi või muu nähtuse esinemise/puudumise või ohtruse prognoosimisel.

PCNM peakoordinaadid on ortogonaalsed ja kirjeldavad ruumilist seost vaatluskohtade vahel või uuritava ala ruumilist struktuuri. Neid saab arvutada iga prognoositava koha kohta. Kui vaatluskohad paiknevad korrapäraselt, on peakoordinaadid sinusoidsed.

Borcard *et al*. (2004) soovitasid enne PCNM analüüsi kontrollida funktsioontunnuse lineaarset trendi, mis viitab uuritavast alast suurema sammuga ruumimustrile. Trendi olemasolul tuleks see andmete teisendamisega eemaldada. PCNM suudab küll trendi kirjeldada, kuid trendi kaasamine hägustab peenemate struktuuride ilmnemist ja omavektorite tähendust.

Morani omavektorkaardid

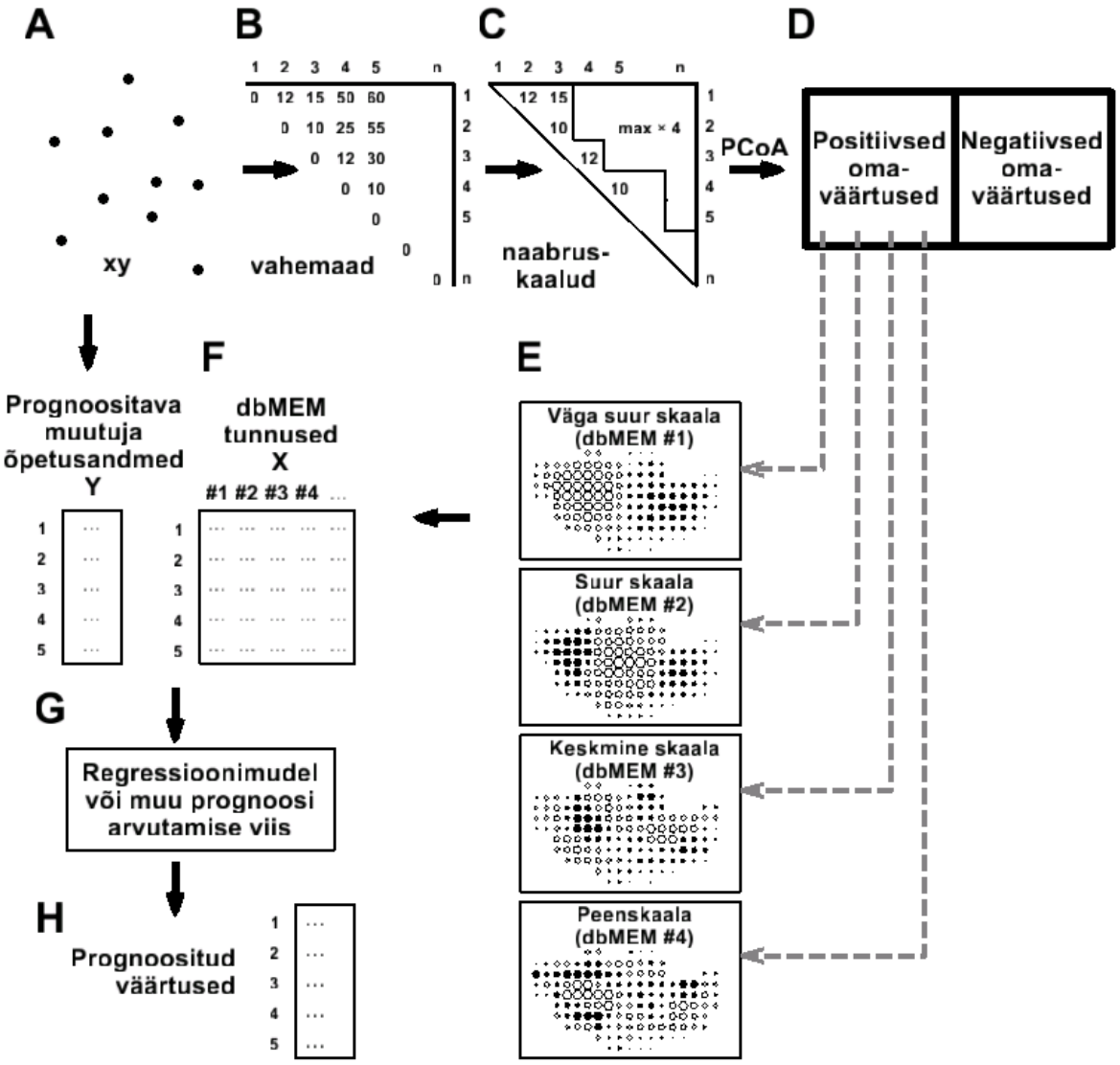
Morani omavektorkaardid (*Moran's eigenvector maps – MEM* ehk *distance-based eigenvector maps – DBEM* ehk *distance-based Moran eigenvector maps – dbMEM*) (Dray *et al*. 2006) on PCNM peakoordinaatide üldistus. Mõlemad kasutavad külgnevuse maatrikseid loomaks teisendatud ortogonaalseid tunnuseid. MEM on paindlikum vaatluskohtade paiknemise ja külgnemise erinevate variantide arvestamisel. Vaatluskohad ei pea paiknema korrapäraselt, juhuslikult paiknev vaatlusvõrk sobib sama hästi (Munoz 2009). Borcard *et al.* (2011) järgi võib PCNM ja MEM sünonüümideks lugeda, MEM on sisuliselt sama meetodi uuem termin.

MEM on variogrammi mudelile tuginevate geostatistika meetodite mitteparameetriline alternatiiv, mis ei eelda ega kasuta ette määratletud ruumilise varieeruvuse mudelit

Omavektorkaartide eesmärk on moodustada teisendatud ortogonaalsed tunnused, mis võtavad arvesse vaatluskohtade paiknemist ja vähendavad lähivaatluste kaalu. Ruumilised omavektorid arvutatakse vaid vaatluskohtade omavahelise paiknemise näitajatest (koordinaatidest või ühendatusest). Ruumilised omavektorid rühmitatakse erinevatesse mõõtkavavahemikesse ja kasutatakse igast mõõtkavavahemikust ühte omavektorit (Lacey *et al*. 2007, Ali *et al*. 2010). Omavektori mõõtkava näeb, kui paigutada omavektori väärtused kaardile – saadakse omavektori kaart. Omavektorite abil teisendatud tunnused arvutatakse igas vaatluskohas ja prognoositavas kohas ning neid kasutatakse kanooniliste seoste leidmisel või prognoosimudelis (joonis 5-7).

Griffith ja Peres-Neto (2006) ühendasid PCNM, MEM ja ruumilise filtreerimise meetodid omaväärtusfunktsioonile tuginevaks analüüsiks (*eigenfunction-based spatial analysis*). Blanchet *et al.* (2008) laiendasid Morani omavektorkaartide kasutust anisotroopsetele andmetele asümmeetriliste omavektorkaartidena (*asymmetric eigenvector maps – AEM*). Jombart *et al.* (2009) rakendasid Morani omavektorkaarte ökoloogilise muutlikkuse erinevas mõõtkavas struktuure arvestavas ordinatsioonis ja selle graafilises esituses, mida nad nimetasid mitmemõõtkavaliseks mustrianalüüsiks (*multi-scale pattern analysis – MSPA*). MSPA kanooniline vorm võimaldab samas teljestikus kujutada nii liikide kui ka keskkonna ordinatsiooni.

Guénard *et al.* (2010) täiendasid asukohaga seotud omavektorkaartide meetodeid mitmemõõtmelise koosmõjuanalüüsiga (*multiscale codependence analysis – MCA*), milles korrelatsioonid funktsioon- ja argumenttunnuste vahel seotakse ruumilist (või ajalist) struktuuri kirjaldavate struktuur-funktsioonidega (siinus või kosiinus, ruumilised peakomponendid, lainekesed, asend regulaarnses vaatlusvõrgustikus). Mitmemõõtkavalise koosmõjuanalüüsi algoritmi kohaselt arvutatakse ruumistruktuuriga seotud koosmõjukordajad (*codependence coefficients*), leitakse suurima absoluutväärtusega kordaja ja sellele vastav struktuur-funktsioon. Arvutatakse struktureeriva funktsiooni mõju olulisus. Kui see on piisav, siis lisatakse see funktsioon struktureerivasse komplekti ja korratakse analüüsi ülejäänud funktsioonidega.



Joonis 5-7. Morani omavektorkaartide tehnoloogiline skeem (Borcard ja Legendre 2002, Lacey *et al.* 2007, Ali *et al.* 2010 järgi muudetult). A – vaatluskohtade ristkoordinaadid, B – vaatluskohtade vahemaade maatriks, C – kärbitud kauguskaalude maatriks ehk ühendusmaatriks, milles on kõigile ühenduses või naabruses mitteolevatele vaatluskohapaaridele omistatud neljakordne minimaalne arvesseminev vahemaa, D – tsentreerimine ja diagonaliseerimine peakoordinaatanalüüsi käigus, mille tulemusel saadakse üksteise suhtes ortogonaalsed omavektorid, mis esindavad erinevas mõõtkavas struktuure, E – erinevas mõõtkavas struktuure esindavad omavektorkaardid (ringi suurus näitab väärtust, seest tühjad ringid esindavad negatiivseid väärtusi), F – Morani omavektortunnused (dbMEM), G – funktsioontunnust prognoosiv (regressiooni)mudel dbMEM tunnustele sobitatud parameetritega, H – prognoositud väärtused.

5.1.2.9. Kategoorilise pinna ruumiline autokorrelatsioon

Ruumilise autokorrelatsiooni näitajatega saab muuhulgas kirjeldada kategoorilise pinna laigulisust. On näidatud, et korrelogrammide maksimumid osutavad keskmisele laigukeskmete vahelisele kaugusele ning et korrelogrammi amplituud on suurim, kui laikudevahelised kaugused on sama suured kui laikude läbimõõt (Radeloff *et al.* 2000). Seejuures võivad erinevad laigulisuse mustrid anda siiski samasuguseid korrelogramme.

Uurimused

Fortin *et al*. (1989) näitasid, et ruumilise autokorrelatsiooni korralikuks kirjeldamiseks peavad vaatluskohtade paiknemises olema esindatud erinevad vahemaad. Lihtsa korrapärase paigutuse puhul ei ole vahemaade esindatus esinduslik.

Brown *et al*. (1995) leidsid, et lindude kogunemine parvedesse sõltub elupaiga sobivusest, maastiku eripärast ja koha asendist levilas. Levila keskel on parvlemise tendents reeglina tugevam kui levila äärtel.

Getis ja Ord (1992), Ord ja Getis (1995) demonstreerisid G statistiku kasutamist epidemioloogiliste andmete najal. Ghimire *et al.* (2010) arvutasid Getise G* statistiku Landsat kujutise iga kanali iga piksli kohta kolmes erinevas raadiuses. Parima klassifikatsioonitäpsuse andis G* statistiku andmekiht, mis oli arvutatud 11 × 11 piksli suuruses aknas. Long *et al.* (2010) näitasid, et LICD sõltub tugevasti arvessemineva naabruse ulatusest, sest maastikumuster ei avaldu väikeses aknas.

Long (1998) tõdes, et nisu saagikuse ruumiline autokorrelatsioon põhjustab regressioonimudeli parameetrite statistilise olulisuse vääralt tugevaid hinnanguid. Autoregressiivses mudelis kasutati funktsioontunnuse väärtusi külgnevate kohtade naabrusmaatriksis, mille alusel määrati funktsioon-tunnuse autoregressiivne mõju.

Escudero *et al*. (2003) esitavad ülevaate taimede geneetiliste tunnuste ruumilise autokorrelatsiooni uuringutest ja meetoditest. Enamik uuringuid on kasutanud Morani *I* kordajat.

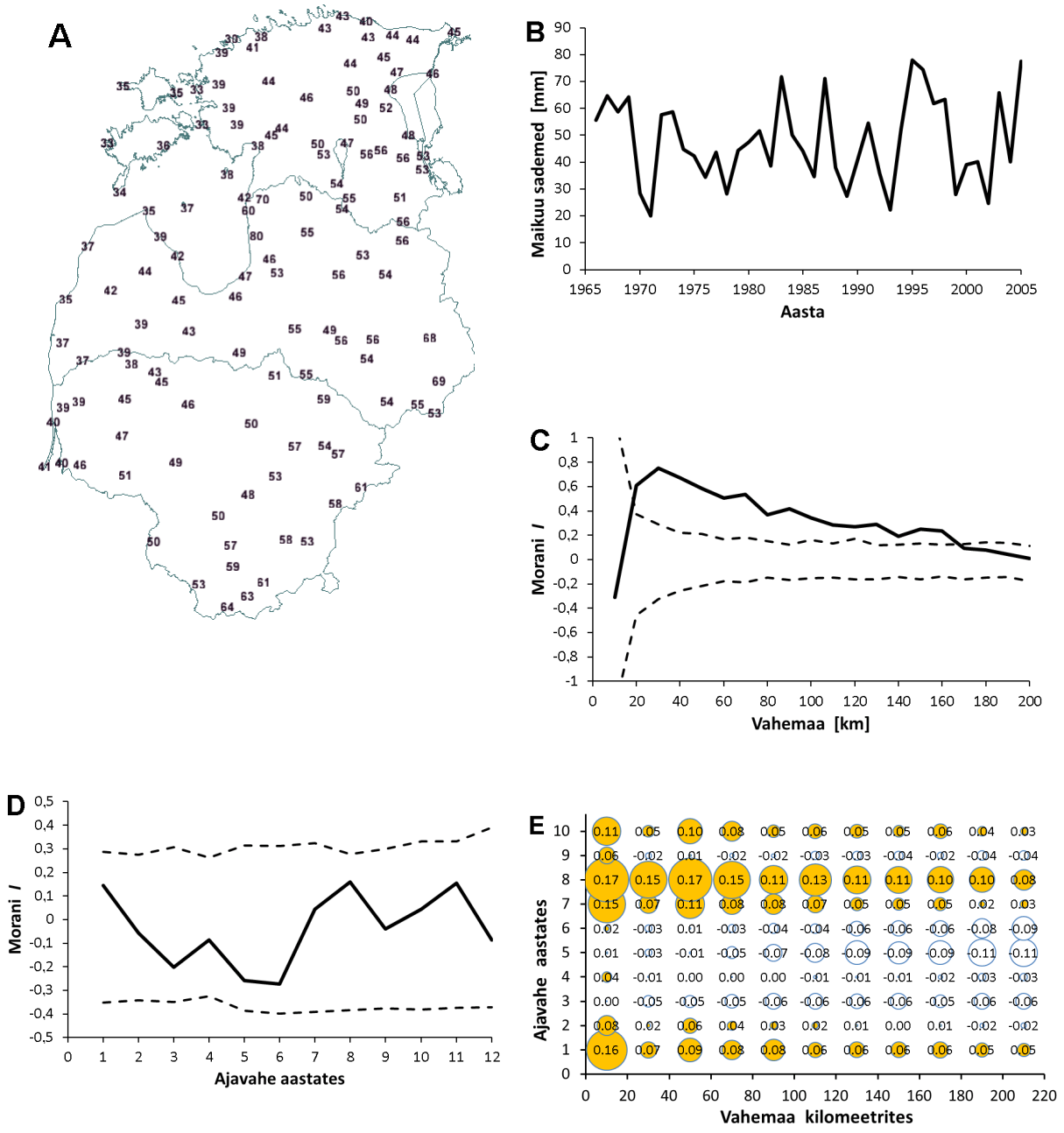
Lokaalselt arvutatud statistikute abil saab moodustada teisendatud väärtuspindasid kas ruumiliste prognooside jaoks või siis kujutise klassifitseerimiseks. Ghimire *et al.* (2010) lisasid maakatteklasside tuvastamise juhumetsa (ptk 5.6.10.2) mudelisse lisaks Landsat kanalite originaalväärtustele ka erinevates raadiustes arvutatud Getis statistiku väärtuspinnad. Parima tulemuse andis 11 × 11 piksli suuruses aknas arvutatud *G* statistik.

Tarkvara

Ajalist autokorrelatsiooni saab arvutada kõigis suuremates statistikapakettides, ruumilise auto-korrelatsiooni arvutamise võimalust pakub rasterandmete puhul näiteks Idrisi moodul AUTOCORR, mis arvutab rasterfailist Morani *I* vaid naaberpiksleid kasutades; punktandmete puhul näiteks SAM (www.ecoevol.ufg.br/sam) ja R paketid *spatial*, *ape*, *PCNM*, *AEM* ja *spacemakeR*. Kui vaatluskohtade omavahelised kaugused on arvutatud, saab autokorrelatsiooni mõõta suvalise lihtsa sarnasuskordaja abil. ArcGis arvutab kauguskaaludega Morani *I*, Anselini lokaalset Morani *I*, lähima naabri keskmise kauguse ning Getis-Ordi G ja *G** statistikut. Lokaalse seose ja Morani jaotusvälja arvutamiseks saab kasutada ka tarkvara GeoDa (Anselin *et al.* 2006).

5.1.3. Autokorrelatsioon aegruumis

Autokorrelatsioon on samal ajal nii ruumiline kui ka ajaline nähtus. Aeg-ruumilisi korrelogramme on koostatud juba 1970ndatel aastatel. Aeg-ruumilise autokorrelatsiooni kirjeldamisel võib käsitleda ajalist ja ruumilist mõõdet eraldi (Bennett 1975) või ühendada need etteantud reegli järgi üheks aeg-ruumiliseks vahemaaks (Martin ja Oeppen 1975). Ajamõõtme ja ruumi eraldi käsitlemisel saadakse kolmemõõtmeline korrelogramm, mida saab kujutada isoliinidega või kolmemõõtmelise mudeli projektsioonina. On andmeid, millest eraldi arvutud ajaline autokorrelatsioon ei ole statistiliselt oluline ja eraldi arvutatud ruumiline autokorrelatsioon ei ole statistiliselt oluline. Kombineeritud ajalis-ruumilisel arvutamisel leitakse aga statistiliselt oluline seos (joonis 5-8).

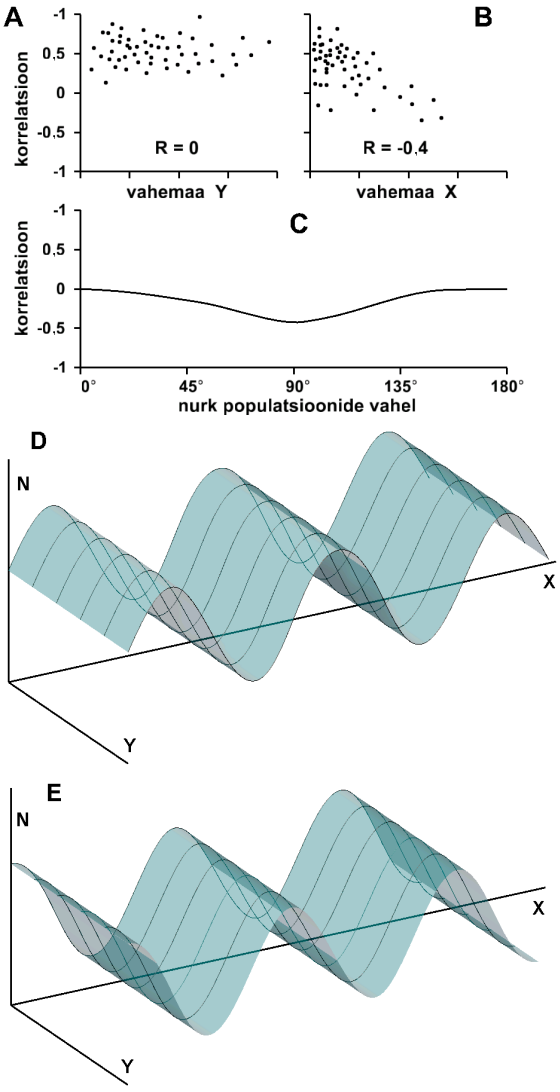


Joonis 5-8. Sademete autokorrelatsioon Baltimaade ilmavaatlusjaamades aastatel 1966 kuni 2005 (andmed Jaagus *et al*. 2010).
A – Baltimaade vaatlusjaamade keskmine maikuu sademete hulk millimeetrites.
B – Baltimaade vaatlusjaamade keskmine maikuu sademete hulk vaatlusaastatel.
C – ruumiline autokorrelogramm vaatlusaastate maikuu keskmisest sademete hulgast vaatlusjaamades juhuslikkuse 95% usaldusvahemikuga. Positiivne ruumiline autokorrelatsioon on statistiliselt oluline kaugusvahemikus 20-170 km. Väiksema vahemaaga on vaid kolm jaamadepaari ning nii vähestest andmetest arvutatud autokorrelatsiooni väärtus on juhuslik.
D – ajaline autokorrelogramm vaatlusjaamade maikuu keskmisest sademete hulgast juhuslikkuse 95% usaldusvahemikuga. Ajaline autokorrelatsioon ei ole statistiliselt oluline ühegi ajavahe puhul.
E – ajalis-ruumiline autokorrelogramm Morani *I* väärtustega. Ajalis-ruumiline autokorrelatsioon on tugevam lähemal paiknevates vaatlusjaamades, esimesel eelneval ja järgneval aastal ning taas umbes kaheksa aastase ajavahemiku järel. Suure vahemaa ja umbes viieaastase erinevuse korral on ajalis-ruumiliselt arvutatud autokorrelatiivne seos negatiivne.

Populatsioonide sünkroonia all mõistetakse eri alade populatsioonide tiheduse ajalist kovarieerumist. Enamik autoreid peab üksteisest kaugel paiknevate populatsioonide dünaamika sünkroonsuse peamiseks põhjuseks keskkonna (eelkõige ilmastiku) ruumilist sünkroonsust (Moran 1953, Lindström *et al.* 1996, Sutcliffe *et al.* 1996, Haydon ja Steen 1997, Koenig ja Knops 1998). Tõsiasja, et laia-alalised keskkonnavapustused võivad sünkroniseerida üksteisest eraldatud lokaalsete populatsioonide tiheduse liigiomast tsüklilist dünaamikat, nimetatakse **Morani efektiks**. Populatsioonide sünkroonia võimalike põhjustena märgitakse veel liigi enda leviku ja rände laineid ja looduslike vaenlaste rändelaineid. Põhiline sünkroniseeriv faktor on siiski keskkonna ajaline muutlikkus (Ripa 2000).

Populatsioonide dünaamika sünkroonsuse mõõtmiseks kasutatakse piirkonna keskmist kahe populatsiooni aegseeriate vahelistest ruumilistest korrelatsioonidest. Kusjuures populatsiooni suuruse aegseeriaid võib kasutada nii tava- kui ka diferentsiaalkujul. Viimasel juhul uuritakse populatsioonide juurdekasvude sünkroonsust. On esitatud populatsioonidünaamika rändavate lainete kiiruse suuna ja statistilise olulisuse määramise teooria, mis seletab populatsioonitsüklite dünaamikat ruumis (Lambin *et al*. 1998, Bjørnstad *et al*. 1999b) (joonis 5-9).

Populatsioonide sünkroonia korrelogrammil kujutatakse ühe piirkonna populatsioonide ajalise dünaamika sarnasuse sõltuvust populatsioonide vahemaast



Joonis 5-9. Populatsioonide dünaamika ruumiline sünkroonsus. A – populatsioonide sünkroonsus (korrelatsioon populatsioonide suuruse vahel erinevatel aastatel) sõltuvalt vahemaast Y telje suunas, B – populatsioonide sünkroonsus sõltuvalt vahemaast X telje suunas, C – populatsioonide sünkroonsuse sõltuvus suunast, D – populatsiooni tiheduse (N) paiknemise mudel ühel aastal, E – populatsiooni tiheduse mudel teisel aastal, (Lambin *et al*. 1998 ja Bjørnstad *et al.* 1999b järgi muudetult).

Uurimused

Ülevaatlikke artikleid ajalis-ruumilisest autokorrelatsioonist populatsioonidünaamikas on kirjutanud W.D. Koenig (1999), Bjørnstad *et al.* (1999a), Hudson ja Cattadori (1999), Lundberg *et al.* (2000). Sama teema kohta on ka raamatuid, näiteks Bascomte ja Solé (1998).

Ranta *et al.* (1997a, b) uurisid valgejänese *Lepus timidus* populatsioonide sünkroonia sõltuvust maakondade vahemaast Soomes ja Põhja-Ameerikas. Leiti, et Ameerika ja Skandinaavia jäneste populatsioonidünaamika ei ole sünkroonis ja vaevalt on seotud päikese aktiivsusega, küll aga on sünkroonis populatsioonidünaamika lähestikku paiknevates maakondades.

Yamanaka *et al.* (2007) uurisid pujulehise ambroosia (*Ambrosia artemisiifolia*) ja poilaste hulka kuuluva mardika *Ophraella communa* esinemist põllul. Kummagi liigi paiknemises omaette ei leitud selget ruumilist struktuuri, kuid leiti tõendeid liikidevahelise ajalise nihkega seose kohta. Ambroosia rohkus soodustas poilase rohkust samas kohas kuu ja kahe kuu pärast.

5.1.4. Autokorrelatsiooni olulisuse testid

Ruumiline autokorrelatsioon mõõdab lähestikku või teatud kaugusel olevate väärtuste vahelist statistilist seost. Statistiliste seoste puhul tuleb otsustada, kuivõrd tõenäoline on sellise tugevusega seose puhtjuhuslik teke. Kui hinnatakse ühe muutuja ruumilise autokorrelatsiooni tugevust ja on põhjust arvata, et etteantud kaugusel on võimalik ainult kas positiivne või negatiivne autokorrelatsioon, siis võib kasutada ühepoolset testi. Muudel juhtudel, ka kahe muutuja vahelise ruumilise seose uurimisel, tuleks kasutada kahepoolset testi, kuna autokorrelatsiooni suund ei ole ette teada.

Lihtsaima variandi korral käsitletakse nii ruumilist muutujat kui ka selle vaatluskohtade asendit kaheväärtuselisena – vaatlused kuuluvad või ei kuulu samasse klassi ja kas külgnevad või ei. Kaheväärtuselise muutuja puhul on pikslite külgnevuse variantide oodatav sagedus arvutatav, valemid on esitanud näiteks Sokal ja Oden (1978a). Tegeliku sageduse ja oodatava sageduse võrdlemine annab hinnangu, kui tõenäoline on vaadeldud paiknemissuhte juhuslikkus.

5.1.4.1. Z statistik

Z statistikut saab ruumilise autokorrelatsiooni olulisuse kontrolliks kasutada juhul, kui autokorrelatsiooni kordaja on eeldatavalt normaaljaotusega ja on arvutatud suurest hulgast andmetest. Näiteks Morani *I* puhul tuleks *Z* statistik arvutada järgmiselt.

$$Z = \frac{I_{obs} - I_{\exp}}{\sqrt{\mathrm{var}(I)}} \qquad [5\text{-}19]$$

Klassikaliste autokorrelatsiooni kordajate ootuse ja dispersiooni arvutamise valemid üsna pikad, neid võib leida raamatutest Cliff ja Ord (1973, 1981) ja näiteks artiklist Bonham ja Reich (1999). (Auto)korrelogrammi statistilise olulisuse võib lugeda tõestatuks, kui vähemalt ühes kaugustsoonis on (auto)korrelatsiooni olulisuse tõenäosus väiksem kui olulisuse nivoo jagatud tsoonide arvuga (Houle 1998).

Morani *I* dispersiooni [*var(I)*] arvutamise valem on üsna pikk, selle võib leida näiteks artiklist Bonham ja Reich (1999). (Auto)korrelogrammi statistilise olulisuse võib lugeda tõestatuks, kui vähemalt ühes kaugustsoonis on (auto)korrelatsiooni olulisuse tõenäosus väiksem kui olulisuse nivoo jagatud tsoonide arvuga (Houle 1998).

Morani *I* dispersiooni saab hinnata ka vaatluste valimiteks jagamisega ja Morani *I* arvutamisega igas valimis eraldi.

5.1.4.2.Järeltestid

Oletame, et dispersioonanalüüsi (ptk 1.5.4) või mõne selle analoogi kasutamisel leiti, et muutuja väärtus sõltub klassidest. Järgmine küsimus on, millised klassid erinevad omavahel olulisemalt? *Post hoc* (ladina k. "pärast seda") testid ehk **järeltestid** võimaldavad võrrelda vaatlusklasside erinevust, võttes arvesse asjaolu, et kaks võrreldavat klassi ei ole ainukesed väljavõtted andmestikust. *Post hoc* võrdlusi kasutatakse nii kirjeldava analüüsi puhul kui ka hüpoteeside testimisel.

Autokorrelatsiooni uurimisel saab *post hoc* testidega määrata, millisel kaugusel on autokorrelatsioon statistiliselt oluline, milliste kaugustsoonide vahel autokorrelatsioon oluliselt erineb või kui kaugele positiivne autokorrelatsioon ulatub. Ruumilise autokorrelatsiooni olulisuse kontrollimiseks on *post hoc* testidest kõige enam kasutatud Bonferroni kriteeriumi, kuna see on üks konservatiivsemaid. Selle puhul loetakse korrelogramm statistiliselt oluliseks, kui vähemalt ühe kaugustsooni puhul on olulisuse tõenäosus väiksem kui olulisuse nivoo jagatud tsoonide arvuga (Oden 1984).

Üksikute tsoonide kaupa vähemkonservatiivse usaldusväärsuse hinnangu annab Holmi (1979) edasiarendus Bonferroni korrektsioonist. Selleks järjestatakse tsoonid olulisuse tõenäosuse järgi ning alustades kõige olulisemat võrreldakse tsoonide olulisuse tõenäosusi olulisuse nivoo väärtusega, mis on jagatud tsoonide arvuga, mille olulisus on konkreetse tsooni olulisusega võrdne või suurem. Sisukas hüpotees võetakse vastu samm-sammuliselt kõigi tsoonide puhul kuni esimeseni, mille puhul tuleb jääda nullhüpoteesi juurde.

5.1.4.3. Mantel test

Mantel test (Mantel 1967, Douglas ja Endler 1982, Manly 1986, 1997) on vahend kahe sarnasusmaatriksi vahel oleva seose tugevuse ja statistilise olulisuse mõõtmiseks. Mantel testi kasutatakse ruumilise autokorrelatsiooni ja ruumilise korrelatsiooni statistilise olulisuse hindamiseks. Ruumiliste seoste hindamisel on ühes maatriksis vaatluste omavahelised vahemaad ja teises vaatluste erinevused. Arvude korduval juhuslikul ümberpaigutamisel ühes maatriksis saadakse juhuslikud vastavused vahemaade ja sarnasuste vahel. Korduval permuteerimisel saadakse vastavuste jaotus nullmudeli korral. Algselt kasutas Nathan Mantel seda testi haiguste leviku ajalise ja ruumilise seose mõõtmiseks.

Vastavuse taset kahe Mantel testi maatriksi vahel mõõdetakse **Mantel statistikuga** (ptk 5.1.2.1), mis on tabelites samal positsioonil olevate läheduse ja sarnasuse mõõtude korrutiste summa, millest jäetakse välja maatriksite diagonaalelemendid. Sarnasuse ja läheduse ristkorrutist standardiseeritakse lahutades kõigist väärtustest korrutise keskväärtus ja jagades need standardhälbega. Mantel testis saab kasutada ka teistsuguseid korrelatsiooni ja sarnasuse mõõdikuid.

Kuna maatriksites olevad arvud ei ole üksteisest sõltumatud, siis tuleb maatriksitevahelise seose statistilise olulisuse hindamiseks ühes maatriksis olevad arvud juhuslikult ümber paigutada (permuteerida). Enamasti on nullhüpoteesi puhul asjakohane eeldada, et iga vaatluse ja naabervaatluste vaheliste seoste struktuur (keskmine ja variatsioon) on sama, mis empiiriliste andmete puhul. Selleks, et mitte ülehinnata nullmudeli usaldatavust, muudetakse juhuslikult vaid maatriksi ridade järjekorda ning veergude järjekorra muutus tehakse vastavalt ridade ümberpaigutumisele. Nii jäävad samas veerus või reas olevate väärtuste komplektid laiali jagamata. Iga permutatsiooni järel arvutatakse Mantel statistik. Korduvate permutatsioonide tulemusel saab hinnata, kui suure tõenäosusega võis maatriksite vahel olev vastavus tekkida juhuslikult. Juhuslike ümberpaigutuste asemel on võimalik ka

järjest läbi proovida kõik võimalikud arvude paigutused maatriksis (*sequental permutation*), aga võimalike paigutuste suure arvu tõttu on see juba 9 väärtuse puhul üsna mahukas, sest 9! = 362 880. Samas soovitatakse, et vaatluste ruumilise paiknemise mõju selgitamiseks peaks vaatlusi olema kindlasti üle 20, vastasel juhul ei pruugi paiknemise mõju avalduda (Fortin ja Gurevitch 1993).

Mantel testiga määratakse kahe kaugusmaatriksi omavahelise vastavuse statistilist olulisust

Mantel testi eelis on eeltingimuste vähesus. Oluline on vaid, et originaalandmetest saaks arvutada läheduse/kauguse maatrikseid ning et vaatluspaarid maatriksites oleksid omavahel vastavuses. Kasutada saab igasuguseid läheduse/kauguse mõõdikuid ükskõik millist tüüpi andmetest. Mantel test sobib vaid testimaks hüpoteesi, mille andmed on esitatud kahe kaugusmaatriksi kujul. Mantel test ei kirjelda seost ega näita seose poolt ärakirjeldatud varieeruvuse osa (Legendre ja Fortin 2010).

Ruumilise (auto)korrelatsiooni hindamisel Mantel testiga saadakse vaatluste vahemaa mõju koondhinnang üle kõigi kaugustsoonide ehk Mantel (auto)korrelatsioon. Seega saab Mantel testiga hinnata korrelogrammi statistilist olulisust. Lihtsa Mantel testi puhul kasutatakse kõiki erinevate vahemaadega vaatusi koos ja saadakse olulisuse hinnang korrelogrammile tervikuna. Korrelogrammi nullmudelile vastava usaldusvahemiku saamiseks igas kaugustsoonis tuleb permuteerida sarnasuse arvutamise lähteandmeid igas kaugustsoonis eraldi.

Mantel osatestiga ehk Mantel partsiaaltestiga ehk mitmese Mantel testiga (*partial Mantel test, multivariate Mantel test*) hinnatakse korraga rohkem kui kahe sarnasusmaatriksi vastavust. Näiteks liigilise koosseisu sarnasust, keskkonnatingimuste sarnasust ja ruumilist kaugust. Eesmärgiks on määrata kauguse mõju pärast seda, kui keskkonnafaktorite mõju vaatluste sarnasusele on eemaldatud või vastupidi. Mantel osatesti arvutamisel võrreldakse maatriksite A ja B vahelise seose jääkide maatriksit maatriksite B ja C vahelise seose jääkide maatriksiga (Fortin ja Gurevitch 1993, Sokal ja Rohlf 1995).

5.1.4.4. Lähteandmete ümberpaigutamine ja randomiseerimine

Mantel testi puhul paigutatakse juhuslikult ringi korrelatsioonimaatriksis olevaid arve. Alternatiivsed võimalused leitud autokorrelatsiooni olulisuse kontrollimiseks on moodustada tegelikule analoogilisi juhuslikke mustreid või olemasolevaid vaatlusandmeid juhuslikult ümber paigutada ja seejärel mõõta autokorrelatsiooni tugevust nendest juhuslikult ümber paigutatud vaatlustes. Korduvate randomisatsioonide tulemusel saadakse juhumustrite korrelogrammi või autokorrelatsiooni indeksi usaldusvahemik. Kui kasutatakse 95% usalduspiire, siis on usaldusvahemik see piirkond mõlemal pool keskmist, mida täidab 95% randomisatsioonidest.

Lähteandmete randomiseerimist saab kasutada ka kahe vaatlustehulga võrdlemiseks. Kui kahe andmekogu seost konkreetse ruumiga ei uurita, siis piisab, kui randomiseerida väärtusi ühes võrreldavas andmestikus. Väärtuste vastavuse mõõtmiseks kahes andmestikus kasutatakse nn **Procrustese sängitust** (*Procrustean superimposion*) (Gower 1971b, Rohlf ja Slice 1990, Bookstein 1996, Peres-Neto ja Jackson 2001). Selle käigus standardiseeritakse kahe vaatluskogumi erinevus nende tunnusruumis nihutamise ja pööramise teel nii, et vastavus oleks maksimaalne ja kummagi kogumi sees väärtuste omavaheliste suhete struktuur ei muutuks.

Procrustese sängitus on ruumiliste andmete standardiseerimise viis

Tarkvara

Tarkvarapaketis Statistica on järgmised *post hoc* testid: Bonferroni test, Scheffé test, Tukey HSD test, erisuurusega valimite HSD test, Newman-Keulsi test, Duncani test, Dunnett' test. Nende kasutamiseks tuleb kõigepealt leida statistiline seos klassidesse (*h*) jagatud väärtuste vahel. Ruumilise autokorrelatsiooni puhul leitakse seosed vaatluste vahel mingis kohas ja kaugusel *h* sellest kohast. Kaugusklassid tuleb andmetabelisse sisestada nominaalse tunnusena. Analüüsi tulemuste menüüst tuleb avada allmenüü *Comps*, klassikoodide tunnus tuleb valida tunnuseks, mille mõju määratakse. Seejärel vajuta nuppu *Specify post-hocs for obs. means* ja vali *post hoc* test.

Uurimused

Jackson ja Harvey (1989) näitasid Mantel testiga, et Ontario piirkonna järvede lähedus määrab kalakoosluse sarnasust rohkem kui järve morfomeetria näitajad. Rodrígues ja Lewis (1997) leidsid, et kalastik sõltus rohkem järve vee läbipaistvusest kui järvedest lähedusest.

Dutilleul *et al*. (2000) näitasid, et kuigi Mantel test ja Pearsoni korrelatsioonikordaja peaksid andma sama olulisustõenäosuse, kui Mantel testis kasutatakse ruuterinevusi ja tunnused on normaaljaotusega, on praktilistes andmetes Mantel test võimsam. P. Legendre ja M.-J. Fortin (2010) väidavad siiski, et lineaarne korrelatsioon, regressioon ja kanooniline korrelatsioon on võimsamad kui Mantel test.

5.1.5. Autokorrelatsiooni mõju vältimine

Autokorrelatsiooni ebasoovitavat mõju saab vältida või eemaldada mitmel erineval viisil.

- Jätta ruumilise autokorelatsiooni ulatuses olevad andmed kasutamata. Eelarvamus, et küllap vaatlused on tehtud piisavalt suure vahemaaga, võib paraku olla petlik (Fortin ja Dale 2009). Autokorrelatsiooni mõju ulatust tuleks määrata eraldi igas üksikus andmestikus. Näiteks võib lähivaatlusi järk-järgult niikaua eemaldada, kuni autokorrelatsioon muutub lähivaatluste kaugusel statistiliselt ebaoluliseks (Legendre 1993). Kui andmete kogumiseks tuleb vaeva näha, siis ei ole see variant soovitatav, sest tähendab loobumist suurest hulgast tehtud tööst.
- Planeerida andmete kogumine nii, et üksikvaatlused oleksid omavahel võimalikult sõltumatud. Absoluutne sõltumatus ökoloogiliste ja maateaduslike vaatluskohtade vahel pole teadagi võimalik. Ka lähestikku paiknevate vaatluste grupeerimine ja vaid gruppide omavahel võrdlemine aitab ruumilise autokorrelatsiooni mõju vähendada.
- Standardhälvete, korrelatsioonikordajate ja *t* testi statistikute korrigeerimine (Cliff ja Ord 1981) või vabadusastmete arvu korrigeerimine (Thomson *et al*. 1996, Clifford *et al*. 1989).
- Rangema olulisusnivoo rakendamine.
- Lülitada ruumiline autokorrelatsioon mudelisse. Ruumilise sõltuvusega regressioonimudelite andmetele lähendamine toimub iteratiivselt ja on arvutusmahukas.
- Andmestiku jagamine piirkondadeks ja piirkonna lülitamine mudelisse kvalitatiivse tunnusena või mudeli sobitamine igas piirkonnas või iga koha ümbruses eraldi (geograafiliselt kaalutud regressioon, ptk 5.2.4).
- Lülitada autokorrelatsioon (laigulisus) null-mudelisse. Seda varianti kasutatakse ruumimustrite omavahelise korreleerumise uurimisel. Eri klassi nähtuste omavahelist paiknemise analüüsil eeldatakse, et ruumiline autokorrelatsioon on kummagi mustri siseasi, mis nende omavahelise paiknemise hindamist mõjutada ei tohi. Küsimus, milline mustri iseloomu

(autokorreleerumise) karakteristik ja kuidas nullmudelisse lülitada, ei pruugi olla lihtsate killast.

- Kasutada autokorreleerunud andmete regressioonanalüüsi asemel andmete dimensionaalsust vähendavaid meetodeid, näiteks külgnevustest arvutatud omaväärtusi (ptk 5.1.2.8).

Ruumiliselt autokorreleerunud muutujate lähestikku paiknevate vaatluste lisamine valimisse ei pruugi olulist lisateavet anda

5.1.5.1. Autokorrelatsioon liikide leviku mudelites

Enamik liikide leviku mudeleid ignoreerib liikide levikuks kuluvat aega eeldades, et liiki võib leida kõikjalt, kus on sobiv elupaik. Vaatlusandmed näitavad aga, et selline eeldus ei kehti. Mingi liigi leidmise tõenäosus on suurem kohtades, mille ümbrusest on liiki juba leitud. Kui liik esineb ohtralt, siis võib teda leida ka ebasobivatest elupaikadest.

Autokorrelatsiooni kaasamiseks mudelisse on kasutatud autoregressioonimudelite variante. Autologistilist mudelit esinemise/puudumise andmetega kasutasid esmalt Smith (1994) ja Augustin *et al*. (1996) ning seejärel mitmed teised (Wu ja Huffer 1997, Osborne *et al*. 2001, Segurado ja Araujo 2004, Luoto *et al*. 2005, Betts *et al*. 2006, Dormann 2007a). Autologistilist meetodit saab laiendada ka teistele jaotustele (Haining 2003).

Autologistilise meetodi üks alternatiiv GEE meetod (*generalized estimating equations*) on üldiste lineaarsete mudelite modifikatsioon (Albert ja McShane 1995) ja on kasutatav R keele vabavarapakettides *gee* ja *geepack*. Ruumilise autokorrelatsiooni mõju arvestamiseks ökoloogilistes uuringutes on seda kasutanud siiski vaid vähesed autorid (Gotway ja Stroup 1997, Gumpertz *et al*. 2000, Augustin *et al*. 2005, Carl ja Kühn 2007).

Teine ja lihtsam võimalus on lisada mudelisse omaette tunnusena kaugusega kaalutud sama liigi leidude hulk ümbruses, mis võib hinnanguid täpsemaks muuta isegi siis, kui levimine ei ole limiteeriv faktor (Allouche *et al*. 2008).

Uurimused

Brito *et al.* (1999) kasutasid juhuslikke mittekülgnevaid UTM-ruute. Li *et al.* (1997) uurisid enne proovide võtmist, milline on piisav vahemaa proovide vahel, et ruumiline autokorrelatsioon ei mõjutaks tulemusi.

Näiteid ruumilise autokorrelatsiooni lülitamist mudelisse võib leida töödest: Haining (1990), Bonham ja Reich (1999), Augustin *et al.* (1996), Syartinilia ja Tsuyuki (2008). Lichstein *et al.* (2002) lisasid liikide levikut prognoosivatesse regressioonimudelitesse funktsioontunnuse ruumilise autokorrelatsiooni. Kahe linnuliigi (*Dendroica pensylvanica* ja *Setophaga caerulescens*) arvukuse mudelis jäi autokorrelatsiooni lisamise järel domineerima liigi seos maastiku tunnustega, kolmanda liigi (*Wilsonia citrina*) puhul oli ruumiline autokorrelatsioon põhiline faktor. *W. citrina* on mõnes kohas suurel hulgal ja mõnes teises kohas teda ei ole mingil põhjusel, mida mudelis olevad maastikutunnused ei kirjelda.

Dormann *et al.* (2007) võrdlesid liikide esinemiskohtade autokorrelatsiooni arvestamise meetodeid:

- regressioon autokorrelatsiooniga (*autocovariate regression*),
- omavektori kaardistus (*spatial eigenvector mapping*),
- üldistatud vähimruutude meetod (*generalised least squares*),
- autoregressiivsed mudelid (*autoregressive models*),
- üldistatud prognoosivõrrandid (*generalised estimating equations – GEE*).

Ruumilise autokorrelatsiooni eiramine põhjustas hinnangutes keskmiselt 25% suurema ebatäpsuse. Võrdluse tulemusena julgustavad autorid kasutama mitut eri tüüpi mudelit paralleelselt. Kõigi meetodite puhul tuleks kas *a priori* või katsetamise teel määrata ümbrusest sõltuvad kaalud või mingid muud autokorrelatsiooni struktuuri kirjeldavad parameetrid.

5.2. Interpoleerimine

Enamik empiirilisi andmeid käib mingi ajahetke või ruumipunkti kohta, kusjuures järeldusi ja oletusi oleks vaja teha tihedama punktide võrgu kohta kui vaatlusvõrk. Muutuja väärtuste hindamist registreeritud tulemustega kohtade vahele nimetatakse **interpoleerimiseks**. Interpoleerida saab nii piki ühte telge (aegread, transektid ehk läbilõiked) kui ka mitmemõõtmelises ruumis (kaart, kõrgusmudel, täis 3-mõõtmeline mudel näiteks veekogust, puuvõrast, maagisoontest vms). Interpoleerimine on vajalik kõikjal, kus andmed on olemas vaid teatud kohtades, järeldusi on vaja teha aga kogu pinna või ruumi kohta. Tüüpilisi näiteid leiab seireandmete töötlusest, aga ka andmeedastuse vallast. Kujutise (kadudega) kokkupakkimise üks võimalik meetod on salvestada väärtused vaid teatud sammuga pikslite võrgustikule ja vahepealsete pikslite väärtused interpoleerida võrgustiku lahtipakkimisel.

Interpoleerimine on väärtuste arvutamine mõõdetud kohtade vahele

Interpoleerimismeetodeid võib jagada järgmiselt:

- interpoleeritava pinna osadeks jagamine ehk tesselatsioon,
- silumine libiseva aknaga,
- interpoleerimine lokaalse või globaalse funktsiooniga, sealhulgas kriging (ptk 5.3.5),
- interpoleerimine ekspertotsuste ja intellektitehnika meetoditega (ptk 3.4.5),
- sarnast struktuuri otsivad ja sobitavad meetodid (ptk 5.2.4 ja 5.2.5),
- interpoleerimine teiste muutujate abil (ptk 5.2.6).

Interpoleerimine kitsamas tähenduses tähendab väärtuste genereerimist kasutades vaid teadaolevaid väärtusi. Teatud hulk kindelpunkte on ette antud ja väärtuspind tuleb mingite reeglite järgi nende punktide vahele genereerida. Lihtsaim näide on libiseva aknaga silumine, mille puhul on mitmeid variante. Üsna sageli on vaatluskohtade vahele jääva ala kohta üht-teist veel teada ja seda teadmist saab interpoleerimisel ära kasutada.

Vaatluste vaheline ala võib jaguneda suhteliselt homogeenseteks ehk ühetaolisteks piirkondadeks. Kui selliste piirkondade ehk plokkide paiknemise piirid on teada, siis ühe ploki piires ühesuguste interpoleeritavate väärtuste eeldamine annab üsna usaldusväärse interpoleeritud pinna. Sellest lähtub näiteks plokk-kriging. Interpoleeritava ala plokkideks jagamine võib toimuda ükskõik millise indikaatortunnuse järgi.

Interpoleerimisalgoritme jagatakse täpseteks ja siluvateks. Esimesed konstrueerivad pinna, mis läbib vaatluspunkte, teised moodustavad pinna, mille väärtus ei pruugi vaatluspunktides vastata seal mõõdetud väärtustele (arvestavad mõõtmisvea võimalusega). Täpsete interpolaatorite hulka kuuluvad näiteks kaalutud kaugusi kasutavad meetodid, kriging ilma eheda variatsioonita ja lähima naabri meetod. Vaatluspunkte ei jälgi regressioonimudeliga interpoleerimine ja kriging eheda variatsiooniga.

Pinna moodustamine vaatluskohtade vahele võib toimuda järk-järguliselt arvestades genereeritava pinna ruumilist struktuuri. Pinna struktuur võib olla kas ette antud, eelnevalt teada või püütakse seda määrata interpoleerimise käigus. Struktuuri sobitavad meetodid on enamasti iteratiivsed ehk järk-järgulised. Järk-järgulisel pinna moodustamisel kohandatakse iga ruumilise üksuse (piksli või areaali) väärtuse lisamise järel teiste üksuste oodatavaid väärtusi arvestades ruumilise struktuuri eripära. Punktobjektide esinemise tõenäosuse autokorrelatiivse iteratiivse interpoleerimise meetod kannab nime Gibbsi sampler (ptk 6.1.5.1). Struktuuri sobitavate meetodite kasutamisel eeldatakse reeglina suurema või väiksema hulga juhuslikkuse olemasolu. Seega võib ka stohhastilise interpoleerimise

lugeda struktuuri sobitamise meetodite hulka.

Peale genereeritava nähtuse enda ruumilist struktuuri saab arvestada ka vaatluste vahele jääva ala teisi omadusi, mille järgi saab interpoleeritavaid väärtusi arvutada. Interpoleeritava pinna loomine täiendavaid andmeid kasutades on lähedane pinna genereerimise ülesandele, mida käsitletakse lähemalt peatükis 6.2.

Interpoleerimise lähtetingimused võivad sisaldada etteantud murdejooni ja tõkkeid. Erinevate autorite poolt on mainitud järgmiste faktorite mõju interpoleerimistulemustele:

- vaatluste tihedus,
- vaatluste paiknemismuster,
- andmete varieeruvus,
- väärtuste vastavus normaaljaotusele,
- kõrvaliste faktorite mõju ja lisateabe kvaliteet.

Uurimused

McBratney *et al*. (2003) ja Scull *et al*. (2003) esitasid ülevaate mullastiku kaardistamisel kasutatud interpoleerimise ja modelleerimise meetoditest ja andmekihtidest.

Li ja Heap (2011) võtsid kokku 53 varasemat interpoleerimismeetodeid võrdlevat uuringut ja nendes kasutatud 72 interpoleerimisviisi. Kõige sagedamini on kasutatud krigingut, pöördkaugusega kaalumist ja kokrigingut. Interpoleerimise tulemusi mõjutab kõige enam andmete varieeruvus – suurema varieeruvuse korral on interpoleeritud pind ebatäpsem.

5.2.1. Tesselatsioon

Tesselatsioon ehk mosaiikimine on kahemõõtmelise pinna jäägita jagamine kindla kujuga mittekattuvateks osadeks. Tesselatsioonipolügoonid võivad olla täiesti korrapärased või vaid mingil määral ühetaolised. Täiesti korrapäraseid tesselatsioone on vaid kolm: võrdkülgsetest kolmnurkadest, ruutudest või võrdkülgsetest kuusnurkadest koosnev pinna jaotus. Kolme- ja enamamõõtmelise ruumi korrapärasteks osadeks jagamise tulemuse nimi on kärg (*honeycomb*).

Pinna osadeks jagamine on kasutatav ka interpoleerimisülesande lahendamisel. Näiteks interpoleerimisel lähima naabri meetodil omistatakse vaatluse väärtus vaatluskoha proksimaalregioonile.

5.2.1.1. Pindade kombineerimine

Interpoleeritavad väärtused võivad olla mitte ainult vaatluspunktide, vaid ka eraldiste atribuudid. Iga laigu võib siduda esinduspunktiga, millele omistatakse ala omadused. Seejuures võivad eraldised olla ebakorrapärase kujuga, kuid võivad ka moodustada korrapärase võrgustiku. Näiteks puistu koosseis metsaeraldises, jahipiirkonnas tegutsevate jahimeeste arv või elanike arv vallas. Selleks, et teada saada tegutsevate jahimeeste suhet iga valla elanike hulgas, tuleks jahipiirkondi kombineerida valla piiridega. Eraldistena esitatud pindade kombineerimist on nimetatud **pindade interpoleerimiseks** (*areal interpolation*).

Interpoleerimisel saab esinduspunktidega asendada nii teadaolevad alad kui ka need alad, millele andmed üle kantakse ja interpoleerida edasi nagu punktandmetega. Teine variant tsentroidide kasutamisel on sihtareaalidest tihedama **ülekandevõrgustiku** kasutamine. Sel juhul interpoleeritakse kõigepealt väärtused ülekandevõrgustiku tsentroididele ja hiljem keskmistatakse ülekandevõrgustiku ühe sihtareaali piiresse jäävad väärtused.

Laikude atribuute saab teise laikude struktuuri üle kanda piisavalt tiheda ülekandevõrgustiku abil

Ruumilisi muutujaid jagatakse **pindalast sõltuvateks** (*spatially intensive*) ja **pindalast sõltumatuteks** (*spatially extensive*). Pindalast sõltuvad tiheduse ja intensiivsuse näitajad, pindalast ei sõltu hulga näitajad. Tiheduse andmeid ei saa eraldiste liitmisel liita, tiheduse arvutamisel liiteraldises tuleb tihedusi eraldiste pindalaga kaaluda. Hulga ja loenduse andmeid saab eraldiste liitmisel liita.

Goodchild ja Lam (1980) esitasid arvutusliku meetodi, mille puhul ei ole vajadust ülekandevõrgustikku kasutada, küll aga on vaja teada lähte- ja sihtareaalide kõigi kombinatsioonide pindalasid. Peale selle on vaja arvestada interpoleeritava nähtuse tüüpi. Kaasaja geoinformaatika tarkvara ja arvutusvõimaluste juures ei ole rasterkujul andmekihi kasutamine ülekandevõrgustiku rollis tülikam kui eraldiste kombineerimine pindala funktsioonide abil.

5.2.2. Silumine

Silumine toimub liikuva aknaga, mis liigub piki andmeid ja arvutab akna piiresse jäävate vaatluste järgi akna keskkoha jaoks parameetri väärtuse. Tulemusi mõjutab nii akna kuju, suurus kui ka akna sisse jäävate andmete keskmistamise algoritm. Akna suurus võib olla määratud nii akna pindala, raadiuse kui ka akna sisse jäävate vaatluste hulgaga. Akna kuju võib olla balansseeritud, nii et igas suunas (sektoris) oleks sama palju või siis sektori pindalaga proportsionaalne arv vaatlusi. Keskmistamine võib olla kaalutud või kaalumata. Silumismeetodit saab kasutada nii korrapäraste kui ka juhuslikult paiknevate punktandmete ja pindade puhul.

Kaalutud keskmistamise korral kasutatakse enamasti kaaludena mingit kauguse funktsiooni. Levinuim on **kauguse pöördväärtus** (*inverse distance weighting – IDW*). Interpoleeritud hinnangu arvutamisel igas kohas kasutatakse kõiki andmeid, kaugemad lähevad arvesse tühise osatähtsusega. Lisaks kauguse pöördväärtusele kasutatakse ka **kauguse pöördväärtuse ruutu**. Sel juhul on osakaal pöördvõrdeline kaugustsooni pindalaga. Kolmes ruumimõõtmes paiknevate vaatluste puhul on analoogselt asjakohane kasutada kauguse pöördväärtuse kuupi. On kasutatud ka keerukamaid kaalude arvutamise eeskirju, näiteks **paraboolset teisendust** (Remm 1989).

$$w_{ij} = 1 - \frac{\left(MX_j - X_i\right)^2}{E^2}, \qquad [5\text{-}20]$$

kus w_{ij} on vaatluse i kaal akna asendi j korral, MX_j on akna keskkoht, $X_i$on vaatluse *i* asukoht ning *E* on akna laius. Kui *i* ja *j* vaheline vahemaa ületab akna laiust, siis omistatakse $w_{ij} = 0$.

Silumise käigus leitakse hinnang kaalutud keskmisena kogu valimist, kaalud sõltuvad hinnatava koha kaugusest ja võivad omada nullväärtust

Kas eelistada kauguse pöördväärtust või mingit muud kaalude arvutamise viisi, tuleb otsustada iga uurimuse korral eraldi, nn *ad hoc* viisil. *Ad hoc* otsusel ei pruugi olla mingit erilist sisulist põhjendust, mõned küsimused tuleb lihtsalt mingil viisil ära otsustada. Kui nähtuse ruumilise autokorrelatsiooni kohta eelnevaid teadmisi ei ole, võib katsetada mitmete akna suurustega ja kaalude arvutamise eeskirjadega.

Kõigi silumismeetodite puhul kipub tekkima raskusi piirkondades, mille lähedal ei ole ühtegi vaatlust. Fikseeritud ja väikese silumisulatusega meetodid ei anna selliste alade kohta üldse hinnangut, suure või piiramata silumisulatuse korral võib saada kahtlase väärtusega tulemuse. Esmapilgul võib tunduda, et andmete puudumisel ei saa tööd jätkata. Siiski, sellises olukorras võib interpoleerimistulemust oluliselt parandada teave interpoleeritava tunnusega korreleeruvate faktorite kohta ja teadmised selle tunnuse väärtuste ruumilisest struktuurist. Näiteks sademete hulga prognoosi aitavad täpsustada andmed maapinna kõrguse ja piirkonna metsasuse kohta; mingi liigi arvukuse interpoleerimisel loenduskohtade vahele on abiks andmed elupaiga hulga ja kvaliteedi kohta.

Interpoleerida saab ka akna miinimumi, maksimumi või muid väärtusi ja teisendusi kasutades. Näiteks miinimumväärtuste trendpinna leidmine. Kui interpoleerimisalgoritm lubab interpoleeritud väärtusi väljaspool aknasiseste andmete haaret, siis oleks toimingut õigem mitte nimetada silumiseks, vaid teisendamiseks või filtreerimiseks.

5.2.2.1. Korduv silumine

Silumist võib teha ühe korraga, piisavalt suure akna ja kaalutud efektidega. Teine võimalus on kasutada väiksemat akent korduvalt – juba silutud pinda silutakse veel. Korduva silumise korral võib ka kaale kasutada, aga viimaste kasutamine ei ole nii oluline kui ühekordsel silumisel. Korduva silumise korral tuleb eelnevalt otsustada silumise lõpetamise kriteerium – kas on see kindel arv kordusi või piisavalt sile tulemus. Piisava silutuse otsustamiseks on taas vaja mingit kriteeriumi. Korduv silumine võib toimuda andmevõrgustiku tihendamisena, aga võib toimuda ka kohe alguses etteantud piisavalt tiheda võrgustiku täitmisena.

5.2.2.2. Suundadega silumine

Suundadega silumisel määratakse iga piksli ümber kontrastsus mitmes eri suunas. Interpoleeritakse vaid minimaalse kontrastsuse suunas. Sellega garanteeritakse kontrastsete servade ja kitsaste taustast erinevate ridade säilumine interpoleeritud kujutises. Mõnikord tekitab see meetod ebaloomulikke struktuure.

5.2.3. Interpoleerimine teiste tunnuste abil

Mullateaduse üheks metoodiliseks probleemiks on, kuidas kõige täpsemalt ja väiksema ressursikuluga hinnata mulla omadusi täpsete mõõtmiste vahele jääval alal. Varasemad mullakaardistused on põhiliselt tuginenud kaardistaja eksperthinnangule, mis põhineb kogetud seostel taimkatte, reljeefi, maapinna iseloomu ja mulla vahel. Liigestatud pinnamoe puhul on hinnang täpsem, kui kasutada ka pinnamoe andmeid. Teiste pidevate tunnuste abil interpoleerimisel on põhiliselt kaks varianti: kas koostada regressioonimudel ja prognoosida vaatluskohtade vahelised väärtused mudelist või klassifitseerida kogu ala mingit tüüpi eraldisteks ja omistada eraldisele kõige tõenäolisem prognoositav väärtus. Teise tunnuse järgi interpoleerides, nagu igasuguse interpolatsiooni korral, võib hinnangule lisada ka teatud hulga juhuslikkust, et saada tõepärasemat mustrit.

Näiteks on teiste tunnustega interpoleerimist kasutatud sademete hulga interpoleerimiseks vaatlusjaamade vahele (Goovaerts 2000). Florinsky *et al.* (2002) leidsid, et oskuslik klassifitseerimine võib mulla omaduste kaardistamisel anda paremaid tulemusi kui regressioonimudel.

5.2.4. Interpoleerimine regressioonimudeliga

Ruumiandmete regressioonimudeli kasutamisel sobitatakse mudeli parameetrid vaatluste järgi ja lahendatakse mudel igas interpoleeritavas kohas kasutades selle koha tunnuseid. Arvutatavad kohad võivad vaatluskohtadest erineda ja interpoleerimisülesande puhul üldiselt erinevadki.

Regressioonimudeli parameetrid saab sobitada globaalselt, kasutades kõiki õpetusandmeid, või lokaalselt, kasutades vaid kohalikke andmeid. Regressioonimudeli lokaalseks sobitamiseks on vähemalt kolm põhjust. Esiteks, vaatlusandmete ebaühtlase juhusliku varieeruvuse ja vaatluskohtade erineva tiheduse tõttu ei ole regressioonimudeli esinduslikkus ruumis konstantne. Teiseks, seos tunnuste vahel võib iseenesest olla ruumis muutuv. Kolmandaks, ükski mudel ei ole täiuslik. Mudel lihtsustab tegelikkust ja ei võta arvesse kõiki mõjufaktoreid. Mudeli lokaalne sobitamine võib anda vihjeid oluliste seletavate tunnuste kohta, mis on mudelis arvesse võtmata.

Lokaalselt sobitatud kordajatega regressioonimudelit nimetatakse **geograafiliselt kaalutud regressiooniks** (*geographically weighted regression – GWR*). GWR lubab regressioonikordajate ruumilist muutlikkust. GWR mudeli parameetrid sobitatakse igas uuritava ruumi osas eraldi, omistades igale vaatlusele kaugusest sõltuva mõjukaalu (Brunsdon *et al*. 1996, 1998, Fotheringham *et al*. 1998, 2005). Enamasti on kaaluks kauguse pöördväärtus, kusjuures etteantud piirkaugusest kaugemal olevate vaatluste kaaluks on null – nende vaatluste mõju ei arvestata. Lokaalse kerneli ulatus võib sejuures olla muutuv (Pàez *et al*. 2002).

Lokaalne regressioon võimaldab kirjeldada statistilise seose ruumilist varieeruvust. Lisanduval vabadusel on alati ka kaasmõju – lokaalse regressiooni tulemus sõltub kasutatud ruumijaotusest. Uuritava ala testmoodi osadeks jagamisel on võimalik, et saadakse teistsugused seosed, milles saab teha teistsuguseid järeldusi.

GWR mudeli võrrandi saab kirja panna järgmiselt:

$$Y_i = B(u_i, v_i)_0 + \sum_k B(u_i, v_i)_k x_{ik}, \qquad [5\text{-}21]$$

kus i on koha indeks, k on tunnuse indeks, B_0 on regressiooni vabaliige, B_k on koha k regressioonikordaja, u_i,v_i on koha i asukohakoordinaadid, x_{ik} on seletava tunnuse x_k väärtus kohas i.

Lokaalse regressiooniga silumise alla kuulub ka silumine splainidega, mida arvutatakse lokaalsete (enamasti kolmanda astme) polünoomidena.

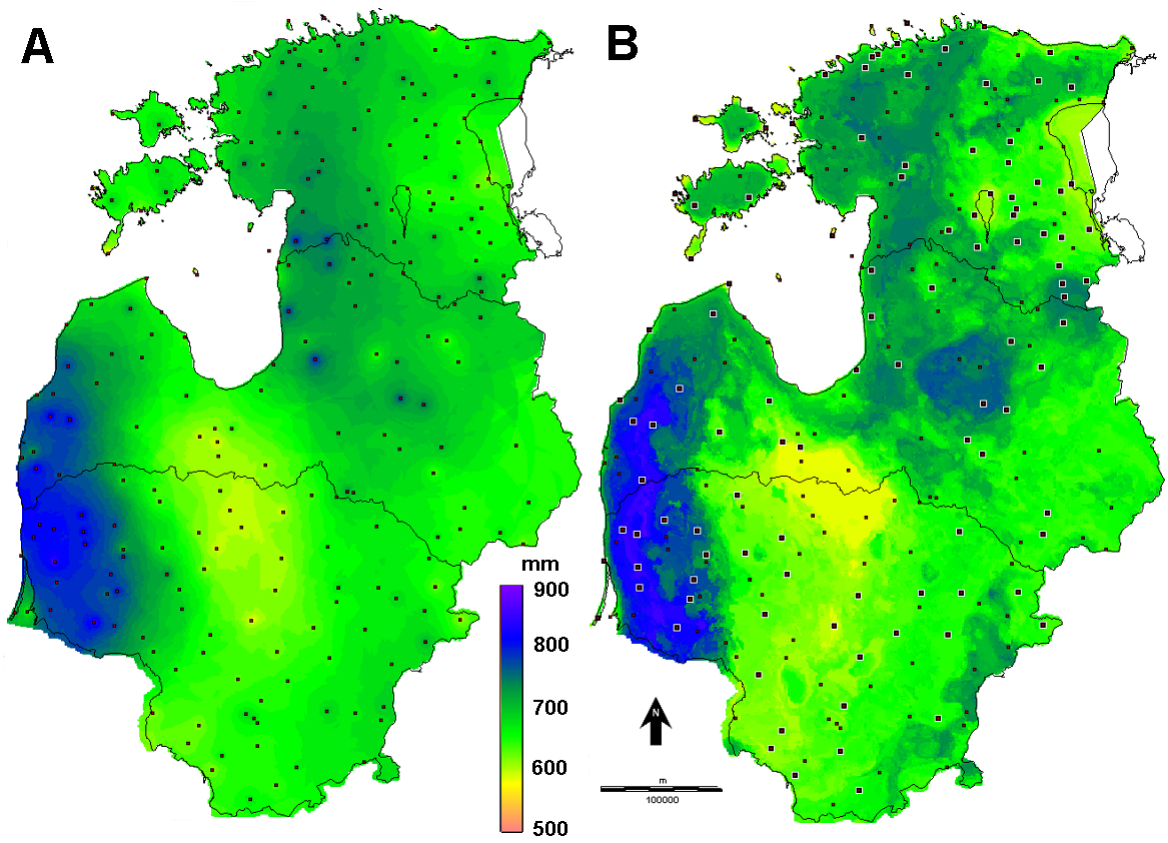
Regressioonimudeliga interpoleerides on võimalik arvesse võtta nii naabervaatluste väärtusi, seoseid täiendavate tunnustega kui ka seoste ruumilist varieeruvust

GWR moodul on tarkvara ArcGis koosseisus. Tehnilisi üksikasju, kasutamisõpetuse ja näidisandmeid saab GWR kodulehelt: http://ncg.nuim.ie/ncg/GWR. Näiteid GWR kasutusest liikide ja elupaikade kaardistamisel on liikide leviku modelleerimise peatükis. Regressioonimudeliga interpoleerimist ja ekstrapoleerimist ehk ruumilist prognoosi kasutatakse sageli klimatoloogilistes uuringutes (Hjort *et al*. 2011).

5.2.5. Interpoleerimine sarnasuse järgi

Sarnasuse järgi interpoleerimine on üks sarnasusele tugineva järeldamise rakendustest (ptk 3.4.6). Enne näidistega sarnasuse järgi interpoleerimist tuleb leida interpoleeritava muutujaga seonduvad tunnused ja soovi või vajaduse korral omistada neile kaalud. Interpoleerimisülesanne erineb näidistega sarnasuse järgi prognoosikaardi arvutamisest (ptk 5.6.8) õpetuskohtade koordinaatide kasutamise poolest. See tagab sarnaste koordinaatidega kohtade, see tähendab lähestikku olevate kohtade, suurema sarnasuse. Näidistega sarnasuse järgi on koostatud näiteks Baltimaade sademete 30 aasta keskmise kaart (Remm *et al*. 2011). Kasutades iga kaardistatava koha sarnasust vaatlusjaamade asukohtadega igat kohta iseloomustavate tunnuste poolest, peaks sademete hulk olema suurem, võrreldes ümberkaudsete vaatlusjaamade andmete kauguskaalutud interpolatsiooniga, metsases Vahe-Eestis ja kõrgustikel ning väiksem rannikul ja suurte järvede kohal (kus vaatlusjaamad paraku puuduvad) (joonis 5-10).

Näidiste järgi interpoleerides omistatakse interpoleeritavale kohale kõige sarnasema näidiskoha väärtus, kusjuures tunnuste hulgas kasutatakse ka asukohakoordinaate



Joonis 5-10. Baltimaade sademete 30 aasta keskmine kauguskaaludega interpoleerimisel (A) ja hinnates näidistega sarnasuse järgi (B) (Remm *et al.* 2011). Mustad ruudukesed tähistavad vaatlusjaamu: suuremad ruudud õpetuskogumit, väiksemad kontrollkogumit.

5.2.5. Struktuuri sobitamine

Struktuuri sobitavad meetodid püüavad interpoleerimise käigus jälgida etteantud struktuuri-parameetreid. Struktuuri sobitamisel kombineerub interpoleerimine mustri genereerimisega.

Lõplike erinevuste meetod (*finite differences method*) lähtub pinna iga koha (piksli) puhul diferentsiaalvõrrandist (Laplace võrrandist)

$$\frac{\delta^2 z}{\delta x^2} + \frac{\delta^2 z}{\delta y^2} = 0. \qquad [5\text{-}22]$$

See tähendab, et kui väärtuste *z* teist järku diferentsiaal koha *i* (mille koordinaadid on x_i ja y_i) ümbruses kaaluda kauguse ruudu pöördväärtusega, siis peaksid need olema iga koha ümbruses tasakaalus ja võrdsed nulliga.

Lihtsaimal juhul jagatakse pind võrdseteks ruudukujulisteks piksliteks ja kasutatakse vaid külgnevaid piksleid. Järgnevalt võetakse etteantud tingimuseks, et iga interpoleeritava piksli väärtus peab olema võrdne külgnevate pikslite väärtuste keskmisega. See tähendab, et teist järku diferentsiaal ehk kumeruse muutus, on null. Kui ühel pool on suuremad väärtused, siis teisel pool peavad olema samavõrd väiksemad. Teadaolevate väärtustega pikslite väärtusi seejuures ümber ei arvutata.

Lõplike erinevuste meetodit on kasutatud näiteks hüdrogeoloogias (põhjavee liikumise mudel MODFLOW http://water.usgs.gov/nrp/gwsoftware/modflow2005/modflow2005.html) ja liikide invasiooni modelleerimisel (Lewis ja Kareiva 1993, Almeida *et al.* 2006).

.

5.3. Geostatistika ja variograafia

Geostatistika on tähistanud igasuguste ruumis muutuvate nähtuste analüüsi ükskõik millise statistilise meetodi abil. Uuemal ajal on aga kinnistunud selle termini kitsama tähendusega kasutus – ruumilise autokorrelatsiooni modelleerimise tähenduses. Kui teised meetodid kas hindavad ruumilise autokorrelatsiooni olemasolu tõenäosust või kirjeldavad seda sarnasuste mõõtmiste abil, siis **variograafias** kirjeldatakse ja modelleeritakse varieeruvust vastavalt variogrammi ja variogrammi mudeli abil (ptk 5.3.4). Seejärel kasutatakse variogrammi mudelit prognoosikaardi arvutamisel. Väiksem uuritava tunnuse varieeruvus mingil kaugusel olevates vaatluspaarides viitab tugevamale autokorrelatsioonile sellel kaugusel. Autokorrelatsiooni puudumisel on variogramm lame.

Üldiselt arvatakse, et variograafia võeti esimesena kasutusele mäeteaduses (Matheron 1963), aga sisuliselt samu ideid publitseeriti tunduvalt varem metsanduses. Langsæter (1926) arvutas takseerandmete varieeruvuse seost vaatluskohtade vahemaaga piki vaatlustrassi ja kujutas seda graafikul, st koostas variogramme. Rootsi metsandusstatistik B. Matérn (1947) arvutas varieeruvust ja kovariatsiooni kindla vahemaaga vaatlusandmete vahel. B. Matérni doktoritöö (Matérn 1960, 1986) on saanud ruumilise varieeruvuse modelleerimise klassikaks. Selles esitati muuhulgas hiljem Matérni nime saanud korrelatsioonimudel ja variogrammi mudelite seeria.

1960ndatest ja 70ndatest aastatest alates kuni kaasajani seostatakse geostatistikat põhiliselt siiski Georges Matheroni töödega. Kuna Matheron ja tema õpilased kirjutasid prantsuse keeles, said tema ideed inglisekeelses teadusmaailmas laiemalt tuntuks alles 1970ndate aastate keskel. Matheroni esimene globaalse levikuga põhjalik teoreetiline artikkel (Matheron 1973) ilmus alles 1973. aastal.

G. Matheron oli professor Pariisi mäeteaduse koolis (*Ecole Normale Superieure des Mines de Paris*). Kui kool Pariisist laiali hajutati, asus Matheron kolledži *Centre de Morphologie Mathematique* direktoriks. Hiljem jagunes see kahe õppekava vahel: matemaatiline morfoloogia ja geostatistika. Põhja-Ameerikasse on geostatistika ideed viinud kaks Matheroni õpilast. Andre Journel asus Stanfordi Ülikooli 1978. Aastal ja Michel David Montreali *Ecole Polytechnique*'i 1977. aastal.

Matemaatikute hulgas ei võetud Matheroni töid eriti hästi vastu, kuna geostatistikat peeti teiste nimede all juba teada olevate meetodite dubleerimiseks. Sellisele suhtumisele aitas kaasa Matheroni tava publitseerida vaid prantsuse keeles ja enamasti asutuse ametialaste sisedokumentidena. Tänapäevaks on enamik mainitud ametkondlikest aruannetest rahvusvahelistes statistikaajakirjades ära trükitud ja geostatistilised meetodid on laia levikuga arvutiprogrammide osad.

Tihti seostatakse geostatistikat vaid interpoleerimisega, mis ei ole päris õige. Enamik interpoleerimismeetodeid olid kasutusel juba enne Matheroni töid. Teisest küljest – traditsioonilised interpoleerimismeetodid, nagu kauguse pöördväärtusega kaalumine või libiseva aknaga silumine, ei modelleeri ruumilist autokorrelatsiooni ega ruumilist hajuvust. Need vaid lähtuvad ruumilise autokorrelatsiooni olemasolust.

Geostatistikale on iseloomulik muutuja ruumilise varieeruvuse või autokorrelatsiooni modelleerimine

Geostatistika esmane eeldus on ruumiliste andmete olemasolu. See tähendab, et iga mõõtmistulemus on seotud kindla asukohaga ruumis. See asukoht võib olla kas punkt või kindla suuruse ja kujuga pind või ka kolmemõõtmeline ruumiosa. Andmetega seotud pinna- või ruumiosa nimetatakse geostatistikas **andmetoeks** (*support*). Oletame, et mingit muutujat (õhu- või mulla temperatuuri, sademete hulka, saasteaine kontsentratsiooni, maagisisaldust mingis lasundis) on mõõdetud kohtades *x1, x2, x3* jne. Muutuja väärtus nendes kohtades oli *Z(x1), Z(x2), Z(x3)* jne. Ülesanne on prognoosida

muutuja väärtus kohas *xn*, kus mõõtmisi ei ole tehtud. Selliselt püstitatud ülesandel ei ole ühest lahendit, täpsemalt – võimalikke lahendeid on lõpmata palju. Üks võimalus kõige usaldusväärsema lahendi leidmiseks on lisada ülesandesse mingi mudel, mis kirjeldab muutuja ruumilist varieeruvust.

Ruumilise varieeruvuse mudel võib olla deterministlik või stohhastiline (statistiline). Mõlemal juhul tuleb eeldada, et prognoosiga kaasneb teatud juhuslikkus. Mitte, et muutuja väärtus kohas *xn* oleks juhuslik, vaid meie teadmised selle koha kohta on ebakindlad ja igasuguste mõõtmistega kaasneb teatud mõõtmisviga. Üks teoreetiline võimalus oleks käsitleda mõõdetud suurusi *Z(x1), Z(x2), Z(x3)* juhuslike muutujatena ja leida nende ja *Z(xn)* ühisjaotusest *Z(xn)* väärtuste tinglik ootus. Tinglik tingimusel, et *Z(x1), Z(x2), Z(x3)* väärtus on konkreetne teadaolev suurus. Paraku aga ei ole võimalik muutuja erinevates kohtades mõõdetud väärtuste ühisjaotust kasutada, sest iga koha kohta on vaid üks mõõtmistulemus ja koha *xn* kohta ei ole üldse mõõtmistulemusi. Seega, ühisjaotuste analüüsile tugineva stohhastilise mudeli kasutamiseks oleks vaja teha samas kohas kordusmõõtmisi.

Geostatistiline analüüs koosneb järgmistest sammudest.

- Varieeruvuse ruumilise struktuuri modelleerimine kas korrelogrammi, kovariatsiooni funktsiooni või variogrammi abil. Otsitakse ruumilist varieeruvust kõige paremini kirjeldav funktsioon.
- Variogrammi mudeli parameetrite määramine. Muuhulgas määratakse mudeli kasutamise ulatus, kuna piiratud ulatus vähendab tunduvalt arvutuste mahtu ja teatud kaugusest edasi on mudeli prognoosiv võime väike. Eeldatakse, et prognoosi arvutamisel teatud ruumipunkti kohta tasub kasutada eelkõige selle koha lähedal tehtud vaatlusi.
- Prognoosi arvutamine nende ruumiosade jaoks, kus otseseid mõõtmisi ei ole tehtud.

Geostatistika meetoditest annavad ülevaate Isaaks ja Srivastava (1989), Cressie (1993), Rossi *et al.* (1992). Ülevaateartikli geostatistika ökoloogilistest ja biogeograafilistest rakendustest on kirjutanud Kent *et al*. (2006), geostatistika kasutusest mullateaduses Goovaerts (1999). Geostatistika meetodeid kasutatakse tänapäeval väga mitmesuguste prognoosikaartide koostamisel (ilmaprognoosid, riskikaardid), ruumiliste nähtuste varieeruvuse, autokorrelatsiooni ulatuse ja suuna modelleerimisel (klimatoloogias, mullateaduses, epidemioloogias, taimekaitses) ja võimalike paiknemismustrite, üksikväärtuste või väärtuspindade genereerimiseks.

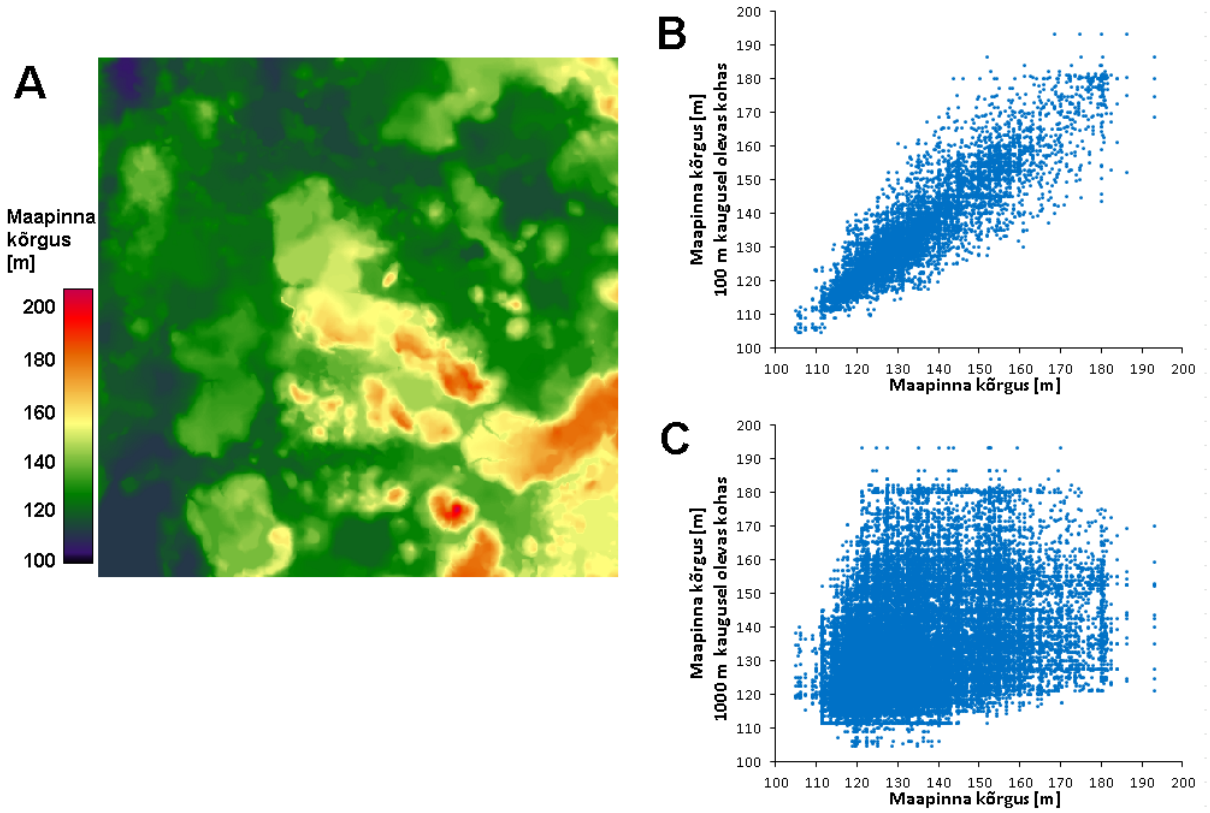
Eestis on geostatistika meetodeid kasutatud näiteks klimatoloogias ja tuuleenergia kaardistamisel (Olev ja Kull 2004, Jaagus ja Kull 2011).

Tarkvara

Spetsiaalselt geostatistika arvutusteks on loodud tarkvarapaketid gstat (Pebesma ja Wesseling 1998, Pebesma 2004, http://www.gstat.org) ja SGeMS (*Stanford Geostatistical Modeling Software*) (Remy *et al*. 2009, http://sgems.sourceforge.net). Kriging interpoleerimise vahendeid on tarkvaras ArcGis ja Surfer.

5.3.1. Autokorrelatsiooniväli

Autokorrelatsiooniväli (*h-scatterplot*) kaugustsoonis on teatud vahemaa ja paarivahelise suunaga vaatluste väärtuste punktdiagramm (joonis 5-11). Autokorrelatsiooniväli näitab etteantud vahemaa ja suunaga väärtuste korreleeruvust, aga ka võimalikke erindeid ja sümmeetrilisust ette antud suuna suhtes. Mida lähedasem on trend üksühesele seosele, seda tugevam on tunnuse autokorreleeritus. Asümmeetria võib näidata kohalikku trendi, mis võib anda näiliselt kõrge autokorrelatsiooni hinnangu.



Joonis 5-11. A – maapinna kõrgus [m] kaardilehel 54344 (Otepää); B – kõrguste autokorrelatsiooniväli kohtades vahemaaga 100 m ja C – vahemaaga 1000 m. Kaardilehe külje pikkus on 5 km. Sajameetrise vahemaaga kohtades on maapinna kõrgus üsna sarnane võrreldes kõrguste varieeruvusega kogu kaardilehel, kaardilehe kõige kõrgemate tippude ümber ei ole 1000 m kaugusel kaardilehe madalama pinnaga alasid ja kõige madalamate kohtade ümber samas raadiuses ei ole kõige kõrgemaid tippe, kuid üldiselt on maapinna kõrguste sarnasus kilomeetrise vahemaa puhul väga nõrk.

5.3.2. Poolhajuvus

Teatavasti on numbrilise muutuja hajuvuse põhiline mõõdik dispersioon ehk keskmine ruuthälve keskväärtuse suhtes. Ühe ja sama tunnuse ruumilise varieeruvuse kirjeldamiseks tuleb võrrelda väärtusi mingi vahemaaga paiknevates vaatluspaarides. Dispersiooni arvutamisel vaatluspaarides on iga vaatlus teise paarilise suhtes korra lähteobjektiks ja korra sihtobjektiks. Paaride arv avaldub kombineerimise puhul üksikvaatluste arvust (*N*) kujul *N(N−1)*. Kuna ruuthälbed arvutatakse ka paaridesisese erinevusena ja iga vaatlus on korra paari esimeseks liikmeks ja korra teiseks liikmeks, siis jagatakse paaridest arvutatud ruuthälvete summa kahega ja dispersioonile analoogilist varieeruvuse mõõtu nimetatakse **semivariatsiooniks** ehk **semidispersiooniks** (*semivariance*) ehk **poolhajuvuseks**. Poolhajuvus avaldub kujul

$$\gamma = \frac{1}{2N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(z_i - z_j\right)^2, \qquad [5\text{-}23]$$

kus *i* ja *j* on vaatluskoha indeksid ning *z* on vaatluskohas mõõdetud väärtus.

Kui dispersioon on keskmine ruuthälve üldkeskmise suhtes, siis semidispersioon näitab keskmist ruuthälvet vaatluspaarides ehk kõrvalekallet erinevuse puudumisest vaatluste vahel. Poolhajuvus on lokaalne näitaja ja selle kasutamisel ei eeldata keskmise ja dispersiooni statsionaarsust, nagu

autokorrelatsiooni mõõtva Morani *I* arvutamisel (ptk 5.1.2.2). Kui autokorrelatsioon võib olla nii positiivne kui ka negatiivne, siis poolhajuvus saab olla vaid positiivne. Geostatistikas ei ole tavaks kontrollida poolhajuvuse statistilist olulisust.

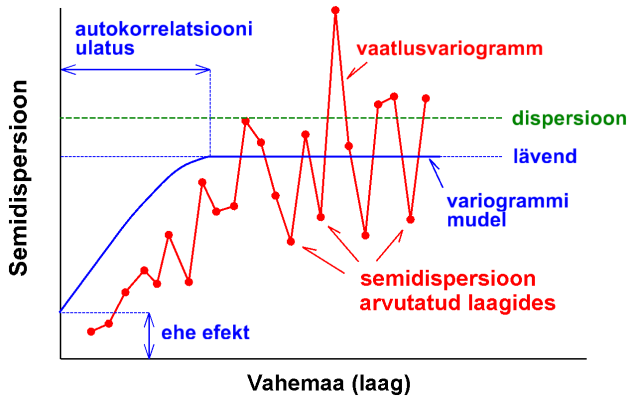
Poolhajuvus on pool keskmisest ruuthälbest teatud vahemaaga vaatluste paarides

5.3.3. Variogramm (semivariogramm)

Variogramm on ruumilise autokorrelatsiooni kirjeldamise vahend, mis näitab mõõtmistulemuste varieeruvuse sõltuvust vaatluste omavahelisest paarikaupa vahemaast (joonis 5-12). Võib ka ütelda, et variogramm on graafik, mis kujutab andmete varieeruvuse jagunemist kaugustsoonidesse. Kuna variogrammil kujutatakse tavaliselt poolhajuvust ehk semivariatsiooni, siis nimetatakse variogrammi ka semivariogrammiks. Kui poolhajuvuse arvutamisel ei ole kasutatud mitte ruuthälbeid, vaid lineaarhäbeid, saadakse **madogramm**. Kui varieeruvust arvutatakse vaatluspaarides, mille üks liige on üks muutuja ja teine liige teine muutuja, saadakse **ristvariogramm** (*cross variogram*). Ristvariogramm kirjeldab tunnustevahelist ruumilist korrelatsiooni, mitte autokorrelatsiooni.

Variogramm kujutab vaatluspaari väärtuste varieeruvuse sõltuvust vahemaast

Tase, millel variogramm suurematel kaugustel stabiliseerub, kannab nime **lävend** (*sill*). Variogrammi (autokorrelatsiooni) **ulatus** (*range*) on vahemaa, mille juures semidispersioon stabiliseerub lävendi tasemele (joonis 5-12).



Joonis 5-12. Vaatlusvariogramm ja variogrammi mudel.

Variogrammi kujutatakse vähemalt lävendi alguseni ja tavaliselt mitte kaugemale poolest vaatlusala suurusest. Variogrammi mudeli lävend näitab mudeli poolt kirjeldatud varieeruvuse osa. Alati ei stabiliseeru variogramm konstantse lävendini, vaid läheneb sellele asümtootiliselt. Sellisel juhul on variogrammi ulatust raske määrata – ulatuseks võib võtta kauguse, kus mudeli lävend siseneb asümptootilise lävendi 95% usalduspiiridesse. Kui poolhajuvused ei stabiliseeru suurematel kaugustel, siis kasutatakse isevaldset (*arbitrary*) ulatust.

Variogrammi funktsiooni väärtust nullkaugusel (laag = 0) nimetatakse **ehedaks variatsiooniks** (*nugget* – ingl. k. maagi- või kullatükk) ehk ehedaks efektiks. Ehe variatsioon on mudelijärgne varieeruvus samast kohast pärinevate vaatluste vahel. Selle põhjusteks on mõõtmisvead, andmete puudumine vaatluste varieeruvuse kohta väikestel vahemaadel ja uuritava nähtuse ebastabiilsus

(johtuvalt näiteks vabalt liikuvate organismide liikuvusest). Ehedat variatsiooni nimetatakse ka valgeks müraks. Andmete kogumisel väljendub ehe efekt selles, et kordusmõõtmised samast kohast ei anna täpselt sama tulemust. Variogrammi konstantsust kõigi vahemaade puhul nimetatakse puhtaks ehedaks variatsiooniks (*pure nugget behaviour*). See viitab autokorrelatsiooni puudumisele andmetes.

Variogrammid võivad olla **igasuunalised** (*omnidirectional*) või **kindlasuunalised** (*directional*). Kindlasuunaline variogramm aitab kujutada suunaga ruumilisi struktuure ja igasuunaline kujutab ruumilist varieeruvust suunast sõltumatult. Ruumiliste struktuuride erinevust eri suundades nimetatakse anisotroopsuseks ehk **anisotroopiaks** (*anisotropy*). Sama ulatusega aga erinevate lävenditega kindlasuunalised variogrammid samadest andmetest viitavad **tsonaalsele anisotroopiale** (*zonal anisotropy*), sarnased lävendid koos suundades erinevate ulatustega näitavad **geomeetrilist anisotroopiat** (*geometric anisotropy*). Arvutuslikult suhteliselt lihtne on vaid ellipsoidja kujuga anisotroopia.

Erinevatest andmestikest pärit variogrammide kuju paremaks võrdlemiseks standardiseeritakse neid andmete dispersiooniga. Standardiseeritud variogramme saab ühel graafikul kujutada isoliinidega. Rossi *et al.* (1992) on selliseid graafikuid nimetanud **võrdlusvariogrammiks** (*exhaustive variogram*). Võrdlusvariogrammiga saab kujutada näiteks õhutemperatuuri ruumilist pidevust eri aastaaegadel.

Kaheväärtuselise muutuja analüüsi geostatistika vahenditega nimetatakse **indikaator-geostatistikaks** (*indicator geostatistics*). Kaheväärtuseliseks saab teisendada igasuguseid tunnuseid. Nominaalse tunnuse puhul on kaheks väärtuseks mingi väärtusklassi esinemine ja puudumine, arvulisi väärtusi saab alati mingi tasemega kahte klassi jagada.

5.3.4. Variogrammi mudel

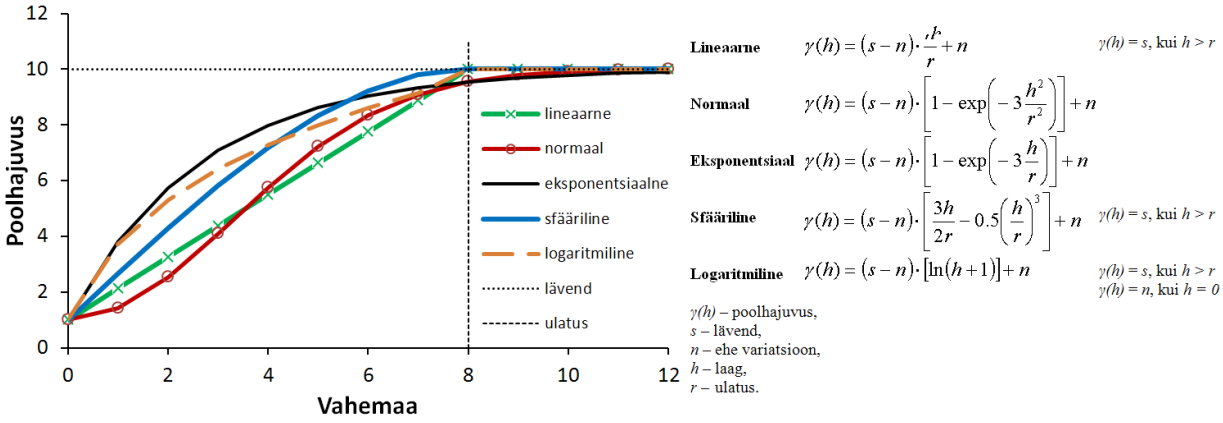
Empiirilised andmed annavad empiirilise semivariatsiooni jaotuse ehk **vaatlusvariogrammi** (*sample variogram*), kus poolhajuvused on esitatud kaugusklasside kaupa. Empiirilisele poolhajuvuse jaotusele saab sobitada mingit teoreetilist funktsiooni ehk **variogrammi mudeli**. Variogrammi mudel ei ole reeglina enam laagideks jagatud, vaid pidev. Nii variogramm kui ka korrelogramm on globaalsed multidistantsed ruumilise autokorrelatsiooni mõõdud. Kui variogrammi funktsioon on mingil kaugusel suhteliselt väike, siis võib eeldada, et selle vahemaaga punktid on suhteliselt sarnased.

Variogrammi mudelid sisaldavad parameetreid, mis tuleb andmestikule sobitada. Lähtudes vigade normaaljaotuse eeldusest, toimub statistiliste mudelite sobitamine ruuthälvete minimeerimise teel. Variogrammi jaoks arvutatakse poolhajuvus kaugustsoonide (h) kaupa vastavalt vahemaale lähtekoha x_i ja sihtkoha x_j vahel. Järgnevas valemis tähistab N_h vaatluspaaride arvu, mille vahemaa on kaugusvahemikus h. Tärn märgib, et tegemist on empiiriliste andmete järgi hinnatud poolhajuvusega.

$$\gamma^*(h) = \frac{1}{2N_h} \sum_{(i,j)\in P_h}^{N_h} \left(z_i - z_j\right)^2, \qquad [5\text{-}24]$$

Sagedamini kasutatavad variogrammi mudelid on **eksponentsiaalne** mudel, mille erikuju on lineaarne mudel astmega üks. Kui aste on <1 siis on variogrammi mudel kumer, kui aste on >1, siis nõgus. **Sfääriline** mudel annab kumera mõjusfääri, normaaljaotuse funktsioon tagab sujuvad üleminekud, **logaritmiline**, sfääriline ja lineaarne mudel nõuavad eraldi funktsiooni variogrammi ulatuse piires ja väljaspool seda. Logaritmiline mudel eeldab nullvahemaade välistamist. Kasutada saab ka variogrammi mudelite positiivseid lineaarkombinatsioone. Enamikel juhtudel sisaldab variogrammi mudel ehedat efekti pluss vähemalt ühte funktsiooni (joonis 5-13).

Kui variogramm on uuritava ala erinevates osades erinev, tuleks neid territooriume eraldi modelleerida, sest variograafia eeldab ruumilise varieeruvuse konstantsust. Ruumilise trendi korral tuleks see eelnevalt eemaldada.



Joonis 5-13. Variogrammi mudelid graafikute ja valemite kujul.

5.3.5. Kriging

Termin **kriging** pärineb D.G. Krige nimest. Meetodit arendati algselt Lõuna-Aafrika kullakaevandusi omavas firmas ja Lõuna-Aafrika Transvaali ja Oranje kaevanduste ametis, sest seniste interpoleerimismeetodite ja trendpindade arvutus ei andnud piisavalt täpseid kulla sisalduse prognoose. D.G. Krige (1966) ise nimetab meetodit **kaalutud liikuvaks keskmiseks** (*weighted moving average*). Kriginguga sisuliselt samatähenduslik on meteoroloogias arendatud optimaalne interpoleerimine Eliassen (1954), Гандин (1963).

Krige (1966) jagab oma meetodi maavarade otsimise kontekstis viieks järjestikuseks sammuks.

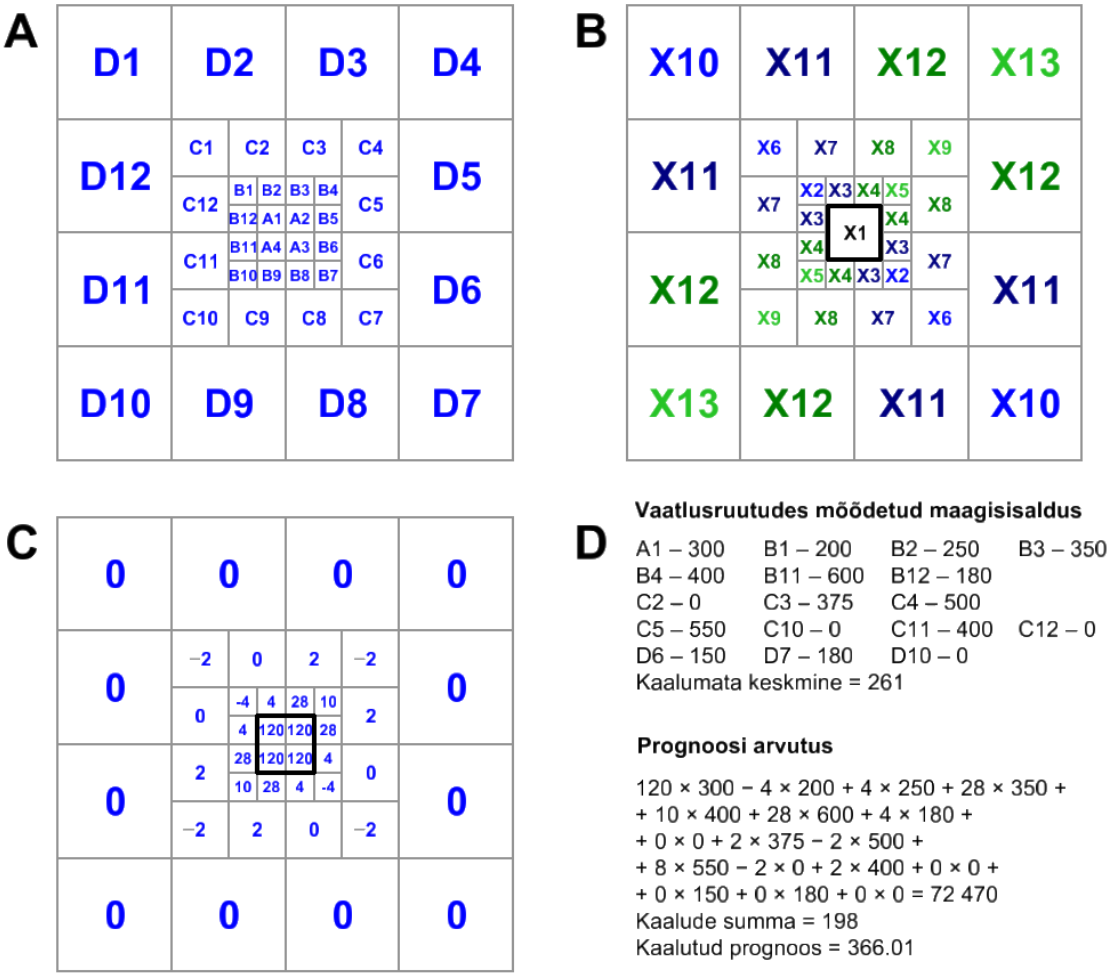
- kasutatavate andmeüksuse suuruse ja konfiguratsiooni määramine,
- kasutatavate hindamisüksuste suuruse valimine,
- interpoleerimiskaale määravate seoste modelleerimine,
- lokaalse kaalutud keskmise arvutamine kasutades regressioonidest saadud kaale,
- usalduspiiride arvutamine kaalutud keskmistele.

Krige (1966) ei seostanud vaatlusandmete varieeruvust otseselt vahemaaga vaatluskohtade vahel ja ei modelleerinud varieeruvust vahemaast sõltuva pideva funktsiooni abil, vaid kasutas naabrust esindavatele ruutudele omistatud kaale ja nende kaaludega kaalutud keskmist (joonis 5-14). Interpoleerimiskaalud määras Krige proportsionaalseks iga vastava ruudu andmete alusel arvutatud logaritmitud hinnangute keskmise veaga. Regressioonimudelite koostamisel kasutati prognoositavast vaatlusruudust mõõdetud muutujat funktsioontunnusena ja seda ümbritsevate ruutude andmeid argumenttunnustena, kusjuures prognoositava ruudu suhtes samal kaugusel, aga vastasasendis olevate ruutude andmed ühendati (joonis 5-14B). Kaalud arvestavad tüüpiliste ruumiliste struktuuride suunda.

Tänapäeval kirjeldatakse paiknemist sõltuvat varieeruvust variogrammiga ja krigingut käsitletakse eelkõige interpoleerimise meetodina. Krigingut saab käsitleda lokaalse vähimruutude autoregressioonina, see tähendab, et interpoleerimishinnang arvutatakse sama muutuja ümbruses olevate väärtuste kombinatsioonina, kasutades variogrammi mudelist saadavaid regressioonikordajaid. Kriging on seega varieeruvuse ruumilise modelleerimise, interpoleerimise ja ruumilise prognoosi meetod. Prognoosimeetod, sest interpoleerimine tähendab alati väärtuste hindamist punktides, kust mõõtmistulemusi ei ole, seetõttu on interpoleerimine ka prognoosiks. Peale selle võimaldab kriging ka variogrammi ulatuses ekstrapoleerimist, mis on samuti prognoos.

Kriging on interpoleerimine andmepunktide suhtelisest asendist sõltuvate kaaludega, mis tuletatakse andmete varieeruvusest selle asendi puhul

Krigingu käigus moodustatakse variogrammi andmete põhjal teatud suuruse, kuju ja kaaludega aken. Akna suuruse määrab variogrammi ulatus, akna kuju on igasuunalise mudeli puhul ring, suunaga variogrammi puhul ringist erinev. Interpoleerimine toimub libiseva akna meetodil, kusjuures akna ulatusse jäävatele vaatlustele omistatakse kaal vastavalt variogrammi funktsioonile. Nagu libiseva akna puhul ikka, võib kriging anda prognoosi akna ulatuses väljapoole empiiriliste andmete piirkonda. Samuti võib kriging anda empiiriliste andmetega hõredalt kaetud piirkonnas ootamatuid ja väheusaldatavaid tulemusi. Kui mingist uuritava ala osast andmed täielikult puuduvad, siis on selle osa kohta usaldusväärset prognoosi raske saada.



Joonis 5-14. A – erineva suurusega vaatlusruudud. B – naabrusandmed keskmises ruudus oleva maagisisalduse seletavate tunnustena. C – regressioonimudelist saadud prognooside logaritmilise vea järgi tuletatud interpoleerimiskaalud. D – kaalutud keskmisena arvutatud prognoosi arvutuskäik keskmise ruudu kohta Krige originaalpublikatsiooni järgi (Krige 1966, muudetud).

5.3.5.1. Tavakriging

Tavakrigingu ehk punktkrigingu (*ordinary kriging*, *point kriging*) puhul on lähteandmeteks uuritava tunnuse vaatluspunktides mõõdetud väärtused ja hinnatakse uuritava tunnuse oodatavat keskmist kogu uuritaval alal, arvestades eelnevalt modelleeritud ruumilist autokorrelatsiooni. Mõnikord eristatakse tavakrigingust **lihtkriging** (*simple kriging*), mille puhul ümbruskonna väärtuste mitteteadmisel ehk nullmudeli kehtimise korral on väärtuste ootus võrdne nulliga.

Tavakriging annab ümbruses olevate väärtuste kaalutud keskmise – kaaludeks on vahemaast sõltuvad väärtused variogrammi mudelis

Tavakrigingu puhul eeldatakse, et normaaljaotusega muutujal Z on kindel, kuid meile teadmata väärtus igas uuritava piirkonna punktis. Uuritaval alal on muutujat mõõdetud n korda kohtades x_i ja on saadud vigadeta mõõtmistulemused z'_i. Eesmärgiks on leida muutuja Z parim lineaarne nihketa hinnang (*best linear unbiased estimate – BLUE*) iga ruumipunkti j jaoks. See on parim ruuthälvete minimeerimise, lineaarne lineaarse autoregressioonimudeli kasutamise (hinnangud on vaatluste kaalutud lineaarkombinatsioonid) ning hälbeta prognooside ja empiiriliste vaatluste keskväärtuste kokkulangevuse mõttes.

Tavakriging eeldab interpoleeritava muutuja normaaljaotust ning selle väärtuste ootuse ja hajuvuse ruumilist statsionaarsust

Kuna iga ruumipunkti puhul lahendatakse ülesanne uuesti, siis tähistatakse seda ruumipunkti, mille jaoks parajasti hinnangut arvutatakse, indeksiga 0. Muutuja väärtuse hinnang (z'_j) kohas j määratakse ümbritsevate vaatluste (z_i) kaalutud keskmisena selle koha ümber

$$z'_0 = \sum_{i=1}^{n} w_i z_i \qquad [5\text{-}25]$$

Vaatluse i kaal w_i sõltub korrelatsiooni tugevusest sellise omavahelise paiknemisega (igasuunalise mudeli puhul vahemaaga) kohtade vahel nagu on interpoleeritaval punktil (fokaalpunkt indeksiga 0) ja vaatlusel (i). Tavakrigingu puhul võrdsustatakse kaalude summa alati ühega.

$$\sum_{i=1}^{n} w_i = 1 \qquad [5\text{-}26]$$

Kaalude määramiseks koostatakse kõigepealt vaatlustevaheliste vahemaade maatriks ja variogrammi mudel. Ruumilised vahemaad teisendatakse variogrammi mudeli järgi vaatlustevahelisteks oodatavateks korrelatsioonideks, mis iseloomustavad vaatlustulemuste keskmist sarnasust selle asendi puhul.

$$\rho_h = \frac{C - \gamma_h}{C}, \qquad [5\text{-}27]$$

kus ρ_h on oodatav korrelatsioon, γ_h on variogrammi mudeli väärtus laagi h puhul ja C on variogrammi mudeli lävend (*sill*).

Iga vaatluspunkti mõju määratakse igas arvutatavas kohas (reeglina moodustavad need kohad regulaarse võrgustiku) eraldi kasutades lineaarsete võrrandite süsteemi

$$\begin{vmatrix} \rho_{11} & \rho_{12} & \rho_{13} & \cdots & \rho_{1n} & 1 \\ \rho_{21} & \rho_{22} & \rho_{23} & \cdots & \rho_{2n} & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{n1} & \rho_{n2} & \rho_{n3} & \cdots & \rho_{nn} & 1 \\ 1 & 1 & 1 & \cdots & 1 & 0 \end{vmatrix} \cdot \begin{vmatrix} w_1 \\ w_2 \\ \cdots \\ w_n \\ \mu \end{vmatrix} = \begin{vmatrix} \rho_{10} \\ \rho_{20} \\ \cdots \\ \rho_{n0} \\ 1 \end{vmatrix}, \qquad [5\text{-}28]$$

kus ρ_{ij} on variogrammi mudeli järgi oodatav korrelatsioon punkti *j* ja punkti *i* vahemaale vastava laagi korral, 0 on parajasti hinnatava koha indeks, 1;...; *n* on teadaolevate väärtustega vaatlused, μ on maatriksis olev muutuja ja Lagrange kordaja ehk määratlemata kordaja, mille lisamine on vajalik kaalude summa võrdsustamiseks ühega ja hälvete minimeerimiseks. Lagrange kordajale omistatakse väärtus, mis tagab minimaalsed hälbed ja kaalude ühega võrdse summa.

Tähistades variogrammi mudelist tuletatud vaatluskohtade korrelatsioonide maatriksit R, kaalude vektorit w ja vaadeldava punkti ülejäänud punktidega eeldatava korreleerumise vektorit D, saab eeltoodud võrrandi kirjutada kujul

$$\mathrm{Rw} = \mathrm{D}, \qquad [5\text{-}29]$$

millest kaalud avalduvad kujul

$$\mathrm{w} = \frac{\mathrm{D}}{\mathrm{R}}. \qquad [5\text{-}30]$$

Variogrammi mudelist ja vahemaast sõltuvate mõjukaaludega arvutatakse interpoleeritava väärtuse hinnang z'_0 iga võrgustiku punkti jaoks ümbritsevate väärtuse kaalutud keskmisena. Seejuures maatriks D väljendab interpoleeritava punkti kaugust vaatluspunktidest ja maatriks R väljendab vaatluste omavahelist grupeerumist. Lähestikku paiknevate vaatluste mõju vähendatakse, kuna need võivad olla pseudoreplikatsioonid.

Kaugusest sõltuvad kriging-kaalud arvutatakse eraldi igas interpoleeritavas kohas

Kriging on väga arvutusmahukas meetod mõõdetud kohtade ja arvutatavate kohtade suure arvu korral, sest kaalud arvutatakse igas ruumipunktis uuesti. See on üks põhjus variogrammi ulatuse piiramiseks. Paraku muutub krigingu lahend vaatluste ebaühtlase paiknemise ja väikese variogrammi ulatuse korral ebastabiilseks. Kui variogrammi ulatuses on vaid üks vaatlus, siis ei saa kaale üldse arvutada. Üldiselt soovitatakse, et hinnangu arvutamiseks kasutataks vähemalt 15–20 naabervaatlust. Vaatlustega hõredalt kaetud piirkondades võib selle saavutamiseks naabrite otsimisraadiust suurendada.

Kriging on interpoleerimismeetod, mille abil prognoositud pind läbib vaatluspunkte. Kui variogrammi mudel sisaldab ehedat varieeruvust, kuid variatsioonimaatriksi peadiagonaalil on nullväärtused, teeb interpoleeritud pind vaatluspunktide juures hüppeid. Need hüpped on seda suuremad, mida suurem on eheda varieeruvuse osa ja mida ebatüüpilisem (interpoleeritud pinna suhtes) on vaatlustulemus.

5.3.5.2. Teised krigingu variandid

Globaalkrigingu (*global kriging*) puhul kasutatakse iga ruumipunkti hinnangu arvutamisel kõiki mõõtmisi. Seetõttu on globaalkriging arvutus- ja ajamahukas. Tavakriging on globaalkrigingule vastandumise mõttes lokaalkriging.

Plokk-kriginguga (*block kriging*) määratakse etteantud piirkondade keskmised. Arvutustes ei kasutata mitte kohtadevahelisi erinevusi, vaid iga ploki ja iga koha vahelist erinevust. Muus osas on plokk-kriging punktkriginguga analoogiline. Oma teedrajavas publikatsioonis kirjeldab Krige just plokkide kaupa arvutust Krige (1966).

Kriging eraldiste piiridega (*kriging with detected edges*, *kriging with barriers*) interpoleerib vaid etteantud piirini, piirist teisel pool olevaid vaatluskohti ei arvestata. Näiteks puidu tagavara hindamisel ei ole mõistlik kasutada mõõtmisi, mis paiknevad teisel pool kaugseireandmetest saadud metsatüübi muutumisjoont (Wallerman *et al*. 2002).

Indikaatorkriging (*indicator kriging*) jagab interpoleeritava muutuja väärtusvahemikesse ja arvutab tõenäosuse, et muutuja on etteantud väärtusest suurem või väiksem. Indikaatorkriging ei nõua muutuja normaaljaotust ja sobib ka kaheväärtuselise tunnuse kaardistamiseks.

Kokriging (*cokriging*) on tavakrigingu mitmetunnuseline variant, mille puhul interpoleeritav väärtus ei sõltu mitte ainult sama muutuja väärtustest ümbruses, vaid ka mingi teise muutuja (lisatunnuse, argumenttunnuse) väärtusest interpoleeritava koha ümbruses olevates vaatluskohtades. Kaugusest sõltuvaid kaale rakendatakse nii ümbruses olevatele sama muutuja väärtustele kui ka lisatunnuse või tunnuste väärtustele. Näiteks sademete hulga kaardistamisel on kokrigingu lisatunnusena kasutatud koha topograafilisi tunnuseid (Diodato *et al*. 2010, Portalés *et al*. 2010).

Välise nihkega kriging (*kriging with external drift*) ehk kriging lokaalselt muutuva keskmisega (*kriging with varying local means*) ehk empiiriline parim lineaarne hälbeta hinnang (*empirical best linear unbiased predictor – E-BLUP*) on kokrigingu variant, mille puhul lisatunnuse väärtused on teada ülepinnaliselt. Igas interpoleeritavas kohas kasutatakse vaid lisatunnuse selle koha väärtust. Kokrigingu puhul on lisatunnused teada vaid vaatluskohtades.

Regressioonkriging (*regression/kriging*, *regression-kriging*) on arvutuslikult esmalt funktsioon-tunnuse prognoos regressioonimudeli abil ja seejärel regressioonijääkide kriging-interpoleerimine. Interpoleerimisest eraldiolev prognoosimudel võimaldab kasutada kõivõimalikke keerukaid prognoosimeetodeid.

Universaalkriging (*universal kriging*) ehk **trendiga kriging** (*kriging with a trend model*) ei nõua interpoleeritava muutuja ootuse ruumilist statsionaarsust, seega võivad andmed sisaldada ruumilist trendi. Trendiga krigingu puhul ei sõltu prognoositav väärtus mitte ainult ümbruses olevatest väärtustest, vaid ka asukoha koordinaatidest. Trendiga kriging on välise nihkega krigingu erijuhtum, mille puhul lisatunnusteks on ruumikoordinaadid.

Ajalis-ruumiline kriging (*spatio-temporal kriging*) on tavakrigingu edasiarendus, mis arvestab nii ruumilist kui ka ajalist vahemaad vaatluste vahel, tuginedes ajalis-ruumilisele variogrammile (Kyriakidis ja Journel 1999). Ühemõõtmelise variaogrammi asemel kasutatakse aeg-ruumilist variogrammi ning prognoositud väärtuste aluseks on nii ajaliselt kui ka ruumiliselt lähedased vaatlustulemused.

5.3.5.3. Krigingu omadused

Krigingu omadused on kokkuvõtlikult järgmised.

- Kriging üritab olla vähimruutude mõttes parim lineaarne hälbeta hinnang.
- Kriging interpoleerib moodustades interpoleeritud pinna, mis läbib andmepunkte.
- Tavakriging eeldab lähteandmete normaaljaotust.
- Kui kasutataval mudelil on ehe variatsioon võrdne nulliga, siis saadakse sujuv pind. Kui ehe variatsioon on mudeli järgi suurem kui null, siis teeb interpoleeritud pind vaatluspunktide kohas jõnkse.
- Mida suurem on mudeli järgne ehe variatsioon, seda vähem varieerub interpoleeritud pind üldkeskmise kõrgusel tasapinna ümber.
- Kriging arvestab interpoleerimisel ruumilise autokorrelatsiooni ulatust ja tugevust.
- Kriging ei nõua andmete ühtlast paiknemist.
- Kriginguga saab hinnata nii uuritava muutuja lokaalseid väärtusi kui ka piirkondade keskmisi.
- Kriging (välja arvatud universaalkriging) eeldab trendi puudumist nii keskmises kui varieeruvuses ehk dispersiooni ja matemaatilise ootuse ruumilist statsionaarsust. Kui trend esineb, siis tuleb see modelleerida ja kasutada interpoleerimisel hälbeid trendist. Samuti võib trendi eemaldamiseks kasutada peakomponentanalüüsi telgi (Hinch *et al.* 1994).
- Kriging prognoosib vaid oodatavat keskväärtust, mitte oodatavate väärtuste varieeruvust. Seetõttu saab krigingu abil teha trendpinna kaarti, aga mitte oodatavate väärtuste realistlikku kujutist.

5.3.5.4. Kriging-interpoleeringu verifikatsioon

Interpoleeritud pinna tõesuse kontrollimiseks saab kasutada ristkontrolli (ptk 3.6.2), mille puhul eemaldatakse osa andmeid valimist ja tehakse selle osaga eraldi kontrollarvutused. Krigingu puhul tähendab see kõigi vaatluspunktide ajutist ükshaaval eemaldamist, iga variandi interpoleerimist ja eemaldatud vaatluse koha jaoks arvutatud interpoleeritud tulemuse võrdlemist tegeliku mõõtmistulemusega. Võrdluseks kasutatakse keskmist ruuthälvet (Isaaks ja Srivastava 1989). Mida väiksem on keskmine ruuthälve, seda usaldusväärsem on krigingu tulemus. Nii saab samade lähteandmete alusel võrrelda erinevaid interpolatsioonimeetodeid (Goovaerts 2000).

Krigingu käigus püütakse minimeerida ruutviga. Kriging-interpolatsiooni jääkide analüüsil tasub meeles pidada, et krigingu tulemused sõltuvad variogrammi mudelist ja enamasti osutavad vaid vajadusele saada hõredalt uuritud piirkondadest täiendavaid andmeid.

Kriging-interpolatsiooni tulemust saab kontrollida ka genereerides korduvalt juhuslikke interpolatsioonipindu ja lisades variogrammi mudeli järgse eheda juhusliku hajuvuse prognoositava pinna igas kohas interpoleeritud väärtusele. Juhusliku komponendiga interpolatsioonipindu korduvalt moodustades saadakse igas punktis lisaks parimale hinnangule ka selle hinnangu oodatav jaotus eheda juhusliku poolhajuvuse korral. Sellest omakorda on lihtne esitada interpolatsiooni usalduspiire.

Uurimused

Pinnavormide korrapära ja mulla omaduste ruumilist struktuuri on variogrammidega kirjeldanud näiteks Oliver ja Webster (1986), Western *et al.* (1998), van Horssen *et al.* (1999). Odeh ja McBratney (2000) võrdlesid mulla savisisalduse kaardistamise meetodeid kasutades väga kõrge lahutusega radiomeetri andmeid ja said parima tulemuse mitmetunnuselise lineaarse regressiooni ja kriging-interpoleerimise kombinatsioonis (*regression/kriging*).

Gething (2006) täpsustas ajalis-ruumilise krigingu abil malaaria levikut Aafrikas.

Walker *et al*. (2008) kaardistasid tavakrigingu ja indikaatorkrigingu abil mõnede linnuliikide esinemistõenäosust. Mõlemad meetodid andsid sarnase tulemuse.

5.3.6. Variogrammile tuginev klassifitseerimine

Variogrammi (ja korrelogrammi) (ptk 5.1.2.5; ptk 5.3.3) eelis teiste struktuuristatistikute ees on komplekssus. Variogramm ühendab kujutise tekstuuri kahte peamist komponenti: lokaalset varieeruvust ja väärtuste paiknemist üksteise suhtes. Variogrammile tugineva tekstuuri klassifitseerimise ideed populariseerisid 1990ndatel F.P. Miranda ja J.R. Carr (Carr 1996, Miranda ja Carr 1994, Miranda *et al*. 1992, 1996). Miranda ja Carri metoodika järgi kasutatakse liikuva akna sees poolhajuvuse hindamiseks väiksemat liikuvat ristkülikut ehk kernelit, mis valib välja pikslid, mille vahemaa ei ületa ristküliku diagonaali pikkust. Poolhajuvus arvutatakse kuni vahemaani, mis võrdub kerneli lühema külje pikkus –1. Seega 7 × 7 piksli suuruse kerneli puhul arvutatakse 6 poolhajuvust, mida kasutatakse otsitavate üksuste äratundmiseks. Kuna poolhajuvused on väga tundlikud üksikute erandliku väärtusega pikslite suhtes, on soovitatud ruumilist varieeruvust mõõta keskmise pikslipaaride erinevuse ruutjuure abil (Cressie and Hawkins 1980, Lark 1996), mis on erindite suhtes vähemtundlik.

Empiirilistest andmetest hinnatud poolhajuvuste asemel on tekstuuride erinevuse hindamiseks kasutatud ka variogrammi mudelite võrdlemist (Herzfeld 1993, Wallace *et al.* 2000). St-Onge ja Cavayas (1995) seostasid metsa struktuuri parameetreid suunaga variogrammi parameetritega. Ülevaateartikli kujutise geostatistilise klassifitseerimise meetoditest on kirjutanud Atkinson ja Lewis (2000). Variogrammide kasutamise peamiste probleemidena tekstuuri eristamisel nimetavad nad tekstuuri sõltuvust kujutise detailsusest ja variogrammi mudelite automaatse vaatlustele lähendamise vähest usaldatavust.

Uurimused

Variogrammi kasutatavust erineva suurusega objektide äratundmisel satelliidipildilt on näidanud Lacaze *et al*. (1994), Chica-Olmo ja Abarca-Hernández (2000), de Bruin (2000). Metsa struktuuri parameetreid on variogrammi abil hinnanud Woodcock *et al*. (1988a), St-Onge ja Cavayas (1997), Lévesque ja King (2003).

Lark (1996) võrdles aerofoto klassifitseerimise täpsust erineva ruumilise varieeruvuse mõõdikute ja erineva suurusega lokaalsete akende puhul.

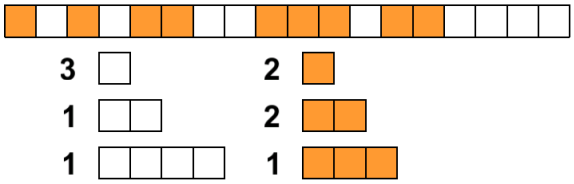
5.3.7. Mitme-punkti geostatistika

Mitme-punkti geostatistika (*multiple-point statistics, multi-point geostatistics – MPS*) on traditsioonilise geostatistika edasiarendus ja alternatiiv mittelineaarse varieeruvusega kolmemõõtmelisse maailma. MPS printsiibid on pärit 1990ndate algusest (Deutsch 1992). Kasutuskõlbulik arvutusalgoritm loodi ligi kümme aastat hiljem (Strebelle 2002). MPS on rakendust leidnud eelkõige geoloogias.

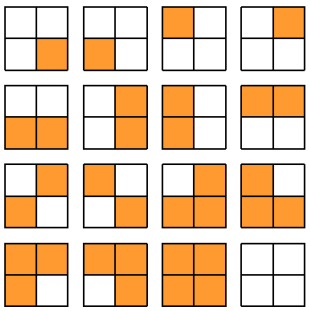
Traditsiooniline geostatistika tugineb variogrammi mudelile, mis kirjeldab varieeruvust kahe vaatluskoha vahel. Sellest termin kahe-punkti statistika. MPS ei loo variogrammi, vaid kasutab mitmes külgnevas punktis määratud väärtuskombinatsioone geoloogiliste struktuuride näidistena ehk templiitidena. Erinevates mõõtkavades näidised võimaldavad iseloomustada igasuguseid ruumilisi struk-

tuure. MPS tugineb treeningkujutisele (*training image*), mis on geoloogilise struktuuri andmebaas, millest tuletatakse mitme-punkti statistikud. Eeldatakse, et nende statistikute abil kirjeldatud struktuurid korduvad uuritavas ruumis ning et templiitide sagedus treeningkujutises esindab nende esinemistõenäost. Kui variograafia eeldas varieeruvuse statsionaarsust uuritaval alal, siis MPS eeldab struktuurset statsionaarsust ja õpetusandmete esinduslikkust.

Mitme-punkti statistiku näide ühemõõtmelisest ruumist on järjestikuste esinemiskordade arvu jaotus (*distribution of runs*) väärtuste reas (joonis 5-15). Kaks korda kaks piksli suuruse templiidi variante kahemõõtmelises ruumis on 16 (joonis 5-16). Iga templiidi jaoks saab treeningkujutisest leida esinemistõenäosuse ja teiste templiitidega kõrvuti paiknemise tõenäosuse.



Joonis 5-15. Sama kategooria järjestikused esinemised ja järjestuste sagedus.



Joonis 5-16. Kahe kategooria võimalikud paiknemiskombinatsioonid vaatlusala neljas alajaotuses (Boisvert *et al*. 2008).

Mittelineaarse ja mitmemõõtmelise varieeruvuse modelleerimiseks on tarvis suuremat hulka õpetusandmeid kui kahe-punkti geostatistikas. Tavapärased geoloogilised puurimised sellist andmehulka ja andmete ruumilist detailsust ei taga. Teiseks on geoloogilised andmed kohaspetsiifilised, see tähendab, et treeningkujutises olevad geoloogilised struktuurid ei pruugi mujal esineda. Üks lahendus on luua treeningkujutis modelleerides geoloogilisi protsesse.

Üksikmaardla realistliku mudeli loomisel võib parimaks meetodiks jääda ekspertarvamuse graafiline kujutis. Parima mudeli saamiseks on soovitatav kaasata mitmeid eriala eksperte.

Tarkvara

Mitme-punkti geostatistika vahendid on avatud koodiga tarkvaras SGeMS (*Stanford Geostatistical Modeling Software*, http://sgems.sourceforge.net).

5.4. Ruumiandmete kovariatsioon

5.4.1. Korrelatsioon väärtuspindade vahel

Statistilist seost kahe numbriliste väärtustega andmekihi vahel saab mõõta tavapäraste seosekordajate abil (ptk 1.5.2) või siis ruumilise seosena. Tavakorrelatsiooni arvutamisel andmekihtidest võrreldakse vaid samal positsioonil olevaid väärtusi ning väärtuste suhtelist asendit ei arvestata, kuigi mõnes tarkvarapaketis nimetatakse seda ruumiliseks korrelatsiooniks. Võrreldakse vaid kas kohakuti olevaid pikslivääruti või kohakuti olevaid eraldisi, asukohakoordinaate korrelatsiooni arvutuses muul moel ei arvestata. Võrreldavate andmekihtide väärtused võib ka tabelisse kanda nii, et võrreldavad kihid on eraldi veergudes ja kohakuti olevad väärtused on tabeli samal real ning siis unustada, et tegemist on ruumiandmetega. Kui arvutatakse korrelatsioon ühe tunnuse ja teise tunnuse vahel, mis on mõõdetud samades vaatluskohtades, siis ei erine see millegi poolest tavalisest korrelatsioonikordaja arvutusest. Sel juhul vaatluskoht on sama, mis vaatlus ja ruumiline aspekt ei oma tähendust.

Igat seost ruumiliste andmete vahel ei tuleks nimetada ruumiliseks seoseks

5.4.2. Ruumiline korrelatsioon

Ruumiline korrelatsioon arvutatakse vaid teatud vahemaaga vaatluste vahel (ruumiline korrelatsioon kaugustsoonis) või omistatakse vaatluspaaridele vahemaast sõltuvad kaalud. Esimesel juhul saadakse korrelatsioonikordaja väärtus iga kaugustsooni jaoks, teisel juhul saadakse üks vahemaast sõltuv tunnustevahelise ruumilise korrelatsiooni kordaja väärtus kogu andmestiku kohta. Ruumiline korrelatsioon on seega kahe muutuja korreleerumine teatud ruumilise vahemaaga vaatluspaarides. Need vaatluspaarid võivad sisaldada kas samal ajal mõõdetud erinevaid tunnuseid või näiteks sama muutuja väärtusi erineval (aasta)ajal. Ruumilise korrelatsiooni mõõtmine on üks osa ruumiliste seoste uurimisest.

Ruumilise korrelatsiooni arvutusesse on kaasatud vaatluskohtade vahemaa

Eristamaks kahe tunnuse vahelist korrelatsiooni autokorrelatsioonist ühe tunnuse sees, nimetatakse esimest ristkorrelatsiooniks. Ruumilist korrelatsiooni mõõdetakse ruumilise ristkorrelatsiooni kordaja (I_{YZ}) abil. I_{YZ} on Morani I (ptk 5.1.2.2) ja Pearsoni R (ptk 1.5.2.2) analoog.

$$I_{YZ} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(y_i - \bar{y})(z_j - \bar{z})}{\sum_{i \neq j} w_{ij} s_y s_z}, \qquad [5\text{-}31]$$

kus N on vaatluspaare moodustavate vaatluste arv, i on lähtekoht ja j on sihtkoht, kusjuures $i \neq j$, w_{ij} on vaatluste vahemaast sõltuv kaal, y_i on muutuja y väärtus kohas i, z_j on muutuja z väärtus kohas j, $\bar{y}$ on muutuja y keskmine, $\bar{z}$ on muutuja z keskmine, s_y on muutuja y standardhälve, s_z on muutuja z standardhälve.

Kui muutujad on eelnevalt tsentreeritud, siis on nende keskmised võrdsed nulliga ja keskmised võib võrrandist ära jätta. Kuna võrrandi nimetajas on standardhälbed (ruutjuur hajuvusest), siis eeldab I_{YZ} muutujate normaaljaotust või vähemalt standardhälbe kasutamise põhjendatust. Ruumilise korrelatsiooni kirjeldamiseks sobib ka üldine ristkorrutis ja mitteparameetrilised korrelatsioonikordajaid. Nagu Morani *I*, on ka I_{YZ} muutumisvahemik homogeense varieeruvuse korral −1 ja +1 vahel, ebaühtlase varieeruvusega muutujate korral võib väärtus sellest vahemikust väljuda (Cliff ja Ord 1981).

Ruumilist korrelatsiooni saab arvutada vaid andmestiku selles osas, kus mõlemad muutujad on mõõdetud. Vastavalt sellele peaksid ka standardhälbed olema arvutatud vaid nendest samadest vaatlustest. Väiksem andmemaht muudab ruumilise korrelatsiooni hinnangud autokorrelatsiooni hinnangutest ebastabiilsemateks. Kui tunnused on vaid osaliselt ülekattes ja kasutada kõigist vaatlustest arvutatud standardhälbeid, on korrelatsioonid stabiilsemad, kuid neid mõjutab andmestiku ülekattest välja jääv osa.

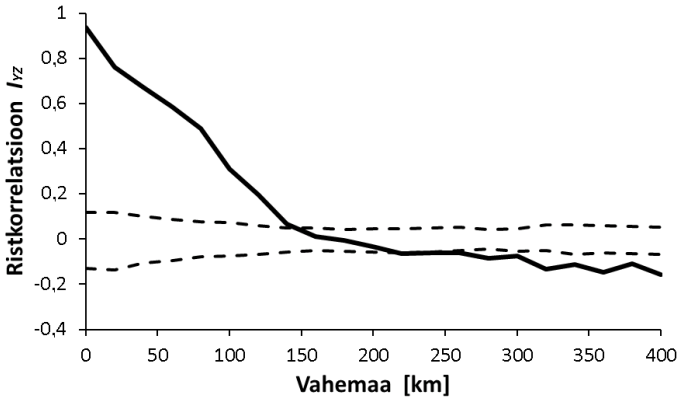
Mitme tunnuse omavaheliste seoste tabelkujul esitamiseks kasutatakse ruumiliste korrelatsioonide maatriksit (Wartenberg 1985). Ruumiliste korrelatsioonide maatriksi diagonaalil on autokorrelatsiooni kordajad, ülejäänud lahtrites ristkorrelatsioonid. Samasse tabelisse saab paigutada nii ruumilised kui ka tavalised korrelatsioonikordajad, paigutades ühed ülespoole ja teised allapoole diagonaali, või siis samasse lahtrisse üksteise alla, nagu esitasid Kalkhan ja Stohlgren (2000) (tabel 8).

Tabel 8. Ruumiliste korrelatsioonide maatriks mõnede kohatunnuste vahel Baltimaade ilmavaatlusjaamades. Diagonaalil on autokorrelatsioonid kaugusel kuni 50 km, lahtri ülemisel real on ruumiline korrelatsioon sama vahemaa korral, alumisel real tavakorrelatsioon. Tavakorrelatsioonid on üldiselt tugevamad kui ruumilised korrelatsioonid ning autokorrelatsioonid tugevamad kui ristkorrelatsioonid.

	Maapinna kõrgus	20km ümbruse metsasus	Kaugus merest	Aastane sademete hulk
Maapinna kõrgus	0,576			
20 km ümbruse metsasus	0,029 0,171	0,489		
Kaugus merest	0,535 0,664	−0,131 −0,037	0,925	
Aastane sademete hulk	−0,030 0,113	0,050 0,361	−0,279 −0,233	0,430

Kahe tunnuse vahelise korrelatsiooni sõltuvust vaatluste vahemaast saab graafiliselt näidata korrelogrammi abil, mis kujutab kahe tunnuse vahelise ristkorrelatsiooni sõltuvust vaatluste vahemaast (joonis 5-17). Arvutada ja graafikul kujutada saab ka tunnustevahelist poolhajuvust. Eristamaks neid autokorrelatsiooni kujutavatest graafikutest, kasutatakse nimetusi ristkorrelogramm ja ristvariogramm.

Korrelogramm võib olla nii kumulatiivne kui ka diskreetne. Esimesel juhul kasutatakse vaatluspaare, mille vahemaa on piirväärtusest väiksem; teisel juhul teatud kaugusvahemikku jäävaid vaatluspaare. Kaugustsoonide kaupa arvutatud ruumiline korrelatsioon on tundlikum ja võib anda seosest detailsema pildi, kuid tsoonide ebaõnnestunud valiku korral või vaatluskohtade ebapiisava hulga korral võib ruumiliselt kumulatiivne korrelatsioon seoseid paremini esile tuua.



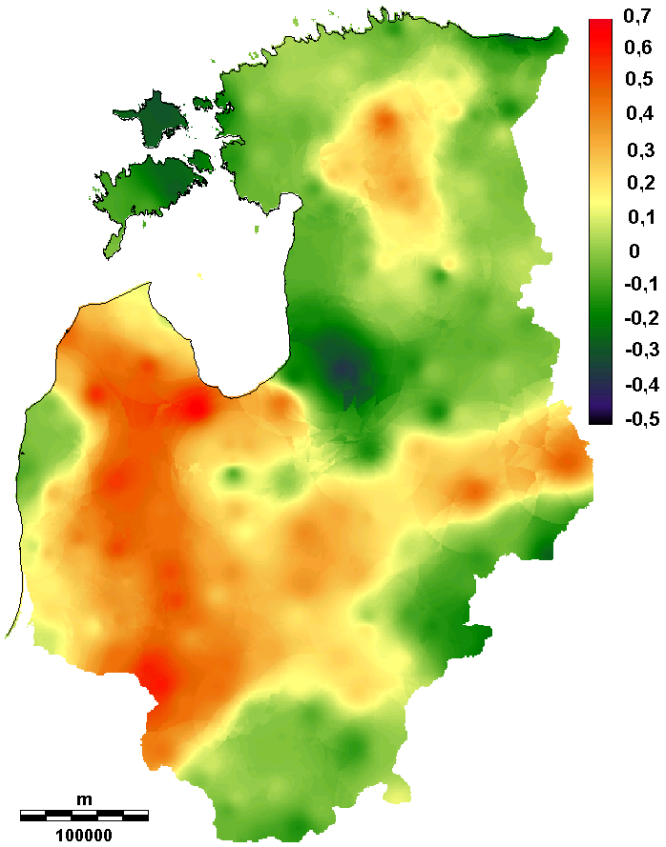
Joonis 5-17. Baltimaade oktoobri ja novembrikuu paljuaastase keskmise sademete hulga ruumiline korrelogramm koos nullmudeli 95% usaldus-piiridega (katkendjooned). Vahemaani 150 km on seos oktoobri ja novembri keskmise sademete hulga vahel statistiliselt olulisel tasemel positiivne, kaugemal kui 250 km negatiivne. Korrelatsioonikordaja samas jaamas mõõdetud oktoobri ja novembri keskmise sademete hulga vahel R = 0,9367. Andmed Remm *et al.* (2011).

Ruumilist korrelatsiooni saab arvutada ka lokaalselt igas uuritava ala kohas eraldi või siis iga vaatluspunkti panusena (Reich *et al*. 1995). Ruumilise korrelatsiooni kaardil kujutamise näitena on esitatud seos paljuaastase keskmise sademete hulga vahel erinevatel kuudel, kus võrreldavad tunnused on oktoobri paljuaastane keskmine ja novembri paljuaastane keskmine sademete hulk vaatlusjaamades (joonis 5-18). Kuna korrelatsioon arvutati kahe tunnuse vahel, siis järelikult on see ristkorrelatsioon, mitte autokorrelatsioon. Ristkorrelatsioon arvutati lokaalselt kilomeetrise vahemaaga võrgustiku igas punktis ehk lokaalses fookuses. Arvesse läksid jaamad, mis olid lokaalsest fookusest kuni 100 kilomeetri kaugusel (muidu ei jääks mõne lokaalse fookuse ümbruses ühtegi jaamadepaari). Seega on arvutati lokaalne ristkorrelatsioon. Lisaks sellele oli aga üks piirang veel – igas arvutatavas kohas kasutati vaid nende vaatlusjaamade andmeid, mille omavaheline vahemaa on kuni 50 km ning üks võrreldav tunnus (oktoobri sademed) võeti ühe jaama vaatlustulemustest ja teise tunnuse (novembri sademed) väärtus teise jaama andmetest. Selliselt arvutatud korrelatsioonikordaja sõltub jaamade omavahelisest vahemaast ja seega on see lokaalselt arvutatud ruumiline ristkorrelatsioon.

Ruumiline ristkorrelatsioon ei näita seost tunnuste vahel samas kohas, vaid teatud vahemaa korral. Samuti nagu ruumiline autokorrelatsioon näitab seost sama tunnuse väärtuste vahel teatud vahemaa korral. Ruumilist ristkorrelatsiooni võib muidugi arvutada ka teistsuguse vahemaa juures, igas kaugustsoonis eraldi või kaugusest sõltuvate kaaludega, nii nagu ruumilist autokorrelatsioonigi. Ruumist korrelatsiooni, ruumilist regressiooni ja ruumilist autoregressiooni selgitatakse ka peatükis 5.5.5 ja joonisel 5-21.

Ruumilise korrelatsiooni korrelogramm kujutab kahe muutuja vahelise korrelatsiooni sõltuvust vaatluskohtade vahemaast

Ruumilise korrelatsioonikordaja olulisuse hindamine ei ole isegi normaaljaotusega andmete puhul triviaalne, kuna vaatlused on sama kaugustsooni vaatluspaarides esindatud erinev arv kordi või siis on nende mõju kaalutud vahemaast sõltuvate kaaludega. Ruumilise korrelatsiooni olulisuse hindamiseks kasutatakse Mantel testi, bootstrap meetodit (Bjørnstad *et al.* 1999a) ja Monte Carlo testi (ptk 3.6.6 ja 4.1.4.2). Kui andmekihtide seost konkreetse ruumiga ei uurita, siis piisab, kui Monte Carlo kordustes juhuslikult ümber paigutada väärtusi ühes võrreldavas andmekihis.



Joonis 5-18. Oktoobri ja novembrikuu paljuaastase keskmise sademete hulga ruumiline korrelatsioon kuni 50 km vahemaaga ilmavaatlusjaamades Baltimaades. Positiivse seose piirkondades on oktoobri ja novembri keskmise sademete hulga vahekord suhteliselt ühetaoline, negatiivse ruumilise korrelatsiooni aladel on kuni 50 km vahemaaga paiknevate jaamade vaatlustulemustes suured erinevused. Andmed Remm *et al*. (2011).

Uurimused

Miller *et al*. (2007) mainivad ruumilise seose kvantifitseerimise vahendina ka uuritava ala osadeks jagamist ja seoste arvutamist igas regioonis eraldi ning ruumikoordinaatide lisamist seletavate tunnuste hulka. Korrelatsiooniks tuleks nimetada siiski vaid seost kahe numbrilise muutuja vahel.

5.4.3. Ruumilise autokorrelatsiooni mõju

Valimid ja üksikvaatlused ruumiliselt autokorreleerunud andmestikest ei ole täiesti sõltumatud, nagu eeldab enamik statistilisi teste. Selle tõttu näitavad statistilised testid reeglina tugevamaid tunnustevahelisi seoseid kui autokorrelatsioonita andmete puhul.

Autokorreleeruvate muutujatega võib toimida kahel viisil: tulemusi korrigeerides või autokorrelatsiooni lülitamisega nullmudelisse (Roxburgh ja Chesson 1998). Ruumilise autokorrelatsiooni nullmudelisse lülitamine tähendab ruumilise struktuuri jäljendamist, mida käsitletakse pindmustrite moodustamise osas (ptk 6.2).

Korrigeerimise variandi puhul arvutatakse tunnustevahelist seost näitav statistik ja siis korrigeeritakse selle väärtust või statistilist olulisust vastavalt autokorrelatsiooni tugevusele (Dale *et al.* 1991). Kasutatakse näiteks Mantel osatesti, mis käsitleb tunnuste ruumilist lähedust täiendava argumenttunnusena. Testis kasutatakse regressioonijääke ja arvutatakse tunnuste partsiaalkorrelatsioonid, nagu mitmese regressiooni puhulgi.

5.4.4. Kriging mitme tunnusega

Kriging varieeruva keskmisega on kahe muutuja analüüs, juhul kui see varieeruv keskmine arvutatakse teise tunnuse või teiste tunnuste järgi oodatava väärtusena, näiteks regressioonimudelist (Goovaerts 1997, 2000). Varieeruva keskmise puhul kasutatakse krigingu kaalude asemel kovariatsioone ja mõõtmistulemuste asemel regressioonijääke.

Kriging lokaalsel regressioonipinnal (*kriging with external drift*) arvutab regressiooni teise muutujaga lokaalselt, see on iga koha ümber uuesti, mitte globaalselt, nagu kriging varieeruva keskmisega (Goovaerts 1997, 2000).

Kokrigingu puhul interpoleeritakse uuritavat muutujat korraga kahe tunnuse järgi. Kui kahe eelmise meetodi korral mõjutab teine tunnus interpoleerimistulemusi kas oma lokaalse või globaalse trendpinna (keskmise ootuse) kaudu, siis kokrigingus kasutatakse teise tunnuse otseseid mõõtmistulemusi.

Uurimused

Goovaerts (2000) modelleeris ja interpoleeris sademete hulka Lõuna-Portugalis kombineeritult kõrgusmudeliga ja leidis, et parima tulemuse andis kriging varieeruva keskmisega. Mainitud meetod võimaldab kasutada krigingut koos regressioonimudeliga. Lokaalsete regressioonimudelite kasutamine prognoosi ei parandanud. Arvatavasti ei ole lokaalsetel regressioonidel piisavalt üldistusjõudu.

5.5. Ümbruse mõju ja ruumiline regressioon

Organismid kasutavad ja mõjutavad erineva suurusega ümbrust erineval määral sõltuvalt eluviisist ja suurusest. Taimede ruumilises populatsiooniökoloogias laialtlevinud lihtsustatud käsitlusele, mille kohaselt isendite tingimused sõltuvad populatsiooni tihedusest, vastandub ruumiliselt detailsem arusaam, et iga isendi edukus sõltub eelkõige tema vahetutest naabritest. **Naabruse mõjudeks** (*neighbourhood effects*) nimetasid Mack ja Harper (1977) faktorit, mis sisaldab naabruses olevate taimede osakaalu, kaugust ja paiknemissuunda. J.F. Addicott *et al.* (1987) nimetasid ala, mille ulatuses organism tegutseb või mis organismi mõjutab, **ökoloogiliseks naabruseks** (*ecological neighbourhood*). Mainitud autorid rõhutavad, et organismidel ei ole vaid üks ja ainus naabrus. Erinevate protsesside ja suhete osas võib ökoloogiline naabrus olla erinev. Condit *et al.* (1994) jagasid naabrusmõjud lokaalseteks ja regionaalseteks, kuid ei suutnud regionaalseid mõjusid tõestada, sest kaugema ümbruse mõju võib teostuda lähinaabruse mõjutamise kaudu.

Ümbruse mõju arvestamiseks tuleb see mõju kõigepealt kas kuidagi ära kirjeldada või kasutada mingit teoreetilist mudelit. Ümbruse mõju saab kirjeldada ja statistilisse mudelisse kaasata vähemalt viiel erineval viisil:

- ruumilise autoregressioonina, mis kirjeldab sama nähtuse mõju ümbrusest (ühepoolne seos);
- ruumilise regressioonina, mis kirjeldab teiste tunnuste mõju ümbrusest (mõju allikas asub ühes kohas, aga ta mõju ulatub mujale);
- ruumilise autokovariatsioonina – kui nähtus on mingil määralgi ruumiliselt pidev, siis sõltub nähtuse väärtus või esinemise tõenäosus mingis punktis selle nähtuse väärtusest või esinemisest ümbruskonnas (vastastikune seos);
- ruumilise kovariatsioonina – vastastikuse statistilise seosena kahe tunnuse vahel, millest üks on mõõdetud ühes kohas ja teine teises kohas (ümbruses);
- hälvete ruumilise muutlikkuse ja hälvete autokovariatsioonina, kui tunnuste mõõtmistäpsuses või mudeli usaldusväärsuses esinevad piirkondlikud tendentsid.

Ümbruse mõju võib mudelis olla eelkõige omaette argumenttunnusena ja/või funktsioontunnuse ruumilise autokorrelatsioonina

Ümbrust kirjeldavad parameetrid võivad olla arvulise muutuja statistikud, mis on arvutatud ümbruses teadaolevatest väärtustest (ümbruse keskmised, varieeruvuse näitajad, samas kohas ja ümbruses mõõdetud väärtuste suhted) – nominaalset muutujat ümbruses kirjeldavad selle statistikud (kategooriate pind, vahekord ja domineerimistase). Kõigi ümbrust kirjeldavate statistikute arvutamisel saab kasutada kaugusest sõltuvaid kaale ning ümbrust saab kirjeldada mistahes raadiuses või kaugustsoonis. Ümbrust kirjeldavate statistikute hulka võivad kuuluda ka tunnustevahelisi seoseid kirjeldavad statistikud ja kõikvõimalikud kombineeritud tunnused.

Uuritava ala serva lähedal annavad naabervaatluste arvu või laikude suurusega seotud näitajad hälbinud hinnanguid, osakaalust ja tihedusest tuletatud indeksite puhul väheneb serva lähedal vaid valimi maht. Etteantud tüüpi ala osakaal vahetus ümbruses ei pruugi olla parim indikaator, kuna lähiümbrus on enamasti üsna samasugune nagu fookuses olev kohtki.

Autokorrelatsiooni osas (ptk 5.1) oli juttu tunnuse endaga sarnanemisest teatud kaugustel. Ümbruse mõju kirjeldamise puhul võib oluline olla nii prognoositava muutuja enda väärtus ümbruses kui ka argumenttunnuste väärtused ümbruses. Nii sõltub liigi tõenäoline ohtrus mingis kohas nii ümbruses olevatest sama liigi naaberpopulatsioonidest kui ka elupaiga hulgast ja kvaliteedist ümbruses. Näiteks

sõltub põdra elupaiga kvaliteet mitte ainult antud koha omadustest, vaid ka ka laiema ümbruse paljudest omadustest vähemalt 10 km raadiuses. Uurida tasuks, millise ulatusega ümbrus on oluline ja kuivõrd sõltub ümbruse mõju muudest faktoritest. Küsimus, kuidas ühe tunnuse väärtus mingis kohas seostub teise tunnuse väärtustega selle koha ümbruses, on modelleeritav ruumilise regressioonina.

Küllastatud ruumilise regressiooni (ptk 5.5.5) mudel sisaldab iga argumenttunnuse mõju igas kaugustsoonis. Kuna küllastatud mudeli parameetrite arv võrdub sel juhul kaugustsoonide arv korda mõjufaktorite arv, siis on ruumilise regressiooni mudelid keerukad ja nõuavad suurt õpetusandmestikku. Ruumilise regressiooni mudeli lihtsustamiseks on võimalik kas integreerida kaugustsoonid, mõjufaktorid või siis mõlemad. Esimesel juhul käsitletakse kõiki kaugustsoone ühtse ümbruskonnana. Teisel juhul ühendab ja asendab igat üksikfaktorit eraldi sobivusindeks, kolmandal juhul esindab iga koha ümbruskonna kõiki omadusi üks sobivusindeks. Ruumilise mõju arvestamise vajadus ja kasutamise viis sõltub muuhulgas uurimuse mõõtkavast.

Ümbrust saab käsitleda ühtsena või eraldi kaugustsoonidena

Uurimused

Ülevaate ruumilist sõltuvust kaasavatest liikide ja elupaikade leviku mudelitest annavad Miller *et al*. (2007).

Milne *et al*. (1989) jagasid uuritavate pikslite ümbrused klasteranalüüsiga kaheks ja kasutasid neid tüüpe täiendava nominaalse argumenttunnusena hirve (*Odocoileus virginianus*) elupaikade kaardistamisel.

Chou ja Soret (1996) näitasid, et ümbruse erinevat mõju erinevatele liikidele tuleks liikide leviku modelleerimisel kindlasti arvestada.

M. Zobel on oma uurimustes näidanud, et niidukoosluste liigirikkus ei sõltu ainult kohalikest tingimustest, vaid ka sellest, millised kooslused on koha ümber. Liikide levikuvõime ja seemnevaru mullas omavad koosluse liigirikkuse ja koosseisu määramisel samavõrd tähtsust kui kohalikud kasvukoha tingimused. Kusjuures koosluse liigirikkus ei paista takistavat uute liikide sissetungi (Cantero *et al*. 1999).

Metsa kasvu ja struktuuri mudelites on naaberpuude tiheduse ja kaugusega seotud puistu dünaamilised aspektid: noorte puude kasvama hakkamine ja puude väljalangemine puistust (He ja Duncan 2000).

Wahl (2001) rõhutas naaberorganismide tähtsust merepõhja koosluste kujunemisel.

5.5.1. Mõjuväljade mudelid

Ümbruskonna mõjude arvestamisel saab igale mõjuvale faktorile eraldi või kõigile kokku omistada mõjupunkti ümbritseva **mõjutsooni** (*influence zone*). Mõjutsoonide kasutamisel arvutatakse iga koha jaoks mõjutsooni piiresse jäävate lähedaste objektide mõjude summa või mõjude muu kombinatsioon. Mõjutsoonide arvutamise mitmesuguseid meetodeid on kasutatud eriti metsanduses puudevahelise konkurentsi ja puude mõjusfääride määramiseks ning seeläbi ka puistu struktuuri modelleerimiseks. Fütogeneetilistest väljadest on kirjutanud juba A. A. Uranov (1965). Üldisema ökoloogiliste väljade teooria on esitanud Wu *et al.* (1985), seda on kasutanud ja edasi arendanud Moeur (1997), Kuuluvainen ja Linkosalo (1998), Saetre (1999), He ja Duncan (2000). Seemnete emataimest eemalelevimise modelleerimise meetoditest võib ülevaate leida Nathan ja Muller-Landau (2000) artiklist.

Geograafias on püütud linnade mõjusfääre modelleerida gravitatsiooniseaduse järgi (Abler *et al.* 1971). Oletatakse, et linna mõju (*I*) on võrdeline tema elanikkonna suurusega (*P*) ja pöördvõrdeline kauguse (*d*) ruuduga.

$$I = K\frac{P}{d^2} \qquad [5\text{-}32]$$

Selles valemis on *K* mudeli sobitamiseks vajalik konstant. Kahe objekti omavaheline mõju avaldub vastavalt gravitatsiooniseadusele.

$$I = K\frac{P_1 P_2}{d_{1,2}^{\ 2}} \qquad [5\text{-}33]$$

Potentsiaalpind on gravitatsioonivälja analoog, mis väljendab kõigi teiste punktide või objektide summaarset mõju (M_j) igale sellel pinnal olevale punktile. Geograafilises kontekstis võib mõju tähendada ka koha kättesaadavust, kommunikatsioonitihedust või turupotentsiaali. Seega koha *i* potentsiaal võrdub

$$V_i = \sum_{j=1}^{n} \frac{M_j}{d_{ij}}. \qquad [5\text{-}34]$$

Selle juures võib murrujoone all oleva objektide vahemaa jätta algkujule, astendada või kasutada mõnda muud teisendust – sõltuvalt nähtust kujundava protsessi omapärast.

5.5.2. Tegutsemisala suurus ja elupaiga valik

Elupaigaeelistuse mudeleid on välja töötatud eelkõige liikuvate loomade jaoks. Traditsiooniliselt eeldatakse mudelis, et kõik elupaigad isendi tegutsemisala piires on talle võrdsel määral kättesaadavad.

Elupaiga valiku indeks näitab elupaigatüübi kasutamise intensiivsust võrreldes selle kättesaadavusega uuritavale liigile

Metapopulatsiooniteooria kohaselt suureneb koha asustatuse tõenäosus sisserände intensiivsusega. Eeldades, et elupaiga lähedaste osade vahel on ränne tõenäolisem ja edukam, võib järeldada elupaiga asustatuse sõltuvust elupaiga teiste osade kaugusest. Üldistatumal kujul võib seda seost väljendada ka nii, et elupaiga sobivus liigi püsimiseks sõltub elupaikade hulgast ja kvaliteedist selle koha ümbruses. Kusjuures vähemalt põdra ja lendorava jaoks on tegelikes maastikes elupaiga hulk ümbruses olulisem kui elupaiga konfiguratsioon (Remm ja Luud 2003, Ritchie *et al*. 2009). Suurem elupaigalaik on enamasti maastikuliselt varieeruvam ja sisaldab rohkem mikroelupaiku ning toetab seetõttu suurema arvu liikide koosesinemist ning tagab populatsioonide suurema vastupidavuse arvukuse kõikumisele ja häiringutele (Honnay *et al.* 1999).

Uurimused

Arthur *et al*. (1996) käsitlesid kasutuskõlbuliku tegutsemisareaali suurust dünaamiliselt sõltuvana isendi asukohast. Hjermann (2000) täiustas elupaiga kättesaadavuse dünaamilise hindamise metoodikat käsitledes elupaiga kättesaadavust pideva muutujana. Elupaigaeelistuste selgitamisel on

kasutatud elupaiga omaduste peakomponentanalüüsi (Garshelis 2000), diskriminantanalüüsi või logistilist regressiooni (Sherburne ja Bissonette 1994, Mladenoff *et al*. 1995, Block *et al*. 1998). Compton *et al*. (2002) võrdlesid elupaiga kasutust mitte kogu võimaliku elupaigaga, vaid kohapunkti paarides, millest üks on looma teadaolev asukoht ja teine on looma tegutsemisraadiuse piires olev juhuslik punkt. Looma teadaoleva asukoha tunnuste väärtustest lahutati sama tunnuse väärtus juhupunktis. Väärtuste vahesid kasutati liigi esinemist/puudumist prognoosivas logistilises regressioonimudelis. Suhteliste väärtustega mudeli interpreteerimine on mõnevõrra keerukam kui absoluutväärtustega mudeli mõistmine. Vaatluste kasutamine vaatluspaaridena võimaldab aga esinemise ja puudumise andmeid tasakaalustada.

5.5.3. Eraldatus ja ühendatus

Elupaikade **ühendatust** (*connectivity*) ja selle pöördväärtust – **eraldatust** (*isolation*) on oluliseks liikide levikut määravaks teguriks peetud juba vähemalt MacArthuri saarte biogeograafia teooriat käsitlevatest töödest alates (MacArthur ja Wilson 1967, Levin 1974, Doak *et al*. 1992, Taylor *et al*. 1993, Lindenmayer ja Possingham 1996, Schumaker 1996, With *et al*. 1997, Hanski 1998, Tischendorf ja Fahrig 2000). Ühendatus tähistab ökoloogias eelkõige ühenduse määra üksikute konkreetsete elupaigalaikude vahel. Ühendatusele lähedane mõiste on sidusus (ptk 4.3.1), mis tähistab pindkategooria üldist ühendatust maastiku tasemel.

Lokaalpopulatsiooni väljasuremise tõenäosus sõltub suuresti populatsiooni (elupaiga) suurusest ja sisserände sagedusest, mis omakorda sõltub elupaiga isoleeritusest. Elupaiga isoleeritus võib sõltuda nii üksikute eraldatud elupaikade või elupaigamaatriksi (*habitat matrix*) omadustest kui ka elupaigalaikude vahelise ala omadustest ja konfiguratsioonist. Kõiki võimalikke parameetreid on raske modelleerida ja seetõttu piirdutakse enamasti maastikustruktuuri sidususe lihtsustatud näitajatega. Tischendorf ja Fahrig (2000) mainivad kahte tüüpi ühendatuse mõõte: lähima naaberelupaiga kaugus ja elupaiga pindala teatud ulatusega ümbruses. Moilanen ja Nieminen (2002) lisavad ühendatuse näitajate kolmanda klassi: näitajad, mis arvestavad kõigi võimalike sisseränne tagavate populatsioonide kaugust.

Puhvertsooni kasutavad indeksid ei arvesta üksikute naaberlaikude erinevat kaugust. Eraldiasuvate elupaigaosade vahemaad saab arvutada nii eraldiste keskkohtade vahel kui ka eraldise servast eraldiste servani. Keskpunktide vahemaade arvutamine on tehniliselt lihtsam, kaugused servast võivad olla ökoloogiliselt sisukamad. Nii lähima naaberlaigu kui ka mitme või kõigi naaberlaikude kauguse indeksit on võimalik korrigeerida vastavalt naaberelupaiga suurusele, mis mõjutab sisserände eeldatavat intensiivsust, ja vastavalt sama elupaiga suurusele, mis on eeldatavasti seotud väljarände intensiivsusega. Reeglina peaks kõigi rändeulatuses olevate sama elupaiga osade kauguse arvestamine andma parema mudeli, kuid sobivate rändeparameetrite leidmine võib osutuda üsna keerukaks. Naaberlaikude kauguse arvestamisel on soovitatud kasutada negatiivset eksponentsiaalfunktsiooni (Moilanen ja Nieminen 2002)

$$S_i = A_i^c \sum_{j \neq i} \exp(-\alpha d_{ij}) A_j^b , \qquad [5\text{-}35]$$

kus S_i on laigu i ühendatus, A_i on laigu i pindala, d_{ij} on vahemaa laikude i ja j vahel, $1/\alpha$ on keskmine rändeulatus, c ja b on sisse- ja väljarände parameetrid.

Populatsioonide ühendatuse mudelitesse on kaasatud ka elupaiga kvaliteedi näitajaid (Verboom *et al*. 1991, van Apeldorn *et al*. 1992, Kindvall 1996, Klok ja De Roos 1998), kuigi põhiliselt kasutatakse elupaiga kvaliteedi näitajaid liikide leviku (ptk 5.6) elupaigasobivuse (ptk 5.6.2) mudelites.

Uurimused

Mitmetes töödes on leitud, et loomade asustustihedus sõltub rohkem elupaiga suurusest või hulgast ümbruskonnas kui ühendatud ökovõrgustikust (Connor *et al*. 2000, Goldingay ja Possingham 1995, Harrison ja Bruna 1999, McIntyre ja Wiens 1999, Trzcinski *et al*. 1999, Verboom *et al*. 2001). Fragmenteerunud elupaiga suures laigus elamine võib anda terve rea eeliseid: teiste elupaigatüüpide röövloomad ohustavad vähem, suurest laigust juhuslik väljakandumine on vähem tõenäoline, suurema elupaigalaigu keskmine kvaliteet on enamasti kõrgem, suures elupaigas kulub vähem energiat ressursside kättesaamiseks, suures elupaigas asub suurem populatsioon, kaaslase leidmine on lihtsam, partnerite valik on suurem ning suurema populatsiooni juhuslik väljasuremine on vähem tõenäoline (Connor *et al*. 2000).

Elupaiga ruumiline struktuur mõjutab elupaiga kvaliteedi varieeruvust elupaiga sees ning liikide levikuvõimalusi, mis omakorda määrab liigi pikaajalise säilimise tõenäosust (Lamberson *et al*. 1992, Akçakaya *et al*. 1995, Lindenmayer *et al*. 1999). Metapopulatsiooni teooria kohaselt on isoleeritud elupaiga osade asustamise tõenäosus väiksem ja nendes oleva populatsiooni väljasuremise tõenäosus suurem.

Ida-Virumaa põtrade asustustiheduse ja elupaikade sobivuse hindamisel osutus kõige informatiivsemaks põdra elupaiga hulk 1…10 km kaugusel kohast (Luud ja Remm 2001, Remm ja Luud 2003). Ruumiliselt pideva elupaiga eraldise suurusega seost ei leitud. Seega paistab, et elupaikade ökoloogilise võrgustiku sidusus ei ole hea liikumise ja levimise võimega põdrale oluline. Põtradele on vajalikud eelkõige suured tuumalad.

5.5.4. Indikaator-ümbrus

Ümbruse mõjude uurimisel on kolm keskset küsimust:

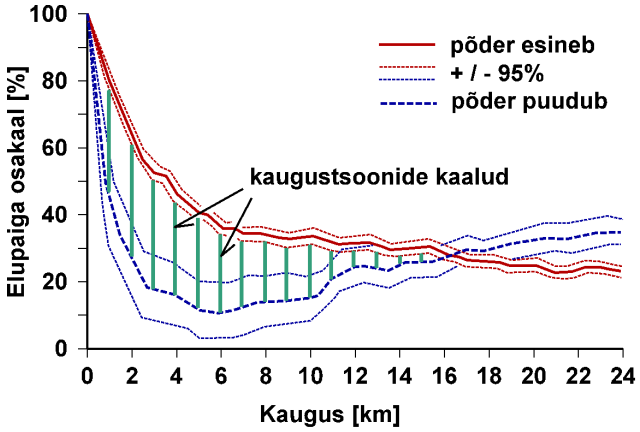
- milliste faktorite mõju on oluline,
- milline on mõju tugevuse sõltuvus vahemaast ehk siis mõjude ruumiline mõõtkava,
- milline on fookuses oleva koha ja ümbruse mõju vahekord.

Ruumianalüüsi seisukohast on kaks viimast küsimust kaugustsoonidele kaalude leidmise probleem. Enamasti võib eeldada, et väga kaugel oleva faktori mõju ei ole oluline. Vahetu naabrus on enamasti samasugune kui antud koht. Seega täiendavad vahetule naabrusele järgneva lähiümbruse andmed sama koha andmeid kõige olulisemal määral, kaugem ümbrus on **kohalikuks taustsüsteemiks** (*local reference*). Kõige suuremat prognoosivat väärtust omavat kaugustsooni on nimetatud **indikaator-ümbruseks** (*indicative neighbourhood*) ehk indikaator-kauguseks (*indicative distance*) (Luud ja Remm 2001, Remm ja Luud 2003, Linder *et al*. 2008).

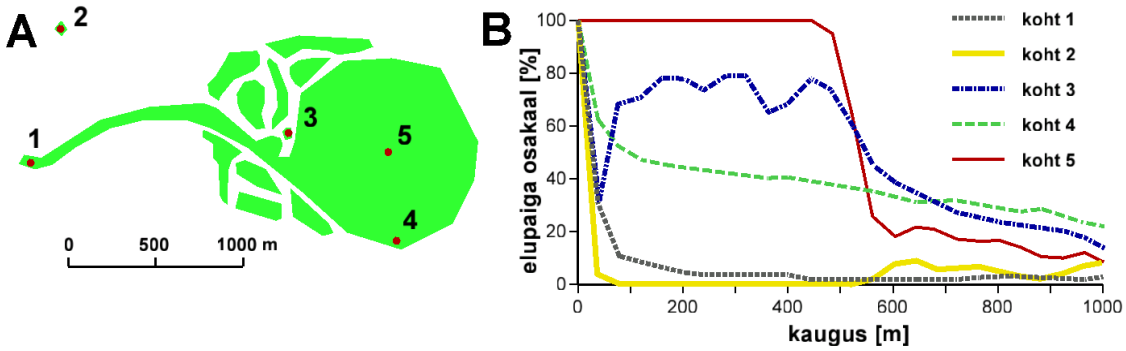
Kaugustsoonide mõju saab võrdlevalt hinnata sõltuva tunnuse erinevaid väärtusi omavate kohtade ümbruse võrdlemise abil. Lihtne on võrrelda ümbrust nendes vaatluskohtades, kus liik esines ja nendes, kus liiki ei leitud (joonis 5-19). Kaugustsoonide mõju määramisest tuleks välja jätta vaatluskohad, mis asusid väljaspool liigi elupaiku, sest nendes kohtades piiravad liigi esinemist sama koha omadused, mitte ümbrus.

Liigi puudumise põhjus mingis kohas, kus ta selle koha omaduste järgi võiks esineda, võib olla

selle koha ümbruse iseloomus. Oluline parameeter võib olla näiteks elupaiga hulk ümbruses, aga ka mingi muu liigile olulise ressursi või välditavate objektide hulk teatud kaugusel (joonis 5-20). Hüpoteesi kontrollimiseks oleks vaja mõõta otsitava muutuja hulka erinevates kaugustsoonides nii nendes proovikohtades, kus liik esines, kui ka nendes, kus liik oleks võinud esineda, aga teda ei leitud. Liigi esinemise kohad võib seejuures liigi ohtrusega läbi kaaluda. Sageli ei ole liigi puudumist võimalik usaldusväärselt hinnata ja vaatluskohtades, kus liik tõenäoliselt tegelikult esineb, kuid varjatud eluviisi ja otsimismeetodi ebapiisavuse tõttu seda ei leitud, on liigi puudumine ajutine või näiv (**näiv-puudumine** – *pseudo absence*). Andmete puudumisel või vähesel usaldusväärsusel liigi mitte-esinemise kohta saab neid asendada juhukohtadega uuritaval alal. Analoogiliselt näivate puudumistega võivad andmestikes olla ka ekslikud või juhuslikud esinemiskorrad (**näiv-esinemine** – *pseudo presence*), kuid nende sagedus on üldjuhul madalam. Need on kohad, mille elupaiga ja ümbruse omadused ei võimalda liigil seda asustada, kuid kus on tuvastatud seal juhuslikel asjaoludel viibinud isendeid. Näiteks rändel olevad või ebatüüpilise häiringu tõttu ringi ekslevad loomad.



Joonis 5-19. Põdra elupaiga keskmise osakaalu erinevus põdra ekskrementidega loendusalade ja põdrata loendusalade vahel Ida-Virumaal 1999. aastal (Remm ja Luud 2003). Indikaator-ümbrus on nende andmete kohaselt kaugusel 0,5 kuni 11 km.



Joonis 5-20. Viis erineva ümbrusega kohta elupaiga ruumilises struktuuris (A) ja elupaiga suhteline hulk samade kohtade ümber (B). Kohtade ümbrused erinevad kõige enam kaugustel 20…560 m.

Uurimused

Remm ja Luud (2003) kasutasid naabruse mõju hindamisel kaugustsoonide kaalutud keskmist mõju. Põdra asustustihedus vaatlusaladel korreleerus kõige enam elupaikade kaalutud hulgaga ümbruses. Kaugustsoonide kaalud elupaikade kaalutud hulga arvutamiseks saadi põdraga ja põdrata vaatlusalade ümbruse võrdlemisel. Erinevus põdraga ja põdrata vaatlusalade vahel oli suurim kaugusel 1 kuni 10 km vaatlusala keskmest. See tähendab, et nendel vaatlusaladel, mis paiknevad suuremate

metsade sees, oli põtrade tihedus suurem kui lageda ümbrusega vaatlusaladel. Seejuures paiknesid kõik vaatlusalad metsas (või raiesmikul) ja ka vaatlusala lähem ümbrus oli reeglina mets, sest maakatteüksustele on omane teatud ruumiline autokorrelatsioon. Põdra elupaikade osakaal põdraga ja põdrata vaatlusalade ümbruses erines vähemalt kuni 16 km kauguseni vaatlusalast.

Remm ja Oja (2001) modelleerisid Otepää piirkonna maa-eluhoonete paiknemist. Regressioonimudelist saadi sobivuspind, mis annab iga koha sobivuse eluhoone seal esinemiseks. Sobivuspinda korrigeeriti sobivusega koha ümbruses, sest eeldati, et kõrgest sobivusest väga väikesel alal ei piisa eluhoone kõrgeks esinemistõenäosuseks. Hoonestatakse eelkõige kohad, kus hoonestuseks sobivat ala jätkub ka õuele ja kõrvalhoonetele. Samuti võib arvata, et head põllu- ja metsamaad on ka hoonestuseks sobivad maad. Seega eriti soodsad peaksid olema kohad, mille ümber säilib kõrge sobivus laiemas raadiuses.

Erinevate kauguste osatähtsuse hindamiseks kasutati prognoositud logit-tõenäosuste standardiseeritud mediaani, mis arvutati linnavälise mineraalmaa pikslitest. Mediaani eelistati keskmisele, kuna see on erindite suhtes vähem tundlik. Selleks, et muuta pikslite naabrused võrreldavateks, lahutati iga kaugustsooni sobivuse mediaanist selle kaugustsooni lähtepiksli sobivus. Kuna kaugemates tsoonides stabiliseerub sobivuse mediaan uuritava ala keskmisele tasemele, siis saadud tulemus jagati kaugemate tsoonide keskmise mediaaniga. Kaugemateks tsoonideks loeti kaugust alates 400 m. Selle kauguse peal on 11 × 5000 juhupunkti ümber sobivuse ruumiline autokorrelatsioon Morani *I* järgi keskmiselt vaid 0,195. Parema piltliku kujutise saamiseks lahutati tulemus ühest. Niimoodi saadi standardiseeritud kõverad, mis algavad ühest (kaugusel null on nähtus iseendaga identne) ja stabiliseeruvad null-taseme lähedal. Selliste standardiseeritud jaotuste kuju võimaldab hinnata nende pikslite naabruste erinevust, kus hoone on, juhupikslitega ja pikslitega, kus hoone ootus on mudeli järgi kõige suurem, kuid kus hoonet tegelikult ei ole.

Kaugustsooni olulisuse kaaludeks võeti juhuslike linnavälisel mineraalmaal paiknevate pikslite kaugustsooni keskmise standardiseeritud sobivuse ühelt poolt ja tegelike hoonete ümbruse kaugustsooni keskmise standardiseeritud sobivuse 95% alumise usalduspiiri vahe. Seega, oluliseks loeti nende kaugustsoonide omadused, mille puhul tegelike hoonete ümbrus erineb oluliselt juhuslike punktide ümbruse keskmisest (on väljaspool tegelike hoonete ümbruse keskmise usalduspiire).

Tegelike hoonete ümbruse erinevust kontrolliti nii juhuslike kui ka nende mudeli järgi parimate pikslite suhtes, kus hoonet tegelikult ei ole. Selgus, et nii juhuslike punktide kui ka parimate hooneta punktide ümbrus sarnaneb lähtepunktile tunduvalt vähem kui hoonet sisaldavate punktide puhul keskmiselt. Kõige suurem on erinevus umbes 20 kuni 120 m kaugusel. Uuritava ala üldisest keskmisest tasemest on hoonete ümbruse standardiseeritud prognoos statistiliselt oluliselt erinev umbes 210 m kauguseni. Järeldame, et hoone esinemiseks on oluline, et suhteliselt kõrge sobivus esineks umbes sellises raadiuses sobiva piksli ümber. Lähemal mõjub ruumiline autokorrelatsioon, mille kohaselt on iga koha oodatav ümbrus selle kohaga sarnane, kaugema ümbrus ei ole juhupunktide ja hoonega pikslite ümber oluliselt erinev. Hoonega pikslite ümber säilib sobivus laiemas raadiuses ja mineraalmaad on rohkem. Erinevus juhukohtade ja hoonete ümbruse vahel võimaldab omistada igale kaugustsoonile kaalu.

Lindenmayer *et al*. (1999) võrdlesid vaatlusala ümbrust 250 ja 500 m raadiuses. Nad leidsid statistiliselt olulisi seoseid pärdiku *Petaurus australis* esinemise ja 500 m raadiuses oleva ümbruse maastikuparameetrite vahel. Samu statistilisi seoseid ei leitud väiksema raadiusega ümbruse puhul.

5.5.5. Ruumiline regressioon ja autoregressioon

Ruumilise regressiooni korral sõltub muutuja väärtus kohas *i* argumenttunnuste väärtustest mingis teises kohas, reeglina naabruses. Seejuures ei pruugi sõltuvus tingimata kõige tugevam olla vahetus naabruses olevate väärtustega (ptk 5.5.4). Ruumilise regressiooni mudel sisaldab asukohast sõltuvaid parameetreid, näiteks kujul

$$Y = bX + \rho WX + \varepsilon, \qquad [5\text{-}36]$$

kus Y on funktsioontunnuste maatriks, b on regressioonikordaja, ε on juhuslik viga, ρ on ruumilise autokovariatsiooni vektor, X on argumenttunnuste maatriks, W on vaatlustevaheliste kauguskaalude maatriks.

Ruumiline regressioonimudel võib olla mitmel erineval kujul (vt näiteks Upton ja Fingleton 1985). Asukohast sõltuvad kaalud võivad olla nii argumenttunnuste, funktsioontunnuse autokovariatsiooni kui ka prognoosijääkide kordajateks. Ruumilise regressiooni tunnusjoon on asukohast sõltuvate kaalude osalus regressioonimudelis. Mitteruumilise regressiooni puhul asukoht mudeli parameetreid ei mõjuta. Enamasti on asukohast sõltuvad kaalud pöördvõrdelised kaugusega prognoositavast kohast. Võimalikud on ka mudelid, kus argumenttunnused on mõõdetud ühes kohas ja funktsioon arvutatakse hoopis teise kaugemal asuva koha kohta. Näiteks mudel, mis kirjeldab, kuidas lume paksus Otepää kõrgustiku eri osades mõjutab Emajõe kevadist veetaset Tartus. Selles mudelis oleksid Otepää kõrgustiku erinevatel osadel asukohast sõltuvad kaalud.

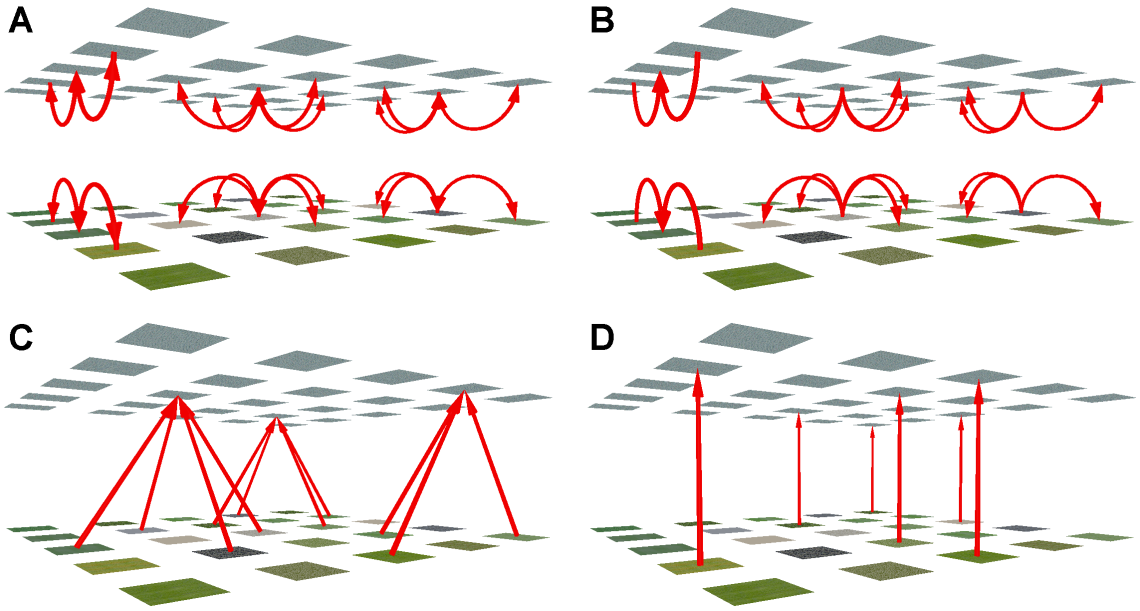
Ruumilise regressiooni mudel sisaldab asukohast sõltuvaid parameetreid

Ruumiline regressioon võib kombineeruda **autokovariatsiooniga**. Sellisel juhul lisanduvad mudelisse autokovariatsiooni kirjeldavad parameetrid. Mõisted autokovariatsioon ja autokorrelatsioon on lähedased. Korrelatsioon tähistab kahepoolset standardiseeritud kujul kirjeldatud seost, kovariatsioon standardiseerimist ei sisalda.

Autoregressiivse mudeli abil arvutatakse muutuja väärtusi sama muutuja väärtuste järgi teisel ajahetkel ja/või teises kohas. Tunnus iseenda väärtusi mõjutava faktorina on autokovariatsioon. Kuna reeglina on muutuja väärtused mingis kohas seotud väärtustega selle koha naabruses, siis ruumilise autokovariatsiooni korral sõltub funktsioontunnuse väärtus kohas *i* funktsioontunnuse enda väärtustest selle koha ümber (joonis 5-21).

Ruumiliselt autoregressiivse mudeli funktsioontunnus varieerub koos iga argumentunnusega ja koos iseenda väärtustega ümbruskonnas

Ruumiliselt autoregressiivse mudeli korral eeldatakse, et funktsioontunnuse prognoositud väärtus Y_i kohas *i* ei sõltu ainult argumenttunnuse väärtusest X_i samas kohas, vaid ka funktsioontunnuse prognoositud või teada olevatest väärtustest koha *i* ümber. Ruumiliselt autoregressiivne mudel eeldab prognoositava tunnuse ruumilist pidevust, mis ei ole argumenttunnuste abil täielikult seletatav. Näiteks sõltub liikumisvõimelise loomaliigi asustustihedus uuritavas kohas asustustihedusest ümbruses. Kui mudelis teisi muutujaid ei ole, siis võrdub funktsioontunnuse ootus sama tunnuse naabruses olevate väärtuste kaalutud keskmisega.



Joonis 5-21. Seosed andmekihtidena esitatud tunnuste vahel: A – ruumiline autokovariatsioon, B – ruumiline autoregressioon, C – ruumiline regressioon ja D – mitteruumiline regressioon.

Autoregressiivseid mudeleid jagatakse samaaegseteks ja suhtelisteks. **Samaaegne autoregressiivne mudel** (*simultaneous autoregressive model – SAR*) sobib selliste seletavate tunnuste puhul, mille mõju ei piirdu vaid uuritava koha ja selle naabrusega. Näiteks meteoroloogilised ja klimaatilised faktorid. Ilm ei muutu ühes kohas, vaid ikka mingil suuremal alal. Kui kasutatakse vaid esimese naabrustsooni andmeid, siis on see esimest järku autoregressiivne mudel, kui kasutatakse mitme kaugustsooni andmeid, siis on mudeli järk kaugustsoonide arvule vastavalt kõrgem.

Klassikalises SAR mudelis on kauguskaalud seotud funktsioontunnusega (muidu poleks põhjust seda autoregressiooniks nimetada)

$$Y = bX + \rho WY + \varepsilon \qquad [5\text{-}37]$$

ehk

$$Y - \rho WY = bX + \varepsilon, \qquad [5\text{-}38]$$

kus ε on juhuslik viga, ρ on ruumilise autokovariatsiooni vektor, b on regressioonikordaja, W on vaatlustevaheliste kauguskaalude maatriks, lihtsamal juhul kauguskaalud naabritel võrdsed ühega ja teistel võrdne nulliga.

Kauguskaaludega võivad olla seotud ka argumenttunnus ja/või prognoosihälbed. Autokovariatsiooni sisaldavat logistilist mudelit nimetatakse autologistiliseks mudeliks (ptk 6.2.4.7).

Suhteline autoregressiivne mudel (*conditional autoregressive model – CAR*) modelleerib erinevust oodatavast väärtusest ehk naabrusest sõltuvaid funktsioontunnuse hälbeid ehk suhtelisi väärtusi. Hälbed võivad olla tingitud näiteks atmosfääri läbipaistvuse varieeruvusest või olla sensorist tulenevad erisused kaugseireandmetes (Griffith ja Layne 1999). CAR eeldab funktsioontunnuse hajuvuse konstantsust ja kaasab vaid esimese astme naabrite mõjusid ning seetõttu sobib eelkõige uurimisala jagunemisel selgepiirilisteks üksusteks.

$$Y = bX + \rho W\left(Y - bX\right) + \varepsilon \qquad [5\text{-}39]$$

ehk

$$Y - \rho WY = bX - \rho WX + \varepsilon \qquad [5\text{-}40]$$

Autokovariatsiooni võib arvutada tunnuse väärtuste keskmisena ümbruskonnas või nende ruumiühikute (pikslite) osakaaluna, kus nähtus esineb. Viimast moodust on kasutanud näiteks Araújo ja Williams (2000). Lihtsamal juhul arvestatakse ümbruskonnaks vaid lähim naaberobjekt (Gellrich *et al*. 2007) või vaid külgnevad pikslid. Detailsema käsitluse korral kasutatakse kaugusest sõltuvaid kaale (Syartinilia ja Tsuyuki 2008).

Autokovariatsiooni sisaldava regressioonimudeli sobitamisel kasutatakse prognoositava muutuja teadaolevaid väärtusi ümbruskonnas. Mudeli järgi prognoosikaardi arvutamine alal, kus on teada vaid argumenttunnuste väärtused ja autokovariatsiooni parameetrid, toimub prognoosi iteratiivse sobitamisena. Prognoosi sobitatakse senikaua, kuni on leitud piisavalt hea pind, mis vastab võimalikult hästi nii seostele argumenttunnustega kui ka autokovariatsiooni mõõdikule. Autoregressiooni arvestamise lihtsustatud viis on ümbruse korrektuuri kasutamine pärast hinnangu arvutust regressioonimudelist.

Uurimused

Frelich *et al*. (1993) seadsid metsa koosseisu muutused Markovi ahela mudelis sõltuvusse metsa koosseisust 30 m raadiuses. Teiste liikide vastastikuse mõju hindamiseks kasutati nende koosesinemise andmeid.

Högmander ja Møller (1995) kasutasid linnuatlase naaberruutude andmeid linnuliigi mingis ruudus esinemise või puudumise modelleerimiseks.

Augustin *et al*. (1996) kasutasid punahirve leviku modelleerimiseks lisaks logistilisse mudelisse hõlmatud keskkonna heterogeensusele ka uuritava nähtuse enda ruumilist autokorrelatsiooni. Liigi esinemist mingis kohas seostati prognoositud esinemistõenäosusega ehk sobivusega selle koha ümber. Viimane lisati mudelisse omaette parameetrina. Augustin *et al*. (1996) mudel oli järgmine:

$$\log\frac{p_i}{1-p_i} = a + b_1 alt_i^{\,2} + b_2 nor_i^{\,2} + b_3 mir_i + b_4 mir_i + b_5 pin_i + b_6 auto\,\mathrm{cov}_i, \qquad [5\text{-}41]$$

kus *alt* tähistab maapinna kõrgust merepinnast, *nor* on laiuskraad, *mir* on soode pindala ümbruses, *east* on pikkuskraad, *pin* on looduslike männikute pindala ümbruses, muutuja *autocov* on asustatud ruutude kaalutud keskmine ruudu *i* naabrite hulgas. Kaalud w_{ij} seati võrdeliseks naabri kauguse pöördväärtustega $1/h_{ij}$, kus h_{ij} on ruutude *i* ja selle ruudu naabri *j* vahemaa.

Kui Augustin *et al*. (1996) ja Osborne *et al*. (2001) kasutasid autokovariatsiooni kaaludena kauguse pöördväärtust, siis Hanski (1994) ja Gu *et al*. (2001) on kasutanud ühendatuse mõõte, mis sisaldab vahemaa astmefunktsiooni. Kaaludena on kasutatud ka erinevust objektiga ja objektita pikslite keskväärtuste vahel kaugustsoonide kaupa (Remm ja Oja 2001, Luud ja Remm 2001, Remm ja Luud 2003).

Gu *et al*. (2001) kasutasid hariliku kopsusambliku (*Lobaria pulmonaria*) esinemise/puudumise modelleerimisel läheduse mõõtu C_i

$$C_i = \sum_{j=1}^{n} p_j \exp(-\alpha d_{ij})(A_j + W_j)^b, \qquad [5\text{-}42]$$

kus *n* on võrgustiku lahtrite arv, p_j on lahtri *j* hõivatuse indikaator, α on vahemaast sõltuva mõju vähenemise kiirust määrav konstant, d_{ij} on vahemaa lahtrite *i* ja j vahel, *A* on haabade hulk (suurusega kaalutud), *W* on remmelgate hulk (suurusega kaalutud), b on sobitamiskonstant.

Dennis *et al.* (2002) katsetasid Prantsusmaa liblikate levikuatlase andmetel mitmeid liikide esinemise ja puudumise ning liigirikkuse prognoosimise mudeleid ning leidsid, et enamasti oli ümbruses olevate departemangude omadusi arvestav mudel parem kui vaid sama departemangu geograafilisi omadusi arvestav mudel.

Lichstein *et al*. (2002) andmetes andsid SAR ja CAR mudel kolme Lõuna-Ameerikast pärit laululinnu USAs esinemise kaardistamisel peaaegu identseid tulemusi.

Overmars *et al*. (2003) kasutasid ruumiliselt autoregressiivset mudelit maakasutuse modelleerimiseks Ecuadoris.

Wallerman (2003) kasutas ruumilist regressioonimudelit puidutagavara hindamisel üksikutes kohtades tehtud välimõõtmiste ja satelliidipiltide abil.

Kissling ja Carl (2008) näitasid, et SAR mudeli võimekus sõltub mudeli tüübist ja kauguskaaludest ja andmetes oleva autokorrelatsiooni omapärast. Kohast sõltuvate hälvetega SAR mudel oli tõhusam kui fikseeritud, funktsioontunnusele omase ruumilise autokorrelatsiooni mudel ja ruumiliselt autokorreleerunud argumenttunnustega mudel.

Syartinilia ja Tsuyuki (2008) modelleerisid autologistilise mudeliga jaava tuttkotka (*Spizaetus bartelsi*) levikut lisades mudelisse esinemistõenäosuse pöördkaugusega kaalutud autokovariatsiooni 450 ...1500 m raadiuses olevas ümbruses.

Saracco *et al.* (2010) modelleerisid CAR mudeliga täpikrästa (*Hylocichla mustelina*) ellujäämust ja paiksustõenäosust rõngastamise ja taasleidude järgi.

5.6. Liikide leviku modelleerimine

Liikide leviku kaardistamise traditsiooniline meetod on piiritleda teadaolevate leiukohtade ala liigi levilaks. Levila kujutab liigi leidude piirkonda, kuid ei anna selgust liigi esinemistõenäosuse varieeruvusest areaali sees ega esinda võimalikke esinemiskohti väljaspool seniste leidude piirkonda. Detailsema hinnangulise leviku kaardi saamiseks tuleb ühel või teisel moel modelleerida liigi ökoloogilist nišši, liigi elupaigaeelistusi ja liigi elupaikadele iseloomulikke tunnuseid.

Liikide leviku modelleerimine sai alguse 1970ndate aastate lõpus (Zimmermann *et al.* 2010). Suurem osa moodsast biogeograafiast on kas geograafiliste piirkondade kvantitatiivne võrdlemine (Crovello 1981) või liikide ja elupaikade leviku modelleerimine. Kusjuures viimasel paarikümnel aastal on liigi ja tema elukeskkonna suhete uurimise valdavaks meetodiks kujunenud statistiline modelleerimine, mille kasutusviisid on kokku võetud ülevaateartiklites (Franklin 1995, 1998, Guisan ja Zimmermann 2000, Austin 2002, 2007, Elith *et al*. 2006) ja raamatus Franklin (2009). Teema on haaranud suurt hulka biolooge, geograafe ja infotehnoloogia spetsialiste ning publikatsioonide arv on jõuliselt kasvanud. Teema otsing infosüsteemis *ISI Web of Science* märksõnadega: *"species distribution models" or "niche models" or "habitat models" or "bioclimatic models"* andis 18. veebruaril 2008 vastuseks 21 973 publikatsiooni (Thuiller *et al.* 2009) ja sama tingimusega 6. septembril 2012 tehtud päring juba 52 019 publikatsiooni, neist viimase kümne aast ajooksul ilmunuid 37 041.

Liikide leviku modelleerimisel võib täheldada peavoolu, milles jälgitakse levikumudeli vastavust ökoloogia teooriale, eelkõige ökoniši kontseptsioonile. Uuritakse liigi nõudeid keskkonna suhtes ja kaardistatakse liigile sobivat ala. Alternatiivsed ja peasuunast hälbivad uuringud keskenduvad metoodikale käsitledes leviku kaardistamist eelkõige andmetöötluse, geoinformaatika, statistilise modelleerimise, intellektitehnika, mustrituvastuse või andmekaevandamise ülesandena.

Liikide ja elupaikade leviku modelleerimine aitab igasuguse territoriaalse planeerimise ja territooriumi majandamise puhul otsuseid langetada, aga on samuti kasulik liikide elupaiganõudluste selgitamisel ja modelleeritavate nähtustega seotud informatsiooni süstematiseerimisel ja üldistamisel. Levikuandmed on vajalikud nii liikide ohustatuse hindamisel, nende kaitse planeerimisel, looduse mitmekesisuse kaardistamisel, kahjurite leviku prognoosimisel, maastiku ja kliima muutumise mõjude prognoosimisel ning liikide esinemistõenäosuse hindamisel väheuuritud aladel. Ka uute metsakultuuride rajamisel tasub arvestada koha sobivust ühe või teise puuliigi kasvamiseks näiteks ennustatavate kliimamuutuste järel (Falk ja Mellert 2011).

Guisan ja Thuiller (2005) loetlevad levikumudelite loomise järgmisi eesmärke:

- liigi ökoniši määramine,
- biogeograafiliste, ökoloogiliste ja evolutsiooniliste hüpoteeside kontrollimine,
- liikide edasise leviku ja ohtruse prognoos,
- väheuuritud kohtadele tähelepanu juhtimine,
- liikide introdutseerimiseks ja taasasustamiseks sobivate kohtade leidmine,
- kaitsealade planeerimine,
- elurikkuse kaardistamine üksikliikide kaupa,
- biogeograafiliste regioonide piiritlemine,
- metapopulatsiooni struktuuri määramine ja geenivoolude modelleerimine.

Liikide leviku modelleerimise alaseid uuringuid võib mitmeti klassifitseerida. Ühest küljest saab eristada liikide ökoloogiliste nõudluste määratlemise suunda koos potentsiaalse levila kaardistamisega globaalses, kontinentaalses või ühe riigi mõõtkavas. Need uurimused toetuvad suures osas klimaatiliste faktorite mõjude modelleerimisele ja sageli seavad eesmärgiks kliima muutumise mõju hindamise.

Sellest suunast erinevad liikide esinemise või puudumise hinnangulised kaardistused detailses mõõtkavas ja väiksemal alal, milles kasutatakse eelkõige kaugseire- ja kaardiandmeid. Kui uurimisalal ei ole maapinna kõrguste suurt erinevust, siis on klimaatiliste tegurite varieeruvus väikesel alal tühine ning sellelt alalt kogutud valim ei võimalda liigi esinemist ja puudumist kliima mõõdikutega seostada. Sellises mõõtkavas ei kirjeldata mitte liigi levilat, vaid pigem selgitatakse liigi isendite ja ohtruse ruumilise paiknemise seaduspärasusi, liigi esinemise ja puudumise tõenäosust ja liigi elupaigasobivust. Seejuures on liigi esinemise/puudumise ja elupaigasobivuse kaardistamine sisuliselt sama, sest enamasti saab mõlemal juhul tulemust tõlgendada liigi esinemistõenäosuse hinnanguna.

Suuremõõtkavaline uuring üritab levila kaarti detailiseerida – püüab modelleerida liigi paiknemist levila piiride sees

Kolmanda suunana võib eristada elupaigatüüpide kaardistamist, mille eesmärk ei ole vaid ühe liigi elupaigasobivuse modelleerimine, vaid erinevate elutingimustega alade üldistatum klassifitseerimine ja kaardistamine. Elupaigatüüpide eristamine ja kaardistamine võib toimuda erineval temaatilise ja ruumilise üldistuse tasemel – bioomidest mikroelupaikadeni.

Lähima aja perspektiivsete uurimissuundadena liikide leviku modelleerimisel on esile toodud järgmiste faktorite põhjalikumat arvestamist ja kaasamist mudelisse (Zimmermann *et al*. 2010, täiendatud):

- liigi ja koha arengulugu,
- liigi nõudluste muutumine,
- biootilised seosed,
- liikide levimise võime,
- levilate muutumine inimtegevuse tõttu,
- ruumiliselt detailne kaardistamine suuremal alal,
- valimite esinduslikkuse tagamine vaatluste planeerimisega.

Liigi levik ruumis on määratud tema ökonišiga, leviku võimaluste ja levikulooga. Liikide esinemise modelleerimist keskkonnatingimuste tunnusruumis nimetatakse ökoniši modelleerimiseks, leviku mudeli väljund on kaart. Liigi **põhinišš** (fundamentaalne ehk teoreetiline nišš) sõltub peamiselt liigi füsioloogilistest nõudmistest ja looduses esinevate tingimuste kompleksist. See on kõigi keskkonnatingimuste kombinatsioonide see osa, mille puhul liik suudab konkurentsi ja stohhastiliste negatiivse mõjuga üksiksündmuste puudumisel piiramatu kestusega püsida.

Realiseerunud nišš on keskkonnatingimuste ja organismidevaheliste suhete see osa, mille juures liik tegelikult esineb. Realiseerunud nišš sõltub keskkonna varasemast arengust ja liigi levikuloost. Ökoloogilise modelleerimise puhul seostub põhiniši ja realiseerunud niši vahe levikuprognoosi lähteandmete päritoluga. Põhiniši mudelid peaksid olema tuletatud uuritava organismi füsioloogilistest nõudlustest ja nõudlusi määravatest faktoritest, empiirilised levikuandmed annavad realiseerunud niši mudeli. Põhiniši mudelist saadud kaart näitab potentsiaalse levila ulatust. Põhinišši on nimetatud ka ökoloogiliseks nišiks ja vastandatud elupaikade esinemise/puudumisega korrigeeritud levikuprognoosile. Liigi esinemist oletatakse piirkonnas, kus klimaatilised tingimused vastavad liigi nõudlustele (ökoloogiline nišš) ja kus on liigi elupaiku (Ortega-Huerta ja Peterson 2004).

Põhinišš näitab liigi nõudlusi, realiseerunud nišš vastab tegelikule levikule

Ökoloogilised mõjufaktorid võib jagada kolme klassi: **ressursid**, **kaudsed mõjurid** ja **otsesed mõjurid** (Austin *et al*. 1984). Ressursse tarbitakse otseselt, otsesed mõjurid loovad elutingimusi ja kaudsed mõjurid on elutingimuste ja ressursi olemasolu indikaatorid. Mudeli prognoosiv võime on eeldatavasti suurem, kui mudel sisaldab otseselt mõjuvaid faktoreid. Paraku on nende mõõtmine enamasti keeruline ja kallis. Isegi sellised lihtsad näitajad, nagu maapinna temperatuur ja niiskus igas uuritava ala punktis, tuleb interpoleerida üksikvaatluste põhjal kasutades maapinna kõrgusmudelit ja maakattetüüpi. Otseste mõjufaktorite kasutamist komplitseerib faktorite koosmõju ja otseselt mõjuvate tingimuste raskestimõõdetavus, mis avaldub näiteks elupaigaeelistuse suhtelisuses. Soomes kasvavad turbasamblad mitte ainult siirde- ja kõrgsoos, vaid ka mineraalmaal ja kaljudel. Lõunapoolsed niidutaimed kasvavad Kagu-Eestis liivikutel.

Liikide leviku kujunemisel on oluline ka teiste liikide mõju, sest liigid elavad kooslustes. Eelkõige tuleneb aga koosluste kaupa leviku modelleerimise vajadus suurest liikide hulgast kooslustes, mille tõttu võib kõigi liikide leviku eraldi modelleerimine täiesti ebareaalne olla. Paraku ei pruugi koosluse leviku modelleerimine nii hästi õnnestuda kui üksikliikide leviku kaardistamine (Beselga ja Araújo 2009, Chapman ja Purse 2011), kuigi on ka vastupidiseid näiteid (Zimmermann ja Kienast 1999). Koosluste koosseis ei ole püsiv – ka palaeoökoloogilised uurimused näitavad, et samad liigid ei pruugi teistes tingimustes samu kooslusi moodustada. Kompromissvariandiks võib olla eluvormide kaupa modelleerimine või vaid dominantliikide eraldi modelleerimine (Lenihan 1993, Austin 1998, Guisan ja Theurillat 2000).

Liikidekaupa ehk individualistlik ehk **Gleasoni käsitlus** (Gleason 1926) ja kooslustekaupa ehk **Clementsi käsitlus** (Clements 1916) on modelleerimise dilemmana juba sajandivanune ja seotud koosluste diskreetsuse ja kontiinumi dilemmaga. Individualistliku käsitluse järgi on igal liigil oma nõudlused ja eelistused ümbruse suhtes ning kooslused on vaid elustiku kirjeldamise abivahendid. Clementsi koolkond nägi kooslusi reaalselt eksisteerivate selgepiiriliste super-organismidena (Whittaker 1962).

Koosluste kaupa käsitlus ei ole sageli tingitud mitte niivõrd selgete koosluste olemasolust looduses, kuivõrd tüüpideks jagamise lihtsustavast abist looduse kirjeldamisel ja toimimise mõistmisel. Koosluste kaupa käsitluse korral on modelleeritavaid üksusi vähem ja lihtsam mudel võib olla õigustatud isegi teadmise juures, et kooslused ei ole selgepiirilised ei oma tunnuste ega ruumiliste piiride poolest. Koosluste leviku liikidekaupa modelleerimine eeldab ka liikidevaheliste suhete arvestamist, mis paljuliigiliste koosluste puhul võib ebareaalselt keerukaks osutuda. Kompromissvariant on modelleerida liigikaupa vaid dominantide levik ja lisada kaasnevate liikide esinemisprognoos kooslusena. Kolmas alternatiiv on kas mingi sihtliigi või üksikliikidest sõltumatu elupaigatüüpide kaardistamine (*habitat mapping*, *ecosystem mapping*) ehk elupaigakaartide loomine.

Elupaigakaartide loomisel on statistilise modelleerimise osa üldiselt väiksem kui liikide leviku modelleerimisel, elupaigakaardid on lähedased maakattekaartidele ja seega kuuluvad elupaikade kaardistamise meetodid ruumiandmete kirjeldamise peatükki. Elupaigakaarte on tehtud rohkem maismaa, kuid ka merepõhja kohta (Brown *et al*. 2011, Robinson *et al*. 2011). Ulatuslikult on elupaiku kaardistatud Euroopa Liidu projekti Natura 2000 raames, aga näiteks ka Hollandis (Klijn *et al*. 1996), Kanadas (MacMillan *et al*. 2010), Suurbritannias (NCC 1990, Blackstock *et al*. 2007, Bradter *et al*. 2011) ja mujalgi. Aktuaalne teema on kaugseireandmete senisest tõhusam kasutamine Natura 2000 elupaikade seires (Vanden Borre *et al*. 2011).

Kaardistada saab nii iga üksikliigi levikut, koosluse levikut kui ka elupaigatüüpide levikut

Liikide leviku mudelid on valdavalt staatilised – need ei kirjelda protsesse ja ei lähtu protsessidest, vaid keskkonnatingimustest ja vaatlusandmetest. Dünaamilise modelleerimise jaoks vajalikud parameetrid ei ole enamasti teada ja neid on raske määrata. Dünaamiliste mudelite ja põhjuslike seoste kasutamist liikide ja koosluste leviku modelleerimisel piirab ka ökoniši realiseerumist puudutava teooria ebapiisav areng ja levikuprotsesside vähene tundmine (Austin 2002). Kuidas täpselt on üks või teine liik praegustesse esinemiskohtadesse jõudnud, on selge vaid inimese poolt teadlikult istutatud, külvatud või levitatud organismide puhul.

Suure ja heterogeense empiirilise andmestiku olemasolu aitab teatud määral kompenseerida staatilisust ja suurendada mudeli väljundi usaldusväärsust. Näiteks potentsiaalse taimkatte modelleerimisel on oluline, et andmestik sisaldaks keskkonnatingimuste kõiki põhilisi kombinatsioone. On näidatud, et liikide potentsiaalse leviku statistilisel modelleerimisel suuremal alal on mõistlik ala osadeks jagada ja kas sobitada mudeli parameetrid igas regioonis eraldi või lisada mudelisse regionaalne faktor (Osborne ja Suárez-Seoane 2002).

Liikide leviku mudelid lähtuvad leviku ja keskkonna stabiilsusest ja olemasolevatest leiuandmetest realiseerunud ökoniši kirjeldajatena

Lisaks teoreetilisele mudelile on liikide ja elupaikade leviku modelleerimise väljundiks kaart, mis võib olla kas tõenäoliste esinemiskohtade punktkaart, kaheväärtuseline esinemise/puudumise pind või pideva väärtuspinna (hinnangulise asustustiheduse, esinemise tõenäosuse või elupaiga sobivuse) kaart. Kaartide arvutamine statistilise mudeli abil kuulub **hinnangulise kaardistamise** (*predictive mapping, estimation mapping*) valdkonda. Hinnangulisi kaarte luuakse lisaks statistilistele mudelitele ka eksperthinnangute alusel, sarnaste analoogide järgi ja intellektitehnika meetodite abil. **Hinnanguline kaugkaardistamine** (*remote predictive mapping*) tugineb valdavalt kaugseireandmetele.

Levikut saab modelleerida nii keskkonnatingimuste suhtes kui ka ruumilise paiknemise prognoosina

Empiiriliste andmete põhjal koostatud mudelite kasutamise võimalused teistsuguste keskkonnatingimustega geograafilistes piirkondades on piiratud. Teoreetiliste eelduste empiirilisse mudelisse lisamine aitab mudelit, vähemalt mõnikord, realistlikumaks ja laiemalt rakendatavaks muuta (Malanson *et al.* 1992). Mitme enam-vähem sama hea statistilise sobivusega mudeli puhul tuleks eelistada seda, mis on teoreetiliste teadmiste järgi realistlikum ka väljaspool vaatlusandmeid (Austin 2002).

Ökoloogiliste nähtuste potentsiaalse leviku prognoosi saab anda nähtuse esinemistõenäosuse kaardina, oodatava või võimaliku ohtruse kaardina või sobivuspinnana. Sobivus (*suitability*) on ühest küljest kompleksne näitaja, millesse saab ühendada mitmete mõõdetavate muutujate väärtused, teisest küljest kajastab sobivus keskkonnatingimuste mõju, kolmandaks võib sobivus olla vahend andmete teisendamiseks nominaalse ja pideva kuju vahel. Näiteks nominaalse keskkonnateguri klassidega (mulla tüüp) võib seostada liigi keskmise esinemissageduse (pidev muutuja) ja edasises analüüsis kasutada mullatüüpide asemel nende sobivushinnanguid.

Suuremat ala hõlmavad liikide esinemise/puudumise mudelid saavutavad enamasti parema vastavuse vaatlusandmetega ning võivad elupaigakasutust paremini esindada kui liikide ohtruse mudelid, sest liikide ohtrus on ebastabiilsem ja sõltub rohkem populatsioonisisestest protsessidest ja

vähem elupaiga sobivusest kui liigi esinemine või puudumine (van Horne 1983, Pulliam 1988, Bonn ja Schröder 2001, Frescino *et al*. 2001, Pearce ja Ferrier 2001, Meggs *et al*. 2004). Ruumiliselt detailse kaardistamise puhul ei pruugi esinemiskohtade prognoos ohtruse hindamisest lihtsam olla, sest iga üksikisendi paiknemist määrab nii palju varasemas ajas toiminud faktoreid, et nende koondmõju võib nimetada juhuslikkuseks.

Liigi esinemist või puudumist on üldiselt lihtsam modelleerida kui ohtrust

Liikide populatsioonitihedusel on teatud ruumiline struktuur. Ühe või teise liigi arvukust mõjutavate tegurite teadmine võimaldab prognoosida selle liigi võimalikku arvukust mingite elupaigategurite kombinatsioonide korral. Liikide arvukus on reeglina suurem levila keskkohas ja väiksem levila servaaladel. Liikide asustustihedusel on teatud ruumiline autokorrelatsioon, mille põhjus on suurel määral elupaiga sobivuse ruumiline autokorrelatsioon ehk laigulisus.

Liikide leviku modelleerimisel on oluline arvestada liigi ja elupaigatingimuste seose mõõtkava. Sobiva elupaiga olemasolu on oluline teatud suurusega ruumis, mis on liigiti erinev. Liikide nõudmised ümbrusele on teatud mõõtkavas, mida on võimalik võrdlusuuringutega määrata. Ka ümbrust iseloomustavad kohatunnused, nagu maastiku mitmekesisus, on teatud mõõtkavas. Kusjuures tugevaim seos teatud mõõtkavas ümbruskonna tunnuse ja liigi arvukuse või olemasolu vahel ei pruugi osutada põhjuslikule seosele (de Knegt *et al*. 2010).

Tänapäevased liikide leviku modelleerimise meetodid on kaugel täiuslikkusest. Arenguruumi on nii ökoniši staatilise käsitluse osas, levimisprotsesside ja koha arenguloo arvestamisel, modelleerimismeetodite osas, kohatunnuste valikul, õpetus- ja kontrollvalimite esinduslikkuse kui ka tulemuste kontrollimise meetodite osas (Araújo ja Guisan 2006). Liikide leviku modelleerimise aktuaalne eesmärk on prognoosida kliimamuutuste mõju elurikkusele. Kliima muutumisel võivad aga tekkida sellised keskkonnatingimuste ja koosluste kombinatsioonid, mida kaasajal olemas ei ole. Sellisel juhul ei pruugi kaasaegse kliima, elupaikade ja nende hõivatuse järgi modelleeritud seaduspärasused uutes oludes kehtida (Fitzpatrick ja Hargrove 2009).

Dormann (2007b) süstematiseerib liikide leviku modelleerimisega seotud raskused järgmiselt. Üldised hädad on järgnevad:

- levikut otseselt määravad asjaolud ei ole mõõdetavad,
- levik ei ole seda mõjutavate faktoritega tasakaalus,
- limiteerivate faktorite mõju varieerub eri kohtade vahel,
- levikut mõjutavad protsessid toimivad erinevates mõõtkavades.

Mudeli kasutamise ja ekstrapoleerimise hädad on järgmised:

- limiteerivad faktorid sõltuvad teiste faktorite väärtustest,
- liikidevahelised suhted sõltuvad teiste faktorite väärtustest,
- keskkonna muutumine võib muuta liikide genofondi,
- statistilised seosed ei pruugi kehtida väljaspool õpetusandmeid,
- keskkonna (kliima) muutumine ei ole kõikjal ühesugune.

Statistilise analüüsi hädad on järgmised:

- seosed ei ole lineaarsed,
- indikaatortunnused on omavahel seotud,
- otsesed mõjurid on omavahel seotud,
- vaatlused ei ole ruumilise autokorrelatsiooni tõttu sõltumatud,

- esinemise või puudumise andmed on elupaigasobivuse modelleerimiseks vähekõnekad,
- vähesed õpetusandmed ja ülelihtsustatud mudelid viivad kasutute mudeliteni.

Uurimused

Liikide ja elupaikade leviku ning elupaigasobivuse modelleerimisel kasutatud meetodid on mitmesugused: ordinatsioonimeetodid (Jongman *et al*. 1995, ter Braak ja Prentice 1988), tehisnärvivõrgud (Folse *et al*. 1989, Davey ja Stockwell 1991, Saarenmaa *et al*. 1988), Bayesi mudelid (kogemuse--eelsete ja kogemusejärgsete tõenäosusjaotuste kombineerimine; Milne *et al*. 1989, Pereira ja Itami 1991, Aspinall 1992, 1994), väärtuskaartide ülekatteoperatsioonid (Jensen *et al*. 1992, Brito *et al.* 1999, Wu ja Smiens 2000), kanooniline vastavusanalüüs (Hill 1991), klassifikatsiooni ja regressioonipuud (De'ath ja Fabricius 2000, De'ath 2002, Walker ja Moore 1988, Walker 1990, Grubb ja King 1991, Lees ja Ritman 1991, Moore *et al.* 1991, Debeljak *et al*. 1999, 2001, Hansen *et al*. 2001, Stankovski *et al*. 1998, Kobler ja Adamic 2000), üldistatud lineaarsed mudelid (Puttock *et al*. 1996, Saveraid *et al*. 2001, Gibson *et al.* 2007), sealhulgas logistiline regressioon (Austin *et al*. 1990, Buckland ja Elston 1993, Brito *et al*. 1999, Cowley *et al.* 2000, Brambilla *et al.* 2009, Estes *et al.* 2011), elupaiga eelistuse ja sobivuse indeksid (Duncan 1983, Hansen *et al*. 2001), üldistatud aditiivsed mudelid (Franklin 1998, Vetaas 2000, Guisan *et al.* 2002, Seoane *et al*. 2004), geneetilised algoritmid (Guinan *et al.* 2009), sarnasusele tuginev järeldamine (Gibson *et al.* 2007, Remm ja Remm 2009). Loetelu teiste autorite poolt kasutatud meetoditest esitavad ülevaateartiklites Guisan ja Zimmermann (2000), Elith *et al*. (2006), Heikkinen *et al.* (2006) ja Aitken *et al.* (2007). Ühe varasema ülevaate intellektitehnika kasutustest ökoloogiliste probleemide lahendamisel võib leida S. Džeroski (2001) artiklist.

Ferrier *et al.* (2002b) modelleerisid kõigepealt iga puuliigi leviku ja arvutasid iga puuliigi esinemistõenäosuse kaardi. Seejärel moodustati pikslite klasteranalüüsiga kooslused ja nendest genereeriti koosluste kaart. Zimmermann ja Kienast (1999) uuringus õnnestus alpiniitude koosluste paiknemist täpsemini modelleerida kui üksikliikide levikut.

Bustamante ja Seoane (2004) leidsid, et röövlindude statistiliselt modelleeritud levilad olid informatiivsemad kui varasemad levikukaardid.

Robinson *et al*. (2011) esitavad ülevaate merepõhja biotoopide kaardistamisest Iiri meres.

5.6.1. Saarte biogeograafia tasakaaluteooria

Saarel elevate liikide arvu ja saare pindala vahel on statistiline seos – mida suurem on saar, seda rohkem on seal liike. Järelikult valitseb häirimata kooslustes tasakaal ühelt poolt immigratsiooni ja liigitekke ning teisalt väljasuremise vahel. Seda tasakaalu on matemaatiliselt modelleeritud. Saareliste alade biogeograafia tasakaaluteooriat omistatakse valdavalt Robert H. MacArthurile ja Edward O. Wilsonile (MacArthur ja Wilson 1963, 1967).

Lomolino *et al.* (2006) järgi kirjutas aga lepidopteroloog Eugene G. Munroe juba 1948. aastal oma doktoridissertatsioonis järgmist. "Korrelatsioon saarelt leitud liikide arvu ja saare pindala logaritmi vahel näitab, et peab olema mingi tasakaaluline liikide suurim võimalik arv igal saarel, mis piirab kohaliku fauna suurust. Seda ülempiiri määravad liigitekke, liikide väljasuremise ja sisserände protsessid." Need Munroe põhiseisukohad on koos seose võrrandiga trükis avaldatud juba aastal 1953 (Munroe 1953). Munroe huvitus eelkõige liblikate süstemaatikast, ta ei levitanud oma uudset ideed aktiivselt, ei arendanud seda ja see ei saanud teadlaste hulgas tuntuks. Samuti ei olnud teadlaskond neljakümnendatel aastatel veel valmis omaks võtma uudset matemaatilist lähenemist. Mitmete

konkureerivate liikide kestvale koosesinemisele juhtis enne MacArthurit ja Wilsoni tähelepanu ka G.E. Hutchinson (1958, 1959). Hutchinson tõstatas probleemi, kuid ei pakkunud lahendust.

MacArthur ja Wilson arendasid välja tervikliku teooria, see oli arusaadav ja võeti omaks ka nende poolt, kes ei olnud matemaatikas eriti tugevad. Aastaks 2002 oli MacArthur ja Wilsoni originaalteost tsiteeritud juba üle 3600 korra.

Saarte biogeograafia tasakaaluteooria aluseks on kaks seaduspärasust. Esiteks, liikide arvu ja iga liigi isendite arvu vahel on enamasti lognormaalne seos, mis tähendab, et histogramm, kus on kujutatud liikide arvu sõltuvus liigi esindatuse (isendite arvu) logaritmist on ligilähedaselt normaaljaotuse tihedusfunktsiooni kujuga. Teiseks, saare pindala logaritmi ja liikide arvu logaritmi vahel on lineaarne seos. Saarte biogeograafia tasakaaluteooria eeldab, et immigratsioon ja väljasuremine on pidevad protsessid, mis on omavahel tasakaalus. Suurte saarte või elupaigalaikude liigifond on rikkalikum eelkõige liikide väiksema väljasuremiste sageduse tõttu võrreldes väikeste saarte või elupaigalaikudega.

Saarte biogeograafia kirjeldab tasakaalu liikide väljasuremise ja taasasustamise vahel elupaigalaikudes

Kuigi kõik teooriad paratamatult lihtsustavad tegelikkust, peetakse saarte biogeograafia teooria peamiseks puuduseks liigset lihtsustamist ja formaliseerimist. Reaalsuses on immigratsioon ja väljasuremine oluliselt keerukamad nähtused kui üksikud ühtlased protsessid. Tasakaaluteooria eeldab, et liigid on ühetaolised ning sisserännet ja väljasuremist võib käsitleda stabiilsete juhuslike protsessidena ja ei arvesta kolme olulist asjaolu:

- immigratsioonitõenäosuse sõltuvust saare suurusest ja kujust,
- isolatsiooni arengut pikemas (geoloogilises) ajaskaalas,
- elupaikade hulka ja eripära (ökoloogilist mahtuvust) saartel.

5.6.2. Elupaigasobivuse hinnangud

Tehakse vahet **elupaiga sobivusel** (*habitat suitability*) ja **elupaiga eelistusel** (*habitat preference*). Elupaiga eelistust võib pidada realiseerunud sobivuseks, sobivust aga potentsiaalseks või teoreetiliseks eelistuseks. Kuna ühe või teise koha eelistamine on aktiivne tegevus ja eeldab organismilt reaalset võimekust kohtade vahel valida, siis saab elupaiga eelistusest rääkida eelkõige vabalt liikuvate loomade puhul. Koha sobivus ei eelda aktiivset valikut. Näiteks seemned ja eosed võivad levida kõikjale, kuid arenema hakkavad vaid neile sobivas keskkonnas. Elupaiga sobivust mõõdetakse enamasti liigi esinemistõenäosuse või potentsiaalse ohtruse kaudu.

Välja on töötatud mitmed elupaiga väärtuse hindamise metoodikad, mille tulemus on mingi numbriline väärtushinnang. Selle korrutis elupaiga pindalaga annab **elupaiga mahtuvuse**. Mingi kindla ala kohta võib arvutada keskmise mahtuvuse.

Eeldades, et sarnase elupaiganõudlusega liike on ühes maastikus palju, saab elupaiga sobivust liikide ökoloogiliste gruppide jaoks hinnata **katusliikide** (*umbrella species*) ehk võtmeliikide (*key species*) ehk sihtliikide (*target species*) abil. Katusliigid peaksid olema mitte liiga haruldased, kergesti äratuntavad, suhteliselt stabiilse esinemisega ja peaksid oma ökoloogilistelt nõudlustelt esindama tervet sarnaste nõudlustega liikide gruppi (Simberloff 1998, Bonn ja Schröder 2001).

5.6.2.1. Eksperthinnangud

Kuna kõigi liikide potentsiaalset levikut ei jõuta statistiliselt modelleerida, kasutatakse sobivuse kaardistamisel ka eksperthinnanguid, mille tulemuse võib edasiseks töötluseks numbrilisele kujule viia. Lihtsama käsitluse puhul loetakse sobivaks elupaigaks ala, kus liigi jaoks teadaolevalt olulisemate faktorite väärtus on optimaalses vahemikus. Põhjalikumas uuringus on ekspertidel palutud välja pakkuda funktsioonid ja funktsioonide parameetrid, mille kaudu hinnatakse keskkonnategurite ja elupaiga konfiguratsiooni mõju elupaiga väärtusele. Eksperthinnangute, geoinformaatika vahendite ja statistilise andmetöötluse kombineerimist **mitmel alusel eksperthinnangu** saamisel (*multicriteria evaluation*) on käsitlenud Store ja Kangas (2001). Üksikfaktorite mõju hinnanguid saab omavahel ühendada aditiivselt (Keeney ja Raiffa 1976, Griffiths *et al*. 2011), geomeetrilise keskmisena (Hansen *et al*. 2001), piiravate tingimustena (Neumann *et al*. 2009), logaritmitud korrutisena (Remm *et al*. 2004) või siis kombinatiivselt (Hopkins 1977). Viimasel juhul seotakse sobivused faktori väärtuste kombinatsioonidega. Üksikfaktorite mõjusid on kasutatud ka transformeeritud kujul (Pukkala *et al*. 1997).

Kaardistamisel käsitletakse elupaiga sobivust enamasti pinna pideva omadusena, kasutades pinna piksliteks jagamist. Pinna ebaühtlase väärtuse arvestamiseni on jõutud ka objektikeskse andmemudeli kasutamisel. Store ja Jokimäki (2003) kasutasid integreeritud sobivusindeksi saamiseks mitme liigi jaoks eraldi arvutatud elupaigasobivuse kaartide kombineerimist üheks sobivuskaardiks.

5.6.2.2. Eristava valiku mudelid

Eristava valiku mudelid (*discrete choice models – DCM*) pärinevad majandusteooriast, kus need kirjeldavad indiviidi rahulolu saadaolevate ressursside vahel valimisel. Eeldatakse, et valikuvõimaluse korral eelistab indiviid varianti, mis tagab talle suurima rahulduse. Ökoloogias on **eristava valiku** ehk ressursivaliku mudeleid rakendatud eelkõige mikroelupaiga valiku modelleerimiseks (Cooper ja Millspaugh 1999, Manly *et al*. 2002, McDonald *et al*. 2006, Telesco ja Van Manen 2006, Thomas ja Taylor 2006), aga ka laiemat ala hõlmavates levikumudelites (Nielsen *et al*. 2009, Smulders *et al*. 2010). Eristava valiku mudeli kohaselt eeldatakse, et olend eelistab piirkonna kõige sobivamat kohta.

Eristava valiku mudel näitab eelistusi eelnevalt piiritletud valikuala suhtes

Klassikalises eristava valiku mudelis arvestatakse valikul *i* ressursi *j* valimise tõenäosust üksikfaktorite *x* sobivusmõjude summana.

$$p_{ij} = \frac{\exp\left(\beta_1 x_{1ij} + \beta_2 x_{2ij} + ... + \beta_n x_{nij}\right)}{\sum_k^m \exp\left(\beta_1 x_{1ik} + \beta_2 x_{2ik} + ... + \beta_n x_{nik}\right)}, \qquad [5\text{-}43]$$

kus β on faktori *x* mõju tugevust näitav kordaja, *n* on faktorite arv. Iga ühiku kõigil üksikvalikutel valimise tõenäosused p_{ij} korrutatakse omavahel, eeldades üksivalikute üksteisest sõltumatust.

Esinemiskohale vastanduvad ja DCM sobivusmudeli parameetrite sobitamiseks vajalikud puudumiskohad genereeritakse juhupunktidena alale, kus sama (tüüpilisel juhul liikumisvõimeline) olend veel viibida võiks. Seda ümbritsevat ala nimetatakse **valikualaks** (*choice set*) ja selle piirid tuleb otsustada eelnevalt. Kui valikuala on kogu uuritava ala suurune, siis taandub DCM meetod elupaigasobivuse kaardistamisele regressioonimudelitega.

5.6.2.3. Elupaigaeelistuse indeksid

Elupaigaeelistuse indeksi algne variant (*comparative grazing intensity – c.g.i.*) näitab elupaiga ühikulisest pindalast pärit leidude osa (Hunter 1954, 1962). Duncan (1983) esitas Hunteri indeksi elupaigaeelistusena (*preference*) (P_{1i}) kujul

$$P_{1i} = \frac{U_i}{A_i}, \qquad [5\text{-}44]$$

kus U_i on vaatlusalalt i leitud isendite osa kõigist leitud isenditest, A_i on vaatlusala i osa uuritava ala kogupinnast.

Kui elupaigas on isendite tihedus kaks korda suurem kui uuritaval alal keskmiselt, siis P_i = 2. Kui tihedus on kaks korda väiksem keskmisest, siis P_i = 0,5. Võrdsustamaks indeksi muutumispiirkonda eelistatud ja välditud elupaikades, soovitab Duncan elupaigaeelistuse indeksile logaritmkuju

$$P_{2i} = \log(P_{1i} + 1). \qquad [5\text{-}45]$$

Logaritmitud kujul indeks P_{2i} saab muutuda vahemikus nullist lõpmatuseni. Null näitab elupaiga täielikku vältimist, väärtused, mis on suuremad kui *log*(2), näitavad elupaiga eelistust. Kümnendlogaritmi puhul on eelistuse/vältimise pöördepunktiks indeksi väärtus 0,3. Selleks, et pöördepunkt oleks 0,5 peab logaritmi aluseks olema 4 ning indeksi väärtus 1 tähistaks seejuures kolmekordset eelistust ja 0,21 tähistaks kolmekordset vältimist. Indeks eeldab, et uurimisala on jagatud elupaikadeks.

Liigi suhet pideva muutuja kujul keskkonnafaktoriga saab kirjeldada ka regressioonivõrrandiga, libiseva keskmisega (Remm 1987) või väärtusvahemike sagedusjaotuste võrdlemise abil.

Elupaigaeelistused võivad ajas muutuda. Eelistuse tugevust erinevatel liikidel, erinavatel ajavahemikel või siis erinevatel uurimisaladel saab mõõta ja võrrelda elupaigaeelistuse indeksi varieeruvuse abil. Varieeruvuse mõõtmiseks on palju võimalusi. Duncan (1983) näiteks defineeris **valivusastme** (*degree of selectivity*).

$$S = \sum |U_i - A_i| \qquad [5\text{-}46]$$

Mikroelupaiga kasutust või eelistust on mõõdetud Manly indeksiga (Chesson 1978, 1983, Guido ja Gianelle 2001).

$$a_i = \frac{\frac{r_i}{n_i}}{\sum_i^m \frac{r_i}{n_i}}, \qquad [5\text{-}47]$$

kus r_i on liigi poolt kasutatud elupaiga i osa, n_i on elupaiga i osa maastikus, m on elupaigatüüpide arv.

Braunisch *et al*. (2008) esitasid põhiniši piirimail elunevate liikide elupaiga osa arvestava sobivuse indeksi (*area-adjusted median + extremum, HMae*), mis on lähedane peatükis 5.6.7 esitatud ökoniši faktoranalüüsi metoodikale. Elupaiga pindala arvestamiseks jagatakse liigi leidudega lahtrite arv selle elupaiga kõigi lahtrite arvuga. Arvestamaks, et liik võib konkurentsi või inimmõju tõttu olla sunnitud elunema vaid ökoniši äärealadel, korrigeeritakse esinemiskohtade jaotust keskonnafaktorite jaotusega.

H_{Mae} indeks on seega järgmisel kujul.

$$H_{Mae} = \left\{ \sum_{i=1}^{c} \frac{S_i}{G_i}, S_{keskm} > G_{keskm} \right. \quad [5\text{-}48]$$

$$H_{Mae} = \left\{ \sum_{i=c}^{N} \frac{S_i}{G_i}, G_{keskm} > S_{keskm} \right. \quad [5\text{-}49]$$

Nendes valemites on i keskkonnafaktori väärtusklass, c on keskonnafaktori keskmine väärtusklass, N on väärtusklasside arv, S_i on liigi leiukohtade osa faktori väärtusklassis i, G_i on kohtade osa faktori väärtusklassis i, S_{keskm} on liigi esinemisvahemiku keskväärtus faktori teljel, G_{keskm} on faktori keskväärtus.

Uurimused

Leathwick (1998) leidis, et Uus-Meremaa *Nothofagus* liikide praegune levik ei vasta bioklimaatilistele optimumidele, mis võib olla tingitud varasemast levikust ja aeglasest edasilevimisest.

Estrada-Peña (2001) esitavad ülevaate puukide elupaigasobivuse ja selle kaudu puukide poolt levitatavate haiguste ohu kaardistamisest.

Eesti iga ruutkilomeetri väärtust ökoloogilise võrgustiku jaoks on arvutatud kasutades ruutkilomeetrite andmebaasi andmeid (Remm ja Mander 2001, Remm *et al.* 2004). Sobivusindeksi arvutamise valemid ja kaalud lähteandmetes olevatele kategooriatele valiti subjektiivselt.

K. Remm ja T. Oja (2001) on arvutanud maastiku sobivust maaeluhoonete jaoks mitmese logistilise regressioonanalüüsiga ja arvestanud seda sobivuspinda hoonestuse paiknemise modelleerimisel. Selles uurimuses kasutatakse esinemistõenäosuse pinda sobivuse pinnana.

Neumann *et al.* (2009) kaardistasid kariloomade tiheduse paiknemist Kesk-Euroopas, seostades seda Corine maakattekaardiga ja kliima andmetega.

5.6.3. Tolerantsipiiride kombineerimine

5.6.3.1. Kattuvusanalüüs

Võib eeldada, et iga liik esineb vaid kohtades, kus ükski keskkonnafaktor ei limiteeri tema esinemist – kõigi faktorite väärtused on liigi tolerantsi piires. Kui kasutada liigi **tolerantsipiire** (*environmental envelopes*) kõigi kaardistatud keskkonnategurite suhtes ning faktorite väärtusi igas maastikupunktis, saab moodustada liigi potentsiaalse leviku kaardi. Tolerantsipiire on määratud nii iga faktori puhul eraldi (BIOCLIM meetod) kui ka kumera kattena tunnusruumis liigi esinemisandmete ümber (Walker ja Cocks 1991).

Kattuvusanalüüsi (*overlap analysis, overlay analysis*) saab teha enamuses geoinformaatika tarkvara pakettides. Selle puhul kontrollitakse liigi esinemiskohtade kattuvust keskkonnategurite väärtusklasside levikuga. Eeldatakse, et kui mingi väärtusklassi puhul liiki kordagi leitud ei ole, siis ka teistes analoogilistes kohtades selle väärtusklassi esinemispiirkonnas liik samuti puudub. Liigi esinemistõenäosuse mudelit tasub koostada vaid selle ala kohta, kus liigi esinemine või puudumine ei ole kindlalt ette teada.

Kattuvusanalüüsi on kasutanud näiteks Jensen *et al*. (1992), Brito *et al.* (1999). Prentice *et al.* (1992) kasutasid vaid tolerantsi alumisi piire, eeldades, et ressursifaktorite suurte väärtuste korral

limiteerib liikide esinemist eelkõige konkurents. Kattuvusanalüüsi on kasutatud kaitsealade paiknemise planeerimisel. Näiteks USAs on föderaalse geoloogiateenistuse (USGS) tasemel rakendatud tühimike analüüsi (*Gap Analysis Program – GAP*, http://gapanalysis.usgs.gov), mille käigus loodi Põhja-Ameerika kohalike selgroogsete levikukaardid, mida saab võrrelda olemasolevate kaitsealade võrgustikuga leidmaks kaitsealadega katmata liike (Rodríguez *et al*. 2007).

5.6.3.2. Spetsiaalsed tarkvaralahendused

BIOCLIM (Busby 1986) kombineerib liigi tolerantsipiire kuni 35 klimaatilise faktori suhtes. Tolerantsipiirid määratakse olemasolevatest leiuandmetest risttahuka kujulise osana tunnusruumist. Algoritm sobib prognoosimaks liikide levila muutusi kliimamuutuste korral. Prognoositud levila sõltub kasutatud tunnuste arvust – mida rohkem tunnuseid kasutatakse, seda kitsamad levilad saadakse (Beaumont *et al*. 2005), sest tunnuseid kasutatakse liigi esinemist välistavate faktoritena. Liigi levilaks prognoositakse vaid ala, kus kõik kliimatunnused on liigi tolerantsi piires. Iga tunnuse äärmuslike väärtuste juures saadud leidude eemaldamisega saab määrata liigi levila tuumikala, mis vastab leidude 5…95% kvantiilile faktorite suhtes. BIOCLIM on välja arendatud Austraalias ning kasutust leidnud eelkõige Austraalias, Lõuna-Aafrikas ja Lõuna-Ameerikas (Finch *et al*. 2006, Tsoar *et al*. 2007).

HABITAT (Walker ja Cocks 1991) moodustab leiuandmete ümber tunnusruumi minimaalse ümbritseva hulknurga. Kõiki tunnuseid ei kasutata, valitakse vaid liigi jaoks olulised tunnused.

SPECIES (Pearson *et al*. 2002, 2004) valib tehisnärvivõrgu abil bioklimaatilise vahemiku kliima, mulla ja maakatte tunnustest alguses suuremas biogeograafilise regiooni mastaabis ja seejärel detailsemas kohalike meta-populatsioonide skaalas.

AquaMaps veebilehel (http://www.aquamaps.org) saab päringute järgi moodustatud kaarte mereorganismide leviku ja mitmekesisuse kohta Maailmameres. Kaartide aluseks on liikide teadaolevad tolerantsipiirid soolsuse, sügavuse, primaarproduktsiooni, temperatuuri ja kalda läheduse suhtes.

Liikide tolerantsipiire kombineerivad mudelid määravad eelkõige ökoniši klimaatilisi piire

5.6.4. Regressioonimudelid ja diskriminantanalüüs

Sobivust mõistetakse enamasti pigem pideva ja mitte diskreetse muutujana ning seetõttu on elupaigasobivuse hindamise esmane statistiline vahend regressioonimudel. Üks sagedamini kasutatud statistiline meetod mingi liigi, elupaiga või muu nähtuse esinemistõenäosuse prognoosimiseks on üldistatud lineaarsete mudelite hulka kuuluv logistiline regressioon (ptk 3.4.1.3). Logistiline regressioonimudel võib sisaldada nii reaalarvulisi kui ka nominaalseid seletavaid tunnuseid. Levinud on elupaigasobivuse modelleerimine liigi esinemistõenäosusena kasutades logistilist regressiooni-mudelit. Logistilisest regressioonist paindlikumad on üldistatud aditiivsed mudelid (GAM) ja mitme tunnuse järgi sobituvad seosejooned ehk splainid (*multivariate adaptive regression splines – MARS*).

Regressioonimudelite kasutamisest liikide leviku ja elupaigasobivuse kaardistamisel oli juttu ka peatükis 5.5.5. Geograafiliselt kaalutud regressioonist (GWR) oli juttu interpoleerimise peatükis (ptk 5.2.3.). GWR puhul sobitatakse regressioonimudeli parameetrid omistades igale vaatlusele igas uuritava ruumi kohas vaatluse kaugusest sõltuv mõjukaal. Mudeli parameetrite väärtused ja mudeli prognoosivat võimet näitav determinatsioonikordaja salvestatakse igas arvutatavas kohas ja nii saab

neid omaette kaartidena kujutada, nagu tegid seda Shi *et al*. (2006) valgesaba-hirve leviku modelleerimisel ja Zhang *et al.* (2004) puude läbimõõdu ja kõrguse vahelise seose kaardistamisel.

Diskriminantanalüüs sobiv liikide esinemise ja puudumise kui kahe eristatava klassi kaardistamiseks juhul, kui kõik seletavad tunnused on pidevad ja eeldatavalt normaaljaotusega.

Uurimused

Logistilist regressiooni on kasutanud näiteks He ja Duncan (2000) puude ellujäämuse analüüsil, Araújo ja Williams (2000) liigirikkuse leviku modelleerimisel, Wilds *et al*. (2000) taimekoosluste leviku modelleerimisel, Cowley *et al*. (2000) liblikate maastiku mastaabis leviku modelleerimiseks, Brito *et al*. (1999) ühe sisalikuliigi leviku modelleerimiseks, Neave *et al.* (1996) ning Mörtberg ja Wallenius (2000) linnuliikide esinemistõenäosuse modelleerimiseks, Lindenmayer *et al*. (1999) pärdikuliikide esinemise tõenäosuse modelleerimiseks, Pew ja Larsen (2001) metsatulekahjude tõenäosuse paiknemise modelleerimiseks, Coops ja Catling (2001) metsa struktuuri keerukusklasside sageduse hindamiseks erinevates suktsessioonijärkudes, Loyn (2001) öökullide kui katusliikide esinemise modelleerimiseks.

GAM-mudeleid on liikide leviku modelleerimisel kasutanud näiteks Yee ja Mitchell (1991), Ferrier *et al*. (2002a), Lehmann *et al*. (2002a). MARS tehnoloogiat on liikide leviku modelleerimisel kasutanud Moisen ja Frescino (2002), Yen *et al*. (2004), Elith ja Leathwick (2007). Leathwick *et al*. (2006) võrdlesid GAM ja MARS mudeleid ja leidsid, et prognoosivõimes olulist vahet ei ole.

Potts ja Elith (2006) võrdlesid viit erinevat tüüpi regressioonimudelit sama taimeliigi ohtruse kirjeldamiseks, parima vastavuse vaatlusandmetega andis tõkkega mudel (*hurdle model*), mis koosneb esinemise/puudumise mudelist ja arvukuse mudelist, millele lisatakse esinemise/puudumise eristamise tõenäosusnivoo.

Stepilõokese (*Melanocorypha calandra*) leviku modelleerimisel Hispaanias andis GWR parimaid tulemusi õpetusandmetes ja õpetusandmetega samast piirkonnast pärit kontrollandmetes, kuid mitte teisest kohast pärit kontrollandmetes (Osborne *et al*. 2007). QWR on võimaldanud näidata kaugseireandmetest arvutatud NDVI indeksi ja sademete hulga vahelise seose ruumilist püsimatust (Propastrin *et al.* 2007).

Li *et al.* (1997) koostasid õnnekure (*Grus japonensis*) esinemistõenäosuse mudeli kasutades logistilist ja peakomponentregressiooni.

T.S. Frescino *et al.* (2001) modelleerisid üldistatud aditiivse mudeliga metsa koosseisu Uinta mägedes USAs.

5.6.5. Tinglikke tõenäosusi kasutavad meetodid

Tõenäosuse tinglikkuse annab mingi eelteadmine, tänu millele korrigeeritakse nähtuse väärtuste tõenäosust ja saadakse tõenäosus eelteadmise kehtimisel. Tinglikke tõenäosusi saab kasutada nii oodatavate väärtuste prognoosimiseks kui ka ökoloogiliste seoste kirjeldamiseks. Tinglikke tõenäosusi arvutatakse Bayesi valemi järgi ([1-1], ptk 1.2.1), mis annab nime kogu metoodilisele lähenemisviisile – *Bayesian inference*. Tõenäosuste kombineerimisel eeldatakse faktorite sõltumatust. Tõenäosusjaotuste ja tinglike tõenäosuste kasutamine sobib eriti suhteliselt harva esinevate liikide (nähtuste) modelleerimiseks erineva detailsusega ruumilise andmete järgi.

Tõendikaalud

Tõendikaalu meetodi puhul kasutatakse tingliku tõenäosuse arvutamise valemites tõenäosuste asemel tunnustevahelise seose tugevust iseloomustavaid kaale. Tõendikaalude meetod võimaldab määrata ja ühendada mingit hüpoteesi toetavate tõendite kaalukust. Terminit **tõendite kaalukus** (*weight of evidence*) kasutatakse sageli kujundlikuna, täpset tähendust määratlemata (Weed 2005).

Geoloogid väidavad, et tõendikaalude meetodi leiutas Bonham-Carter (Bonham-Carter *et al*. 1989, Agterberg *et al*. 1990) 1980ndate teisel poolel, kuigi termin on õigusemõistmisel ja meditsiinis ammu tuntud, kus tõendikaalude meetodit on rakendatud haiguste ja riskide diagnoosimisel sümptomite järgi.

Tõendikaalude meetodit saab kasutada kõikjal, kus tõendite ja teadaolevate suhteliste sageduste järgi on tarvis leida kategoorilise muutuja variantide tõenäosust: õigusemõistmisel, äriliste otsuste langetamisel, arheoloogias, metsatulekahjude ohu hindamisel (Dilts *et al*. 2009), maalihete ohu prognoosimisel (Dahal *et al*. 2007, Mathew *et al*. 2007), maavarade otsimisel (He *et al*. 2010), elupaigasobivuse hindamisel (Romero-Calcerrada ja Luque 2006), loomafarmide kaardistamisel (Emelyanova *et al*. 2009).

Tõendikaalud (*weights-of-evidence*) tuginevad tinglikele tõenäosustele. Esinemise tõendikaalud (W_+) ja puudumise tõendikaalud (W_-) arvutatakse eraldi. Iga tõendi kaal määratakse eraldi ja seega ei arvesta faktorite kombineerumisel tekkivaid lisamõjusid. Koosmõjude arvestamiseks peaksid teada olema tõenäosused tõendite kombinatsioonides.

$$W_+ = \ln \frac{P(B/D)}{P(B/D')} \qquad [5\text{-}50]$$

$$W_- = \ln \frac{P(B'/D)}{P(B'/D')} \qquad [5\text{-}51]$$

P(B/D) on nähtuse (liigi) *B* esinemistõenäosus tunnuse *D* esinemisel, *P(B/D')* on nähtuse *B* esinemistõenäosus tunnuse *D* puudumisel, *P(B'/D)* on nähtuse B puudumistõenäosus kohatunnuse *D* esinemisel, *P(B'/D')* on nähtuse *B* puudumistõenäosus tunnuse *D* puudumisel. Tinglikud tõenäosused saadakse olemasolevatest andmetest.

Poolt tõendite kaalud ja vastu tõendite kaalud summeeritakse eraldi. Võrreldavaid variante võib olla ka rohkem kui kaks. Erinevate tõendite järgi arvutatud tõendikaalude summeerimisel on oluline jälgida tõendite sõltumatust.

Tõendi kaal on sündmuse tõenäosus tõendi olemasolu korral

Tõendite kontrast on poolt ja vastu hinnangukaalude vahe, mille märk (positiivne või negatiivne) väljendab nähtuse tõenäolisemat esinemist või puudumist ning absoluutväärtus näitab hinnangu kindlust.

WhyWhere

Ökoniši modelleerimise tarkvara *WhyWhere* otsib liigi leviku kirjeldamiseks sobivaid indikaator-tunnuseid paljudest saadaolevatest kaugseire ja kliimaandmetest, mida on kokku üle tuhande erineva globaalse andmekihi. Kõiki andmekihte käsitletakse kujutistena, mille pikslivääruused jagatakse klasteranalüüsi meetodite abil võrdse mahuga klassidesse ehk tüüpidesse. Leidude hulga ja tüübi levikuala suuruse alusel määratakse liigi esinemise tõenäosus igas kohatüübis. Liigi esinemis-

tõenäosuse määramiseks uuritavas kohas leitakse selle koha tüüp ja kasutatakse prognoosina tüübile vastavat liigi esinemistõenäosust. Tunnuste grupeerimise kaudu suudab *WhyWhere* arvestada faktorite koosmõjusid. Aastaks 2007 oli *WhyWhere* andmebaasis üle tuhande andmekihi. Parima mudeli otsingut saab taasalustada niipea, kui uus andmekiht lisandub vabalt saada olevate andmekihtide hulka. Enamasti kasutab süsteem prognoosi arvutamiseks vaid kahte-kolme formaalset indikaator-tunnust (Stockwell 2006, 2007).

WhyWhere valib sadadest kohatunnustest paar indikaatortunnust ja seostab liigi esinemistõenäosuse nende tunnuste väärtusvahemikega

Süsteemi kirjelduse kohta on ilmunud kriitiline artikkel (Peterson 2007), milles väidetakse, et kõikide saadaolevate andmekihtide kasutamiseks ei ole vajadust, paarist indikaatortunnusest liigi nõudluste kirjeldamiseks ja leviku modelleerimiseks aga ei piisa ning et *WhyWhere* ei anna usaldusväärseid levikukaarte.

Maxent

Statistilisest mehhaanikast pärit **maksimumentroopia printsiip** (*the principle of maximum entropy*) on üldine meetod järelduste tegemiseks ebatäielikest andmetest. Maksimumentroopia printsiibi kohaselt tuleks jaotuse lähendamisel etteantule esmalt jälgida kõiki etteantud piiranguid ning nende täitmisel eelistada võimalikult ühtlast (maksimaalse entroopiaga) jaotust. See on ainus hälbeta hinnang, sellest kõrvalekaldumine tähendaks meile mitte teadaoleva informatsiooni eeldamist (Jaynes 1957). Kuni teave jaotuse kohta puudub, siis on kõik variandid võrdse tõenäosusega – jaotus on ühtlane. Maksimumentroopia printsiip on vastavuses parsimoonireegliga – ilma tõenditeta ei ole põhjust lähendatavat jaotust muuta. Maksimumentroopia printsiip on aktuaalne teema küberneetikas ja tehisõppes. Selle kohta on veel vähe õppekirjandust, kasutamisjuhiseid ja rakendamise tarkvara.

Maxent on firmas AT&T Labs ja Princetoni ülikoolis loodud vabavara (Phillips *et al*. 2004, 2006), mis on vabalt alla laaditav aadressilt http://www.cs.princeton.edu/~schapire/maxent.

Maxent teisendab tunnused elementaartunnusteks (*feature*) ja võrdleb elementaartunnuste väärtuste jaotust uurimisalal *f(z)* (taustjaotus) ja samade tunnuste jaotust esinemiskohtades *f1(z)*. Taustjaotus on tunnuse esinemiskohtade hulgas jaotumise nullmudel. See on tunnuste tinglik jaotus nähtuse esinemise tingimusel. Kui ühtegi leidu ei ole, tuleb eeldada esinemiskohtade juhuslikku paiknemist. Leiuandmete olemasolu korral on *f(z)* ja tunnuse väärtuste jaotus leiukohtade hulgas *f1(z)* erinevad (joonis 5-22). Taustajaotust *f(z)* korrigeeritakse tunnustest sõltuvalt hoides jaotuse *f1(z)* keskväärtuse tegelikuga vastavuses.

$$f1(z) = f(z)e^{n(z)} \qquad [5\text{-}52]$$

$$\ln[f1(z)] = \ln[f(z)] + n(z) \qquad [5\text{-}53]$$

$$\eta(z) = \alpha + \beta h(z), \qquad [5\text{-}54]$$

Valemid 5-52 ja 5-53 on samatähenduslikud; $\eta(z)$ on taustjaotuse korrektuur, mida Maxent iteratiivselt optimeerib, z tähistab tunnusvektorit (kõigi elementaartunnuste väärtusi) prognoositavas kohas. α on konstant, mis hoiab *f1(z)* tõenäosuste summa võrdse ühega, $\beta h(z)$ on tunnuse väärtusest sõltuvad aditiivsed kaalud.

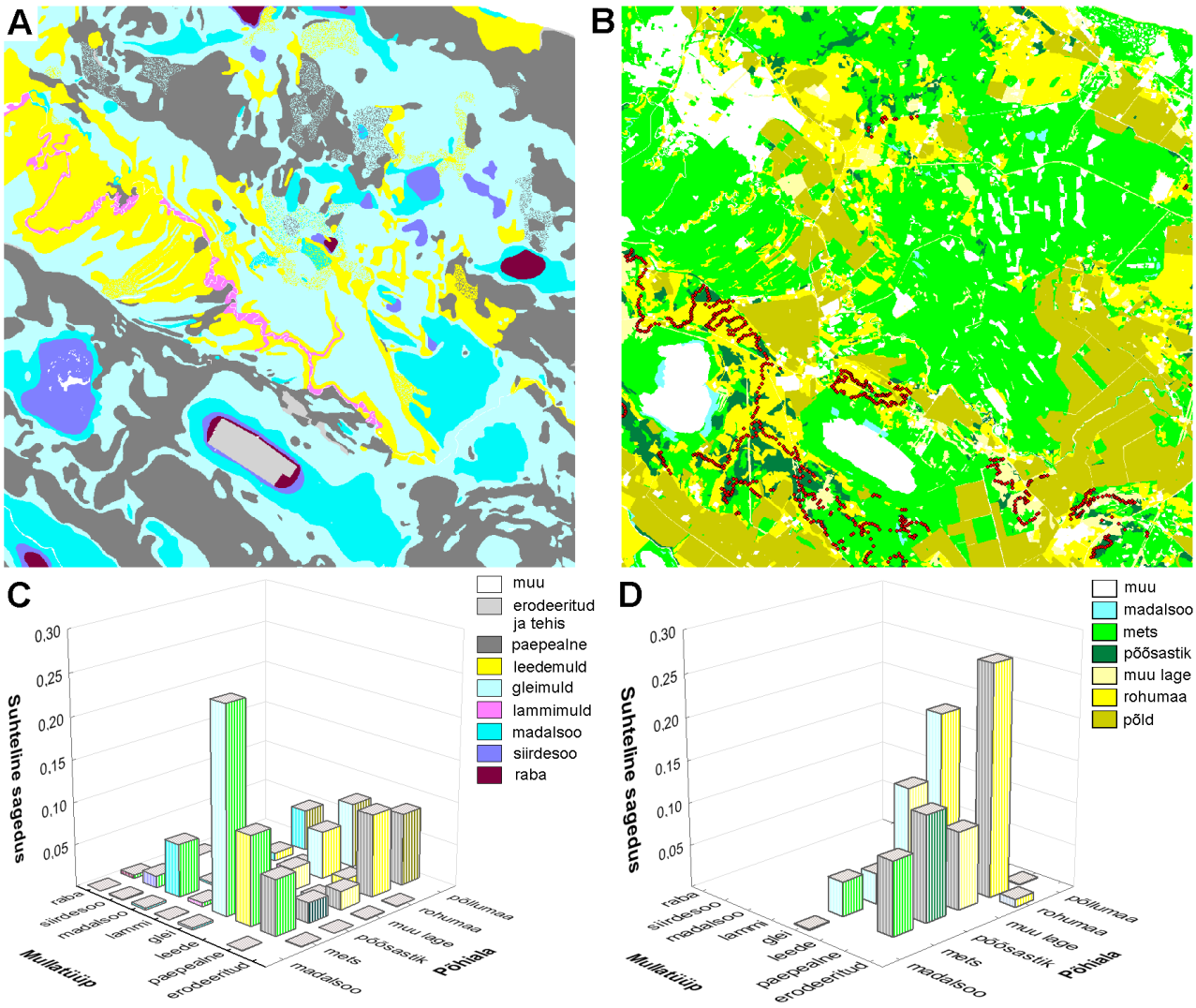
Maxent alustab kaalude ükshaaval muutmist tõenäosuste ühtlase jaotusega. Kuna kaalude nihutamisel ei kasutata juhuslikkust, on tulemus samadest andmetest alati sama. Maxent hinnangud sõltuvad mitte ainult leiukohtadest, vaid ka taustjaotusest – see tähendab sõltuvad uuritava ala piiridest või tausta iseloomustavate kohtade valikust juhul, kui see jäetakse uurija teha. Samad leiuandmed teistsuguse taustjaotuse suhtes annavad mõnevõrra teistsuguse levikukaardi. Liikide leviku modelleerijatele suunatud Maxent algoritmi põhjalik selgitus on artiklis Elith *et al.* (2011).

Maxent vastandab kohatunnuste väärtuste sagedusjaotuse uurimisalal ja leiukohtades ning seostab nende jaotuste erinevuse tunnuste eksponentsiaalse kujuga

Maxent on:

- on kasutatav ka väheste (<10) leiukohtade puhul,
- on hea eristamisvõimega võrreldes teiste meetoditega,
- suudab kasutada nii numbrilisi kui ka kategoorilisi tunnuseid,
- ei vaja puudumiskohtade teavet, kuid võib ka neid kasutada,
- ei sea mingeid eeltingimusi tunnuste jaotuse ja seoste tüübi osas,
- annab tulemuseks esinemistõenäosuse vahemikus 0 kuni 1.

Maxent on mitmetes leviku modelleerimise uuringutes osutunud teistest meetoditest tõhusamaks, eriti suhteliselt haruldaste liikide puhul (Hernandez *et al.* 2006, 2008, Phillips *et al.* 2006, Wisz *et al.* 2008).



Joonis 5-22. Üldistatud mullatüüpide (A) ja põhikaardi põhialade (B) levik Keilast kirdepool oleval kaardilehel 6382 koos põõsasmarana registreeritud esinemiskohtadega (punased täpid), põhialade ja mullatüüpide kombinatsioonide suhteline sagedus kogu kaardilehel (C) ja põõsasmarana esinemiskohtades (D). Põõsasmaran ei esine raba- ja siirdesoomullal ning on harva põllumaal.

Uurimused

Griffiths *et al.* (1993) üritasid mitmese regressiooniga prognoosida mõnede linnuliikide arvukuse sõltuvust maakatte ja paiknemise parameetritest. Edukam oli liikide esinemise/puudumise vahekorra prognoos klassifitseeritud andmetest. Iga väärtusklassi kohta arvutati liigi esinemise tinglik tõenäosus. Tinglikud tõenäosused kombineeriti nendes kohtades, kus liigi leviku kohta andmed puudusid, üheks lõpptõenäosuseks Bayesi valemi abil. Liigi esinemise algtõenäosuseks võeti tundmatutes ruutudes 0,5.

Tinglike tõenäosuste kasutamist liikide leviku kaardistamisel on kirjeldanud Aspinall (1992, 1993, 1994) ja Skidmore *et al*. (1996).

Yang *et al.* (2006) modelleerisid heinikute (*Tricholoma spp*) esinemiskohti tõenäosusi kombineeriva ekspertsüsteemi abil ja leidsid, et see andis paremaid tulemusi kui logistiline esinemise/puudumise mudel.

Termansen *et al*. (2006) kasutasid tinglikke tõenäosusi ökoniši mudelite moodustamiseks.

Romero-Calcerrada ja Luque (2006) kasutasid tõendikaalu (*weights-of-evidence*) meetodit

kolmvarvas-rähni pesitsuskohtade modelleerimiseks.

Maxent tarkvara on liikide leviku modelleerimisel lisaks Maxent alapeatükis mainitud publikatsioonidele kasutanud näiteks Elith *et al*. (2006), Gibson *et al.* (2007), Guisan *et al*. (2007), Parolo *et al.* (2008), Gontier *et al.* (2010), Gogol-Prokurat (2011). Maxent tarkvara abil liikide leviku mudeleid loonud uurimustest annavad ülevaate Elith *et al*. (2011).

Remm (1989) kirjeldas ekspertsüsteemi, milles vee troofsuse väärtusvahemike tinglikud tõenäosused olid teisendatud suhtelisteks tõenäosusteks. Koondhinangu saamisel erinevate tunnuste põhjal saadud suhtelised tõenäosused korrutati. Korrutamine võimaldab arvesse võtta ühe või teise faktori limiteerivat mõju, sest nulliga korrutamisel on tulemus null. Süsteem võimaldas kaasata olemasolevaid teadmisi prognoositava tunnuse sagedusjaotuse kohta ja arvestada rannikumere piirkonda. Kahjuks ei leidnud see Lääne-Eesti rannikumere andmetele tuginev Apple arvutile programmeeritud ekspertsüsteem kasutamist, sest järgnesid pöördelised ajad ja Zooloogia ja Botaanika Instituudi merebioloogia uurimisrühm siirdus jõgesid uurima. Programmi Turbo Pascal keeles koodi väljatrükk seisuga september 1990 on hoiul TÜ geoinformaatika õppetoolis.

5.6.6. Klassifikatsiooni- ja regressioonipuud

Klassifikatsiooni- ja regressioonipuudest üldiselt on juttu peatükis 3.4.2. Klassifikaatori võimendamist (*boosting*) ja sellele tuginevat klassifitseerimismeetodit võimendatud klassifikatsioonipuud (*boosted classification trees)* kirjeldatakse koos teiste ansamblimeetoditega peatükis 5.6.10.1.

5.6.6.1. CART

CART (*Classification And Regression Trees*) võib tähistada nii klassifikatsiooni ja regressioonipuude andmetöötlusmetoodikat kui ka eraldiseisvat tarkvarapaketti või ühte osa suuremas tarkvarapaketis. Klassifikatsioonipuu ülesanne leviku modelleerimisel on kohtade jagamine liigi arvatavateks esinemis- ja puudumiskohtadeks, regressioonipuu ülesanne on pideva muutuja – liigi ohtruse või esinemistõenäosuse hindamine. CART ei sea eeltingimusi seletavate tunnuste tüübile ega väärtuste jaotusele, kuid ei paku võimalust mudeli olulisuse hindamiseks. Regressioonipuud on lihtsasti mõistetavad, kuid tundlikud väikeste muutuste suhtes andmetes. Vaid veidi muudetud õpetusandmed võivad anda täiesti erineva ülesehitusega otsustepuu.

Kaubamärk CART kuulub firmale *California Statistical Software, Inc*, tarkvararakenduse litsentsi omab *Salford Systems,* mille kodulehelt saab programmi CART tasuta prooviversiooni ja kasutamisjuhiseid alla laadida. Programmis Statistica on klassifikatsiooni ja regressioonipuude moodul nime all C&RT.

5.6.6.2. GARP

Erinevalt klassikalistest otsuste puudest opereerib geneetilisi algoritme kasutav tarkvara **GARP** (*Genetic Algorithm for Rule Set Production*) mitte üksikkriteeriumitega, vaid reeglistikega (*rule set*), mis on suvalises järjekorras kasutatavate eri tüüpi reeglite komplektid (Stockwell ja Noble 1992, Stockwell ja Peters 1999). Samast andmestikust saab genereerida väga palju erinevaid reeglistikke. GARP alustab mingist reeglistikust ja hakkab selle parameetreid muutma. Iga muutuse järel kontrollitakse reeglistiku abil saadud prognooside vastavust vaatlusandmetele. Kui reeglistik kuulub prognoositäpsuselt seni parimate hulka, siis imiteerides looduslikku evolutsiooniprotsessi säilitatakse

selle reeglistiku parameetrid. Seejärel moodustatakse järgmise põlvkonna reeglistikud kasutades heuristilisi operaatoreid: lisades uusi juhuslikke reegleid, parameetrite juhuslikku muutmist (mutatsiooni) ja reeglite ühendamist. Sama põlvkonna reeglistikke moodustatakse ja testitakse paralleelselt. Kuna geneetilised algoritmid sisaldavad juhuslikkust, siis on samadest andmetest korduvalt arvutatud tulemused iga kord mõnevõrra erinevad.

GARP-mudel, nagu teisedki tehisintellekti meetodid on osutunud kasulikuks paljutunnuseliste, mürarohkete, mitme optimumiga ja ebakorrapärase kujuga optimume sisaldavate andmestike puhul. Eriti sobib GARP hajusarvutust võimaldavatele süsteemidele. On leitud, et GARP on vähemtundlik ülesobitumise ja õpetusandmete vähesuse suhtes kui logistiline regressioon (Stockwell ja Peterson 2002).

GARP tarkvara on saadaval aadressil http://www.nhm.ku.edu/desktopgarp/.

Uurimused

Ülevaate otsuste puu meetodist annavad D.M. Moore *et al.* (1991), kes kasutasid seda metsakoosluste kaardistamisel. Otsuste puuga on modelleeritud ka punahirve ja pruunkaru elupaigasobivust (Kobler ja Adamic 2000, Debeljak *et al.* 2001, Stankovski *et al*. 1998).

G. De'ath ja K. Fabricius (2000) tutvustavad regressioonipuude kasutamist koralliliikide leviku elupaigatingimustest sõltuvuse näitel. Regressioonipuid on bioloogias kasutanud veel Staub *et al.* (1992), Baker *et al*. (1993) ja Rejwan *et al*. (1999).

F. Huettmann ja A.W. Diamond (2001) selgitasid kõigepealt mitmefaktorilise logistilise regressiooniga, millised tegurid mõjutavad oluliselt merelindude pesilate paiknemist ja siis kasutasid neid tegureid regressioonipuu klassifikaatoritena.

J. Miller ja J. Franklin (2002) katsetasid klassifikatsioonipuu ja indikaatorkrigingu kombinatsiooni taimkatte kaardistamisel Mojave kõrbes. Igas kohas olev kõige tõenäolisem taimkatteüksus prognoositi selle koha tunnuste (kõrgus, keskmine õhutemperatuur, nõlvakalle, pinnavorm, niiskuseindeks) järgi.

G. De'athi (2002) ülevaateartikkel selgitab ühe- ja mitmemõõtmeliste regressioonipuude kasutamist. Näiteid regressioonipuude ja lähedaste meetodite kasutamisest on S. Džeroski (2001) artiklis.

L.R. Iverson *et al*. (1999) kasutasid regressioonipuudega analüüsi *Pinus virginiana* leviku ja keskkonnatingimuste vahel. Seoseid kasutati koos puude paiknemismustri seaduspärasustega ja stohhastilise migratsioonimudeliga puu leviku prognoosimisel kliimamuutuste järel.

Samblikuliikide esinemist on klassifikatsioonipuuga prognoosinud Edwards *et al*. (2006).

Programmiga CART arvutatud lihtsaid regressioonipuid on kasutatud metsakahjustuste tõenäosuse hindamisel (Coops *et al*. 2006).

Geneetilised algoritmid programmis GARP on leidnud kasutust näiteks korallide leviku kaardistamisel (Guinan *et al*. 2009).

5.6.7. Ökoniši faktoranalüüs

Ökoniši faktoranalüüs (*ecological niche factor analysis – ENFA*) on algselt N. Perrini (1984) poolt välja pakutud elupaigasobivuse ja liikide oodatavate esinemiskohtade prognoosimise metoodika, mille on Lausanne ülikoolis Šveitsis detailsemalt läbi töötatud, nime andnud ja tarkvaraks Biomapper vorminud A.H. Hirzel (Hirzel 2001, Hirzel *et al.* 2002, http://www.unil.ch/biomapper).

ENFA kasutab kohatunnuste faktoranalüüsi abil saadud üldistatud telgi, mida saab käsitleda statistiliselt sõltumatute faktoritena. Ökoniši kirjeldamiseks kasutatakse vaid liigi esinemiskohti või

levikuatlase ruutude või muid eraldisi, kus liiki on registreeritud võrreldes neid kogu uurimisalaga. Liigi teadaolevate või oletatavate puudumiskohtade tunnuseid ei kasutata. Faktoranalüüsil saadud telgedest esimene (*marginality factor*) paigutatakse andmestiku tunnusruumi nii, et see läbiks nii kogu uurimisala esindava valimi kui ka liigi esinemiskohti esindava valimi keset (joonis 5-23A). Iga järgmine telg paigutatakse kõigi eelmistega risti nii, et see maksimaalselt esindaks valimi jääk-varieeruvust (*specialisation factors*).

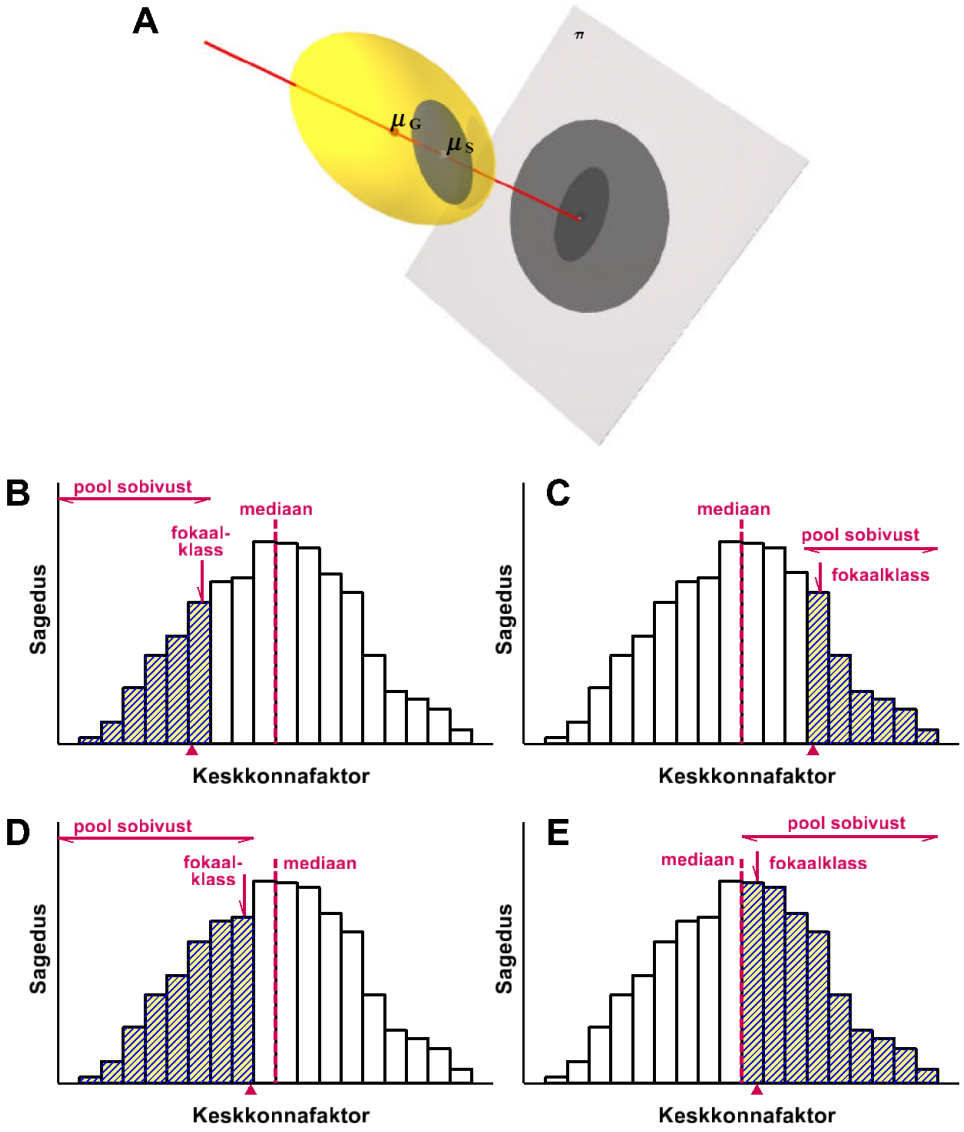
Ökoniši faktoranalüüs üldistab kohatunnused omavahel sõltumatuteks faktoriteks ja võrdleb faktorite väärtuste jaotust liigi leiukohtades ja kõigis uuritava ala kohtades

Ökoniši faktoranalüüsil saadud telgi kasutatakse tarkvaras Biomapper elupaigasobivuse indeksi arvutamiseks ja elupaigasobivuse kaartide loomiseks selle indeksi väärtuste alusel. Elupaigasobivuse indeks arvutatakse esmalt eraldi iga faktori suhtes. Selleks jagatakse faktori väärtused liigi leiukohtades väärtusvahemikesse nii, et mediaan jääks kahe väärtusklassi piirile. Siis loendatakse nende leiukohtade hulk, mis on samas väärtusklassis või jaotuse samal poolel kaugemal faktori mediaanist ja korrutatakse see kahega (joonis 5-23B-E). Arvestada tuleb vaid mediaanist samal poolel olevaid sagedusi ja neid kahega korrutada, kuna empiirilised sagedusjaotused enamasti ei ole sümmeetrilised. Seejärel normeeritakse mediaanist sama kaugel või kaugemal olevate leiukohtade kahekordne summa nii, et sobivusindeks oleks alati vahemikus nullist üheni. Selleks jagatakse saadud vahetulemus leiukohtade koguarvuga.

Koha sobivuse koondhinnang saadakse osahinnangute kaalumata või kaalutud keskmisena. Kaaludeks võivad olla faktorite omaväärtused. Elupaik on selle liigi jaoks maksimaalselt sobiv, kui koha omadused vastavad öko-geograafilise faktori väärtuste mediaanile liigi leiukohtades. Koht on täiesti ebasobiv, kui mediaanist samas suunas niivõrd erinevate tingimuste korral ei ole liiki kordagi leitud.

Ökoniši faktoranalüüs kasutab vaid liigi leiukohtade ja kogu uuritava ala iga koha tunnuseid, liikide puudumisandmed ei ole vajalikud. Registreeritud esinemiskohad ei pruugi aga olla esinduslik valim liigi tegelikust esinemisalast. Samuti ei pruugi uuritav ala adekvaatselt esindada keskkonnatingimuste võimalike väärtuste sagedust. Seega, nagu teistegi mudelite puhul, sõltuvad ökoniši faktoranalüüsi tulemused kasutatavate andmete esinduslikkusest ja uuritava ala piiritlemisest.

Kui liigi asustustihedus ületab optimumi ja liik esineb ka väljaspool sobivaid elupaiku, võib logistiline regressioon liigi puudumise andmete kasutamisega anda täpsemaid hinnanguid kui ENFA (Hirzel *et al.* 2001). Liigi esinemise sagedusjaotuse kasutamine võib viia eksiteele ka siis, kui liik on uuritavat ala alles asustamas või kui kõik vaatlusandmed pärinevad väljastpoolt liigile sobivate keskkonnatingimuste optimumi. Sellisel juhul klassifitseeruvad liigile kõige sobivamad kohad sobimatuteks, sest tõeliselt sobivatest kohtadest pole uuritava liigi esinemist kordagi registreeritud (Sachot 2002). Liigi puudumisandmete kasutamine on eriti oluline geograafiliselt laialt levinud ja elupaiga suhtes vähenõudlike liikide puhul. Ubikvistide leviku prognoosikaartide täpsust on raske hinnata sõltumata prognoosimeetodist (Brotons *et al.* 2004).



Joonis 5-23. A – ökoniši faktoranalüüsi kolmemõõtmeline tõlgendus (enamasti on kasutatavaid tunnuseid paarkümmend). Kollane pilv esindab kaardistatava ala kõiki kohti, kollase pilve sees olev hall pilv liigi leiukohti. Nende pilvede keskkohti (μG ja μS) läbiv sirge on marginaalsusfaktor. Selle faktoriga seotud kohatunnuste varieeruvust iseloomustavad pilvede projektsiooni ellipsid sirgega risti oleval tasandil π.
B – elupaigasobivuse arvutamine liigi esinemiskohtade sageduse järgi, mis sõltub leiukohti iseloomustavast keskkonnafaktorist. Keskkonnafaktor jagatakse klassidesse nii, et mediaan jääks klasside piirile. Kaardistatava ala igas kohas leitakse faktori seda kohta iseloomustavale väärtusele (kolmnurk horisontaalteljel) vastav klass (fokaalklass). Liigi esinemissageduse jaotuses mediaanist sama kaugel või samas suunas kaugemal olevate klasside sagedused (viirutatud tulpade pind) liidetakse, korrutatakse kahega ja jagatakse kõigi sageduste summaga (histogrammi kogupinnaga). Saadakse koha sobivus sellele liigile selle keskkonnafaktori osas. Sobivusindeksi väärtused joonisel olevates näidetes: A – 0,44; B – 0,41 C – 0,79; D – 1,00. (Hirzel 2001, Hirzel *et al*. 2002, muudetud).

Uurimused

Ökoniši faktoranalüüsi abil on kaardistatud näiteks mägikitse (*Capra ibex*) (Hirzel 2001, Hirzel *et al*. 2002), hiirte (Reutter *et al*. 2003), Uus-Meremaa sõnajalaliikide (Zaniewski *et al*. 2002) ja metsise (Sachot 2002), Kataloonias pesitsevate lindude (Brotons *et al*. 2004), ninasarvikmardikate (Chefaoui *et al*. 2005) ja pandakaru (Viña *et al*. 2008), lataste rästiku (*Vipera latastei*) (Santos *et al*. 2006) elupaigasobivust.

5.6.8. Leviku kaardistamine sarnasuse järgi

Sarnasusele tuginev hinnanguline kaardistus eeldab, et otsitav liik või muu objekt esineb sagedamini kohtades, mis on sarnased kohtadele, kus seda on juba varem registreeritud, võrreldes kohtadega, kust leide ei ole. Sarnasusele tuginevates hinnangutes ei vajata kõiki vaatlusandmeid. Osa vaatlusi dubleerivad üksteist, osa on erandlikud, osa aegunud ja mõned ehk mitte piisavalt usaldusväärsed. Sarnasuse hindamisel tuleb alati arvestada küsimust: sarnasus mille poolest? Sarnasus ühtede tunnuste poolest ei ole sama, mis sarnasus teiste tunnuste poolest. Kõik tunnused ei ole vajalikud, osa tunnuseid dubleerivad üksteist, osa ei seostu otsitava nähtusega ning mõned andmekihid on osaliselt aegunud. Kõik kaardid sisaldavad üldistust, aga üldistustase ja üldistamise viis on erinevad, kaugseireandmed sõltuvad sensorist, atmosfääri ja taimkatte seisundist. Seega, kõik kaugseireandmed ja kaardid võivad sisaldada juhuslikke vigu.

Sarnasuse järgi hindamisel tuleb alati otsustada, milliste tunnuste poolest sarnasust hinnatakse

Erineval faktoranalüüsi ja regressioonanalüüsi tüüpi mudelitest ei eelda leiukohtadega võrdlemine ökoniši pidevust tunnusruumis. Samuti sõltub sarnasuse järgi hindamine suhteliselt vähe ökoniši teooriast, mille puudulikkust ja mittearvestamist leviku modelleerimisel on korduvalt kurdetud (Araújo ja Guisan 2006, Austin *et al.* 2006, Austin 2007).

5.6.8.1 DOMAIN

DOMAIN (Carpenter *et al.* 1993) omistab elupaigasobivuseks koha sarnasuse liigi leiukohtadest kõige sarnasemaga. Uuritavad tunnused standardiseeritakse jagades nende väärtused tunnuse väärtuste haardega uurimisala piires. Seejärel kasutatakse sarnasuse mõõduna negatiivset vastandväärtust keskmisest tunnuste erinevusest kahe võrreldava koha vahel. Hinnang varieerub üldjuhul vahemikus 0...1, kuid võib olla negatiivse väärtusega vaadeldavast alast väljas pool olevate kohtade puhul, mida ei ole kasutatud tunnuse ulatuse hinnangus.

5.6.8.2. D^2

Mahalanobise vahemaa ehk kaugus (D^2) on hajuvustega standardiseeritud ruuterinevus muutujate vahel. D^2 on sarnasusele tuginev meetod, mis mõõdab iga koha sarnasust leiukohtade statistilise keskmisega, mitte üksiknäidistega, nagu näidistele tugineva järeldamise puhul tarkvarades DOMAIN ja Constud (ptk 3.4.6.1 ja 5.6.8.3). Erinevalt eukleidilisest vahemaast on vahemaa Mahalanobise ruumis sihist sõltuv – ühikulise vahemaa piki tunnuse telge määrab tunnuse hajuvus. Statistilise vahemaa määramine eeldab tunnuste pidevust, kuid ei sea eeldusi tunnuste jaotustele. D^2 näitab elupaiga sarnasust (*habitat similarity*), mitte liigi esinemistõenäosust antud kohas.

D^2 mõõdab koha sarnasust liigi leiukohtade statistilise keskmisega

D^2 sõltub tunnuste hulgast ja tunnuste omavahelisest korreleerumisest. Tunnuste liiasust ja interkorreleerumist on püütud eemaldada Mahalanobise vahemaa aditiivseteks osadeks jagamisega vastavalt tunnuste peakomponentanalüüsil leitud omaväärtustele (Browning *et al.* 2005, Rotenberry *et al.* 2006).

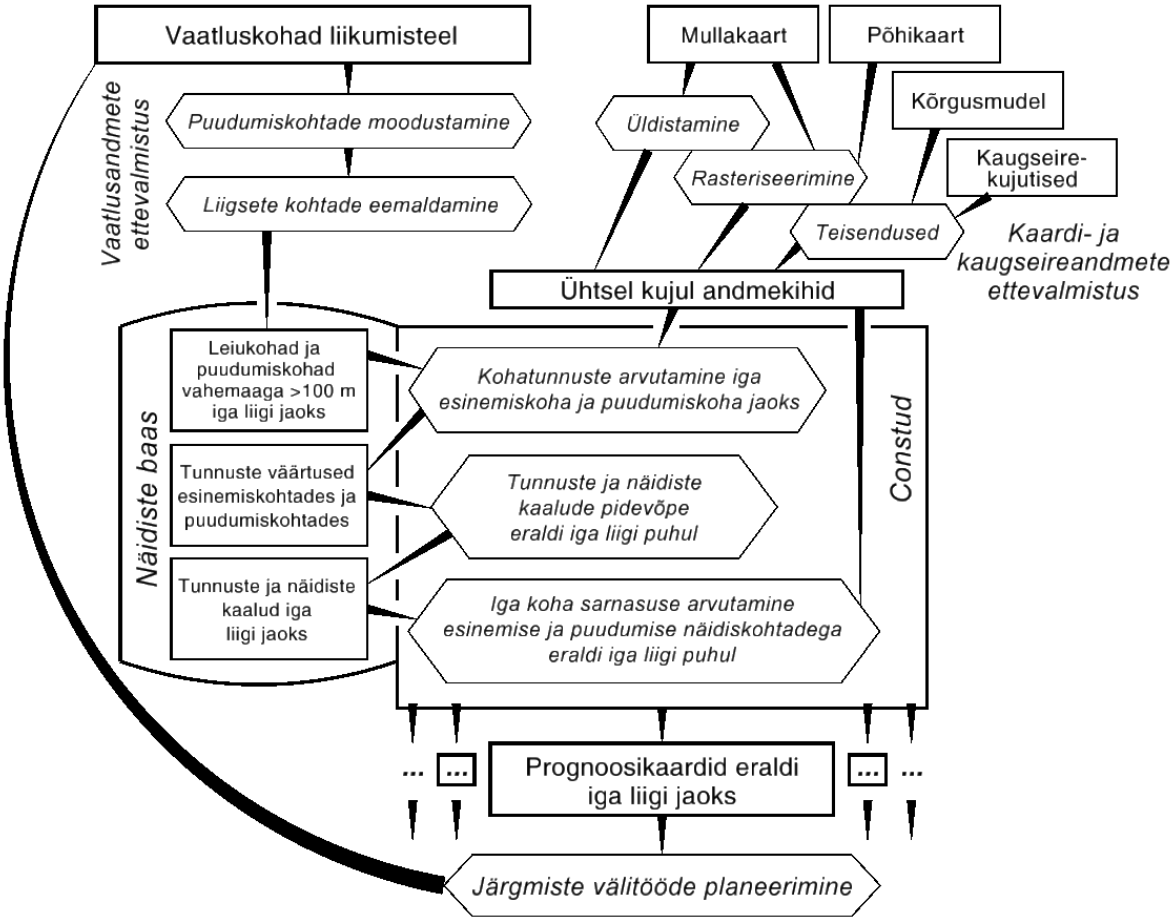
Maakatteüksuste eristamine kaugseires suurima tõepära meetodil (ptk 1.2 ja 3.6.1) kombineeritakse Mahalanobise vahemaad klasside etteantud tõenäosustega. Kui liigi levikut käsitleda kahe kategooria (esinemine ja puudumine) eristamisena vaid selliste pikslivääртuste järgi, mida saab kasutada kui normaaljaotusega muutujaid, siis on suurima tõepära meetod liikide leviku kaardistamisel hästi rakendatav ja on osutunud mõnes uuringus suhteliselt tõhusaks (Pasher *et al*. 2006).

5.6.8.3. Constud

Tartu Ülikoolis loodud tarkvarasüsteem Constud (varasem eestikeelne versioon nimega Pidevstuudium) (Remm ja Linder 2007; Remm ja Remm 2008, Remm ja Kelviste 2011a, b, c) võimaldab arvutada kohta ja selle ümbrust kirjeldavaid kaardi- ja pildimustri indekseid, otsida optimaalseid kaale tunnustele ja vaatlustele, arvutada sarnasusele tuginevaid hinnanguid ja sarnasusi andmebaasi tabelitesse või rasterkaardile (vaata ka ptk 3.4.6.1). Parimaid prognoose andvat tunnuste ja näidiste kaalude komplekti valitakse Constud süsteemis iteratiivse sobitamise ehk tehisõppe abil. Süsteem võimaldab genereerida hinnangulisi kaarte ja sarnasuskaarte ning uusi vaatlusandmeid ja tunnuseid töö käigus juurde lisada. Kaheväärtuselise tunnuse puhul eeldab süsteem nii esinemise kui ka puudumise vaatluste olemasolu õpetusandmestikus. Kui kasutaja soovib, et esinemise/puudumise hinnangu künnis oleks muul tasemel kui 50%, siis saab esinemise/puudumise hinnangu tuletada esinemiskohtadega sarnasusest, mille Constud väljastab vahemikus 1…100%.

Constud genereerib lisaks prognoositava muutuja kaardile ka kasutatud näidistega sarnasuse kaardi või etteantud klassiga sarnasuse kaardi. Ühel juhul kujutatakse igas prognoositud kohas selle koha sarnasust hinnangu aluseks olnud näidistega. Kui sarnasustase seniste kõige sarnasemate näidistega on madal, siis tuleks näidistebaasi täiendada. Teisel juhul on kaardil iga koha sarnasus etteantud klassiga, näiteks kuivõrd sarnaneb iga piksel andmebaasis olevate uuritava liigi esinemiskohtade näidistest kõige sarnasematega.

Sarnasusele tuginev hinnanguline kaardistamine peaks sobima paljude ülesannete lahendamiseks ja peaks olema pidev protsess, kuna maastikud ja liikide levik muutub, kaardi- ja kaugseireandmed täienevad ning vaatlusandmed on aina ebapiisavad. Sarnasusele tugineva meetodi eelis on paindlikkus andmete lisandumisel – mudeli sobitamist andmetele ei pea nullist alustama, sest tunnuste ja vaatluste olemasolevad kaalud säiluvad andmebaasis. Iga järgmise välivaatluste perioodi vaatluskohad saab planeerida senise hinnangulise kaardi alusel, arvestades senise hinnangu usaldusväärsust. Näiteks eelkõige kohtadesse, mis sarnanevad seniste esinemiskohtadega, kuid kus liiki pole leitud ega pole ka tõsiselt otsitud (joonis 5-24).



Joonis 5-24. Saranasusele tugineva kaardistuse kui pideva protsessi tehnoloogiline skeem (Remm ja Remm 2009 muudetud).

Uurimused

USA ökoregioonide kaart võimaldab esitada iga koha sarnasust mingi etteantud kohaga või tunnuste kombinatsiooniga (Hargrove ja Hoffmann 2005).

D^2 meetodit elupaigasobivuse kaardistamisel on kasutanud (Knick ja Dyer 1997, Corsi *et al*. 1999, Browning *et al*. 2005, Thompson *et al*. 2006, Tsoar *et al*. 2007, Griffin *et al*. 2010).

Tarkvarasüsteemi Constud on kasutatud käpaliste leviku hinnanguliseks kaardistamiseks (Remm *et al*. 2009, Remm ja Remm 2009).

5.6.9. Tugivektormasinad ja tehisnärvivõrgud

Intellektitehnika meetodite hulka loetavaid tehisnärvivõrke (*artificial neural networks – ANN*) on liikide leviku ja elupaigasobivuse modelleerimisel kasutatud sagedamini kui tugivektormasinaid (*support vector machines – SVM*). Teoreetiliselt on tugivektormasinad siiski sobivam vahend modelleerimaks ökonišši kui suvalise kujuga paljumõõtmelist osa liigi nõudluste hüperruumis (Drake *et al*. 2006).

Uurimused

Tehisnärvivõrgud (ANN) on sageli üks meetod meetodeid võrdlevates uuringutes (Pearson *et al*. 2002, Thuiller 2003, Segurado ja Araújo 2004). Näiteks Shan *et al*. (2006) võrdlesid järgmisi intellektitehnika meetodeid: otsuste puu, tehisnärvivõrku, tugivektormasinat ja geneetilisi algoritme Austraalia kukrulise *Isoodon obesuslus* leviku modelleerimisel. Tugivektormasinad ja tehisnärvivõrk andsid veidi parema tulemuse, kuid otsuste puu ja geneetilise algoritmi tulemuse eelis on parem interpreteeritavus.

Guo *et al*. (2005) modelleerisid SVM abil tammede surma põhjustava kahjuri *Phytophthora ramorum* levikut Californias.

ANN ja SVM andsid kõige täpsemaid taimeliikide leviku hinnanguid Põhja-Hispaanias (Bedia *et al*. 2011).

SVM oli täpsem kui GARP troopilise puuliigi *Miconia calvescens* tuvastamisel kohatunnuste ja kaugseireandmete järgi (Pouteau *et al*. 2011).

Zuo *et al*. (2008) on loonud tarkvarapaketi SVM meetodi kasutamiseks liikide potentsiaalse leviku modelleerimiseks.

5.6.10. Ansamblimeetodid ja konsensusmeetodid

Sama ülesannet paralleelselt lahendavate mudelite kogu nimetatakse ansambliks (*ensemble*) või komiteeks (*committee*). Samadest lähteandmetest mudelite ansambli abil paralleelseid lahendeid genereerivaid meetodeid nimetatakse **ansamblimeetoditeks** (*ensemble methods*). Paralleelset arvutust kasutatakse ka geneetilistes algoritmides (ptk 3.4.5.3), ansamblimeetodid eristuvad viimastest mitme lahendi poolest. Ansamblimeetodite idee ulatub 19ndasse sajandisse. Kaasaegses tähenduses sai prognooside kombineerimise idee tuntuks Bates ja Granger (1969) artikliga, kuigi ka need autorid viitavad varasemale sarnasele publikatsioonile (Stone *et al*. 1942). Ansamblimeetodite kasutust liikide leviku hinnangulisel kaardistamisel on seletanud Araújo ja New (2007). Ansamblimeetodi tulemusel saadakse mitu lahendit, mis on tarvis üheks koondada. Erineval viisil saadud tulemuste ühendamiseks on konsensusmeetodid.

Ansamblimeetod loob paralleelsed lahendid sama mudelitüübi abil, konsensusmeetod otsib üksmeelt erinevate meetodite tulemustes

5.6.10.1. Klassifikaatori võimendamine

Nõrga **klassifikaatori võimendamine** (*boosting*) kasutab paljude lihtsate otsusepuude koostamist tunnuste juhuslikest valimitest. Iga järgneva puu puhul arvestatakse eelmise puu rakendamisel saadud tulemusi. Valestiklassifitseerunud (raskestiklassifitseeruvate) vaatluste ja õigemaid otsuseid andvate (seletatava tunnusega tugevamalt korreleeritud) tunnuste kaalu suurendatakse nii, et neil on suurem tõenäosus sattuda järgmise otsusepuu koostamisse. Tänu vaatluste kaalu muutmisele ei kaldu võimendusmeetodid mudelit ülesobitama (Wu *et al.* 2008).

Võimendusmeetodid suurendavad igas järgnevas korduses eelmise korduse tulemustes valesti klassifitseerunud vaatluste kaalu

Kõige tuntum ansamblimeetodite võimendamismeetodite algoritm on AdaBoost (*Adaptive Boosting*) (Freund ja Schapire 1997). Selle autorid Schapire ja Freund said 2003. aastal Nobeli

preemia teoreetilise arvutiteaduse alal. AdaBoost algoritmi võib leida muuhulgas raamatust Bishop (2006, lk. 658–659). Võimendamine on rakendust leidnud paljudes tarkvaralahendustes ja sellele on antud palju erinevaid nimesid. Algsetes publikatsioonides on nimed gradiendi võimendusmasin (*gradient boosting machine – GBM*) ja gradiendi juhuslik võimendamine (*stochastic gradient boosting – SGB*) (Friedman 2001, 2002). Klassifitseerimisülesande puhul on kasutatud nimesid võimendatud klassifikatsioonipuud *boosted classification trees – BCT*, *gradient tree boosting* ja *TreeBoost*; pideva muutuja puhul kasutatakse nime võimendatud regressioonipuud (*boosted regression trees – BRT*) ja gradiendi võimendamine (*stochastic gradient boosting – SGB*). Kasutusel on ka nimed: *generalized boosting model, functional gradient boosting, gradient boosted models, gradient boosted decision trees, gradient boosted regression trees, multiple additive regression trees, TreeNet.*

Praktikas on võimendamine protsess, kus kasutaja poolt määratud parameetrid omavad olulist tähtsust – ühelt poolt saab määrata puu maksimaalse keerukuse (*tree complexity*) ja teisalt õppimiskiiruse (*learning rate*) ehk kaalu, mille saab iga järgnev puu juba olemasolevate puude ansambliga liitmisel. Puude suur keerukus nõuab üldjuhul väiksemat õppimiskiirust, sest vastasel juhul võib juba juba külllalt väikese puude arvu korral leida aset ülesobitumine, kuna puud koostatakse aga juhuslike valimite põhjal, siis ei ole liiga väikesed ansamblid soovitatavad. Väga aeglane õppimiskiirus välistab küll ülesobitumise, ent optimaalsete tulemuste jaoks on vaja väga mahukat ansamblit.

Klassifikaatori võimendamist kasutavad Interneti otsingumootorid Yahoo ja Yandex. Ülevaate võimendamismeetoditest ja nende arendamisloost on kirjutanud Ridgeway (1999).

5.6.10.2. Juhumets

Ansamblimeetodit, mille puhul genereeritakse esmalt palju otsuste puid, kasutades iga puu loomisel tagasipanekuga valimit tunnustest, seejärel kombineeritakse lõpptulemuseks otsuste puudest parimad, nimetatakse **juhumetsaks** (*random forest*). Iga puu järgi arvutatakse hinnang ning lõpphinnang tuletatakse üksikhinnangutest hääletamise teel. Otsuste puudest võib moodustada ka mitu **salu** (*grove*) ning võrrelda ja kombineerida erinevates saludes saadud tulemusi. Parima mudeli valimisel paljudest sama tüüpi mudelivariantidest või valimitest (hääletamine klassifitseerimisülesande puhul, keskmistamine regressiooniülesande puhul) on inglise keeles nimi *bagging* (*bootstrap aggregating*), mida eesti keeles võiks mitme mudeli tulemuste ühendamise puhul nimetada **tulemuste koondamiseks** ja valimitest saadud tulemuste puhul **valimite koondamiseks**.

Juhumetsa iga üksiku puu aluseks on valim tunnustest, valimite koondamise puhul valim vaatlustest

Juhumetsa paljude eraldi koostatud puude abil saadud suhteliselt sõltumatute tulemuste varieeruvuse järgi saab määrata hinnangu ebakindlust ja iga ruumipiksli või objekti igasse etteantud klassi kuulumise tõenäosust.

5.6.10.3. Konsensusmeetodid

Kui ansamblimeetod annab paralleelsed lahendid sama mudelitüübi abil, siis **konsensusmeetodite** puhul otsitakse üksmeelt erinevate uurimismeetodite tulemustes. Modelleerimise puhul koostatakse paralleelsed mudelid erinevate meetodite abil. Kõige tõepärasemaks loetakse selline tulemus, mida toetab enamik meetodeid. Konsensusmeetodeid kasutatakse eelkõige probleemide puhul, kus puudub selge lahend ja üksmeel. Kolm tuntumat konsensusmeetodit on Delfi protsess (Delphi oraakli järgi),

ekspertpaneel ja konsensuse kujundamise konverents, aga konsensuse leidmise vahendid on ka keskmise, üksikmudeli täpsusega kaalutud keskmise, moodi ja mediaani arvutamine.

Delfi protsessi etapid on järgmised.

- Kutsutakse kokku eksperdid.
- Ekspertide arvamused grupeeritakse piiratud arvuks teemadeks ja väideteks.
- Osalejad hindavad oma nõusolekut iga väitega ja järjestavad väited selle alusel.
- Väidete astakud summeeritakse.
- Osalejad hindavad oma nõusolekut iga väitega uuesti arvestades nüüd ka astakute summasid.
- Väidete astakud summeeritakse ja hinnatakse, kas piisav konsensus on saavutatud.

Kui üksmeel ei ole piisav, siis korratakse nõustumistasemete hindamist.

Konsensusmeetod on vahend eriarvamuste summutamiseks

Ekspertpaneeli puhul panevad eksperdid lühidalt kirja oma seisukohad. Iga ekspert esitab koosoleku juhatajale ühe mõtte. Sarnased arvamused grupeeritakse kokku. Iga osaleja annab omaette igale arvamusele punkte. Hindamise tulemus summeeritakse. Koondhinnangut diskuteeritakse ning seejärel hinnatakse uuesti kuni piisava konsensuse saavutamiseni.

Konsensuse kujundamise **konverents** on laia osalejaskonnaga ja erinevate reeglitega üritus enamuse osalejate üksmeelele leidmiseks antud ajahetkel.

Konsensusmeetodeid on kõige enam uuritud seoses meditsiiniliste otsustega (Fink *et al*. 1984, Campbell *et al*. 2001). Liikide leviku modelleerimisel on konsensusmeetod sobiv vahend erinevatest mudelitest ja kliima muutumise erinevate stsenaariumite korral saadud hinnanguliste levikukaartide ühendamiseks ja kaartidevahelise üksmeele mõõtmiseks.

5.6.10.4. BIOMOD

BIOMOD (*BIOdiversity MODelling*) on tarkvararaamistik, mis ühendab R tarkvarakeskkonnas liikide leviku modelleerimise meetodeid (Thuiller 2003, Thuiller *et al*. 2009). BIOMOD tarkvaraarnduskeskkonna R veebileht on aadressil http://r-forge.r-project.org/projects/biomod.

5.6.10.5. Lifemapper

Lifemapper (http://www.lifemapper.org) on ülemaailmne liikide levilate hinnanguliste kaartide elektrooniline atlas (Stockwell *et al.* 2006). Levikukaartide genereerimiseks kasutatakse GARP geneetilist algoritmi (ptk 5.6.6.2). Kuna geneetilised algoritmid suurte andmehulkadega on arvutusmahukad, siis kasutatakse hajusarvutust – muust tööst vabal ajal osalevad paljude vabatahtlike arvutid üle maailma. Iga arvuti arvutab mingi liigi kaarti mingite lähteparameetritega. Sama liigi levikukaartide erinevad versioonid koondatakse keskusesse, neid võrreldakse ja valitakse parameetrid järgmiste arvutusülesannete jaoks. Geneetiline algoritm on keerukas, tavaliselt kulub rahuldava lahenduseni jõudmiseks 500 kuni 1000 kordust. Korraga on kasutusel neli reeglite taset. Atomaarne tase võimaldab prognoosi siduda keskkonnafaktori üksikväärtusega, BIOCLIM tase moodustab tolerantsipiire kõigi faktorite järgi, piirkonna tase moodustab tolerantsipiire osade faktorite järgi, logitmudeli tasemel sobitatakse andmetele logistilisi esinemise ja puudumise mudeleid.

Leiuandmed pärinevad muuseumikollektsioonidest ja tasuta saada olevatest keskkonna-andmestikest. Andmedoonorina on Eestist projektiga liitunud projekti veebilehel oleva teabe kohaselt

vaid TTÜ Geoloogia Instituut. Kliima 0,5° lahutusega andmed rahvusvahelisest kliimamuutuste paneelist (*Intergovernmental Panel on Climate Change* http://www.ipcc.ch andmed aadressil http://www.ipcc-data.org/obs). Kasutatakse aastate 1961–1990 keskmisi ja aasta esimese poole sama perioodi kuude keskmisi tunnuseid: pilvisus, päevane temperatuuri amplituud, miinustemperatuuride sagedus maapinnal, maksimaalne, keskmine ja minimaalne temperatuur, sademete hulk, päikesekiirgus, õhuniiskus, vesiste päevade sagedus ning tuuled. Reljeefi andmetest kasutatakse nõlva suunda, nõlva kallet, voolu akumulatsiooni ja maapinna niiskuse indeksit. Kõik kohatunnuste andmekihid on süsteemi Lifemapper jaoks teisendatud 1 km küljega piksliteks.

5.6.10.6. OpenModeller

Tarkvarasüsteem openModeller on avatud koodiga vabavara, mis pakub ühtset sisend- ja väljundsüsteemi kasutamaks erinevaid liikide leviku modelleerimise meetodeid ja andmekihte. Liikide leviku modelleerimise meetodeid on mitmeid ja iga üksik tarkvara eeldab andmete ettevalmistamist kindlal kujul. Mitme erineva algoritmi rakendamise võimalus on kommertstarkvara pakettides, vaba lähtekoodiga sama põhjalikult väljaarendatud liikide leviku modelleerimise lahendusi teadaolevalt ei ole. Süsteemi väljaarendajad rõhutavad, et openModeller on loodud liikide potentsiaalse, mitte tegeliku levila kaardistamiseks (Muñoz *et al.* 2011).

Lähteandmete ühtse vormingu tagavad interneti kaudu saada olevad leidude registrid (Guralnick 2007) ning klimaatiliste ja topograafiliste andmete kihid. Meetodeid ja andmeid integreeriva süsteemina suudab openModeller lugeda ja teisendada erinevates koordinaatsüsteemides olevaid andmeid, lugeda kohatunnuseid erineva lahtrisuurusega rasterkihtidest, lugeda ja kirjutada erinevaid rasterformaate. Samuti suudab kasutada erinevaid modelleerimise algoritme ja võimaldab paralleelset kasutust erinevates op-süsteemides. Mudeli loomisel saab openModeller versioon 1.2 puhul kasutada järgmisi algoritme: ANN, AquaMaps, Bioclim, *Climate Space Model*, ökoniši faktoranalüüs, *Envelope Score*, vahemaa tunnusruumis, juhumets, tugivektormasinad. Enamikku neist algoritmidest on kirjeldatud ka selles raamatus. OpenModeller veebilehel http://openmodeller.sourceforge.net on vabalt saadaval desktop lahendus, algoritmide kirjeldus, tarkavarakood, süsteemi kirjeldus ja muu asjakohane teave.

5.6.10.7. NeuralEnsembles

NeuralEnsembles (http://purl.oclc.org/NeuralEnsembles) on vabavaraline Windows tarkvara, mis arvutab hinnanguid tehisnärvivõrkude ansambli abil (O'Hanley 2009).

5.6.10.8. BioEnsembles

Brasiilias arendatud liikide leviku modelleerimise meetodite integreeritud keskkond BioEnsembles sisaldab meetodeid: BIOCLIM, Eukleidiline kaugus, Mahalanobise vahemaa, GARP, GLM, RF, Maxent, (Rangel *et al*. 2009, Diniz-Filho *et al.* 2009, 2010).

Uurimused

Juhumets võib anda täpsemad liikide leviku mudelid kui lihtsad otsuste puud (CART) ja tehisnärvivõrgud (ANN) (Garzón *et al*. 2006).

BRT meetod on võimaldanud puuliikide levikut täpsemalt hinnata kui lihtsad klassifikatsioonipuud või GAM (Moisen *et al*. 2002) ning olnud üks paremaid meetodeid liikide leviku modelleerimise meetodite võrdlustes (Elith *et al*. 2006, Guisan *et al*. 2007, Elith ja Graham 2009, Zurell *et al*. 2009).

BRT oli ka kõige usaldusväärsemate meetodite hulgas meetodite võrdluses Baltimaade sademete pikaajalise keskmise modelleerimiseks ja interpoleerivaks kaardistamiseks (Remm *et al*. 2011). RF ja BRT andsid kõige täpsemaid tulemusi tavaliste puuliikide leviku modelleerimisel USAs (Prasad *et al*. 2006).

Thuiller (2004) määras Euroopa taimeliikide 40 erineval viisil aastaks 2050 prognoositud leviku konsensuse peakomponentanalüüsiga ja leidis, et konsensus 40 erineva üksiktulemuse vahel oli 56,1%. Üksmeeletulemuseks valiti üksiktulemus, mis korreleerus kõige tugevamini peakomponentanalüüsi konsensuse teljega.

Marmion *et al*. (2009) võrdlesid kaheksa erineva modelleerimismeetodi abil saadud leviku-kaartidest konsensuskaardi saamise mooduseid. Esiteks, konsensuskaart vastas kontrollandmetele alati paremini kui ükski üksikmeetodiga saadud kaart. Teiseks, üksmeelele jõudmise viisidest oli efektiivseim mudeli täpsust näitava toimimiskõvera aluse pinnaga kaalutud keskmistamine.

Jeong *et al*. (2011) valisid sobivaimat *Sterna albifrons* pesapaiagaeelistuse mudelit 2000 kandidaatmudeli hulgast algul ruutkeskmise vea järgi ja seejärel ekspertotsustega.

5.6.12. Leviku modelleerimine puudumisandmeteta

Enamik liikide leiuandmeid sisaldavatest andmestikest kajastavad ebaühtlase tiheduse ja intensiivsusega tehtud vaatluste tulemusi. Enamasti puudub kindel teave ühe või teise liigi puudumiskohtadest, mis raskendab liikide esinemistõenäosuse hindamist (Hirzel *et al*. 2002, Zaniewski *et al*. 2002, Anderson ja Mártinez-Meyer 2004, Ottaviani *et al*. 2004, Pearce ja Boyce 2006). Liigi mitteleidmine kohas võib olla küll tingitud sellest, et koht on liigile sobimatu, kuid liik võib jääda tuvastamata ka kohtades, kus ta tegelikult esineb või kohtades, mis on küll sobivad, kuid liik puudub sealt ajutiselt. Liigi tuvastamine on üldiselt märksa kindlam kinnitus koha sobivusest, kuigi erandlikult võib liike kohata ka kohtades, mis on neile pikaajaliselt sobimatud.

Vaadeldud puudumiskohtade puudumisel võib liigi puudumine olla esindatud kaudselt: juhuslike kohtadega uurimisalal (Stockwell ja Noble 1992, Osborne *et al*. 2001, Ferrier *et al*. 2002a, Stockwell ja Peterson 2002, Anderson *et al*. 2003, Olivier ja Wotherspoon 2006, Irfan-Ullah *et al*. 2007), juhuslike kohtadena vältides vaadeldava koha lähedust (Remm ja Remm 2009, Mateo *et al*. 2010a,b), juhuslike punktidena elupaiga osas, kust liiki ei ole kunagi leitud (Syartinilia ja Tsuyuki 2008), ökoloogiliste faktorite keskmiste ja amplituudide kaudu (Hirzel *et al*. 2002), teiste liikide leiukohtadena, kus uuritavat liiki ei ole registreeritud (Zaniewski *et al*. 2002, Dudík *et al*. 2006, Elith ja Leathwick 2007), kohtadena, kus on registreeritud piisav arv teisi liike (Parsons *et al*. 2009), eelmise mudeli abil prognoositud puudumisalale (Ward *et al*. 2009), juhupunktidena piki vaatlusteed (Gibson *et al*. 2004), leiukohtadest piisavalt kaugel kogu alal (Gibson *et al*. 2007) või vaatlusteel (Remm *et al*. 2009, Remm ja Remm 2009). Vaatlustrassi arvestamist lihtsustab tänapäevane GPS tehnoloogia, mis võimaldab salvestada suvalist kõverjoonelist välivaatluste liikumisteed.

Elith ja Leathwick (2007) nimetavad juhuslikena genereeritud puudumiskohti **juhuslikeks näiv-puudumisteks** (*random pseudo-absences*) vastandades neid kohtadele, kus vaatleja on kohal olnud, kuid pole liiki registreeritud – **näiv-puudumised vaatlustes** (*inventory pseudo-absences*). Mateo *et al*. (2010a) eristavad veel **sihtgrupiga seotud puudumiskohti** (*target-group absences*), millesse arvatakse kohad, kus lähedasi liike on registreeritud, kuid uuritavat liiki mitte.

Puudumiskohad on näivad, sest liigi mitteregistreeritud esinemine ei ole esinemiskohaga samaväärne – liiki võib olla ei märgatud või liiki parajasti ei olnud kohal

Vaadeldud puudumiskohtade kasutamine annab täpsemad või vähemalt paremini interpreteeritavad levikumudelid võrreldes näiv-puudumise juhukohtadega (Václavík ja Meentemeyer 2009). Juhuslikud puudumiskohad võivad sisaldada registreerimata esinemiskohti ning ebaühtlane vaatlusintensiivsus mõjutab samal määral nii leiukohtade kui ka vaadeldud puudumiskohtade esindatust. Paraku enamasti ei kaasne kasutada olevate leiuandmetega puudumisandmed, küll aga on võimalik piiritleda uurimisala ehk **taustaandmed** (*background data*), mille suhtes esinemist modelleeritakse. Kogu uurimisalale genereeritud juhuslikult paiknevad punktid on hälbeta valim uurimisala suhtes ja ka haruldaste liikide puudumiskohtade üsna headeks esindajateks. Laialt levinud liigi puhul annab aga juhuslikult paiknevate puudumiskohtadega **naiivne mudel** hälbega hinnangu.

Puudumiskohti esindavaid juhupunkte genereeritakse enamasti sama palju, kui on esinemiskohti. Kitsa nõudlusega liikide puhul on kohatunnuste varieeruvus liigi esinemisalal aga palju väiksem kui puudumisalal. Seetõttu võib puudumisala suurem esindatus mõnikord põhjendatud olla. On ka soovitatud, et juhuslikke kohti tuleks kasutada vaid juhul, kui sihtgrupiga seotud puudumiskohti ei ole piisavalt. Juhuslike puudumiskohtade genereerimisel tuleks leiukohtade ümber moodustada puhvrid, kuhu juhuslikud puudumiskohad ei lange (Mateo *et al*. 2010b).

Kahjuks ei ole isegi registreeritud puudumiskohad täiesti usaldusväärsed. Uuritav objekt võis jääda vaatlejale märkamatuks, liikuv objekt ei pruukinud vaatluse ajal parajasti vaatluskohas viibida, samuti võis objekt välja jääda liiga väikese vaatlusala tõttu. Seepärast võib arvata, et ka registreeritud puudumisandmete hulgas on väärnegatiivseid. Koha määramine ekslikult liigi puudumiskohaks on tõenäolisem haruldaste ja varjatud eluviisiga liikide puhul. On arvutatud, et koha klassifitseerimiseks puudumiskohaks 95% kindlusega on tarvis seda eri aegadel mitukümmend korda külastada ja liiki otsida (Kéry 2002). Mõned liikide levikut modelleerivad autorid (näiteks Jiménez-Valverde *et al*. 2008 ja Gogol-Prokurat 2011) peavad väärnegatiivsete puudumisandmete riski tõttu õigemaks puudumiskohtade teabest üldse loobuda.

Sisutihedam lahendus puudumiskohtade liiasuse vähendamiseks on neile väiksema kaalu omistamine võrreldes esinemiskohtadega. Puudumiskohtade kaalud, mis vastavad esinemiskohtade ja puudumiskohtade suhtele, tagavad esinemiskohtade ja puudumiskohtade võrdse esindatuse ja samas säilitavad puudumiskohtade varieeruvuse ulatuslikuma esindatuse.

Vaatlusintensiivsuse ebaühtlusest mõjutatud leiuandmed kombinatsioonis ühtlaselt või juhuslikult paiknevate puudumiskohtadega annavad nihkega hinnanguid

Asukoha valiku mudelites (*discrete choice models – DCM*) genereeritakse puudumiskohad juhupunktidena leiukohta ümbritsevale valikualale (*choice set*), mille suuruse määrab uurija vastavalt uuritava liigi liikumis- ja levimisvõimele (Cooper ja Millspaugh 1999). Eeldatakse, et vabalt liikuv organism eelistab olla piirkonna kõige sobivamas kohas.

Liigi puudumine mingis kohas vaatlushetkel ei tähenda selle koha sobimatust selle liigi jaoks

Samaväärselt puudumiskohtade mitteteadmisega või vähese usaldusväärsusega on regressiooni-mudelite kasutamisel probleemiks puudumiskohtade üleesindatus ehk **nullide liiasus** (*zero inflated data*). Nullide liiasuseks nimetatakse juhtu, kus loendites on rohkem nulle kui peaks olema Poissoni jaotuse järgi. Selle tulemusena ei ole selle jaotuse eeldusel leitud standardvea hinnangud tõesed.

Nullide liiasuse vältimiseks soovitasid Mullahy (1986), Heilbron (1994), Welsh *et al*. (1996) kõigepealt modelleerida esinemist/puudumist logit-lingiga ja seejärel modelleerida esinemisohtrus Poissoni või negatiivse binoomjaotusega. Barry ja Welsh (2002) soovitasid samuti kaheastmelist modelleerimist, aga üldistatud aditiivsete mudelitega, mis ei sea tunnuste jaotusele eeltingimusi.

Uurimused

Nullvaatlusi on ökoloogias eraldi modelleerinud näiteks Remm (1989), Welsh *et al*. (1996), Barry ja Welsh (2002), Pearce ja Ferrier (2001).

Mitmete levikumudelite koostamisel on liigi puudumiskohad olemasoleva teadmise alusel suunatud kindlama puudumise alale. Näiteks Engler *et al*. (2004) genereerisid juhuslikud puudumiskohad vaid alale, kus varasema mudeli järgi on uuritava liigi esinemistõenäosus < 0,3. Le Maitre *et al*. (2008) genereerisid kaks komplekti puudumiskohti. Üks sisaldas juhupunkte kogu uurimisalal, teine juhupunkte väljaspool liigile klimaatiliselt sobivat ala. Teise puudumiskohtade komplektiga loodud mudelid vastasid kontrollandmetele paremini. Huang *et al*. (2011) võrdlesid puudumiskohtade juhuslikul ja suunatud juhuslikul genereerimisel saadud levikumudeleid ja leidsid, et kõik ebasobivale alale suunatud juhupunktide kasutamisel andsid erinevat tüüpi mudelid täpsemaid tulemusi kui lihtsalt juhupunktidele tuginevad mudelid. Mateo *et al*. (2010a,b) kasutasid vaid puudumiskohti, mis olid esinemiskohtadest vähemalt 30 km kaugusel. Puudumiskohti on genereeritud ka vastavalt koha sarnasusele seniste leiukohtadega (Zaniewski *et al*. 2002, Chefaoui ja Lobo 2008). Kui puudumiskohad genereeriti eelistatult leiukohtadest erinevatesse kohtadesse, siis saadi leviku laiem hinnang – kui eelistatult sarnastesse kohtadesse, siis kitsam levila.

Lütolf *et al*. (2006) modelleerisid liblikate levikut Šveitsis ja võrdlesid puudumiskohtade moodustamise viise. Kaasagset levikut esindas kõige paremini mudel, mis oli sobitatud puudumiskohtadega alal, kus uuritavat liiki ei olnud viimasel 10 aastal leitud. Koha uuritust ja varasemaid leide ei arvestatud. Autorid tõdesid, et ajalooliste leiukohtade vari (*the ghost of past occurrence*) segab kaasaegse leviku modelleerimist.

Gibson *et al.* (2007) genereerisid 5000 juhuslikku puudumiskohta kogu uurimisalale tingimusega, et ükski puudumiskoht ei tohi olla uuritava maapapagoi teadaolevale esinemiskohale lähemal kui 400 meetrit. Kuna genereeritud puudumiskohad ei olnud kindlad puudumiskohad, siis omistati neile esinemis- ja puudumiskohtade mahtu võrdsustav kaal *n*/5000, kus *n* on esinemiskohtade arv. Nii tagati esinemis- ja puudumiskohtade võrdne esindatus ning välditi nende kattuvust.

Maggini *et al*. (2006) leidsid, et puudumiskohtade suuremat hulka kompenseerivad kaalud parandavad metsatüüpide leviku modeleid.

Tsoar *et al*. (2007) võrdlesid levikumudeleid, mis kasutavad vaid liikide esinemiskohti. Parimaks osutusid GARP ja Mahalanobise kaugus (*Mahalanobis Distance* – MD).

5.6.13. Kohatunnused liikide leviku mudelites

Liikide leviku mudelites kasutatud keskkonnaparameetrid on üsna erinevad (tabel 9). Kohta iseloomustavate tunnuste valikul ei ole ühtseid põhimõtteid, mis raskendab mudelite võrdlemist. Üldprintsiipe, et enamiku taimede elupaigasobivus sõltub valguse hulgast, temperatuurist, toitainetest, veest, süsihappegaasist, häiringutest ja biootiliste tegurite komplektist, on levikumudelites keerukas rakendada, sest organismidel otseselt mõjuvad faktorid ei ole ülepinnaliselt kaardistatud. Liikide potentsiaalse leviku globaalset või maailmajao taset hõlmavates uuringutes kasutatakse eelkõige klimaatilisi tunnuseid. Detailsemate kaardistuste puhul on olulised mulla ja maakatte tunnused, mille

kohta saadakse teavet olemasolevatelt kaartidelt ja kaugseirest. Suure ala kaardistamisel saab enamasti kasutada vaid üldistatud mõõtkavas andmekihte, mis on ruumiliselt palju vähem detailsed kui faktorite mõjud. Ülevaate liikide leviku mudelites kasutatud tunnustest annavad Franklin (2009) ning Austin ja Van Niel (2011).

Levikumudeli kalibreerimiseks ja levikukaardi loomiseks mudeli abil saab kasutada kaardistatud tunnuseid. Paraku on vaid vähestel juhtudel andmed liikide levikut määravate faktorite kohta ülepinnaliselt olemas. Seega tuleb liikide leviku kaardistamisel suures osas tugineda kaudsetele, otseste mõjufaktoritega korreleeruvatele tunnustele. Näiteks võivad maapinna kõrgusmudelist tuletatud pinnavormid ja asend reljeefil asendada mullakaarti (Bridge ja Johanson 2000).

Enamasti on liikide leviku mudelitesse kaasatud vaid abiootilised tunnused, kuigi elusorganismid esinevad üksteisest sõltuvate organismide kooslustena. Mitmed autorid on märkinud, et enamik ökoloogilisi gradiente on ühes otsas füüsiliste tingimuste poolest karmid (limiteerivad) ja teises otsas biootiliste tingimuste (konkurentsi) poolest karmid (MacArthur 1972, Dobzhansky 1950, Brown *et al.* 1996, Guisan *et al.* 1998). Liikide loodusliku levila üldisel paiknemisel on biootiliste faktorite mõju suhteliselt väike (Huntley *et al*. 1995, Pearson ja Dawson 2003, Araújo ja Luoto 2007). Esinemise ja puudumise kaardistamisel detailsemas mõõtkavas on teiste liikide esinemise ja hulga tunnustel oluline indikaatorväärtus (Heikkinen *et al*. 2007, Remm ja Remm 2009, Meier *et al*. 2010). Liikidevaheline seos võib olla otsene konkurentsi- või sümbioosisuhe, ühise elukeskkonna mõjutamine või siis elupaiga indikaatorina sobivale või sobimatule kohale osutamine. Teiste liikide teadaolevad leiukohad võivad olla head indikaatorid näitamaks ühte või teist tüüpi elupaikade paiknemist. Näiteks Remm ja Remm (2009) uuringus osutusid kõige paremateks käpaliste esinemise/puudumise indikaatoriteks kindla kasvukohaeelistusega ja keskmise esinemisagedusega teised käpalisteliigid.

Liikide praegused levilad ei ole määratud vaid kaasaegsete keskkonnatingimustega, neid on mõjutanud ja mõjutavad teised liigid ning maastiku arengulugu. Seetõttu ei ole kaasaegsed leiuandmed eriti usaldusväärne pidepunkt kaardistamaks levilate muutumist kliima muutumisel (Davis *et al.* 1998).

Kohta iseloomustavate tunnuste arv mingi liigi leviku modelleerimisel võib olla üsna suur, näiteks Remm ja Remm (2009) valmistasid ette 161 kohatunnust, millest hinnangute arvutamisel kasutati korraga vaid kümmekonda. Kohatunnuste hulgas olid ka teiste liikide leidude suhtelised tihedused ümbruskonnas.

Kui oletada, et kliimat saab kirjeldada 24 parameetriga, siis nende parameetrite võimalike kombinatsioonide arv ulatub 17 miljonini. Milliseid parameetrite kombinatsioone kasutada, tuleb vähemalt osaliselt mingite mudeliväliste lisateadmiste alusel otsustada. Näiteks valides vaid liigi jaoks eeldatavasti kriitiliste kuude ilmastiku andmed.

Kohatunnuste hulgas on kasutatud ka asukoha koordinaate, mis kirjeldavad üldisi klimaatilisi tingimusi. Sarnasuse järgi otsustamisel esindavad vaatluskohtade koordinaadid ka funktsioontunnuse ruumilist autokorrelatsiooni. Lähedaste kohtade sarnasus on suurem, sest koordinaadid on sarnased (Remm *et al*. 2011). Juhul, kui huvi ei paku ette teadaoleva suunaga gradiendid, tuleks koordinaadid mudelisse kaasata nii, et need moodustaksid komplekstunnuse (näiteks kaugusmaatriksina). Vastasel juhul võib nii ristkoordinaatide kui ka geograafiliste koordinaatide kasutamisel tekkida olukord, kus kaugus põhi-ilmakaarte suunas omandab erineva kaalu kui sama kaugus vahe-ilmakaarte suunas. Ümbruse tunnuseid käsitletakse ka peatükis 5.5.

Tabel 9. Oluliseks osutunud kohatunnused mõnedes liikide esinemise/puudumise ja ohtruse mudelites. Lühend e/p tähistab esinemist või puudumist funktsioontunnusena.

Kohatunnus	Funktsioontunnus	Publikatsioonid
Geograafilised koordinaadid	Punahirve e/p	Buckland ja Elston 1993, Augustin *et al.* 1996
Regioon	3 linnuliigi ohtrus	Saveraid *et al.* 2001
Kõrgus merepinnast	*Cornus florida* tüvede tihedus	Wilds *et al.* 2000
	Lacerta schreiberi esinemine /puudumine	Brito *et al.* 1999
	Linnuliikide e/p	Griffiths *et al.* 1993
	Parnassius mnemosyne levik ja ohtrus	Luoto *et al.* 2001
	Chioglossa lusitanica e/p	Teixeira *et al.* 2001
	Karibuu elupaigasobivus	Hansen *et al*. 2001
Nõlvakalle	*Petauroides volans* populatsiooni tihedus	Whigham 2000
	Chioglossa lusitanica e/p	Teixeira *et al*. 2001
	Karibuu elupaigassobivus	Hansen *et al*. 2001
	Petaurus volans e/p	Lindenmayer *et al*. 1999
	Parnassius mnemosyne levik ja ohtrus	Luoto *et al*. 2001
Pinnavormide indeksid	*Cornus florida* tüvede tihedus	Wilds *et al*. 2000
	Chrysococcyx lucidus ja *Cuculus pyrrhophanus* esinemine /puudumine	Neave *et al*. 1996
	Linnuliikide e/p	Griffiths *et al.* 1993
Pinnavormide tüübid	*Arnica montana* levik	Parolo *et al*. 2008
Liigi esinemine koha naabruses	*Spizaetus bartelsi* levik	Syartinilia ja Tsuyuki 2008
	Punahirve e/p	Augustin *et al*. 1996
	Parnassius mnemosyne levik ja arvukus	Luoto *et al*. 2001
Sarnase nõudlusega liikide leiutihedus 500 m ulatuses	12 käpaliseliigi e/p	Remm ja Remm 2009
Maakattetüüp	*Arnica montana* levik	Parolo *et al*. 2008
	Grus japonensis esinemine /puudumine	Li *et al.* 1997
	Parnassius mnemosyne levik ja arvukus	Luoto *et al*. 2001
Ümbritseva maastiku metsasus	*Seiurus aurocapillus* arvukus	Lee *et al.* 2002
	4 linnuliiki	Saveraid *et al*. 2001
	Petaurus volans ohtrus	Lindenmayer *et al*. 1999
Puistu katvus	*Clemmys insculpta* esinemine /puudumine	Compton *et al.* 2002
Puistu tüüp	*Petaurus volans* ohtrus	Whigham 2000
Puistu koosseis	Punahirve e/p	Buckland ja Elston 1993, Augustin *et al*. 1996
Puistu vanus	Karibuu elupaigasobivus	Hansen *et al*. 2001
Puistu kõrgus	*Malurus cyaneus* esinemine /puudumine	Neave *et al*. 1996
Lehtmetsa osa	*Vireo olivaceus* arvukus	Lee *et al*. 2002
Elupaiga hulk piir-konnas	Põdrapopulatsiooni tihedus	Puttock *et al*. 1996
Lehepinnaindeks (LAI)	Kahe linnuliigi ohtrus 2	Saveraid *et al*. 2001
Põõsarinde biomass	5 linnuliigi ohtrus	Saveraid *et al*. 2001
Põõsarinde katvus	Jooksiklaste liikide esinemine /puudumine	Bonn ja Schröder 2001
Rohurinde katvus	Jooksiklaste liikide esinemine /puudumine	Bonn ja Schröder 2001
Võsa osa ümbruses	*Leipoa ocellata* levik	Parsons *et al.* 2009
Maastikumeetrika indeksid	*Parnassius mnemosyne* levik ja ohtrus	Luoto *et al*. 2001
	Karibuu elupaigasobivus	Hansen *et al*. 2001
Mullaeraldiste tihedus	12 käpaliseliigi e/p	Remm ja Remm 2009

Elupaigalaigu suurus	Linnuliikide *Hylocichla mustelina, Seiurus aurocapillus, Vireo olivaceus* arvukus	Lee *et al*. 2002
	Karibuu elupaigasobivus	Hansen *et al*. 2001
	Parnassius mnemosyne levik ja ohtrus	Luoto *et al.* 2001
Hoonestusala pind	Liblikate levik Soomes	Luoto *et al.* 2006
Kaugus lähima metsani	5 linnuliigi arvukus	Saveraid *et al*. 2001
Kaugus lähima maanteeni	*Petauroides volans* populatsiooni tihedus	Whigham 2000
Kaugus lähima maanteeni	*Grus japonensis* esinemine /puudumine	Li *et al.* 1997
Kaugus lähima veekoguni	*Clemmys insculpta* esinemine /puudumine	Compton *et al*. 2002
Inimasustuse tihedus	Enamik suurimetajaid Indias	Karanth *et al.* 2009
	Lymantria dispar püsiasustus	Lippitt *et al*. 2008
Kultuuriline sallimatus	*Antilope cervicapra, Gazella gazella, Boselaphus tragocamelus* Indias	Karanth *et al.* 2009
Rahvastiku ränne	*Lymantria dispar* püsiasustus	Lippitt *et al*. 2008
Raiete ja põlengute aeg	*Cornus florida* tüvede tihedus	Wilds *et al*. 2000
Domineeriv taimekooslus	*Cornus florida* tüvede tihedus	Wilds *et al.* 2000
Katmata mulla osa	Jooksiklaste liikide esinemine /puudumine	Bonn ja Schröder 2001
Mulla tekstuur (savisus)	*Leipoa ocellata* levik	Parsons *et al.* 2009
Madalsoomulla osa 100 m raadiuses	12 käpaliseliigi e/p	Remm ja Remm 2009
Mulla tüüp	*Lacerta schreiberi* esinemine /puudumine	Brito *et al*. 1999
Mulla niiskus ja pH	Jooksiklaste liikide esinemine /puudumine	Bonn ja Schröder 2001
Märgala soolsus	*Grus japonensis* esinemine /puudumine	Li *et al.* 1997
Vee sügavus	*Coregonus zenithicus* elupaik	Naumann ja Crawford 2009
	Laminaria hyperborea katvus	Bekkby *et al*. 2009
Eksponeeritus lainetele	*Laminaria hyperborea* katvus	Bekkby *et al.* 2009
Keskmine päi-kesekiirgus	*Lacerta schreiberi* esinemine /puudumine	Brito *et al.* 1999
Päeva keskmine õhu-temperatuur	*Lacerta schreiberi* esinemine /puudumine	Brito *et al.* 1999
Külmima kuu keskmine õhutemperatuur	*Lichenostomus chrysopsi* e/p	Neave *et al*. 1996
Kõige soojema kuu keskmine õhutempera-tuur	*Chioglossa lusitanica* e/p	Teixeira *et al*. 2001
Pakasekuude arv	*Chioglossa lusitanica* e/p	Teixeira *et al.* 2001
Külmima kuu miinimum	Põõsaliikide levik Californias	Franklin 1998
Tuule kiirus	*Parnassius mnemosyne* levik ja ohtrus	Luoto *et al*. 2001
Päikselisus	*Antechinus minimus maritimus* esinemis-tõenäosus	Gibson *et al*. 2004
	Kardemoniseene *Phytophtora cinnamomi* e/p	Wilson *et al.* 2003
Õhuniiskus	*Lacerta schreiberi* esinemine /puudumine	Brito *et al*. 1999
Suvine sademete hulk	*Graellsia isabelae* levik	Chefaoui ja Lobo 2008
Aastane sademete hulk	*Lacerta schreiberi* esinemine /puudumine	Brito *et al*. 1999
Aastane sademete hulk	*Chioglossa lusitanica* e/p	Teixeira *et al*. 2001
Klimaatilised äärmused	11 hariliku puuliigi levik Šveitsis	Zimmermann *et al.* 2009
Lume paksus	Põdrapopulatsiooni tihedus	Puttock *et al*. 1996

Uurimused

R.T. Clarke *et al.* (1998) jõudsid liblika *Maculinea rebeli* populatsiooni modelleerides järeldusele, et populatsiooni arvukust ja eksistentsi ei mõjuta ainult elupaiga ruumiline konfiguratsioon, vaid ka erineva sobivusega piirkondade paiknemine üksikute elupaigalaikude piires.

Tian *et al*. (2005) näitasid, et rahvastiku tihedust Hiinamaa igas ruutkilomeetris saab üsna hästi hinnata järgnevate kaardiandmete alusel: maakatteüksuste vahekord, teede jõgede ja linnade lähedus, maapinna kõrgus ja kliimaandmed. Koostati 1 km lahutusega Hiina rahvastiku paiknemise mudel, milles linnavälise ala rahvastiku tihedus saadi kaardiandmete järgi kaalutud lineaarsetest mudelitest.

A. Bartel (2000) prognoosis kosmosefotode järgi kahepaiksete esinemist.

E.H. Saveraid *et al.* (2001) rõhutavad, et mäginiitude lindude leviku prognoosimisel võib satelliidi-info olla kasulik, kuid usaldatava prognoosi saamiseks tuleb seda kombineerida detailsemate andmetega elupaiga omaduste kohta, sealhulgas väliuuringutel hangitavatega.

Remm ja Remm (2009) leidsid, et paremad teiste käpaliseliikide esinemise indikaatorid olid keskmise sagedusega käpalised *Ophrys insectifera, Malaxis monophyllos, Gymnadenia conopsea, Dactylorhiza russowii.*

5.7. Elurikkuse kaardistamine

Lihtsate ruumiliste muutujate, nagu näiteks maastikuüksuste mitmekesisuse kaardistamist käsitlesime ruumimustri kirjeldamise osas (peatükk 4). Elurikkus on kompleksne ja abstraktne muutuja, mida on raske otseselt mõõta ja veel raskem usaldusväärsel tasemel ülepinnaliselt kaardistada. Elurikkuse kaartide saamiseks tuleb seosed kaardistatava mitmekesisusnäitaja ja kaudsete, kuid ülepinnaliselt reaalselt mõõdetavate argumenttunnuste vahel modelleerida või esindada teabebaasi kantud näidistena. Seejärel on võimalik koostada elurikkuse ülepinnaline hinnang mudeli või näidiste abil.

Elurikkuse traditsiooniline näitaja on liigirikkus, kuid viimasel ajal on aktuaalseks muutunud liigisise ja maastiku tasemel geneetilise mitmekesisuse kaardistamine (Vandergast *et al*. 2008, Gömöry *et al*. 2010). Tähelepanuväärne on, et taimede geneetiline varieeruvus seostub taimtoiduliste loomade ja röövloomade mitmekesisusega (Crutsinger *et al*. 2006).

Elurikkust saab kaardistada mitmel viisil.Peamisi metoodilisi lähenemisi võib liigitada järgmiselt.

- Interpoleerides üksikkohtades mõõdetud elurikkust (Kumar *et al.* 2006).
- Seostades mitmekesisuse näitajat eelnevalt kaardi- ja kaugseireandmetest arvutatud kohatunnustega (Cardillo *et al.* 1999, Lehmann *et al.* 2002a, b, Bruun *et al.* 2003, Foody 2004, Maes *et al.* 2005, Dogan ja Dogan 2006, Foody ja Cutler 2006, St-Louis *et al.* 2006, Mellin *et al.* 2007, Lin *et al.* 2008, Altamirano *et al.* 2010).
- Seostades elurikkust maastiku tasemel mitmekesisusega (Dauber *et al.* 2003, Honnay *et al.* 2003, Ewers *et al.* 2005, Remm 2005, Matlock ja Edwards 2006, Kumar *et al.* 2006).
- Indikaatorliikide järgi (Chiarucci *et al.* 2007, Pearman *et al.* 2011, Pearman ja Weber 2007).
- Sarnase nõudlusega liikide koosluste levikut modelleerides (McKenzie *et al.* 1989, Underwood et al. 2004).
- Seostades taimestiku gradientide ruumis muutumise kiirust kohatunnustega (Feilhauer ja Schmidtlein 2009).
- Elupaiga- või maakattetüüpide kaardistamise ja neile iseloomuliku elurikkuse järgi (Moore *et al.* 1991, Keith and Bedward 1999, Hernandez-Stefanoni ja Ponce-Hernandez 2004, 2006).
- Kompleksse maastikulise mitmekesisuse mõõdikuna, mis arvestab nii maastikutüüpide rohkust, ökosüsteemide unikaalsust, liigirikkust kui ka häiringute taset (Roy ja Tomar 2000).
- Modelleerides eraldi mitmete liikide levikut ja hinnates liigirikkust prognoositud liikide arvu järgi igas kohas (Parviainen *et al.* 2009, Remm *et al.* 2009). Ükshaaval modelleeritud liikide levilate järgi saab lisaks mitmekesisuse näitajatele eristada nii kooslusi (Lehmann *et al.* 2002a, Ferrier *et al.* 2002b) kui ka sarnase nõudlusega liikide elupaigatüüpe (Lenihan 1993, Austin 1998, Cawsey *et al.* 2002).
- DNA järjestuse vöötkoodide (triipkoodide) andmebaaside alusel (Guralnick ja Hill 2009).

Ferrier ja Guisan (2006) esitasid kümme elurikkuse kaardistamise viisi ja koondasid need kolmeks. põhivariandiks:

- etteantud koosluste, liigirühmade või ordinatsioonitelgede kaardistamine (*assemble first, predict later*),
- esmalt liikide leviku kaardistamine, seejärel erineva liigirikkusega koosluste, liigirühmade või ordinatsioonitelgede tuletamine (*predict first, assemble later*),
- korraga prognoosimine ja rühmitamine (*assemble and predict together*).

Elurikkus võib olla iseseisev funktsioontunnus või erinevate üksikhinnangute koondtulemus

Seosed elurikkuse ja keskkonnafaktorite vahel ei ole ruumiliselt statsionaarsed (see tähendab ei ole erinevates geograafilistes piirkondades samad). Kuna elurikkuse komponendid vahelduvad piirkonniti, siis on põhjust arvata, et ka seosed elurikkuse ja keskkonnafaktorite vahel ei ole eri piirkondades samad. Elurikkuse hinnangulisel kaardistamisel on seoste kohalik omapära olulisem kui üksikliikide leviku puhul. Seoste kohast sõltavat muutlikkust tuleks arvestada seoseid kirjeldavate meetodite valikul. Näiteks tuleks liigirohkuse ja kohatunnuste vahelise seose kirjeldamiseks ja liigirohkuse prognoosimiseks kasutada geograafiliselt kaalutud regressiooni (Foody 2004) või tuleks mudelisse kaasata liigirohkuse ruumiline autokorrelatsioon (Bacaro *et al.* 2011).

Seosed elurikkuse ja kohatunnuste vahel ei ole ruumiliselt statsionaarsed

Koosluste erinevuse kirjeldamiseks on soovitatud **üldistatud erinevuse modelleerimist** (*generalized dissimilarity modelling – GDM*), mis on maatrikskorrelatsiooni või maatriksregressiooni variant, kus seosed esitatakse erinevustena koosluste vahel. Erinevused seostatakse geograafilise vahemaaga või keskonnatingimuste erinevuse ja kaugseireandmete kombinatsiooniga (Ferrier *et al.* 2002b, 2007). Maatriksite erinevuse olulisuse kontrolliks sobib Mantel test. Tulemuste visualiseerimiseks sobib mitmemõõtmeline skaleerimine ning koosluste gradiendi kujutamine sujuvalt muutuvate värvidega.

Elurikkuse modelleerimise tulemuste kohaselt on olulisemad elurikkust mõjutavad tunnused ja faktorid: kliima (paljudes uuringutes esindab kliimafaktoreid maapinna kõrgus) (De Mas 2009, Altamirano *et al.* 2010), maastiku mitmekesisus (mida väljendab ka kaugseire kujutisest arvutatud näitajate varieeruvus) (Bacaro *et al.* 2011), maakasutuse intensiivsus ja maakasutuse muutumise kiirus (Lütolf *et al*. 2009, Vicente *et al*. 2010) ja muud elupaiga eripärad (Mellin *et al.* 2007, De Mas 2009). Lokaalne liigirikkus sõltub tugevasti ka piirkonna liigifondist ehk liigivarust (*species pool*). Liigirikkus on üldiselt suurem tingimustes, mis on eelneva evolutsioonilise aja jooksul olnud püsivamad ning seega on evolutsiooniprotsessidel olnud pikemalt aega toimida (Zobel 1992, 1997, Zobel *et al*. 2011).

Mitmed autorid on leidnud, et haruldaste liikide esinemiskohad ei ole kõige liigirikkamates kohtades (Benayas ja de la Montaña 2003, Chiarucci *et al.* 2007), pigem on liigirikkuse tulipunktid (*hotspots*) seal, kus esinevad paljud sagedased liigid (Pearman and Weber 2007).

Uurimused

Grace ja Pugesek (1997) sidusid vaatlusala taimede liigirohkuse häiringute, abiootiliste faktorite ja samade liikide biomassiga. Nad näitasid, et üldistatud abiootilisted tingimused, nagu merevee üleujutused, võivad mõjutada liigirohkust mulla soolsuse ja taimede biomassist tulenevate valgustingimuste kaudu.

M. Luoto kasutas Landsat TM kujutisi, kombineeritult maapinna kõrgusmudeliga, taimkatte liigilise mitmekesisuse modelleerimiseks. Liigirikkuse varieeruvusest õnnestus seletada 57%. Tugevaimad seosed olid nõlvadel oleva lehtmetsa ja looduslike rohumaade ja hoonestusala hulgaga. Prognoositi ka uurimata ala haruldaste liikide rohkuse tulipunkte (Luoto 2000).

Stockwell ja Peterson (2003) leidsid, et Mehhiko lindude üksikliikide mudelitest koostatud mitmekesisuse kaart on palju detailsem ja usaldusväärsem kui lihtne liikide leiuandmete koondamine või mitmekesisuse prognoos taimkatte järgi.

Remm (2005) arvutas korrelatsioone puistu mitmekesisuse ja ümbruses oleva maastiku mitmekesisuse vahel sõltuvalt ümbruse kaugustsoonist. Seosed olid kõige tugevamad veidi kaugema tsooniga, mitte kõige lähema ümbruse maastikuga.

St-Louis *et al.* (2006) seostasid lindude liigirohkust kaugseirekujutiste tekstuuriga. Tekstuuri parameetrid üksi suutsid seletada kuni 63% linnuliikide arvust, kaks tekstuuri parameetrit koos üldistatud elupaigatüübiga kuni 76%.

Wiegand *et al.* (2007a) kaardistasid puude liigirohkust troopilises vihmametsas ja seostasid seda kindlat liiki üksikpuude lähedusega. Enamiku liikide üksikpuu mõju teiste puude liigirikkusele ulatus vaid 20 m kauguseni. Nad leidsid, et puuliigid võivad olla mitmekesisuse tõrjujad või koondajad.

Altamirano *et al.* (2010) leidsid metsa liigilist mitmekesisust modelleerides ja kaardistades, et Tšiili kaitsealad ei paikne (regressioonimudeli järgi arvutatud) kõige liigirikkamates kohtades. Metsa põõsarinde mitmekesisus ei seostunud puurinde mitmekesisusega ega kasutatud kohatunnustega.

Estes *et al.* (2010) uurisid seoseid kaugseirekujutisest arvutatud tunnuste ja puistu struktuurse keerukuse indeksi vahel, mis kirjeldab mikroelupaikade tihedust.

Bacaro *et al.* (2011) modelleerisid linnuliikide hulka üldistatud lineaarse mudeliga, millesse kaasati ruumiline autokorrelatsioon. Algul arvutati mitmekesisus ruumilist autorrelatsiooni arvestamata. Saadi regressioonijäägid, mida modelleeriti variogrammiga.

Mereorganismide leviku ja mitmekesisuse kaarte on AquaMaps veebilahenduses http://www.aquamaps.org.

5.8. Puistu andmete hinnanguline kaardistamine

Puistu takseertunnuste (katvus, juurdekasv, tagavara, biomass, rinnaspind, vanusklass, puude tihedus, kõrgus) hinnanguline kaardistamine on põhiliselt toimunud regressioonimudelite ja *k* lähima naabri (*kNN*) meetodi abil, mida on kirjeldatud peatükis 3.4.6. *kNN* meetodi puhul saadakse hinnang ümbruses olevate sarnaste näidisalade järgi. Sarnased, kuid teatud vahemaast kaugemal olevaid näidisalasid ei kasutata. Näidisala panust hinnangusse kaalutakse vastavalt sarnasusele ja kaugusele.

Soome riiklikus metsakorralduses on kaugseireandmeid kNN meetodit rakendades kasutatud 1990ndate aastate algusest (Tomppo ja Katila 1991, Tomppo ja Heikkinen 1999). *kNN* meetodi üksikasju ja erinevate lähteandmete kasutatavust on soomlased uurinud mitme kandi pealt. Kilpeläinen ja Tokola (1999) kasutasid olemasolevaid metsakorraldusandmeid satelliidipildi järgi puistu tagavara hindamisel. Prognoosis kasutati pikslite kaalumist sarnasuse järgi, arvestati eraldiste piire ja eraldiste piiril olevaid piksleid töödeldi eraldi. Tulemuse täpsus sõltus tugevasti metsakorralduse eraldiste piiride ja satelliidipildi geomeetrilise ühildamise täpsusest. Teises uuringus (Tokola ja Kilpeläinen 1999) võrdlesid nad puidu tagavara prognoosi täpsust eraldiste piiril olevaid segapiksleid kasutades ja kasutamata. Eraldise piirialade kohta oli prognoos vähetäpne, kuid segapikslite väljajätmine testandmetest põhjustas nihkega hinnangu. Parimaks lahenduseks osutus külgnevate vaatlusalade andmete kombineerimine. T. Tokola (2000) näitas, et puidu tagavara hindamisel näidisalade ja Landsat TM kujutise järgi peaksid näidisalad paiknema mitte kaugemal kui 20 km prognoositavast kohast, kuna fenoloogiline areng, metsa eripära ja pildiomadused on ruumis varieeruvad. Tokola *et al*. (2001) kasutasid kõrgusmudelit Nepaali metsade inventeerimisel Landsat TM pildi järgi. M. Katila ja E. Tomppo (2001) leidsid, et näidisalad peaksid olema 40…50 km raadiuses arumetsade ja 60…90 km raadiuses soometsade puhul. A. Pekkarinen (2002) mõõtis metsa takseertunnuste hinnangu paranemist, kui kasutada detailsemat, 1,6 m pikslikülgega eelnevalt eraldisteks segmenteeritud kaugseirepilti. H. Mäkelä ja A. Pekkarinen (2001) võrdlesid puidutagavara hinnangu täpsust kaugseirepildi piksli eri suurusega naabruse kasutamisel koos eelneva pildi segmenteerimisega ja ilma selleta. Metsa segmenteerimine eraldisteks andis hinnagutäpsuste mõningase tõusu. J. Heikkilä *et al*. (2002) kasutasid satelliidipilte ja aerofotosid metsakahjustuste hindamisel.

5.8.1. Puistu tunnuste kaugkaardistuse täpsus

Puistu takseertunnuseid on kaugseireandmete järgi hinnatud erineva edukusega. Puistu klassifitseerimine põhiliste enamuspuuliikide järgi on õnnestunud sõltuvalt liigist 50…84% täpsusega (Reese *et al*. 2002), 65…87%, kuid segapuistude puhul väiksema täpsusega (Dymond *et al*. 2002), tasemel $\kappa = 0{,}41$ Landsat kujutise järgi ja $\kappa = 0{,}38$ ortofoto järgi (Tamm ja Remm 2009), viis klassi 38% täpsusega (Luther *et al*. 2006), Otepää looduspargis 10 klassi 79% täpsusega, $\kappa = 0{,}58$ (Remm 2004), väiksemal uurimisalal on saavutatud ka 80…90% täpsust (Gong *et al*. 1997, Carleer ja Wolff 2004, Chubey *et al*. 2006). Wolter *et al*. (1995) saavutasid 13 metsatüübi eristamisel 80% täpsuse, Bauer *et al*. (1994) eristasid 11 metsa klassi 80% täpsusega, Franco-Lopez *et al*. (2001) eristasid 13 puistutüüpi 52% õigsusega ja kolm tüüpi 64% õigsusega ($\kappa = 0{,}45$), Chubey *et al*. (2006) puistu liituvuse kolme klassi 85% ja männi osakaalu nelja klassi 86% täpsusega. Metzler ja Sader (2005) saavutasid kokkulangevuse metsakorralduskaartidel olevate metsatüüpide ja Landsat ETM+ andmete abil koostatud hinnangute 76...79% kokkulangevuse. Okas-, leht- ja segametsa eristamisel on saavutatud kokkulangevust välivaatlustega: 56…62%, seejuures lehtmets üle 90% täpsusega (Oviir *et al*. 2008), 65% (Franklin *et al*. 2000), 80…90% (Lewinsky 2006).

Puistu katvuse ja võrade liituvuse kaardistamisel kaugseire andmete abil on saavutatud korrelatsioon R = −0,64 kahest erinevast kuupäevast pärit Landast viienda kanali kiirgusväärtuste erinevuse ja välivaatluse vahel (Gemmell *et al*. 2001), R^2 = 0,74 välivaatluse ja Landast kanalite 3, 4 ja 5 väärtusi hõlmavast mudelist saadud prognoosi vahel (Carreiras *et al*. 2006), R^2 = 0,74 ja RMSE = 12% (Cohen *et al*. 2001), R^2 = 0,57 ja RMSE = 12% (Hall *et al*. 2006).

Puidu tagavara ja biomassi hindamisel kaugseire andmete järgi on tulemused enamasti vähem usaldusväärsed. Hyyppä *et al.* (2000) saavutasid hinnangute suhtelise standardvea 12%, Muinonen *et al*. (2001) 18...27%, Franco-Lopez *et al*. (2001) RMSE = 51,6 m^3/ha, Katila ja Tomppo (2001) tulemustes on puidumahu hindamise suhteline RMSE 64...295%, Wallerman *et al*. (2002) 41...58%, Reese *et al.* (2003) 58...69%, Mäkelä ja Pekkarinen (2001) töös 79...83%, Mäkelä ja Pekkarinen (2004) saavutasid puistu tasemel suhtelise RMSE = 48%, kuuskede puhul 81%, männi ja lehtpuude puhul ületas viga 100%, Tamm ja Remm (2009) jõudsid Landsat andmete abil tulemuseni 36% ja ortofotode järgi 41%. *kNN* meetod ja regressioonimudelid on puidu tagavara kaardistamisel andnud sarnaseid tulemusi (Tomppo *et al*. 2002). Muukkonen ja Heiskanen (2005), suutsid puistu maapealset biomassi kaugseire andmete järgi modelleerida keskmise suhtelise veaga RMSE = 44,7%, Heiskanen (2006) suhtelise veaga RMSE = 41%.

Puistu struktuuri, katvuse ja metsa maapealse biomassi hindamisel on edukalt kasutatud satelliitidel paiknevate mitmesageduslike mitme polarisatsiooniga radarsensorite andmeid. Radarsensorite eelis on võime näha läbi pilvede ja sõltuvalt kiirguse sagedusest ka eri sügavusele puude võradesse. Tänu mitmerindelisele sondeerimisele on saadud metsa biomassi hinnanguid täpsusega 10…15 kg/m^2 (Ranson ja Sun 1994, 1997, Ranson *et al*. 2001).

Puistu rinnaspinna modelleerimisel profileeriva radari andmete järgi saavutasid Hyyppä *et al.* (2000) standardvea SE = 6,07 m^2/ha, suhteline viga 32%, R^2 = 0,59, Franco-Lopez *et al*. (2001) SE = 7,2 m^2/ha.

Ülevaate lidartehnoloogia kasutamisest metsa struktuuri kaardistamisel on kirjutanud Lefsky *et al.* (2002). Üksikute puuvõrade eristamist lidari andmete abil on katsetanud Pouliot *et al.* (2002), Heurich *et al*. (2004). Brandtberg *et al*. (2003) üritasid ka puuliike tuvastada. Lefsky *et al*. (2005) uurisid lidari kasutatavust puistu maapealse biomassi ja vanuse hindamiseks, Hudak *et al.* (2008) määrasid lidari andmete järgi rinnaspinda ja puude tihedust. Bütler ja Schlaepfer (2004) tuvastasid üksikuid kuivanud kuuski infrapuna-aerofotodelt. Uuttera *et al*. (1998) ja Pouliot *et al*. (2002) eristasid üksikuid puuvõrasid aerofotolt. Leckie *et al*. (2003) saavutasid üksikute puude liigi äratundmisel mitmekanalilise spektromeetriga madalal lennul saadud kujutisest vaid 7,25% vea. Liikide kaupa varieerus viga piires 5,7…13,3%.

Noorendike vanust õnnestub ootuspäraselt hinnata väiksema absoluutveaga kui raieküpset metsa, kus hinnangu keskmine viga võib olla üle 40 aasta (Reese *et al.* 2003).

Ülevaate geostatistiliste meetodite kasutamisest metsa kaugseires esitasid Zawadzki *et al*. (2005).

5.9. Indikatsioon

Bioindikatsioon on keskkonnatingimuste kaudne hindamine bioloogiliste näitajate järgi. Modelleerimisülesande puhul on küsimus, millist indikaatortunnuse väärtust (k. a. esinemine/puudumine) võib oodata etteantud keskkonnatingimuste korral. Indikatsiooniülesande puhul aga vastupidi: milline võiks olla ümbruskond, kui tuvastatakse indikaator? **Indikatsioon** on seega indikaatortunnuste järgi sihttunnuste väärtuste kaudne hindamine. Indikatsiooni üks osa on **bioindikatsioon** ehk indikatsioon bioloogiliste liikide esinemise või ohtruse järgi. Bioindikaatoritena kasutatakse väga erinevaid organisme ja bioloogiliste protsesside parameetreid. Sagedamini kasutatavatest võiks mainida taimi, samblikke, ränivetikaid, veekogude põhjaloomastikku, vihmausse ja lülijalgseid ning protsessidest näiteks primaarproduktsiooni intensiivsust.

Milleks kasutada kaudset mõõtmist ehk indikaatoreid, kui võiks ka otseselt mõõta? Indikatsioon võib olla õigustatud, sest:

- mõõteriistad on tihti liiga tundlikud, indikaator võib anda pikema aja jooksul või suuremalt alalt akumuleeritud efekti,
- vajalik sihttunnus on kompleksne, ei ole ühe parameetri kaudu otseselt mõõdetav,
- puudub tehnoloogia, mis võimaldaks uuritavat tunnust otseselt mõõta,
- indikaatori kasutamine võib olla odavam,
- indikaatortunnuse kasutamisel võib hinnangu saada kiiremini,
- otseseid mõõtmisi saab teha vaid reaalajas. Ei ole võimalik minna minevikku või tulevikku ja seal vajalikke mõõtmisi teha. Indikaatortunnustel võib olla mälu või tuleviku prognoosiv võime.
- Tuleviku ennustamiseks võib otsida analoogilisi (klimaatilisi) olukordi minevikust. Andmed mineviku kohta aitavad hinnata ka inimmõju ja looduslike protsesside vahekorda kaasajal.

Bioindikatsiooniga seotud olulisemad mõisted on järgmised (van Straalen 1998).

Bio-katse (*bio-assay*) – ökotoksikoloogiline test, katseorganismide ja nende füsioloogiliste ning geneetiliste tunnuste kasutamine keskkonnaseisundi hindamisel.

Bio-marker (*bio-marker*) – ökotoksikoloogiline või molekulaarbioloogiline indikaator.

Bio-reporter (*bio-reporter*) – molekulaarbioloogiline süsteem keemilise aine olemasolu muutmiseks hästi mõõdetavaks signaaliks, näiteks luminestsentsiks.

Bio-sensor (*bio-sensor*) – molekulaarbioloogiline seade keemilise signaali muutmiseks elektriliseks signaaliks.

Signaalliik – kvantitatiivne indikaator.

Seire ehk monitooring ehk keskkonnaseisundi või muude muutuvate nähtuste pikaajaline sama metoodikaga jälgimine. Seire võib toimuda nii otseste mõõtmiste kui ka kaudsete tunnuste ehk indikaatorite abil. Bioloogiliste indikaatorite kasutamisel nimetatakse **biomonitooringuks**.

Kalibreerimine – üldiselt metoodika või mõõteriista etaloniga või õpetusandmetega vastavusse viimine. Bioindikatsioonis tähistab kalibreerimine indikatsioonimeetodi sobitamist ehk parima vastavuse leidmist indikaatortunnuste ja sihttunnuste väärtuste vahel. Sobivate indikaatortunnuste leidmine ja nende kalibreerimine on keerukaim andmetöötluslik ülesanne indikaatorite kasutamisel. Indikaatortunnuste sobivus tähendab sobivust ülesande kohasuse mõttes, mõõdetavuse mõttes, konstantsuse mõttes (indikaatortunnus peaks olema ajas samatähenduslik või siis ajas kalibreeritav), kindluse mõttes (kui kindlalt võimaldab järeldusi teha), tundlikkuse mõttes (sobiv tundlikkus sõltub uurimisteemast: mida väiksem objekt, seda rohkem sõltub ta ümbrusest. Globaalsete muutuste jälgimiseks sobivad Antarktika liustikud ja Maailmameri. C.J.F. ter Braak (1995, 1988a,b) nimetab

kalibreerimiseks kogu bioindikatsiooni, mis hõlmab ka konkreetse prognoosi koostamist.

Kõige pikemad bioindikatsiooni traditsioonid on hüdrobioloogias. Juba A.N. Hassal (1850) hindas vee kvaliteeti füto- ja zooplanktoni liikide järgi. Vee mikrobioloogilise analüüsi alusepanijaks peetakse F. Cohni (Cohn 1870). Esimese põhjalikuma loetelu vee puhtuse või reostatuse indikaatoritest esitas C. Mez (1898). Mezi süsteemi arendasid edasi Kolkwitz ja Marsson (1902), kellelt pärinevad terminid **saprobiont** ja **katarobiont**, vastavalt orgaaniliste ainetega reostatud vett ja täiesti puhast vett eelistavate organismide kohta, ning ka polüsaproob ja oligosaproob. Hilisemates töödes täiendasid Kolkwitz ja Marsson korduvalt indikaatorliikide nimekirju (Kolkwitz 1911). Kui Kolkwitz ja Marsson klassifitseerisid organisme vastavalt nende sabroobsuseelistustele, siis hilisemad autorid Pantle ja Puck (1955) ning Sládeček (1973, 1983) arvutasid veeorganismide järgi veekogu saproobsuse numbrilist väärtust (saproobsusindeksit – *S*). Pantle ja Pucki järgi on iga organism seotud vaid ühe saproobsustsooniga ning veekogu saproobsushinnang saadakse iga leitud liigi esinemissageduse järgi kaalutud saproobsusindeksi keskmisena.

$$S = \frac{\sum_{i=1}^{n} h_i s_i}{\sum_{i=1}^{n} h_i}, \qquad [5\text{-}55]$$

kus h_i on liigi *i* ohtruse klass, s_i on liigi *i* indikaatorkaal ja *n* on teadaolevate indikaatorliikide arv.

M. Zelinka ja P. Marvani (1961) metoodika kohaselt hinnatakse veekogu saproobsust mitte ühe arvuga, vaid saproobsete valentside jaotusena. Iga saproobsusklassi valents arvutatakse Zelinka ja Marvani järgi selle klassi indikaatorliikide indikaatorkaalude ja sageduste korrutise osana kõigi indikaatorliikide vastavate korrutiste summas.

$$A_k = \frac{\sum_{i=1}^{n} a_{ki} h_i g_i}{\sum_{i=1}^{n} h_i g_i}, \qquad [5\text{-}56]$$

kus A_k tähistab saproobset klassi *k*, a_{ki} on liigi *i* väärtus klassi *k* indikaatorina, h_i on liigi *i* ohtruse näitaja, g_i on liigi *i* indikaatorkaal ja *n* on teadaolevate indikaatorliikide arv.

Liigid esinevad reeglina rohkem kui ühe saproobsusklassi elupaikades, omades iga elupaigatüübi suhtes erinevat indikaatorväärtust. Kitsa ökoloogilise amplituudiga liik on usaldusväärsem ja tema indikaatorkaal suurem.

Liigi **indikaatorväärtus** (*indicator value*) ehk indikaatorvalents on liigile (tunnusele) omistatud subjektiivne või kogemuslik väärtus, mis väljendab liigi suhtelist sagedust erinevate keskkonnatingimuste korral. Liigi indikaatorväärtus on seega liigi optimum mingil keskkonnateguri teljel. Lisaks indikaatorväärtusele omistatakse liikidele ja vaatlustele usaldusväärsuse ehk indikaatorväärtuse kaalud.

Indikaatorväärtus näitab indikaatorliigi eelistust, indikaatorkaal indikaatori usaldusväärsust

Botaanikas pärineb indikaatorväärtuste ja indikaatorkaalude kasutamine Saksa geobotaanikult Heinz Ellenbergilt (Ellenberg 1948, Ellenberg *et al.* 1992), kes kasutas taimeliikide esinemist mullatingimuste indikaatorina. Ellenbergi indikaatorväärtuste (*Zeigerwerte*) süsteemi kasutatakse Euroopas seniajani. Ühe või teise taimeliigi indikaatorväärtust saab vaadata ka internetist (http://statedv.boku.ac.at/zeigerwerte/?#Skalierung). Liikide tasemelt indikaatorkoosluste kasutamiseni on jõutud nii hüdrobioloogias (Fjerdingstad 1964) kui ka maismaaökoloogias.

Cajo ter Braak (ter Braak *et al.* 1993, ter Braak ja Juggins 1993) on indikaatorväärtuste käsitlust edasi arendanud **kaalutud vähimruutude** metoodikaks (*weighted averaging partial least squares – WA-PLS*). **Vähimruutude osaregressioon** (*PLS*) kasutab regressioonimudelis originaaltunnuste lineaarkombinatsioone, mis arvutatakse argument- ja funktsioontunnuste vahelise kovariatsiooni abil. WA-PLS on välja töötatud unimodaalse vastavusfunktsiooni ja binaarse indikaatortunnusega kalibratsiooni jaoks. Seega sobib see meetod vaid indikaatorliigi esinemise või puudumise andmetele ning eeldusega, et igal liigil on üks esinemise optimum iga keskkonnafaktori suhtes. Hinnang antakse liikide indikaatorväärtuste kaalutud keskmisena. Liikide optimumid aga hinnatakse õpetusandmetest selliselt, et nende kaalutud keskmine hindaks kõige täpsemalt keskkonnafaktorite väärtusi.

Otseselt raskesti mõõdetav on ka elurikkus. Elurikkuse ja ökosüsteemide kaudse hindamise indikaatorite enimkasutatud nimekirjad on koostatud Euroopa Liidu projekti *Streamlining European* 2010 *Biodiversity Indicators* – SEBI2010 raames. Ülevaade nendest nimekirjadest on publikatsioonis Petrou ja Petrou (2011).

Spetsiaalset tarkvara palaeoökoloogiliste indikaatorite analüüsimiseks on kirjutanud Stephen Juggins Newcastle ülikoolist (http://www.staff.ncl.ac.uk/staff/stephen.juggins).

5.9.1. Statistiline kalibreerimine

Indikatsiooniülesanne on lahendatav statistilise mudeli, tüüpilisel juhul regressioonimudeli abil. Nagu statistilise modelleerimise puhul ikka, peab meetodi ja mudeli valiku uurija ise otsustama. Üks esimesi küsimusi on, kas kasutada indikaatortunnuseid ühekaupa ja üksteisest sõltumatult või korraga. Kui kasutatakse mitut indikaatorit eraldi, siis tuleb üksikute indikaatortunnuste alusel saadud hinnangud kuidagi ühendada. Näiteks üksikhinnangute kaalutud keskmisena üksikute tõenäosusjaotuste ühendamisel Bayesi valemi järgi.

Kui prognoositavaid muutujaid on üks, siis saab kasutada mitmetunnuselist regressiooni, lülitades regressioonimudelisse mitu indikaatortunnust. Nii võib näiteks kliima rekonstruktsioonide puhul keskmist temperatuuri ja sademete hulka eraldi hinnata. Kui aga eelistatakse prognoositavaid muutujaid käsitleda ühtse komplektina, tuleb kas funktsioontunnuste väärtuskombinatsioonid tüüpideks jagada või kasutada mitmemõõtmelisi meetodeid, nagu näiteks kaalutud vähimruudud.

Mitmemõõtmelise indikatsiooni kalibreerimise korral moodustavad indikaatortunnused (liigid) vektori Y ja prognoositavad tunnused (keskkonnategurid) vektori X. Kummalgi vektoril on mitu komponenti. Kasutame siinkohal Y ja X tähistust selliselt, kuna eeldame põhjuslikku sidet keskkonnalt indikaatortunnustele ehk X-lt Y-le. Klassikaline indikatsiooniülesanne on leida kõigepealt, kuidas Y sõltub X-st ja seejärel leitud seost teistpidi rakendada. Kuna regressioonid ei ole pööratavad, siis on sellise pööramise tulemusel saadud hinnang nihkega ja ligikaudne. Klassikaline indikatsiooniülesanne lähtub järgmisest mudelist (ε on prognoosiviga):

$$\mathrm{Y} = f(\mathrm{X}) + \varepsilon. \qquad [5\text{-}57]$$

Pöördmeetodi (*inverse method*) korral püütakse leida indikaatortunnuste väärtuste kombinatsioonile Y kõige paremini vastavaid X väärtusi. Näiteks, kui leitakse setetest mingi hulk mingite liikide õietolmu, siis püütakse leida koht ja keskkonnatingimused, kus esineb sellisele õietolmu koosseisule vastav taimekooslus tänapäeval. Parimad tänapäevased analoogid näitavadki minevikus tõenäoliselt olnud keskkonnatingimusi. Pöördmeetod tugineb enamasti sarnasusele ja sel juhul regressiooniseost keskkonnategurite ja indikaatortunnuste vahel ei arvutata. Pöördülesanne lähtub mudelist

$$\mathrm{X} = g(\mathrm{Y}) + \varepsilon. \qquad [5\text{-}58]$$

Ei saa kindlalt väita, et üks lähenemisviis oleks teisest parem. Väidetakse, et õpetusandmetega hästi kaetud muutumispiirkonnas on pöördmeetod enamasti täpsem, aga klassikalised meetodid annavad parema tulemuse, kui seoseid on vaja ekstrapoleerida õpetusandmetega katmata või vähekaetud muutumispiirkonda (ter Braak 1995).

Kalibreerimismeetodeid võib jagada ka lineaarseteks ja mittelineaarseteks, parameetrilisteks ja mitteparameetrilisteks (viimased kasutavad mingit lokaalset silumist), dimensionaalsuse järgi (kas suur hulk tunnuseid koondatakse väiksemaks arvuks üldistatud näitajateks või ei) (tabel 10). Lihtsad kalibreerimismeetodid kasutavad lineaarset mudelit. Mittelineaarsed kalibreerimismeetodid lähtuvad **Shelfordi tolerantsireeglist**. See on Liebigi miinimumireegli edasiarendus, mille järgi on organismile limiteeriv see faktor, mis kõige enam läheneb organismi tolerantsipiirile. Sellest tulenevalt eeldatakse, et iga liigi esindajatel on kõige soodsam vältida konkurentsi ja eelistada optimumilähedaste keskkonnatingimuste kombinatsiooni ehk Hutchinsoni nišši.

Vastavuspinna meetod indikatsioonis on seletavate tunnuste väärtuste sobitamine funktsioon-tunnuse etteantud väärtusele vastavaks. Vastavusfunktsioon (*response function*, *response curve*) kujutab seost seletava tunnuse ja funktsioontunnuse vahel (ptk 3.4.1). Vastavuspind (*response surface*) seost mitme seletava tunnuse korral. Indikatsiooni pöördülesande puhul püütakse vastavuspinna väärtuse järgi teha järeldusi seda tinginud faktorite kohta. Vastavuspinda saab arvutada mingit regressioonimudelit andmetele sobitades või mitteparameetrilise lokaalse sobitamisega (silumisega).

Tabel 10. Kalibreerimismeetodid C.J.F. ter Braak (1995) järgi täiendatult.

Meetod	Viited	Tüüp	Mudel	Transf.	Sobitamine	Dimens.
SRS	Prentice *et al.* 1991	klassikaline	mittelineaarne	ei	lokaalne	täis
GLR	Birks *et al.* 1990, ter Braak ja van Dam 1989	klassikaline	unimodaalne	ei	globaalne	täis
MLM	Ter Braak ja van Dam 1989, ter Braak *et al.* 1993	klassikaline	unimodaalne	E(y)	globaalne	täis
kNN	Overpeck *et al.* 1985, Quiot 1990	pöörd	mittelineaarne	ei	lokaalne	täis
S-IR	Huntley ja Prentice 1988	pöörd	mittelineaarne	ei	lokaalne	täis
PLS	Borggaard ja Thodberg 1992	pöörd	lineaarne	ei	globaalne	vähendatud
WA-PLS	Ter Braak *et al.* 1993, ter Braak ja Barendregt 1986	pöörd	unimodaalne	jah	globaalne	vähendatud
SRS	Remm 1989	pöörd	mittelineaarne	jah	lokaalne	täis

Transf. on lineariseeriva transformatsiooni olemasolu (jah, ei, *E(y)* – oodatavad väärtused). Globaalsed meetodid sobitavad parameetrilist funktsiooni, lokaalsed meetodid kasutavad lokaalset mitteparameetrilist silumist. SRS on silutud vastavuspind (*smooth response surface approach via local averaging*). GLR on probitregressioon. MLM on multinominaalne logitregressioon. kNN on näidistele tuginev järeldamine *k* lähima naabri järgi. S-IR on segmenteeritud lineaarne pöördregressioon. PLS on vähimruutude osaregressioon (*partial least squares*). WA-PLS kaalutud vähimruudud (*weighted averaging partial least squares*).

Uurimused

Oostermeijer ja van Swaay (1998) leidsid seoseid Ellenbergi süsteemis troofsus-, happesus- ja niiskusväärtuse ning liblikaliikide esinemise vahel Hollandis. Enamik statistiliselt olulisi seoseid näitas selge keskkonnatingimuste optimumi olemasolu. Leitud seoseid saab kasutada nii keskkonna muutuste mõju hindamiseks liblikakooslusele kui ka keskkonnaomaduste hindamiseks liblikate kui indikaatorite järgi.

Caroll ja Pearson (1998) näitasid, et liblikate liigirohkust saab määrata liivikate (*Cicindelidae*) liigirohkuse järgi.

Sajanditetaguse aja ilmastiku kohta hoiavad teavet aastarõngad vanade ehitiste palkides. Tuntud on seos männi aastarõngaste laiuse ja suve keskmise temperatuuri vahel Soomes (Helama *et al*. 2010). Eesti geograafilises piirkonnas on leitud positiivne seos aastarõnga laiuse ja kevadtalvise perioodi temperatuuri vahel (Läänelaid ja Eckstein 2003).

De Cáceres ja Legendre (2009) kirjeldavad 12 indeksit indikaatorliikide leidmiseks ja nende indikaatorväärtuse mõõtmiseks.

5.9.2. Tõenäosuslik indikatsioon

Tõenäosusliku indikatsiooni puhul kasutatakse vastavuspinna või sarnaste analoogide asemel tõenäosusjaotusi. Mõnikord vastab üksiku indikaatortunnuse väärtusele mitu täiesti erinevatest keskkonnatingimustest pärinevat näidist. Sellisel juhul tuleks otsida täiendavaid tõendeid ja indikaatoreid ning õigeks lugeda see prognoos, millele viitavad mitmed indikaatorid. Variandid, mida teised tunnused ei kinnita, tuleb lugeda väärlahenditeks. Näiteks männi kui kasvukohatüübi indikaatori rohke esinemine võib viidata nii nõmmele kui ka rabale. Kui samast kohast on leitud ka vaevakaske, on vaatluskohas ilmselt raba.

Remm (1989) kasutas zooplanktoni kooseisule ja liikide ohtrusele sarnaste varasemate vaatluste klorofülli sisalduse tõenäosusjaotuste kaalutud keskmisi rannikumere eutrofeerumise indikatsioonis.

5.10. Vigade allikad ruumilistes hinnangutes

Statistiliste mudelite puhul eeldatakse tavapäraselt, et argumenttunnused on omavahel sõltumatud, ei sisalda vigu ega autokorrelatsiooni. Reaalsuses selline eeldus paraku enamasti ei kehti. Diskussiooni vigade modelleerimisest ja edasikandumisest geoinformaatilistes analüüsides alustas M.F. Goodchild (1993). Ülevaated ruumiandmete kvaliteedi temaatikast võib leida raamatutest Shi *et al.* (2002) ning Devillers ja Jeansoulin (2006). R. Aspinall (1992) mainib nelja põhilist ruumilise prognoosi vigade allikat:

- ebatäpsused lähteandmetes,
- mudeli üldistatus ja ebatäpsus,
- vigade võimendumine andmete kombineerimisel,
- tulemuste üldistamine.

Liikide ja elupaikade kaardistamise ebakindlust ja tulemuste määramatust on jagatud episteemiliseks ja lingvistiliseks (Regan *et al*. 2002). Episteemiline määramatus on tunnetuslik, lingvistiline määramatus on seotud keelega ja sellest arusaamisega. Tunnetuse ebakindlus sisaldab: juhuslikke mõõtmisvigu, süstemaatilisi hälbeid, objekti muutumist, objektile omast juhuslikkust, mudeli hägusust, subjektiivseid otsuseid. Lingvistiline määramatus sisaldab: mõistete ja terminite hägusust, mõistete tähenduse sõltuvust kontekstist, sõnade mitmetähenduslikkust, üldsõnalisust.

Barry ja Elith (2006) loetlevad järgmisi vigade ja määramatuse allikaid liikide leviku mudelites jagades need andmevigadeks ja mudeli vigadeks.

Andmevigade allikad on:

- puuduvad seletavad tunnused,
- ebapiisav valimi maht,
- tendentslik valim,
- liigi mitte-esinemise kohtade puudumine andmetes,
- vead ja ebatäpsused tunnustes.

Mudelist tulenavad vead jagunevad:

- ebasobivaks mudeli tüübiks ja
- ebasobivaks mudeli parameetrite määramise viisiks.

Subjektiivsete otsuste mõju hinnangulistele kaartidele on vähe uuritud. Ometi on subjektiivsetel otsustel oluline osa näiteks looduses oleva kontiinumi liigitamisel kasvukohatüüpidesse või taime-kooslustesse, mitmesuguste hindepunktide andmisel, vaatluskohtade valikul, lähteandmete ja andmetöötluse nii ruumilise mõõtkava kui ka temaatilise detailsuse valikul, samuti meetodi ja kontrollstatistiku valimisel ja väheusaldatavate vaatluste eemaldamisel. Subjektiivsete otsuste mõju kandub edasi mudelitesse, hinnangutesse, teisendatud andmetesse ja analüüside tulemustesse (Ray ja Burgman 2006).

Samade andmete alusel loodud erinevate modelleerimismeetodite tulemused võivad omavahel üsna palju erineda. Kaasaegset levikut üldistavad mudelid võivad oodatavale kliimamuutusele vastavate muudetud lähteandmete kasutamisel prognoosida levila piiride muutumisel lausa vastassuunalisi arengutendentse (Pearson *et al*. 2006).

5.10.1. Valimi esinduslikkus

Kaugseirepõhises maakattekaardistuses on levinud arusaam, et tõese hinnangu saamiseks peab õpetuspikslite arv igas eristatavas kategoorias olema vahemikus 10 kuni 30 korda suurem kui kasutatud satelliidi (või muu sensori) salvestatud kujutise kiirgusvahemike ehk kanalite arv (Mather 1999). Pal ja Mather (2003) järgi piisab otsuste puu abil klassifitseerimisel 300 pikslist klassi kohta ja 15–30 tunnusest. Van Niel *et al*. (2005) jõudsid katsetulemuste alusel järeldusele, 95% täpsuse saavutamiseks piisab igas eristatavas kategoorias õpetusalade hulgast, mis on vahemikus 2 kuni 4 korda suurem kui kanalite arv. Hjort ja Marmion (2008) võrdlesid erinevate kaugseire andmete järgi igikeltsa leviku modelleerimismeetodeid erineva õpetusandmete mahuga ja leidsid, et alates 200 õpetuskohast annavad mudelid enam vähem sama täpsusega tulemusi.

Ka liikide leviku mudelite usaldatavus sõltub õpetusandmete hulgast (Hirzel ja Guisan 2002, Reese *et al*. 2005). Realistliku mudeli loomiseks vajalik minimaalne leiukohtade arv on erinevate autorite järgi erinev: 5 (Hernandez *et al*. 2006; Pearson *et al*. 2007), 10 (Stockwell ja Peterson 2002), 15 (Papeş ja Gaubert 2007), 18-20 (Mateo *et al*. 2010b), 40 (Drake *et al*. 2006), üle 30 (Wisz *et al*. 2008), 50 kuni 75 (Kadmon *et al*. 2003), 300 (Cumming 2000). Coudun ja Gégout (2006) väidavad, et aktsepteeritava täpsusega logistilise mudeli jaoks peaks olema vähemalt 50 leiukohta, 95% leiukohtade tuvastamiseks peaks olema 500–1000 vaatlust. Hirzel *et al*. (2006) leidsid, et alla 50 esinemiskoha puhul vaid esinemisandmeid kasutava mudeli prognoosivõime langeb, kuid puudumisandmeid kasutav mudel säilitab prognoosivõime. Mateo *et al*. (2010b) leidsid, et liikide leviku modelleerimisel vajalike esinemiskohtade arv sõltub andmete kvaliteedist, uuritava ala heterogeensusest ja liigi nõudlustest. Kitsa ökoloogilise nõudlusega liikide puhul piisab vähemast, ubikvistide leviku modelleerimiseks on tarvis suuremat õpetusandmestikku. Suurem õpetusandmestik võimaldab üldjuhul luua täpsema mudeli, kuid kätkeb ka liigdetailse ülesobitunud mudeli loomise ohtu (Verbyla ja Litvaitis 1989).

Taimeliigi või muu kaardistatava objekti leidmise tõenäosus sõltub vaatlusala suurusest. Liiga väikestelt vaatlusaladelt kogutud õpetusandmed vähendavad oluliselt esinemise/puudumise mudelist saadavate prognooside täpsust (Pandit *et al*. 2010).

Olulisem kui vaatluste hulk, on siiski valimi esinduslikkus keskkonnafaktorite ja uuritava ala kogu mitmekesisuse suhtes (Albert *et al*. 2010). Mittepiisava esinduslikkusega õpetusandmetest saadakse näiliselt usaldusväärsed prognoosid, mis ei pruugi siiski kehtida väljaspool õpetusandmeid. Seetõttu tuleb liikide leviku mudeleid ikka ja alati sõltumatu andmestikuga kontrollida (Manel *et al*. 1999, Edwards *et al*. 2006). Kadmon *et al*. (2003) leidsid vastupidiselt oodatule, et klimaatiliste tunnuste kombinatsioonide ühtlasem esindatus pigem vähendas mudeli formaalset tõhusust kontrollandmetes, mitte ei suurendanud seda. Mudeli universaalsus võis seejuures siiski paraneda.

Valimi esinduslikkuse olulisust näitavad ka uuringud, kus ühes piirkonnas paiknevate õpetusandmete järgi koostatud mudelit rakendatakse teises geograafilises piirkonnas. Kui kasutatud õpetusandmestik ei ole mingis olulises osas teise piirkonna suhtes esinduslik, siis ei anna mudel usaldusväärseid tulemusi. Näiteks Šveitsi andmetel koostatud liikide leviku mudelid andsid Austrias 20% väiksema täpsusega ja Austria andmetel koostatud levikumudelid Šveitsis 13% väiksema täpsusega tulemusi võrreldes kasutusega samas piirkonnas (Randin *et al*. 2006). Vähe on tähelepanu pööratud asjaolule, et ka sõltumatu kontrollandmestik peaks olema uuritava piirkonna suhtes esinduslik.

Kui liigi leviku modelleerimisel kasutatavad andmed kajastavad vaid ökoloogilise amplituudi keskosa, siis modelleerimise tulemuste ekstrapoleerimine ei õnnestu (Arntzen 2006). Kuna haruldaste ja kitsa nõudlusega liikide puhul on liigi puudumiskohad palju mitmekesisemad kui liigi esinemiskohad, siis võiks hinnangu arvutamise aluseks olevate vaatluste hulgas puudumiskohti

esinemiskohtadest rohkem olla (Remm *et al.* 2009). Esinemise ja puudumise esindatuse mahu võrdsustamiseks, säilitades seejuures varieeruvama variandi muutlikkuse piisavat esindatust, saab kasutada kaalumist alltoodud valemiga (Gibson *et al.* 2007, Platts *et al*. 2008):

$$W_{-} = \frac{P}{A}, \qquad [5\text{-}59]$$

kus W_{-} on puudumiskoha kaal, P on esinemiskohtade arv, A on puudumiskohtade arv.

Nii õpetusandmestik kui ka kontrollandmestik peaksid olema uuritava ala, uuritava objekti ja uuritava probleemi suhtes esinduslikud

Üsna tüüpiliselt selgub mudeli valideerimisel, et formaalse koondnäitaja järgi toimib mudel suurepäraselt, kuid liigi esinemiskohti suudetakse siiski üsna ebakindlalt tuvastada. Näiteks Aitken *et al.* (2007) töös saadi üldiseks kontrolltäpsuseks >96%, esinemiskohtade tuvastamise täpsuseks siiski vaid 60%. Sageli põhjustab nii suurt erinevust mudeli sobivuse näitajates liigi puudumiskohtade täiesti õige tuvastamine uuritud ala suuremas osas. Kui suurema osa uurimisalast moodustavad elupaigad, kus haruldane liik kunagi ei esine, siis õnnestub liigi puudumist selles piirkonnas väga kindlalt modelleerida. Näiteks ei ole raske saavutada vaid metsas eluneva liigi tõeseid esinemise või puudumise hinnanguid põllul ja meres. Sellise liigi hinnanguline levikukaart vähese metsaga piirkonnas võib olla peaaegu kogu kaardi ulatuses täiesti täpne. Liigi puudumine on täiesti õigesti prognoositud pea kõikjal. Formaalselt on kõik õige, aga sellised leviku mudelid ja kaardid ei lisa kaugeltki mitte sel määral uut teavet, kui võiks kõrge üldvastavuse järgi arvata. Näiteks Luoto *et al.* (2001) suutsid mustlaik-apollo levikut kaardistada 93,6% täpsusega suuresti tänu sellele, et umbes pool uurimisalast oli põllumaa, mis ei ole selle liblika elupaik.

Ebasobiva ala eemaldamist liikide levikumudelite ja elupaigasobivuse mudelite õpetusandmete hulgast on kasutatud (näiteks Titeux *et al*. 2007), kuid see ei ole üldlevinud tava. Ometi peaks modelleerimise esmane eesmärk olema hinnangute saamine olukorras, kus prognoositav muutuja ei ole ette teada. Olemasolevat teadmist kinnitaval mudelil on vähe mõtet.

Juba teadaoleva puudumise või esinemise ala suur osa tagab prognoosikaardi kõrge koondvastavuse, kuid ei lisa samaväärselt uut teavet

Liikide leviku modelleerimise aluseks oleva valimi moodustamisel eelistatakse enamasti mingi keskkonnafaktori või geograafilise piirkonna järgi stratifitseerimist ehk eelklassifitseerimist. Õpetusvalimi stratifitseerimine võib tagada usaldusväärsemad prognoosid (Hirzel ja Guisan 2002), aga ei pruugi seda teha (Reese *et al.* 2005). Eelklassifitseerimine asukoha järgi muudab vaatlusvõrgu korrapärasemaks. Korrapärase vaatlusvõrgu eeliseid juhusliku ja subjektiivse vaatluskohtade valiku ees on näidanud Hirzel ja Guisan (2002) ning Edwards *et al*. (2006). Korrapärase vaatlusvõrgu puudus on korrapära valiku subjektiivsus ja võimalik resonants mõne mudelisse kaasatud tunnusega. Peale selle on ruumilise autokorrelatsiooni arvestamiseks tarvis nii lähestikku kui ka sõredalt paiknevaid vaatluskohti.

Liikide leviku andmebaasides ja levikukaartidel olevatel leiuandmetel on kolm olulist puudust. Esiteks, need kannatavad ebaühtlase vaatluste tiheduse all, mis ei mõjuta järeldusi ainult üksikliikide levila kohta, vaid ka liigilise mitmekesisuse hinnanguid (Dennis *et al*. 1999, 2002, Dennis ja Shreeve 2003, Reese *et al*. 2005). Teiseks, liikide esinemiskohad on enamikus andmebaasides suuremõõt-kavalise kaardistamise jaoks ebapiisava asukohatäpsusega. Kolmandaks, liikide puudumisandmeid

(vaatluste aega ja kohta, kui liiki ei leitud) ei ole enamasti üldse registreeritud (Hirzel *et al*. 2002, Ottaviani *et al*. 2004).

Liikide leviku registrid ei kajasta registreeritud puudumisandmeid, leiukohad ei ole täpse asukohaga ja vaatlemisintensiivsus on ruumiliselt väga erinev

Mõne liigi esinemist/puudumist õnnestub paremini prognoosida, mõnda vähem ja seda erinevust on raske kõrvaldada. Üldiselt õnnestub paremini modelleerida tavaliste, silmatorkavate ja samas kitsa elupaiganõudlusega liikide esinemiskohti heterogeenses maastikus. Vastupidisel juhul – kui liigi isendid on raskesti leitavad, väikesed, ubikvistid ning uuritav ala on ühetaoline – vaatlustele hästi vastavaid prognoose ei saada (Hirzel *et al*. 2001, Kéry 2002, Seoane *et al*. 2005). Ei aita ei andmete suur hulk ega vaatluste hea planeering. Ka suure vaatluste hulga puhul jääb harulduste leiuandmeid nende liikide leviku seaduspärasuste modelleerimiseks väheks. Võimalikud lahendused on kaasata eksperthinnanguid või prognoosida haruldaste liikide levikut neile tüüpilise koosluse leviku järgi.

Uurimused

Vaatlusjaamade võrgu optimeerimine on aktuaalne meteoroloogias ja klimatoloogias. Amorim *et al.* (2011) uurisid Portugali ilmavaatlusjaamade paiknemist aasta keskmise temperatuuri määramise jaoks. Vaatlusjaamu jäeti ükshaaval välja ja lisati juurde. Hinnangu ebakindlus iga variandi puhul kaardistati indikaatorkriging meetodi abil.

Remm ja Remm (2009) läbisid mitme suve jooksul 300 km^2 suurusel alal käpaliste esinemist vaadeldes jalgsi 1161 kilomeetrit. Kui arvestada rohurinde vaatlustrassi keskmiseks toimivaks laiuseks 4 m, siis moodustab vaadeldud ala siiski vaid 1,46% uuritud alast ja uuritud ala katab vaid 0,6% Eesti territooriumist. Seega on kogu maa oma silmaga üle vaatamine üsna lootusetu isegi Eesti-suguses väikeriigis.

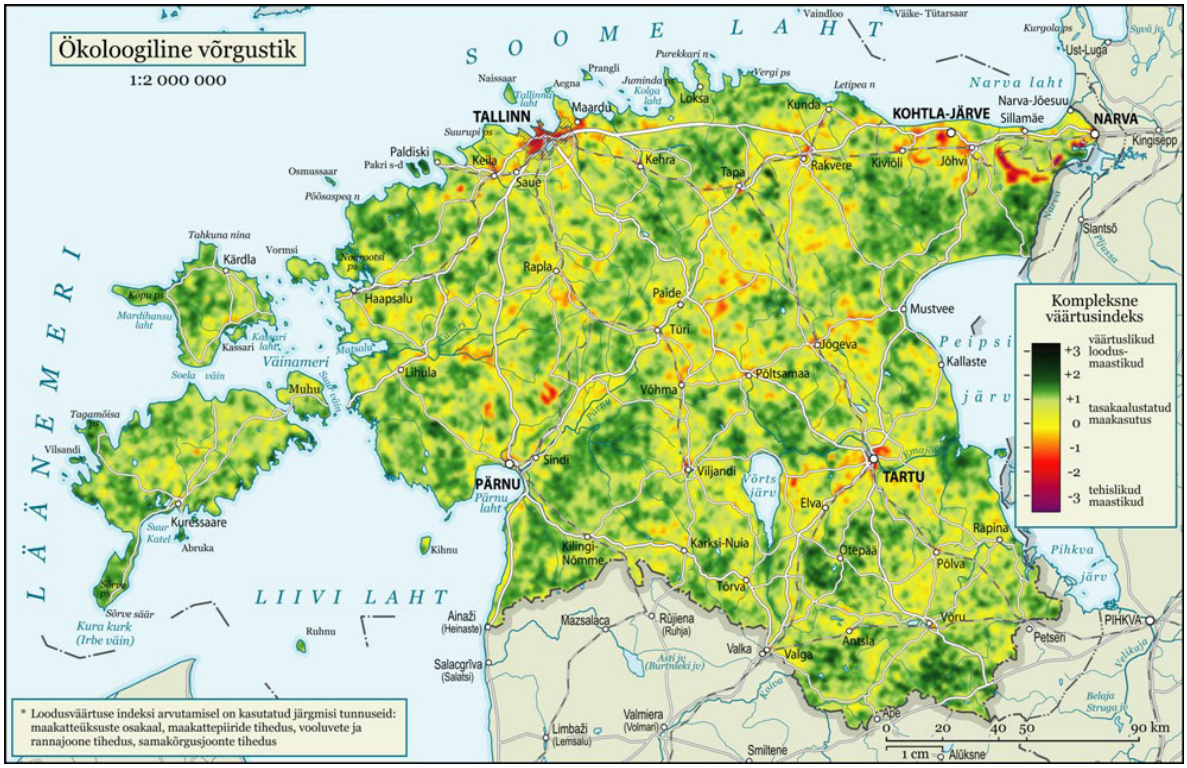
Mateo *et al.* (2010b) soovitasid väheste leiuandmete korral moodustada juhuslikud tagasipanekuga valimid, genereerida mudel ja arvutada prognoos iga valimi alusel eraldi ning leida siis eraldiarvutatud prognooside mood ehk konsensus igas prognoositavas kohas.

Graham *et al.* (2008) katsetasid liikide leiukohtade asukoha ebatäpsuse mõju liikide leviku modelleerimise tulemustele. Enamikel juhtudel leiukohtade juhuslik nihutamine nihke standard-hälbega 5 km vähendas mudeli edukust, kuid mitte suurem määral.

5.10.2. Kaardistatava nähtuse subjektiivsus ja äratuntavus

Kaardistaja otsusest sõltub nii kaardistatavate üksuste määratluse tõlgendamine, looduses äratundmine kui ka õpetusaladena piiritlemine. Tulemuse subjektiivne sõltuvus üksikisiku otsustest on üldine igasugusel kaardistamisel, kus otsustamisel osaleb inimene. Eriti terav on probleem, kui eristatavad üksused võivad üksteiseks sujuvalt teiseneda ning nende vahel ei ole selgeid ruumilisi piire, nagu see on taimekoosluste või detailsete maakatteüksuste eristamisel suhteliselt looduslikus maastikus.

Subjektiivsel faktoril on eriti suur osa nähtuste ja väärtushinnangute kaardistamisel, mida ei ole seotud range määratlusega, näiteks elupaigasobivus ja sobivus ökovõrgustiku jaoks (joonis 5-25) (Remm ja Mander 2001). Kuidas hägus, otseselt mittemõõdetav nähtus olemasolevate andmekihtidega siduda, otsustab ekspert ning igal eksperdil võib olla teistest erinev arvamus. Elupaigasobivuse hinnangu subjektiivne mitmevariandilisus kandub edasi nendesse otsustesse, mis elupigasobivuse kaardi alusel tehakse.



Joonis 5-25. Sobivus üleeuroopalise ökovõrgustiku jaoks (Aunap *et al.* 2007).

Mudeli koostamisel kasutatud õpetusandmete esinduslikkus sõltub muuhulgas kaardistatava nähtuse looduses tuvastamise tõenäosusest. Näiteks on peaaegu kõiki loomi ja linde, kuid ka kõiki muid objekte, kergem märgata lagedal kui metsas. Kui vaatlustrass läbib nii metsa kui lagedat, võivad metsas olevad esinemiskohad saada alaesindatud. Nende andmete alusel antud hinnangud on siis kindlasuunalise hälbega. Kaardistatava objekti tuvastamise tõenäosust ning selle kaudu õpetusandmete usaldusväärsust ja mudeli täpsust mõjutab ka objekti suurus. Näiteks on leitud, et suurema kehaga liblikate levikumudelid on täpsemad kui väiksematel (Newbold *et al.* 2010).

Kaheldud on kaugseirepõhise maakattekaardistuse usaldusväärsuses ja on püütud leida selleks paremaid lähteandmeid ja meetodeid. Sellegipoolest ei ole märgata paranemistendentsi publitseeritud kaugseirekaardistuste väliandmetele vastavuses, mis on kapa kordaja järgi keskmiselt 0,66 (Wilkinson 2005). Kaugseireandmeid kasutava maakattekaardistuse puhul on eesmärgiks saavutada vähemalt 85% vastavus kapa kordaja järgi tõeseks peetava kontrollvalimiga (Wulder *et al*. 2006). Nii kõrge kokkulangevus võib olla saavutatav eesmärk kultuurmaastikus või muude hästieristuvate klassidega selgepiiriliste kõlvikute puhul. Sujuvate üleminekutega loodusmaastikul ning teadusuuringutes, kus eesmärgiks on midagi uut avastada või leiutada, nii head vastavust reeglina ei saavutata.

Briti uurijad teevad välikaardistuste vähesest kokkulangevusest järelduse, et sealkasutatavad taimkatte kaardistamise üksuste klassifikatsioonid ei sobi välikaardistuseks (Cherrill ja McClean 1999a, Hearn *et al.* 2011). Ka Eestis läbi viidud uuringu kohaselt on detailsete kaardistusüksuste kasutamisel välikaardistajad enamiku kohtade klassifitseerimisel eriarvamusel. Probleem on ilmselt üldisem. Looduslike alade taimkate on suurel määral kontiinum, mida ei ole võimalik nii edukalt selgepiirilistesse etteantud kategooriatesse jagada kui haritava maa kõlvikuid. Rohkem kui kümne eristatava taimkatteüksuse korral loodusmaastikus on lootusetu iseseisvalt mõtlevate inimeste vahel üksmeelele jõuda. Kümnekonna üksusega kaarti võivad kõik õigeks pidada, aga see on pigem

maakattekaart, mitte taimestikukaart.

Takistuseks kaugseirepõhiste kaartide ja välivaatluste suurema vastavuse saavutamisel on eelpool mainitud klassifitseeritava nähtuse enda hägusus ja klassifitseerimise subjektiivsus ja kontrollvalimi ebapiisav esinduslikkus, kuid ka kapa kordaja järgi vastavuse mõõtmise algelisus (Foody 2008). Kapa ei arvesta kolme asjaolu:

- osa piksleid kaugseirekujutises on paratamatult segapikslid, mis ei saa ühtegi etteantud klassi täpselt esindada;
- nii kaugseirekujutises kui ka maapealsetes andmetes võib olla asukohavigu ja moonutusi;
- maapealsetes andmetes võib olla klassifitseerimisvigu ja subjektiivsust, eriti juhul, kui eristatavaid klasse on palju.

Isegi sama nimega klassid võivad olla erinevates andmestikes erinevalt määratletud. Suur osa kaartide mittevastavusest võib olla tingitud erinevalt piiritletud eraldistest. Näiteks Suurbritannia maakattekaardi vastavus kontrollandmetele suurenes eraldiste piirivööndite arvestusest väljajätmise järel 46 protsendilt 71 protsendini (Fuller *et al.* 1994).

Kaugseire-kaardistuse puhul lisatakse kaardile vähemalt selle vastavuse hinnang (kapa või kokkulangevus kontrollandmetega). Muude kaardistuste puhul ei ole klassifitseerimise vigade ja subjektiivsuse määra enamasti üldse näidatud. Ometi eeldatakse teiste kaartide põhjal kaugseire-kaardistuse tulemuste hindamisel kaardistusvigade puudumist.

Kaugseire-põhiste kaartide usaldusväärsuse hindamisel tõeseks loetavad kontrollandmed võivad sisaldada vigu ja subjektiivseid valikuid

Uurimused

Suurbritannias on tehtud katseid, kus sama ala taimkatet kaardistasid sama juhendi järgi teistest sõltumatult mitu välitöötajat. Cherrill ja McClean (1995) võrdluse aastase vahega tehtud kaartidel oli taimkate identselt klassifitseeritud 44,4% ulatuses. 55,6% erinevusest sai maastiku muutumisele omistada maksimaalselt 14,4%. Ülejäänud 41,2% tulenes valdavalt taimestiku erinevast interpreteerimisest kaardistajate poolt.

Stevens *et al*. (2004) leidsid 76% vastavuse kaardistaja ja kontrollija määrangu vahel, kusjuures kõige väiksem oli vastavus pool-looduslike koosluste osas. Enamik erinevusi tulenes taimkatte erinevast klassifitseerimisest.

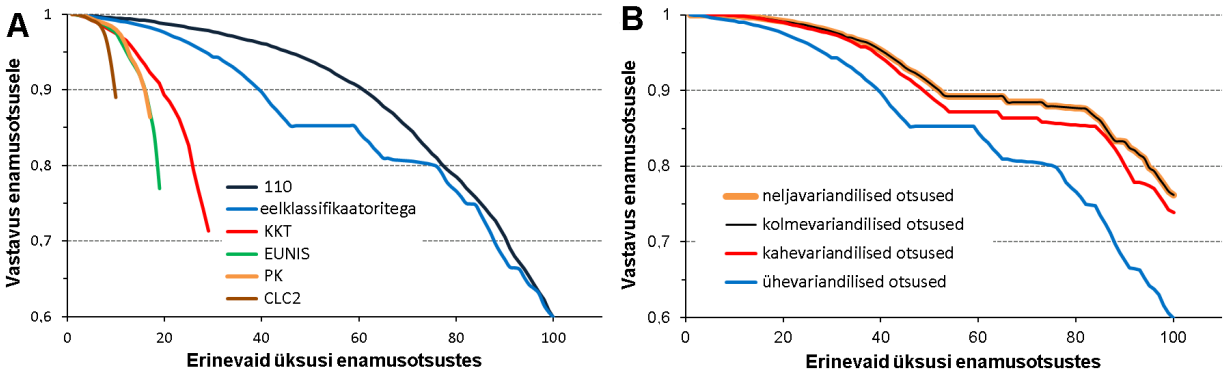
Cherrill ja McClean (1999a,b) katses oli 6 kaardistajat ja kokku 36 taimkatteüksust. Keskmine kokkulangevus kahekaupa võrreldud tulemustes oli 25,6% (17,3...38,8%). Vaid 7,9% kaardistatud ala pinnast olid kõik kuus kaardistajat üksmeelel. Sarnasus esinduskohtades kirja pandud liikide nimekirjades varieerus 18,8 ja 63,7% vahel. Sama instituudi töötajate kaardid olid omavahel sarnasemad. Mõlemas uurimuses leiti, et eraldiste piiritlemise erinevus põhjustab vaid tühise osa kaartide mittevastavusest. Valdav osa erimeelsusi tulenes ühiste liikidega taimkatteüksuste erinevast klassifitseerimisest, mitte asupaigahälvetest.

Hearn *et al.* (2011) uuringus oli 7 kaardistajat, kaardistatud alal vähemalt korra registreeritud üksuste arv oli vastavalt klassifikatsiooni detailsusele 6, 20 ja 34. Keskmine vastavus kaartide vahel vastavalt taimkatteüksuste detailsusele oli 77,6% (6 üksuse puhul), 34,2% (20 üksuse puhul) ja 18,5% (34 üksuse puhul).

Rapp *et al.* (2005) hindasid ortofoto interpreteerimise ja välikaardistuse kombinatsioonis loodud taimkattekaardi täpsust Uus-Inglismaal USAs ja said vastavuseks 46%. Vastavuse parandamiseks

soovitatakse täpsustada üksuste määratlusi, lisada segaüksusi, kasutada üldisemat klassifikatsioonitaset, kasutada aerofotol paremini eristuvaid üksusi ja detailsemaid aerofotosid ning suurendada välitöö osa.

Remm *et al.* (2010) võrdlevas uuringus osales neli kaardistajat. Kaardistatavas piirkonnas esinevate ja kaardistajatele ette antud üksuste koguarv oli 110. Üksmeele määra kaardistajate vahel hinnati üksikkaardistaja esimese valiku vastavusena enamusotsusele. Enamusotsustes esines 101 kaardistusüksust. Üksikvaatleja valik vastas enamusotsusele 60% juhtudest. Neid detailseid üksusi ühendati üldisemateks klassideks mitmel viisil. Kui üksusi liita viisil, mis tagab selles andmestikus igal liitmise järel võimalikult suure üksmeele (kuid ei arvesta klassifikatsiooni seesmist loogikat ega rakendatavust mujal), võiks 60 eristatava üksuse juures saavutada üksikkaardistaja valitud üksuse keskmiselt 0,9 tasemel vastavuse enamusotsusega (joonis 5-26A). Kui üksuste liitmisel jälgida Corine maakattekaardi teise taseme üksusi, kasvukohatüüpe, Euroopa elupaigatüüpide klassifikatsiooni (EUNIS) suuri üksusi ja Eesti põhikaardi põhialasid eelklassifikaatoritena, siis võiks selles andmestikus üheksakümneprotsendilise üksmeele saavutada umbes 40 üksuse juures. Paraku on formaalsete reeglite järgi moodustatud üksuste maht väga erinev ja klassifikatsiooni loogika nõrk.



Joonis 5-26. Kaardistajatevahelise üksmeele suurenemine kaardistatavate taimkatteüksuste arvu vähenemisel (A) ja mitmevariandiliste otsuste lubamisel (B). 110 – 110 suhteliselt detailset taimkatteüksust, KKT – kasvuklassid (29 üksust uuritaval alal), EUNIS – Euroopa elupaigatüüpide klassifikatsioon (19 üksust), PK – Eesti põhikaardi põhialad (17 üksust), CLC2 – Corine teise taseme maakatteüksused (10 üksust). Remm *et al.* (2010 muudetud).

Erimeelsuste osa jääb suureks isegi mitmevariandiliste otsuste lubamisel ja eeldusel, et kaardistaja otsus loetakse õigeks, kui ükskõik milline kaardistaja valitud variant langeb kokku enamusotsusega. Üle 50 võimaliku üksuse lubamisel ei õnnestuks üheksakümneprotsendilise üksmeeleni jõuda isegi mitmevariandilisel kaardistamisel. Neljanda variandi lubamine lisaks kolmandale ei suurenda üksmeelt nähtaval määral, sest kaardistajad kasutasid neljanda variandi ülesmärkimise võimalust vaid vähestes kohtades (joonis 5-26B).

5.10.3. Prevalents

Liikide realiseerunud leviku kaardistamisel on leitud, et laia levikuga ja elupaiga suhtes vähenõudlikke liikide levikut õnnestub modelleerida vähema täpsusega kui kitsalt spetsialiseerunud ja haruldasi liike (Brotons *et al*. 2004, Segurado ja Araújo 2004, Luoto *et al*. 2005, Franklin *et al*. 2009, Zurell *et al.* 2009). Jiménez-Valverde *et al*. (2009) leidsid konstrueeritud andmetega katsetades, et valimi mahu mõju on tugevam kui ühe variandi prevalentsi mõju, mis on oluline alles siis, kui esinemis- ja puudumiskohtade vahekord on tasakaalust väga kaugel (<0,01 või >0,99). McPherson *et*

al. (2004) näitasid, et mudeli ülesobitumise oht on 50% lähedase prevalentsi puhul palju väiksem kui esinemise/puudumise ebavõrdsuse korral.

Kitsalt või laialt spetsialiseerumine sõltub faktorist, mille suhtes spetsialiseerumist käsitletakse. Elupaiga füüsiliste omaduste suhtes vähespetsialiseerunud liik võib olla tundlik haigustekitajate ja parasiitide suhtes või määravad selle liigi levikut kaasaegsed või varasemad levikutõkked. Tundub Sellise liigi esinemistõenäosuse modelleerimisel elupaiga tunnuste alusel ei ole usaldusväärset tulemust loota.

Haruldaste liikide loodusest leidmine nõuab ulatuslikke välivaatlusi. Välitöö tulemuslikkuse tõstmiseks tuleks vaatluskohad planeerida eelkõige liigi esialgse levikumudeli järgi oodatava levila piirkonda. Täiendatud andmete järgi saab koostada uue ja täpsema levikumudeli ja seda taaskord kasutada välivaatluste asukoha planeerimisel. Välivaatluste suunatud planeerimine võib anda kuni neljakordse välitöö efektiivsuse tõusu võrreldes vaatluskohtade juhupaigutusega (Guisan *et al.* 2006).

5.10.4. Mudeli ja seletavate tunnuste valik

Ruumiliste muutujate väärtuste hindamisel võib tulemus oluliselt sõltuda valitud mudelist. On näidatud, et liikide leviku erinevad modelleerimisviisid põhjustavad tulemustes suuremaid erinevusi kui atmosfääri tsirkulatsiooni muutumise ja kasvuhoonegaaside emissiooni erinevad variandid (Diniz-Filho *et al*. 2009, Buisson *et al.* 2010). Suure hulga taimeliikide ja mudelitega tehtud katseseerias korreleerusid erinevat tüüpi mudelite abil saadud hinnangulised esinemistõenäosuse kaardid omavahel keskmiselt vaid 60% ulatuses (Syphard ja Franklin 2009). Üldiselt on esinemistõenäosuse kaardistamine vigade suhtes robustsem kui esinemise ja puudumise kaardistamine kaheväärtuselise muutujana (Heuvelink 1998).

Seletavate tunnuste ja andmekihtide valik ning kvaliteet on kasutatavate vaatlusandmete hulga, usaldatavuse ja esinduslikkuse kõrval veel üks oluline faktor, mis mõjutab mudeli prognoosimisvõimet. Andmete kvaliteedi ja mudeli valikust tulenevate hälvete eristamine on keerukas. On väidetud, et andmevigade mõju prognoosile võib olla suurem kui ebasobiva mudeli või selle keerukustaseme mõju (Berg *et al*. 2004). Kohta iseloomustavad tunnused on aina ebatäielikud ning ajas ja ruumis ebaühtlase tiheduse ja kvaliteediga ning prognoositava tunnuse jaoks olulisi faktoreid kirjeldavaid tunnuseid ei ole pahatihti üldse mõõdetud.

Liikide leviku ja selle võimaliku muutumise modelleerimisel kasutatakse liikide esinemise andmeid. Liigi esinemist või puudumist mingis kohas määrab väga palju faktoreid, sealhulgas koha ja selle ümbruse arengulugu ning liigi evolutsiooni käik, teiste liikide esinemine ja ohtrus. Seetõttu ei pruugi seosed ühelt poolt liigi kaasaegse leviku ning teisalt kaasaegsete keskkonnatingimuste ja teiste liikide mõju vahel olla tulevikku ekstrapoleeritavad.

Loodusnähtuste modelleerimisel on sageli probleemiks otseste mõjurite ja varasemal ajal mõjunud faktorite puudumine seletavate tunnuste hulgast

Nii uuritava nähtuse asukoha määramisel kui ka seletavates andmekihtides olevate asukohavigade mõju sõltub ruumilise autokorrelatsiooni ulatusest igas andmekihis. Kui lähedalolevad kohad on sarnased, siis väike asukohaviga tunnuse väärtust oluliselt ei muuda (Osborne ja Leitão 2009). Mudeli parameetrid on lähteandmete asukohavigade suhtes tundlikumad kui mudelist saadud hinnangud ning parameetrite otsese tõlgendamisega (näiteks seletavate tunnuste mõju tugevuse üle otsustamine regressioonikordajate järgi) tuleb olla ettevaatlik (Johnson ja Gillingham 2008).

Enamikule liikide ja elupaikade leviku hinnangulistele kaartidele ei ole lisatud prognoositäpsuse

hinnangut ega tõenäolise vea suuruse paiknemise andmekihti, kuigi see on vajalik tulemuste usutavuse hindamiseks (Elith *et al*. 2002, Barry ja Elith 2006). Kui mudeli vead ei ole ruumis ühtlaselt jaotunud, näitab see mingi olulise faktori puudumist mudelist ja annab vihjeid mudeli parandamiseks. Liikide, mille kohta on vähe esinemise/puudumise vaatlusi, saab leviku modelleerimisel ära kasutada teiste, sealhulgas kaasnevate liikide esinemisandmeid (Lehmann *et al*. 2002a). Teiste liikide prognoositud või vaadeldud esinemine võib olla kas seletavaks tunnuseks vähe esindatud liigi esinemisootuse hindamisel või siis prognoositakse mitme liigi esinemist korraga, ühe mudeliga.

Uurimused

Faktorite multikollineaarsuse mõju vähendamiseks lülitasid Kok ja Veldkamp (2001) maakasutuse mudelisse omavahel enam kui 0,5 tasemel korreleerunud faktoritest vaid ühe.

Glenn ja Ripple (2004) näitasid, kuivõrd sõltub taimkattekaart kaugseireandmete saamise ja teisendamise üksikasjadest, mis pahatihti ei ole publikatsioonide puhul korralikult kirja pandud.

Fleming *et al*. (2004) soovitasid kontrollida oma analüüside tulemuste tundlikkust maakatteandmete vigade suhtes. Nad kasutasid juhuslike vigade lisamist maakatteandmetele ja vaatasid seejärel kuivõrd muutub seejärel valgesaba-hirve (*Odocoileus virginianus*) elupaigasobivuse indeksi kaart.

Van Cauter *et al*. (2005) võrdlesid suurimetajate arvukuse hinnangute sõltuvust taimkattepiiride kaardistuse täpsusest Lõuna-Aafrikas.

Van Horssen *et al*. (2002) näitasid, et ruumiline prognoos logistilisest mudelist võib olulisel määral sõltuda uuritava tunnuse lähteandmete ebamäärasusest ja regressioonikordajate ehk seletavate tunnuste määramise vigadest. Kaardil on võimalik kujutada nende kahe vigadeallika vahekorda uuritava ala erinevates osades.

Lees *et al*. (2008) leidsid, et maapinna kõrguse teisendamisel täiendavateks tunnusteks, nagu asend nõlval, nõlva suund ja kaldenurk ei ole mõtet, sest mägisel alal on koha absoluutne kõrgus põhiline. Kuni 10% vigade lisamine kõrgusmudelisse taimkatte hinnangulist kaarti oluliselt ei mõjutanud, kuid mõju erinevate klasside eristamisse on erinev.

Liblikate leviku muutumise hinnangut kliima muutumise korral Zurell *et al*. (2009) uuringus mõjutas tugevasti parasitoidi kaasamine dünaamilisse levikumudelisse.

5.10.5. Hinnangute ebakindluse kaardistamine

Mudeli **ebakindlus** väljendub mudeli abil saadud hinnangute hajuvust ja usaldusväärsust. **Otsusekindlus** on ebakindluse vastand. Otsusekindlus ehk mudeli ebakindlus ei ole sama, mis mudeli **tundlikkus** (*sensitivity*). Otsusekindlus kirjeldab mudelist saadud hinnangute varieeruvust, tundlikkus mudeli erinevat mõjutatavust erinevate lähteandmete poolt. Mudeli tundlikkuse analüüs selgitab, mil määral mõjutavad hälbed mudeli sisendites mudeli väljundit, millised hälbed mõjutavad tulemust rohkem, millised vähem. Ruumiliste mudelite tundlikkuse analüüsi meetoditest annavad lühiülevaate Lilburne ja Tarantola (2008). Põhjalikum käsitlus on raamatutes (Saltelli *et al*. 2004, 2008). Frysinger (2002) jagab keskkonnamudelite ebamäärasuste põhjused nelja liiki:

- andmete vähesus,
- vead andmetes,
- mudeli sobimatus,
- modelleeritavate nähtuste tõenäosuslik loomus.

Näiteks samakõrgusjoonte alusel moodustatud kõrgusmudel sisaldab kasutatud samakõrgusjoonte ebatäpsusi, samakõrgusjoonte digitaliseerimise vigu, interpoleerimisel tekkiva üldistuse ebatäpsust ning maapinna kõrguse üldistamisest tulenevat ebatäpsust. Kõrgusmudelid ei peagi kujutama igat üksikut mätast.

Mudeli ebakindlus põhjusteb prognooside määramatust, mudeli tundlikkus sisendfaktorite erinevat panust ebakindlusesse

Ebakindluse kaart (*uncertainty map*) ja **otsusekindluse kaart** (*confidence map*) näitavad tulemuste hajuvust, määramatust, mudelist saadud prognooside varieeruvust või prognoosi aluseks olnud lokaalsete andmete hulka ja usaldusväärsust. Näidiste järgi klassifitseerimise (ptk 5.6.8) otsusekindluse kaart eristab piirkondi, mille klassifitseerimiseks on õpetusandmetes piisavalt eeskujusid ning alasid, mille varieeruvust õpetuskogum korralikult ei esinda. Sarnasuse alusel arvutatud prognoosi usaldusväärsust näitab otsuse langetamiseks kasutatud näidiskohtade sarnasus prognoositava kohaga. Ansamblimeetodite puhul saab otsusekindluse prognoositud kategooria suhtelisest sagedusest ansambli liikmete hinnangutes. Võimalike ebatäpsuste paiknemise kaart võimaldab vähendada välitöökulusid, kuna maastikus tuleb esmajärjekorras üle vaadata vaid ebakindlalt otsustatud kohad.

Otsusekindluse kaart mitte ainult ei too esile väheesindatud piirkonnad, vaid võib osutada ka etteantud klassifikatsiooni puudulikkusele, see tähendab osutada mingite omanäoliste klasside puudumisele eristatavate kategooriate hulgas või õpetusandmetes. Etteantud klassifikatsioon võib olla ebatäiuslik või ebaesinduslik ka klassidevaheliste üleminekute ja segaklasside esinemise poolest. Hinnangu määramatus võib olla tingitud ka seletavate tunnuste puudulikkusest – mõne kategooria eristamiseks vajalik tunnus on puudu.

Otsusekindluse kaart näitab alaesindatud uurimisala piirkondi ja tunnuste väärtusvahemikke

Mudeli ebakindluse mõõtmiseks tuleb lisada juhuslikke hälbeid lähteandmetesse või hinnata lähteandmete usaldusväärsust. Mudeli tundlikkuse määramiseks muudetakse mudeli parameetreid ja sisendfaktoreid ükshaaval või plaanipärastes kombinatsioonides.

Uurimused

Journel (1996) kasutas ebamäärasuse modelleerimiseks geostatistika vahendeid.

Goovaerts (2001) eristab lokaalset määramatust ühe vaatluskoha ümber ja ruumilist (mitme punkti) määramatust. Ta võrdleb määramatuse kaardistamist kriging meetodi abil ja stohhastilise jäljendusmudeli abil. Kriging kui ühekordne arvutus on vähem töömahukas, aga jäljendusmudelil on mitmeid eeliseid: esiteks, see annab määramatuse mudeli, teiseks, võimaldab erineval üldistustasemel väljundeid, kolmandaks, võimaldab uurida määramatuse edasikannet andmete teisendamise käigus.

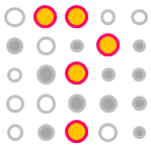
Stum *et al*. (2010) lisasid juhumetsa meetodil saadud maakattekaardile prognoositud klassi sageduse iga üksiku otsusepuu abil saadud hinnangute hulgas. 7,7% kaardistatud ala pikslitel oli tõenäosus ≤0,20 kuuluda mõnda etteantud klassi. Mudeli ebakindluse põhjusena loetlevad mainitud autorid järgmist:

- etteantud klassifikatsiooni puudulikkus,
- klassidevahelisi üleminekud looduses,

- mitme klassi koosesinemine,
- seletavate tunnuste puudulikkus,
- õpetusvaatluste puudulikkus.

Tarkvarasüsteem Constud (ptk 3.4.6.1 ja 5.6.8.3) arvutab koos sarnasusele tugineva hinnanguga ka hinnangu otsusekindluse, mis on hinnangu arvutamiseks valitud näidiste keskmine sarnasus prognoositava koha või objektiga. Otsusekindlus saadakse nii hinnangulise kaardi loomisel kui ka prognooside arvutamisel tabelisse (Remm *et al.* 2011, Remm ja Kelviste 2011a, b, c, Remm ja Linder 2007, Remm ja Remm 2008, 2009). Constud süsteemis arvutatud Baltimaade sademete hinnangu ebakindluse kaart osutab kõige suuremale teadmatusele Peipsi kohal sadavate sademete hulga osas (joonis 5-10), sest ükski Baltimaade ilmavaatlusjaam ei asu suure siseveekogu keskosas.

Diniz-Filho *et al.* (2009, 2010) jagavad ansamblimeetodil saadud hinnangute hajuvuse peakomponentideks, et eristada ja kaardistada hinnangute hajuvuse allikaid.



Küsimused

1. Mis muutujad on autokorrelogrammi telgedel?

2. Mis vahe on ristkorrelogrammil, korrelogrammil ja autokorrelogrammil?

3. Kas rohkem varieeruvad suurema või väiksema aknaga silutud andmed?

4. Kas silumine kauguse pöördväärtusega võib anda interpoleeritud väärtuse, mis on väljaspool originaalandmete haaret?

5. Kumb annab siledama pinna, kas silumine kauguse pöördväärtuse proportsionaalsete kaaludega või kauguse pöördväärtuse ruuduga proportsionaalsete kaaludega?

6. Too mõni negatiivse ruumilise autokorrelatsiooni näide.

7. Kas ruumiline autokorrelatsioon on alati seotud kindla vahemaaga?

8. Mis mõttes on näivkordus näiv?

9. Mille suhtes hälbed kasutatakse Morani *I* arvutamisel ja mille suhtes Pearsoni korrelatsioonikordaja arvutamisel?

10. Kas positiivne ruumiline autokorrelatsioon suurendab esimest või teist liiki statistilise vea tõenäosust? Põhjenda.

11. Millise eelduse kehtimisel on Morani *I* võimalike väärtuste vahemik –1…+1?

12. Mis vahe on lokaalsel autokorrelatsioonil ja autokorrelatsioonil naabervaatluste vahel?

13. Kui suur on kogu uuritaval alal konstantse väärtusega muutuja ruumiline autokorrelatsioon Morani *I* järgi?

14. Millega põhjendaksid kahega jagamist semidispersiooni arvutamisel ehk ruumilise varieeruvuse mõõtmist poolhajuvusena, mitte ruumilise hajuvusena?

15. Mis mõttes on risttabel ja ristkorrelogramm risti?

16. Kas kriging sobib liigi esinemise või puudumise interpoleerimiseks vaatluskohtade vahele?

17. Mille poolest on sarnased näidistele tuginev järeldamine ja mitme punkti geostatistika?

18. Mille poollest erineb ruumiline korrelatsioon klassikalisest mitteruumilisest korrelatsioonist?

19. Kuidas seletada negatiivset ruumilist korrelatsiooni oktoobrikuu ja novembrikuu paljuaastase keskmise sademete hulga vahel üle 250 km vahemaaga Baltimaade ilmavaatlusjaamades?

20. Miks on indikaator-ümbrus üldjuhul lähimast naabrusest mõnevõrra kaugemal?

21. Mille poolest erineb ruumiline regressioon ruumilisest korrelatsioonist?

22. Miks ei sobi BIOCLIM mudel põõsasmarana leviku modelleerimiseks Loode-Eestis?

23. Mis on liikide leviku ükshaaval kaardistamise ja kooslustekaupa kaardistamise peamine puudus?

24. Millega põhjendada mediaanist sama palju või rohkem erineva väärtusega leiukohtade sageduste summeerimist vaid ühel pool leiukohtade sagedusjaotuse mediaani, nagu see toimub tarkvaras Biomapper?

25. Mis on sarnasusele tugineva järeldamise peamised eelised ja puudused liikide leviku kaardistamisel?

26. Miks tegeletakse liikide leviku modelleerimisel aktiivselt meetoditega, mis tuginevad vaid leiuandmetele?

28. Kas metsakorralduses kasutatava *kNN* meetodi puhul kasutatakse geograafilises ruumis lähedasi või kaugseireandmete järgi sarnaseid näidisalasid?

29. Nimeta puistu takseertunnuseid, mida on kaugseireandmete järgi kaardistatud.

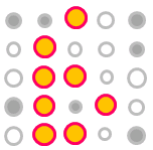
30. Miks mudelite prognoosiv võime reeglina ei parane, kui mudelisse lülitada ülemäära palju argumenttunnuseid?

31. Mille suhtes esinduslikkust peaks arvestama liikide leviku modelleerimise aluseks olevate andmete valimisel?

32. Milleks on kasulikud ruumiliste prognooside ebakindluse kaardid?

33. Mis on toimimiskõvera telgedel haruldase taimeliigi esinemise ja puudumise kaardistamistäpsuse hindamisel?

6. Paiknemismustri loomine



Maastiku mõõtkavas nähtuste paiknemismustreid luuakse mitmetel eesmärkidel. Neist olulisemad on:

- metoodika uuringud proovivõtuskeemide, mudelite, analüüsimeetodite ja indeksite testimiseks;
- nullhüpoteesile vastavate paiknemismustrite moodustamine hüpoteeside testimisel;
- ruumiliselt ilmutatud mudelite optimeerimine ja nende käitumise läbimängimine;
- andmete interpoleerimine ja ekstrapoleerimine;
- bioloogiliste invasioonide modelleerimine;
- ajalooliste maastikumustrite rekonstruktsioon mittetäielike andmete järgi;
- võimalike maastiku arengute ja maastikumustrite prognoos;
- ilmaprognoos ja muude loodusnähtuste (ka loodusõnnetuste) leviku prognoos;
- inimtegevuse (ka sõjaliste konfliktide) ajalis-ruumilise dünaamika prognoos;
- arvutimängud ja õppesimulaatorid.

Lihtsamad paiknemismustrit loovad algoritmid kasutavad **juhuslikku jäljendust** (*stochastic simulation*), mille käigus luuakse alternatiivseid samatõenäolisi juhuslikke punktide paiknemismustreid või rasterpindu. Keerukamad ruumilised jäljendused on aegruumilised, mitmefaktorilised ja objekt-orienteeritud. Objekt-orienteeritud mudel käsitleb objekte individuaalsetena, objektid mõjutavad üksteist ja oma ümbrust, objektide käitumine sõltub ümbrusest, objektide käitumine võib olla nii eesmärgipärane kui ka spontaanne.

Õppesimulaatoritena võiks mainida virtuaalse linna ja virtuaalsete elanike konstrueerimise vahendeid SimCity (http://www.simcity.com/) ja SIMS (http://thesims.ea.com/us) ja 3D visualiseerimisvahendite komplekti *World Constriction Set – WCS* (http://www.3dnature.com/index.html). Ilmamudelitest osaleb Eesti Meteoroloogia ja Hüdroloogia Instituut keskmisemõõtkavalise ilmamudeli HIRLAM (*HIgh Resolution Limited Area Model*) konsortsiumis (http://hirlam.org).

6.1. Punktmustri loomine

Punktmustri loomisel eeldatakse, et punktobjektide tihedus ja tiheduse paiknemismuster on tingitud mingist mustrit kujundavast protsessist. Stohhastilised punkte genereerivad protsessid (*stochastic point processes*) loovad mustrit mingi reegli järgi, mis kasutab juhuslikke arve. Juhuslike arvude saamise võimalusi on mainitud valikumeetodite peatükis (ptk 1.1.2).

Punktmustri loomise lähteparameetrid jagunevad nelja klassi:

- mustri oodatav tihedus ehk protsessi intensiivsus;
- objektide esinemise tõenäosus pinna eri osades (sobivuspind, mis võib olla kas pidev või diskreetne);
- punktide tiheduse ruumiline autokorrelatsioon (objektide koondumine või korrapära, klastrite paiknemise või suuruse või kuju korrapära);
- punktide kadumise intensiivsus vastavalt asukohale ja naabrite paiknemisele.

Loodava punktmustri oodatavat tihedust on üldiselt lihtne otsustada. Iga koha sobivuse saab arvutada iga koha tunnuste põhjal pinna logistilisest sobivuse mudelist. Punktobjektide tiheduse ruumilist autokorrelatsiooni saab arvestada juhuslikult paigutatud punkti alles jättes, eemaldades vastavalt mingile kontrollstatistikule või leides uuele punktile sobivaimat kohta vastavalt olemasolevate naabrite paiknemisele (Gibbsi sampler, naabritiheduse jaotus).

Punktmustrite genereerimise algoritme on jagatud neljaks (Upton ja Fingleton 1985):

- lihtne ehk ühtlane ehk homogeenne juhuslik protsess ehk Poissoni mets (*homogeneous Poisson process*),
- ebaühtlased ehk mittehomogeensed juhuslikud protsessid (*nonhomogeneous Poisson process*),
- võreprotsessid (*lattice-based process*),
- takistusega ehk inhibitsiooniga protsessid (*inhibition process*).

Märgitud protsessis (*marked process*) omistatakse objektidele klassikuuluvus või muud omadused punkti genereerimise hetkel või enne seda. **Märkimata protsessi** (*unmarked process*) käigus tekkivatele punktobjektidele kohe atribuute ei omistata, küll aga võib atribuudid omistada mustri moodustamise järel.

Ülevaate punktprotsessidest võib leida Diggle (1979b, 1983, 2003), Stoyan ja Penttinen (2000) ning Comas ja Mateu (2007) töödest.

6.1.1. Homogeenne juhuslik protsess

Homogeenne juhuslik protsess on ühetaoline lihtprotsess, mis moodustab täieliku juhuslikkuse hüpoteesile vastavaid mustreid. Homogeensel juhuslikul protsessil on kolm iseloomulikku omadust:

- protsessi intensiivsus (λ) on ajas ja ruumis konstantne;
- ükskõik millise suurusega (A) proovialadel leiduvate punktsündmuste arv on Poissoni jaotusega, mille keskmine võrdub $A \cdot \lambda$;
- punktsündmused on omavahel sõltumatud.

Homogeense juhusliku protsessi ainus parameeter on protsessi intensiivsus

Homogeense juhusliku protsessi erijuht on nn Poissoni mets (*Poisson forest process*) ehk juhuslik punktprotsess kahemõõtmelisel pinnal (ptk 4.1.1). Gelfand *et al.* (2010) eristavad Poissoni protsessi ja **täielikku ruumilist juhuslikkust** (*complete spatial randomness*). Selle liigituse järgi on punktsündmuste omavaheline sõltumatus täieliku ruumilise juhuslikkuse, aga mitte Poissoni protsessi eeldus.

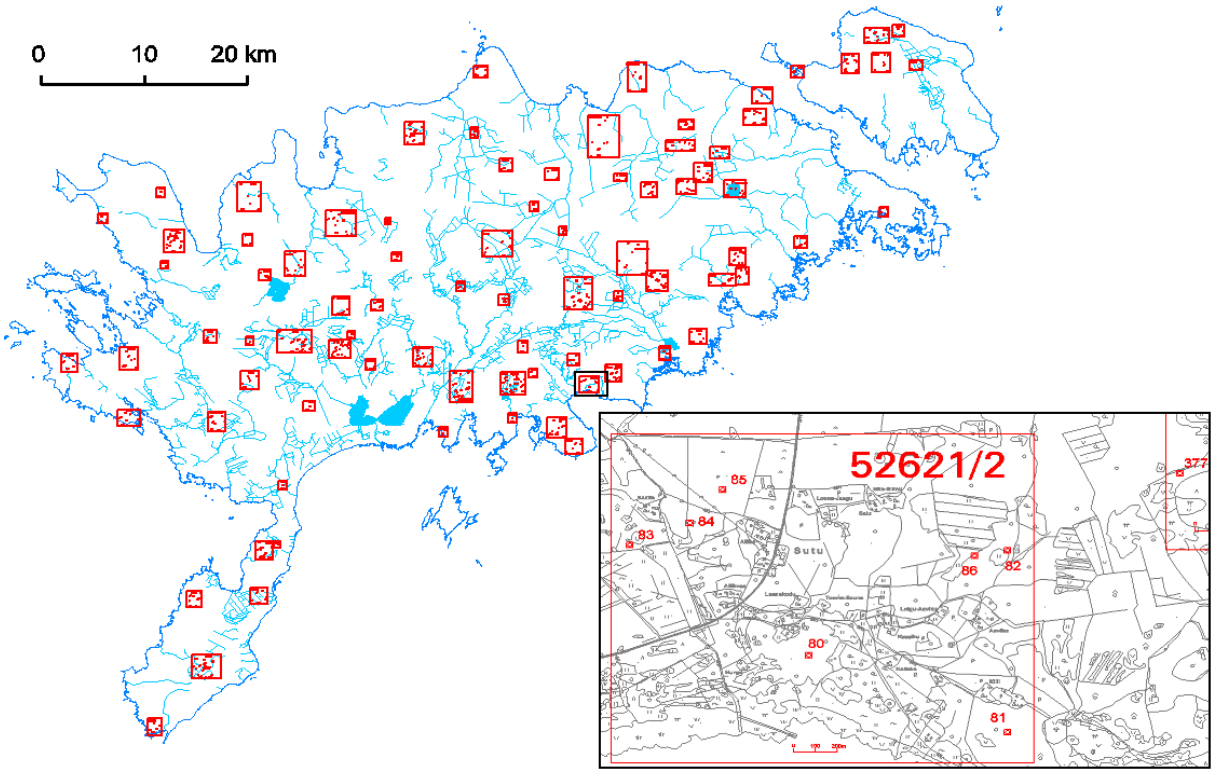
6.1.2. Liitprotsessid

Samal pinnal võib toimuda rohkem kui üks punktprotsess. Sel juhul on üksikprotsessid liitprotsessi komponendid. Liitprotsessi komponendid võivad olla omavahel sõltumatud või üksteist mõjutada, võivad luua sama või erinevat liiki sündmusi. Mitmeliigiliste punktmustrite genereerimist kasutatakse enamasti liikidevaheliste mõjude ja koosesinemise uurimiseks (ptk 4.1.3). Naabrite liigikuuluvust arvestava mitmeliigilise punktmustri moodustamiseks tuleb mudelisse lülitada liikide omavahelised mõjud, mis ei pruugi olla sümmeetrilised. Populatsiooni dünaamika mudel peaks sisaldama ka viljakuse sõltuvust asustustihedusest.

6.1.2.1. Juhuslik liitprotsess

Juhuslike liitprotsesside (*Poisson cluster process*) käigus genereeritakse kõigepealt pinnale esimest järku sündmused (*parent events*). Teist järku sündmused (*daughter events*) paigutatakse esimest järku sündmuste suhtes juhuslikus suunas ja juhuslikule kaugusele. Tulemusena tekib juhuslik reeglina agregeeritud muster. Dünaamilise liitprotsessi korral on esimest ja teist järku protsesside intensiivsus ajas muutuv. Juhuslikku liitprotsessi saab kasutada vaatluskohtade valimi moodustamiseks, kui ühest punktist teise liikumise vahemaad ja aega on vaja kokku hoida (joonis 6-1).

Juhuslikud liitprotsessid koosnevad rohkem kui ühest juhuslikust protsessist



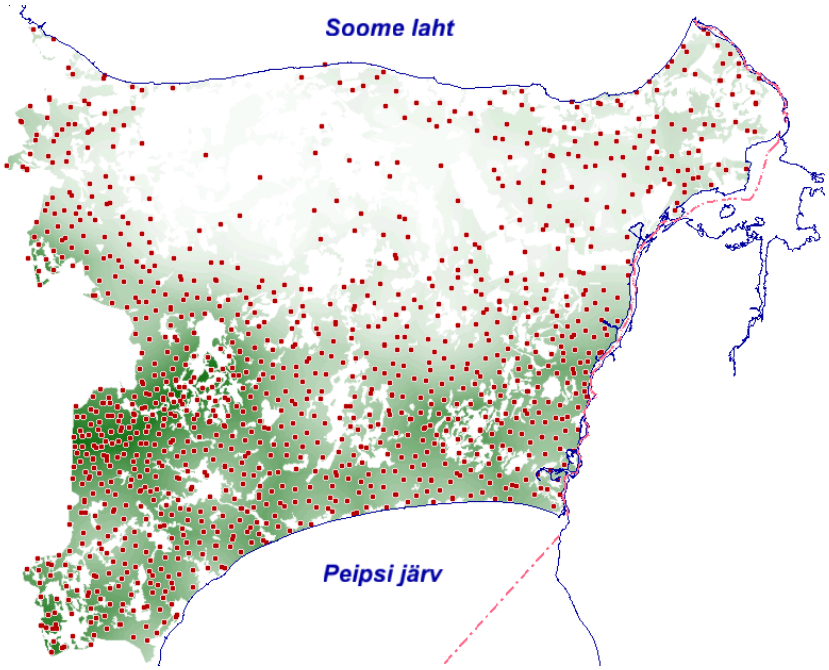
Joonis 6-1. Juhuslike piirkondade ja kohtade liitprotsessi tulemusel saadud vaatluskohtade paiknemine Saare maakonnas.

6.1.2.2. Heterogeenne juhuslik protsess

Heterogeense ehk ebaühtlase Poissoni protsessi ehk topeltstohhastilise protsessi (*doubly stochastic process*) ehk **Coxi protsessi** (*Cox process*) puhul on protsessi intensiivsus ruumiliselt varieeruv. Protsessi modelleerimise põhiprobleemiks on intensiivsuse muutlikkust kirjeldava funktsiooni leidmine. Protsessi intensiivsus võib olla määratud koha paiknemisega mingite protsessi jaoks oluliste nähtuste (näiteks reostusallikate) suhtes või pinna omadustega (näiteks elupaiga sobivusega) (joonis 6-2).

Ebaühtlase sobivusega pinnal toimiv juhuslik lihtprotsess võib anda täpselt samasuguse tulemuse kui heterogeenne juhuslik protsess või juhuslik liitprotsess. Seetõttu ei saa vaid ühe punktmustri järgi teha kindlaid järeldusi mustrit kujundava protsessi tüübi kohta. Kui pinna ebaühtlastest omadustest tingitud punktide koondumine on tingitud protsessivälistest teguritest, siis nimetatakse seda **näiliseks koonduvuseks** (*apparent contagion*).

Lihtsa juhusliku protsessi poolt ebaühtlasele pinnale tekitatud punktmuster võib paista sama moodi agregeerituna kui juhusliku liitprotsessi tulemus



Joonis 6-2. Põtrade elupaigasobivus (valge – sobimatu, tumedam roheline – kõige sobivam) ja üksikpõtrade prognoositud paiknemine Ida-Viru maakonnas (punased täpid). Andmed Luud ja Remm (2001).

6.1.2.3. Neyman-Scotti protsess

Homogeenne **Neyman-Scotti protsess** on liitprotsess, mille puhul eeldatakse klastrite omavahelist sõltumatust ning juhuslike sündmuste sama intensiivsust ja oodatavat jaotust igas klastris. Neyman-Scotti protsess toimub järgmiste reeglite järgi:

- esimest järku sündmused (vanemad) tekivad homogeense juhusliku protsessi tulemusel (neid ei ole võimalik vaadelda),
- iga esimest järku sündmuse oodatav järglaste arv on juhuslik,
- järglaste oodatav paiknemine emapunkti suhtes on kahemõõtmelise normaaljaotusega. Kahemõõtmelisus tähendab siinjuures normaaljaotust nii pinna *x*-telje kui ka *y*-telje suunas ehk kellukakujulist tihedusfunktsiooni.

Neyman-Scotti protsessi korral on järglaste kaugus vanemast normaaljaotusega

6.1.2.4. Võreprotsess

Võreprotsessi puhul on punktid seotud korrapärase võrestikuga. Punktidel on lubatud paikneda kas ainult etteantud lõplikus arvus kohtades või siis juhuslikult etteantud kohtade ümber vastavalt mingitele reeglitele. Sisuliselt on võremustrid kõik juhuslikud mustrid, mille komponentide koordinaate väljendatakse kas täisarvudes või kümnendmurdudes lõpliku komakohtade arvuga.

Võreprotsessid on seotud etteantud asukohtadega

Uurimused

Moody *et al.* (1997) jäljendasid meriharakate (*oystercatchers*) ruumilist paiknemist austripangal ringikujuliste alade juhusliku mittekattuva paigutamisega. Lindude statistiliselt oluline üksteise vältimise raadius saadi tegeliku paiknemismustri võrdlemisel juhusliku paiknemisega. Vältimisraadiuse suurust seostati keskkonnafaktoritega.

Coomes *et al.* (1999) kasutasid ühe taime järglaste arvu modelleerimiseks järgmist reeglit: esimesed eksemplarid paigutati pinnale juhuslikult, nende järglased paigutati juhuslikult raadiuse *r* ulatuses.

$$F = \frac{F_i}{1 + \sum_j \alpha_{ij} N_{ij}}, \qquad [6\text{-}1]$$

kus F_i on liigi *i* maksimumaalne viljakus, α_{ij} on liigi *i* ja *j* vahelise mõju koefitsient ning N_{ij} on liigi *j* isendite arv raadiuses *r* liigi *i* isendist.

Reich *et al.* (2004) modelleerisid kanakulli pesade paiknemist arvestades elupaiga sobivust ja kanakullide territoriaalset käitumist. Territoriaalsust esindas mudelis pesade vahemaast sõltuv tõmbumise/tõukumise suhteline potentsiaal.

6.1.3. Harvendusega protsessid

Harvendusega juhuslik protsess (*thinned stochastic process*) sisaldab teatud osa eelnevalt genereeritud sündmuste eemaldamist. Eemaldamine võib olla puhtjuhuslik või sõltuda ajast, kohast, koha ümbrusest ja/või genereeritud punkti vanusest. Harvendusega protsesse kasutatakse puistu arengu modelleerimisel, kus harvendus imiteerib puude hukkumist.

Harvendusega algoritm kujundab objektide paiknemismustrit osade objektide eemaldamise abil

6.1.4. Markovi protsessid

Markovi protsessis sõltub koha seisund ainult sama koha ja selle ümbruse seisundist eelmisel ajahetkel. Markovi punktprotsessi puhul sõltub sündmuse suhteline tõenäosus naabruses juba toimunud sündmustest. Tõkestava mõjuga Markovi protsessi nimetatakse **inhibitsiooniga protsessiks**, mille puhul vähendavad juba toimunud sündmused või olemasolevad objektid uute sündmuse või objektide tõenäosust enda lähedal. Ümbruse mõju võib punktmustri moodustamisel arvutada iga koha (piksli) jaoks lähtudes kas ümbruse konstantsetest omadustest või lisades nähtuse esinemise/mitteesinemise tõenäosusele selles kohas juba paigutatud objektide mõju.

Inhibitsiooniga protsessides vähendavad varasemad sündmused uute sündmuse tõenäosust enda lähedal

Lihtsad inhibitsiooniga protsessid (*simple inhibition processes*) kasutavad konstantset mõju tugevust ja ulatust. **Keerukates inhibitsiooniga protsessides** (*nonsimple inhibitory processes*) reguleerib uute sündmuste tõenäosust lähikonnas olevate sündmuste tihedus. Arvestada saab ka iga punktsündmuse omadusi (näiteks puu liiki ja suurust) ning iga lähikonnas oleva objekti kaugust.

Naabruses olevate objektide hulka ja nende omadusi saab kokku võtta mingisse ühte reaalarvulisse sobivusindeksisse. Nomiaalsete atribuutandmete korral võib uute punktide lisamisel olla oluline naabruses olevate objektide klassikuuluvuse kombinatsioon. **Paarilise toimega protsessis** (*pairwise interaction process*) mõjutab sündmuse tõenäosust vaid lähim naaber.

6.1.5. Gibbsi protsessid

Erinevalt inhibitsiooniga protsessidest võivad **Gibbsi protsessis** naabruses olevad tõenäosused mõjuda nii stimuleerivalt kui ka inhibeerivalt. Kõige tõenäolisem objektide paiknemine leitakse iteratiivsel mustri sobitamisel, mitte ühekordselt. Seega sisaldab Gibbsi protsess ka harvendust ja on eelkõige paiknemismustri sobitamise, mitte vaid struktuuri loomise vahend. Mustrit moodustatakse senikaua, kuni see vastab piisavalt hästi etteantud parameetritele.

Juba genereeritud objektide nihutamine iga uue objekti lisamise järel eristab Gibbsi protsesse teistest tagasisidemega protsessidest

Gibbsi protsesse hakati esmalt kasutama statistilises füüsikas. Praeguseks on neid rakendatud ka bioinformaatikas, kujutiste analüüsil, lünklike andmete visualiseerimisel, näiteks reljeefikujutise loomisel üksikute kõrgusandmete järgi ja fraktalite genereerimisel. Klassikaline Gibbsi protsess on homogeenses ruumis olevate, ühetaoliste, lõpmata väikeste, üksteise paiknemist mõjutavate osakeste asukoha mudel. Protsessi käivitava sündmuse asukoht võib olla juhuslik.

Keerukamate Gibbsi protsesside puhul võivad osakeste omadused ja mõju erineda (märgitud protsess) või loetakse ruumi omadused ebaühtlasteks (Mateu *et al.* 1998). **Straussi protsessi** puhul sõltub protsessi intensiivsus vaid lähedaste, protsessi parameetriga ettantud raadiuses olevate, naabrite arvust (Strauss 1975).

6.1.5.1. Gibbsi sampler

Gibbsi protsessi abil iteratiivselt sobitatav struktuur võib olla nii punkt- kui ka pindmuster. Mustrit Gibbsi protsessina sobitavat generaatorit nimetatakse **Gibbsi sampleriks** (*Gibbs' sampler*). Gibbsi samplerit saab kasutada nii klassifitseeritud kui ka pideva muutuja väärtuspinna prognoositavat struktuuri sobitava protsessina. Gibbsi sampler võimaldab lisada oodatava struktuuri kohta teadaolevaid andmeid paiknemisprognoosi mudelisse.

Sampler käivitatakse kas punktide juhusliku paiknemisega või mingi mudeli, punktandmete puhul enamasti logistilise, järgi eeldatud paiknemisetõenäosusega. Nähtuse esinemine või puudumine uurimata kohas prognoositakse nähtuse ümbruses esinemise ja puudumise ning nähtuse üldise sageduse või paiknemismudeli järgi. Protsessi korratakse senikaua, kuni nähtuse esinemise sageduse vahekord kohtades ja nende ümbruses enam ei parane. Seega otsitakse paiknemismustrit, kus punktide esinemine/puudumine vastab kõige paremini nii sama koha parameetritele kui ka tõenäolisele esinemisele naabruses. Gibbsi sampler on üks Markovi juhuslike väljade (*Markov random fields – MRF*) rakendusi.

Gibbsi sampler on iteratiivne meetod muutuja väärtuste või tõenäosuste pinna moodustamiseks sama muutuja väärtuste järgi ümbritsevas ruumis ja teadaolevate paiknemisseaduspärasuste järgi

Gibbsi samplerit on muuhulgas kasutatud pildi tõlgitsuse ülesannetes. Igale pikslile kõige tõenäolisemalt vastava maakatteklassi leidmine on statistilise otsuse probleem, mille lahendamisel saab kombineerida väärtuste tõenäosusi, kontekstist (ümbrusest) sõltuvaid korrektuure ja iteratiivset lähendamist. Protsess täpsustab piksli prognoositavat klassi ümbruses olevate pikslite omaduste järgi senikaua, kuni prognoos stabiliseerub (Atkinson ja Lewis 2000, Schröder *et al.* 2000).

Tarkvara

Gibbsi sampler sisaldub tarkavaralahendustes JAGS (http://sourceforge.net/projects/mcmc-jags/files) ja WinBUGS, mis on projekti BUGS (*Bayesian inference Using Gibbs Sampling*) (http://www.mrc-bsu.cam.ac.uk/bugs) käigus loodud Microsoft Windows tarkvara.

Uurimused

Augustin *et al.* (1996) kasutasid liigi oodatava esinemise või puudumise prognoosimiseks uurimata ruutudes nelja meetodit:

- logistiline mudel (sobivusväli),
- autologistiline mudel (logistilisele mudelile lisatakse autokovariatsioon),
- autologistiline mudel koos Gibbsi sampleriga (uurimata ruutude täitmiseks kasutatakse lisaks eeltoodule ka esinemise/puudumise iteratiivset juhuslikkust),
- autologistiline mudel koos modifitseeritud Gibbsi sampleriga, (arvestab mitte esinemise puudumise kui kaheväärtuselise tunnuse autokovariatsiooni, nagu tavalises Gibbsi sampleris, vaid reaalarvulist prognoositud esinemistõenäosust).

See kiirendas tunduvalt arvutusi ja andis realistlikumaid tulemusi. Gibbsi samplerit kasutati iteratiivselt juhuslikke alguspikslitega. Tulemuseks oli nähtuse esinemise/puudumise hinnang uurimata ala igas osas.

Pacala (1987) modelleeris liikide koosesinemist/väljatõrjumist erineva laigulisusega keskkonnas.

Maastikuelementide paiknemismustrite dünaamilise modelleerimisega on palju tegeletud metsanduses. Frelich *et al.* (1993) võtsid kahe puuliigi vahekorra pikaajalisel modelleerimisel aluseks nende liikide esindajate praeguse paiknemise teise liigi suhtes. Dünaamilise mudeli lähteseisuks võeti looduses esineva tiheduse ja vanuselise struktuuriga kujuteldav kaheliigiline (suhkruvaher ja kanada kuusk) mets. Leiti, et 3000 aasta järel peaksid sama liigi esindajad koonduma laikudesse. Kuna tegelikult need liigid siiski metsas laikudena ei esine, siis peab sellel mingi põhjus olema. Näiteks mingi häiriv faktor,keskkonna või liigi levila muutumine, või siis ei ole laigulisus veel jõudnud tekkida.

Hanus *et al.* (1998) seadsid metsa simulatsioonis juba genereeritud puude inhibitsioonitsooni vastavusse puu suurusega. Puude suuruste jaotus saadi väliuuringust. Maapinna omaduste erinevust modelleerimisel ei arvestatud.

Varasematesse puistu mudelitesse, mis ei kujuta üksikuid puid, on juhuslikkust lülitatud ilma seda ajas ja ruumis varieerimata. Puudekaupa mudelitesse on eelkõige lülitatud ajalist juhuslikkust (Miina 1993, Stage ja Wykoff 1993, Kangas 1997, 1998). Puudekaupa metsa modelleerimise puhul on lisaks sellele puu omadusi seotud naaberpuude ja koha omadustega. Tuleviku mudelid peaksid kajastama nii ajalisi kui ka ruumilisi juhuslikke struktuure (Fox *et al.* 2001).

Goreaud *et al.* (2002) lisasid varasematesse puistu struktuuri määravatesse liikidevahelise konkurentsi mudelitesse üksikpuude omavahelised suhted ja kasutasid Gibbsi samplerit puistu

oodatavate ruumiliste struktuuride moodustamiseks. Üksikpuu ellujäämus ja kasv seati sõltuvaks mulla omadustest ja puistu koosseisust selle puu ümbruses. Autorid näitasid mudelitega, kuidas puuliikide ruumiline koondumine aitab kaasa konkurentsis nõrgemate liikide püsimisele. Nii samast kui ka teistest liikidest naaberpuude tiheduse mõju hindamiseks kasutati Ripley *L(r)* statistikut.

Remm ja Luud (2003) kasutasid põtrade paiknemismustri modelleerimisel naabrite tiheduse jaotust. Naabrite tihedust käsitleti pinna omadusena – see võib olla optimaalne ja koha sobivust parandada, samuti võib see optimumist tugevasti erineda ja sellega koha sobivust järgmise punkti jaoks vähendada.

6.1.7. Dünaamilised mustrid

Ökoniši käsitlusele tugineva mudeli kohaselt peaks liigi esinemise sagedus nii ajas kui ka ruumis olema proportsionaalne liigile sobiva elupaiga hulgaga. Seega ökoniši mudelid populatsiooni paiknemise muutumist ei kirjelda.

Difusiooniprotsess käsitleb populatsiooni liikumist maastikul abstraktse juhusliku protsessina, mille puhul üksikindiviidide liikumist ei modelleerita. Üksikute isendite juhusliku liikumise modelleerimiseks kasutatakse juhuretkede ehk **juhujalutajate** (*random walks*, *random walkers*) meetodit (Pearson 1905), mille kohaselt on liikuva objekti iga järgmise sammu suund ja pikkus juhuslik. Juhujalutajate meetodi kasutust ökoloogias vaata näiteks Marsh ja Jones (1988), Bergman *et al.* (2000), Codling *et al.* (2008).

Metapopulatsiooni mudelid eeldavad, et leviku ja arvukuse nii ajaline kui ka ruumiline struktuur kujunevad elupaiga eraldi asuvate osade juhusliku asustamise ja lokaalpopulatsioonide juhusliku väljasuremise tulemusel (Levins 1969, Hanski 1982, 1994, 1997). Metapopulatsiooni käsitluse järgi võib liikide arvukus ja konstantsus nii ajas kui ruumis juhuslikult varieeruda isegi statsionaarsete keskkonnatingimuste korral. Konstantsed ja juhuslikud liigid võivad metapopulatsioonide käsitluse järgi vahetuda. Metapopulatsioonimudelites on iga laik omaette üksus, seetõttu on metapopulatsiooni mudelid arvutuslikult efektiivsed vaid siis, kui laike ei ole väga palju. Maastiku struktuuri ja populatsiooni liikumise seostamiseks metapopulatsiooni käsitlus hästi ei sobi, sest selles käsitletakse igat elupaigalaiku individuaalselt (Baker 1996), küll aga on see laialt rakendatud populatsioonide dünaamika ja teiste ökoloogiliste protsesside modelleerimiseks (Wu ja Levin 1994, 1997).

Geograafilist aegruumi, ajalist kaugust ja liikumistakistust kasutakse optimaalse liikumise modelleerimiseks. Eeldatakse, et iga indiviid peab oma liikumisi olelusvõitluses edukas olemise tarvis optimeerima. On näidatud, et liikide ruumis ja ajas leviku struktuur võib erinevates mõõtkavades olla erinev. Ka vaatlusalade suurus on oluline. Väikeste proovialade korral on liikide leidmine/mitte leidmine juhuslikum.

6.1.8. Punktprotsessi verifitseerimine

Genereeritud punktmustri tegelikkusele vastavust saab kontrollida mingite testaladel mõõdetud kontroll-statistikute järgi. Keskväärtused ja esinemissagedused ei iseloomusta paiknemist ja seetõttu ei ole neist mustri verifikatsioonis palju abi. Punktmustrite verifikatsioonis on kõige enam kasutatud Ripley *K(t)* funktsiooni (Hanus *et al.* 1998, Mateu *et al.* 1998, Getis ja Franklin 1987), lähima naabri kaugust (Gappa *et al.* 1997, Mateu *et al.* 1998), bootstrap meetodit (Augustin *et al.* 1996) ja liigendnoameetodit (*jack-knife*) (ptk 3.6.4 ja 3.6.5).

Punktmustrite hindamiseks kasutatakse objektide omavahelise paiknemise statistikuid ja osavalimite moodustamist

Mitme nähtuse koosesinemise mustri tegelikkusele vastavust kontrollitakse eelkõige kokkulangevuse ulatuse järgi. Ruumilist struktuuri saab kontrollida mitmesuguste struktuuriparameetrite abil.

6.1.9. Vaatluskohtade kavandamine

Paljudel juhtudel ei ole kogu uuritavalt alalt andmete kogumine võimalik ja piirdutakse mingi põhimõtte järgi tehtud valimiga. Korrapärane valim võib seejuures sattuda sünkrooni mõne tunnuse ruumilise mustri korrapäraga ja põhjustada valimi ebaesinduslikkust. Juhuvalimi puhul on probleemiks vaatluste paiknemise ebaühtlus – mõned vaatlused paiknevad lähestikku, mõned teistest eraldi. Lähedal olevad vaatlused ei ole sõltumatud, üksikuna paiknevad vaatlused peavad esindama ebaproportsionaalselt suurt piirkonda. Vaatluste juhusliku paiknemise korral ei ole kõik uuritava ala osad võrdselt esindatud.

Harv ei ole olukord, kus uuritav ala on suur ja seetõttu oleks kasulik vähendada ühest kohast teise liikumist. Samas on tarvis vältida vaatluspunktide paiknemist liiga lähestikku ja tagada klasside teadaolevale pinnale vastav esindatus. Sellisel juhul sobib kasutada ruumiliselt juhuslikku liitjaotust, kus klastrikeskmed (emapunktid) paiknevad juhuslikult ja tütarpunktid paiknevad emapunktidest juhuslikus suunas ja juhuslikul kaugusel; kuid seejuures mitte kaugemal kui maksimaalne lubatud kaugus ja üksteisele mitte lähemal kui minimaalne lubatud kaugus (joonis 6-1). Mustri moodustamise käigus aktsepteeritakse vaid neid punkte, mis ei vii etteantud klasside esindatust tasakaalust välja.

Olulisematele faktoritele järgi kihilise valimi moodustamist kombineeritud eraldiste meetodil käsitleti peatükis 3.3.1.

6.2. Väärtuspinna moodustamine

Väärtuspinna loomise ülesanne eeldab vähemalt mõne pinna omaduse modelleerimist. Interpoleerimisel, mis on pinna moodustamine etteantud väärtuste alusel, on aluseks teadolevad väärtused ja nende järgi pinna moodustamise reegel. Juhuslike pindade puhul eeldatakse väärtuste juhuslikkust, kusjuures juhuslikkuse kohta peaks teada olema jaotus ja selle parameetrid (ühtlase jaotuse korral muutumisvahemik, normaaljaotuse korral keskväärtus ja dispersioon, Poissoni jaotuse korral keskväärtus). Pideva ruumilise muutuja puhul luuakse selle muutuja muutuvate väärtuste pinda, nominaalse muutuja puhul laigulisust.

Pindade moodustamise rakendused esinevad mitmes valdkonnas:

- väärtuste prognoosimine, mille hulka kuulub ka interpoleerimine ja ekstrapoleerimine,
- nullmudelile vastavate pindade moodustamine,
- mustrite paiknemisvastavuse mõõtmine,
- visualiseerimisülesanded.

Pindade genereerimise ülesande hulka kuulub näiteks liigi oodatava esinemise või ohtruse pinna moodustamine kas teadaolevate elupaigatingimuste järgi või potentsiaalse mustri moodustamine uurija poolt etteantud tingimustel. Võimalike elupaikade ja elustiku mudelid aitavad ette näha maastikumuutuste võimalikke tagajärgi. Maastiku nullmudel on võrdluseks mingi sisuka hüpoteesi alusel genereeritud mustrile või looduses esinevale mustrile.

Väärtuspinna moodustamise ja maastiku modelleerimise ülesannet võib ruumikäsitluse tehnika seisukohast lahendada punktobjektide (sealhulgas alade tsentroidide), korrapärase võrgustiku (rastri), ebakorrapärase võrgustiku (tesselatsiooni) või laikude (eraldiste) abil. Dünaamilistele laikudele tugineva maastikumodelleerimise kontseptsiooni on esitanud Wu ja Levin (1997). Chen *et al.* (2011) mainivad järgmisi ökosüsteemi aeg-ruumilise modelleerimise suundi: diffusioonimudelid, rakk-automaat ja indiviidipõhised mudelid.

Kui vektorpolügoonide kuju muutumise modelleerimine on jäänud keerukaks ülesandeks, siis pindstruktuuri moodustamine korrapärase võrgustiku abil on tehniliselt suhteliselt lihtne. Iga pikslivääruse lisamise või muutmise järel kontrollitakse tekkinud mustri vastavust etteantud parameetritele (McElhany *et al.* 1995). Meetod suudab genereerida tegelikuga väga sarnaseid juhumustreid, kuid on suurte pindade puhul arvutuslikult ebaefektiivne (Real ja McElhany 1996).

Roxburgh ja Chesson (1998) genereerisid laigulisust juhuslike kõrvutiolevate vastandvärvi pikslite värvide äravahetamise teel. Kontroll-statistikuna kasuti lihtsaid külgnemiskontaktide statistikuid (servakontaktide arv *E*, nurgakontaktide arv *C*, tühjad pikslid *O*, täis pikslid *S*). Tekkiva mustri vastavust etteantule kontrolliti funktsiooniga

$$\phi = \left|\frac{E_{null}}{E_{obs}} - 1\right| + \left|\frac{C_{null}}{C_{obs}} - 1\right| + \left|\frac{O_{null}}{O_{obs}} - 1\right| + \left|\frac{S_{null}}{S_{obs}} - 1\right|. \qquad [6\text{-}2]$$

Kui genereeritud mustri struktuur vastab täielikult etteantule, on $\varphi = 0$. Juhusliku mustri puhul on see kuskil 0 ja 3 vahel. Kui külgnevate pikslite väärtuste vahetamise järel φ vähenes, siis jäeti muutus jõusse, kui ei vähenenud, siis taastati vahetuseelne seis ja valiti uus juhuslik paar. Protsessi korrati senikaua, kuni φ oli piisavalt väike.

Mustreid on moodustatud ka ökoloogilistest andmetest (Chiarello ja Barrat-Segretain 1997, Palmer 1992, Watkins ja Wilson 1992, Mangel ja Adler 1994, Palmer ja van der Maarel 1995, Wilson 1995).

6.2.1. Neutraalsed maastikumudelid

Neutraalseteks mudeliks nimetatakse ökoloogias minimaalsest reeglite arvust koosnevat nullmudelit, mis jäljendab uuritava protsessi puudumisel oodatavat struktuuri. Neutraalseid mudeleid on ökoloogias kasutatud eelkõige liikidevahelise konkurentsi ning koosluste ja maastiku struktuuri uurimisel (Hansson 1984, Begon *et al.* 1990, Pearson ja Gardner 1997). Mõisted neutraalne mudel ja nullmudel on väga lähedased. Nullmudel on rohkem seotud juhuslikkuse eeldusega kui neutraalne mudel.

Maastikuökoloogias eelistatakse terminit **neutraalne maastikumudel**. Maastikumudelid on eriti olulised, sest eksperimendid on maastiku mõõtkavas harva võimalikud ning nende asemel tuleb kasutada arvutatud jäljendusi. Neutraalsed maastikumudelid on mänginud olulist rolli teoreetilise maastikuökoloogia arengus (Gardner *et al.* 1987, Turner *et al.* 1989, With ja King 1997, Hagen-Zanker ja Lajoie 2008). Neutraalse mudeli abil saadud prognoosi võrreldakse kas empiiriliste andmetega või nende mudelite prognoosidega, mis sisaldavad uuritavat protsessi.

Neutraalne mudel ei pea olema maastiku realistlik jäljendus, ka lihtne juhuslik mudel võib olla väärtuslik uurimuse lähtekohana. Lihtsast juhuslikust mudelist genereeritud maastikumustrid on kaardipildina ebarealistlikud, kuid need ei pruugi olla ebarealistlikud igas mõttes ja kõigi parameetrite osas. Näiteks elupaiga osade ühendatuse võrdlemiseks ei pea eraldiste kuju tingimata realistlik olema. Teiseks, kõik tuntud teoreetilised mudelid on detailides ebarealistlikud, sest eeldavad teatud ideaalseid tingimusi. Sellegipoolest kasutatakse neid lähtekohana uue ja täpsema teabe saamiseks ja formuleerimiseks.

Neutraalses maastikumudelis asendatakse uurimuse aspektist kõrvalised omadused juhuslikkusega

Empiiriliste andmete ja neutraalse mudeli vastavuse korral ei tuleks siiski järeldada, et neutraalne mudel on õige. Sellisel juhul ei ole lihtsalt põhjust keerukamat mudelit otsima hakata. See ei tähenda ka, et neutraalsest mudelist välja jäetud faktor üldse ei mõjuks. Lihtsalt nende andmete ja selle metoodika järgi ei õnnestunud mõju tõestada. Nii nagu nullhüpoteesi ei saa tõestada, nii ei tohiks kunagi väita neutraalset mudeli kehtivust. Kui olla päris järjekindel, siis ei tohiks kunagi midagi väga kindlalt väita. Teadustöö tulemused annavad vaid tõendeid ühe või teise seisukoha või mudeli kehtivuse kohta teatud tingimustel.

Neutraalne mudel on abstraktsioon, mis ei kehti kunagi; selle juurde jäädakse, kuni muud ei ole õnnestunud tõestada

Ülevaate neutraalsetest maastikumudelitest ja nende kasutusest annavad With ja King (1997), Johnson *et al.* (1999) ning Hagen-Zanker ja Lajoie (2008). Neutraalseid maastikumudeleid kasutatakse With ja King (1997) järgi järgmistel eesmärkidel.

- Ruumimustreid kirjeldavate indeksite võrdlemine.
- Ökoloogiliste näitajate kriitiliste väärtuste hindamine. Elupaikade ühendatuse kriitilise väärtuse juures muutub näiteks hüppeliselt organismide levimisvõime, elupaiga kriitiliselt vähese hulga juures võib populatsioon kiiresti välja surra.
- Liikide elupaigastruktuuri nõudluste selgitamine. Maastiku struktuuri liigi esinemise ja liigi puudumise kohtades võrreldakse neutraalsete mudelitega.

- Maastiku keerukuse üldistatud mudelite väljatöötamine. Nii ökovõrkude kui ka metapopulatsioonide teooria eeldab maastiku jagamist selgepiirilisteks elupaigalaikudeks ja ühendusteedeks. Eraldiste kuju, omavahelise paiknemise ja omaduste kõikvõimalikke variante reeglina ignoreeritakse. Maastikku pikslitena käsitlevad mudelid võimaldavad ökoloogilisi protsesse paindlikumalt modelleerida.
- Ruumilise heterogeensuse lülitamisel populatsioonide ja koosluste ökoloogilisse teooriasse. Näiteks aegruumilised populatsioonilained ja liikide koosesinemise teooria.

Esimesed neutraalsed mudelid olid perkolatsiooniteooriast pärinevad **binaarsed juhumustrid**, kus iga piksel on teistest pikslitest sõltumatu ja omab kaheväärtuselise muutuja väärtusi konstantse tõenäosusega. Kui mingit kategooriat on palju, moodustab see kogu maastiku ulatuses ühenduses oleva maatriksi. Kategooria osakaalu langemisel fragmenteerub juhuslik maatriks laikudeks. Millise osakaalu juures fragmenteerumine toimub, sõltub rastri geomeetriast, kategooriate ruumilisest jaotusest ja sellest, kuidas perkolatsiooni modelleeritakse. Kui ühe klassi tõenäosus on >0,5928, siis reeglina ei teki selle klassi eraldatud laigud, vaid servast servani ulatuv maatriks ja liikumine selle kategooria pinnal ühest rastri servast teise on võimalik.

Binaarse maastikumustri edasiarendus oli multinomiaalne muster (Gardner *et al.* 1987). Gardneri algoritm oli järgmine:

- jaga kaart piksliteks;
- omista juhusliku valiku põhjal igale pikslile maakatte klass, jälgides klasside pindalalist proportsiooni;
- moodusta külgnevatest sama klassi pikslitest laigud.

Gardneri mudel on neutraalne sedavõrd, kuivõrd juhuslikkus sobib nullmudeliks. See esindab uuritava protsessi puudumist.

Lihtsat juhuslikku kaheväärtuselist mustrit nimetatakse ka **perkolatsioonikaardiks**. Lihtsad pikslite kaupa juhuslikud kaardid ei ole realistlikud maastikumudelid, kuna ei sisalda ruumilist autokorrelatsiooni. Laikude arv ja servajoonte pikkus on lihtsas juhuslikus mudelis suurem kui looduses tavaks. Lihtsa juhusliku mudeli kasutamisest ei tule siiski järeldada, et maastikuökoloogid oleksid uskunud sellise mudeli reaalsusesse. Säästvusreegli kohaselt ei ole aga keerukamat mudelit põhjust enne kasutada, kui juhuslikkusel baseeruv nullmudel on kehtetuks tunnistatud.

O'Neill *et al.* (1992) ja Lavorel *et al.* (1993, 1995) lisasid neutraalsete mudelite kontseptsioonile kihtide või tasemetega määratud hierarhilise mõõtkava. Iga eelmisel tasemel edukas piksel jagatakse järgneval tasemel osadeks ja osadel on mingi edu tõenäosus.

Johnson *et al.* (1999) rakendasid hierarhilise mõõtkavaga neutraalset mudelit mitmekategoorialisele maastikumudelile kasutades Markovi ahela metoodikat, kus ajasammu asemel kasutati detailiseerimissammu. Kuna laikude kuju ja suurust ei modelleeritud, siis saadi nullmudel, mitte realistlik maastiku kujutis. Johnsoni algoritm on järgmine:

- alusta kõige suuremate pikslitega ja omista igale pikslile pinnaklassi kategooria vastavalt klasside esinemise tõenäosustele;
- jaga iga emapiksel $m \cdot n$ tütarpiksliks;
- omista igale tütarpikslile pinnaklassi kategooria vastavalt tütarpikslite klasside tõenäosustele selle emapiksli klassi korral selles mõõtkavas (Markovi ahela üleminekutõenäosuste analoogid);
- korda samme 2 ja 3.

Vajalike tasemete hulga võib otsustada vastavalt vajalikule tulemuse detailsusele. Üleminekumaatriksis olevaid tõenäosusi võib valida vastavalt soovitavale maastikumustrile või arvutada empiirilistest andmetest. Kui klassi erinevatel tasemetel iseendaks jäämise tõenäosus on suurem, siis saab suuremate laikudega maastiku. Johnson *et al.* (1999) kasutasid näiteks pinnaklasside iseendaks jäämise tõenäosust 0,45. Kuna üleminekumaatriksite servasumma peab võrduma ühega, siis jagati ülejäänud tõenäosusest 0,45 vastavalt sagedamate pinnaklasside tegelikule sagedusele uuritavas piirkonnas ja 0,1 võrdselt kõigi haruldaste klasside vahel.

Juhuslikult laigulise neutraalse maastikumudeli Gibbsi protsessina genereerimise põhimõtted esitasid Gaucherel *et al.* (2006). Juhuslikku mustrit saab genereerida ka tegelikult esinenud laike juhuslikult kaardile paigutades (Hargis *et al.* 1998).

Nullmudelile vastava pinna moodustamise meetodite hulka kuulub olemasoleva mustri toroidnihutamine juhusliku nihke võrra. Meetod on kasutatav laigulisuse võrdlemiseks teise laigulisusega või juhustruktuuriga, kuid tuleb tähele panna, et toroidnihutamisel tekib mustrisse ebanormaalne hüppekoht (joonis 4-21). Toroidnihutus ei ole rakendatav ebaühtlase sobivusega pinna puhul, sest nihutamisel muutub mustri suhe pinna sobivusse.

Keskpunkti asendamise (*midpoint displacement*) metoodi puhul kasutatakse pideva muutuja väärtusvälja moodustamiseks fraktaalsust genereerivaid algoritme (Palmer 1992). Pideva väärtusvälja jagamisel väärtuskategooriateks saadakse reaalsetele sarnased maastikukaardid. Meetodit on kasutanud näiteks With ja King (2001) pesitsus-elupaikade ruumistruktuuri mõju analüüsil. Lähteparameetritena kasutati postuleeritud minimaalset elupaiga nõudlust ja kriitilist vahemaad eraldiste vahel. Järeldati, et populatsioonilätete säilitamiseks on fragmenteerunud maastike puhul olulisim säilitada elupaikade laigulisus pigem väheste, aga piisavalt suurte laikudena.

Taimkatte uuringutes sobivad neutraalseks mudeliks potentsiaalse loodusliku taimkatte kaardid (Ricotta *et al.* 2002). Elupaikade potentsiaalse leviku kaardid on olnud heaks aluseks liikide leviku modelleerimise meetodite võrdluses (Hoffmann *et al.* 2010).

Hagen-Zanker ja Lajoie (2008) märgivad, et maastiku muutumise hindamisel ei sobi klassikalised neutraalsed mudelid, sest muutuste algseis on varasem maastik, aga mitte juhuslik maastikumuster. Muutuste modelleerimisel on petlik ka saadud tulemuse vastavus oodatavale, sest mida vähem muutusi, seda lihtsam on saavutada kõrge vastavus oodatava kaardi ja vaadeldud kaardi vahel.

Gardner ja Urban (2007) teevad ettepaneku eraldada maastiku püsiv ja muutuv osa ning rakendada nende suhtes erinevat neutraalset mudelit.

Hagen-Zanker ja Lajoie (2008) rõhutavad, et maastikumuutuste jälgimisel tuleks kasutada spetsiaalseid **maastiku neutraalseid muutumismudeleid** (*neutral models of landscape change*) ehk lihtsalt neutraalseid muutumismudeleid, mis ei lähtu juhuslikkust lähtemustrist, vaid muutumisele eelnenud maastikust. Viimatimainitud artiklis on kahe neutraalse muutumise mudeli algoritmi kood. Juhuslik piirangutele vastav (*random constraint match*) mudel jälgib muutuste etteantud mahtu iga maakattekategooria osas, kuid valib muutuste kohad juhuslikult. Klastripiiride muutmise mudel (*growing clusters*) piirab juhuslike muutuste asukohti muutuvate kategooriate laikude piiril olevate kohtadega.

6.2.2. Maastikusimulaatorid ja metsa arengu mudelid

Kui neutraalsed mudelid, mis genereerivad mingite mõjurite puudumisel oodatavaid ruumilisi struktuure, ei püüagi jäljendada tegelikke maastikke, siis maastikusimulaatorite eesmärk on tegelikku või mingitel tingimustel oodatavat maastikku võimalikult realistlikult jäljendada. Laias mõistes võib maastikusimulaatorite hulka lugeda ka liikide või koosluste levimisprotsessi imiteerivaid mudeleid. Ülevaade metsa ja maastiku arengut jäljendavatest mudelitest on publikatsioonides Zavala ja Burkey (1997) ja Scheller ja Mladenoff (2007). Ülevaateartikli maakattemuutuste prognoosidest on kirjutanud Irwin ja Geoghegan (2001). Ülevaate segametsade juurdekasvumudelitest, millest osa arvestab paiknemissuhteid, võib leida artiklist Porté ja Bartelink (2002). Mõned maastikumudelid on lühidalt iseloomustatud alljärgnevalt.

Chiarello ja Barrat-Segretain (1997) imiteerisid taimestiku taaslevikut puhastatud alale **juhusliku püramiidprotsessina** (*stochastic pyramid*). Pilditöötluses tuntud püramiidprotsesssi puhul tekib uus kujutis eelmisest kindlate reeglite järgi (Jolion ja Rosenfeld 1994). Juhusliku püramiidprotsessi (Meer 1989, Meer ja Connely 1989) puhul on lähtekujutiseks mingites etteantud kohtades paiknevad objektid. Järgnevas laienemise staadiumis mõned objektid kaovad ja teised laiendavad oma ala juhuslikesse naaberpikslitesse. Juhitava püramiidprotsessi puhul sõltub objektide laienemine ka juba olemasoleva objekti konfiguratsioonist.

Saura ja Martínez-Millán (2000) töötasid välja modifitseeritud **juhuslike klastrite meetodi** (*modified random clusters – MRC*), mida saab kasutada korrapärases võrgustikus (rastris). Meedodi esimeseks etapiks on lihtsa juhusliku kaheväärtuselise kaardi moodustamine, mille puhul pikslite klassikuuluvuse tõenäosus on ette antud. Kui ühe klassi tõenäosus on >0,5928, siis reeglina ei teki selle klassi eraldatud laigud, vaid servast servani ühendatud maatriks. Teine etapp on juhuslikult tekkinud klastrite leidmine. Klastrisse kuuluvateks loetakse pikslid, mis on etteantud kriteeriumi järgi naabrid (4 naabrit, 8 naabrit, naabrid teatud suunas). Klastritele omistatakse kategooria vastavalt kategooriate etteantud pindalalisele vahekorrale. Klastritevahelise ala pikslid klassifitseeritakse sellesse klassi, mis esineb 8 naaberpiksli puhul kõige sagedamini. Klassifitseeritud naabrite puudumisel või mitme klassi võrdse naaberpikslite hulgas esindatuse korral omistatakse klass pikslile juhuslikult. Seejuures arvestatakse klasside etteantud vahekorda. Seega kontrollivad Saura ja Martínez-Milláni meetodi puhul mustri tekitamist järgmised parameetrid: laikude lähtetõenäosus, naabruse arvestamise viis, kategooriate arv ja iga kategooria tõenäosus.

Maastikusimulaatorit **LSPA** on kirjeldatud töödes Li ja Reynolds (1993) ning Li *et al.* (1993). See on välja töötatud eelkõige metsaraiete dünaamika jäljendamiseks. LSPA modifitseeritud ja üldisem variant **SHAPC** (*Spatial Heterogeneity Analysis Program for Categorical Maps*) arvutab genereeritud maastike jaoks ka heterogeensusindeksid (Li ja Reynolds 1994). Mainitud maastiku jäljendajate kasutamisel tuleb eelnevalt fikseerida ruumilise heterogeensuse viis komponenti:

- kategooriate arv,
- iga kategooria osakaal,
- laikude paiknemine,
- laikude kuju,
- laikude naabrussuhted.

Kaks esimest parameetrit on lihtsalt arvuliselt fikseeritavad. Laikude paiknemist jagasid Li ja Reynolds juhuslikuks, korrapäraseks ja agregeerituks, laikude kuju jagati ruudukujuliseks, ringi-kujuliseks ja juhuslikuks. Naabrussuhteid kontrolliti naabruse kontrasti maatriksi abil, mis kirjeldab pikslite külgnevust sama kategooria või teise kategooria piksliga. Laikude keskmine suurus on küll

paljude ruumilist heterogeensust kirjeldavate indeksite koostises, kuid seda võib asendada ka naabruse kontrast, sest suuremate laikude korral on naabrus ühetaoline suuremal alal. Ruumilise heterogeensuse parameetrite hulka võiksid kuuluda ka anisotroopiat, laikude ühendatust ja suuruse jaotust kirjeldavad parameetrid, kuid SHAPC simulaator neid ei kasutanud.

Maakasutuse muutumise simulaator **LUDAS** (*Land-Use Dynamic Simulator*) (Le *et al.* 2008, 2010) käsitleb nii inimpopulatsiooni kui ka maastikuosiseid üksteist mõjutavate iseorganiseeruvate agentidena. Mudeli ülesehituse põhilised osad on:

- majandamisviisi alusel tüüpidesse jagatud ja vastavalt tüübile otsuseid langetav talurahvas;
- eraldisteks jagatud maastik, kus igal eraldisel on atribuudid, mis iseloomustavad inimmõju, loodusmõjude ja omaenda arengu seaduspära tõttu muutumise viisi ja muutumise tõenäosust;
- maakasutust mõjutavad administratiivsed otsused;
- maakasutusotsusteni jõudmise seaduspärasused, mis sõltuvad majapidamise tüübist, maastiku komponentidest ja majanduse administreerimisest.

Maastikumudel **LANDIS** (Mladenoff *et al.* 1996, Mladenoff 2004) on algselt koostatud prognoosima muutusi metsa struktuuris ja liigilises koosseisus nii loomuliku arengu käigus kui ka raietööde järel (Gustafson *et al.* 2000). Mudeli uuem versioon LANDIS-II (Scheller *et al.* 2007, http://www.landis-ii.org) jäljendab metsa arengut, häiringuid, kliima muutumist ja seemnete levimist suuremal alal (tüüpiliselt 1...2000 km^2). LANDIS-II käsitleb eraldi iga puu ja põõsaliigi levikut muudetavas ruumilises ja ajalises detailsuses. LANDIS-II mudelile viitavate publikatsioonide loetelu on veebiaadressil http://www.landis-ii.org/documentation/PublicationsPage. LANDIS mudeli visualiseerimise vahendit kirjeldavad Birt *et al.* (2009). LANDIS mudeli uusim versioon on 4.0 (http://web.missouri.edu/~umcsnrlandis)

SELES (*Spatially Explicit Landscape Event Simulator*) (Fall ja Fall 2001) on pigem struktureeritud modelleerimiskeel kui valmis mudel.

TreeMig on põhiliselt metsa arengu mudel, mille igas ruumilises üksuses modelleeritakse metsa kasvu, konkurentsi puude vahel ja puude surma puistu igas rindes eraldi (Lischke *et al.* 2006). Eeldatakse puude juhuslikku paiknemist lahtri sees. Mudel on kasutatav erinevates geograafilistes ulatustes. Puistu koosseisu vaheldumise modelleerimine aitab genereerida ja selgitada maastikumustreid.

SORTIE (http://www.sortie-nd.org) (Pacala *et al.* 1993, 1996, Ménard *et al.* 2002) on erinevalt varasematest metsa dünaamika mudelitest nii ruumiliselt ilmutatud kui ka indiviididele tuginev, see tähendab iga indiviidi käitumine ja saatus sõltub tema asukohast ja iga indiviidi käsitletakse eraldi üksusena. Nii konkurentsi kui ka levimist käsitletakse pidevate kaugusest sõltuvate funktsioonidena. Metsa dünaamika tekib üksikute puude valguse pärast ja omavahelise konkureerimise tulemusena. Põhjalikult on töötatud tulemuste visualiseerimise ja animatsioonide kallal.

BTELSS (*Barataria-Terrebonne ecological landscape spatial simulation model*) (Reyes *et al.* 2000) jäljendas Louisiana ranniku märgalade muutumist. Mudelit kalibreeriti 1956., 1978. ja 1988. aastast pärit looduskaitse kaartide järgi. Mudel näitas, et muutlik ilmastik põhjustab suuremat elupaikade hävimist kui üksikud äärmuslikud aastad.

Metsa arengu mudel **COMMIX** sisaldab metsa kasvu nelja ruumiliselt ilmutatud protsessi: 1) päikesekiirguse püüdmine iga puu poolt, 2) orgaanilise aine moodustumine, 3) sünteesitud biomassi jagunemine puu osade vahel, 4) puu üleminek uude seisundisse. Kasvuprotsessiga koos toimub loomulik harvendus, mille tõenäosus sõltub puu omadustest ja asendist. COMMIX suudab jäljendada metsa arengut nii monikultuurides kui ka mitmeliigilise puistus (Bartelink 2000).

Maastiku arengu mudel **TELSA** (*Tool for Exploratory Landscape Scenario Analyses*)

(http://essa.com/tools/telsa) jäljendab taimkatte loomulikku arengut, juhuslikke häiringuid ja inimtegevust. TELSA mudelit on kasutatud suurte alade arengu jäljendamiseks aastakümnete ja aastasadade jooksul Põhja-Ameerikas. Mudeli mooduleid seob MS Access andmebaas, väljundid on graafikutena, kaartidena ja tabelitena (Kurz *et al*. 2000).

Uurimused

Maakatteklasside dünaamikat on maastikuökoloogias modelleeritud peamiselt Markovi ahelana (Aaviksoo 1993, Aaviksoo 1993, 1995, Aaviksoo *et al.* 1994, Logston *et al.* 1996). Markovi ahelatest oli juttu statistilise modelleerimise peatükis (ptk 3.4.4). Markovi mudelit on kasutatud ka kujutiste moodustamiseks naabruse muutumise tõenäosuste abil. Eeldatakse, et pikslile *i* omistatud väärtusklass määrab teiste väärtusklasside tõenäosuse naabruses. Protsessi võib alustada korrapärase vahemaaga pikslitele väärtuste omistamisest ja seejärel genereerida väärtused naabruses olevatele pikslitele.

Korduvalt on Euroopast pärineva asustuse eelse Ameerika maastiku rekonstrueerimiseks kasutatud Ameerika Ühendriikide maamõõtjate märkmeid, kes aastatel 1785…1919 omanikuta maad kruntideks mõõtsid (Kapp 1978, Grimm 1984, Iverson 1988, White ja Mladenoff 1994, Delcourt ja Delcourt 1996, Radeloff *et al.* 1999). Selliste, küllaltki ebajärjekindlate andmete järgi taimkatte rekonstrueerimise usaldusväärsust on kontrollinud Manies ja Mladenoff (2000). Need autorid leidsid, et kui üksikvaatluste interpoleerimisega on võimalik enam-vähem õigesti hinnata valdavaid metsatüüpe, siis ühe või teise taimkatteklassi pindala hindamine on ebatäpne. Interpoleerimise täpsust aitaks tõsta mulla ja reljeefiandmete kasutamine.

Lepš ja Kindlmann (1987) mudel näitas, kuidas puude paiknemine puistu kasvades juhuslikumaks muutub.

Taimkatte, põhiliselt metsa simulaatoreid on kasutatud aerofotodel ja suure lahutusega kosmosefotodel olevate mustrite näidiskujutiste loomiseks (Coops ja Culvenor 2000). Modelleeritakse etteantud struktuuriga (tihedus, laigulisus, puude kõrguserinevused) virtuaalne 3D mets, sellest moodustatakse etteantud valgustingimustega pealtvaate kujutis, mida võrreldakse tegelike fotodega. Struktuuri võrdlemiseks kasutatakse kõige enam lokaaldispersioone. Puistu kirjelduse alusel loodud 3D mudelit on kasutatud ka võrade liituvuse ja häilude paiknemismustri jäljendamiseks (Silbernagel ja Moeur 2001).

Külgnevuse või laigulisuse parameetritega on teadlikult manipuleeritud veel mitmetes uurimustes (Gardner ja O'Neill 1991, Gustafson ja Parker 1992, O'Neill *et al.* 1992). Maakatte kategooriate piirialasid arvestasid kõige tõenäolisemateks muutuste kohtadeks Coops ja Catling (2001) modelleerides metsa struktuuri keerukust ajas tagasi.

6.2.3. Mittejuhuslikud protsessid

6.2.3.1. Fraktalid

Fraktal koosneb lõputust arvust üksteise sees olevatest struktuuridest. Kui Eukleidilise geomeetria järgi saab objektil olla vaid täisarvuline dimensioon (punkti puhul 0, joone puhul 1, pinna puhul 2), siis fraktaalse objekti dimensioon ehk võime täita eukleidilist ruumi on murdarvuline. Tavalise kõverjoone pikkust on võimalik mõõta, fraktaalse joone pikkust mitte. Fraktaalne joon täidab mingi osa pinnast ja on dimensiooniga 1 ja 2 vahel. Fraktaalsel joonel olevate punktide kogumi dimensioon on 0 ja 1 vahel, Fraktaalsel pinnal olevate punktide dimensioon on 0 ja 2 vahel. Fraktaalne dimensioon on ideaaljuhul mõõtkavast sõltumatu, see tähendab, et jääb samaks igas mõõtkavas.

Mittejuhuslikud lihtprotsessid toimuvad kindlate juhuslikkust mittesisaldavate reeglite järgi.

Keeruka mustri moodustamiseks ei ole aga juhuslik komponent tingimata vajalik, fraktaalseid mustreid saab moodustada kindlate reeglite järgi. Fraktalid ei ole reeglina juhuslikud, kuid eri mõõtkavades fraktaalselt korduvana saab modelleerida ka juhuslikke protsesse (Halley 1996).

Fraktalitel on iseennast erinevates mõõtkavades kordav struktuur

Fraktaalse korrelatsiooni all mõistetakse fraktalite (objektide kuju) sarnasust teatud omavahelisel kaugusel. Fraktaalne korrelatsioon on seega autokorrelatsiooni üks esinemisvorm.

Fraktaalsuse ehk eri mõõtkavades iseendaga sarnasuse idee juured ulatuvad 17. sajandi teadusesse. Sõna fraktal on tuletatud ladinakeelsest sõnast *fractus*, mis tähendab murdunud. Fraktaalsetena on käsitletud rannajoont (Mandelbrot 1967), elupaiga mustrit, puuokste hargnemist, jäälilli klaasil, loomade liikumist ja liikide levikut (Cole 1995). Ülevaate fraktalitega seotud metoodikate kasutamisest ökoloogias on avaldanud Halley *et al.* (2004).

6.2.3.2. Rakk-automaat

Rakk-automaat (*cellular automata*) on dünaamilise modelleerimise meetod, mille puhul piksli järgmise ajahetke seisund tuleneb naaberpikslite ja sama piksli seisundist antud ajahetkel. Rakk-automaat käsitleb aega diskreetsena ja eeldab lähteseisundi ning üleminekutõenäosuste teadmist. Enamasti käsitletakse ka süsteemi seisundeid diskreetsetena, sest see lihtsustab üleminekutõenäosuste kasutamist. Rakk-automaat võib sisaldada ka juhuslikkust ja asukoha mõjusid (sobivuse pinda, atraktiivsuspinda, trendpinda, regionaalseid eeldusi) (Molovsky 1994, Loibl 2000). Klassikaline rakk-automaat kasutab vaid esimese astme naaberpikslite seisundit, kuid võimalikud on ka keerukamad mudelid.

Rakk-automaat tuletab koha väärtuse koha naabruses olevate pikslite väärtustest eelneval ajahetkel

Rakk-automaat on kasutust leidnud maakattemuutuste modelleerimisel (Feng *et al.* 2011), maastikustruktuuri ja elupaigakasutuse seoste teoreetilisel modelleerimisel (Hiebeler 2000) ja ruumiliselt ilmutatud ökoloogilistes mudelites, sealhulgas bioloogiliste populatsioonide dünaamika modelleerimisel, näiteks limuste *Macomona* paiknemine liivamadalikul (McArdle *et al.* 1997) ja pikaokkalise männi (*Pinus palustris*) rohtlas paiknemise modelleerimisel (Drake ja Weishampel 2001).

Uued suunad on rakk-automaadi ja geneetiliste algoritmide sidumine (Mitchell *et al.* 1996, Liu ja Feng 2010), mitmemõõtkavaline rakk-automaat (Hoekstra *et al.* 2008) ja pidev-automaat (*continuous automata*) (MacLennan 1990). Pidev-automaat ei käsitle aega ja ruumi diskreetsena, vaid pidevana. Pidev-automaadis on üleminekutõenäosuste asemel diferentsiaalvõrrandid.

6.2.4. Ruumistruktuuride stohhastiline modelleerimine

Erinevalt neutraalsetest maastikumudelitest genereerivad **ruumiliselt ilmutatud** (*spatially explicit*) maastikumudelid mustrit mudelisse lülitatud protsesside alusel. Ruumiliselt ilmutatud stohhastilisi ehk juhuslikkust sisaldavaid mudeleid on ökoloogias senini suhteliselt vähe kasutatud. Pikemad stohhastilise modelleerimise traditsioonid on inseneriteaduses, meteoroloogias ja okeanograafias (vt ka ptk 3).

Pindade puhul on modelleeritavaks tunnuseks **ruumiliselt muutuv omadus** (*spatially distributed*

attribute), näiteks niiskuse sisaldus mullas. Kõigis ruumipunktides ei ole võimalik seda mõõta. Mõõtmisi tehakse vaid proovikohtades. Enamasti on muutuja väärtuse hinnang vaja anda kogu uurimisala kohta. Seda saab teha interpoleerimismeetodite abil või väärtuste paiknemismustrit jäljendades. Interpoleerimismeetodid ja juhuslik jäljendus annavad samadest lähteandmetest erineva tulemuse (tabel 11).

Tabel 11. Empiirilise pinna, interpoleerimisimudeli ja genereeritud pinna parameetrite võrdlus. Rasvases kirjas on näitajad, mis erinevad interpoleerimismudelil ja stohhastilisel jäljendusel.

Empiirilised andmed	Interpoleerimismudel	Stohhastiline jäljendus
Muutuja mõõdetud väärtused üksikutes punktides	Muutuja arvutatud väärtused mõõdetud punktide vahelisel alal	Muutuja genereeritud väärtused mõõdetud punktide vahelisel alal
Vaatluste sagedusjaotus	Interpoleeritud väärtuste sagedus-jaotus (suurte ja väikeste väärtuste sagedus on väiksem)	Genereeritud väärtuste sagedusjaotus
Vaatluste keskväärtus	Interpoleeritud väärtuste keskväärtus	Genereeritud väärtuste keskväärtus
Vaatluste dispersioon	Interpoleeritud väärtuste dispersioon (< vaatluste dispersioon)	Genereeritud väärtuste dispersioon
Vaatlus-variogramm	Hinnangute variogramm (väiksem semidispersioon väike-matel vahemaadel)	Genereeritud väärtuste variogramm

Oluline on, et interpoleerimismeetodid prognoosivad õigesti vaid keskväärtust, mitte väärtuste varieeruvust. Interpoleeritud prognoosi ruumiline varieeruvus on väiksem kui lähteandmetel, stohhastiliselt genereeritud väärtuste statistikud on aga empiiriliste andmetele lähedased. Stohhastiline mudel suudab prognoosida muutuja detailset (suuremõõtkavalist, peenstruktuurset) varieeruvust, prognoosides seejuures uuritava muutuja kõrgete ja madalate väärtuste piirkondi sama hästi kui interpoleerimismeetodid, mis varieeruvust prognoosida ei suuda. Mitmed meetodid kombineerivad interpoleerimist ja etteantud jaotusega juhuslikkust.

Interpoleerimisel saadud väärtused on vähem varieeruvad kui mõõdetud väärtused

Stohhastilist modelleerimist tuleks interpoleerimisele eelistada alati, kui lisaks keskväärtusele on oluline tõepäraselt kaardistada ka väärtuspinna ruumilist struktuuri ja väärtuste varieeruvust. Oluline on siiski tähele panna stohhastiliste simulatsioonide identsena mittekorratavust. Igal stohhastilisel simulatsioonil saadakse erinev tulemus ka siis, kui lähteandmed ja simulatsiooni tingimused on täpselt samad. Stohhastilise simulatsiooni tulemusi nimetatakse seetõttu näideteks või protsessi üksik-realisatsioonideks. Erinevate stohhastiliste mudelite võrdlemiseks tuleb mustreid palju kordi genereerida. Kui eesmärgiks on hinnata vaid juhusliku komponendi tugevust, ei ole põhjust stohhas-tilisi pindu genereerida. Juhusliku komponendi osa saab hinnata ka variogrammi eheda efekti alusel.

Stohhastilise jäljenduse iga üksikkordus annab erineva tulemuse

Stohhastilised simulatsioonid annavad lisaks väärtuspinna ja selle ruumilise varieeruvuse prognoosile ka prognoosi usaldusväärsuse hinnangu. Kui teame ühte või teist kriitilist piiri ületava väärtuse ühes või teises kohas esinemise tõenäosust ja teame ühe või teise väärtuse või eksimuse

hinda, siis saame hinnata ka prognoosiga kaasnevaid riske. Näiteks mäenduses on maagi paiknemise prognoosi varieeruvusel kõrge nii majanduslik kui ka keskkonnakaitseline hind.

Juhuslikke interpoleerimispindu saab genereerida kasutades variogrammi mudeli järgi hinnatud ehedat hajuvust. Viimase alusel lisatakse igas prognoositavas punktis kriginguga interpoleeritud prognoosile juhuslik komponent. Juhusliku komponendiga interpoleerimispindu korduvalt genereerides saadakse igas punktis lisaks parimale hinnangule ka selle hinnangu oodatav jaotus. Sellest omakorda on lihtne tuletada interpoleerimise usalduspiire. Variogrammi mudelit kasutavaid pinna stohhastilise moodustamise meetodeid nimetatakse **geostatistilisteks** simulatsioonimeetoditeks.

Geostatistiline jäljendus kaasab muutuja ruumilist pidevust ja väärtuste jaotust

Variogrammi mudelit ja autokorrelogrammi saab kasutada ka pikslikaupa arvutatud prognoosile ruumilise autokorrelatsiooni lisamiseks (de Bruin 2000). Suurima tõepära (ptk 1.2 ja 3.6.1) hinnangul prognoositud maakatteklasside ruumistruktuur ei vasta enamasti reaalsusele. Maakatteklassi tõenäosust pikslites, mille lähedal on selle maakatteklassi esinemise tõenäosus suur, tuleks realistlikuma pildi saamiseks tõsta. Kui palju ja millises ruumilises ulatuses maakatteklasside esinemise tõenäosusi korrigeerida, sõltub iga maakatteklassi ruumilisest autokorrelatsioonist.

6.2.4.1. Tõenäosusvälja jäljendus

Tõenäosusvälja jäljendus (*probability field simulation*) (Srivastava, 1992) on üldine metoodika võrdselt tõenäoliste kujutiste moodustamiseks, mis ei ole muutuja tüübi ega teoreetiliste eeldustega piiratud. Tõenäosusvälja jäljendus võimaldab lisateavet paindlikult kasutada. Selle algoritm koosneb kahest osast:

- lokaalsete tõenäosusjaotuste määramine igas hinnatavas kohas,
- juhuslike väärtuste genereerimine lokaalsetest jaotustest vastavalt väärtuste tõenäosusele selles kohas ja arvestades naabruses olevate tõenäosusjaotuste korreleerumist.

Lokaalsed tõenäosusjaotused esindavad seejuures ka kõigi teiseste faktorite mõjusid. Meetod on arvutuslikult suhteliselt kiire, kuid võib tõenäosusi täpsustavate täiendavate andmekihtide kasutusel anda hälbinud hinnanguid ning ei arvesta lähedal olevate vaatlustulemustega (Pyrcz ja Deutsch 2001, de Almeida 2010).

6.2.4.2. Normaaljaotusele tuginevad jäljendused

Normaaljaotusele tuginev jäljendus (*gaussian simulation*) eeldab, et jäljendatava muutuja väärtused on oodatavalt normaaljaotusega ja juhuslikkust sisaldavas jäljenduses peab etteantud parameetritega jaotus säilima. Mitmemõõtmeline normaaljaotusele tuginev jäljendus (*plurigaussian simulation*) jäljendab korraga mitut muutujat või nomiaalse muutuja kategooriat. Ruumiandmete normaaljaotusele tuginevad jäljendused lähtuvad variogrammi mudelist, väärtuse ootus vaatluspunktide vahelisel alal saadakse kriging-interpoleerimise teel. Interpoleeritud väärtusele lisatakse juhuslik komponent.

Teiste muutujate mõju saab jäljendusmudelisse kaasata kas järelkorrektuurina või pidevalt järjestikku iga koha väärtuse arvutamise järel. Normaaljaotusele tuginev **jäljendus järelkorrektuuriga** (*gaussian simulation plus posterior conditioning*) kasutab järelkorrektuuri vastavalt teiste muutujate väärtustest sõltuvatele tinglikele tõenäosustele. Viimane eeldab, et ka

korrigeeritud (tinglikud) jaotused vastavad normaaljaotusele. Erinevalt järjestikusest jäljendusest rakendatakse järelkorrektuuri algse jäljenduse valmimise järel, mitte hinnangulise väärtuse arvutamise järel igas jäljendatava ala kohas.

Jäljendust mõjutavaid faktoreid kaasatakse kas järelkorrektuurina või jäljenduse arvutamise käigus

Kärbitud normaaljaotusele tuginev jäljendus (*truncated gaussian simulation*) seob nomiaalse muutuja iga kategooria esinemistõenäosuse normaaljaotuse tihedusfunktsiooni aluse pinnaga mingite piirväärtuste vahel.

Mitmemõõtmeline kärbitud normaaljaotusele tuginev jäljendus (*truncated plurigaussian simulation*) on mitmemõõtmelise ja kärbitud jäljenduse kooskasutus, mis ei nõua modelleeritava muutuja kategooriate järjestamist. Igat kategooriat käsitletakse omaette muutujana.

6.2.4.3. Järjestikused jäljendused

Järjestikune jäljendus (*sequential simulation*) on geostatistiline stohhastiline algoritm, mille käigus sobitatakse väärtuste paiknemismustrit järjestikku uuritava ala igas kohas, kasutades eelmistes kohtades arvutatud väärtusi, otseseid vaatlusandmeid ja kohast sõltuvaid tunnuseid. Arvutatud väärtuste jaotus on tinglike tõenäosuste jaotus just kohast sõltuvate tunnuste suhtes.

Järjestikuses jäljenduses arvestatakse eelmistes kohtades saadud hinnanguid

Järjestikuse jäljenduse eesmärk on moodustada ühe kõige tõepärasema hinnangukaardi asemel võrdselt võimalike väärtuspindade komplekt, mis vastab vaatlustulemustele näidates samal ajal muutuja väärtuste realistlikku paiknemismustrit. Järjestikune jäljendus võimaldab esitada jäljendatud tunnuse hinnangute ebakindlust.

Järjestikuse jäljenduse algoritm on järgmine:

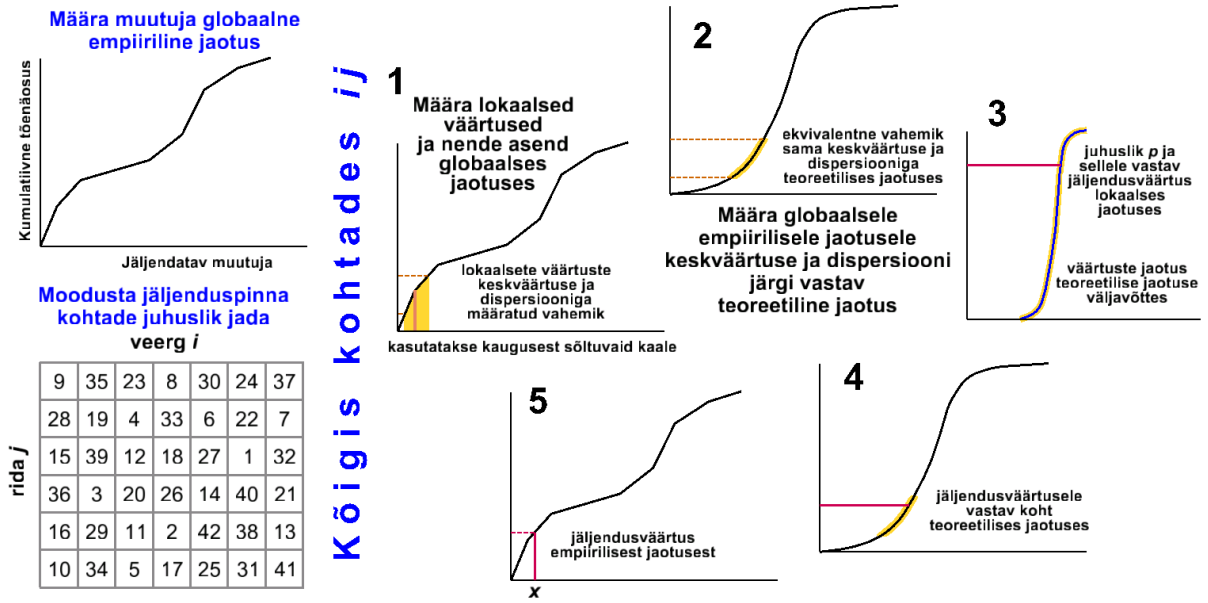
- moodusta jäljendatava pinna arvutuskohtade võrgustiku sõlmede (kohtade) järgnevuse juhuslik jada ja sea $i = 1$;
- määra muutuja lokaalne kumulatiivne tinglik tõenäosusjaotus kohas i originaalandmete ja eelmistes kordustes saadud väärtuste suhtes;
- võta täpsustatud kohalikust jaotusest üks juhuslik väärtus;
- pöördu järgmise koha juurde ja jätka punktist 2, kuni kõik kohad on punktis määratud järjekorras läbitud.

Normaaljaotusele tugineva järjestikuse jäljenduse (*sequential gaussian simulation*) puhul saadakse kohalik jaotus kriging-interpoleerimise käigus hinnatud lokaalse keskmise ja lokaalse standardhälbe järgi. Seega käsitletakse tinglikke jaotusi normaaljaotuse kohaselt. Kui väärtused ei ole normaaljaotusega, siis tuleks neid teisendada.

Järjestikune indikaatorjäljendus (*sequential indicator simulation*) kasutab indikaatorkrigingut, seega teisendab muutuja väärtusvahemikesse.

Järjestikuse otsese jäljenduse (*direct sequential simulation – DSS*) ja **koosjäljenduse** (*direct sequential cosimulation – CODSS*) eelis normaaljaotusele tugineva jäljenduse ees on normaaljaotuse piirangust vabanemine. Otsese jäljenduse algoritmi esitas Soares (2001). Selle kohaselt ei defineerita

lokaalset tinglikku tõenäosusjaotust, vaid võetakse väärtusi globaalse tõenäosusjaotuse lokaalselt määratud vahemikust. Globaalse jaotuse vahemiku määramisel viiakse väärtuste lokaalne ja globaalne jaotus omavahel vastavusse keskväärtuse ja hajuvuse abil. Keskväärtust ja hajuvust ei kasutata tõenäosuste määramisel ei lokaalses ega globaalses jaotuses (joonis 6-3).



Joonis 6-3. Otsese järjestikuse jäljenduse algoritm. Kollane taust tähistab ühte ja sama väärtusvahemikku. Jäljendatud väärtus on juhuslik, kuid seotud jäljendatava muutuja lokaalse ja globaalse jaotusega.

Otseses jäljenduses kasutatakse lokaalseid keskväärtusi ja dispersioone jäljendatava väärtuse võtmiseks globaalsest empiirilisest jaotusest

Koosjäljendus võimaldab korraga jäljendada mitut üksteisest sõltuvat muutujat. Muutujaid jäljendatakse ükshaaval kindlas järjekorras, kasutades iga kord eelmiste jäljenduste tulemusi. Koosjäljenduse tarkvara on kirjeldanud Hansen ja Mosegaard (2008, 2011).

6.2.4.4. Mitme-punkti jäljendus

Mitme-punkti jäljendus (*multiple-point simulation*) kasutab jäljendusmudelite loomisel mitme-punkti geostatistikat (Strebelle 2002) (ptk 5.3.7). Õpetusandmetest arvutatakse kahe või enama mõõtmelisi struktuure kirjeldavad statistikud ja püütakse neid struktuure jäljendusmudelites sarnasel kujul taasluua. Suure arvutusmahu tõttu on suurtel aladel metoodikat tülikas kasutada.

Mitme-punkti jäljendust saab kombineerida statistilise mustrituvastuse meetoditega (Caers 2001). Üks variant on seejuures käsitleda 2D või 3D rasterkujutises olevate mustrite näidiseid objektidena, mitte piksligruppidena (Arpat ja Caers 2007).

6.2.4.5. Jäljendatud karastamine

Karastamine tähendab metallurgias metalli tugevdavat kuumutamist ja jahutamist, mille käigus metalli aatomid leiavad kristallvõres stabiilsema asendi. **Jäljendatud karastamise** (*simulated annealing*) algoritmid loovad esialgsest kanditaat-lahendusest juhuslikul määral muudetud teisendusi (analoogne metalli kuumutamisega) ja kontrollivad iga teisenduse sobivust õpetusandmetega. Parim

muudetud lahend saab uueks kanditaatlahendiks. Protsessi käigus vähendatakse lubatud muutuste määra (analoogne temperatuuri alandamisega). Jäljenduse karastamine on ühe fikseeritud prognoosi juhuslik lõdvestamine, mis aitab vältida lahendi otsimise protsessi sumbumist lokaalselt hea, kuid mitte parima lahendi juurde. Meetodi kirjeldasid Kirkpatrick *et al.* (1983).

Karastamine on struktuuri juhuslik lõdvestamine leidmaks veelgi sobivamat struktuuri

Esimesel pilgul tundub karastamise algoritm lootusetult ebaefektiivne. Olemasolevast parema lahendi leidmiseks võib kuluda miljoneid juhuslikkust lisavaid katseid. Õnneks saab ruumimustri jäljendamisel parimat lahendit otsida iga koha ümbrusest eraldi, mis võimaldab rakendada paralleelarvutust.

Jäljendatud ruumilise struktuuri karastamise puhul lisatakse juhuslikkust struktuuri lähestikku paiknevate osade (pikslite) asukohtade vahetamisega. Karastamise algoritm on juhitud protsess, mida määravad mitmed kasutaja valitud parameetrid (Deutsch ja Cockerham 1994).

6.2.4.6. Pikslivahetus

Pikslivahetuse algoritmis (*cell swapping*) genereeritakse algul suvalise (enamasti juhusliku) struktuuriga lähtepind, milles kategooriate vahekord vastab etteantule. Etteantud parameetritega ruumimustri variantide moodustamiseks vahetatakse paariviisi juhuslike pikslite asukohti. Kui vahetuse järel vastab pinna struktuur etteantud parameetritele paremini kui enne vahetust, siis vahetus teostub. Meetod võimaldab jäljendada igasugustele parameetritele vastavaid pinnastruktuure. Etteantud parameeter võib olla ka kompleksne, näiteks ruumilise autokorrelatsiooni korrelogramm. Meetodit on kasutanud vigade analüüsis (Goodchild 1990, Veregin 1997, Ehlschlaeger 2000).

6.2.4.7. Autologistiline mudel

Genereeritavaks väärtuspinnaks võib olla mingi liigi esinemise tõenäosuse kaart, mida tavapäraselt modelleeritakse logistilise regressioonimudeliga. Asukohateavet saab logistilisse mudelisse kaasata täiendava tunnusena, mille väärtus sõltub sama tunnuse väärtustest ümbruses. Sellist tunnust, mis samaaegselt esineb nii funktsioon- kui ka argumenttunnusena nimetatakse **isekaasuvaks muutujaks** (*autocovariate*). Autokovariatsiooni lisamisel mudelisse muutub ka mudeli nimi – logistiline regressioonimudel saab autokovariatsiooni lisamisel nimeks **autologistiline mudel.** Autologistilise mudeli valikul on suhteliselt raske otsustada autokovariatsiooni mõju funktsiooni ja ruumilise ulatuse üle.

Autologistilist mudelit saab andmetele sobitada iteratiivselt, näiteks Gibbsi sampleriga (ptk 6.1.5) järgmise algoritmi abil, mida kasutasid Augustin *et al.* (1996).

1. Prognoosi logistilise mudeli abil liigi esinemise tõenäosus igas pikslis.
2. Arvuta esinemistõenäosuse autokovariatsioon iga piksli kohas.
3. Sobita autologistiline mudel näidisalade andmetele (leia näidisalade andmetel parim autologistiline mudel).
4. Arvuta autologistilise mudeli järgi iga piksli kohas korrigeeritud esinemistõenäosus.
5. Korda järgnevaid samme teatud arv kordi või piisavalt hea tulemuseni (Gibbsi sampler):
   a) arvuta kõigi pikslite autokovariatsioon,
   b) sobita autologistiline mudel näidisalade andmetele,

c) vali juhuslik lähtepiksel,

- arvuta esinemistõenäosuse autokovariatsioon selle kohas ümbruskonnas olevate pikslite järgi,
- korrigeeri esinemistõenäosust

d) võta järgmine juhuslik korrigeerimata piksel, kuni esinemistõenäosus on korrigeeritud kõigi pikslite kohas.

6. Salvesta liigi esinemistõenäosused kogu prognoositaval alal.

Tarkvara

Ruumiliste struktuuride jäljendamise kõige tuntum vabavara on olnud GSLIB (*Geostatistical Software Library*, http://www.gslib.com). Selle nime all vaba kasutusõigusega tarkvara on 15 aastat arendatud Stanfordi ülikoolis (Deutsch ja Journel 1998). Uuemal ajal on Stanfordi ülikoolis arendatav geostatistika tarkvara pakutud üldnimega SGeMS (*Stanford Geostatistical Modeling Software*) (http://sgems.sourceforge.net) (Remy *et al.* 2009, Liu 2006, Liu ja Journel 2009).

gOcad (http://www.gocad.org) (*Geological Object Computer Aided Design*) on 1989. aastal Nancy ülikoolis professor Jean-Laurent Mallet algatatud projekti ja selle raames loodud geoloogiliste struktuuride modelleerimise ja visualiseerimise tarkvara algne nimi. Praeguseks on tarkvara arendamise ja müügi õigused üle antud firmale Paradigm Geophysical (http://www.pdgm.com).

Ruumiliste struktuuride jäljendamise tarkvara Plurigau (Dowd *et al.* 2003) kood on saadaval aadressil http://www.iamg.org/documents/oldftp/VOL29/v29-02-02.zip.

Uurimused

Li *et al.* (1993) jäljendasid erinevate raiekavade järgi tekkivat maastiku fragmenteeritust.

Augustin *et al.* (1996) kasutasid naaberpikslite kaalu määramisel kauguse pöördväärtust. See tähendab piksli j mõju kaal w_{ij} piksli i jaoks on $w_{ij} = 1/h_{ij}$, kus h_{ij} on vahemaa pikslite i ja j vahel. Autokovariatsiooni arvutasid mainitud autorid naaberpikslite kaalutud keskmisena

$$auto\,\mathrm{cov}_i = \frac{\sum_{j=1}^{k_i} w_{ij}\, p_j}{\sum_{j=1}^{k_i} w_{ij}}, \qquad [6\text{-}3]$$

kus p_j on liigi esinemise tõenäosus, i on lähtepiksli indeks, j on naaberpiksli indeks, w_{ij} on piksli j kaal piksli i jaoks ning k_i on piksli i naaberpikslite arv.

Dungan (1998) võrdles stohhastilist tõenäosusvälja jäljendust regressioonimudeliga ja kokrigingguga taimkatte ohtruse hindamisel. Oluline erinevus on eeldatavate prognoosivigade jaotuses: regressioonimudeli puhul tuletatakse need algandmete väärtustest, krigingu puhul lähedal olevate väärtuste paiknemisest, jäljenduse puhul saadakse varieeruvuse kujutis, mis arvestab nii väärtuste jaotust kui ka ruumilist paiknemist.

Franco *et al.* (2006) lisasid mulla reostuse interpoleeritud kaardile otsese järjestikuse jäljenduse abil saadud juhusliku komponendi, et saada realistlikum reostuse paiknemismuster. Carvalho *et al.* (2006) kasutasid järjestikust koosjäljendust puuliikide katvuse kaardistamisel. Kriginginterpoleerimise kaasamine tagas ruumiliselt pidevama tulemuse. Koosjäljendus võimaldas arvestada puuliikide koosesinemise seaduspärasusi.

Durão *et al.* (2010) kaardistasid otseste järjestikuste jäljenduste abil maastikupõlengute riski Portugalis.

J.A. de Almeida (2010) võrdles nomiaalse muutuja jäljendusmeetodeid. Tõenäosusvälja jäljendus ja järjestikune indikaatorjäljendus on suhteliselt lihtsa teoreetilise taustaga ja arvutuslikult efektiivsemad kui jäljenduse karastamist sisaldavad meetodid.

6.2.5. Detailiseerimine

Kujutise või prognoositud väärtuste pinna detailsemaks muutmine ehk **detailiseerimine** ehk lähem esitus (*downscaling*) on lähedane interpoleerimisülesandele. Erinevus on lähteandmetes – ruumilise interpoleerimise lähteandmeteks on kohtvaatlused, ajalise interpoleerimise puhul ajavahemikega vaatlusandmed. Ruumilise detailiseerimise puhul on esmane andmeallikas üldisem väärtuspind, mis on küll enamasti esitatud korrapärase suuresilmalise võrgustikuna, aga sisuliselt käsitletakse seda väärtuspinnana. Ajalise detailiseerimise puhul lähtutakse pidevana käsitletavast üldistatud aegreast. Seega detailiseerimine eeldab mingi üldisema hinnangu olemasolu.

Terminiga *downscaling* tähistatakse ka kaalus alla võtmist ja hinna alandamist. Termin *upscaling* tähistab klimatoloogias üldistatud trendide tuletamist kohalikest vaatlusandmetest (üldistamine ja ekstrapoleerimine) või kaudsetest andmetest (indikatsioon). Pilditöötluses tähistab *upscaling* madala lahutuse signaali või kujutise teisendamist detailsemasse vormingusse.

Detailiseerimine võib tugineda kas ainult objektide ja pikslivääртuste paiknemise ja kuju seaduspärasustele olemasolevas robustses kujutises või siis detailsema mõõtkavaga lisaandmetele. Mõlemal juhul on detailsema prognoosipinna saamiseks tarvis moodustada kas mingid reeglid või mudelid.

Detailiseerimise meetodid võib jagada kolme tüüpi:

- tüüpjuhtude klassifitseerimine ja kõige sarnasema näidise rakendamine,
- regressioonimudelid,
- protsessi simulaatorid (Wilby *et al.* 2004). Viimased võivad jäljendada nii lokaalseid protsesse kui ka juhuslikkust.

Detailiseerimine on levinud meteoroloogias ja klimatoloogias. Globaalsed ja suuremat piirkonda katvad kliima- ja ilmamudelid ei arvesta kohaliku maastiku üksikasju ja nende väljund on suurema võrgusammuga kui üksikud vihmapilved. Kohaliku ilma ennustamiseks detailiseeritakse suure mudeli väljundit, võttes arvesse kohalikke andmeid. Näiteks ümbruse metsasusest, veekogude lähedusest või reljeefist tingitud mõju. Detailiseerimine on oluline ka liikide leviku kaardistamisel kasutades globaalseid klimaatilisi andmeid, mis on paarisajameetrise võrgustikusammuga. Liikide levik ja elupaigasobivus sõltuvad detailsemas skaalas palju ilmastikuoludest.

Uurimused

Detailiseerimisprobleemidega on palju tegeletud kaugseires ja pilditöötluses. Aplin *et al.* (1999) kasutasid rasterkujutise detailiseerimisel põhikaardi alade piire. Lisaks töötlemata rasterkujutisele on detailiseerimiseks kasutatud ka eelnevalt klassifitseeritud kujutist. Atkinson (1997) kasutas piksliosade klassifikatsiooni täpsustamiseks klasside ruumilise paiknemise seaduspärasusi. Gavin ja Jennison (1997) kasutasid klassitõenäosuste stohhastilist mudelit, mis eelistas kompaktsema kujuga eraldisi. Verhoeye ja De Wulf (2000) lahendasid detailiseerimisülesannet lineaarse optimeerimise vahenditega. Tatem *et al.* (2001, 2002) kasutasid pikslisiseste üksuste paiknemise modelleerimiseks Hopfieldi tehisnärvivõrku (Hopfield 1984).

Üldistatumas skaalas õpetusandmete järgi detailsemas skaalas hinnanguid andvaid liikide

levikumudeleid on koostanud Collingham *et al.* (2000), Barbosa *et al.* (2003) ja Araújo *et al.* (2005). Cowley *et al.* (2000) uurisid seoseid liblikaliikide leiuandmete ning 500 × 500 m ruutudes oleva elupaiga hulga ja kvaliteedi järgi ennustatud arvukuse vahel. Liikide esinemine oli elupaiga hulgaga erineval määral seotud. Uuriti eelkõige liikide seotust elupaigaga, aga samalaadse metoodikaga saaks ka näiteks 10 × 10 km suuruse ruuduga levikuatlaste andmeid detailiseerida.

Tian *et al.* (2005) detailiseerisid maakatte kaartide järgi Hiina ja Linard *et al.* (2010) Aafrika rahvastiku paiknemist.

Neumann *et al.* (2009) kaardistasid kariloomade tiheduse paiknemist Kesk-Euroopas, detailiseerides regionaalseid statistilisi andmeid Corine maakattekaardi, kliima ja kõrgusmudeli abil.

Euroopa Liidu rahastatud projekti ENSEMBLES tulemusel loodud kliimatunnuste detailiseerimise veebiportaali kirjeldavad Cofiño *et al.* (2007).

6.2.6. Üldistamine

Selle alapeatüki kirjutas **Tiiu Kelviste**.

Üldistamine ehk generaliseerimine ehk *up-scaling* tähendab objektide hulga vähendamist või vormi lihtsustamist mõõtkava või resolutsiooni muutmisel. Üldistamine on vajalik nii andmemahtude vähendamiseks kui ka väljundi – kaardi või graafiku loetavuse parandamiseks.

Diskussioonid generaliseerimise teemal ulatuvad tagasi 19. sajandi algusesse. Kontseptuaalselt võib generaliseerimist käsitleda üsna mitmeti. Ratajski (1967) jagas generaliseerimise kaheks:

- kvantitatiivne generaliseerimine kui järkjärguline kaardi info vähendamine lähtuvalt mõõtkava muutusest,
- kvalitatiivne generaliseerimine kui kaardi sümboolika abstraktsemaks teisendamine.

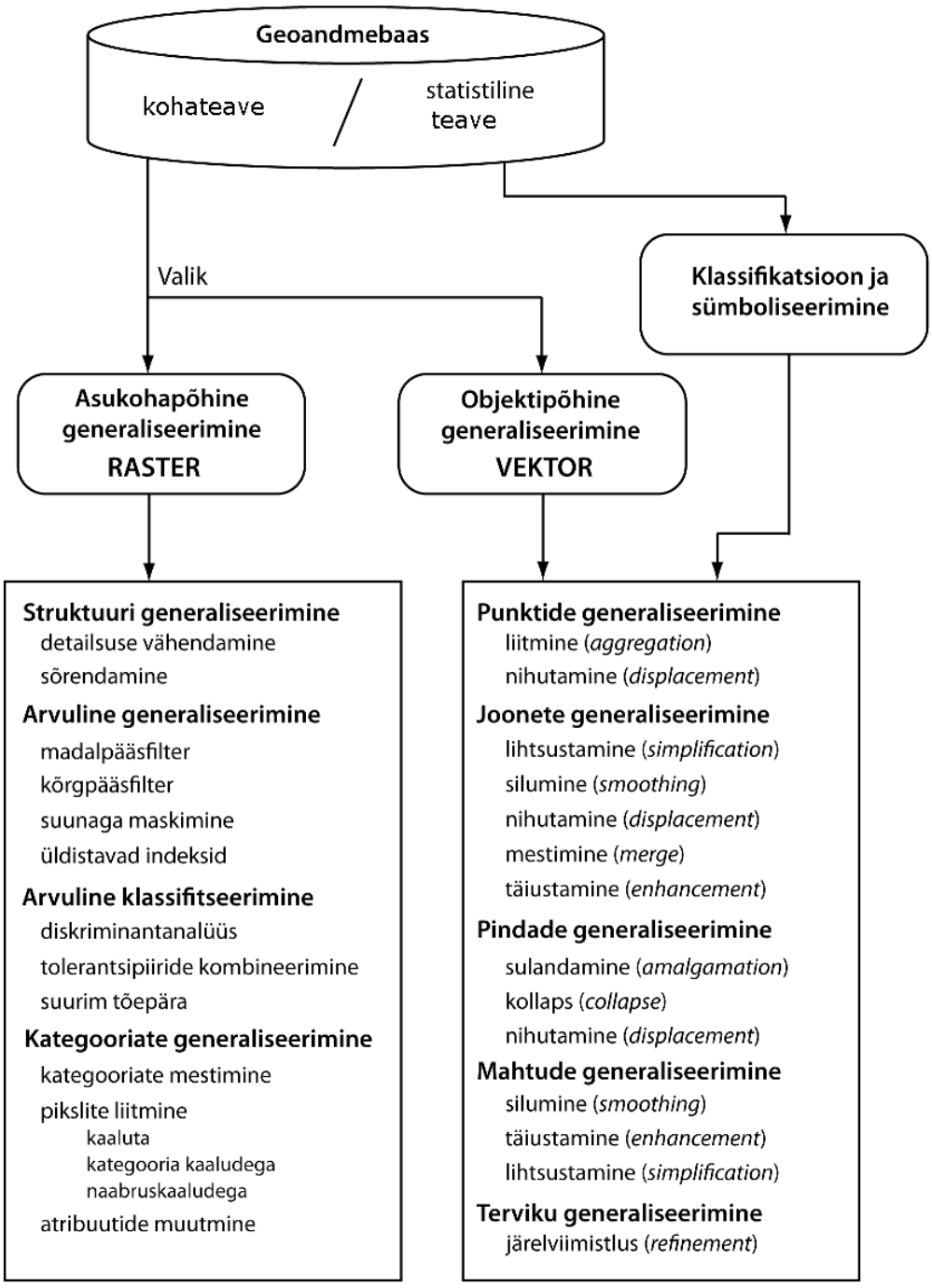
Morrison (1974) lähtus generaliseerimisprotsessi defineerimisel neljast sel ajal väljatöötatud üldistamise viisist milleks olid: lihtsutamine, klassifitseerimine, sümboliseerimine ja induktsioon. Morrisoni järgi käivitab generaliseerimisprotsessi klassifitseerimine, mille käigus kartograaf teadlikult valib kaardilt teatud hulga elemente, mille kujutusviisi ta lihtsustab ja muudab sümboleid, saades induktsiooni tulemusena uue üldistatud kaardipildi. Brassel ja Weibel (1988) olid ühed esimesed, kes vaatasid generaliseerimisprotsessi andmebaasipõhiselt, uurides automaatse generaliseerimise võimalusi. Generaliseerimisprotsess hõlmas andmete struktuuri eristamist, protsessi eristamist, protsessi modelleerimist, protsessi teostamist ning lõpptulemusena andmete kuvamist. Digitaalsete protsesside juures eristasid nad statistilist ja kartograafilist generaliseerimist. Statistiline generaliseerimine on vaadeldav filtreerimisprotsessina, mille põhieesmärk on andmete ühildatavus ja statistiline analüüs. Seevastu kartograafilise generaliseerimise eesmärk on modifitseerida kaardi struktuuri ja väljanägemist, et parandada loetavust.

Başaraner (2002) on lähtuvalt geoinfosüsteemide rakendustest jaganud generaliseerimisprotsessi kolme klassi:

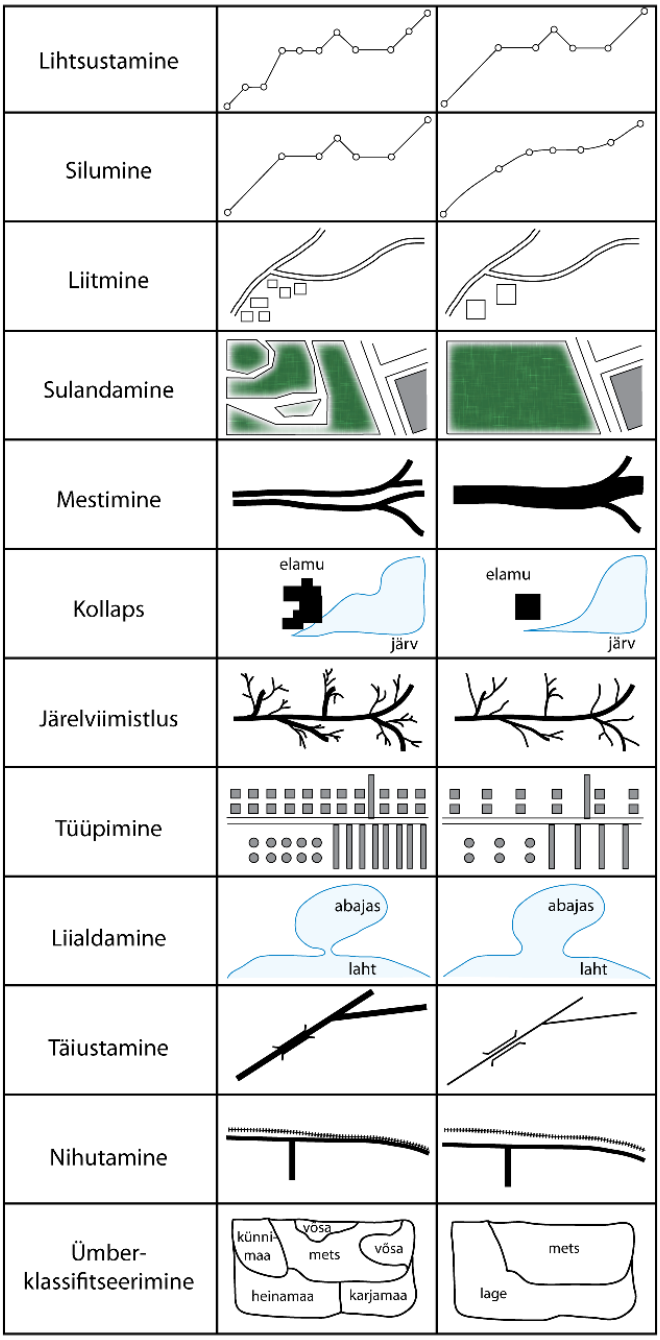
- objekti generaliseerimine kui üldistamise protsess andmebaasi defineerimise ja loomise etapis, kus toimub kogutud andmete valik ja reduktsioon,
- mudeli generaliseerimine kui matemaatiliselt formuleeritud üldistamine tagamaks andmete kontrollitud vähendamise mitme-mõõtkavaliste (*multi-resolution*) andmebaaside operatiivseks kasutamiseks,
- kartograafiline generaliseerimine kui andmete graafiline sümboliseerime nähtuste visualiseerimiseks.

Harjumuspäraseim viis andmetöötlejate ja teiste kasutajate jaoks on käsitleda generaliseerimisprotsessi lähtuvalt andmetele rakendatavast meetodist. Generaliseerimise meetodid pärinevad suuresti kartograafide käsitööprotsessidest, mida on aegade jooksul korduvalt matemaatiliste vahenditega täiendatud. Väljatöötatud meetodid on loodud samaaegselt nii ruumilisteks kui ka atribuut-teisendusteks.

McMaster ja Shea (Shea ja McMaster 1989, McMaster ja Shea 1992) poolt väljatöötatud raamistiku järgi rakendatakse generaliseerimist kas nähtuse geograafilisele asendile või atribuutandmetele, seda olenemata andmemudelist (vektorandmed või rasterandmed) (joonis 6-4 ja 6-5).



Joonis 6-4. Generaliseerimismeetodite raamistik (McMaster ja Shea 1992, muudetult).



Joonis 6-5. Enamlevinud kartograafilised generaliseermisvõtted (Shea ja McMaster 1989, muudetud).

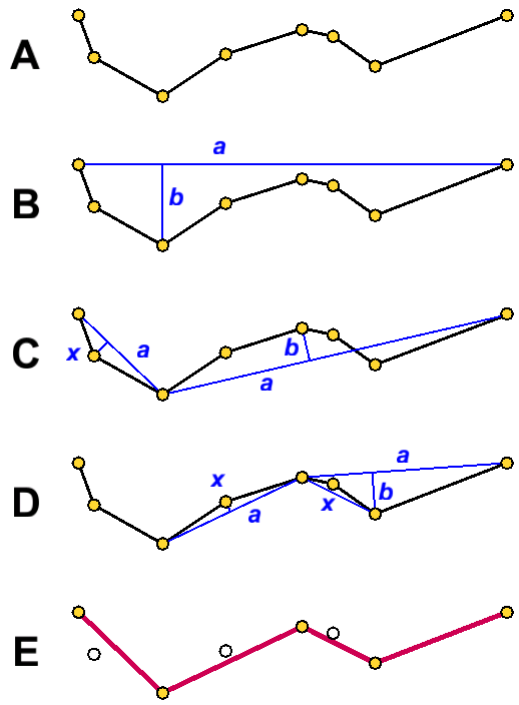
Seda, kas kaardi mõõtkava vähendamise tulemusena peaks üldistamist rakendama või mitte, võiks hinnata läbi kuue situatsiooni (Shea ja McMaster 1989):

- kuhjumine (*congestion*) on olukord, kus kaardil on kujutatud nähtused on liiga tihedasti koos,
- kattumine (*coalescence*) on olukord, kus leppemärgid/nähtused hakkavad üksteist katma,
- konflikt (*conflict*) on olukord, kus kaks või enam nähtust satuvad omavahel sisu ja/või kujutusviisi poolest vastuollu,
- komplikatsioon (*complication*) olukord, kus sobiva generaliseerimismeetodi valikut mõjutab rohkem kui üks faktor (andmete sisu, valik, tolerants),
- seosetus (*inconsistency*) viitab ebaühtlaselt rakendatud generaliseerimismeetoditele ühe kaardi ala piires, mille tõttu võib esineda nihkeid objektide vahel,
- vaevumärgatavus (*imperceptibility*) olukord, kus teatud objektid on muutnud alamõõduliseks.

Iga üldistamismeetod on olemuslikult kas **sõltumatu** (*independent operator*) või **kontekstitundlik** (*contextual operator*) (Bader 2001). Sõltumatu meetod käsitleb objekte ühekaupa, kontekstitundlikku meetodit saab rakendada mitmele objektile korraga. Iga meetodi poolt on kindlaks määratud teisendamiste lõpptulemus, mille saavutamiseks on kasutada erinev hulk algoritme, mida kasutatakse kindlas järjekorras. Kasutamiseks sobivad algoritmid on tugevasti seotud andmete struktuuri, esitusviisi ja mõõtkava muutusega.

Enim algoritme on väljatöötutud sõltumatute generaliseerimismeetodite jaoks sealjuures joonte geomeetria lihtsustamiseks (tuntuim on **Ramer-Douglas-Peuckeri algoritm**) (Ramer 1972, Douglas ja Peucker 1973) (joonis 6-6). Konteksti arvestavate meetodite vallas on enim tegeletud nihutamisalgoritmide arendusega, ülejäänud on alles algfaasides (Bader *et al*. 1999, Bader 2001).

Rasterandmete töötluses on pindade üldistamise põhimeetod lähestikku olevate pikslite liitmine ehk agregeerimine suuremateks piksliteks või regioonideks. Pikslite liitmisel tuleb uuele agregaatpikslile omistada mingi lähtepikslitest sõltuv väärtus. Kasutatakse näiteks keskmise piksli väärtust, keskväärtust või mediaani. Nende meetodite kasutamise tulemus sõltub piksli suuruse ja ruumilise autokorrelatsiooni ulatuse vahekorrast. Kui piksel on ruumilisest autokorrelatsioonist väiksem, siis ruumiline muster säilub ja keskmise piksli kasutamine säilitab ka kujutise kontrasti. Suurema piksli korral annab keskmise piksli kasutamine suuri hälbeid ja keskväärtuse kasutamine ühtlaselt häguse tulemuse (Bian ja Butler 1999).



Joonis 6-6. Joone lihtsustamine Ramer-Douglas-Peuckeri algoritmi järgi. A – algne joon esiletõstetud nurkadega. B – joone algus ja lõpp säilitatakse igal juhul. Nende vahele konstrueeritakse sirge (*a*) ja mõõdetakse algse joone kõige kaugema nurga kaugus (*b*) sellest sirgest. Kui see kaugus on suurem kui etteantud piirväärtus, siis see nurk säilitatakse. C, D – jätkatakse sirgete (*a*) konstrueerimist säilitatud nurkade vahele ja algse joone nurkade vahemaa mõõtmist konstrueeritud sirgest. Säilitatakse algse joone nurgad, mis on piirväärtusest kaugemal (*b*) ja eemaldatakse nurgad, mis on piirväärtusele lähemal (*x*). E – lihtsustatud joon ning algse joone säilitatud ja eemaldatud nurgad.

6.2.6.1. Horemaatika

Horemaatikaks nimetatakse geograafiliste andmete skemaatilise modelleerimise meetodit, milles kasutatakse üldistatud ruumilisi üksusi ehk horeeme. Horeemid võeti kasutusele prantslasest kartograafi Roger Brunet poolt (Brunet 1980). Ta arendas välja süsteemi ruumiliste struktuuride ja protsesside kujutamiseks graafilise tõlgendamisskeemi abil. Sõna **horeem** (prantsuse keeles *chorème*) tuleneb kreekakeelsest sõnast *χώρα*, mis tähistab territooriumi või ruumi üldisemalt ja kreekakeelsele mõistele iseloomulikust sõnalõpust *-ημα*.

Brunet (1987) sõnul piisab horeemina esitatava mudeli loomiseks seitsmest kujundist:

- punkt (kujutamaks asukohta, poolsust),
- joon (kontakt, piir, sidemed jne),
- pind (ulatus, kuju),
- voog (liikumine, sümmeetria, intensiivsus),
- ühendus (sillad, ülekäigud, tunnelid),
- polaarsus (fookuspunkt),
- gradient (ebakorrapärasus, tõukumine, tõmbumine).

Kombineerides omavahel nelja vormi (punkt, joon, pind, võrgustik) ja seitset ruumilist nähtust (kontuur, muster, külgetõmme, kontakt, voolavus, dünaamika, hierarhia) eristas Brunet (1987) kokku 28 horeemi, mida ta nimetas ruumiliseks alfabeediks (joonis 6-7).

Brunet ruumimudelite väljakujundamist mõjutasid tugevasti nii prantsuse strukturalism, konstruktivism, süsteemi teooria, küberneetika kui ka semiootika. Seetõttu on teooria läbipõimunud pigem mõttelistest seostest kui matemaatilistest funktsioonidest. Unikaalne on see lähenemine selle poolest, et keskendub pigem ruumiliste protsesside ja mitte niivõrd struktuuri edasiandmisele (Haggett 2001). Brunet on ise öelnud, et eesmärgiks pole aluskaarti üle generaliseerida, vaid pigem muuta seda nii, et selle pinnal tõuseks olulisemaks edasiantavate nähtuste struktuur ja dünaamika (Brunet 1987).

Olemuslikult viitab horeem väikseimale mõttelisele ühikule, mida graafiliselt esitatakse. Tainz (2001) on horeeme kirjeldanud ka kui vahendit keeruliste ruumisituatsioonide struktuurseks ja sümboolseks esitamiseks. Laiemas kontekstis võib horeeme käsitleda teemakaartide jõuliselt generaliseeritud ja seetõttu äärmiselt abstraktse tüübina (Ormeling 1992). Vajalikke protsesse lahutatakse osadeks senikaua, kuni neid saab kujutada võimalikult väheste graafiliste vahenditega (joonis 6-8). Horeemkaardi eeliseks tavapärase teemakaardi ees on esimese tugevam seos geograafilise ruumi mentaalse mudeliga, mis inimesel kaarti vaadates tekib ning seetõttu on ta võimeline kaarti kiiremini tõlgendama (Reimer ja Fohringer 2010).

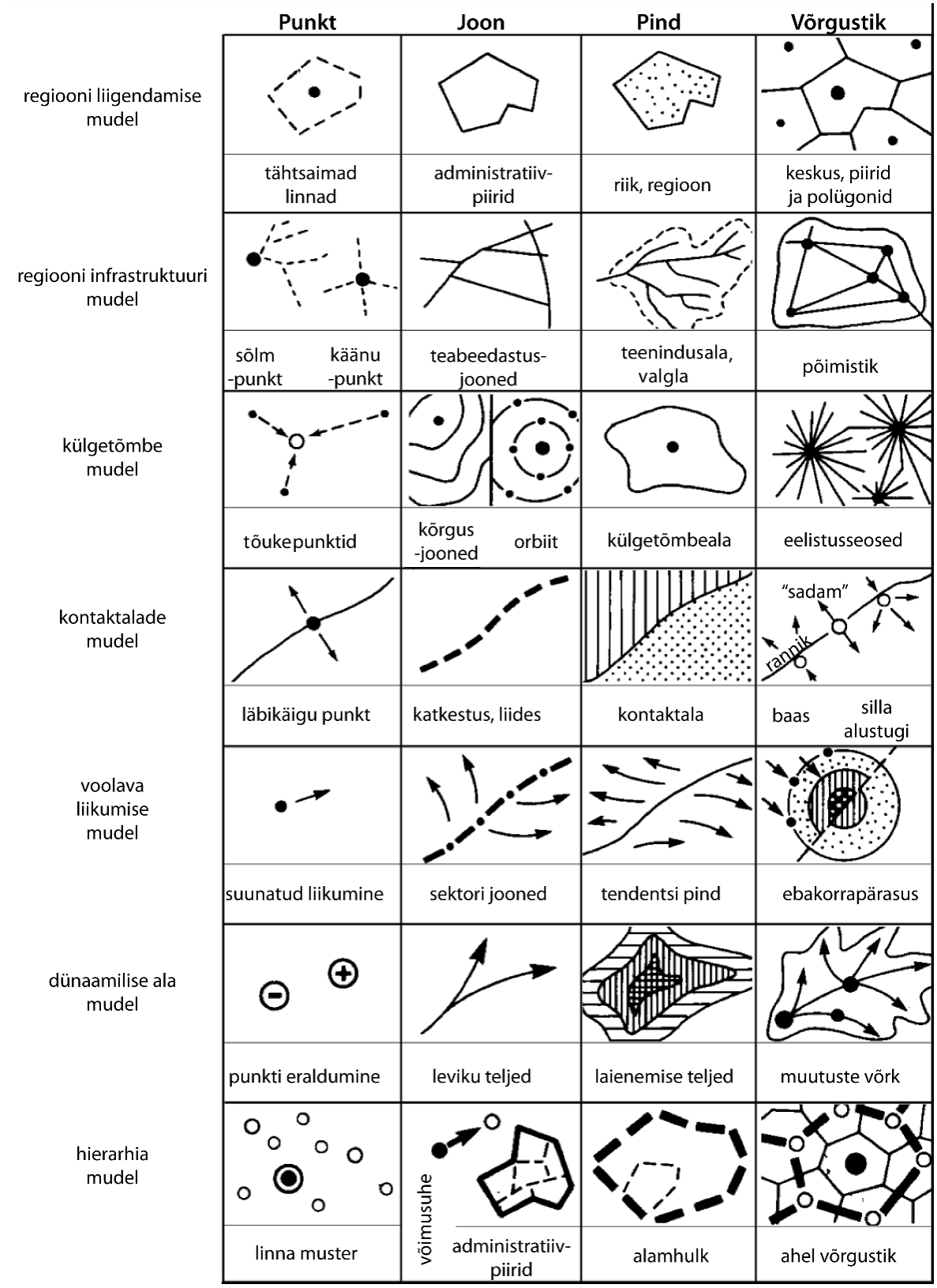
Semantiliselt on horeeme jaotatud kolme kategooriasse (Del Fatto *et al*. 2008):

- geograafiline horeem (*geographic choreme*) esitab geograafilist infot koos geomeetriliselt lihtsate kujunditega (punkt, joon, pind) moodustades sidusvõrgustikke;
- olemuslik horeem (*phenomenal choreme*) hõlmab ühte või mitut geograafilist horeemi ja kirjeldab nende suhteid ja muutumisi;
- tekstiline horeem (*annotation choreme*) abitekstid andmete esitamiseks või lisateabeks.

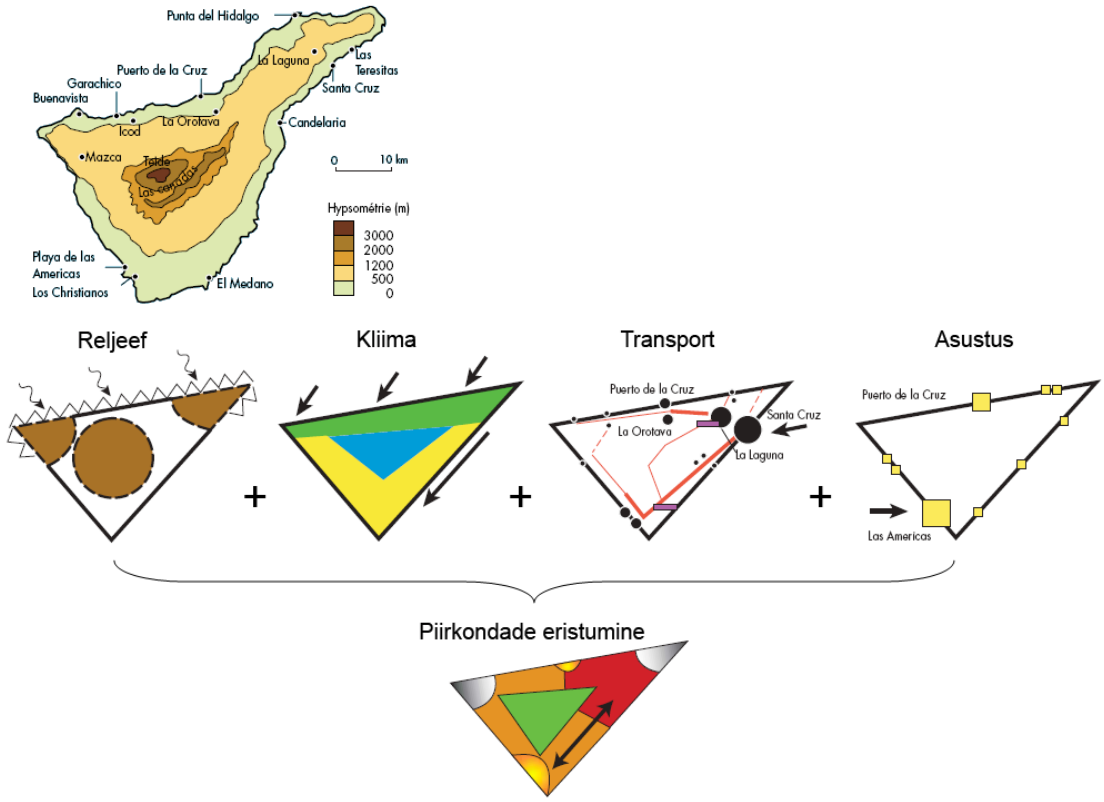
Horeemkaartide konteksti on põhjalikumalt käsitlenud Reimer (2010) artiklis, kus ta toob välja horeemide ühe võimaliku taksonoomia arvestades nende ajaloolist tausta ja tänapäevaste suurte ruumiandmemahtude visualiseerimise keerukust.

Horeemide kognitiivset mõju ruumiinfo edasi andmisele on senini uuritud üsna vähe. Küll aga on uuritud horeemide rakendatavust liikumistrajektooride kaardistamiseks (Klippel 2003, 2011). Klippel nimetab piiratud hulka liikumismustrite mõttelisi kontseptsioone **liikuvhoreemideks** (*movement choremes*), mis on tajutavad ühetaolistena. Eesmärk on esitada kaardil olevat ruumilist teavet mõtestatult ja inimtajule omasemalt.

Horeemide teooria pole kartograafias siiani suurt kasutust leidnud, üheks põhjuseks võib pidada minevikus olnud tugevat vastasseisu Brunet poolt rajatud GIP-RECLUS (GIP – *Public Interset Group* ehk huvigrupp, Reclus – prantsuse keeles *Réseau d'etudes des changements dans les localisations et les unités spatiale*, eesti keeles – ruumistruktuuri ja paiknemise muutuste uurimise ühendus) koolkonnale prantslastest geograafide seas. Vastasseis viis Brunet´nimetamiseni kogunisti pseudo-teadlaseks, kuna ta õhutas ruumi liigsele lihtsustamisele (Lacoste 1993). Tänaseni puudub üldtunnustatud horeemide genereerimise algoritm. Varem koostati ja kujundati horeeme käsitsi andmeid subjektiivselt tõlgendades. Huvi horeemide automaatse genereerimise vastu üha kasvab, seda on tinginud on tinginud massiivsete andmekogude kasv ja nende visualiseerimise vajadus. Horeemide abil modelleerimise juures on peamine omada selget arusaama protsessidest, mis uuritavat objekti mõjutavad, kuna horeemkaardina saadav üldistus toob esile üksnes tugevaimad ruumilised seosed (Cabus 2000).



Joonis 6-7. Horeemide klassifikatsioon (Brunet 1987 järgi).



Joonis 6-8. Üldistatud horeem-kaart (Jadé 2000).

Uurimused

Tang ja Hurni (2009) uurimuses hinnatakse horeemide sobivust planeeringukaartide koostamisel erinevate teemade nagu linnastumine, maakate ja transport kaardistamiseks. Uurimuses võrreldi senitoodetud Hiina ja Šveitsi traditsioonilisi planeeringukaarte nii kujutusviiside kui ka ideoloogia taustal, rõhutades seejuures kultuuriruumide eripärasid erinevate nähtuste edasiandmisel.

Cabus (2000) kasutas horeemide teooriat Belgia autotööstuse strateegia visualiseerimisel näitamaks ettevõtlusvõrgustike kujunemist (suurettevõtte ja tarnija vahelisi suhteid). Traditsioonilisele kartodiagrammile lisaks koostatud horeemkaart andis selgema ülevaate varustajate ruumijaotusest kui multitsentrilisest võrgustikust. Selgesti eristusid autotootjate ja varustajate vahelised industriaaltsoonid, mis viitab sellele, et ruum reageerib ettevõte strateegiale kulude minimaliseerimise suunas ja välja kujuneb selge ruumiline lahusus.

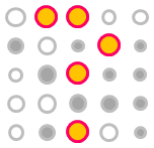
Horeemide teooriat on püütud rakendada ka ilukirjanduslikele tekstidele (Ligozat *et al.* 2007). Eesmärgiks oli luua skemaatiline kujutis ruumis aset leidnud sündmustest ja nähtustest.

Viimaste aastate aktuaalne teema on horeemide automaatne genereerimine. Horeemkaarti on automaatselt loodud näiteks läbi kolme etapi: territoriaalse piirjoone üldistamine, pindobjektide agregeerimine või liialdamine ja kattuvate elementide poolt põhjustatud liigse graafilise müra kaotamine arvestades kihtide hierarhiat (Reimer ja Fohringer 2010).

Senini enim mainitud horeeme automaatselt loov süsteem kannab nime ChEViS (*Chorem Extraction and Visualization System*) (Del Fatto 2009, Del Fatto *et al.* 2007, 2008). ChEViS kasutab horeemkaartide genereerimiseks spetsiaalselt selleks otstarbeks välja töötatud keelt ChorML (põhineb XML keelel), mille abil eeltöödeldakse andmebaas kasutades erinevaid andmekaevandamise ja

generaliseerimise algoritme ning seejärel kuvatakse soovitud horeemina. Süsteemi on kasutatud näiteks Itaalia makroregioonide vahelise elanike rände analüüsiks (Chiara 2011).

Ökoloogiliste andmete esitamiseks pole horeeme teadolevalt kasutatud, kuid horemaatika võiks olla kohandatav ka ökoloogiliste ajas ja ruumis muutuvate seoste esitamiseks.



Küsimused

1. Mille poolest erinevad harvendusega protsessid ja inhibitsiooniga protsessid?

2. Mis on vaatluskohtade lihtsa juhusliku paigutuse puudused?

3. Milleks kasutatakse neutraalseid maastikumudeleid?

4. Mis vahe on nullhüpoteesil, nullmudelil, neutraalsel mudelil ja neutraalsel maastikumudelil?

5. Mis on maastiku neutraalse muutumismudeli lähteandmed?

6. Kas fraktaalne muster on juhuslik?

7. Mis on stohhastilise jäljenduse eelis interpoleerimise ees?

8. Mis on stohhastilise jäljenduse puudus võrreldes interpoleerimisega?

9. Milleks karastatakse jäljendust?

10. Mille poolest erineb Gibbsi sampler ja järjestikune jäljendus?

11. Mis vahe on logistilisel regressioonimudelil ja autologistilisel mudelil?

12. Kas detailiseerimine eeldab täiendavate detailsemate tunnuste kasutamist või saab detailiseerida ka vaid sama andmekihi järgi?

13. Millise näitaja järgi eemaldatakse nurki Ramer-Douglas-Peuckeri algoritmi kohaselt?

14. Mis on kartograafilise generaliseerimise peamine eesmärk?

Viidatud kirjandus

Aas K., Eikvil L., Huseby R.B. 1999. Application of hidden Markov chains in image analysis. *Pattern Recognition* 32(4), 703–713.
Aaviksoo K. 1993. *Application of Markov models in investigation of vegetation and land use dynamics in Estonian mire landscapes*. Tartu, TÜ kirjastus.
Aaviksoo K. 1995. Simulating vegetation dynamics and land use in a mire landscape using a Markov model. *Landscape and Urban Planning* 31(1-3), 129–142.
Aaviksoo K., Ilomets M., Zobel M. 1994. Dynamics of mire communities: a Markovian approach (Estonia). Patten B.C. *et al*. (toim). Wetlands and Shallow Continental Water Bodies. Volume 2. Case studies. The Hague, SPB Academic Publishing, 23–43.
Abler R., Adams J.S., Gould P. 1971. *Spatial organization: the geographers view of the world*. New Jersey, Enlewood Cliffs.
Addicott J.F., Aho J.M., Antolin M.F., Padilla D.K., Richardson J.S., Soluk D.A. 1987. Ecological neighbourhoods: scaling environmental problems. *Oikos* 49(3), 340–346.
Agresti A. 1996. *An introduction to categorical data analysis.* New York etc. Wiley.
Agterberg F.P., Bonham-Carter G.F., Wright D.F. 1990. Statistical pattern integration for mineral exploration. Gaal G., Merriam D.F. (toim). Computer Applications in Resource Estimation: Prediction and Assessment for Metals and Petroleum. Oxford, Pergamon Press, 1–21.
Aha D.W. 1998. The omnipresence of case-based reasoning in science and application. *Knowledge-Based Systems* 11(5-6), 261–273.
Aitken M., Roberts D.W., Shultz L.M. 2007. Modeling distributions of rare plants in the Great Basin, western North America. *Western North American Naturalist* 67(1), 26–38.
Akçakaya H.R, McCarthy M.A., Pearce J. 1995. Linking landscape data with population viability analysis: management options for the helmeted honeyeater. *Biological Conservation* 73(2), 169–176.
Akçakaya H.R. 2001. Linking population-level risk assessment with landscape and habitat models. *The Science of the Total Environment* 274(1-3), 283–291.
Albert C.H., 2010. Sampling in ecology and evolution – bridging the gap between theory and practice. *Ecography* 33(6), 1028–1037.
Albert P.S., McShane L.M. 1995. A generalized estimating equations approach for spatially correlated binary data: applications to the analysis of neuroimaging data. *Biometrics* 51(2), 627–638.
Ali G.A., Roy A.G., Legendre P. 2010. Spatial relationships between soil moisture patterns and topographic variables at multiple scales in a humid temperate forested catchment. *Water Resources Research* 46(10), W10526.
Allouche O., Steinitz O., Rotem D., Rosenfeld A., Kadmon R. 2008. Incorporating distance constraints into species distribution models. *Journal of Applied Ecology* 45(2), 599–609.
Allouche O., Tsoar A., Kadmon R. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology* 43(6), 1223–1232.
Almeida R.C., Delphim S.A., da S. Costa M. 2006. A numerical model to solve single-species invasion problems with Allee effects. *Ecological Modelling* 192(3-4), 601–617.
Almon S. 1965. The distributed lag between capital appropriations and expenditures. *Econometrica* 33(1), 178–196.
Altamirano A., Field R., Cayuela L., Aplin P., Lara A., Rey-Benayas J.M. 2010. Woody species diversity in temperate Andean forests: The need for new conservation strategies multiple regression to develop models to predict woody species richness. *Biological Conservation* 143(9), 2080–2091.
Amorim A.M.T., Gonçalves A.B., Nunes L.M., Sousa A.J. 2011. Optimizing the location of weather monitoring stations using estimation uncertainty. *International Journal of Climatology*, DOI: 10.1002/joc.2317.
Andréfouët S., Claereboudt M. 2000. Objective class definitions using correlation of similarities between remotely sensed and environmental data. *International Journal of Remote Sensing* 21(9), 1925–1930.
Anderson R.P., Lew D., Peterson A.T. 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modelling* 162(3), 211–232.
Anderson R.P., Mártinez-Meyer E. 2004. Modeling species' geographic distributions for preliminary conservation assessments: an implementation with the spiny pocket mice (*Heteromys*) of Ecuador. *Biological Conservation* 116(2), 167–179.
Anselin L. 1995. Local indicators of spatial association – LISA. *Geographical Analysis* 27(2), 93–115.
Anselin L., Syabri I., Kho Y. 2006. GeoDa: An Introduction to Spatial Data Analysis. *Geographical Analysis* 38(1), 5–22.
Apeldorn R.C. van, Oostenbrink W.T., van Winden A., van der Zee F.F. 1992. Effects of habitat fragmentation on the bank vole, *Clethrionomys glareolus*, in an agricultural landscape. *Oikos* 65(2), 265–274.

Aplin P., Atkinson P.M., Curran P.J. 1999. Fine spatial resolution simulated satellite sensor imagery for land cover mapping in the United Kingdom. *Remote Sensing of the Environment* 68(3), 206–216.
Arai K. 1993. A classification method with a spatial-spectral variability. *International Journal of Remote Sensing* 14(4), 699–709.
Araújo M.B., Guisan A. 2006. Five (or so) challenges for species distribution modelling. *Journal of Biogeography* 33(10), 1677–1688.
Araújo M.B., Luoto, M. 2007. The importance of biotic interactions for modelling species distributions under climate change. *Global Ecology and Biogeography* 16(6), 743–753.
Araújo M.B., New M. 2007. Ensemble forecasting of species distributions. *Trends in Ecology and Evolution* 22(1), 42–47.
Araújo M.B., Thuiller W., Williams P.H., Reginster I. 2005. Downscaling European species atlas distributions to a finer resolution: implications for conservation planning. *Global Ecology and Biogeography* 14(1), 17–30.
Araújo M.B., Williams P.H. 2000. Selecting areas for species persistence using occurrence data. *Biological Conservation* 96(3), 331–345.
Arntzen J.W. 2006. From descriptive to predictive distribution models: a working example with Iberian amphibians and reptiles. *Frontiers in Zoology* 3(8), http://www.frontiersinzoology.com/content/3/1/8.
Arpat G.B., Caers J. 2007. Conditional simulation with patterns. *Mathematical Geology* 39(2), 177–203.
Arthur S.M., Manly B.F., McDonald L.L., Garner G.W. 1996. Assessing habitat selection when availability changes. *Ecology* 77(1), 215–227.
Aspinall R.J. 1992. An inductive modelling procedure based on Bayes' theorem for analysis of pattern in spatial data. *International Journal of Geographical Information Systems* 6(2), 105–121.
Aspinall R. 1993. Habitat mapping from satellite imagery and wildlife survey using Bayesian modeling procedure in a GIS. *Photogrammic Engineering and Remote Sensing* 59(4), 537–543.
Aspinall R.J. 1994. Exploratory spatial analysis in GIS: generating geographical hypotheses from spatial data. Worboys M.F. (toim). Innovations in GIS 1: selected papers from the First National Conference on GIS Research UK. Taylor and Francis, 139–147.
Atkinson P.M. 1997. Mapping sub-pixel boundaries from remotely sensed images. Bristol P.A. (toim). Innovations in GIS 4. London, Taylor and Francis, 166–180.
Atkinson P.M., Danson F.M. 1988. Spatial resolution for remote sensing of forest plantations. Proceedings of IGARSS '88 Symposium, Edinburgh, Scotland, 13–16 September, ESA SP-284, IEEE88CH2497-6, 221–223.
Atkinson P.M., Lewis P. 2000. Geostatistical classification for remote sensing: an introduction. *Computers & Geosciences* 26(4), 361–371.
Augustin N.H., Kublin E., Metzler B., Meierjohann E., von Wuhlisch G. 2005. Analyzing the spread of beech canker. *Forest Science* 51(5), 438–448.
Augustin N.H., Mugglestone M.A., Buckland S.T. 1996. An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology* 33(2), 339–347.
Aunap R., Ainsaar L., Rattas M., Mardiste H., Jaagus J., Kull A., Ahas R., Aasa A., Järvet A., Rooma I., Tarre A., Kuusk V., Arold I., Remm K., Timm U., Tammaru T., Kask I., Pae T., Liiber Ü., Rõivas T., Thomson H., Vessin U., Kurs O. 2007. *Eesti atlas* (3., täiendatud ja parandatud trükk). Tartu, Avita.
Austin M.P. 1987. Models for the analysis of species response to environmental gradients. *Vegetatio* 69(1-3), 35–45.
Austin M.P. 1998. An ecological perspective on biodiversity investigations: examples from Australian eucalypt forests. *Annals of the Missouri Botanical Garden* 85(1), 2–17.
Austin M.P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling* 157(2-3), 101–118.
Austin M.P., Van Niel K.P. 2011. Improving species distribution models for climate change studies: variable selection and scale. *Journal of Biogeography* 38(1), 1–8.
Austin M.P, Belbin L., Meyers J.A., Doherty M.D., Luoto M. 2006. Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory. *Ecological Modelling* 199(2), 197–216.
Austin M.P. 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling* 200(1-2), 1–19.
Austin M.P., Heyligers P.C. 1989. Vegetation survey design for conservation: gradsect sampling of forests in north-east New South Wales. *Biological Conservation* 50(1-4), 13–32.
Austin M.P., Meyers J.A. 1996. Current approaches to modelling the environmental niche of Eucalypts: implications for management of forest biodiversity. *Forest Ecology and Management* 85(1-3), 95–106.
Austin M.P., Cunningham R.B., Fleming P.M. 1984. New approaches to direct gradient analysis using environmental scalars and statistical curve-fitting procedures. *Vegetatio* 55(1), 11–27.
Austin M.P., Nicholls A.O., Margules C.R. 1990. Measurement of the realized quantitative niche: environmental niche of five Eucalyptus species. *Ecological Monographs* 60(2), 161–177.

Baatz M., Schäpe A. 1999. Object-oriented and multi-scale image analysis in semantic networks. Proceedings of the 2nd International Symposium on Operationalization of Remote Sensing. August 16th-20th 1999. Enschede. ITC.

Baatz M., Schäpe A. 2000. Multiresolution segmentation: an optimization approach for high quality multiscale image segmentation. Strobl J., Blaschke T. (toim). Angewandte Geogr. Informationsverarbeitung, vol. XII. Heidelberg, Wichmann, 12– 23.

Bacaro G., Santi E., Rocchini D., Pezzo F., Puglisi L., Chiarucci A. 2011. Geostatistical modelling of regional bird species richness: exploring environmental proxies for conservation purpose. *Biodiversity and Conservation* 20(8), 1677–1694.

Bader M. 2001. *Energy Minimization Methods for Feature Displacement in Map Generalization*. Dissertation zur Erlangung der naturwissenschaftlichen Doktorwürde, Universität Zürich.

Bader M., Barrault M., Regnauld N., Mustiére S., Duchêne C., Ruas A., Fritsch E., Lecordix F., Barillot X. 1999. *AGENT Workpackage D2 – Selection of Basic Algorithms*. Technical report, Department of Geography, University of Zurich.

Baddeley A.J., Turner R. 2005. Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software* 12(1), 1–42.

Baddeley A.J., Møller J., Waagepetersen R. 2000. Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. Statistica Neerlandica 54(3), 329–350.

Bagan H., Wang Q.X., Watanabe M., Yang Y.H., Jianwen M. 2005. Land cover classification from MODIS EVI times-series data using SOM neural network. *International Journal of Remote Sensing* 26(22), 4999−5012.

Baker B.D. 1996. Landscape pattern, spatial behavior, and a dynamic state variable model. *Ecological Modelling* 89(1-3), 147–160.

Baker F.A., Verbyla D.L., Hodges C.S.Jr., Ross E.W. 1993. Classification and regression tree analysis for assessing hazard of pine mortality caused by Heterobasidion annosum. *Plant Disease* 77(2), 136–139.

Barbosa A.M., Real R., Olivero J., Vargas J.M. 2003. Otter (*Lutra lutra*) distribution modeling at two resolution scales suited to conservation planning in the Iberian Peninsula. *Biological Conservation* 114(3), 377–387.

Barnett J.L., How R.A., Humphreys W.F. 1978. The use of habitat components by small mammals in eastern Australia. *Australian Journal of Ecology* 3(3), 277–285.

Barnsely M., Barr S.L. 1996. Inferring urban land use from satellite sensor images using kernel-based spatial re-classification. *Photogrammetric Engineering and Remote Sensing* 62(8), 949–958.

Barot S., Gignoux J., Menaut J.-C. 1999. Demography of a savanna palm tree: predictions from comprehensive spatial pattern analyses. *Ecology* 80(6), 1987–2005.

Barry S., Elith J. 2006. Error and uncertainty in habitat models. *Journal of Applied Ecology* 43(3), 413–423.

Barry S.C., Welsh A.H. 2002. Generalized additive modelling and zero inflated count data. *Ecological Modelling* 157(2-3), 179–188.

Bartel A. 2000. Analysis of landscape pattern: towards a 'top down' indicator for evaluation of landuse. *Ecological Modelling* 130(1-3), 87–94.

Bartelink H.H. 2000. Effects of stand composition and thinning in mixed-species forests: a modeling approach applied to Douglas-fir and beech. *Tree Physiology* 20(5-6), 399–406.

Bartlein P.J., Whitlock C. 1993. Paleoclimatic interpretation of the Elk Lake pollen record. Bradbury J.P., Dean W.E. (toim). Elk Lake, Minnesota: evidence for rapid climate change in north-central United States, Geological Society of America, Special Paper 276. Boulder, CO, 275–293.

Bartlett M.S. 1964. Spectral analysis of two dimensional point processes. *Biometrika* 51(3-4), 299–311.

Başaraner M. 2002. Model generalization in GIS. International Symposium on GIS, September 23-26, 2002. Istanbul-Turkey.

Bascomte J., Solé R.V. 1998. *Modeling spatiotemporal dynamics in ecology.* Springer. Berlin and Heidelberg.

Bates J.M., Granger C.W. 1969. The combination of forecasts. *Operational Research Quarterly* 20(4), 451–468.

Batista J.L.F., Maguire D.A. 1998. Modeling the spatial structure of tropical forest. *Forest Ecology and Management* 110(1-3), 293–314.

Bauer E.M., Burk E.T., Ek R.A., Coppin R.P., Lime D.S., Walsh A.T., Walters K.D., Befort W., Heinzen F.D. 1994. Satellite inventory of Minnesota forest resources. *Photogrammetric Engineering and Remote Sensing* 60(3), 287–298.

Beals E.W. 1969. Vegetation change along altitudinal gradients. *Science* 165(3897), 981–985.

Beaumont L.J., Lesley L., Poulsen M. 2005. Predicting species distributions: use of climatic parameters in BIOCLIM and its impact on predictions of species' current and future distributions. *Ecological Modelling* 186(2), 250–269.

Bedia J., Busqué J., Gutiérrez J.M. 2011. Predicting plant species distribution across an alpine rangeland in northern Spain. A comparison of probabilistic methods. *Applied Vegetation Science* 14(3), 415–432.

Begon M., Harper J.L., Townsend C.R. 1990. *Ecology, individuals, populations and communities.* Second edition. Boston, Blackwell Scientific Publications.
Beh E.J. 2004. Simple correspondence analysis: a bibliographic review. *International Statistical Review* 72(2), 257–284.
Bekkby T., Rinde E., Erikstad L., Bakkestuen V. 2009. Spatial predictive distribution modelling of the kelp species *Laminaria hyperborea*. *ICES Journal of Marine Science* 66(10), 2106–2115.
Benayas J.M.R., de la Montaña E. 2003. Identifying areas of high-value vertebrate diversity for strengthening conservation. *Biological Conservation* 114(3), 357–370.
Bennet R.J. 1975. Dynamic systems modelling of the North-west region: 1. Spatio - temporal representation and identification. *Environment and Planning A* 7(5), 525–538. Cit. Cliff ja Ord 1981.
Bennie J., Anderson K., Wetherelt A. 2011. Measuring biodiversity across spatial scales in a raised bog using a novel paired-sample diversity index. *Journal of Ecology* 99(2), 482–490.
Berens P. 2009. CircStat: A MATLAB Toolbox for Circular Statistics. *Journal of Statistical Software* 31(10), 1–21.
Berg Å., Gärdenfors U., von Proschwitz T. 2004. Logistic regression models for predicting occurrence of terrestrial molluscs in southern Sweden – importance of environmental data quality and model complexity. *Ecography* 27(1), 83–93.
Bergelson J. 1990. Life after death: site pre-emption by the remains of *Poa annua*. *Ecology* 71(6), 2157–2165.
Berger J.O., Sellke T. 1987. Testing a point hypothesis: the irreconcilability of *P* values and evidence. *Journal of the American Statistical Association* 82(397), 112–139.
Bergman C.M., Schaefer J.A, Luttich S.N. 2000. Caribou movement as a correlated random walk. *Oecologia* 123(3), 364–374.
Besag J.E., Gleaves J.T. 1973. On the detection of spatial pattern in plant communities. *Bulletin of the International Statistical Institute* 45, 153–158.
Beselga A., Araújo M.B. 2009. Individualistic vs community modelling of species distributions under climate change. *Ecography* 32(1), 55–65.
Betts M.G., Diamond A.W., Forbes G.J., Villard M.-A., Gunn J.S. 2006. The importance of spatial autocorrelation, extent and resolution in predicting forest bird occurrence. *Ecological Modelling* 191(2), 197–224.
Beveridge J.R., Griffith J., Kohler R.R., Hanson A.R., Riseman E.M. 1989. Segmenting images using localized histograms and region merging. *International Journal of Computer Vision* 2(3), 311–347.
Bian L., Butler R. 1999. Comparing effects of aggregation methods on statistical and spatial properties of simulated spatial data. *Photogrammetric Engineering and Remote Sensing* 65(1), 73–84.
Binns M.R., Nyrop J.P. Werf W. van der 2000. Sampling and Monitoring in Crop Protection: The Theoretical Basis for Designing Practical Decision Guides. *CABI Publishing*.
Birks H.J.B. 1993. Quaternary palaeoecology and vegetation science – current contributions and possible future developments. *Review of Paleaobotany and Palynology* 79(1–2), 153–177.
Birks H.J.B., Line J.M., Juggins S., Stevenson A.C., ter Braak C.J.F. 1990. Diatioms and pH reconstruction. *Philosophical Transactions of the Royal Society of London, Series B* 327(1240), 263–278.
Birt A.G., Xi W., Coulson R.N. 2009. LANDISVIEW: a visualization tool for landscape modeling. *Environmental Modelling & Software* 24(11), 1339–1341.
Bishop C.M.2006. *Pattern recognition and machine learning*. New York. Springer.
Björnsson H., Venegas S.A. 1997. *A manual of EOF and SVD analyses of climatic data*. McGill University, CCGCR Report No. 97-1. Montréal, Québec, 52. http://brunnur.vedur.is/pub/halldor/PICKUP/eof.pdf.
Bjørke J.T., Nilsen S. 2003. Wavelets applied to simplification of digital terrain models. *International Journal of Geographical Information Science* 17(7), 601–621.
Bjørnstad O.N., Stenseth N.C., Saitoh T. 1999a. Synchrony and scaling in dynamics of voles and mice in northern Japan. *Ecology* 80(2), 622–637.
Bjørnstad O.N., Ims R.A., Lambin X. 1999b. Spatial population dynamics: analysing patterns and processes of population synchrony. *Trends in Ecology and Evolution* 14(11), 427–432.
Blackstock T.H., Burrows C.R., Howe E.A., Stevens D.P., Stevens J.P. 2007. Habitat inventory at a regional scale: a comparison of estimates of terrestrial Broad Habitat cover from stratified sample field survey and full census field survey for Wales, UK. *Journal of Environmental Management* 85(1), 224–231.
Blanchet P.G., Legendre P., Borcard D. 2008. Modelling directional spatial processes in ecological data. *Ecological Modelling* 215(4), 325–336.
Block W.M., Morrison M.L., Scott P.E. 1998. Development and evaluation of habitat models for herpetofauna and small mammals. *Forest Science* 44(3), 430–437.
Bloom S.A. 1981. Similarity indices in community studies: potential pitfalls. *Marine Ecology, Progress Series* 5(2), 125–128.
Blum A. L., Langley P. 1997. Selection of relevant features and examples in machine learning. *Artificial*

Intelligence 97(1-2), 245–271.
Bocquet-Appel J.P., Sokal R.R. 1989. Spatial autocorrelation analysis of trend residuals in biological data. *Systematic Zoology* 38(4), 333–341.
Boisvert J.B., Pyrcz M.J., Clayton V., Deutsch C.V. 2008. Multiple-point statistics for training image selection. *Natural Resources Research* 16(4), 313–321.
Bonham C.D., Reich R.M. 1999. Influence of spatial autocorrelation on a fixed-effect model used to evaluate treatment of oil spills. *Applied Mathematics and Computation* 106(2-3), 149–162.
Bonham-Carter G.F., Agterberg F.P., Wright D.F. 1989. Weights of evidence modelling: a new approach to mapping mineral potential. Agterberg F.P., Bonham-Carter G.F. (toim). Statistical Applications in the Earth Sciences. Geology Survey of Canada 89(9), 171–183.
Bonn A., Schröder B. 2001. Habitat models and their transfer for single and multi species groups: a case study of carabids in an alluvial forest. *Ecography* 24(4), 483–496.
Bookstein F.L. 1996. Biometrics, biomathematics and the morphometric synthesis. *Bulletin of Mathematical Biology* 58(2), 313–365.
Boots B. 2003. Developing local measures of spatial association for categorical data. *Journal of Geographical Systems* 5(2), 139–160.
Boots B. 2006. Local configuration measures for categorical spatial data: binary regular lattices. *Journal of Geographical Systems* 8(1), 1–24.
Borcard D., Gillet F., Legendre P. 2011. *Numerical ecology with R.* New York. Springer.
Borcard D., Legendre P. 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling* 153(1-2), 51–68.
Borcard D., Legendre P., Avois-Jacquet C., Tuomisto H. 2004. Dissecting the spatial structure of ecological data at multiple scales. *Ecology* 85(7), 1826–1832.
Borcard D., Legendre P., Drapeau P. 1992. Partialling out the spatial component of ecological variation. *Ecology* 73(3), 1045–1055.
Borggaard C., Thodberg H.H. 1992. Optimal minimal neural interpretation of spectra. *Analytical Chemistry* 64(5), 545–551.
Borrough P.A. 1983. Fractal dimensions of landscapes and other environmental data. *Nature* 294, 240–242.
Botequilha Leitão A., Ahern J. 2002. Applying landscape ecological concepts and metrics in sustainable landscape planning. *Landscape and Urban Planning* 59(2), 65–93.
Box G.E.P., Cox D.R. 1964. An analysis of transformations. *Journal of the Royal Statistical Society, Series B* 26(2), 211–252.
Boyle T.B., Smillie G.M., Anderson J.C., Beeson D.R. 1990. Research Journal of the Water Pollution Control Federation 62(6), 749–762.
Braak C.J.F. ter 1988a. Partial canonical correspondence analysis. Bock H.H. (toim). Classification and Related Methods of Data Analysis, North-Holland, Amsterdam, 551–558.
Braak C.J.F. ter 1988b. CANOCO–an extension of DECORANA to analyze species-environment relationships. *Vegetatio* 75(3), 159–160.
Braak C.J.F. ter 1995. Non-linear methods for multivariate statistical calibration and their use in palaeoecology: a comparison of inverse (k-nearest neighbours, partial least squares and weighted averaging partial least squares) and classical. *Chemometrics and Intelligent Laboratory Systems* 28(1), 165–180.
Braak C.J.F. ter, Barendregt L.G. 1986. Weighted averaging of species indicator values: its efficiency in environmental calibration. *Mathematical Biosciences* 78(1), 57–72.
Braak C.J.F. ter, van Dam H. 1989. Inferring pH from diatoms: a comparison of old and new calibration methods. *Hydrobiologia* 178(3), 209–223.
Braak C.J.F. ter, Juggins S. 1993. Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from species assemblages. *Hydrobiologia* 269(1), 485–502.
Braak C.J.F. ter, Juggins S., Birks H.J.B., van der Voet H. 1993. Weighted averaging least squares regression (WA-PLS): definition and comparison with other methods for species-environment calibration. Patil G.P., Rao C.R. (toim). Multivariate Environmental Statistics, North-Holland, Amsterdam, 525–560.
Braak C.J.F. ter, Prentice I.C. 1988. A theory of gradient analysis. *Advances in Ecological Research* 18, 271–317.
Bradley A.P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30(7), 1145–1159.
Bradshaw G.A., Spies T.A. 1992. Characterizing canopy gap structure in forests using wavelet analysis. *Journal of Ecology* 80(2), 205–215.
Bradter U., Thom T.J., Altringham J.D., Kunin W.E., Benton T.G. 2011. Prediction of National Vegetation Classification communities in the British uplands using environmental data at multiple spatial scales, aerial images and the classifier random forest. *Journal of Applied Ecology* 48(4), 1057–1065.
Brambilla M., Casale F., Bergero V., Crovetto G.M., Falco R., Negri I., Siccardi P., Bogliani G. 2009. GIS-

models work well, but are not enough: habitat preferences of *Lanius collurio* at multiple levels and conservation implications. *Biological Conservation* 142(10), 2033–2042.
Brandtberg T., Warner T.A., Landenberger R.E., McGraw J.B. 2003. Detection and analysis of individual leaf-off tree crowns in small footprint, high sampling density lidar data from the eastern deciduous forest in North America. *Remote Sensing of Environment* 85(3), 290–303.
Brassel K.E., Weibel R. 1988. A review and conceptual framework of automated map generalization. *International Journal of Geographical Information Systems* 2(3), 229–244. Cit. McMaster ja Shea (1992).
Braunisch V., Bollmann K., Graf R.F., Hirzel A.H. 2008. Living on the edge–modelling habitat suitability for species at the edge of their fundamental niche. *Ecological Modelling* 214(2-4), 153–167.
Bray J.R., Curtis J.T. 1957. An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs* 27(4), 326–349.
Bregt A.K., Wopereis M.C.S. 1990. Comparison of complexity measures for choropleth maps. *The Cartographic Journal* 27(2), 85–91.
Breiman L., Friedman J.H. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association* 80(391), 580–619.
Breiman L., Friedman J.H., Olshen R.A., Stone C.G. 1984. *Classification and regression trees*. Boca Raton etc. Chapman & Hall.
Bridge S.R.J., Johnson E.A. 2000. Geomorphic principles of terrain organization and vegetation gradients. *Journal of Vegetation Science* 11(1), 57–70.
Brin S, Page L. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks* 30(1-7), 107–117.
Brito J.C., Crespo E.G., Paulo O.S. 1999. Modelling wildlife distributions: logistic multiple regression vs overlap analysis. *Ecography* 22(3), 251–260.
Brotons L., Thuiller W., Araújo M.B., Hirzel A.H. 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography* 27(4), 437–448.
Brown C.J., Smith S.J., Lawton P., Anderson J.T. 2011. Benthic habitat mapping: a review of progress towards improved understanding of the spatial ecology of the seafloor using acoustic techniques. *Estuarine, Coastal and Shelf Science* 92(3) 502–520.
Brown D.G. 1994. Predicting vegetation types at treeline using topography and biophysical disturbance variables. *Journal of Vegetation Science* 5(5), 641–656.
Brown J.H., Mehlman D.W., Stevens G.C. 1995. Spatial variation in abundance. *Ecology* 76(7), 2028–2043.
Brown J.H., Stevens G.C., Kaufman D.M. 1996. The geographic range: size shape, boundaries, and internal structure. *Annual Review of Ecology and Systematics* 27(1), 597–623.
Browning D.M., Beaupre S.J., Duncan L. 2005. Using partitioned Mahalanobis $D^2(k)$ to formulate a GIS-based model of timber rattlesnake hibernacula. *Journal of Wildlife Management* 69(1), 33–44.
Brunet R. 1980. La composition des modèles dans l'analyse spatiale. *L'Espace Géographique* 9(4), 253–265. Cit. http://fr.wikipedia.org/wiki/Chor%C3%A9matique.
Brunet R. 1987. *La carte: Mode d'Emploi.* Fayard-Reclus, Paris.
Brunsdon C.A., Fotheringham A.S., Charlton M.E. 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis* 28(4), 281–298.
Brunsdon C.A., Fotheringham A.S., Charlton M.E. 1998. Geographically weighted regression-modelling spatial non-stationarity. *The Statistician* 47(3), 431–443.
Bruun H.H., Moen J., Angerbjörn A. 2003. Environmental correlates of meso-scale plant species richness in the province of Härjedalen, Sweden. *Biodiversity and Conservation* 12(10), 2025–2041.
Buckland S.T., Elston D.A. 1993. Empirical models for the spatial distribution of wildlife. *Journal of Applied Ecology* 30(3), 478–495.
Buckland S.T., Anderson D.R., Burnham K.P., Laake J.L. 1993. *Distance sampling: Estimating Abundance of Biological Populations.* London, Chapman and Hall.
Buisson L., Thuiller W., Casajus N., Lek S., Grenouillet G. 2010. Uncertainty in ensemble forecasting of species distribution. *Global Change Biology* 16(4), 1145−1157.
Bunce R.G.H., Barr C.J., Clarke R.T., Howard D.C., Lane A.M.J. 1996. Land classification for strategic ecological survey. *Journal of Environmental Management* 47(1), 37–60.
Burges C.J.C. 2009. *Dimension Reduction: A Guided Tour*. MSR Tech Report MSR-TR-2009-2013. http://research.microsoft.com/en-us/um/people/cburges/tech_reports/msr-tr-2009-2013.pdf
Burges C.J.C. 2010. Dimension reduction: a guided tour. *Foundations and Trends® in Machine Learning* 2(4), 275–365.
Busby J.R. 1986. A biogeographical analysis of *Nothofagus cunninghamii* (Hook.) Oerst. in southeastern Australia. *Australian Journal of Ecol*ogy 11(1), 1–7.
Bustamante J., Seoane J. 2004. Predicting the distribution of four species of raptors (Aves: Accipitridae) in southern Spain: statistical models work better than existing maps. *Journal of Biogeography* 31(2), 295–306.

Bütler R., Schlaepfer R. 2004. Spruce snag quantification by coupling colour infrared aerial photos and a GIS. *Forest Ecology and Management* 195(3), 325–339.
Cabus P. 2000. Modelling spatial relationships between Belgian car manufacturers and their suppliers using choremes. *Tijdschrift voor Economische en Sociale Geografie* 91(1), 3–19.
Caers J. 2001. Geostatistical reservoir modelling using statistical pattern recognition. *Journal of Petroleum Science and Engineering* 29(3-4), 177–188.
Camarero J.J., Gutiérrez E., Fortin M.-J. 2006. Spatial patterns of plant richness across treeline ecotones in the Pyrenees reveal different locations for richness and tree cover boundaries. *Global Ecology and Biogeography* 15(2), 182–191.
Campbell D.J. 1992. Nearest-neighbour graphical analysis of spatial pattern and a test for competition in populations of singing crickets (*Teleogryllus commodus*). *Oecologia* 92(4), 548–551.
Campbell D.J. 1995. Detecting regular spacing in patchy environments and estimating its density using nearest-neighbour graphical analysis. *Oecologia* 102(2), 133–137.
Campbell D.J. 1996. Aggregation and regularity: an inclusive one-tailed nearest-neighbour analysis of small spatially patchy populations. *Oecologia* 106(2), 206–211.
Campbell D.J., Shipp E. 1979. Regulation of spatial pattern in populations of the field cricket *Teleogryllus commodus* (Walker). *Zeitschrift für Tierpsychologie* 51(3), 260–268.
Campbell S.M., Econ M.A., Cantrill J.A. 2001. Consensus methods in prescribing research. *Journal of Clinical Pharmacy and Therapeutics* 26(1), 5–14.
Canny J. 1986. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8(6), 679–698.
Cantero J.J., Pärtel M., Zobel M. 1999. Is species richness dependent on the neighbouring stands? An analysis of the commumity patterns in mountain grasslands of central Argentina. *Oikos* 87(2), 346–354.
Cardillo M., Macdonald D.W., Ruchton S.P. 1999. Predicting mammal species richness and distributions: testing the effectiveness of satellite-derived land cover data. *Landscape Ecology* 14(5), 423–435.
Carl G., Kühn I. 2007. Analyzing spatial autocorrelation in species distributions using Gaussian and logit models. *Ecological Modelling* 207(2-4), 159–170.
Carleer A., Wolff E. 2004. Exploitation of very high resolution satellite data for tree species identification. *Photogrammetric Enginering & Remote Sensing* 70(1), 135–140.
Caroll S.S. Pearson D.L. 1998. Spatial modeling of butterfly species richness using tiger beetles (Cicindelidae) as bioindicator taxon. *Ecological Applications* 8(2), 531–543.
Carpenter G., Gillison A.N., Winter J., 1993. DOMAIN: a flexible modeling procedure for mapping potential distributions of plants, animals. *Biodiversity and Conservation* 2(6), 667–680.
Carpenter G.A., Gopal S., Macomber S., Martens S., Woodcock C.E., Franklin J. 1999. A neural network method for efficient vegetation mapping. *Remote Sensing of the Environment* 70(3), 326–338.
Carr J.R. 1996. Spectral and textural classification of single and multiple band digital images. *Computers & Geosciences* 22(8), 849–865.
Carreiras J.M.B., Pereira J.M.C., Pereira J.S. 2006. Estimation of tree canopy cover in evergreen oak woodlands using remote sensing. *Forest Ecology and Management* 223(1-3), 45–53.
Carvalho J., Soares A., Bio A. 2006. Improving satellite images classification using remote and ground data integration by means of stochastic simulation. *International Journal of Remote Sensing* 27(16), 3375–3386.
Cauter A. van, Kerley G.I.H., Cowling R.M. 2005. The consequence of inaccuracies in remote-sensed vegetation boundaries for modelled mammal population estimates. *South African Journal of Wildlife Research* 35(2), 155–161.
Cawsey E.M., Austin M.P., Baker B.L. 2002. Regional vegetation mapping in Australia: a case study in the practical use of statistical modelling. *Biodiversity and Conservation* 11(12), 2239–2274.
Céréghino R., Santoul F., Compin A., Mastrorillo S. 2005. Using self-organizing maps to investigate spatial patterns of non-native species. *Biological Conservation* 125(4), 459–465.
Chakraborty K., Mehrotra K., Mohan C.K., Ranka S. 1992. Forecasting the behavior of multivariate time series using neural networks. *Neural Networks* 5(6), 961–970.
Chapin F.S.III, McGraw J.B.III, Shaver G.R. 1989. Competition causes regular spacing of alder in Alaska shrub tundra. *Oecologia* 79(3), 412–416.
Chapman D.S., Purse B.V. 2011. Community versus single-species distribution models for British plants. *Journal of Biogeography* 38(8), 1524–1535.
Chatterjee S., Laudato M., Lynch L.A. 1996. Genetic algorithms and their statistical applications: an introduction. *Computataional Statistics & Data Analysis* 22(6), 633–651.
Cheetham A.H., Hazel J.E. 1969. Binary (presence-absence) similarity coefficients. *Journal of Paleontology* 43(5), 1130–1136.
Chefaoui R.M., Hortal J., Lobo J.M. 2005. Potential distribution modelling, niche characterization and conservation status assessment using GIS tools: a case study of Iberian *Copris* species. *Biological Conservation*

122(2), 327–338.
Chefaoui R.M., Lobo J.M. 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling* 210(4), 478–486.
Chen J., Bradshaw G.A. 1999. Forest structure in space: a case study of old growth spruce-fir forest in Changbaishan Natural Reserve, PR China. *Forest Ecology and Management* 120(1-3), 219–233.
Chen Q., Han R., Ye F., Li W. 2011. Spatio-temporal ecological models. *Ecological Informatics* 6(1), 37–43.
Cherrill A., McClean C. 1995. An investigation of uncertainty in field habitat mapping and the implications for detecting land cover change. *Landscape Ecology* 10(1), 5–21.
Cherrill A., McClean C. 1999a. Between observer variation in the application of a standard method of habitat mapping by environmental consultants in the UK. *Journal of Applied Ecology* 36(6), 989–1008.
Cherrill A., McClean C. 1999b. The reliability of „Phase 1" habitat mapping in the UK: the extent and types of observer bias. *Landscape and Urban Planning* 45(2) 131–143.
Chesson J. 1978. Measuring preference in selective predation. *Ecology* 59(2), 211–215.
Chesson J. 1983. The estimation and analysis of preference and its relationship to foraging models. *Ecology* 64(5), 1297–1304.
Chevan A., Sutherland M. 1991. Hierarchical partitioning. *The American Statistician* 45(2), 90–96.
Chiara D.D., Fatto V.D., Laurini R., Sebillo M., Vitiello G. 2011. A chorem-based approach for visually analyzing spatial data. *Journal of Visual Languages & Computing* 22(3), 173–193.
Chiarello E., Barrat-Segretain M.H. 1997. Recolonization of cleared patches by macrophytes: modelling with point processes and random mosaics. *Ecological Modelling* 96(1-3), 61–73.
Chiarucci A., D'Auria F., Bonini I. 2007. Is vascular plant species diversity a predictor of bryophyte species diversity in Mediterranean forests? *Biodiversity and Conservation* 16(2), 525–545.
Chica-Olmo M., Abarca-Hernández F. 2000. Computing geostatistical image trexture for remotely sensed data classification. *Computers and Geosciences* 26(4), 373–383.
Chou Y.-H, Soret S. 1996. Neighborhood effects in bird distributions, Navarre, Spain. *Environmental Management* 20(5), 675–687.
Chou J., Weger R.C., Ligtenberg J., Kuo K.S., Welch R.M., Breeden P. 1994. Segmentation of polar scenes using multi-spectral texture measures and morphological filtering. *International Journal of Remote Sensing* 15(5), 1019–1036.
Chubey M.S., Franklin S.E., Wulder M.A. 2006. Object-based analysis of Ikonos-2 imagery for extraction of forest inventory parameters. *Photogrammetric Engineering and Remote Sensing* 72(4), 383–394.
Clark P.J., Evans F.C. 1954. Distance to nearest neighbour as a measure of spatial relationship in populations. *Ecology* 35(4), 445–453.
Clark J.D., Dunn J.E., Smith K.G. 1993. A multivariate model of female black bear habitat use for a geographic information system. *Journal of Wildlife Management* 57(3), 519–526.
Clarke K.R. 1993. Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology* 18(1), 117–143.
Clarke K.R., Ainsworth M. 1993. A method of linking multivariate community structure to environmental variables. *Marine Ecology Progress Series* 92(1), 205–219.
Clarke R.T., Thomas J.A., Elmes G.W., Wardlaw J.C., Munguira M.L., Hochberg M.E. 1998. Population modelling of the spatial interactions between *Maculinea rebeli*, their initial foodplant *Gentiana cruciata* and Myrmica ants within a site. *Journal of Insect Conservation*. 2(1), 29–37.
Clements F.E. 1916. *Plant succession: an analysis of the development of vegetation*. Washington, Carnegie Institution of Washington. http://www.archive.org/stream/cu31924000531818.
Cleveland W.S. 1979. Robust locally weighted regression and smoothing scatter plots. *Journal of American Statistical Association* 74(368), 829–836.
Cleveland W.S., Devlin S.J. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* 83(403), 596–610.
Cleveland W.S., McGill R. 1985. Graphical perception and graphical methods for analyzing scientific data. *Science* 229(4716), 828–833.
Cliff A.D., Ord J.K. 1973. *Spatial autocorrelation*. London, Pion.
Cliff A.D., Ord J.K. 1981. *Spatial processes: models & applications*. London, Pion.
Clifford P., Richardson S., Hémon D. 1989. Assessing the significance of the correlation between two spatial processes. *Biometrics* 45(1), 123–134.
Codling E.A., Plank M.J., Benhamou S. 2008. Random walk models in biology. *Journal of the Royal Society Interface* 5(25), 813–834.
Cofiño A.S, San-Martín D., Gutiérrez J.M. 2007. A web portal for regional projection of weather forecast using GRID middleware. Shi Y. *et al*. (toim). ICCS 2007, Part III, LNCS 4489. Springer, 82–89.
Cohen J.A. 1960. Coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1), 37–46.

Cohen W.B., Maiersperger T.K., Spies T.A., Oetter D.R. 2001. Modelling forest cover attributes as continuous variables in a regional context with Thematic Mapper data. *International Journal of Remote Sensing* 22(12), 2279–2310.
Cohen W.B., Spies T.A., Bradshaw G.A. 1990. Semivariograms of digital imagery for analysis of conifer canopy structure. *Remote Sensing of the Environment* 34(3), 167–178.
Cohn F. 1870. Über den Brunnenfaden (*Crenothrix polyspora*) mit Bemerkungen über die mikroskopische Analyse des Brunnenwassers. *Beiträge zur Biologie der Pflanzen* 1, 346–357. Cit. Абакумов (1981).
Colby J.D. 1991. Topographic normalization in rugged terrain. *Photogrammetric Engineering and Remote Sensing* 57(5), 531–537.
Cole L.C. 1954. Some features of random population cycles. *Journal of Wildlife Management* 18(1), 2–24.
Cole B.J. 1995. Fractal time in animal behaviour – the movement activity of *Drosophila*. *Animal Behaviour*, 50(5), 1317–1324.
Collingham Y.C., Wadsworth R.A., Huntley B., Hulme P.E., 2000. Predicting the spatial distribution of non-indigenous riparian weeds: issues of spatial scale and extent. *Journal of Applied Ecology* 37(s1), 13–27.
Comas C., Palahí M., Pukkala T., Mateu J. 2009. Characterising forest spatial structure through inhomogeneous second order characteristics. *Stochastic Environmental Research and Risk Assessment* 23(3), 387–397.
Comas C., Mateu J. 2007. Modelling forest dynamics: a perspective from point process methods. *Biometrical Journal* 49(2), 176–196.
Compton B.W., Rhymer J.M., McCollough M. 2002. Habitat selection by wood turtles (*Clemmys insculpta*): an application of paired logistic regression. *Ecology* 83(3), 833–843.
Condit R., Ashton P.S., Baker P., Bunyavejchewin S., Gunatilleke S., Gunatilleke N., Hubbell S.P., Foster R.B., Itoh A., La-Frankie J.V., Seng Lee H., Losos E., Manokaran N., Sukumar R., Yamakura T. 2000. Spatial patterns in the distribution of tropical tree species. *Science* 288(5470), 1414–1418.
Condit R., Stephen P., Hubbell S.P, Foster R.B. 1994. Density dependence in two understory tree species in a neotropical forest. *Ecology* 75(3), 671–680.
Congalton R.G., Green K. 1999. *Assessing the accuracy of remotely sensed data: principles and practices*. Lewis Publishers.
Connor E.F., Courtney A.C., Yoder J.M. 2000. Individuals-area relationships: the relationship between animal population density and area. *Ecology* 81(3), 734–748.
Coomes D.A., Rees M., Turnbull L. 1999. Identifying aggregation and association in fully mapped spatial data. *Ecology* 80(2), 554–565.
Cooper A.B., Millspaugh J.J. 1999. The application of discrete choice models to wildlife resource selection studies. *Ecology* 80(2), 566–575.
Coops N., Culvenor D. 2000. Utilizing local variance of simulated high spatial resolution imagery to predict spatial pattern of forest stands. *Remote Sensing of the Environment* 71(3), 248–260.
Coops N.C., Catling P.C. 1997. Predicting the complexity of habitat in forests from airborne videography for wildlife management. *International Journal of Remote Sensing* 18(12), 2677–2686.
Coops N.C., Catling P.C. 2001. Prediction of historical forest habitat patterns using binomial distributions and simple Boolean logic from high spatial resolution remote sensing. *Computers & Geosciences* 27(7), 795–805.
Coops N.C., Culvenor D., Preston R., Catling P.C. 1998. Procedures for predicting habitat and structural attributes in eucalypt forests using high spatial resolution remotely sensed imagery. *Australian Forestry* 61(4), 244–252.
Coops N.C., Wulder M.A., White J.C. 2006. Integrating remotely sensed and ancillary data sources to characterize a mountain pine beetle infestation. *Remote Sensing of Environment* 105(2), 83–97.
Cormack R.M. 1979. Spatial aspects of competition between individuals. Cormack R.M., Ord J.K. (toim). Spatial and temporal analysis in ecology. Fairland, Maryland, USA, International Co-operative Publishing House, 151–212.
Cornelius J.M., Reynolds J.F. 1991. On determining the statistical significance of discontinuities within ordered ecological data. *Ecology* 72(6), 2057–2070.
Corsi F., Dupre E., Boitani L. 1999. A large-scale model of wolf distribution in Italy for conservation planning. *Conservation Biology* 13(1), 150–159.
Cosmopoulos P., King D.J. 2004. Temporal analysis of forest structural condition at an acid mine site using multispectral digital camera imagery. *International Journal of Remote Sensing* 25(12), 2259–2275.
Couclelis H. 1992. People manipulate objects (but cultivate fields): beyond the raster-vector debate in GIS. Frank A.U., Campari I., Formentini U. (toim). Theories and Methods of Spatio-Temporal Reasoning in Geographic Space. International Conference GIS – From Space to Territory: Theories and Methods of Spatio-Temporal Reasoning Pisa, Italy, September 21–23, 1992. Berlin, Springer, 65–77.
Coudun C., Gégout J.-C. 2006. The derivation of species response curves with Gaussian logistic regression is sensitive to sampling intensity and curve characteristics. *Ecological Modelling* 199(2), 164–175.

Couteron P., Seghieri J., Chadœuf J. 2003. A test for spatial relationships between neighbouring plants in plots of heterogeneous plant density. *Journal of Vegetation Science* 14(2), 163–172.
Couteron P., Kokou K. 1997. Woody vegetation spatial patterns in a semi-arid savanna of Burkina Faso, West Africa. *Plant Ecology* 132(2), 211–227.
Cowley M.J.R., Wilson R.J., León-Cortés J.L., Gutiérrez D., Bulman C.R., Thomas C.D. 2000. Habitat-based statistical models for predicting the spetial distribution of butterflies and day-flying moth in a fragmented landscape. *Journal of Applied Ecology* 37(s1), 60–72.
Cox F. 1971. *Dichtebestimmung und Strukturanalyse von Pflanzenpopulationen mit Hilfe von Abstands-messungen: Ein Beitrag zur methodischen Weiterentwicklung von Verfahren für Verjüngungsinventuren*. Göttingen. Georg-August-Universität Göttingen. Cit. Neumann ja Starlinger (2001).
Cressie N. 1993. *Statistics for spatial data*. New York, Wiley.
Cressie N., Hawkins D.M. 1980. Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology* 12(2), 115–125.
Crovello T.J. 1981. Quantitative biogeography: an overview. *Taxon* 30(3), 563–575.
Crutsinger G.M., Collins M.D., Fordyce J.A., Gompert Z., Nice C.C., Sanders N.J. 2006. Plant genotypic diversity predicts community structure and governs an ecosystem process. *Science* 313(5789), 966–968.
Csillag F., Fortin M.-J., Dungan J.L. 2000. On the limits and extensions of the definition of scale. *Ecological Society of America Bulletin* 81(3), 230–232.
Czekanowski J. 1913. Zarys metod statystycznych w zastosowaniu do antropologii. *Prace Towarzystwa Naukowego Warszawskiego III. Wydzial nauk matematycznych i przyrodniczych* No 5. 1–228.Warszawa, Towarzystwo Naukowe Warszawskie.
Cumming G.S. 2000. Using between-model comparisons to fine-tune linear models of species ranges. *Journal of Biogeography* 27(2), 441–455.
Curran P.J., Milton E.J., Atkinson P.M., Foody G.M. 1998. Remote sensing: from data to understanding. Longley P., Brooks S., Macmillan W., McDonnell R. (toim). Geocomputation: a primer. Chichester. Wiley, 33–59.
Cushman S.A., McGarigal K., Neel M.C. 2008. Parsimony in landscape metrics: strength, universality, and consistency. *Ecological Indicators* 8(5), 691–703.
D'heygere T., Goethals P.L.M., De Pauw N. 2003. Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates. *Ecological Modelling* 160(3), 291–300.
Dahal R.K., Hasegawa S., Nonomura A., Yamanaka M., Masuda T., Nishino K. 2007. GIS-based weights-of-evidence modelling of rainfall-induced landslides in small catchments for landslide susceptibility mapping. *Environmental Geology* 54(2), 311–324.
Dale M.R.T., Blunton D.J., MacIsaac D.A., Thomas A.G. 1991. Multiple species effects and spatial autocorrelation in detecting species associations. *Journal of Vegetation Science* 2(5), 635–642.
Dale M.R.T., MacIsaac D.A. 1989. New methods for the analysis of spatial pattern in vegetation. *Journal of Ecology* 77(1), 78–91.
Dale M.R.T., Mah M. 1998. The use of wavelets for spatial pattern analysis in ecology. *Journal of Vegetation Science* 9(6), 805–814.
Dale M.R.T., Powell R.D. 2001. A new method for characterizing point patterns in plant ecology. *Journal of Vegetation Science* 12(5), 597–608.
Dauber J., Hirsch M., Simmering D., Waldhardt R., Otte A., Wolters V. 2003. Landscape structure as an indicator of biodiversity: matrix effects on species richness. *Agriculture, Ecosystems and Environment* 98(1-3), 321–329.
Davey S.M., Stockwell D.R.B. 1991. Incorporating wildlife habitat into an AI environment: concepts, theory and practicalities. *A.I. Applications in Natural Resource Management* 5(2), 59–104.
David F.N., Moore P.G. 1954. Notes on contagious distributions in plant populations. *Annals of Botany* 18(1), 47–53.
Davis A.J., Lawton J.H., Shorrocks B., Jenkinson L.S. 1998. Individualistic species responses invalidate simple physiological models of community dynamics under global environmental change. *Journal of Animal Ecology* 67(4), 600–612.
Davis J.H., Howe R.W., Davis G.J. 2000. A multi-scale spatial analysis method for point data. *Landscape Ecology* 15(2), 99–114.
de Almeida J.A. 2010. Stochastic simulation methods for characterization of lithoclasses in carbonate reservoirs. *Earth-Science Reviews* 101(3-4), 250–270.
de Bruin S. 2000. Predicting the areal extent of land-cover types using classified imagery and geostatistics. *Remote Sensing of the Environment* 74(3), 387–396.
De Cáceres M., Legendre P. 2009. Associations between species and groups of sites: indices and statistical inference. *Ecology* 90(12), 3566–3574.
de Knegt H.J., van Langevelde F., Coughenour M.B., Skidmore A.K., de Boer W.F., Heitkönig I.M.A.,

Knox N.M., Slotow R., van der Waal C., Prins H.H.T. 2010. Spatial autocorrelation and the scaling of species–environment relationships. *Ecology* 91(8), 2455–2465.
De Mas E., Chust G., Pretus J.L., Ribera C. 2009. Spatial modelling of spider biodiversity: matters of scale. *Biodiversity and Conservation* 18(7), 1945–1962.
De'ath G. 2002. Multivariate regression trees: a new technique for modeling species-environment relationships. *Ecology* 83(4), 1105–1115.
De'ath G., Fabricius K.E. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* 81(11), 3178–3192.
Debeljak M., Džeroski S., Adamič M. 1999. Interactions among the red deer (*Cervus elaphus*, L.) population, meteorological parameters and new growth of the natural regenerated forest in Snežnik, Slovenia. *Ecological Modelling* 121(1), 51–61.
Debeljak M., Dzeroski S., Jerina K., Kobler A., Adamic M. 2001. Habitat suitability modelling for red deer (*Cervus elaphus* L.) in South-central Slovenia with classification trees. *Ecological Modelling* 138(1), 321–330.
Deichman U., Anselin L. 1994. Eploratory spatial data analysis of categorical variables: an application to African farming systems data. EGIS/MARI. Conference Proceedings. Fifth European Conference and Exhibition on Geographical Information Systems. Vol. 1. Utrecht, Amsterdam, 2107–2116.
Dekruger D., Hunt B.R. 1994. Image-processing and neural networks for recognition of cartographic area features. *Pattern Recognition* 27(4), 461–483.
Del Fatto V. 2009. *Visual summaries of Geographic Databases by Chorems*. Ph.D Thesis. University of Salerno, Italy. INSA of Lyon, France. http://liris.cnrs.fr/Documents/Liris-4346.pdf.
Del Fatto V., Laurini R., Lopez K., Loreto R., Milleret-Raffort F., Sebillo M., Sol-Martinez D., Vitiello G., 2007. Potentialities of Chorems as Visual Summaries of Geographic Databases Contents. Qiu G. *et al*. (toim.) Visual 2007, 9th Conference on Visual Information Systems, LNCS 4781, 537–548.
Del Fatto V., Laurini R., Lopez K., Sebillo M., Vitiello G. 2008. A chorem-based approach for visually synthesizing complex phenomena**.** *Information Visualization* 7(3-4), 253–264.
Delcourt H.R., Delcourt P.A. 1996. Presettlement landscape heterogeneity: evaluating grain of resolution using General Land Office Survey data. *Landscape Ecology* 11(6), 363–381.
Dennis R.L.H., Shreeve T.G. 2003. Gains and losses of French butterflies: tests of predictions, under-recording and regional extinction from data in a new atlas. *Biological Conservation* 110(1), 131–139.
Dennis R.L.H., Shreeve T.G., Sparks T., Lhonore J.E. 2002. A comparison of geographical and neighbourhood models for improving atlas databases. The case of the French butterfly atlas. *Biological Conservation* 108(2), 143–159.
Dennis R.L.H., Sparks T.H., Hardy P.B. 1999. Bias in butterfly distribution maps: the effects of sampling effort. *Journal of Insect Conservation* 3(1), 33–42.
Deutsch C.V. 1992. *Annealing techniques applied to reservoir modeling and the integration of geological and engineering (well test) data*. Doctoral dissertation, Stanford University. http://www.ualberta.ca/~cdeutsch/Phdthesis.pdf.
Deutsch C.V., Cockerham P.W. 1994. Practical considerations in the application of simulated annealing to stochastic simulation. *Mathematical Geology* 26(1), 67–82.
Deutsch C.V., Journel A.G. 1998. *GSLIB: geostatistical software library and user's guide*. New York, Oxford University Press.
Devillers R., Jeansoulin R. 2006. *Fundamentals of spatial data quality*. ISTE Ltd.
Diamond J. 1975. The island dilemma – lessons of modern biogeographic studies for the design of nature reserves. *Biological Conservation* 7(2), 129–145.
Dice L.R. 1945. Measures of the amount of ecological association between species. *Ecology* 26(3), 297–302.
Diggle P.J. 1979a. On parameter estimation and gooness-of-fit testing for spatial point patterns. *Biometrics* 35(1), 87–101.
Diggle P.J. 1979b. Statistical methods for spatial point patterns in ecology. Cormack R.M., Ord J.K. (Eds). Spatial and Temporal Analysis in Ecology. International Co-operative Publishing House, Fairland, Maryland, USA. 95–150.
Diggle P.J. 1983. *Statistical analysis spatial point patterns*. London. Academic Press.
Diggle P.J. 2003. *Statistical analysis of spatial point patterns* (2nd ed). London. Arnold.
Dilts T.E., Sibold J.S., Biondi F. 2009. A weights-of-evidence model for mapping the probability of fire occurrence in Lincoln county, Nevada. *Annals of the Association of American Geographers* 99(4), 712–727.
Diniz-Filho J.A.F., Bini L.M., Hawkins B.A. 2003. Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology & Biogeography* 12(1), 53–64.
Diniz-Filho J.A.F., Bini L.M., Rangel T.F., Loyola R.D., Hof C., Nogués-Bravo D., Araújo M.B. 2009. Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. *Ecography* 32(6), 897−906.
Diniz-Filho J.A.F., Malaspina O.F.F. 1995. Evolution and population structure of Africanized honey bees in

Brazil: evidence from spatial analysis of morphometric data. *Evolution* 49(6), 1172–1179.
Diniz-Filho J.A.F., Nabout J.C., Bini L.M., Loyola R. D., Rangel T.F., Nogues-Bravo D., Araújo M.B. 2010. Ensemble forecasting shifts in climatically suitable areas for *Tropidacris cristata* (Orthoptera: Acridoidea: Romaleidae). *Insect Conservation and Diversity* 3(3), 213−221.
Diodato N., Tartari G., Bellocchi G. 2010. Geospatial rainfall modelling at eastern Nepalese Highland from ground environmental data. *Water Resource Management* 24(11), 2703–2720.
Dirnböck T., Dullinger S., Gottfried M., Ginzler C., Grabherr G. 2003. Mapping alpine vegetation based on image analysis, topographic variables and Canonical Correspondence Analysis. *Applied Vegetation Science* 6(1), 85−96.
Dixon P.M. 1994. Testing spatial segregation using a nearest-neighbor contingency table. *Ecology* 75(7), 1940–1948.
Doak D.F, Marino P.C., Kareiva P.M. 1992. Spatial scale mediates the influence of habitat fragmentation on dispersal success: implications for conservation. *Theoretical Population Biology* 41(3), 315–336.
Dobzhansky T. 1950. Evolution in the tropics. *American Scientist* 38(2), 209–211.
Dogan H.M., Dogan M. 2006. A new approach to diversity indices – modeling and mapping plant biodiversity of Nallihan (A3-Ankara/Turkey) forest ecosystem in frame of geographic information systems. *Biodiversity and Conservation* 15(3), 855–878.
Dormann C.F. 2007a. Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography* 16(2), 129–138.
Dormann C.F. 2007b. Promising the future? Global change projections of species distributions. *Basic and Applied Ecology* 8(5), 387–397.
Dormann C.F., McPherson J.M., Araújo M.B., Bivand R., Bolliger J., Carl G., Davies R.G., Hirzel A., Jetz W., Kissling W.D., Kühn I., Ohlemüller R., Peres-Neto P.R., Reineking B., Schröder B., Schurr F.M., Wilson R. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* 30(5), 609–628.
Douglas D., Peucker T. 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *The Canadian Cartographer* 10(2), 112–122.
Douglas M.E., Endler J.A. 1982. Quantitative matrix comparisons in ecological and evolutionary investigations. *Journal of Theoretical Biology* 99(4), 777–795.
Dowd P.A., Pardo-Igúzquiza E., Xu C. 2003. Plurigau: a computer program for simulating spatial facies using the truncated plurigaussian method. *Computers and Geosciences* 29(2), 123–141.
Drake J.M., Randin C., Guisan A. 2006. Modelling ecological niches with support vector machines. Journal of *Applied Ecology* 43(3), 424–432.
Drake J.B., Weishampel J.F. 2001. Simulating vertical and horizontal multifractal patterns of a longleaf pine savanna. *Ecological Modelling* 145(2-3), 129–142.
Draper D. 1995. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Society Series B* 57(1), 45–97.
Dray S., Legendre P., Peres-Neto P.R. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological Modelling* 196(3-4), 483–493.
Džeroski S. 2001. Applications of symbolic machine learning to ecological modelling. *Ecological Modelling* 146(1-3), 263–273.
Džeroski S, Drumm D. 2003. Using regression trees to identify the habitat preference of the sea cucumber (*Holothuria leucospilota*) on Rarotonga, Cook Islands. *Ecological Modelling* 170(2-3), 219–226.
Dudík M., Schapire R., Phillips S. 2006. Correcting sample selection bias in maximum entropy density estimation. Weiss Y., Schölkopf B., Platt J. (toim). Advances in Neural Information Processing Systems 18. Cambridge, MA: MIT Press, 323–330.
Duncan P. 1983. Determinants of the use of habitat by horses in a Mediterranean wetland. *Journal of Animal Ecology* 52(1), 93–109.
Duncan R.P. 1991. Competition and the coexistence of species in a mixed podocarp stand. *Journal of Ecology* 79(4), 1073–1084.
Duncan R.P. 1995. A correction for including competitive asymmetry in measure of local interference in plant populations. *Oecologia* 103(3), 393–396.
Dungan J. 1998. Spatial prediction of vegetation quantities using ground and image data. *International Journal of Remote Sensing* 19(2), 267–285.
Dungan J.L., Perry J.N., Dale M.R.T., Legendre P., Citron-Pousty S., Fortin M.-J., Jakomulshka A., Miriti M., Rosenberg M.S. 2002. A balanced view of scale in spatial analysis. *Ecography* 25(5), 626–640.
Durão R.M., Pereira M.J., Branquinho C., Soares A. 2010. Assessing spatial uncertainty of the Portuguese fire risk through direct sequential simulation. *Ecological Modelling* 221(1), 27–33.
Dutilleul P., Stockwell J.D., Frigon D., Legendre P. 2000. The Mantel test versus Pearson's correlation analysis: assessment of the differences for biological and environmental studies. *Journal of Agricultural,*

Biological, and Environmental Statistics 5(2), 131–150.
Dymond C.C., Mladenoff D.J., Radeloff V.C. 2002. Phenological differences in Tasseled Cap indices improve deciduous forest classification. *Remote Sensing of Environment* 80(3), 460–472.
Eberhart L.L. 1967. Some developments in 'distance sampling'. *Biometrics* 23(2), 207–216.
Edwards G., Landry R., Thompson K.P.B. 1988. Texture analysis of forest regeneration sites in high-resolution SAR imagery. Proceedings of the International Geosciences and Remote Sensing Symposium (IGARSS 88), ESA SP-284. Paris, European Space Agency, 1355–1360.
Edwards T.C.Jr., Cutler D.R., Zimmermann N.E., Geiser L., Moisen G.G. 2006. Effects of sample survey design on the accuracy of classification tree models in species distribution models. *Ecological Modelling* 199(2), 132–141.
Ehlschlaeger C.R. 2000. Representing uncertainty of area class maps with a correlated inter-map cell swapping heuristic. *Computers, Environment and Urban Systems* 24(5), 451–469.
Eliassen A. 1954. *Provisional report on calculation of spatial covariance and autocorrelation of the pressure field*. Report 5. Videnskaps-Akademiet, Institut for Vaer och Klimaforskning. Oslo, Norway. Cit. Gelfand *et al.* 2010.
Elith J., Burgman M.A., Regan H.M. 2002. Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecological Modelling* 157(2-3), 313–329.
Elith J., Graham C.H. 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography* 32(1), 66–77.
Elith J., Graham C.H., Anderson R.P., Dudík M., Ferrier S., Guisan A., Hijmans R.J, Huettmann F., Leathwick J.R., Lehmann A., Li J., Lohmann L.G., Loiselle B.A., Manion G., Moritz C., Nakamura M., Nakazawa Y., Overton J.McC., Peterson A.T., Phillips S.J., Richardson K., Scachetti-Pereira R., Schapire R.E., Soberón J., Williams S., Wisz M.S., Zimmermann N.E. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29(2), 129–151.
Elith J., Leathwick J. 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions* 13(3), 265–275.
Elith J., Phillips S., Hastie T., Dudik M., Chee Y., Yates C. 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions* 17(1), 43–57.
Ellenberg H. 1948. Unkrautgesellschaften als Mass für den Säuregrad, die Verdichtung und andere Eigenschaften des Ackerbodens. *Berliner Landtechnik* 4, 130–146.
Ellenberg H., Weber H.E., Dull R., Wirth V., Werner W., Paulissen D. 1992. Zeigerwerte von Pflanzen in Mitteleuropa (2. Auflage). *Scripta Geobotanica* 18, 1–258.
Emelyanova I.V., Donald G.E., Miron D.J., Henry D.A., Garner M.G. 2009. Probabilistic modelling of cattle farm distribution in Australia. *Environmental Modeling and Assessment* 14(4), 449–465.
Engler R., Guisan A., Rechsteiner L. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology* 41(2), 236–274.
Escudero A., Iriondo J.M., Torres E.M. 2003. Spatial analysis of genetic diversity as a tool for plant conservation. *Biological Conservation* 113(3), 351–365.
Estes L.D., Reillo P.R., Mwangi A.G., Okin G.S., Shugart H.H. 2010. Remote sensing of structural complexity indices for habitat and species distribution modeling. *Remote Sensing of Environment* 114(4), 792–804.
Estes L.D., Mwangi A.G., Reillo P.R., Shugart H.H. 2011. Predictive distribution modeling with enhanced remote sensing and multiple validation techniques to support mountain bongo antelope recovery. *Animal Conservation* 14(5), 521–532.
Estrada-Peña A. 2001. Forecasting habitat suitability for ticks and prevention of tick-borne diseases. *Veterinary Parasitology* 98(1-3), 111–132.
Ewers R.M., Didham R.K., Wratten S.D., Tylianakis J.M. 2005. Remotely sensed landscape heterogeneity as a rapid tool for assessing local biodiversity value in a highly modified New Zealand landscape. *Biodiversity and Conservation* 14(6), 1469–1485.
Falk W., Mellert K.H. 2011. Species distribution models as a tool for forest management planning under climate change: risk evaluation of *Abies alba* in Bavaria. *Journal of Vegetation Science* 22(4), 621–634.
Fall A., Fall J. 2001. A domain-specific language for models of landscape dynamics. *Ecological Modelling* 141(1-3), 1–18.
Feilhauer H., Schmidtlein S. 2009. Mapping continuous fields of forest alpha and beta diversity. *Applied Vegetation Science* 12(4), 429–439.
Feng Y., Liu Y., Tong X., Liu M., Deng S. 2011. Modeling dynamic urban growth using cellular automata and particle swarm optimization rules. *Landscape and Urban Planning* 102(3), 188–196.
Fernandez-Duque E. 1997. Comparing and combining data across studies: alternatives to significance testing. *Oikos* 79(3), 616–618.

Ferrier S. 2002. Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology* 51(2), 331–363.
Ferrier S., Drielsma M., Manion G., Watson G. 2002b. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling. *Biodiversity and Conservation* 11(12), 2309–2338.
Ferrier S., Guisan A. 2006. Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology* 43(3), 393–404.
Ferrier S., Manion G., Elith J., Richardson K. 2007. Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity and Distributions* 13(3), 252–264.
Ferrier S., Watson G., Pearce J., Drielsma M. 2002a. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. *Biodiversity and Conservation* 11(12), 2275–2307.
Ferris R., Peace A.J., Humphrey J.W., Broome A.C. 2000. Relationships between vegetation, site type and stand structure in coniferous plantations in Britain. *Forest Ecology and Management* 136(1-3), 35–51.
Finch J.M., Samways M.J., Hill T.R., Piper S.E., Taylor S. 2006. Application of predictive distribution modelling to invertebrates: Odonata in South Africa. *Biodiversity and Conservation* 15(13), 4239–4251.
Fink A., Kosecoff J., Chassin M., Brook R.H. 1984. Consensus methods: characteristics and guidelines for use. *American Journal of Public Health* 74(9), 979–983.
Fischer M.M., Getis A. (toim.) 2010. *Handbook of applied spatial analysis. Software tools, methods and applications*. Springer.
Fisher R.A., Thornton H.G., Mackenzie W.A. 1922. The accuracy of the plating method of estimating the density of bacterial populations, with particular reference to the use of Thornton's agar medium with soil samples. *Annals of Applied Botany* 9(3-4), 325–359.
Fisher W.D. 1958. On grouping for maximum homogeneity. *Journal of American Statistical Association* 53(284), 789–798.
Fitzpatrick M.C., Hargrove W.W. 2009. The projection of species distribution models and the problem of non-analog climate. *Biodiversity and Conservation* 18(8), 2255–2261.
Fjerdingstad E. 1964. Pollution of streams estimated by benthal phytomicro-organisms I. A saprobic system based on communities of organisms and ecological factors. *Internationale Revue der gesamten Hydrobiologie und Hydrographie* 49(1), 63–131.
Fleming K.K., Didier K.A., Miranda B.R., Porter W.F. 2004. Sensitivity of a white-tailed deer habitat-suitability index model to error in satellite land-cover data: implications for wildlife habitat-suitability studies. *Wildlife Society Bulletin* 32(1), 158–168.
Florinsky I.V., Eilers R.G., Manning G.R., Fuller L.G. 2002. Prediction of soil properties by digital terrain modelling. *Environmental Modelling & Software* 17(3), 295–311.
Flower R.J., Juggins S., Battarbee R.W. 1997. Matching diatom assemblages in lake sediment cores and modern surface sediment samples: the implications for lake conservation and restoration with special reference to acidified systems. *Hydrobiologia* 344(1), 27–40.
Folse L.J., Packard J.M., Grant W.E. 1989. AI modelling of animal movements in a heterogeneous habitat. *Ecological Modelling* 46(1-2), 57–72.
Foody G.M. 1999. Applications of the self-organizing feature map neural network in community data analysis. *Ecological Modelling* 120(2-3), 97–107.
Foody G.M. 2004. Spatial nonstationarity and scale-dependency in the relationship between species richness and environmental determinants for the sub-Saharan endemic avifauna. *Global Ecology and Biogeography* 13(4), 315–320.
Foody G.M. 2008. Harshness in image classification accuracy assessment. *International Journal of Remote Sensing* 29(11), 3137–3158.
Foody G.M., Cutler F.M.J. 2006. Mapping the species richness and composition of tropical forests from remotely sensed data with neural networks. *Ecological Modelling* 195(1-2), 37–42.
Foody G.M., Mathur A. 2004. Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification. *Remote Sensing of Environment* 93(1-2), 107–117.
Forman R.T.T., Godron M. 1986. *Landscape Ecology*. New York, John Wiley and Sons.
Fortin M.-J. 1994. Edge detection algorithms for two-dimensional ecological data. *Ecology* 75(4), 956–965.
Fortin M.-J. 1999. Effects of sampling unit resolution on the estimation of spatial autocorrelation. *Ecoscience* 6(4), 636–641.
Fortin M.-J., Dale M.R.T. 2009. Spatial autocorrelation in ecological studies: a legacy of solutions and myths. *Geographical Analysis* 41(4), 392–397.
Fortin M.-J., Drapeau P., Legendre P. 1989. Spatial auto-correlation and sampling design in plant ecology. *Vegetatio* 83(1-2), 209–222.
Fortin M.-J., Gurevitch J. 1993. Mantel tests: spatial structure in field experiments. Sheiner S.M., Gurevitch J.

(toim). Design and Analysis of Ecological Experiments. Chapman & Hall, 342–359.
Fotheringham A., Brunsdon C., Charlton M. 2005. *Quantitative geography: perspectives on spatial data analysis*. London, Sage Publications.
Fotheringham A., Charlton M., Brunsdon C. 1998. Geographically weighted regression: a natural extension of the expansion method for spatial data analysis. *Environment and Planning A* 30(11), 1905–1928.
Fox J.C., Ades P.K., Bi H. 2001. Stochastic structure and individual-tree growth models. *Forest Ecology and Management* 154(1-2), 261–276.
Foxall R., Baddeley A. 2002. Nonparametric measures of association between a spatial point process and a random set. *Journal of the Royal Statistical Society Series C Applied Statistics* 51(2), 165–182.
Franco C., Soares A., Delgado J. 2006. Geostatistical modelling of heavy metal contamination in the topsoil of Guadiamar river margins (S Spain) using a stochastic simulation technique. *Geoderma* 136(3-4), 852–864.
Franco-Lopez H., Ek A.R., Bauer M.R. 2001. Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. *Remote Sensing of Environment* 77(3), 251–274.
Franklin J. 1995. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography* 19(4), 474–499.
Franklin J. 1998. Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. *Journal of Vegetation Science* 9(5), 733–748.
Franklin J. 2009. *Mapping species distributions. Spatial inference and prediction.* Cambridge etc., Cambridge University Press.
Franklin J., Wejnert K., Hathaway S., Rochester C., Fisher R. 2009. Effect of species rarity on the accuracy of species distribution models for reptiles and amphibians in southern California. *Diversity and Distributions* 15(1), 167–177.
Franklin S.E., Hall R.J., Moskal L.M., Maudie A.J., Lavigne M.B. 2000. Incorporating texture into classifcation of forest species composition from airborne multispectral images. *International Journal of Remote Sensing* 21(1), 61–79.
Freeman E., Ford E.D. 2002. Effects of data quality on analysis of ecological pattern using the K(d) function. *Ecology* 83(1), 35–46.
Frelich L.E., Calcote R.R., Davis M.B., Pastor J. 1993. Patch formation and maintenance in an old-growth hemlock-hardwood forest. *Ecology* 74(2), 513–527.
Frescino T.S., Edwards T.C.Jr., Moisen G.G. 2001. Modelling spatially explicit forest structural attributes using generalized additive models. *Journal of Vegetation Science* 12(1), 15–26.
Freund Y., Schapire R.E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55(1), 119–139.
Friedman J.H. 1991. Multivariate adaptive regression splines. *Annals of Statistics* 19(1), 1–141.
Friedman J.H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 29(5), 1189–1232.
Friedman J.H. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38(4), 367–378.
Friedman J.H., Kohavi R., Yun Y. 1996. Lazy decision trees. Proceedings of the Thirteenth National Conference on Artificial Intelligence. Portland, AAAI Press, 717–724.
Frysinger S.P. 2002. Integrative environmental modeling. Clarke K.C., Parks B.O., Crane M.P. (toim). Geographic Information Systems and Environmental Modeling. Prentice Hall Series in Geographic Information Science. New Jersey, Prentice Hall, 211–222.
Fuller R.M., Groom G.B., Jones A.R. 1994. The land cover map of Great Britain: an automated classification of Landsat Thematic Mapper data. *Photogrammetric Engineering and Remote Sensing* 60(5), 553–562.
Galton F. 1892. *Finger Prints.* London, Macmillan.
Gappa J.L., Calcagno J.A., Tablado A. 1997. Spatial pattern in a low-density population of barnacle *Balanus amphritrite* Darwin. *Hydrobiologia* 357(1-3), 129–137.
Gardner R.H., O'Neill R.V. 1991. Pattern, process, and predictability: the use of neutral models for landscape analysis. Turner M.G., Gardner R.H. (toim). Quantitative Methods in Landscape Ecology. New York, Springer, 289–307.
Gardner R.H., Milne B.T., Turner M.G., O'Neill R.V. 1987. Neutral models for the analysis of broad-scale landscape pattern. *Landscape Ecology* 1(1), 19–28.
Gardner R.H., Urban D.L. 2007. Neutral models for testing landscape hypotheses. *Landscape Ecology* 22(1), 15–29.
Garshelis D.L. 2000. Delusions in habitat evaluation: measuring use, selection, and importance. Boitani L., Fuller T.K. (toim). Research Techniques in Animal Ecology: Controversies and Consequences. New York, Columbia University Press, 111–164.
Garzón M.B., Blazek R., Neteler M., Sánchez de Dios R., Ollero H.S., Furlanello C. 2006. Predicting habitat suitability with machine learning models: the potential area of *Pinus sylvestris* L. in the Iberian Peninsula. *Ecological Modelling* 197(3-4), 383–393.

Gaucherel C. 2007. Multiscale heterogeneity map and associated scaling profile for landscape analysis. *Landscape and Urban Planning* 82(3), 95–102.
Gaucherel C., Fleury D., Auclair D., Dreyfus P. 2006. Neutral models for patchy landscapes. *Ecological Modelling* 197(1-2), 159–170.
Gavin J., Jennison C. 1997. A subpixel image restoration algorithm. *Journal of Computational and Graphical Statistics* 6(2), 182–201.
Geary R.C. 1947. Testing for normality. *Biometrika* 34(3/4), 209–242.
Geary R.C. 1954. The contiguity ratio and statistical mapping. *Incorporated Statistician* 5(3), 115–145.
Gelfand A. E., Diggle P. J., Fuentes M., Guttorp P. (toim.) 2010. *Handbook of spatial statistics*. Boca Raton. Chapman & Hall/CRC Press.
Gellrich M., Baur P., Koch B., Zimmermann N.E. 2007. Agricultural land abandonment and natural forest re-growth in the Swiss mountains: a spatially explicit economic analysis. *Agriculture, Ecosystems and Environment* 118(1-4), 93–108.
Gemmell F., Varjo J., Strandstrom M. 2001. Estimating forest cover in a boreal forest test site using thematic mapper data from two dates. *Remote Sensing of Environment* 77(2), 97–211.
Gething P.W., Noor A.M., Gikandi P.W., Ogara E.A.A., Hay S.I., Nixon M.S., Snow R.W., Atkinson P.M. 2006. Improving imperfect data from health management information systems in Africa using space–time geostatistics. *PLoS Medicine* 3(6), 825–831.
Getis A. 1984. Interaction modeling using second-order analysis. *Environment and Planning A* 16(2), 173–183.
Getis A., Franklin J. 1987. Second-order neighbourhood analysis of mapped point patterns. *Ecology* 68(3), 473–477.
Getis A., Ord J.K. 1992. The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24(3), 189–206.
Getis A., Ord J.K. 1996. Local spatial statistics: an overview. Longley P.A. ja Batty M. Spatial analysis Modelling in a GIS Environment. John Wiley. Cambridge, New York, 261–277.
Getzin S., Dean C., He F., Trofymow J.A., Wiegand K., Wiegand T. 2006. Spatial patterns and competition of tree species in a Douglas-fir chronosequence on Vancouver Island. *Ecography* 29(5), 671–682.
Getzin S., Worbesa M., Wiegand T., Wiegand K. 2011. Size dominance regulates tree spacing more than competition within height classes in tropical Cameroon. *Journal of Tropical Ecology* 27(1), 93–102.
Ghimire B., Rogan J., Miller J. 2010. Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the Getis statistic. *Remote Sensing Letters* 1(1), 45–54.
Gibson L., Barrett B., Burbidge A. 2007. Dealing with uncertain absences in habitat modelling: a case study of a rare ground-dwelling parrot. *Diversity and Distributions* 13(6), 704–713.
Gibson L.A., Wilson B.A., Aberton J.G. 2004. Landscape characteristics associated with species richness and occurrence of small native mammals inhabiting a coastal heathland: a spatial modelling approach. *Biological Conservation* 120(1), 75–89.
Gill D. 1970. Application of a statistical zonation method to reservoir evaluation and digitized log analyses. *Bulletin of the American Association of Petroleum Geologists* 54(5), 719–729.
Gillison A.N., Brewer K.R.W. 1985. The use of gradient directed transects or gradsects in natural resource surveys. *Journal of Environmental Management* 20, 103–127.
Giorgi G.M. 2005. Gini's scientific work: an evergreen. *METRON - International Journal of Statistics* 63(3), 299–315.
Gleason H.A. 1926. The Individualistic Concept of the Plant Association. *Bulletin of the Torrey Botanical Club* 53(1), 7–26.
Glenn E.M., Ripple W.J. 2004. On using digital maps to assess wildlife habitat. *Wildlife Society Bulletin* 32(3), 852–860.
Gogol-Prokurat M. 2011. Predicting habitat suitability for rare plants at local spatial scales using a species distribution model. *Ecological Applications* 21(1), 33–47.
Goldingay R., Possingham H. 1995. Area requirements for viable populations of the Australian gliding marsupial, *Petaurus australis*. *Biological Conservation* 73(2), 161–167.
Gong P., Pu R., Yu B. 1997. Conifer species recognition: an exploratory analysis of in situ hyperspectral data. *Remote Sensing of Environment* 62(2), 189–200.
Gontier M., Mörtberg U., Balfors B. 2010. Comparing GIS-based habitat models for applications in EIA and SEA. *Environmental Impact Assessment Review* 30(1) 8–18.
Goodall D.W. 1954. Objective methods for the classification of vegetation. III. An essay in the use of factor analysis. *Australian Journal of Botany* 2(3), 304–324.
Goodall D.W. 1974. A new method for the analysis of spatial pattern by random pairing of quadrats. *Vegetatio* 29(2), 135–146.
Goodchild M.F. 1986. Spatial autocorrelation, concepts and techniques in modern geography. Norwich,

GeoBooks.
Goodchild M.F. 1990. Algorithm 9: simulation of autocorrelation for aggregate data. *Environment and Planning A* 12(9), 1073–1081.
Goodchild M.F. 1993. Data models and data quality: problems and prospects. Goodchild M.F., Parks B.O., Steyaert L.T. (toim). Environmental Modelling with GIS. New York, Oxford University Press, 94–103.
Goodchild M.F., Lam N.S.N. 1980. Areal interpolation, a variant of the traditional spatial problem. *Geo-Processing* 1, 297–312.
Goovaerts P. 1997. *Geostatistics for natural resources evaluation.* Oxford University Press.
Goovaerts P. 1999. Geostatistics in soil science: state-of-the-art and perspectives. *Geoderma* 89(1-2),1–45.
Goovaerts P. 2000. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology* 228(1-2), 113–129.
Goovaerts P. 2001. Geostatistical modelling of uncertainty in soil science. *Geoderma* 103(1-2), 3–26.
Goreaud F., Pelissier R. 1999. On explicit formulas of edge effect correction for Ripley's K-function. *Journal of Vegetation Science* 10(3), 433–438.
Goreaud F., Loreau M., Millier C. 2002. Spatial structure and the survival of an inferior competitor: a theoretical model of neighbourhood competition in plants. *Ecological Modelling* 158(1-2), 1–19.
Gotway C.A., Stroup W.W. 1997. A generalized linear model approach to spatial data analysis and prediction. *Journal of Agricultural, Biological & Environmental Statistics* 2(2), 157–178.
Gower J.C. 1971a. A general coefficient of similarity and some of its properties. *Biometrics* 27(4), 857–871.
Gower J.C. 1971b. Statistical methods of comparing different multivariate analyses of the same data. Hodson F.R., Kendall D.G., Tautu P. (toim). Mathematics in the archaeological and historical sciences. Edinburgh, Edinburgh University Press, 138–149.
Grace J.B., Pugesek B.H. 1997. A structural equation model of plant species richness and its application to a coastal wetland. *American Naturalist* 149(3), 436–460.
Graham C.H., Elith J., Hijmans R., Guisan A., Peterson A.T., Loiselle B.A., NCEAS species distribution modelling group 2008. Evaluating the influence of spatial uncertainty in locality points for species distributional modeling. *Journal of Applied Ecology* 45(1), 239–247.
Grau H.R. 2000. Regeneration patterns of *Cedrela lilloi* (*Meliaceae*) in northwestern Argentina subtropical montane forests. *Journal of Tropical Ecology* 16(2), 227–242.
Grebby S., Cunningham D., Naden J., Tansey K. 2010. Lithological mapping of the Troodos ophiolite, Cyprus, using airborne LiDAR topographic data. *Remote Sensing of Environment* 114(4), 713–724.
Green R.H. 1966. Measurement of non-randomness in spatial distributions. *Researches on Population Ecology* 8(1), 1–7.
Greenacre M.J. 2007. *Correspondence analysis in practice, second edition.* Boca Raton, Chapman & Hall.
Greig-Smith P. 1952. The use of random and contiguous quadrats in the study of the structure of plant communities. *Annals of Botany* 16(2), 293–316.
Greig-Smith P. 1961. Data on pattern within plant communities: I. The analysis of pattern. *Journal of Ecology* 49(3), 695–702.
Greig-Smith P. 1964. *Qantitative plant ecology*. London. Butterworth.
Griffin S.C., Taper M.L., Hoffman R., Mills L.S. 2010. Ranking Mahalanobis distance models for predictions of occupancy from presence-only data. *Journal of Wildlife Management* 74(5), 1112–1121.
Griffith D., Layne L. 1999. *A casebook for spatial statistical analysis. A compilation of analyses of different thematic data sets.* New York, Oxford University Press.
Griffith D.A., Peres-Neto P.R. 2006. Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. *Ecology* 87(10), 2603–2613.
Griffiths G.H., Smith J.M., Veitch N., Aspinall R. 1993. The ecological interpretation of satellite imagery with special reference to bird habitats. Haines-Young R., Green D.R., Cousins S. (toim). Landscape Ecology and Geographic Information Systems. London, New York, Philadelphia, Taylor & Francis 255–272.
Griffiths G.H., Vogiatzakis I.N., Porter J.R., Burrows C. 2011. A landscape scale spatial model for semi-natural broadleaf woodland expansion in Wales, UK. *Journal for Nature Conservation* 19(1), 43–53.
Grimm E.C. 1984. Fire and other factors controlling the Big Woods vegetation of Minnesota in the mid-nineteenth century. *Ecological Monographs* 54(3), 291–311.
Grubb T.G., King R.M. 1991. Assessing human disturbance of breeding bald eagles with classification tree models. *Journal of Wildlife Management* 55(3), 500–511.
Gu W.-D., Kuusinen M., Konttinen T., Hanski I. 2001. Spatial pattern in the occurrence of lichen *Lobaria pulmonaria* in managed and virgin boreal forests. *Ecography* 24(2), 139–150.
Guénard G., Legendre P., Boisclair D., Bilodeau M. 2010. Multiscale codependence analysis: an integrated approach to analyze relationships across scales. *Ecology* 91(10), 2952–2964.
Guido M., Gianelle D. 2001. Distribution patterns of four Orthoptera species in relation to microhabitat heterogeneity in an ecotonal area. *Acta Oecologica* 22(3), 175–185.

Guinan J., Brown C., Dolan M.F.J., Grehan A.J. 2009. Ecological niche modelling of the distribution of cold-water coral habitat using underwater remote sensing data. *Ecological Informatics* 4 (2), 83–92.
Guisan A., Broennimann O., Engler R., Yoccoz N.G., Vust M., Zimmermann N.E., Lehman A. 2006. Using niche-based models to improve the sampling of rare species. *Conservation Biology* 20(2), 501–511.
Guisan A., Edwards T.C.Jr., Hastie T. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling* 157(2-3), 89–100.
Guisan A., Graham C., Elith J., Huettmann F., the NCEAS Species Distribution Modelling Group 2007. Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions* 13(3), 332–340.
Guisan A., Zimmermann N.E. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling* 135(2-3), 147–186.
Guisan A., Theurillat J.-P. 2000. Equilibrium modelling of alpine plant distribution: how far can we go? *Phytocoenologia* 30(3-4), 353–384.
Guisan A., Theurillat J.-P., Kienast F. 1998. Predicting the potential distribution of plant species in an alpine environment. *Journal of Vegetation Science* 9(1), 65–74.
Guisan A., Thuiller W. 2005. Predicting species distributions: offering more than simple habitat models. *Ecology Letters* 8(9), 993–1009.
Gullison J.J., Bourque C.P.-A. 2001. Spatial prediction of tree and shrub succession in a small watershed in Northern Cape Breton Island, Nova Scotia, Canada. *Ecological Modelling* 137(2-3), 181–199.
Gumpertz M.L., Wu C., Pye J.M. 2000. Logistic regression for southern pine beetle outbreaks with spatial and temporal autocorrelation. *Forest Science* 46(1), 95–107.
Guo Q., Kelly M., Graham C.H. 2005. Support vector machines for predicting distribution of Sudden Oak Death in California. *Ecological Modelling* 182(1), 75–90.
Guralnick R.P., Hill A.W. 2009. Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. *Bioinformatics* 25(4), 421–428.
Guralnick R.P., Hill A.W., Lane M. 2007. Towards a collaborative, global infrastructure for biodiversity assessment. *Ecological Letters* 10(8), 663–672.
Gustafson E.J., Parker G.P. 1992. Relationships between landcover proportion and indices of landscape spatial pattern. *Landscape Ecology* 7(2), 101–110.
Gustafson E.J., Shifley S.R., Mladenoff D.J., Nimerfro K.K., He H.S. 2000. Spatial simulation of forest succession and timber harvesting using LANDIS. *Canadian Journal of Forest Research* 30(1), 32–43.
Gömöry D., Longauer R., Paule L., Krajmerová D., Schmidtová J. 2010. Across-species patterns of genetic variation in forest trees of Central Europe. *Biodiversity and Conservation* 19(7), 2025–2038.
Haase P. 1995. Spatial pattern analysis in ecology based on Ripley's K-function: introduction and methods of edge correction. *Journal of Vegetation Science* 6(4), 575–582.
Hagen-Zanker A. 2006. Map comparison methods that simultaneously address overlap and structure. *Journal of Geographical Systems* 8(2), 165–185.
Hagen-Zanker A. 2009. An improved Fuzzy Kappa statistic that accounts for spatial autocorrelation. *International Journal of Geographical Information Science* 23(1), 61–73.
Hagen-Zanker A., Lajoie G. 2008. Neutral models of landscape change as benchmarks in the assessment of model performance. *Landscape and Urban Planning* 86(3-4), 284–296.
Hagen-Zanker A., Straatman B., Uljee I. 2005. Further developments of a fuzzy set map comparison approach. *International Journal of Geographical Information Science* 19(7), 769–785.
Haggett P. 2001. *Geography: a Global Synthesis*. Harlow, Prentice Hall.
Haining R.P. 1990. *Spatial data analysis in Social and Environmental Sciences*. Cambridge, Cambridge University Press.
Haining R. 2003. *Spatial data analysis. Theory and practice*. Cambridge, Cambridge University Press.
Hall R.J., Skakun R.S., Arsenault E.J., Case B.S. 2006. Modeling forest stand structure attributes using Landsat ETM+ data: Application to mapping of aboveground biomass and stand volume. *Forest Ecology and Management* 225(1-3), 378–390.
Halley J.M. 1996. Ecology, evolution and l/f-noise. *TREE* 1(1), 33–37.
Halley J.M., Hartley S., Kallimanis A.S., Kunin W.E, Lennon J.J., Sgardelis S.P. 2004. Uses and abuses of fractal methodology in ecology. *Ecology Letters*, 7(3), 254–271.
Hanley J.A., McNeil B.J. 1982. The meaning and use of the area under reciever operating characteristics (ROC) curve. *Radiology* 143(1), 29–36.
Hanley J.A., McNeil B. 1983. A method comparing the areas under reciever operating characteristic curves derived from the same cases. *Radiology* 148(3), 839–843.
Hannachi A., Jolliffe I.T., Stephenson D.B. 2007. Empirical orthogonal functions and related techniques in atmospheric science: a review. *International Journal of Climatology* 27(9), 1119–1152.
Hansson S. 1984. Competition as a factor regulating the geographical distribution of fish species in a Baltic

Archipelago: a neutral model analysis. *Journal of Biogeography* 11(5), 367–381.
Hansen M.J., Franklin S.E., Woudsma C.G., Peterson M. 2001. Caribou habitat mapping and fragmentation analysis using Landsat MSS, TM, and GIS data in the North Columbia Mountains, British Columbia, Canada. *Remote Sensing of the Environment* 77(1), 50–65.
Hansen T.M., Mosegaard K. 2008. VISIM : sequential simulation for linear inverse problems. *Computers and Geosciences* 34(1), 53–76.
Hansen T.M., Mosegaard K. 2011. Corrigendum to ‘‘VISIM: sequential simulation for linear inverse problems’’ [Comput. Geosci. 34(1) (2008) 53–76]. *Computers and Geosciences* 37(7), 973–974.
Hanski I. 1982. Dynamics of regional distribution: the core and satellite species hypothesis. *Oikos* 38(2), 210–221.
Hanski I. 1994. A practical model of metapopulation dynamics. *Journal of Animal Ecology* 63(1), 151–162.
Hanski I. 1997. Predictive and practical population models: the incidence function approach. Tilman D., Kareiva P. (toim). Spatial ecology: the role of space in population dynamics and interspecific interactions. Monographs in Population Biology 30. Princeton University Press. Princeton, New Jersey, USA, 21–25.
Hanski I. 1998. Metapopulation dynamics. *Nature* 396(6706), 41–49.
Hanski I. 1999. *Metapopulation Ecology.* Oxford Series in Ecology and Evolution. Oxford, Oxford University Press.
Hanus M.L., Hann D.W., Marshall D.D. 1998. Reconstructing the spatial pattern of trees from routine stand examination. *Forest Science* 44(1), 125–133.
Hao Z., Zhang J., Song B., Te J., Li B. 2007. Vertical structure and spatial associations of dominant tree species in an old-growth temperate forest. *Forest Ecology and Management* 252(1-3), 1–11.
Haralick R.M. 1979. Statistical and structural approaches to texture. *The Proceedings of the IEEE* 67(5), 786–804.
Haralick R.M., Shapiro L.G. 1985. Image segmentation techniques. *Computer Vision, Graphics and Image Processing* 29(1), 100–132.
Haralick R.M., Shanmugam K., Dinstein I. 1973. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*. SMC-3(6), 610–621.
Harell F.E., Lee K.L., Mark D.B. 1996. Multivariable prognostic models: issues in developing models, evalutating assumptions, and measuring and reducing errors. *Statistics and Medicine* 15(4), 361–387.
Hargis C.D., Bissonette J.A., David J.L. 1998. The behaviour of landscape metrics commonly used in the study of habitat fragmentation. *Landscape Ecology* 13(1992), 167–186.
Hargrove W.W., Hoffman F.M. 2005. The potential of multivariate quantitative methods for delineation and visualization of ecoregions. *Environmental Management* 34(s1), 39–60.
Harkness R.D., Isham V.S. 1983. A bivariate spatial point pattern of ants' nests. *Applied Statistics* 32(3), 293–303.
Harrison S., Bruna E. 1999. Habitat fragmentation and large-scale conservation: what do we know for sure? *Ecography* 22(3), 225–232.
Hassal A.N. 1850. *Microscopic examination of the water supplied to the inhabitants of London.* London, Samuel Highley. Cit. Абакумов (1981).
Hastie T., Tibshirani R. 1990. *Generalized Additive Models*. Chapman and Hall, London.
Hay G. J., Castilla G., Wulder M. A., Ruiz J. R. 2005. An automated object-based approach for the multiscale image segmentation of forest scenes. *International Journal of Applied Earth Observation and Geoinformation* 7(4), 339–359.
Hay G.J., Castilla G. 2008. Geographic object-based image analysis (GEOBIA): a new name for a new discipline. Blaschke T., Lang S., Hay G. (toim). Object Based Image Analysis. Springer, Heidelberg, Berlin, New York, 93–112.
Haydon D., Steen H. 1997. The effects of large- and small-scale random events on the synchrony of metapopulation dynamics: a theoretical analysis. *Proceedings of the Royal Society of London Series B* 264(1386), 1375–1381.
He B., Chen C., Liu Y. 2010. Gold resources potential assessment in eastern Kunlun mountains of China combining weights-of-evidence model with GIS Spatial Analysis Technique. *Chinese Geographical Science* 20(5), 461–470.
He F., Duncan R.P. 2000. Density-dependent effects on tree survival in an old-growth Douglas fir forest. *Journal of Ecology* 88(4), 676–688.
He H.S., DeZonia B.E., Mladenoff D.J. 2000. An aggregation index (AI) to quantify spatial patterns of landscapes. *Landscape Ecology* 15(7), 591–601.
He H.S., Mladenoff D.J. 1999. Spatially explicit and stochastic simulation of forest-landscape fire disturbance and succession. *Ecology* 80(1), 81–99.
He H.S., Mladenoff D.J., Boeder J. 1999. An object-oriented forest landscape model and its representation of tree species. *Ecological Modelling* 119(1), 1–19.

Hearn S.M., Healey J.R., McDonald M.A., Turner A.J., Wong J.L.G., Stewart G.B. 2011. The repeatability of vegetation classification and mapping. *Journal of Environmental Management* 92(4), 1174–1184.
Heidke P. 1926. Berechnung des Erfolges und der Gute der Windstarkevorhersagen im Sturmwarnungsdienst. *Geografiske Annaler* 8, 301–349.
Heikkilä J., Nevalainen S., Tokola T. 2002. Estimating defoliation in boreal coniferous forests by combining Landsat TM, aerial photograps and field data. *Forest Ecology and Management* 158(1-3), 9–23.
Heikkinen R.K., Luoto M., Araújo M.B., Virkkala R., Thuiller W., Sykes M.T. 2006. Methods and uncertainties in bioclimatic envelope modelling under climate change. *Progress in Physical Geography* 30(6), 751–777.
Heikkinen R.K., Luoto M., Virkkala R., Pearson R.G., Körber J.-H. 2007. Biotic interactions improve prediction of boreal bird distributions at macro-scales. *Global Ecology and Biogeography* 16(6), 754–763.
Heilbron D. 1994. Zero-altered and other regression models for count data with added zeros. *Biometrical Journal* 36(5), 531–547.
Heiskanen J. 2006. Estimating aboveground tree biomass and leaf area index in a mountain birch forest using ASTER satellite data. *International Journal of Remote Sensing* 27(6), 1135–1158.
Helama S., Läänelaid A., Tietäväinen H., Macias Fauria M., Kukkonen I.T., Holopainen J., Nielsen J.K.,Valovirta I. 2010. Late Holocene climatic variability reconstructed from incremental data from pines and pearl mussels – a multi-proxy comparison of air and subsurface temperatures. *Boreas* 39, 734–748.
Hennenberg K.J., Goetze D., Kouame L., Orthmann B., Porembski S. 2005. Border and ecotone detection by means of vegetation composition along continuous forest-savanna transects in north-eastern Ivory Coast. *Journal of Vegetation Science* 16(3), 301–310.
Herfindahl O.C. 1950. *Concentration in the U.S. Steel Industry*. Doctoral dissertation, Columbia University, Department of Economics. Cit. http://en.wikipedia.org/wiki/Diversity_index.
Hernandez P.A., Franke I., Herzog S.K., Pacheco V., Paniagua L., Quintana H.L., Soto A., Swenson J.J., Tovar C., Valqui T.H., Vargas J., Young B.E. 2008. Predicting species distributions in poorly-studied landscapes. *Biodiversity and Conservation* 17(6), 1353–1366.
Hernandez P.A., Graham C.H., Master L.L., Albert D.L. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography* 29(5), 773–785.
Hernandez-Stefanoni J.L., Ponce-Hernandez R. 2004. Mapping the spatial distribution of plant diversity indices in a tropical forest using multi-spectral satellite image classification and field measurements. *Biodiversity and Conservation* 13(14), 2599–2621.
Hernandez-Stefanoni J.L., Ponce-Hernandez R. 2006. Mapping the spatial variability of plant diversity in a tropical forest: comparison of spatial interpolation methods. *Environmental Monitoring and Assessment* 117(1-3), 307–334.
Herzfeld U.C. 1993. A method for seafloor classification using directional variograms, demonstrated for data from the western flank of the Mid-Atlantic Ridge. *Mathematical Geology* 25(7), 901–924.
Heurich M., Schadeck S., Weinackerc H., Krzystek P. 2004. Forest Parameter Deriviation from DTM/DSM Generated from Lidar and Digital Modular Camera (DMC). In: ISPRS, Istanbul, Turkey, XXth Congress. Commission 2, 84–89.
Heuvelink G.B.M. 1998. *Error propagation in environmental modelling with GIS*. London, Bristol. Taylor & Francis.
Hiebeler D. 2000. Populations on fragmented landscapes with spatially structured heterogeneities: landscape generation and local dispersal. *Ecology* 81(6), 1629–1641.
Hilbert D.W., Ostendorf B. 2001. The utility of artificial neural networks for modelling the distribution of vegetation in past, present and future climates. *Ecological Modelling* 146(1-3), 311–327.
Hill M.O. 1973. The intensity of spatial pattern in plant communities. *Journal of Ecology* 61(1), 225–236.
Hill M.O. 1991. Patterns of species distribution in Britain elucidated by canonical correspondence analysis. *Journal of Biogeography* 18(3), 247–255.
Hill M.O., Gauch H. 1980. Detrended correspondence analysis. An improved ordination technique. *Vegetatio* 42(1-3), 47–58.
Hinch S.G., Somers K.M., Collins N.C. 1994. Spatial autocorrelation and assessment of habitat-abundance relationships in littoral zone fish. *Canadian Journal of Fisheries and Aquatic Sciences* 51(3), 701–712.
Hirschfeld H.O. 1935. A connection between correlation and contingency. *Proceedings of the Cambridge Philosophical Society* 31(4), 520–524.
Hirschman A.O. 1964. The Paternity of an Index. *The American Economic Review* 54(5), 761.
Hirzel A. 2001. *When GIS come to life. Linking landscape- and population ecology for large population management modelling: the case of Ibex (Capra ibex) in Switzerland*. Doctoral dissertation, University of Lausanne, Institute of Ecology, Laboratory for Conservation Biology.
Hirzel A., Guisan A. 2002. Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling* 157(2-3), 331–341.

Hirzel A.H., Helfer V., Metral F. 2001. Assessing habitat-suitability models with a virtual species. *Ecological Modelling* 145(2-3), 111–121.
Hirzel A.H., Hausser D., Chessel D., Perrin N. 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology* 83(7), 2027–2036.
Hirzel A., Le Lay G., Helfer V. 2006. Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling* 199(2), 142–152.
Hjermann D.Ø. 2000. Analyzing habitat selection in animals without well-defined home ranges. *Ecology* 81(5), 1462–1468.
Hjort J., Marmion M. 2008. Effects of sample size on the accuracy of geomorphological models. *Geomorphology* 102(3-4), 341–350.
Hjort J., Suomi J., Käyhkö J. 2011. Spatial prediction of urban–rural temperatures using statistical methods. *Theoretical and Applied Climatology* 106(1-2), 139–152.
Hodgson M.E. 1998. What size window for image classification? A cognitive perspective. *Photogrammetric Engineering and Remote Sensing* 64(8), 797–807.
Hoekstra A.G., Falcone J.-L., Caiazzo A., Chopard B. 2008. Multi-scale modeling with cellular automata: the complex automata approach. *Lecture Notes in Computer Science* 5191/2008, 192–199.
Hoffmann J.D., Aguilar-Amuchastegui N., Tyre A.J. 2010. Use of simulated data from a process-based habitat model to evaluate methods for predicting species occurrence. *Ecography* 33(4), 656–666.
Holdridge L.R. 1967. *Life zone ecology*. San José, Costa Rica, Tropical Science Center.
Holgate P. 1965. Some new tests of randomness. *Journal of Ecology* 53(2), 261–266.
Holgate P. 1972. The use of distance methods for the analysis of spatial distributions of points. Lewis P.A.W. (toim). Stochastic Point Processes. New York etc. Wiley, 122–135.
Holland J.H. 1992a. *Adaptation in natural and artifitial systems*. Cambridge, MIT Press.
Holland J.H. 1992b. Genetic algorithms. *Scientific American* 267(1), 66–72.
Holm S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 65–70.
Honnay O., Endels P., Vereecken H., Hermy M. 1999. The role of patch area and habitat diversity in explaining native plant species richness in disturbed suburban forest patches in northern Belgium. *Diversity and Distributions* 5(4), 129–141.
Honnay O., Piessens K., Van Landuyt W., Hermya M., Gulinck H. 2003. Satellite based land use and landscape complexity indices as predictors for regional plant species diversity. *Landscape and Urban Planning* 63(4), 241–250.
Hopfield J. 1984. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences of the USA* 81(10), 3088–3092.
Hopkins L. 1977. Methods for generating land suitability maps: a comparative evaluation. *Journal of the American Institute of Planners* 43(4), 386–400.
Hopkins B., Skellam J.G. 1954. A new method for determining the type of distribution of plant individuals. *Annals of Botany* 18(70), 213–226.
Horne B. van 1983. Density as misleading indicator of habitat quality. *Journal of Wildlife Management* 47(4), 893–901.
Horssen P.W. van, Pebesma E.J., Schot P.P. 2002. Uncertainties in spatially aggregated predictions from a logistic regression model. *Ecological Modelling* 154(1-2), 93–101.
Horssen P.W. van, Schot P.P., Barendregt A. 1999. A GIS-based plant prediction model for wetland ecosystems. *Landscape Ecology* 14(3), 253–265.
Houle G. 1998. Seed dispersal and seedling recruitment of Betula alleghaniensis: spatial inconsistency in time. *Ecology* 79(3), 807–818.
Huang Z., Brooke B., Li J. 2011. Performance of predictive models in marine benthic environments based on predictions of sponge distribution on the Australian continental shelf. *Ecological Informatics* 6(3-4), 205–216.
Hubálek Z. 1982. Coefficients of association and similarity based on binary (presence-absence) data. *Biological Reviews* 57(4), 669–689.
Hudak A.T., Crookston N.L., Evans J.S., Hall D.E., Falkowski M.J. 2008. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. *Remote Sensing of Environment* 112(5), 2232–2245.
Hudak A.T., Wessman C.A. 1998. Textural analysis of historical aerial photography to characterize woody plant encroachment in South African savanna. *Remote Sensing of the Environment* 66(3), 317–330.
Hudson P.J., Cattadori I.M. 1999. The Moran effect: a cause of population synchrony. *Trends in Ecology and Evolution* 14(1), 1–2.
Huettmann F., Diamond A.W. 2001. Seabird colony locations and environmental determination of seabird distribution: a spatially explicit breeding seabird model for the Northwest Atlantic. *Ecological Modelling* 141(1-3), 261–298.

Huisman J., Olff H., Fresco L.F.M. 1993. A hierarchical set of models for species response analysis. *Journal of Vegetation Science* 4, 37–46.
Hunter R.F. 1954. The grazing of hill pasture sward types. *The Journal of the British Grassland Society* 9(3), 195–208.
Hunter R.F. 1962. Hill sheep and their pasture; a study of sheep grazing in South East Scotland. *Journal of Ecology* 50(3), 651–680.
Huntley B., Berry P.M., Cramer W., McDonald A.P. 1995. Modeling present and potential future ranges of some European higher plants using climate response surfaces. *Journal of Biogeography* 22(6), 967–1001.
Huntley B., Bartlein P.J., Prentice I.C. 1989. Climatic control of the distribution and abundance of beech (*Fagus* L.) in Europe and North America. *Journal of Biogeography* 16(6), 551–560.
Huntley, B., Prentice, C. 1988. July temperatures in Europe from Pollen Data, 6000 years before present. *Science* 241(4866), 687–690.
Hurlbert S.H. 1971. The nonconcept of species diversity: a critique and alternative parameters. *Ecology* 52(4), 577–586.
Hutchinson G.E. 1958. Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology* 22, 415–427.
Hutchinson G.E. 1959. Homage to Santa Rosalia, or why are there so many kinds of animals? *American Naturalist* 93(870), 145–159.
Högmander H., Møller J. 1995. Estimating distribution maps from atlas data using statistical methods of image analysis. *Biometrics* 51, 393–404.
Hyyppä J., Hyyppä H., Inkinen M., Engdahl M., Linko S., Zhy Y.H. 2000. Accuracy comparison of various remote sensing data sources in the retrieval of forest stand attributes. *Forest Ecology and Management* 128(1-2), 109–120.
Indurkhya N., Weiss S.M. 1995. Using case data to improve on rule-based function approximation. Proceedings of the First International Conference on Case-Based Reasoning. Portugal, Sesimbra, Springer, 217–228.
Irfan-Ullah M., Amarnath G., Murthy M.S.R., Peterson A.T. 2007. Mapping the geographic distribution of Aglaia bourdillonii Gamble (Meliaceae), an endemic and threatened plant, using ecological niche modeling. *Biodiversity and Conservation* 16(6), 1917–1925.
Irwin E.G., Geoghegan J. 2001. Theory, data methods: developing spatially explicit economic models of land use change. *Agriculture, Ecosystems and Environment* 85(1-3), 7–23.
Isaaks E.H., Srivastava R.M. 1989. *An Introduction to Applied Geostatistics*. New York, Oxford, Oxford University Press.
Iverson L.R. 1988. Land-use changes in Illinois, U.S.A.: the influence of landscape attributes on current and historic land use. *Landscape Ecolology* 2(1), 45–61.
Iverson L.R., Prasad A., Schwartz M.W. 1999. Modeling potential future individual tree-species distributions in the eastern United States under a climate change scenario: a case study with Pinus virginiana. *Ecological Modelling* 115(1), 77–93.
Jaagus J., Briede A., Rimkus E., Remm K. 2010. Precipitation pattern in the Baltic countries under the influence of large-scale atmospheric circulation and local landscape factors. *International Journal of Climatology* 30(5), 705–720.
Jaagus J., Kull A. 2011. Changes in surface wind directions in Estonia during 1966–2008 and their relationships with large-scale atmospheric circulation. *Estonian Journal of Earth Sciences* 60(4), 220–231.
Jackson D.A., Harvey H.H. 1989. Biogeographic associations in fish assemblages: local vs. regional processes. *Ecology* 70(5), 1472–1484.
Jacquemyn H., Endels P., Honnay O., Wiegand T. 2010. Evaluating management interventions in small populations of a perennial herb *Primula vulgaris* using spatio-temporal analyses of point patterns. *Journal of Applied Ecology* 47(2), 431–440.
Jadé E. 2000. Organisation spatiale de l'île de Ténériffe. *Mappemonde* 60(4), 29–32.
Jamars P.A., Thistle D., Jones M.L. 1977. Detecting two-dimensional spatial structure in biological data. *Oecologia* 28(2), 109–123.
Jammalamadaka S.R., SenGupta A. 2001. *Topics in circular statistics.* Series in multivariate analysis. Vol 5. Singapore. World Scientific.
Janson S, Vegelius J. 1981. Measures of ecological association. Oecologia, 49(3), 371–376.
Jaynes E.T. 1957. Information theory and statistical mechanics. *The Physical Review* 106(4), 620–630.
Jarvis C.H., Stewart N. 1996. The sensitivity of a neural network for classifying remotely sensed imagery. *Computers and Geosciences* 22(9), 959–967.
Jenks G.F. 1967. The data model concept in statistical mapping. *International Yearbook of Cartography* 7, 186–190.
Jenks G.F. 1977. *Optimal data classification for choropleth maps*. Occasional paper No. 2. Department of

Geography, University of Kansas. Lawrence, Kansas.
Jensen J.R., Narumalani S., Weatherbee O., Morris K.S. 1992. Predictive modelling of Cattail and Waterlily distribution in a South Carolina Reservoir using GIS. *Photogrammetric Engineering and Remote Sensing* 58(11), 1561–1568.
Jensen J.R., Qiu F., Ji M.H. 1999. Predictive modelling of coniferous forest age using statistical and artificial neural network approaches applied to remote sensor data. *International Journal of Remote Sensing* 20(14), 2805–2822.
Jeong K.-S., Jang J.-D., Kim D.-K., Joo G.-J. 2011. Waterfowls habitat modeling: simulation of nest site selection for the migratory Little Tern (*Sterna albifrons*) in the Nakdong estuary. *Ecological Modelling* 222(17), 3149−3156.
Ji C.Y. 2000. Land-use classification of remotely sensed data using Kohonen Self-Organizing Feature Map neural networks. *Photogrammetric Engineering and Remote Sensing* 66(12), 1451−1460.
Jianwen M., Bagan H. 2005. Land-use classification using ASTER data and selforganized neutral networks. *International Journal of Applied Earth Observation and Geoinformation* 7(3), 183−188.
Jiménez-Valverde A., Lobo J.M., Hortal J. 2008. Not as good as they seem: the importance of concepts in species distribution modelling. *Diversity and Distributions* 14(6), 885–890.
Jiménez-Valverde A., Lobo J.M., Hortal J. 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. *Community Ecology* 10(2), 196–205.
Johansen K., Bartolo R., Phinn S. 2010. SPECIAL FEATURE – Geographic Object-Based Image Analysis. *Journal of Spatial Science* 55(1), 3–7.
Johnson C.J., Gillingham M.P. 2008. Sensitivity of species-distribution models to error, bias, and model design: an application to resource selection functions for woodland caribou. *Ecological Modelling* 213(2), 143–155.
Johnson D.H. 1999. The insignificance of statistical significance testing. *Journal of Wildlife Management* 63(3), 763–772.
Johnson G., Myers W.L., Patil G.P. 1999. Stochastic generating models for simulating hierarchically structured multi-cover landscapes. *Landscape Ecology* 14(5), 413–421.
Johnson R.B., Zimmer W.J. 1985. A more powerful test for dispersion using distance measurements. *Ecology* 66(5), 1669–1675.
Jolion J.M., Rosenfeld A. 1994. *A Pyramid Framework for Early Vision*. London etc. Kluwer Academic Publishers. Cit. Chiarello ja Barrat-Segretain (1997).
Jombart T., Dray S., Dufour A.-B. 2009. Finding essential scales of spatial variation in ecological data: a multivariate approach. *Ecography* 32(1), 161–168.
Jones D., Matloff N. 1986. Statistical hypothesis testing in biology: a contradiction in terms. *Journal of Economic Entomology* 79(5), 1156–1160.
Jongman R.H.G., ter Braak C.F.J., Tongeren O.F.R. 1995. *Data Analysis in Community and Landscape Ecology. Second edition.* Cambridge, Cambridge University Press.
Jost L. 2006. Entropy and diversity. *Oikos* 113(2), 363–375.
Jost L. 2007. Partitioning diversity into independent alpha and beta components. *Ecology* 88(10), 2427–2439.
Journel A.G. 1996. Modelling uncertainty and spatial dependence: stochastic imaging. *International Journal of Geographical Information Systems* 10(5), 517–522.
Jurasinski G., Retzer V., Beierkuhnlein C. 2009. Inventory, differentiation and proportional diversity: a consistent terminology for quantifying species diversity. *Oceologia* 159(1), 15–26.
Kadmon R., Farber O., Danin A. 2003. A systematic analysis of factors affecting the performance of climatic envelope models. *Ecological Applications* 13(3), 853–867.
Kalbermatten M., Van De Ville D., Turberg P., Tuia D., Joost S. 2012. Multiscale analysis of geomorphological and geological features in high resolution digital elevation models using the wavelet transform. *Geomorphology* 138(1) 352–363.
Kalkhan M.A., Stohlgren T.J. 2000. Using multi-scale sampling and spatial cross-correlation to investigate patterns of plant species richness. *Environmental Monitoring and Assessment* 64(3), 591–605.
Kane V.R., Bakker J. D., McGaughey R.J., Lutz J.A., Gersonde R.F., Franklin J.F. 2010. Examining conifer canopy structural complexity across forest ages and elevations with LiDAR data. *Canadian Journal of Forest Research* 40(4), 774–787.
Kangas A.S. 1997. On the prediction bias and variance in long-term growth projections. *Forest Ecolology and Management* 96(3), 207–216.
Kangas A.S. 1998. Uncertainity in growth and yield projections due to annual variation of diameter growth. *Forest Ecolology and Management* 108(3), 223–230.
Kapp R.O. 1978. Presettlement forests of the Pine River watershed (central Michigan) based on original land survey records. *The Michigan Botanist* 17, 3–15.
Karanth K.K., Nichols J.D., Hines J.E., Karanth K.U., Christensen N.L. 2009. Patterns and determinants of

mammal species occurrence in India. *Journal of Applied Ecology* 46(6), 1189–1200.
Katila M., Tomppo E. 2001. Selecting estimation parameters for the Finnish multisource National Forest Inventory. *Remote Sensing of Environment* 76(1), 16–32.
Keeney R., Raiffa H. 1976. *Decisions with multiple objectives: preferences and value trade-offs.* New York, Wiley.
Keith D.A., Bedward M. 1999. Native vegetation of the South East Forests region, Eden, New South Wales. *Cunninghamia* 6(1), 1–218.
Kent M., Moyeed R.A., Reid C.L., Pakeman R., Weaver R. 2006. Geostatistics, spatial rate of change analysis and boundary detection in plant ecology and biogeography. *Progress in Physical Geography* 30(2), 201– 231.
Kent M., Coker P. 1992. *Vegetation description and analysis. A practical Approach*. Chichester, Wiley.
Keramitsoglou I., Kontoes C., Sifakis N., Mitchley J., Xofis P. 2005. Kernel based re-classification of Earth observation data for fine scale habitat mapping. *Journal for Nature Conservation* 13(2-3), 91–99.
Keramitsoglou I., Sarimveis H., Kiranoudis C.T., Kontoes C., Sifakis N., Fitoka E. 2006. The performance of pixel window algorithms in the classification of habitats using VHSR imagery. *ISPRS Journal of Photogrammetry & Remote Sensing* 60(4), 225–238.
Kershaw K.A. 1960. The detection of pattern and association. *Journal of Ecology* 48(1), 233–242.
Kéry M. 2002. Inferring the absence of a species: a case study of snakes. *Journal of Wildlife Management* 66(2), 330–338.
Kilpeläinen P., Tokola T. 1999. Gain to be achieved from stand delineation in LANDSAT TM image-based estimates of stand volume. *Forest Ecology and Management* 124(2-3), 105–111.
Kimes D.S., Holben B.N., Hickeson J.E., McKee W.A. 1996. Extracting forest age in a Pacific North west forest from Thematic Mapper and topographic data. *Remote Sensing of the Environment* 56(2), 133–140.
Kimes D.S., Nelson R.F., Salas W.A., Skole D.L. 1999. Mapping secondary tropical forest and forest age from SPOT HRV data. *Journal of Remote Sensing* 20(18), 3625–3640.
Kindvall O. 1996. Habitat heterogeneity and survival in a bush cricket metapopulation. *Ecology* 77(1), 207–214.
Kirkpatrick S., Gelatt C.D. Jr., Vecchi M.P. 1983. Optimization by simulated annealing. *Science* 220(4598), 671–680.
Kissling W.D., Carl G. 2008. Spatial autocorrelation and the selection of simultaneous autoregressive models. *Global Ecology and Biogeography* 17(1), 59–71.
Klijn F., Groen C.L.G., Witte J.P.M. 1996. Ecoseries for potential site mapping, an example from the Netherlands. *Landscape and Urban Planning* 35(1), 53–70.
Klippel A. 2003. Wayfinding Choremes. Kuhn W., Worboys M.F., Timpf S. (toim.) Spatial Information Theory: Foundations of Geographic Information Science, 320–334. Berlin, Springer.
Klippel A. 2011. Movement choremes: bridging cognitive understanding and formal characterizations of movement patterns. *Topics in Cognitive Science* 3(4), 722–740.
Klok C., De Roos A.M. 1998. Effects of habitat size and quality on equilibrium density and extinction time of Sorex araneus populations. *Journal of Animal Ecology* 67(2), 195–209.
Knick S.T., Dyer D.L. 1997. Distribution of black-tailed jackrabbit habitat determined by GIS in southwestern Idaho. *Journal of Wildlife Management* 61(1), 75–85.
Knox E.G., Bartlet M. S 1964. The detection of space-time interactions. *Applied Statistics* 13(1), 25–29.
Kobler A., Adamic M. 2000. Identifying brown bear habitat by combined GIS and machine learning method. *Ecological Modelling* 135(2-3), 291–300.
Kobler A., Džeroski S., Keramitsoglou I. 2006. Habitat mapping using machine learning-extended kernel-based reclassification of an Ikonos satellite image. *Ecological Modelling* 191(1), 83–95.
Koch R.L., Burkness E.C., Hutchison W.D. 2006. Spatial distribution and fixed-precision sampling plans for the ladybird *Harmonia axyridis* in sweet corn. *BioControl* 51(6), 741–751.
Koenig W.D. 1999. Spatial autocorrelation of ecological phenomena. *Trends in Ecology and Evolution* 14(1), 22–26.
Koenig W.D., Knops J.M.H. 1998. Testing for spatial autocorrelation in ecological studies. *Ecography* 21(4), 423–429.
Kohonen T. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics* 43(1), 59–69.
Kohonen T. 1984. *Self-organization and associative memory.* Berlin, Springer.
Kok K., Veldkamp A. 2001. Evaluating impact of spatial scales on land use pattern analysis in Central America. *Agriculture, Ecosystems and Environment* 85(1-3), 205–221.
Kolkwitz R. 1911. Die Beziehungen des Kleinplanktons zum Chemismus der Gewässer. *Mitteilungen aus der königlichen Prüfungsanstalt für Wasserversorgung und Abwässerbeseitigung* 24, 145–215.
Kolkwitz R., Marsson M. 1902. Grundsätze für die biologische Beurteilung des Wassers nach seiner Flora und Fauna. *Mitteilungen aus der königlichen Prüfungsanstalt für Wasserversorgung und Abwässerbeseitigung* 1, 33–

72. Cit. Sládeček (1973).
Krige D.G. 1966. Two-dimensional weighted moving average trend surfaces for ore-evaluation. *Journal of the South Africa Institute of Mining and Metallurgy* 66(1), 13–38.
Kulldorff M. 2006. Tests of spatial randomness adjusted for an inhomogeneity: a general framework. *Journal of the American Statistical Association* 101(475), 1289–1305.
Kumar S., Stohlgren T.J., Chong G.W. 2006. Spatial heterogeneity influences native and nonnative plant species richness. *Ecology* 87(12), 3186–3199.
Kurz W.A., Beukema S.J., Klenner W., Greenough J.A., Robinson D.C.E., Sharpe A.D., Webb T.M. 2000. TELSA: the Tool for Exploratory Landscape Scenario Analyses. *Computers and Electronics in Agriculture* 27(1-3), 227–242.
Kuuluvainen T., Linkosalo T. 1998. Estimation of a spatial tree-influence model using iterative optimization. *Ecological Modelling* 106(1), 63–75.
Kyriakidis P.C., Journel A.G. 1999. Geostatistical space–time models: a review. *Mathematical Geology* 31(6), 651–684.
Labrecque S., Fournier R.A., Luther J.E., Piercey D. 2006. A comparison of four methods to map biomass from Landsat-TM and inventory data in western Newfoundland. *Forest Ecology and Management* 226(1-3), 129–144.
Lacaze B., Rambal S., Winkel T. 1994. Identifying spatial patterns of Mediterranean landscapes from geostatistical analysis of remotely-sensed data. *International Journal of Remote Sensing* 15(12), 2437–2450.
Lacey R.W.J., Legendre P., Roy A.G. 2007. Spatial-scale partitioning of in situ turbulent flow data over a pebble cluster in a gravel-bed river. *Water Resources Research* 43(3), W03416.
Lacoste Y. 1993. Chorèmatique et gèopolitique. *Herodote* 69/70, 224–259. Cit. Reimer (2010).
Lambert D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34(1), 1–14.
Lamberson R.H., McKelvey R., Noon B.R., Voss C. 1992. A dynamic analysis of northern spotted owl viability in fragmented forest landscape. *Conservation Biology* 6(4), 505–512.
Lambin X., Elston D.A., Petty S.J., MacKinnon J.L. 1998. Spatial asynchrony and periodic travelling waves in cyclic populations of field voles. *Proceedings of the Royal Society of London. Biological Sciences* 265(1405), 1491–1496.
Lande R. 1996. Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos* 76(1), 5–13.
Langsæter A. 1926. Om beregning av middelfeilen ved regelmessige linjetakseringer. *Meddelanden fra det Skogsforsoksvesen* 2(7), 5–47.
Lark R.M. 1996. Geostatistical description of texture on an aerial photograph for discriminating classes of land cover. *International Journal of Remote Sensing* 17(11), 2115–2133.
Lavorel S., Gardner R.H., O'Neill R.V. 1993. Analysis of patterns in hierarchically structured landscapes. *Oikos* 67(3), 521–528.
Lavorel S., Gardner R.H., O'Neill R.V. 1995. Dispersal of annual plants in hierarchically structured landscapes. *Landscape Ecology* 10(5), 277–289.
Law R., Illian J., Burslem D.F.R.P., Gratzer G., Gunatilleke C.V.S., Gunatilleke I.A.U.N. 2009. Ecological information from spatial patterns of plants: insights from point process theory. *Journal of Ecology* 97(4), 616–628.
Le Q.B., Park S.J., Vlek P.L.G., Cremers A.B. 2008. Land-Use Dynamic Simulator (LUDAS): a multi-agent system model for simulating spatio-temporal dynamics of coupled human–landscape system. I. Structure and theoretical specification. *Ecological Informatics* 3(2), 135–153.
Le Q.B., Park S.J., Vlek P.L.G. 2010. Land Use Dynamic Simulator (LUDAS): a multi-agent system model for simulating spatio-temporal dynamics of coupled human–landscape system 2. Scenario-based application for impact assessment of land-use policies. *Ecological Informatics* 5(3), 203–221.
Le Maitre D.C, Thuiller W., Schonegevel L. 2008. Developing an approach to defining the potential distributions of invasive plant species: a case study of *Hakea* species in South Africa. *Global Ecology and Biogeography* 17(5), 569–584.
Leathwick J.R. 1998. Are New-Zealand's *Notofagus* species in equilibrium with their environment? *Journal of Vegetation Science* 9(5), 719–732.
Leathwick J.R., Elith J., Hastie T. 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling* 199(2), 188–196.
Leathwick J.R., Overton J.M., McLeod M. 2003. An environmental domain analysis of New Zealand, and its application to biodiversity conservation. *Conservation Biology* 17(6), 1612–1623.
Leckie D., Gougeon F.A., Walsworth N., Paradine D. 2003. Stand delineation and composition estimation using semi-automated individual tree crown analysis. *Remote Sensing of Environment* 85(3), 355–369.

Lee M., Fahrig L., Freemark K., Currie D. 2002. Importance of patch scale vs landscape scale on selected forest birds. *Oikos* 96(1), 110–118.
Lees B.G., Ritman K. 1991. Decision-tree and rule induction approach to integration of remotely sensed and GIS data in mapping vegetation in disturbed hilly environments. *Environmental Management* 15(6), 823–831.
Lees B.G., Zhi H., van Niel K., Laffan S.W. 2008. The impact of DEM error on predictive vegetation mapping. Advances in digital terrain analysis. Lecture notes in geoinformation and cartography. Section 4. Springer, 349–362.
Lefsky M.A., Cohen W.B., Parker G.G., Harding D.J. 2002. Lidar remote sensing for ecosystem studies. *BioScience* 52(1), 19–30.
Lefsky M.A., Turner D.P., Guzy M., Cohen W.B. 2005. Combining lidar estimates of aboveground biomass and Landsat estimates of stand age for spatially extensive validation of modeled forest productivity. *Remote Sensing of Environment* 95(4), 49–558.
Legendre P. 1993. Spatial autocorrelation: trouble or new paradigm? *Ecology* 74(6), 1659–1673.
Legendre P., Fortin M.-J. 1989. Spatial pattern and ecological analysis. *Vegetatio* 80(2), 107–138.
Legendre P., Fortin M.-J. 2010. Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular Ecology Resources* 10(5), 831–844.
Legendre P., Legendre L. 1998. *Numerical ecology*. Second English edition. Amsterdam, Elsevier Science BV.
Lehmann A., Leathwick J.R., Overton J.McC. 2002a. Assessing New Zealand fern diversity from spatial predictions of species assemblages. *Biodiversity and Conservation* 11(12), 2217–2238.
Lehmann A., Overton J.M., Leathwick J.R. 2002b. GRASP: generalized regression analysis and spatial prediction. *Ecological Modelling* 157(2–3), 189–207.
Lek S., Guégan J.F. 1999. Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling* 120(2-3), 65–73.
Lek S., Delacoste M., Baran P., Dimopoulus I., Lauga J., Aulagnier S. 1996. Application of neural networks to modelling non linear relationships in ecology. *Ecological Modelling* 90(1), 39–52.
Lenihan J.M. 1993. Ecological response surfaces for North American boreal tree species and their use in forest classification. *Journal of Vegetation Science* 4(5), 667–680.
Lennon J.J. 2000. Red-shifts and red herrings in geographical ecology. *Ecography* 23(1), 101–113.
Lepš J., Kindlmann P. 1987. Models of the development of spatial pattern of an even-aged plant population over time. *Ecological Modelling* 39(1-2), 45–57.
Les M., Maher C. 1998. Measuring diversity: choice in local housing markets. *Geographical Analysis* 30(2), 172–190.
Leung Y., Mei C.-L., Zhang W.-X. 2003. Statistical test for local patterns of spatial association. *Environment and Planning A* 35(4), 725–744
Lévesque J., King D.J. 2003. Spatial analysis of radiometric fractions from high-resolution multispectral imagery for modelling individual tree crown and forest canopy structure and health. *Remote Sensing of Environment* 84(4), 589–602.
Levin S.A. 1974. Dispersion and population interactions. *American Naturalist* 108(960), 207–228.
Levin S.A. 1992. The problem of pattern and scale in ecology. *Ecology* 73(6), 1943–1967.
Levins R. 1969. Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bulletin of the Entomological Society of America* 15(3), 237–240.
Lewinsky S. 2006. Applying fused multispectral and panchromatic data of Landsat ETM+ to object oriented classification. Proceedings of the 26th EARSeL Symposium, New Developments and Challenges in Remote Sensing, May 29-June 2, Warsaw, Poland.
Lewis M.A., Kareiva P. 1993. Allee dynamics and the spread of invading organisms. *Theoretical Population Biology* 43(2), 141–158.
Li B.-L. 2000. Fractal geometry applications in description and analysis of patch patterns and patch dynamics. *Ecological Modelling* 132(1-2), 33–50.
Li H., Reynolds J. 1993. A new contagion index to quantify spatial patterns of landscapes. *Landscape Ecology* 8(3), 155–162.
Li H., Reynolds J. 1994. A simulation experiment to quantify spatial heterogeneity in categorical maps. *Ecology* 75(8), 2446–2455.
Li H., Franklin J.F., Swanson F.J., Spies T.A. 1993. Developing alternative forest cutting patterns: a simulation approach. *Landscape Ecology* 8(1), 63–75.
Li J., Heap A.D. 2011. A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors. *Ecological Informatics* 6(3-4), 228–241.
Li L., Huang Z., Ye W., Cao H., Wei S., Wang Z., Lian J., Sun I.-F., Ma K., He F. 2009. Spatial distributions of tree species in a subtropical forest of China. *Oikos* 118(4), 495–502.
Li W., Wang Z., Ma Z., Tang H. 1997. A regression model for the spatial distribution of red-crown crane in

Yancheng Biosphere Reserve, China. *Ecological Modelling* 103(2-3), 115–121.
Lichstein J.W., Simons T.R, Shriner S.A., Franzreb K.E. 2002. Spatial autoregressive term to model fine-scale spatial autocorrelation. *Ecological Monographs* 72(3), 445–463.
Lieshout M.N.M. van, Baddeley A.J. 1996. A nonparametric measure of spatial interaction in point patterns. *Statistica Neerlandica* 50(3), 344–361.
Ligozat G., Nowak J., Schmitt D. 2007. From language to pictorial representations. Poznańskie W. (toim.) Proceedings of the Language and Technology Conference (L&TC'07), Poznan, Poland, 5-7 September, 2007. http://archives.limsi.fr/Individu/ligozat/DOCUMENTS/poznan_2007.pdf.
Lilburne L., Tarantola S. 2008. Sensitivity analysis of spatial models. *International Journal of Geographical Information Science* 23(2), 151–168.
Lim K., Treitz P., Wulder M., St-Onge B., Flood M. 2003. LiDAR remote sensing of forest structure. *Progress in Physical Geography* 27(1), 88–106.
Lin Y.-P., Yeh M.-S., Deng D.-P., Wang Y.-C. 2008. Geostatistical approaches and optimal additional sampling schemes for spatial patterns and future sampling of bird diversity. *Global Ecology and Biogeography* 17(2), 175–188.
Linard C., Gilbert M., Tatem A.J. 2010. Assessing the use of global land cover data for guiding large area population distribution modelling. *GeoJournal* 75(5), 525–538.
Lindenmayer D.B., Possingham H.P. 1996. Modellingthe inter-relationship between habitat patchness, dispersal capability and metapopulation persistence of the endangered species, Leadpeater's possum, in South-eastern Australia. *Landscape Ecology* 11(2), 79–105.
Lindenmayer D.B., Cunningham R.B., McCarthy M.A. 1999. The conservation of arboreal marsupials in the mountine ash forests of the central highlands of Victoria, south-eastern Australia. VIII. Landscape analysis of the occurrence of arboreal marsupials. *Biological Conservation* 89, 83–92.
Linder M., Remm K., Absalon E. 2009. Tehisõppesüsteemi Pidevstuudium/CONSTUD kasutused. Mander, Ü. Uuemaa, E. Pae, T. (toim). Uurimusi eestikeelse geograafia 90. aastapäeval. Publicationes Instituti Geographici Universitatis Tartuensis 108. Tartu, Tartu Ülikooli Kirjastus, 52–62.
Linder M., Remm K., Proosa H. 2008. The application of the concept of indicative neighbourhood on Landsat ETM+ images and orthophotos using circular and annulus kernels. SDH 2008. The 13th International Symposium on Spatial Data Handling. 23rd to 25th June, 2008 Montpellier, France, 147–162.
Lindström J., Ranta E., Lindén H. 1996. Large-scale synchrony in the dynamics of capercaillie, black grouse and hazel grouse populations in Finland. *Oikos* 76(2), 221–227.
Lippitt C., Rogan J., Toledano J., Sangermano F., Eastman J., Mastro V., Sawyer A. 2008. Incorporating anthropogenic variables into a species distribution model to map gypsy moth risk. *Ecological Modelling* 210(3), 339–350.
Lischke H., Zimmermann N., Bolliger J., Rickebusch S., Löffler T. 2006. TreeMig: a forest-landscape model for simulating. *Ecological Modelling* 199(4), 409–420.
Liu C., Berry P.M., Dawson T.P., Pearson R.G. 2005. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* 28(3), 385–393.
Liu Y. 2006. Using the Snesim program for multiple-point statistical simulation. *Computers and Geosciences* 32(10), 1544–1563.
Liu Y., Feng Y. 2010. An optimised cellular automata model based on adaptive genetic algorithm for urban growth simulation. Guilbert E., Lees B., Leung Y. ISPRS. Proceedings of the Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science. XXXVIII, Part 2. Hong Kong. ISPRS Technical Commission II, 45–50.
Liu Y., Journel A.G. 2009. A package for geostatistical integration of coarse and fine scale data. *Computers and Geosciences* 35(3), 527–547.
Lloyd M. 1967. Mean crowding. *Journal of Animal Ecology* 36(1), 1–30.
Lloyd M., Ghelardi R.J. 1964. A table for calculating the 'equitability' component of species diversity. *Journal of Animal Ecology* 33(2), 217–225.
Logan T.L., Strahler A.H., Woodcock C.E. 1979. Use of a standard deviation based texture channel for Landsat classification of forest strata. Machine Processing of Remotely Sensed Data Symposium. Indiana, Purdue University, 395–404.
Logsdon M.G., Bell E.J., Westerlund F.V. 1996. Probability mapping of land use change: A GIS interface for visualizing transition probabilities. *Computers, Environment and Urban Systems* 20(6), 389–398.
Loibl W. 2000. Modellierung der Siedlungsdynamik mit einem GIS-basierten Zellular Automaten – Konzeption, GIS-Integration und erste Ergebnisse. Strobl J., Blaschke T., Griesebner G. (toim). Angewandte Geographische Informationsverarbaitung XII: Beitrage zum AGIT-Symposium Salzburg 2000. Heidelberg, Wichmann, 297–303.
Lomolino M.V., Riddle B.R., Brown J.H. 2006. *Biogeography*. Third edition. Sunderland, Massachusetts, Sinauer Associates.

Long D.S. 1998. Spatial autoregression modeling of site-specific wheat yield. *Geoderma* 85(2-3), 181–197.
Long J.A., Nelson T.A., Wulder M.A. 2010. Local indicators for categorical data: impacts of scaling decisions. *The Canadian Geographer* 54(1), 15–28.
Lorenz E.N. 1956. *Empirical orthogonal functions and statistical weather prediction*. Scientific Report No. 1, Statistical Forecasting Project, M.I.T., Cambridge.
Lorenz M.O. 1905. Methods of measuring concentrations of wealth. *Publications of the American Statistical Association* 9(40), 209–219.
Loyn R.H., McNabb E.G, Volodina L., Willig R. 2001. Modelling landscape distributions of large forest owls as applied to managing forests in north-east Victoria, Australia. *Biological Conservation* 97(3), 361–376.
Ludwig J.A., Reynolds J.F. 1988. *Statistical ecology. A primer on methods and computing*. New York etc. Wiley.
Lundberg P., Ranta E., Ripa J., Kaitala V. 2000. Population variability in space and time. *Trends in Ecology and Evolution* 15(11), 460–464.
Luoto M. 2000. *Landscape ecological analysis and modelling of habitat and species diversity in agricultural landscapes using GIS.* Academic dissertation, Turun Yliopiston Julkaisuja, AII(141), Biologica - Geographica - Geologica.
Luoto M. Heikkinen R.K., Pöyry J., Saarinen K. 2006. Determinants of the biogeographical distribution of butterflies in boreal regions. *Journal of Biogeography* 33(10), 1764–1778.
Luoto M., Kuussaari M., Rita H., Salminen J., von Bondsdorff T. 2001. Determinants of distribution and abundance in the clouded apollo butterfly: a landscape ecological approach. *Ecography* 24(5), 601–617.
Luoto M., Poyry J., Heikkinen R.K., Saarinen K. 2005. Uncertainty of bioclimate envelope models based on the geographical distribution of species. *Global Ecology and Biogeography* 14(6), 575–584.
Luther J.E., Fournier R.A., Piercey D.E., Guindon L., Hall R.J. 2006. Biomass mapping using forest type and structure derived from Landsat TM imagery. *International Journal of Applied Earth Observation and Geoinformation* 8(3), 173–187.
Luud A., Remm K. 2001. Impact of human activities on the value of moose habitats on the example of spatial distribution of moose population in Ida-Viru county, Estonia. *Publicationes Instituti Geographici Universitatis Tartuensis* 92, 460–465.
Lõhmus E. 1984. Eesti metsakasvukohatüübid. Tallinn.
Läänelaid A., Eckstein D. 2003. Development of a tree-ring chronology of scots pine (*Pinus sylvestris* L.) for Estonia as a dating tool and climate proxy. *Baltic Forestry* 9(2), 76–82.
Lütolf M., Guisan A., Kienast F. 2009. History matters: relating land-use change to butterfly species occurrence. *Environmental Management* 43(3), 436–446.
Lütolf M., Kienast F., Guisan A. 2006. The ghost of past species occurrence: improving species distribution models for presence-only data. *Journal of Applied Ecology* 43(4), 802–815.
Lynch H.J., Moorcroft P. 2008. A spatiotemporal Ripley's K-function to analyze interactions between spruce budworm and fire in British Columbia, Canada. *Canadian Journal of Forest Research* 38(12), 3112–3119.
Maa-amet 2002. Eesti põhikaardi 1: 10 000 digitaalkaardistuse juhend (v. 2006). http://geoportaal.maaamet.ee/est/Andmed-ja-kaardid/Topograafilised-andmed/Eesti-Pohikaart-110-000/Juhendid-ja-abifailid-p130.html.
Maarel E. van der 1974. Small-scale vegetation boundaries: on their analysis and typology. Sommer W.H. ja Tüxen R. (toim). Tatsachen und Probleme der Grenzen in der Vegetation: Bericht über das Internationale Symposion der Internationalen Vereinigung für Vegetationskunde in Rinteln 8.-11. April 1968, 75–80. Cit. Maarel (1976)
Maarel E. van der 1976. On the establishment of plant community boundaries. *Berichte der Deutschen Botanischen Gesellschaft* 89, 415–443.
Mac Nally R., Bennett A.F., Horrocks G. 2000. Forecasting the impacts of habitat fragmentation. Evaluation of species-specific predictions of the impact of habitat fragmentation on birds in the box-ironbark forests of central Victoria, Australia. *Biological Conservation* 95(1), 7–29.
MacArthur R.H. 1957. On the relative abundance of bird species. *Proceedings of National Academy of Sciences* 43(3), 293–295.
MacArthur R.H., Wilson E.O. 1963. An equilibrium theory of insular zoogeography. *Evolution* 17(4), 373-387.
MacArthur R.H. 1972. *Geographical ecology: patterns in the distribution of species*. New York, Harper & Row.
MacArthur R.H., Wilson E.O. 1967. *The theory of island biogeography. Monographs in population biology 1*. Princeton, Princeton University Press.
Mack R.N., Harper J.L. 1977. Interference in dune annuals: spatial pattern and neighbourhood effects. *Journal of Ecology* 65(2), 345–363.
Mackey B.G., Berry S.L., Brown T. 2008. Reconciling approaches to biogeographical regionalization: a

systematic and generic framework examined with a case study of the Australian continent. *Journal of Biogeography* 35(2), 213–229.
MacLennan B. 1990. *Continuous spatial automata*. Department of Computer Science Technical Report CS-90-121.
MacMillan R.A., Moon D.E., Coupé R.A., Phillips N. 2010. Chapter 27. Predictive Ecosystem Mapping (PEM) for 8.2 million ha of forestland, British Columbia, Canada. Boettinger J.L. *et al*. (toim). Digital Soil Mapping, Progress in Soil Science 2, DOI 10.1007/978-90-481-8863-5_27. Springer Science+Business Media B.V.
Maegher T.R., Burdick D.S. 1980. The use of nearest neighbor frequency analysis in studies of association. *Ecology* 61(5), 1253–1255.
Maes D., Bauwens D., de Bruynt L., Anselin A., Vermeersch G., van Landuyt W., de Knijf G., Gilbert M. 2005. Species richness coincidence: conservation strategies based on predictive modelling. *Biodiversity and Conservation* 14(6), 1345–1364.
Maggini R., Lehmann A., Zimmermann N.E., Guisan A. 2006. Improving generalized regression analysis for the spatial prediction of forest communities. *Journal of Biogeography* 33(10), 1729–1749.
Malanson G.P. 1985. Spatial autocorrelation and distributions of plant species on environmental gradients. *Oikos* 45(2), 278–280.
Malanson G.P., Westman W.E., Yan Y.-L. 1992. Realized versus fundamental niche functions in a model of chaparral response to climatic change. *Ecological Modelling* 64(4), 261–277.
Mandelbrot B.B. 1967. How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science*, 156(3775), 636-638.
Manel S., Dias J.-M., Ormerod S.J. 1999. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling* 120(2-3), 337–347.
Mangel M., Adler F.R. 1994. Construction of multidimensional clustered patterns. *Ecology* 75(5), 1289–1298.
Manies L.K., Mladenoff D.J. 2000. Testing methods to produce landscape-scale presettlement vegetation maps from the U.S. public land survey records. *Landscape Ecology* 15(8), 741–754.
Manly B.F.J. 1986. Randomization and regression methods for testing for associations with geographical, environmental and biological distances between populations. *Researches on Population Ecology* 28(2), 201–218.
Manly B.F.J. 1997. *Randomization, bootstrap and Monte Carlo methods in biology*. Second edition. London, Weinheim, New York, Tokyo, Melbourne, Madras, Chapman & Hall.
Manly B.F.J., McDonald L.L., Thomas D.L., McDonald T.L., Erickson W.P. 2002. *Resource selection by animals: statistical analysis and design for field studies*. Second edition. Boston, Kluwer.
Mantel N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research* 27(2), 209–220.
Marceau D.J. 1999. The scale issue in the social and natural sciences. *Canadian Journal of Remote Sensing* 25(4), 347–356.
Margalef D.R. 1957. La teoría de la información en Ecología. *Memorias de la Real Academia de Ciencias y Artes de Barcelona* 32, 373–449. Translated into English and published as: Margalef R. 1958. Information theory in ecology. *General Systems* 3, 36–71.
Marmion M., Parviainen M., Luoto M., Heikkinen R.K., Thuiller W. 2009. Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions* 15(1), 59–69.
Marsh L.M., Jones R.E. 1988. The form and consequences of random walk movement models. *Journal of Theoretical Biology* 133(1), 113–131.
Martin R.L., Oeppen J.E. 1975. The identification of regional forecasting models using space-time correlation functions. *Transactions of the Institute of British Geographers* 66, 95–118.
Mas J.F., Puig H., Palacio J.L., Sosa-López A. 2004. Modelling deforestation using GIS and artificial neural networks. *Environmental Modelling & Software* 19(5), 461–471.
Mateo R.G., Croat T.B., Felicísimo A.M., Muñoz J. 2010a. Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections. *Diversity and Distributions* 16(1), 84–94.
Mateo R.G., Felicísimo A.M., Muñoz J. 2010b. Effects of the number of presences on reliability and stability of MARS species distribution models: the importance of regional niche variation and ecological heterogeneity. *Journal of Vegetation Science* 21(5), 908–922.
Matérn B. 1947. Metoder att uppskatta noggrannheten vid linje- och provytetaxering. *Meddelanden från Statens Skogsforskningsinstitut* 36(1), 1–138.
Matérn B. 1960. Spatial variation. *Meddelanden från Statens Skogsforskningsinstitut* 49(5). Uppsala. Cit. Gelfand *et al.* (2010).
Matérn B. 1986. Spatial variation. 2nd Edition. *Lecture Notes in Statistics* 36. New York, Springer. Cit. Gelfand *et al.*(2010).

Mateu J., Uso J.L., Montes F. 1998. The spatial pattern of a forest ecosystem. *Ecological Modelling* 108(1-3), 163–174.
Mather P.M. 1999. *Computer processing of remotely-sensed images. Second edition.* Chichester, Wiley.
Matheron G. 1963. Principles of geostatistics. *Economic Geology* 58(8), 1246–1266.
Matheron G. 1973. The intrinsic random functions and their application. *Advances in Applied Probability* 5(3), 439–468.
Mathew J., Jha V.K., Rawat G.S. 2007. Weights of evidence modelling for landslide hazard zonation mapping in part of Bhagirathi valley, Uttarakhand. *Current Science* 92(5), 628–638.
Matlock R.B.Jr., Edwards P.J. 2006. The influence of habitat variables on bird communities in forest remnants in Costa Rica. *Biodiversity and Conservation* 15(9), 2987–3016.
McArdle B.H., Hewitt J.E., Thrush S.F. 1997. Pattern from process: it is not as easy as it looks. *Journal of Experimental Marine Biology and Ecology* 216(1-2), 229–242.
McBratney A.B., Mendonça Santos M.L., Minasnya B. 2003. On digital soil mapping. *Geoderma* 117(1-2), 3–52.
McDonald T.L., Manly B.F.J., Nielson R.M., Diller L.V. 2006. Discrete-choice modeling in wildlife studies exemplified by northern spotted owl nighttime habitat selection. *The Journal of Wildlife Management* 70(2), 375–383.
McElhinny C., Gibbons P., Brack C., Bauhus J. 2005. Forest and woodland stand structural complexity: its definition and measurement. *Forest Ecology and Management* 218(1), 1–24.
McElhinny C., Gibbons P., Brack C. 2006. An objective and quantitative methodology for constructing an index of stand structural complexity. *Forest Ecology and Management* 235(1), 54–71.
McElhany P.L., Real L.A., Power A.G. 1995. Vector preference and disease dynamics: a study of barley yellow dwarf virus. *Ecology* 76(2), 444–457.
McIntire E.J.B., Fajardo A. 2009. Beyond description: the active and effective way to infer processes from spatial patterns. *Ecology* 90(1), 46–56.
McIntyre N.E., Wiens J.A. 1999. Interactions between habitat abundance and configuration: experimental validation of some predictions from percolation theory. *Oikos* 86(1), 129–137.
McKenzie N.L., Belbin L., Margules C.R., Keighery G.J. 1989. Selecting representative reserve systems in remote areas: a case study in the Nullarbor Region, Australia. *Biological Conservation* 50(1-4), 239–261.
McLaughlin J.F., Rouchgarden J. 1992. Predation across spatial scales in heterogenous environments. *Theoretical Population Biology* 41(3), 277–299.
McMaster R.B., Shea K.S. 1992. *Generalization in digital cartography*. Washington, D.C. Association of American Geographers.
McPherson J.M., Jetz W., Rogers D.J. 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology* 41(5), 811–823.
McRoberts R., Holden G.R., Nelson M.D., Liknes G.C., Gormanson D.D. 2006. Using satellite imagery as ancillary data for increasing the precision of estimates for the Forest Inventory and Analysis program of the USDA Forest Service. *Canadian Journal of Forest Research* 36(12), 2968–2980.
Meer P. 1989. Stochastic image pyramids. *Computer Vision, Graphics, and Image Processing* 45(3), 269–264.
Meer P., Connely S. 1989. A fast parallel method for synthesis of random patterns. *Pattern Recognition* 22(2), 189–204.
Meggs J.M., Munks S.A., Corkey R., Richards K. 2004. Development and evaluation of predictive habitat models to assist the conservation planning of a threatened lucanid beetle, *Hoplogonus simsoni*, in north-east Tasmania. *Biological Conservation* 118(4), 501–511.
Meier E.S., Kienast F., Pearman P.B., Svenning J.-C., Thuiller W., Araújo M.B., Guisan A., Zimmermann N.E. 2010. Biotic and abiotic variables show little redundancy in explaining tree species distributions. *Ecography* 33(6), 1038–1048.
Mellin C., Andréfouët S., Ponton D. 2007. Spatial predictability of juvenile fish species richness and abundance in a coral reef environment. *Coral Reefs* 26(4), 895–907.
Ménard A., Dubé P., Bouchard A., Canham C.D., Marceau D.J. 2002. Evaluating the potential of the SORTIE forest succession model for spatio-temporal analysis of small-scale disturbances. *Ecological Modelling* 153(1-2), 81–96.
Merchant J.W. 1984. Using spatial logic in classification of Landsat TM data. Proceedings of the 9th Annual Pecora Symposium, Sioux Falls, South Dakota, 378–385.
Mez C. 1898. *Microscopische Wasseranalüse. Anleitung zur Untersuchung des Wassers mit besonderer Berücksichtigung von Trink- und Abwasser*. Berlin. Springer. Cit. Sládeček (1973).
Metzger M.J., Bunce R.G.H., Jongman R.H.G., Mücher C.A., Watkins J.W. 2005. A climatic stratification of the environment of Europe. *Global Ecology and Biogeography* 14(6), 549–563.
Metzler J.W., Sader S.A. 2005. Agreement assessment of spatially explicit regression-derived forest cover and traditional forest industry stand type maps. *Photogrammetry Engineering Remote Sensing* 71(11), 1303–1309.

Miina J. 1993. Residual variation in diameter growth in a stand of Scots pine and Norway spruce. *Forest Ecology and Management* 58(1-2), 111–128.
Miller J.N., Brooks R.P., Croonquist M.J. 1997. Effects of landscape patterns on biotic communities. *Landscape Ecology* 12(3), 137–153.
Miller J., Franklin J. 2002. Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. *Ecological Modelling* 157(2-3), 227–247.
Miller J., Franklin J., Aspinall R. 2007. Incorporating spatial dependence in predictive vegetation models. *Ecological Modelling* 202(3-4), 225–242.
Milne B.T. 1988. Measuring the fractal geometry of landscapes. *Journal of Applied Mathematics and Computing* 27(1), 67–79.
Milne B.T. 1992. Spatial aggregation and neutral models in fractal landscapes. *The American Naturalist* 139(1), 32–57.
Milne B.T., Johnston K.M., Forman R.T.T. 1989. Scale-dependent proximity of wildlife habitat in a spatially-neutral Bayesian model. *Landscape Ecology* 2(2), 101–110.
Minnaert M. 1941. The reciprocity principle in lunar photometry. *Astrophysical Journal* 93, 403–410.
Miranda F.P., Macdonald J.A., Carr J.R. 1992. Application of the semivariogram textural classifier (STC) for vegetation discrimination using SIR-B data of Borneo. *International Journal of Remote Sensing* 13(12), 2349–2354.
Miranda F.P., Fonseca L.E.N., Carr J.R., Taranic J.V. 1996. Analysis of JERS-1 (FUYO-1) SAR data for vegetation discrimination in Northwestern Brazil using the semivariogram textural classifier (STC). *International Journal of Remote Sensing* 17(17), 3523–3529.
Miranda F.P., Carr J.R. 1994. Application of the semivariogram textural classifier (STC) for vegetation discrimination using SIR-B data of the Guiana Shield, Northwestern Brazil. *Remote Sensing Reviews* 10(1), 155–168.
Mitchell M., Crutchfield J.P., Das R. 1996. Evolving cellular automata with genetic algorithms: a review of recent work rule. http://web.cecs.pdx.edu/~mm/evca-review.pdf
Miyadokoro T., Nishimura N., Yamamoto S. 2003. Population structure and spatial patterns of major trees in a subalpine old-growth coniferous forest, central Japan. *Forest Ecology and Management* 182(1), 259–272.
Mladenoff D.J. 2004. LANDIS and forest landscape models. *Ecological Modelling* 180(1), 7–19.
Mladenoff D.J., Host G.E., Boederand J., Crow T.R. 1996. LANDIS: a spatial simulation model of forest landscape disturbance, management and succession. Goodchild M.F. *et al.* GIS and environmental modeling: progress and research issues, 175–179.
Mladenoff D.J., Sickley, T.A., Haight R.G., Wydeven A.P. 1995. A regional landscape analysis and prediction of favorable gray wolf habitat in the northern Great Lakes region. *Conservation Biology* 9(2), 279–294.
Moellering H., Tobler W. 1972. Geographical variances. *Geographical Analysis* 4(1), 34–64.
Moeur M. 1997. Spatial models of competition and gap dynamics in old-growth *Tsuga heterophylla/Thuja plicata* forests. *Forest Ecology and Management* 94(1-3), 175–186.
Moilanen A., Nieminen M. 2002. Simple connectivity measures in spatial ecology. *Ecology* 83(4), 1131–1143.
Moisen G.G., Freeman E.A., Blackard J.A., Frescino T.S., Zimmermann N.E., Edwards T.C.Jr. 2002. Predicting tree species presence and basal area in Utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. *Ecological Modelling* 199(2), 176–187.
Moisen G.G., Frescino T.S. 2002. Comparing five modelling techniques for predicting forest characteristics. *Ecological Modelling* 157(2-3), 209–225.
Molovsky J. 1994. Population dynamics and pattern formation in theoretical populations. *Ecology* 75(1), 30–39.
Moody A.L., Thompson W.A., de Bruijn B., Houston A.I., Goss-Custard J.D. 1997. The analysis of spacing of animals, with an example based on oystercatchers during the tidal cycle. *Journal of Animal Ecology* 66(5), 615–628.
Moore P.G. 1954. Spacing in plant populations. *Ecology* 35(2), 222–227.
Moore A.D., Noble I.R. 1993. Automatic model simplification: the generation of replacement sequences and their use in vegetation modelling. *Ecological Modelling* 70(1–2), 137–157.
Moore D.M., Lees B.G., Davey S.M. 1991. A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. *Environmental Management* 15(1), 59–71.
Moran P.A.P. 1950. Notes on continous stochastic phenomena. *Biometrika* 37(1-2), 17–23.
Moran P.A.P. 1953. The statistical analysis of the Canadian lynx cycle. II. Synchronization and meteorology. *Australian Journal of Zoology* 1(3), 291–298.
Morisita M. 1959a. Measuring of the dispersion of individuals and analysis of the distributional patterns. *Memoirs of the Faculty of Science Kyushu University, Series E* 2, 215–235.
Morisita M. 1959b. Measuring of interspecific association and similarity between communities. *Memoirs of the Faculty of Science Kyushu University, Series E* 3, 65–80.
Morrison J.L. 1974. A theoretical framework for cartographic generalization with emphasis on the process of symbolization. *International Yearbook of Cartography* 14, 115–127. Cit. McMaster ja Shea (1992).

Mountford M.D. 1962. An index of similarity and its application to classificatory problems. Murphy P.W. (toim). Progress in soil zoology. Papers from a colloquium on research methods organized by the soil zoology committee of the international Society of soil science held at Rothamsted experimental station Hertfordshire 10-14th July, 1958. London, Butterworths, 43–50.
Muchoney D.M., Strahler A.H. 2002. Pixel- and site-based calibration and validation methods for evaluating supervised classification of remotely sensed data. *Remote Sensing of Environment* 81(2-3), 290–299.
Mugglestone M.A., Renshaw E. 1996. A practical guide to the spectral analysis of spatial point processes. *Computational Statistics & Data Analysis* 21(1), 43–65.
Muinonen E., Maltamo M., Hyppänen H., Vainikainen V. 2001. Forest stand characteristics estimation using a most similar neighbor approach and image spatial structure information. *Remote Sensing of the Environment* 78(3), 223–228.
Mullahy J. 1986. Specification and testing of some modified count data models. *Journal of Econometrics* 33(3), 341–365.
Munoz F. 2009. Distance-based eigenvector maps (DBEM) to analyse metapopulation structure with irregular sampling. *Ecological Modelling* 220(20), 2683–2689.
Muñoz M.E.S, Giovanni R., de Siqueira M.F., Sutton T., Brewer P., Pereira R.S., Canhos D.A.L., Canhos V.P. 2011. OpenModeller: a generic approach to species' potential distribution modelling. *Geoinformatica* 15(1), 111–135.
Munroe E.G. 1953. The size of island faunas. Proceedings of the Seventh Pacific Science Congress of the Pacific Science Association, vol IV, Zoology. Auckland, New Zealand, 52–53. Cit. Lomolino *et al.* (2006).
Muukkonen P., Heiskanen J. 2005. Estimating biomass for boreal forests using ASTER satellite data combined with standwise forest inventory data. *Remote Sensing of Environment* 99(4), 434–447.
Murtaugh P.A. 1996. The statistical evaluation of ecological indicators. *Ecological Applications* 6(1), 132–139.
Mäkelä H., Pekkarinen A. 2001. Estimation of timber volume at the sample plot level by means of image segmentation and Landsat TM imagery. *Remote Sensing of the Environment* 77(1), 66–75.
Mäkelä H., Pekkarinen A. 2004. Estimation of forest stand volumes by Landsat TM imagery and stand-level field-inventory data. *Forest Ecology and Management* 196(2-3), 245–255.
Männil P., Veeroja R., Tõnisson J. 2011. *Ulukiasurkondade seisund ja küttimissoovitus 2011*. *Status of Game populations in Estonia and proposal for hunting in 2011*. Tartu, Keskkonnateabe Keskus. http://www.keskkonnainfo.ee/failid/ULUKITE_SEIREARUANNE_2011.pdf.
Mörtberg U., Wallenius H.-G. 2000. Red-listed forest bird species in an urban environment – assessment of green space corridors. *Landscape and Urban Planning* 50(4), 215–226.
Myers W., Patil G.P., Joly K. 1996. *Echelon approach to areas of concern in synoptic regional monitoring*. Technical Report Number 96-0601. Technical Reports and Reprints Series. Center for Statistical Ecology and Environmental Statistics. Department of Statistics. The Pennsylvania State University.
Myers W., Patil G.P., Joly K. 1997. Echelon approach to areas of concern in synoptic regional monitoring. *Environmental and Ecological Statistics* 4(2), 131–152.
Myers W., Patil G.P., Taillie C. 1995. Comparative paradigms for biodiversity assessment. Boyle T.J.B., Boontawee B. (toim). Measuring and Monitoring Biodiversity in Tropical and Temperate Forests. Proceedings of IUFRO Symposium held at Chiang Mai, Thailand, Aug. 27 – Sept. 2, 1994. CIFOR Center for International Forestry Research, Bogor, Indonesia, 67–85.
Nathan R., Muller-Landau H.C. 2000. Spatial patterns of seed dispersal, their determinants and consequences for recruitment. *Trends in Ecology and Evolution* 15(7), 278–285.
Naumann B.T., Crawford S.S. 2009. Is it possible to identify habitat for a rare species? Shortjaw Cisco (*Coregonus zenithicus*) in Lake Huron as a case study. *Environmental Biology of Fishes* 86(2), 341–348.
NCC. 1990. *Handbook for Phase 1 habitat survey—a technique for environmental audit.* Nature Conservancy Council, Peterborough. http://jncc.defra.gov.uk/pdf/pub90_HandbookforPhase1HabitatSurvey.pdf
Neave H.M., Cunningham R.B., Norton T.W., Nix H.A. 1996. Biological inventory for conservation evaluation III. Relationships between birds, vegetation and environmental attributes in southern Australia. *Forest Ecology and Management* 85(1-3), 197–218.
Nel E.M., Wessman C.A., Veblen T.T. 1994. Digital and visual analysis of Thematic Mapper imagery for differentiating old growth from younger spruce-fir stands. *Remote Sensing of the Environment* 48(3), 291–301.
Neldner V.J., Crossley D.C., Cofinas M. 1995. Using Geographic Information Systems (GIS) to determine the adequacy of sampling in vegetation surveys. *Biological Conservation* 73(1), 1–17.
Neumann K., Elbersen B.S., Verburg P.H., Staritsky I., Pérez-Soba M., Vries W., Rienks W.A. 2009. Modelling the spatial distribution of livestock in Europe. *Landscape Ecology* 24(9), 1207–1222.
Neumann M., Starlinger F. 2001. The significance of different indices for stand structure and diversity in forests. *Forest Ecology and Management* 145(1), 91–106.
Newbold T., Reader T., El-Gabbas A., Berg W., Shohdi W.M., Zalat S., El Din S.B., Gilbert F. 2010. Testing the accuracy of species distribution models using species records from a new field survey. *Oikos* 119(8),

1326–1334.
Nicotra A.B., Chazdon R.L., Iriarte S.V.B. 1999. Spatial heterogeinity of light and woody seedling regeneration in tropical wet forests. *Ecology* 80(6), 1908–1926.
Niel T.G. van, McVicar T.R., Datt B. 2005. On the relationship between training sample size and data dimensionality: Monte Carlo analysis of broadband multi-temporal classification. *Remote Sensing of Environment* 98(4), 468–480.
Nielsen S.E., Cranston J., Stenhouse G.B. 2009. Identification of priority areas for grizzly bear conservation and recovery in Alberta, Canada. *Journal of Conservation Planning* 5, 38–60.
North M., Greenberg J. 1998. Stand conditions associated with truffle abundance in western hemlock/Douglas-fir forests. *Forest Ecology and Management* 112(1-2), 55–66.
Nuske R.S., Sprauer S., Saborowski J. 2009. Adapting the pair-correlation function for analysing the spatial distribution of canopy gaps. *Forest Ecology and Management* 259(1), 107–116.
Nyrop J.P., Binns N.J.P 1992. Algorithms for computing operating characteristic and average sample number functions for sequential sampling plans based on binomial count models and revised plans for european red mite (Acari: Tetranychidae) on apple. *Journal of Economic Entomology* 85(4), 1253–1273.
Odeh I.O.A, McBratney A.B. 2000. Using AVHRR images for spatial prediction of clay content in the lower Namoi Valley of eastern Australia. *Geoderma* 97(3–4), 237–254.
Oden N.L. 1984. Assessing the significance of spatial correlogram. *Geographical Analysis* 16(1), 1–16.
Oden N.L., Sokal R.R., Fortin M.-J., Goebl H. 1993. Categorical wombling; detecting regions of significant change in spatially located categorical variables. *Geographical Analysis* 25(4), 315–336.
Odum H.T. 1973. Energy, ecology and economics. *Ambio* 2(6), 220–227.
O'Hanley J.R. 2009. NeuralEnsembles: a neural network based ensemble forecasting program for habitat and bioclimatic suitability analysis. *Ecography* 32(1), 89–93.
Ohser J. 1983. On estimators for the reduced second moment measure of point process. *Statistics* 14(1), 63–71.
Ohser J., Stoyan D. 1981. On the second-order and orientation analysis of planar stationary point process. *Biometrical Journal* 23(6), 253–333.
Okabe A., Boots B., Sugihara K. 1992. *Spatial tesselations: concepts and applications of Voronoi diagrams*. New York, John Wiley.
Okabe A., Boots B., Sugihara K. 1994. Nearest neighbourhood operations with generalized Voronoi diagrams: a review. *International Journal of Geographic Information Systems* 8(1), 43–71.
Oksanen J., Minchin P.R. 2002. Continuum theory revisited: what shape are species responses along ecological gradients? *Ecological Modelling* 157(2-3), 119–129.
Olev R., Kull A. 2004. Tuulikute energiatoodangu modelleerimine ja paigutuse optimeerimine tuuleparkides. *Publicationes Instituti Geographici Universitas Tartuensis* 89, 421–436.
Oliver M.A., Webster R. 1986. Semi-variograms for modelling the spatial pattern of landform and soil properties. *Earth Surface Processes and Landforms* 11(5), 491–504.
Oliver M.A., Webster R. 1989. A geostatistical basis for spatial weighting in multivariate classification. *Mathematical Geology* 21(1), 15–35.
Olivier F., Wotherspoon S.J. 2006. Modelling habitat selection using presence-only data: case study of a colonial hollow nesting bird, the snow petrel. *Ecological Modelling* 195(3-4), 187–204.
O'Neill R.V., Krummel J.R., Gardner R.H., Sugihara G., Jackson B., DeAngelis D.L., Milne B.T., Turner M.G., Zygmunt B., Christensen S.W., Dale V.H., Graham R.L. 1988. Indices of landscape pattern. *Landscape Ecology* 1(3), 153–162.
O'Neill R.V., Gardner R.H., Turner M.G. 1992. A hierarchical neutral model for landscape analysis. *Landscape Ecology* 7(1), 55–61.
Onof C., Chandler R.E., Kakou A., Northrop P., Wheater H.S., Isham V. 2000. Rainfall modelling using Poisson-cluster processes: a review of developments. *Stochastic Environmental Research and Risk Assessment* 14(6), 384–411.
Oostermeijer J.G.B., van Swaay C.A.M. 1998. The relationship between butterflies and environmental indicator values: a tool for conservation in a changing landscape. *Biological Conservation* 86(3), 271–280.
Openshaw S. 1984. *The modifiable areal unit problem*. Concepts and Techniques in Modern Geography No. 28. Geo Books. Norwich.
Openshaw S., Taylor P.J. 1979. A million of so correlation coefficients: three experiments on the modifiable areal unit problem. N. Wrigley (toim). Statistical Applications in the Spatial Sciences. Pion, London, 127–144. Cit. Marceau (1999).
Ord J.K., Getis A. 1995. Local spatial autocorrelation statistics: distributional issues and application. *Geographical Analysis* 27(4), 286–306.
Ormeling F. 1992. Brunet and the revival of French geography and cartography. *The Cartographic Journal* 29(1), 20–24.
Ortega-Huerta M.A., Peterson A.T. 2004. Modelling spatial patterns of biodiversity for conservation

prioritization in North-eastern Mexico. *Diversity and Distributions* 10(1), 39–54.
Osborne J.M, Brearley D.R. 2000. Completion criteria – case studies considering bond relinquishment and mine decommissioning: Western Australia. *International Journal of Surface Mining, Reclamation and Environment* 14(3), 193–204.
Osborne P.E., Alonso J.C., Bryant R.G. 2001. Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. *Journal of Applied Ecology* 38(2), 458–471.
Osborne P.E., Foody G.M., Suárez-Seoane S. 2007. Non-stationarity and local approaches to modelling the distributions of wildlife. *Diversity and Distributions* 13(3), 313–323.
Osborne P.E., Leitão P.J. 2009. Effects of species and habitat positional errors on the performance and interpretation of species distribution models. *Diversity and Distributions* 15(4), 671–681.
Osborne P.E., Suárez-Seoane S. 2002. Should data be partitioned spatially before building large-scale distribution models? *Ecological Modelling* 157(2-3), 249–259.
Ottaviani D., Lasinio G.J., Boitani L. 2004. Two statistical methods to validate habitat suitability models using presence-only data. *Ecological Modelling* 179(4), 417–443.
Overpeck T.J., Webb T.I., Prentice I.C. 1985. Quantitative interpretation of fossil pollen spectra: dissimilarity coefficients and the method of modern analogs. *Quaternary Research* 23(1), 87–108.
Overmars K.P., de Koning G.H.J., Veldkamp A. 2003. Spatial autocorrelation in multi-scale land use models. *Ecological Modelling* 164(2-3), 257–270.
Oviir M., Remm K., Linder M. 2008. Eesti põhikaardi okas-, sega- ja lehtmetsa eristatavus kaugseire-andmete ja mullakaardi järgi, kasutades näidistele tuginevat järeldamist ja tehisõpet. Väljataga K., Kaukver K. (toim). Kaugseire Eestis. Tallinn, Tartu Observatoorium, Keskkonnaministeeriumi Info- ja Tehnokeskus, 69–77.
Pacala S.W. 1987. Neighbourhood models of plant population dynamics 3. Models with spatial heterogeneity in the physical environment. *Theoretical Population Biology* 31(3), 359–392.
Pacala S.W., Canham C.D., Silander J.A.J. 1993. Forest models defined by field measurements: I. The design of a northeastern forest simulator. *Canadian Journal of Forest Research* 23(10), 1980–1988.
Pacala S.W., Canham C.D., Silander J.A.J., Kobe R.K., Ribbens E. 1996. Forest models defined by field measurements: estimation, error analysis and dynamics. *Ecological Monographs* 66(1), 1–43.
Pàez A., Uchida T., Miyamoto K. 2002. A general framework for estimation and inference of geographically weighted regression models. 1. Location-specific kernel bandwidths and a test for locational heterogeneity. *Environment and Planning A* 34(4), 733–754.
Pal M., Mather P.M. 2003. An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sensing of Environment* 86(4), 554–565
Palmer M.W. 1992. The coexistance of species in fractal landscapes. *American Naturalist* 139(2), 375–397.
Palmer M.W., van der Maarel E. 1995. Variance in species richness, species association, and niche limitation. *Oikos* 73(2), 203–213.
Pandit S.N., Hayward A., de Leeuw J., Kolasa J. 2010. Does plot size affect the performance of GIS-based species distribution models? *Journal of Geographical Systems* 12(4), 389–407.
Pantle R., Puck H. 1955. Die biologische Überwachung der Gewässer und die Darstellung der Ergebnisse. *Gas und Waserfach* 96(18), 604S. Cit. Sládeček (1973).
Papeş M., Gaubert P. 2007. Modelling ecological niches from low numbers of occurrences: assessment of the conservation status of poorly known viverrids (Mammalia, Carnivora) across two continents. *Diversity and Distributions* 13(6), 890–902.
Parolo G., Rossi G., Ferrarini A. 2008. Toward improved species niche modelling: *Arnica montana* in the Alps as a case study. *Journal of Applied Ecology* 45(5), 1410–1418.
Parsons B., Short J., Roberts J.D. 2009. Using community observations to predict the occurrence of malleefowl (*Leipoa ocellata*) in the Western Australian wheatbelt. *Biological Conservation* 142(2), 364–374.
Parviainen M., Marmion M., Luoto M., Thuiller W., Heikkinen R.K. 2009. Using summed individual species models and state-of-the-art modelling techniques to identify threatened plant species hotspots. *Biological Conservation* 142(11), 2501–2509.
Pasher J., King D. 2011. Development of a forest structural complexity index based on multispectral airborne remote sensing and topographic data. *Canadian Journal of Forest Research* 41(1), 44–58.
Pasher J., King D., Lindsay K. 2006. Modelling and mapping potential hooded warbler (*Wilsonia citrina*) habitat using remotely sensed imagery. *Remote Sensing of Environment* 107(3), 471–483.
Patton D.R. 1975. A diversity index for quantifying habitat edge. *Wildlife Society Bulletin* 3(4), 171–173.
Pearce J.L., Boyce M.S. 2006. Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology* 43(3), 405–412.
Pearce J., Ferrier S. 2001. The practical value of modelling relative abundance of species for regional conservation planning: a case study. *Biological Conservation* 98(1), 33–43.
Pearman P.B., Guisan A., Zimmermann N. 2011. Impacts of climate change on Swiss biodiversity: an indicator taxa approach. *Biological Conservation* 144(2), 866–875.

Pearman P.B., Weber D. 2007. Common species determine richness patterns in biodiversity indicator taxa. *Biological Conservation* 138(1-2), 109–119.
Pearson K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2(6), 559–572.
Pearson K. 1905. The problem of the Random Walk. *Nature* 72(1867), 294.
Pearson R.G., Dawson T.P. 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography* 12(5), 361–371.
Pearson R.G., Dawson T.P., Berry P.M., Harrison P.A. 2002. SPECIES: a spatial evaluation of climate impact on the envelope of species. *Ecological Modelling* 154(3), 289–300.
Pearson R.G., Raxworthy C.J., Nakamura M., Peterson A.T. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography* 34(1), 102–117.
Pearson R.G., Terence P., Dawson T.P., Liu C. 2004. Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. *Ecography* 27(3), 285–298.
Pearson R.G., Thuiller W., Araujo M.B., Martinez-Meyer E., Brotons L., McClean C., Miles L., Segurado P., Dawson T.P., Lees D.C. 2006. Model-based uncertainty in species range prediction. *Journal of Biogeography* 33(10), 1704–1711.
Pearson S.M., Gardner R.H. 1997. Neutral models: useful tools for understanding landscape patterns. Bissonette J.A. (toim). Wildlife and landscape ecology, effects of pattern and scale. New York, Springer-Verlag, 215–230.
Pebesma E.J. 2004. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences* 30(7), 683–691.
Pebesma E.J., Wesseling C.G. 1998. Gstat: a program for geostatistical modelling, prediction and simulation. *Computers & Geosciences* 24(1), 17–31.
Pekkarinen A. 2002. Image segment-based spectral features in the estimation of timber volume. *Remote Sensing of Environment* 82(2-3), 349–359.
Penttinen A., Stoyan D., Henttonen H.M. 1992. Marked point-processes in forest statistics. *Forest Science* 38(4), 806–824.
Pereira J.M.C., Itami R.M. 1991. GIS-based habitat modelling using logistic multiple regression: a case study of the Mt Graham Red Squirrel. *Photogrammetric Engineering and Remote Sensing* 57(11), 1475–1486.
Peres-Neto P.R., Jackson D. 2001. How well do multivariate data sets match? The advantages of a Procrustean superimposion approach over the Mantel test. *Oecologia* 129(2), 169–178.
Perrin N. 1984. *Contribution à l'écologie du genre Cepaea (Gastropoda): Approche descriptive et expérimentale de l'habitat et de la niche écologique.* Doctoral dissertation, University of Lausanne. Cit. Guisan ja Zimmermann (2000).
Perry G.L.W., Miller B.P, Enrigh N.J. 2006. A comparison of methods for the statistical analysis of spatial point patterns in plant ecology. *Plant Ecology* 187(1), 59–82.
Perry J.N. 1995. Spatial analysis by distance indices. *Journal of Animal Ecology* 64(3), 303–314.
Perry J.N. 1998. Measures of spatial pattern for counts. *Ecology* 79(3), 1008–1017.
Perry J.N., Hewitt M. 1991. A new index of aggregation for animal counts. *Biometrics* 47(4), 1505–1518.
Perry J.N., Liebhold A.M., Rosenberg M.S., Dungan J., Miriti M., Jakomulska A., Citron-Pousty S. 2002. Illustrations and guidelines for selecting statistical methods for quantifying spatial pattern in ecological data. *Ecography* 25(5), 578–600.
Peterson A.T. 2007. Why not WhyWhere: The need for more complex models of simpler environmental spaces. *Ecological Modelling* 203(3-4), 527–530.
Peterson A.T., Papeş M., Soberón J. 2008. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling* 213(1), 63–72.
Peterson C.J., Squiers E.R. 1995. An unexpected change in spatial pattern across 10 years in an aspen-white pine forest. *Journal of Ecology* 83(5), 847–855.
Petrou Z., Petrou M. 2011. A review of remote sensing methods for biodiversity assessment and bioindicator extraction. 2nd International Conference on Space Technology. Athens, Greece, DOI:10.1109/ICSpT.2011.6064679.
Pew K.L., Larsen C.P.S. 2001. GIS analysis of spatial and temporal patterns of human-caused wildfires in the temperate rain forest of Vancouver Island, Canada. *Forest Ecology and Management* 140(1), 1–18.
Pielou E.C. 1959. The use of point to plant distances in the study of the pattern of plant populations. *Journal of Ecology* 47(3), 607–613.
Pielou E.C. 1960. A single mechanism to account for regular, random and aggregated populations. *Journal of Ecology* 48(3), 575–584.
Pielou E.C. 1961. Segregation and symmetry in two-species populations as studied by nearest-neighbour relationships. *Journal of Ecology* 49(2), 255–269.

Pielou E.C. 1977. *Mathematical ecology*. New York etc. Wiley-Interscience.
Pielou E.C. 1984. *The interpretation of ecological data. A primer on classification and ordination*. New York etc. John Wiley.
Pietikäinen M., Rosenfeld A. 1981. Image segmentation by texture using pyramid node linking. *IEEE Transactions on Systems, Man, and Cybernetics* 11, 822–825.
Pinto R., Patrício J., Baeta A., Fath B., Neto J.M., Marques J.C. 2009. Review and evaluation of estuarine biotic indices to assess benthic condition. *Ecological Indicators* 9(1), 1–25.
Phillips S.J., Anderson R.P., Schapire R.E. 2006. Maximum entropy modelling of species geographic distributions. *Ecological Modelling* 190(3-4), 231–259.
Phillips S.J., Dudík M., Schapire R.E. 2004. A maximum entropy approach to species distribution modeling. Proceedings of the Twenty-First International Conference on Machine Learning, Banff, Alberta, Canada, 655–662.
Platt T., Denman K.L. 1975. Spectral analysis in ecology. *Annual Review of Ecology and Systematics* 6(1), 189–210.
Platts P.J., McClean C.J., Lovett J.C., Marchant R. 2008. Predicting tree distributions in an East African biodiversity hotspot: model selection, data bias and envelope uncertainty. *Ecological Modelling* 218(1-2), 121–134.
Plotnick R.E., Gardner R.H., O'Neill R.V. 1993. Lacunarity indices as measures of landscape texture. *Landscape Ecology* 8(3), 201–211.
Podani J., Miklós I. 2002. Resemblance coefficients and the horseshoe effect in principal coordinates analysis. *Ecology* 83(12), 3331–3343.
Pollard J.H. 1971. On distance estimators of density in randomly distributed forests. *Biometrics* 27(4), 991–1002.
Pontius R.G. 2000. Quantification error versus location error in comparison of categorical maps. *Photogrammetric Engineering and Remote Sensing* 66(8), 1011–1016.
Portalés C., Boronat N., Pardo-Pascual J.E., Balaguer-Beser A. 2010. Seasonal precipitation interpolation at the Valencia region with multivariate methods using geographic and topographic information. *Interantional Journal of Climatology* 30(10), 1547–1563.
Porté A., Bartelink H.H. 2002. Modelling mixed forest growth: a review of models for forest management. *Ecological Modelling* 150(1-2), 141–188.
Potts J.M., Elith J. 2006. Comparing species abundance models. *Ecological Modelling* 199(2), 153–163.
Pouteau R., Meyer J.-Y., Stoll B. 2011. A SVM-based model for predicting distribution of the invasive tree Miconia calvescens in tropical rainforests. *Ecological Modelling* 222(15), 2631–2641.
Pouliot D.A., King D.J., Bell F.W., Pitt D.G. 2002. Automated tree crown detection and delineation in high-resolution digital camera imagery of coniferous forest regeneration. *Remote Sensing of Environment* 82(2-3), 322–334.
Prasad A.M., Iverson L.R., Liaw A. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* 9(2), 181–199.
Prentice I.C., Bartlein P.J., Webb T.III. 1991. Vegetation and climate change in eastern North America since the last glacial maximum. *Ecology* 72(6), 2038–2056.
Prentice I.C, Cramer W., Harrison S.P., Leemans R., Monserud R.A., Solomon A.M. 1992. A global biome model based on plant physiology and dominance, soil properties and climate. *Journal of Biogeography* 19(2), 177–134.
Proosa H. 2008. Optimaalse kerneli ulatus maakatteüksuste eristamiseks digitaalsest rastrist. Väljataga K., Kaukver K. (toim). Kaugseire Eestis. Tallinn, Tartu Observatoorium, Keskkonnaministeeriumi Info- ja Tehnokeskus, 91–105.
Propastrin P., Kappas M., Erasmi S. 2007. Application of geographically weighted regression to investigate the impact of scale on prediction uncertainty by modelling relationship between vegetation and climate. *International Journal of Spatial Data Infrastructures Research* 3, 73–94.
Pukkala T., Kangas J., Kniivilä M., Tianen A.M. 1997. Integrating forest-level and compartment-level indices of species diversity with numerical forest planning. *Silva Fennica* 31(4), 417–429.
Pulliam H.R. 1988. Sources, sinks, and population regulation. *American Naturalist* 132(5), 652–661.
Puttock G.D., Shakotko P., Rasaputra J.G. 1996. An empirical model for moose, *Alces alces*, in Algonquin Park, Ontario. *Forest Ecology and Management* 81(1), 169–178.
Pärn J., Remm K., Mander Ü. 2010. Correspondence of vegetation boundaries to redox barriers in a Northern European moraine plain. *Basic and Applied Ecology* 11(1), 54–64.
Pyrcz M.J., Deutsch C.V. 2001. Two artifacts of probability field simulation. *Mathematical Geology* 33(7), 775–799.
Quinlan J.R. 1986. Induction of decision trees. *Machine Learning* 1(1), 81–106.
Quiot J. 1990. Methodology of the last climatic cycle reconstruction in France from pollen data.

Palaeogeography, Palaeoclimatology, Palaeoecology 80(1), 49–69.
Radeloff V.C., Miller T.F., He H.S., Mladenoff D.J. 2000. Periodicity in spatial data and geostatistical models: autocorrelation between patches. *Ecography* 23(1), 81–91.
Radeloff V.C., Mladenoff D.J., He H.S., Boyce M.S. 1999. Forest landscape change in the northwestern Wisconsin Pine Barrens from pre-European settlement to the present. *Canadian Journal of Forest Research* 29(11), 1649–1659.
Rajamoney S.A., Lee H.-Y. 1991. Prototype-based reasoning: an integrated approach to solving large novel problems. AAAI-91 Proceedings, 34–39.
Ramer U. 1972. An iterative procedure for the polygonal approximation of plane curves. *Computer Graphics and Image Processing* 1(3), 244–256.
Randin C.F., Dirnböck T., Dullinger S., Zimmerman N.E., Zappa M., Guisan A. 2006. Are species distribution models transferable in space? *Journal of Biogeography* 33(10), 1689–1703.
Rangel T.F., Diniz-Filho J.A.F., Bini L.M. 2010. SAM: a comprehensive application for Spatial Analysis in Macroecology. *Ecography* 33(1), 46–50.
Rangel T.F.L.V.B., Diniz-Filho J.A.F., Araújo M.B. 2009. *BIOENSEMBLES 1.0. Software for Computer Intensive Ensemble Forecasting of Species Distributions Under Climate Change*. Goiás, Madrid, Évora. Cit. Diniz-Filho *et al.* (2010).
Rangel T.F.L.V.B., Diniz-Filho J.A.F., Bini L.M. 2006. Towards an integrated computational tool for spatial analysis in macroecology and biogeography. *Global Ecology and Biogeography* 15(4), 321–327.
Ranson K.J., Sun G. 1994. Northern forest classification using temporal multifrequency and multipolarimetric SAR images. *Remote Sensing of the Environment* 47(2), 142–153.
Ranson K.J., Sun G. 1997. An evaluation of AIRSAR and SIR-C/X-SAR images for maping northern forest attributes in Maine, USA. *Remote Sensing of the Environment* 59(2), 203–222.
Ranson K.J., Sun G., Kharuk V.I., Kovacs K. 2001. Characterization of forests in Western Sayani Mountains, Siberia from SIR-C SAR data. *Remote Sensing of the Environment* 75(2), 188–200.
Ranta E., Kaitala V., Lindstrom J., Helle E. 1997a. The Moran effect and synchrony in population dynamics. *Oikos* 78(1), 136–142.
Ranta E., Lindström J., Kaitala V., Kokko H., Lindén H., Helle E. 1997b. Solar activity and hare dynamics: cross-continental comparison. *American Naturalist* 149(4), 765–775.
Rao C.R. 1982. Diversity and dissimilarity coefficients: a unified approach. *Theoretical Population Biology* 21(1), 24–43.
Rapp J., Wang D., Capen D., Thompson E., Lautzenheiser T. 2005. Evaluating error in using the national vegetation classification system for ecological community mapping in northern New England. *Natural Areas Journal* 25(1), 46–54.
Ratajski L. 1967. Phénomènes des points de généralisation. *International Yearbook of Cartography* 7, 143–151. Cit. McMaster ja Shea (1992).
Ray N., Burgman M.A. 2006. Subjective uncertainties in habitat suitability maps. *Ecological Modelling* 195(3-4), 172–186.
Rayburn A.P., Schiffers K., Schupp E.W. 2011. Use of precise spatial data for describing spatial patterns and plant interactions in a diverse Great Basin shrub community. *Plant Ecology* 212(4), 585–594.
Rayburn A.P., Wiegand T. 2012. Individual species – area relationships and spatial patterns of species diversity in a Great Basin, semi-arid shrubland. *Ecography* 35(), 341–347.
Reader S. 2000. Using survival analysis to study spatial point patterns in geographical epidemiology. *Social Science & Medicine* 50(7-8), 985–1000.
Real L.A., McElhany P. 1996. Spatial pattern and process in plant-pathogen interactions. *Ecology* 77(4), 1011–1025.
Recknagel F. 2001. Applications of machine learning to ecological modelling. *Ecological Modelling* 146(1-3), 303–310.
Reese G.C., Wilson K.R., Hoeting J.A., Flather C.H. 2005. Factors affecting species distribution predictions: a simulation modeling experiment. *Ecological Applications*, 15(2), 554–564.
Reese H., Nilsson M., Pahlén T.G., Hagner O., Joyce S., Tingelöf U., Egberth M., Olsson H. 2003. Countrywide estimates of forest variables using satellite data and field data from the National Forest Inventory. *Ambio* 32(8), 542–548.
Reese H.M., Lillesand T.M., Nagel D.E., Stewart J.S., Goldmann R.A., Simmons T.E., Chipman J.W., Tessar P.A. 2002. Statewide land cover derived from multiseasonal Landsat TM data. A retrospective of the WISCLAND project. *Remote Sensing of Environment* 82(2-3), 224–237.
Regan H.M., Colyvan M., Burgman M.A. 2002. A taxonomy and treatment of uncertainty for ecology and conservation biology. *Ecological Applications* 12(2), 618–628.
Reggiani A., Nijkamp P., Sabella E. 2001. New advances in spatial network modelling: towards evolutionary algorithms. *European Journal of Operational Research* 128(2), 385–401.

Reich R.M., Czaplewski R.L., Bechtold W.A. 1995. Spatial cross-correlation of undisturbed natural shortleaf pine stands in northern Georgia. *Environmental and Ecological Statistics* 1(3), 201–217.
Reich R.M., Joy S.M., Reynolds R.T. 2004. Predicting the location of northern goshawk nests: modeling the spatial dependency between nest locations and forest structure. *Ecological Modelling* 176(1–2), 109–133.
Reimer A.W. 2010. Understanding chorematic diagrams: towards a taxonomy. *The Cartographic Journal* 47(4), 330–350.
Reimer A., Fohringer J. 2010. Towards constraint formulation for chorematic schematisation tasks – work in progress. Geographic Information on Demand 13th Workshop of the ICA commission on Generalisation and Multiple Representation, Zürich.
Rejwan C., Collins N.C., Brunner L.J., Shuter B.J., Ridgway M.S. 1999. Tree regression analysis on the nesting habitat of smallmouth bass. *Ecology* 80(1), 341–348.
http://edoc.gfz-potsdam.de/gfz/get/15984/0/69f6e49369e936e066db19bc5e1cdb6a/15984.pdf
Remm K. 1976. *Putukakoosluste uurimise metoodikast. Võistlustöö*. Käsikiri Tartu Ülikooli Ökoloogia ja Maateaduste Instituudis.
Remm K. 1987. A statistical method for the assessment of ecological relationships on the example of *Filinia longiseta* (Her.) (*Rotatoria*). *Proceedings of the Academy of Sciences of the Estonian SSR. Biology* 36(4), 319–326.
Remm K. 1989. A zooplankton bioindication system for the Matsalu Bay: a probabilistic approach. *Proceedings of the Academy of Sciences of the Estonian SSR. Biology* 38(1), 61–71.
Remm K. 2000. Eesti ruutkilomeetrite andmebaas. Frey T. (toim). Kaasaegse ökoloogia probleemid. Tartu; 26.-27. aprill, 2000. Tartu, Teadusühing IM SAARE, 241–247.
Remm K. 2002. Otepää looduspargi taimkatte kasvukohatüüpide kaart. Frey T. (toim). Eesti süsinikubilansi ökoloogiast ja ökonoomikast. Eesti XIII ökoloogiapäev 03. mail 2002 Tartus EPMÜ aulas. Tartu, 62–76.
Remm K. 2004. Case-based predictions for species and habitat mapping. *Ecological Modelling* 177(3-4), 259–281.
Remm K. 2005. Correlations between forest stand diversity and landscape pattern in Otepää Nature Park, Estonia. *Journal for Nature Conservation* 13(2-3), 137–145.
Remm K., Jaagus J., Briede A., Rimkus E., Kelviste T. 2011. Interpolative mapping of mean precipitation in the Baltic countries by using landscape characteristics. *Estonian Journal of Earth Sciences* 60(3), 172–190.
Remm K., Kelviste T. 2011a. *Constud Tutorial*. Tartu, University of Tartu, Chair of Geoinformatics and Cartography.
Remm K., Kelviste T. 2011b. *Constud Tutorial*. Tartu, University of Tartu, Chair of Geoinformatics and Cartography. http://hdl.handle.net/10062/18192.
Remm K., Kelviste T. 2011c. *Tarkvarasüsteemi Constud kasutamisõpetus*. Tartu, Tartu Ülikool, geoinformaatika ja kartograafia õppetool. http://hdl.handle.net/10062/18193.
Remm K., Külvik M., Mander Ü., Sepp K. 2004. Design of the Pan-European ecological network: a national level attempt. Jongman R.H.G., Pungetti G. (toim). Ecological networks and Greenways. Concept, Design, Implementation. Cambridge University Press, 151–170.
Remm K., Linder M. 2007. Prognoosisüsteemi Pidevstuudium tutvustus. *Geodeet* 34, 37–43.
Remm K., Linder M., Remm L. 2009. Relative density of finds for assessing similarity-based maps of orchid occurrence. *Ecological Modelling* 220(3), 294–309.
Remm K., Luud A. 2003. Regression and point pattern models of moose distribution in relation to habitat distribution and human influence in Ida-Viru county, Estonia. *Journal for Nature Conservation* 11(3), 197–211.
Remm J., Lõhmus A., Remm K. 2006. Tree cavities in riverine forests: what determines their occurrence and use by hole-nesting passerines? *Forest Ecology and Management* 221(1-3), 267–277.
Remm K., Mander Ü. 2001. Euroopa ökoloogilise võrgustiku kujundamine Eesti näitel. Frey T. (toim). Eesti loodus ja Euroopa Liit: Eesti XII ökoloogiapäev, 27. aprill 2001, Tartus, 70–83.
Remm K., Oja T. 2001. Stepwise modelling of rural housing pattern near Otepää, Estonia using neighbourhood corrections. *Publicationes Instituti Geographici Universitatis Tartuensis* 92, 157–162.
Remm K., Palo A., Linder M., Pärn J. 2010. Mitmevariandilisus taimkatte välikaardistusel. http://kalleremm.ee/Doc/Valikaardistuse_esinduslikkus_16-11-2010.pdf.
Remm K., Remm L. 2009. Similarity-based large-scale distribution mapping of orchids. *Biodiversity and Conservation*. 18(6), 1629–1647.
Remm M., Remm K. 2008. Case-based estimation of the risk of enterobiasis. *Artificial Intelligence in Medicine* 43(3), 167–177.
Remmel T.K., Csillag F. 2003. When are two landscape pattern indices significantly different? *Journal of Geographical Systems* 5(4), 331–351.
Remy N., Boucher A., Wu J. 2009. *Applied geostatistics with SGeMS: a users guide*. New York, Cambridge University Press.
Renkonen O. 1938. Statistisch-ökologische Untersuchungen über die Terrestrische Käferwelt der Finnischen

Bruchmoore. *Suomalaisen Eläin- ja Kasvitieteellisen Seuran Vanamon Eläintieteelisiä Julkaisuja* 6(1), 1–231.
Reutter B.A., Helfer V., Hirzel A.H., Vogel P. 2003. Modelling habitat-suitability using museum collections: an example with three sympatric Apodemus species from the Alps. *Journal of Biogeography* 30(1), 581–590.
Reyes E., White M.L., Martin J.F., Kemp G.P., Day J.W., Aravamuthan V. 2000. Landscape modeling of coastal habitat change in the Mississippi delta. *Ecology* 81(8), 2331–2349.
Ricotta C. 2002. Bridging the gap between ecological diversity indices and measures of biodiversity with Shannon's entropy: comment to Izsák and Papp. *Ecological Modelling* 152(1), 1–3.
Ricotta C., Carranza M.L., Avena G., Blasi C. 2002. Are potential natural vegetation maps a meaningful alternative to neutral landscape models? *Applied Vegetation Science* 5(2), 271–275.
Ricotta C., Corona P., Marchetti M., Chirici G., Innamorati S. 2003. LaDy: software for assessing local landscape diversity profiles of raster land cover maps using geographic windows. *Environmental Modelling Software* 18(4), 373–378.
Ridgeway G. 1999. The state of boosting. *Computing Sciences and Statistics* 31, 172–181.
Riitters K.H., O'Neill R.V., Hunsaker C.T., Wickham J.D., Yankee D.H., Timmins S.P., Jones K.B., Jackson B.L. 1995. A factor analysis of landscape pattern and structure metrics. *Landscape Ecology* 10(1), 23–39.
Ripa J. 2000. Analysing the Moran effect and dispersal: their significance and interaction in synchronous population dynamics. *Oikos* 89(1), 175–187.
Ripley B.D. 1976. The second-order analysis of stationary point processes. *Journal of Applied Probability* 13(2), 225–266.
Ripley B.D. 1977. Modelling spatial patterns. *Journal of the Royal Statistical Society B* 39(2), 172–212.
Ripley B.D. 1981. *Spatial statistics*. New York, Wiley.
Ripley B.D. 1982. Edge effects in spatial stochastic processes. Ranneby B. (toim). Statistics in theory and practice. Essays in honour of Bertil Matérn. Umeå, Swedish University of Agricultural Sciences, 247–262.
Ripley B.D. 1985. Analyses of nest spacings. Lecture notes in statistics. 29. Statistics in ornithology. Berlin, Heidelberg, New York, Tokyo, Springer, 151–158.
Ritchie L.E., Betts M.G., Forbes G., Vernes K. 2009. Effects of landscape composition and configuration on northern flying squirrels in a forest mosaic. *Forest Ecology and Management* 257(9), 1920–1929.
Roberts D.W. 1996. Landscape vegetation modelling with vital attributes and fuzzy systems theory. *Ecological Modelling* 90(2), 175–184.
Robinson K.A., Ramsay K., Lindenbaum C., Frost N., Moore J., Wright A.P., Petrey D. 2011. Predicting the distribution of seabed biotopes in the southern Irish Sea. *Continental Shelf Research* 31(2), 120–131.
Rodrígues M.A., Lewis W.M.Jr. 1997. Structure of fish assemblages along environmental gradients in floodplain lakes of the Orinico River. *Ecological Monographs* 67(1), 109–128.
Rodríguez J.P., Brotons L., Bustamante J., Seoane J. 2007. The application of predictive modelling of species distribution to biodiversity conservation. *Diversity and Distributions* 13(3), 243–251.
Rohlf F.J., Slice D.E. 1990. Extensions of the Procrustes method for the optimimal superimposion of landmarks. *Systematic Zoology* 39(1), 40–59.
Romero-Calcerrada R., Luque S. 2006. Habitat quality assessment using weights-of-evidence based GIS modelling: The case of *Picoides tridactylus* as species indicator of the biodiversity value of the Finnish forest. *Ecological Modelling 196*(1-2), 62–76.
Romme W.H. 1982. Fire and landscape diversity in subalpine forests of Yellowstone National Park. *Ecological Monographs* 52(2), 199–221.
Rosenberg M.S. 2004. Wavelet analysis for detecting anisotropy in point patterns. *Journal of Vegetation Science* 15(2), 277–284.
Rosenberg M.S., Anderson C.D. 2011. PASSaGE: Pattern Analysis, Spatial Statistics and Geographic Exegesis. Version 2. *Methods in Ecology and Evolution* 2(3), 229–232.
Rosenblatt F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* 65(6), 386–408.
Rosenzweig M.L. 1995. Species diversity in space and time. Cambridge University Press.
Rossi R.E., Mulla D.J., Journel A.G., Franz E.H. 1992. Geostatistical tools for modeling and interpreting ecological spatial dependence. *Ecological Monographs* 62(2), 277–314.
Rotenberry J.T., Preston K.L., Knick S.T. 2006. GIS-based niche modeling for mapping species habitat. *Ecology* 87(6), 1458–1464.
Roxburgh S.H., Chesson P. 1998. A new method for detecting species associations with spatially autocorrelated data. *Ecology* 79(6), 2180–2192.
Roy P.S. Tomar S. 2000. Biodiversity characterization at landscape level using geospatial modelling technique. *Biological Conservation* 95(1), 95–109.
Ružička M. 1958. Anwendung mathematisch-statischer Methoden in der Geobotanik (synthetische Bearbeitung von Aufnahmen). *Biológia (Bratislava)* 13(9), 647–661.

Ryherd S., Woodcock C. 1996. Combining spectral and texture data in the segmentation of remotely sensed images. *Photogrammetric Engineering and Remote Sensing* 62(2), 181–194.
Saarenmaa H., Stone N.D., Folse L.J., Packard J.M., Grant W.E., Makela M.E., Coulson R.N. 1988. An artificial intelligence modelling approach to simulating animal/habitat interactions. *Ecological Modelling* 44(1-2), 125–141.
Sabol D.E.Jr., Gillespie A.R., Adams J.B., Smith M.O., Tucker C.J. 2002. Structural stage in Pacific Northwest forests estimated using simple mixing models of multispectral images. *Remote Sensing of the Environment* 80(1), 1–16.
Sachot S. 2002. *Viability and management of an endangered capercaillie (Tetrao urogallus) metapopulation.* Doctoral dissertation, Lausanne, Institute of Ecology.
Saetre P. 1999. Spatial patterns of ground vegetation, soil microbial biomass and activity in a mixed spruce-birch stand. *Ecography* 22(2), 183–192.
Salas F., Marcos C., Neto J.M., Patrício J., Pérez-Ruzafa A., Marques J.C. 2006. User-friendly guide for using benthic ecological indicators in coastal and marine quality assessment. *Ocean and Coastal Management* 49 (5-6), 308–331.
Saltelli A., Ratto M., Andres T., Campolongo F., Cariboni J., Gatelli D., Saisana M., Tarantola S. 2008. *Global Sensitivity Analysis. The Primer*. Wiley. Cit. http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470059974.html.
Saltelli A., Tarantola S., Campolongo F., Ratto M. 2004. *Sensitivity Analysis in Practice. A Guide to Assessing Scientific Models*. Wiley.
Santos X., Brito J.C., Sillero N., Pleguezuelos J., Llorente G., Fahd S., Parellada X. 2006. Inferring habitat-suitability areas with ecological modelling techniques and GIS: a contribution to assess the conservation status of *Vipera latastei*. *Biological Conservation* 130(3), 416–425.
Saracco J.F., Royle A.J., DeSante D.F., Gardner B. 2010. Modeling spatial variation in avian survival and residency probabilities. *Ecology* 91(7), 1885–1891.
Sazanova L., Osipov G., Godovnikov M. 1999. Intelligent system for fish stock prediction and allowable catch evaluation. *Environmental Modelling & Software* 14(5), 391–399.
Saunders S.C., Chen J., Drummer T.D., Gustafson E.J., Brosofske K.D. 2005. Identifying scales of pattern in ecological data: a comparison of lacunarity, spectral and wavelet analyses. *Ecological Complexity* 2(1), 87–105.
Saura S., Martínez-Millán J. 2000. Landscape patterns simulation with a modified random clusters method. *Landscape Ecology* 15(7), 661–678.
Saveraid E.H., Debinski D.M., Kindscher K., Jakubauskas M.E. 2001. A comparison of satellite data and landscape variables in predicting bird species occurrences in the Greater Yellowstone Ecosystem. *Landscape Ecology* 16(1), 71–83.
Schaefer J.A., Morellet N., Pépin D., Verheyden H. 2008. The spatial scale of habitat selection by red deer. *Canadian Journal of Zoology* 86(12), 1337–1345.
Scheller R.M., Domingo J.B., Sturtevant B.R., Williams J.S., Rudy A., Gustafson E.J., Mladenoff D.J. 2007. Design, development, and application of LANDIS-II, a spatial landscape simulation model with flexible spatial and temporal resolution. *Ecological Modelling* 201(3-4), 409–419.
Scheller R.M., Mladenoff D.J. 2007. An ecological classification of forest landscape simulation models: tools and strategies for understanding broad-scale forested ecosystems. *Landscape Ecology* 22(4), 491–505.
Schiffers K., Schurr F.M., Tielbörger K., Urbach C., Moloney K., Jeltsch F. 2008. Dealing with virtual aggregation—a new index for analyzing heterogeneous point patterns. *Ecography* 31(5), 545–555.
Schleiter I.M., Borcchardt D., Wagner R., Dapper T., Schmidt K.-D., Schmidt H.-D., Werner H. 1999. Modelling water quality, bioindication and population dynamics in lotic ecosystems using neural networks. *Ecological Modelling* 120(2-3), 271–286.
Schröder B., Seppelt R. 2006. Analysis of pattern–process interactions based on landscape models–overview, general concepts, and methodological issues. *Ecological Modelling* 199(4), 505–516.
Schröder M., Walessa M., Rehrauer H., Seidel K., Datcu M. 2000. Gibbs random field models: a toolbox for information extraction. *Computers & Geosciences* 26(4), 423–432.
Schuerholz G. 1974. Quantitative evaluation of edge from aerial photographs. *Journal of Wildlife Management* 38(4), 913–920.
Schumaker N.H. 1996. Using landscape indices to predict habitat connectivity. *Ecology* 77(4), 1210–1225.
Scull P., Franklin J., Chadwick O.A., McArthur D. 2003. Predictive soil mapping: a review. *Progress in Physical Geography* 27(2), 171–197.
Segurado P., Araujo M.B. 2004. An evaluation of methods for modelling species distributions. *Journal of Biogeography* 31(10), 1555–1568.
Seoane J., Bustamante J., Díaz-Delgado R. 2004. Are existing vegetation maps adequate to predict bird distributions? *Ecological Modelling* 175(2), 137–149.

Seoane J., Carrascal L.M., Alonso C.L., Palomino D. 2005. Species-specific traits associated to prediction errors in bird habitat suitability modelling. *Ecological Modelling* 185(2-4), 299–308.
Seppelt R., Voinov A. 2002. Optimization methodology for land use patterns using spatially explicit landscape models. *Ecological Modelling* 151(2-3), 125–142.
Seppelt R., Voinov A. 2003. Optimization methodology for land use patterns – evaluation based on multiscale habitat pattern comparison. *Ecological Modelling* 168(3), 217–231.
Shan Y., Paull D., McKay R.I. 2006. Machine learning of poorly predictable ecological data. *Ecological Modelling* 195(1-2), 129–138.
Shandley J., Franklin J., White T. 1996. Testing the Woodcock-Harward image segmentation algorithm in an area of southern California chapparral and woodland vegetation. *International Journal of Remote Sensing* 17(5), 983–1004.
Shannon C. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27, 379–423.
Shea K.S., McMaster R.B. 1989. Cartographic generalization in a digital environment: when and how to generalize. *Autocarto* 9, 56–67.
Sheail J., Bunce R.G.H. 2003. The development and scientific principles of an environmental classification for strategic ecological survey in the United Kingdom. *Environmental Conservation* 30(2), 147–159.
Sherburne S.S., Bissonette J.A. 1994. Marten subnivean access point use: response to subnivean prey levels. *Journal of Wildlife Management* 58(3), 400–405.
Shi H., Laurent E.J., LeBouton J., Racevskis L, Hall K.R., Donovan M., Doepker R.V, Walters M.B., Lupi F., Liu J. 2006. Local spatial modeling of white-tailed deer distribution. *Ecological Modelling* 190(1-2), 171–189.
Shi W., Fisher P.F., Goodchild M.F. 2002. *Spatial data quality*. Taylor & Francis.
Shimatani K. 2001. Multivariate point process and spatial variation of species diversity. *Forest Ecology and Management* 142(1-3), 215–229.
Silbernagel J., Moeur M. 2001. Modeling canopy openness and understory gap patterns based on image analysis and mapped tree data. *Forest Ecology and Management* 149(1-3), 217–233.
Simberloff D. 1998. Flagships, umbrellas, and keystones: is single-species management passé in the landscape era? *Biological Conservation* 83(3), 247–257.
Simberloff D., Abele L.G. 1976. Refuge design and island biogeography theory: effects of fragmentation. *American Naturalist* 120(1), 40–50.
Simpson E.H. 1949. Measurement of diversity. *Nature* 163(4148), 688.
Skidmore A.K., Gauld A., Walker P. 1996. Classification of kangaroo habitat distribution using three GIS models. *International Journal of Geographical Information Science* 10(4), 441–454.
Sládeček V. 1973. System of water quality from the biological point of view. *Archiv für Hydrobiologie Beiheft Ergebnisse der Limnologie* 7(1-4), 1–218.
Sládeček V. 1983. Rotifers as indicators of water quality. *Hydrobiologia* 100, 169–201.
Smith P.A. 1994. Autocorrelation in logistic regression modelling of species' distributions. *Global Ecology and Biogeography Letters* 4(2), 47–61.
Smulders M., Nelson T.A., Jelinski D.E., Nielsen S.E., Stenhouse G.B. 2010. A spatially explicit method for evaluating accuracy of species distribution models. *Diversity and Distributions* 16(6), 996–1008.
Snelder T., Leathwick J.R., Dey K. 2007. A procedure for making optimal selection of input variables for multivariate environmental classifications. *Conservation Biology* 21(2), 365–375.
Snelder T.H., Leathwick J.R., Dey K.L., Rowden A.A., Weatherhead M.A., Fenwick G.D., Francis M.P., Gorman R.M., Grieve J.M., Hadfield M.G., Hewitt J.E., Richardson K.M., Uddstrom M.J., Zeldis J.R. 2006. Development of an ecologic marine classification in the New Zealand region. *Environmental Management* 39(1), 12–29.
Snelder T.H., Lehmann A., Lamouroux N., Leathwick J.R., Allenbach K. 2009. Strong influence of variable treatment on the performance of numerically defined ecological regions. *Environmental Management* 44(4), 658–670.
Snelder T., Lehmann A., Lamouroux N., Leathwick J., Allenbach K. 2010. Effect of classification procedure on the performance of numerically defined ecological regions. *Environmental Management* 45(5), 939–952.
Soares A. 2001. Direct sequential simulation and cosimulation. *Mathematical Geology* 33(8), 911–926.
Sokal R.R., Oden N. 1978a. Spatial autocorrelation in biology. 1. Methodology. *Biological Journal of the Linnean Society* 10(2), 199–228.
Sokal R.R., Oden N. 1978b. Spatial autocorrelation in biology. 2. Some biological implications and four applications of evolutionary and ecological interest. *Biological Journal of the Linnean Society* 10(2), 229–249.
Sokal R.R., Oden N., Thomson B.A. 1998. Local spatial autocorrelation in biological variables. *Biological Journal of the Linnean Society* 65(1), 41–62.
Sokal R.R., Rohlf F.J. 1995. *Biometry, third edition*. New York, Freeman.
Sokal R.R., Unnasch R.S, Thomson B.A. 1991. *Pemphigus* revisited: changes in geographic variation but

constancy in variability and covariation. *Evolution* 45(7), 1585–1605.
Sołtysiak A., Jaskulski P. 1999. Czekanowski's diagram. A method of multidimensional clustering. Barceló J.A., Briz I., Vila A. (toim). New Techniques for Old Times. CAA 98. Computer Applications and Quantitative Methods in Archaeology. Proceedings of the 26th Conference, Barcelona, March 1998, BAR International Series 757. Oxford, 175–184.
Somers G.L., Oderwald R.G. 1988. Estimating and constructing confidence intervals for spatial patterns between random and regular. *Ecological Modelling* 44(1-2), 57–72.
Sørensen T. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter / Kongelige Danske Videnskabernes Selskab* 5(4), 1–34.
Spann M., Wilson R. 1985. A quad-tree approach to image segmentation which combines statistical and spatial information. *Pattern Recognition* 18(3-4), 257–269.
Srivastava R.M. 1992. Reservoir characterisation with probability field simulation. SPE Annual Conference and Exhibition, paper 24753. Washington, 927–938.
Szava-Kovats R. 2001. *Assessment of stream sediment contamination by median sum of weighted residuals regression.* Doctoral dissertation, University of Tartu.
Szwagrzyk J. 1990. Small-scale spatial patterns of trees in a mixed *Pinus sylvestris-Fagus sylvatica* forest. *Forest Ecology and Management* 51(4), 301–315.
Szwagrzyk J., Czerwczak M. 1993. Spatial patterns of trees in natural forests of East-Central Europe. *Journal of Vegetation Science* 4(4), 469–476.
St-Louis V., Pidgeon A.M., Radeloff V.C., Hawbakwer T.J., Clayton M.K. 2006. High-resolution image texture as a predictor of bird species richness. *Remote Sensing of Environment* 105(4), 299–312.
St-Onge B.A., Cavayas F. 1995. Estimating forest stand structure from high-resolution imagery using the directional variogram. *International Journal of Remote Sensing* 16(11), 1999–2021.
St-Onge B.A., Cavayas F. 1997. Automated forest structure mapping from high resolution imagery based on directional semivariogram estimates. *Remote Sensing of the Environment* 61(1), 82–85.
Stage A.R., Wykoff W.R. 1993. Calibrating a model of stochastic effects on diameter increment for individual-tree simulations of stand dynamics. *Forest Science* 39(4), 692–705.
Staub J.E., Knerr L.D., Holder D.J., May B. 1992. Phylogenetic relationships among several African Cucumis species. *Canadian Journal of Botany* 70(3), 509–517.
Stankovski V., Debeljak M., Bratko I., Adamič M. 1998. Modelling the population dynamics of red deer (*Cervus elaphus* L.) with regard to forest development. *Ecological Modelling* 108(1-3), 145–153.
Statsoft Inc. 2011. *Electronic Statistics Textbook.* Tulsa, OK: Statsoft. http://www.statsoft.com/textbook.
Sterner R.W., Ribie C.A., Schatz G.E. 1986. Testing for life historical changes in spatial patterns of four tropical tree species. *Journal of Ecology* 74(3), 621–633.
Stevens J.P., Blackstock T.H., Howe E.A., Stevens D.P. 2004. Repeatability of Phase 1 habitat survey. *Journal of Environmental Management* 73(1), 53–59.
Stewart-Oaten A. 1996. Goals in environmental monitoring. Schmitt R.J., Osenberg C.W. (toim). Detecting ecological impacts: concepts and applications in coastal habitats. San Diego, California, USA, Academic Press, 17–27.
Stockwell D. 2007. *Niche modeling: predictions from statistical distributions.* Chapman & Hall/CRC.
Stockwell D.R.B. 2006. Improving ecological niche models by data mining large environmental datasets for surrogate models. *Ecological Modelling* 192(1-2), 188–196.
Stockwell D.R.B., Beach J.H., Stewart A., Vorontsov G., Vieglais D., Scachetti Pereira R. 2006. The use of the GARP genetic algorithm and Internet grid computing in the Lifemapper world atlas of species biodiversity. *Ecological Modelling* 195(1-2), 139–145.
Stockwell D.R., Noble I.R. 1992. Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. *Mathematics and Computers in Simulation* 33(5-6), 385–390.
Stockwell D.R., Peters D. 1999. The GARP modeling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science* 13(2), 143–158.
Stockwell D.R.B., Peterson A.T. 2002. Effects of sample size on accuracy of species distribution. *Ecological Modelling* 148(1), 1–13.
Stockwell D.R.B., Peterson A.T. 2003. Comparison of resolution of methods used in mapping biodiversity patterns from point-occurrence data. *Ecological Indicators* 3(3), 213–221.
Stone J.R.N., Champernowne D.G., Meade J.E. 1942. The precision of national income estimates. *The Review of Economic Studies* 9(2), 111–125.
Store R., Jokimäki J. 2003. A GIS-based multi-scale approach to habitat suitability modeling. *Ecological Modelling* 169(1), 1–15.
Store R., Kangas J. 2001. Integrating spatial multi-criteria evaluation and expert knowledge for GIS-based habitat suitability modelling. *Landscape and Urban Planning* 55(2), 79–93.

Stoyan D. 1988. Thinnings of point processes and their use in the statistical analysis of a settlement pattern with deserted villages. *Statistics* 19(1), 45–56.
Stoyan D., Penttinen A. 2000. Recent applications of point process methods in forestry statistics. *Statistical Science* 15(1), 61–78.
Stoyan D., Stoyan H. 1998. Non-homogeneous Gibbs process models for forestry – a case study. *Biometrical Journal* 40(5), 521–531.
Straalen N.M. van 1998. Evaluation of bioindicator systems derived from soil arthropod communities. *Applied Soil Ecology* 9(1-3), 429–437.
Straus D.M., Krishnamurthy V. 2007. The preferred structure of the interannual Indian monsoon variability. *Pure and Applied Geophysics* 164(8-9), 1717–1732.
Strauss D.J. 1975. A model for clustering. *Biometrica* 62(2), 467–475.
Strebelle S. 2002. Conditional simulation of complex geological structures using multiplepoint statistics. *Mathematical Geology* 34(1), 1–22.
Stum A.K., Boettinger J.L., White M.A., Ramsey R.D. 2010. Chapter 15. Random forests applied as a soil spatial predictive model in arid Utah. J.L. Boettinger et al. (toim.). Digital soil mapping. Progress in soil science 2. pp. 179–189. Springer Science.
Stümer W., Kenter B., Köhl M. 2010. Spatial interpolation of in situ data by self-organizing map algorithms (neural networks) for the assessment of carbon stocks in European forests. *Forest Ecology and Management* 260(3), 287–293.
Sutcliffe O.L., Thomas C.D., Moss D. 1996. Spatial synchrony and asynchrony in butterfly population dynamics. *Journal of Animal Ecology* 65(1), 85–95.
Syartinilia, Tsuyuki S. 2008. GIS-based modeling of Javan Hawk-Eagle distribution using logistic and autologistic regression models. *Biological Conservation* 141(3), 756–769.
Syphard A.D., Franklin J. 2009. Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. *Ecography* 32(6), 907–918.
Zaniewski A.E., Lehmann A., Overton J.McC. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling* 157(2-3), 261–280.
Zar J.H. 1996. *Biostatistical analysis, 3rd ed*. Prentice Hall, Upper Saddle River, NJ.
Zavala M.A., Burkey T.V. 1997. Application of ecological models to landscape planning: the case of the Mediterranean basin. *Landscape and Urban Planning* 38(3-4), 213–227.
Zawadzki J., Cieszewski C.J., Zasada M., Lowe R.C. 2005. Applying geostatistics for investigations of forest ecosystems using remote sensing imagery. *Silva Fennica* 39(4), 599–617.
Zelinka M., Marvan P. 1961. Zur Präzisierung der biologischen Klassifikation der Reinheit fließender Gewässer. *Archiv für Hydrobiologie* 57(3), 389–407.
Zenner E.K. 2005. Investigating scale-dependent stand heterogeneity with structure-area-curves. *Forest Ecology and Management* 209(1-2), 87–100.
Zenner E.K., Hibbs D.E. 2000. A new method for modeling the heterogeneity of forest structure. *Forest Ecology and Management* 129(1-3), 75–87.
Zhang J., Yim Y-S., Yang J. 1997. Intelligent selection of instances for prediction functions in Lazy learning algorithms. *Artificial Intelligence Review* 11(1-5), 175–191.
Zhang L., Bib H., Cheng P., Davis C.J. 2004. Modeling spatial variation in tree diameter–height relationships. *Forest Ecology and Management* 189(1-3), 317–329.
Zhang Z.-H., Hu G., Zhu J.-D., Luo D.-H., Ni J. 2010. Spatial patterns and interspecific associations of dominant tree species in two old-growth karst forests, SW China. *Ecological Research* 26(6), 1151–1160.
Zimmermann N.E., Edwards T.C.Jr., Graham C.H., Pearman P.B., Svenning J.-C. 2010. New trends in species distribution modelling. *Ecography* 33(6), 985–989.
Zimmermann N.E., Kienast F. 1999. Predictive mapping of alpine grasslands in Switzerland: species versus community approach. *Journal of Vegetation Science* 10(4), 469–482.
Zimmermann N.E., Yoccoz N.G., Thomas C., Edwards T.C.Jr., Meier E.S., Thuiller W., Guisan A., Schmatz D.R., Pearman P.B. 2009. Climatic extremes improve predictions of spatial patterns of tree species. *Proceedings of the National Academy of Sciences of the United States of America* 106(s2), 19723–19728.
Zobel M. 1992. Plant species coexistence – the role of historical, evolutionary and ecological factors. *Oikos* 65(2), 314–320.
Zobel M. 1997. The relative role of species pools in determining plant species richness: an alternative explanation of species coexistence? *Trends in Ecology and Evolution* 12(7), 266–269.
Zobel M., Rüdiger O., Laanisto L., Naranjo-Cigala A., Pärtel M., Fernández-Palacios J.M. 2011. The formation of species pools: historical habitat abundance affects current local diversity. *Global Ecology and Biogeography* 20(2), 251–259.
Zuo W., Lao N., Geng Y., Ma K. 2008. GeoSVM: an efficient and effective tool to predict species' potential distributions. *Journal of Plant Ecology* 1(2), 143–145.

Zurell D., Jeltsch F., Dormann C.F., Schröder B. 2009. Static species distribution models in dynamically changing systems: how good can predictions really be? *Ecography* 32(5), 733–744.
Zweig M.H., Campbell G. 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* 39(4), 561–577.
Tainz P. 2001. *Chorème*. Lexikon der Kartographie und Geomatik, Vol 1. Heidelberg, Spektrum, 118–119.
Tan P-N., Steinbach M., Kumar V. 2006. *Introduction to Data Mining*. Boston, Pearson Education.
Tang X., Hurni L. 2009. Regional spatial planning maps: a sino-swiss comparison of cartographic visualization methodologies. Proceedings of the 24th International Cartographic Conference (ISBN 978-1-907075-02-5). Chile: Santiago.
Tamm T., Remm K. 2009. Estimating the parameters of forest inventory using machine learning and the reduction of remote sensing features. *International Journal of Applied Earth Observation and Geoinformation* 11(4), 290–297.
Tappeiner U., Tasser E., Tappeiner G. 2001. Modelling vegetation patterns using natural and anthropogenic influence factors: preliminary experience with GIS based model applied to an Alpine area. *Ecological Modelling* 113(1-3), 225–237.
Tatem A.J., Lewis H.G., Atkinson P.M., Nixon M.S. 2001. Multiple-class land-cover mapping at the sub-pixel scale using a Hopfield neural network. *International Journal of Applied Earth Observation and Geoinformation* 3(2), 184–190.
Tatem A.J., Lewis H.G., Atkinson P.M., Nixon M.S. 2002. Super-resolution land cover pattern prediction using a Hopfield neural network. *Remote Sensing of the Environment* 79(1), 1–14.
Taylor L.R. 1961. Aggregation, variance and the mean. *Nature* 189(4766), 732–735.
Taylor P.D., Fahrig L., Henein K., Merriam G. 1993. Connectivity is a vital element of landscape structure. *Oikos* 68(3), 571–572.
Teixeira J., Ferrand N., Arntzen J. 2001. Biogeography of the golden-striped salamander *Chioglossa lusitanica*: a field survey and spatial modelling approach. *Ecography* 24(5), 618–624.
Telesco D.J., Van Manen F.T. 2006. Do black bears respond to military weapons training? *The Journal of Wildlife Management* 70(1), 222–230.
Termansen M., McClean C.J., Preston C.D. 2006. The use of genetic algorithms and Bayesian classification to model species distributions. *Ecological Modelling* 192(3-4), 410–424.
Thomas C.D., Kunin W.E. 1999. The spatial structure of populations. *Journal of Animal Ecology* 68(4), 647–657.
Thomas D.L., Taylor E.J. 2006. Study designs and tests for comparing resource use and availability II. *Journal of Wildlife Management* 70(2), 324–336.
Thompson H.R. 1956. Distribution of distance to nth neighbor in a population of randomly distributed individuals. *Ecology* 37(2), 391–394.
Thompson L.M., van Manen F.T., Schlarbaum S. E., M. DePoy. 2006. A spatial modeling approach to identify potential Butternut restoration sites in Mammoth Cave National Park. *Restoration Ecology* 14(2), 289–296.
Thomson J., Weiblen G., Thomson B., Alfaro S., Legendre P. 1996. Untangling multiple factors in spatial distributions: lilies, gophers and rocks. *Ecology* 77(6), 1698–1715.
Thuiller W. 2003. Biomod - optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology* 9(10), 1353–1362.
Thuiller W. 2004. Patterns and uncertainties of species' range shifts under climate change. *Global Change Biology* 10(12), 2020–2027.
Thuiller W., Lafourcade B., Engler R., B. Araújo M. 2009. BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography* 32(3), 369–373.
Tian Y., Yue T., Zhu L., Clinton N. 2005. Modeling population density using land cover data. *Ecological Modelling* 189(1-2), 72–88.
Tiit E. 1972. *Matemaatilise statistika tabelid. II*. Tartu, TRÜ.
Tischendorf L., Fahrig L. 2000. On the usage and measurement of landscape connectivity. *Oikos* 90(1), 7–19.
Titeux N., Dufrene M., Radoux J., Hirzel A.H., Defourny P. 2007. Fitness-related parameters improve presence-only distribution modelling for conservation practice: the case of the red-backed shrike. *Biological Conservation* 138(1-2), 207–223.
Tokola T. 2000. The influence of field sample data location on growing stock volume estimation in Landsat TM-based forest inventory in Eastern Finland. *Remote Sensing of the Environment* 74(3), 422–431.
Tokola T., Kilpeläinen P. 1999. The forest stand margin area in the interpretation of growing stock using Landsat TM imagery. *Canadian Journal of Forest Research* 29(3), 303–309.
Tokola T., Sarkeala J., Van Der Linden M. 2001. Use of topographic correction in Landsat TM based forest interpretation in Nepal. *International Journal of Remote Sensing* 22(4), 551–563.
Tokola T., Shrestha S.M. 1999. Comparison of cluster-sampling techniques for forest inventory in southern

Nepal. *Forest Ecology and Management* 116(1-3), 219–231.
Tomppo E. 1986. *Models and methods for analyzing spatial patterns of trees.* Communicationes Instituti Forestalis Fenniae No. 138. Helsinki, Finland, The Finnish Forest Research Institute.
Tomppo E., Katila M. 1991. Satellite image-based national forest inventory of Finland. Proceedings of IGARSS'91, Remote Sensing: Global Monitoring for Earth Management. 1991 International Geoscience and Remote Sensing Symposium. Helsinki University of Technology, Espoo, Finland, June 3-6, 1991. Vol. III. 1141–1144.
Tomppo E., Goulding C., Katila M. 1999. Adapting Finnish multisource forest inventory techniques to the New Zealand preharvest inventory. *Scandinavian Journal of Forest Research* 14(2), 182–192.
Tomppo E., Heikkinen 1999. National Forest Inventory of Finland—past, present and future. Alho J. (toim). Statistics, Registries, and Science—Experiences from Finland 89–108. Helsinki, Statistics Finland.
Tomppo E., Nilsson M., Rosengren M., Aalto P., Kennedy P. 2002. Simultaneous use of Landsat-TM and IRS-1C WiFS data in estimating large area tree stem volume and aboveground biomass. *Remote Sensing of Environment* 82(1), 156–171.
Trakhtenbrot A., Kadmon R. 2005. Environmental cluster analysis as a tool for selecting complementary networks of conservation sites. *Ecological Applications* 15(1), 335–345.
Trexler J.C., Travis J. 1993. Nontraditional regression analyses. *Ecology* 74(6), 1629–1637.
Trotter C.M., Dymond J.R., Goulding C.J. 1997. Estimation of timber volume in a coniferous plantation forest using Landsat-TM. *International Journal of Remote Sensing* 18(10), 2209–2223.
Tryon R.C. 1939. *Cluster analysis*. Ann Arbor, MI, Edwards Brothers. Cit. Tryon ja Bailey (1970).
Tryon R.C., Bailey D.E. 1970. *Cluster analysis*. New York etc., McGraw-Hill.
Trzcinski M.K., Fahrig L., Merriam G. 1999. Independent effects of forest cover and fragmentation on the distribution of forest breeding birds. *Ecological Applications* 9(2), 586–593.
Tsoar A., Allouche O., Steinitz O., Rotem D., Kadmon R. 2007. A comparative evaluation of presence only methods for modelling species distribution. *Diversity and Distributions* 13(4), 397–405.
Tukey J.W. 1977. *Exploratory data analysis.* Reading, MA, Addison-Wesley.
Turner M.G., Constanza R., Sklar F.H. 1989. Methods to evaluate the perfomance of spatial simulation models. *Ecological Modelling* 48(1-2), 1–18.
Turner S.J., O'Neill R., Conley W., Conley M., Humphries H.C. 1991. Pattern and scale: statistics for landscape ecology. Turner M.G., Gardner R.H. (toim). Quantitative methods in landscape ecology. The analysis and interpretation of landscape heterogeneity. Ecological Studies 82. New York etc., Springer, 17–49.
Ulrich W., Buszko J. 2003. Species-area relationships of butterflies in Europe and species richness forecasting. *Ecography* 26(3), 365–373.
Underwood E., Klinger R., Moore P. 2004. Predicting patterns of non-native plant invasions in Yosemite National Park, California, USA. *Diversity and Distributions* 10(5-6), 447–459.
Uranov A.A. 1965. Fitogennoe pole. Lavrenko E.M. (toim). Problemy sovremennoj botaniki. 1. Nauka, Moscow, 251–254.
Urban D.L., Bonan G.B., Smith T.M., Shugart H.H.F. 1991. Spatial applications of gap models. *Forest Ecology and Management* 42(1-2), 95–110.
Upton G., Fingleton B. 1985. *Spatial data analysis by example. Vol. 1. Point pattern and quantitative data*. New York, Wiley.
Upton G., Fingleton B. 1989. *Spatial data analysis by example. Vol 2. Categorical and directional data.* New York, Wiley.
Uuemaa E., Antrop M., Roosaare J., Marja R., Mander Ü. 2009. Landscape metrics and indices: an overview of their use in landscape research. *Living Reviews in Landscape Research* 3. http://www.livingreviews.org/lrlr-2009-1.
Uuttera J., Haara A., Tokola T., Maltamo M. 1998. Determination of the spatial distribution of trees from digital aerial photographs. *Forest Ecology and Management* 110(1-3), 275–282.
Václavík T., Meentemeyer R.K. 2009. Invasive species distribution modeling (iSDM): are absence data and dispersal constraints needed to predict actual distributions? *Ecological Modelling* 220(23), 3248–3258.
Vanden Borre J., Paelinckx D., Mucher C.A., Kooistra L., Haest B., De Blust G., Schmidt A.M. 2011. Integrating remote sensing in Natura 2000 habitat monitoring: Prospects on the way forward. *Journal for Nature Conservation* 19(2), 116–125.
Vandergast A.G., Bohonak A.J., Hathaway S.A., Boys J., Fishera R.N. 2008. Are hotspots of evolutionary potential adequately protected in southern California? *Biological Conservation* 141(6), 1648–1664.
Vapnik V.N. 1995. *The nature of statistical learning theory*. New York, Springer. Cit. Vapnik (1998).
Vapnik V.N. 1998. *Statistical learning theory*. New York, Wiley-Intersciense.
Veech J.A., Summerville K.S., Crist T.O., Gering J.C. 2002. The additive partitioning of species diversity: recent revival of an old idea. *Oikos* 99(1), 3–9.
Verboom J., Schotman A., Opdam P., Metz A.J. 1991. European nuthatch metapopulations in a fragmented

agricultural landscape. *Oikos* 61(2), 149–156.
Verboom J., Foppen R., Chardon P., Opdam P., Luttikhuizen P. 2001. Introducing the key patch approach for habitat networks with persistent populations: an example for marshland birds. *Biological Conservation* 100(1), 89–101.
Verbyla D.L., Litvaitis J.A. 1989. Resampling methods for evaluating classification accuracy of wildlife habitat models. *Environmental Management* 13(6), 783–787.
Veregin H. 1997. The effects of vertical error in digital elevation models on the determination of flow-path direction. *Cartography and Geographic Information Systems* 24(2), 67–79.
Verhoeye J., De Wulf R. 2000. Sub-pixel Mapping of Sahelian Wetlands using Multi-temporal SPOT VEGETATION Images. Proceedings of Vegetation 2000, April 3–6 2000, Lago Magiore, Italy, 96–104.
Verhulst P.-F. 1838. Notice sur la loi que la population poursuit dans son accroissement. *Correspondance Mathématique et Physique de l'Observatoire de Bruxelles* 10: 113–121. http://books.google.ee/books?hl=fr&id=8GsEAAAAYAAJ&jtp=113&redir_esc=y#v=onepage&q&f=false
Vetaas O.R. 2000. Comparing species temperature response curves: population density versus second-hand data. *Journal of Vegetation Science* 11(5), 659–666.
Vicente J., Alves P., Randin C., Guisan A., Honrado J. 2010. What drives invasibility? A multi-model inference test and spatial modelling of alien plant species richness patterns in northern Portugal. *Ecography* 33(6), 1081–1092.
Viña A., Bearer S., Zhang H., Ouyang Z., Liu J. 2008. Evaluating MODIS data for mapping wildlife habitat distribution. *Remote Sensing of Environment* 112(5), 2160–2169.
Vliet J. van, Bregt A.K., Hagen-Zanker A. 2011. Revisiting Kappa to account for change in the accuracy assessment of land-use change models. *Ecological Modelling* 222(8), 1367–1375.
Wahl M. 2001. Small scale variability of benthic assamblages: biotic neighbourhood effects. *Journal of Experimental Marine Biology and Ecology* 258(1), 101–114.
Walker J.S., Balling R.C., Briggs J.M., Katti M., Warren P.S., Wentz E.A. 2008. Birds of a feather: Interpolating distribution patterns of urban birds. *Computers, Environment and Urban Systems* 32(19), 19–28.
Walker P.A. 1990. Modelling wildlife distributions using a geographic information system: kangaroos in relation to climate. *Journal of Biogeography* 17(3), 279–289.
Walker P.A., Cocks K.D. 1991. HABITAT: a procedure for modelling a disjoint environmental envelope for a plant or animal species. *Global Ecology and Biogeography Letters* 1(4), 108–118.
Walker P.A., Moore D.M. 1988. SIMPLE: an inductive modelling and mapping tool for spatially-oriented data. *International Journal of Geographical Information Systems* 2(4), 347–363.
Walker S., Wilson J.B., Steel J.B., Rapson G.L., Smith B., King W.M., Cottam Y.H. 2003. Properties of ecotones: evidence from five ecotones objectively determined from a coastal vegetation gradient. *Journal of Vegetation Science* 14(4), 579–590.
Wallace C.S.A., Watts J., Yool S. 2000. Characterizing the spatial structure of vegetation communities in the Mojave Desert using geostatistical techniques. *Computers & Geosciences* 26(4), 397–410.
Wallerman J. 2003. *Remote Sensing Aided Spatial Prediction of Forest Stem Volume.* Doctoral dissertation, Acta Universitatis Agriculturae Sueciae Silvestria 271.
Wallerman J., Joyce S., Vencatasawmy C.P., Olsson H. 2002. Prediction of forest stem volume using kriging adapted to detected edges. *Canadian Journal of Forest Research* 32(3), 509–518.
Walley W.J., Fontama V.N. 1998. Neural network predictors of average score per taxon and number of families at unpolluted river sites in Great Britain. *Water Research* 32(3), 613–622.
Walters S. 2001. Landscape pattern and productivity effects on source-sink dynamics of deer populations. *Ecological Modelling* 143(1-2), 17–32.
Ward G., Hastie T., Barry S., Elith J., Leathwick J.R. 2009. Presence-only data and the EM algorithm. *Biometrics* 65(2), 554–563.
Ward J.H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58(301), 236.
Ward J.S., Parker G.R., Ferrandino F.J. 1996. Long-term spatial dynamics in an old-growth deciduous forest. *Forest Ecology and Management* 83(3), 189–202.
Warrick A., Myers D., Nielsen D. 1986. Geostatistical methods applied to soil science. Methods of Soil Analysis, Part 1. Physical and mineralogical methods. 2nd ed. American Society for Agronomy and Soil Science Society of America, Agronomy Monograph No. 9, 53–82.
Wartenberg D. 1985. Canonical trend surface analysis: a method for describing geographic patterns. *Systematic Zoology* 34(3), 259–279.
Watkins A.J., Wilson J.B. 1992. Fine-scale community structure of lawns. *Journal of Ecology* 80(1), 15–24.
Webster R. 1973. Automatic soil-boundary location from transect data. *Journal of the International Association of Mathematical Geology* 5(1), 27–37.

Weed D.L. 2005. Weight of evidence: a review of concept and methods. *Risk Analysis* 25(6), 1545–1557.
Weisberg S. 1980. *Applied linear regression*. New York. Wiley. Cit. Guisan ja Zimmermann (2000).
Welsh A.H., Cunningham R.B., Donnelly C.F., Lindenmayer D.B. 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling* 88(1-3), 297–308.
Wessels K.J., Van Jaarsveld A.S., Grimbeek J.D., Van der Linde M.J. 1998. An evaluation of the gradsect biological survey method. *Biodiversity and Conservation* 7(8), 1093–1121.
Western A.W., Bloschl G., Grayson R.B. 1998. How well do indicator variograms capture the spatial connectivity of soil moisture? *Hydrological Processes* 12(12), 1851–1868.
Wettschereck D., Aha D.W. 1995. Weighting features. Veloso M.M., Aamodt A. (toim). Case-Based Reasoning Research and Development, First International Conference, ICCBR-95, Sesimbra, Portugal, October 23-26, 1995, Proceedings. Lecture Notes in Computer Science 1010. Springer, 347–358.
Wettschereck D., Aha D.W., Mohri T. 1997. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review* 11(1-5), 273–314.
Whigham P.A. 2000. Induction of a marsupial density model using genetic programming and spatial relationships. *Ecological Modelling* 131(2-3), 299–317.
Whigham P.A. 2005. Applying case-based reasoning to explore freshwater phytoplankton dynamics. Lek S., Scardi M., Verdonschot P.F.M., Descy J.-P., Park Y.-S. (toim). Modelling Community Structure in Freshwater Ecosystems, Springer, Berlin Heidelberg New York, 263–272.
White M.A., Mladenoff D.J. 1994. Old growth forest landscape transitions from pre-European settlement to present. *Landscape Ecology* 9(3), 191–205.
Whittaker R.H. 1956. Vegetation of the Great Smoky Mountains. *Ecological Monographs* 26, 1–80.
Whittaker R.H. 1960. Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs* 30(3), 279–338.
Whittaker R.H. 1962. Classification of Natural Communities. *Botanical Review* 28(1), 1–239.
Whittaker R.H. 1972. Evolution and measurement of species diversity. *Taxon* 21(2-3), 213–251.
Wiegand T., Kissling W.D., Cipriotti P.A., Aguiar M.R. 2006. Extending point pattern analysis for objects of finite size and irregular shape. *Journal of Ecology* 94(4), 825-837.
Wiegand T., Martínez I., Huth A. 2009. Recruitment in tropical tree species: revealing complex spatial patterns. *The American Naturalist* 174(4), E106-E140.
Wiegand T., Moloney K.A. 2004. Rings, circles and null-models for point pattern analysis in ecology. *Oikos* 104(2), 209–229.
Wiegand T., Moloney K.A., Naves J., Knauer F. 1999. Finding the missing link between landscape structure and population dynamics: a spatially explicit perspective. *The American Naturalist* 154(6), 605–627.
Wiegand T., Gunatilleke C.V.S., Gunatilleke I.A.U.N., Huth A. 2007a. How individual species structure diversity in tropical forests. *Proceedings of the National Academy of Sciences of the United States of America* 104(48), 19029–19033.
Wiegand T., Gunatilleke S., Gunatilleke N. Okuda T. 2007b. Analyzing the spatial structure of a Sri Lankan tree species with multiple scales of clustering. *Ecology* 88(12), 3088–3102.
Wiens J.A. 1989. Spatial scaling in eology. *Functional Ecology* 3(4), 385–397.
Wiens J.A. 1997. Metapopulation dynamics and landscape ecology. Hanski I., Gilpin M. (toim). Metapopulation Biology: Ecology, Genetics and Evolution. London, Academic Press, 43–62.
Wilby R.L., Charles S.P., Zorita E., Timbal B., Whetton P., Mearns L.O. 2004. *Guidelines for Use of Climate Scenarios Developed from Statistical Downscaling Methods.* Data Distribution Centre of the Intergovernmental Panel on Climate Change. http://www.narccap.ucar.edu/doc/tgica-guidance-2004.pdf
Wilds S., Boetsch J., van Manen F.T., Clark J.D., White P.S. 2000. Modeling the distributions of species and communities in Great Smoky Mountains National Park. *Computers and Electronics in Agriculture* 27(1-3), 389–392.
Wilkinson G.G. 2005. Results and implications of a study of fifteen years of satellite image classification experiments. *IEEE Transactions on Geoscience and Remote Sensing* 43(3), 433–440.
Wilson B.A., Lewis A., Aberton J. 2003. Spatial model for predicting the presence of cinnamon fungus (*Phytophthora cinnamomi*) in sclerophyll vegetation communities in south-eastern Australia. *Austral Ecology* 28(2), 108–115.
Wilson J.B. 1995. Variance in species richness, niche limitation, and vindication of patch models. *Oikos* 73(2), 277–279.
Wilson S.D. 1991. Variation in competition in eucalypt forests: the importance of standardization in pattern analysis. *Journal Vegetation Science* 2(5), 577–586.
Wilson D.R., Martinez T.R. 2000. Reduction techniques for instance-based learning algorithms. *Machine Learning* 38(3), 257–286.
Wilson M.V., Mohler C.L. 1983. Measuring compositional change along gradients. *Vegetation* 54(3), 129–141.
Wissel C. 1992. Aims and limits of ecological modeling examplified by island theory. *Ecological Modelling*

63(1-4), 1–12.
Wisz M.S., Hijmans R.J., Li J., Peterson A.T., Graham C.H., Guisan A., the NCEAS Species Distribution Modelling Group. 2008. Effects of sample size on the performance of species distribution models. *Diversity and Distributions* 14(5), 763–773.
With K.A., King A.W. 1997. The use and misuse of neutral landscape models in ecology. *Oikos* 79(2), 219–229.
With K.A., King A.W. 2001. Analysis of landscape sources and sinks: the effect of spatial pattern on avian demography. *Biological Conservation* 100(1), 75–88.
With K.A., Gardner R.H., Turner M.G. 1997. Landscape connectivity and population distributions in heterogenous environments. *Oikos* 78(1), 151–169.
Wolda H. 1981. Similarity indices, sample size, and diversity. *Oecologia* 50(3), 296–302.
Wolter P.T., Mladenoff D.J., Host G.E., Crow T.R. 1995. Improved forest classification in the Northern Lake states using multi-temporal Landsat imagery. *Photogrammetric Engineering and Remote Sensing* 61(9), 1129–1143.
Womble W.H. 1951. Differential systematics. *Science* 114(2961), 315–322.
Wooda S.N., Augustin N.H. 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling* 157(2-3), 157–177.
Woodcock C., Harward V.J. 1992. Nested-hierarchical scene models and image segmentation. *International Journal of Remote Sensing* 13(16), 3167–3187.
Woodcock C.E., Strahler A.H. 1987. The factor of scale in remote sensing. *Remote Sensing of the Environment* 21(3), 311–332.
Woodcock C., Strachler A.H., Jupp D.L.B. 1988a. The use of variograms in remote sensing: I. Scene models and simulated images. *Remote Sensing of the Environment* 25(3), 323–348.
Woodcock C., Strachler A.H., Jupp D.L.B. 1988b. The use of variograms in remote sensing: II. Real digital image. *Remote Sensing of the Environment* 25(3), 349–379.
Wu H., Huffer F.W., 1997. Modeling the distribution of plant species using the autologistic regression model. *Environmental and Ecological Statistics* 4(1), 49–64.
Wu J., Jelinski D.E., Luck M., Tueller P.T. 2000. Multiscale analysis of landscape heterogeneity: scale variance and pattern metrics. *Geographic Information Sciences* 6(1), 6–19.
Wu J., Levin S.A. 1994. A spatial patch dynamic modeling approach to pattern and process in an annual grassland. *Ecological Monographs* 64(4), 447–464.
Wu J., Levin S.A. 1997. A patch-based spatial modeling approach: conceptual framework and simulation scheme. *Ecological Modelling* 101(2–3), 325–346.
Wu X., Kumar V., Quinlan J.R., Ghosh J., Yang Q., Motoda H., McLachlan G.J., Ng A., Liu B., Yu P.S., Zhou Z.-H., Steinbach M., Hand D.J., Steinberg D. 2008. Top 10 algorithms in data mining. *Knowledge and Information Systems* 14(1), 1–37.
Wu X.B., Smiens F.E. 2000. Multiple-scale habitat modeling approach for rare plant conservation. *Landscape and Urban Planning* 51(1), 11–28.
Wu H., Sharpe P.J.H., Walker J., Penridge L.K. 1985. Ecological field theory: a spatial analysis of resource interference among plants. *Ecological Modelling* 29(1-4), 215–243.
Wulder M.A., Franklin S.E., White J.C., Linke J., Magnussen S. 2006. An accuracy assessment framework for large-area land cover classification products derived from medium-resolution satellite data. *International Journal of Remote Sensing* 27(4), 663–683.
Wulf M. 2004. Plant species richness of afforestations with different former use and habitat continuity. *Forest Ecology and Management* 195(1-2), 191–204.
Wunderle A.L., Franklin S.E., Guo X.G. 2007. Regenerating boreal forest structure estimation using SPOT-5 pan-sharpened imagery. *International Journal of Remote Sensing* 28(19), 4351–4364.
Özesmi S.L., Özesmi U. 1999. An artificial neural network approach to spatial habitat modelling with interspecific interaction. *Ecological Modelling* 116(1), 15–31.
Xu X. 2003. Considerations for the use of SADIE statistics to quantify spatial patterns. *Ecography* 26(6), 821–830.
Yamanaka T., Tanaka K., Otuka A., Bjørnstad O.N. 2007. Detecting spatial structures and interactions in the ragweed (*Ambrosia artemissifolia* L.) and the ragweed beetle (*Ophraella communa* LeSage) populations. *Ecological Research* 22(2), 185–196.
Yang X., Skidmore A.K., Melick D.R., Zhou Z., Xu J. 2006. Mapping non-wood forest product (matsutake mushrooms) using logistic regression and a GIS expert system. *Ecological Modelling* 198(1-2), 208–218.
Yee T.W., Mitchell N.D. 1991. Generalized additive models in plant ecology. *Journal of Vegetation Science* 2(5), 587–602.
Yen P., Huettmann F., Cooke F. 2004. A large-scale model for the at-sea distribution and abundance of Marbled Murrelets (*Brachyramphus marmoratus*) during the breeding season in coastal British Columbia,

Canada. *Ecological Modelling* 171(4), 395–413.
Yoshioka P.M. 2008. Misidentification of the Bray-Curtis similarity index. *Marine Ecology Progress Series*. 368, 309–310. doi: 10.3354/meps07728.
Yu H., Wiegand T., Yang X., Ci L. 2009. The impact of fire and density-dependent mortality on the spatial patterns of a pine forest in the Hulun Buir sandland, Inner Mongolia, China. *Forest Ecology and Management* 257(10), 2098–2107.

Абакумов В.А. 1981. К истории контроля качества вод по гидробиологическим показателям.
Абакумов В.А. (toim). Научные основы контроля качества вод по гидробиологическим показателям: труды всесоюзной конференции, Москва, 1-3 ноября 1978 г. Ленинград. Всесоюзное гидробиологическое общество АН СССР, Гидрометеоиздат.
Гандин Л. С. 1963. *Объективный анализ метеорологических полей*. Ленинград. Гидрометеоиздат.

Mõistete register

E

F

G

H

I

J

K

L

M

N

O

P

R

S

T

U

V

W

Õ

Ä

Ö

Ü

Lisa 1. Kreeka tähestik

A	α	alfa	I	ι	ioota	P	ρ	roo
B	β	beeta	K	κ	kapa	Σ	σ	sigma
Γ	γ	gamma	Λ	λ	lambda	T	τ	tau
Δ	δ	delta	M	μ	müü	Y	υ	üpsilon
E	ε	epsilon	N	ν	nüü	Φ	φ	fii
Z	ζ	dzeeta	Ξ	ξ	ksii	X	χ	hii
H	η	eeta	O	o	omikron	Ψ	ψ	psii
Θ	θ	teeta	Π	π	pii	Ω	ω	oomega

Lisa 2. Eesti põhikaardi topoloogilised põhialad.

Eesti põhikaardi topoloogilised põhialad ja nende värvikood suletud areaalidena salvestamisel (Maa-amet 2002).

Põhiala	Värvikood
Vundament	109
Elu-, ühiskondlik hoone	236
Kõrvalhoone	237
Vare	110
Kasvuhoone	238
Muu ehitis	232
Katusealune	229
Meri	58
Muu veekogu	59
Vooluveekogu	57
Järv	48
Kalmistu	135
Haljasala	155
Eraõu	66
Tootmisõu	254
Põld	108
Rohumaa	106
Aiamaa	107
Põõsastik	95
Noor mets	67
Muu lage	104
Mets	64
Jäätmaa	103
Raba	102
Madalsoo	62
Raskestiläbitav soo	101
Mahajäetud turbaväli	100
Turbaväli	99
Teed, tänavad ja platsid	30
Piiritagune ala	250

Statistical Analysis of Earth and Ecological Data

By Kalle Remm, Jaanus Remm, Ants Kaasik

Summary

This textbook originally written in Estonian language gives a wide overview of basic statistical methods and methods of spatial analysis used in ecology and earth sciences. The book consists of six chapters: basic principles of statistical analysis, exploratory data analysis, statistical modelling, descriptive spatial analysis, spatial models, and pattern generation. Every chapter ends with test questions, all together 171. The text includes 339 formulae and 109 figures. The book includes a list of references containing 1395 items, a list of terms includes 854 items explained in the text.

The authors have personal experience in using methods for: similarity quotients, similarity based reasoning, description of point patterns, directional analysis, modelling spatial autocorrelation, neighbourhood effects, habitat suitability modelling, predictive distribution mapping in a detailed spatial scale, creating similarity based system for predictive mapping and subjectiveness of vegetation field mapping. Therefore these subchapters include original details and are somewhat longer than the other issues.

The text is based on a study of literature, however, the authors' effort to present methods in a hierarchical order and to express their own understanding should be evident. In addition, several figures present original examples of statistical analyses of empirical data with explanation and interpretation in legend.