TARTU ÜLIKOOL

MATEMAATIKA-INFORMAATIKATEADUSKOND

MATEMAATILISE STATISTIKA INSTITUUT

Kairiin Kütt

Leibkonnad ja perekonnad registripõhises rahva ja eluruumide loenduses

Magistritöö

Juhendaja: Mare Vähi

Tartu 2014

Leibkonnad ja perekonnad registripõhises rahva ja eluruumide loenduses

Järgmine rahva ja eluruumide loendus Eestis on planeeritud toimuma registripõhiselt. Leibkonnad ja perekonnad ei ole andmekogudest otsesel kujul kättesaadavad, vaid need tuleb kokku panna kaudse info põhjal. Raskuskese on (lasteta) vabaabielupartnerite määratlemisel – eelmise loenduse andmete peal hinnatakse võimalust partnerite tuvastamiseks logistilise regressioonimudeli abil. Mudelist paremaid tulemusi annab aga metoodika, mida rakendatakse Soome Statistikaametis – ka seda lähenemisviisi on töös tutvustatud ning sellele tuginedes on koostatud IML (*interactive matrix language*) programm, milles määratletakse lasteta vabaabielupartnerid. Koostatud on ka programmikood leibkonnaliikmete jagamiseks perekondadesse vastavalt definitsioonile. IML koodid on ette nähtud kasutamiseks ja vajadusel täiendamiseks käesoleva aasta sügisel, mil registripõhisele loendusele ülemineku raames viiakse läbi pilootloendus, mille väljundiks on muuhulgas erinevad leibkonna ja perekonna koosseisu iseloomustavad tunnused.

Märksõnad: rahvaloendused, leibkonnad, perekonnad, plokkskeemid, logistiline regressioon.

Households and families in register-based census

The next population and housing census in Estonia is intended to be register-based. Since information about households and families is not directly obtainable from relevant databases, their existence can only be ascertained from indirect data. The hardest challenge is identifying partners in a consensual union (in absence of mutual children). This work investigates two possible determination methods - a logistic regression model based on data from the previous census and the algorihtmic method currently used by Statistics Finland. The latter approach was found to be more accurate and has been used as a basis for creating an IML (*Interactive Matrix Language*) program for determining partners in a consensual union. Another program was also created for the purpose of dividing members of a household into nuclear families, according to definition. Both programs are intended to be used and refined during the register-based pilot census in autumn of 2014, with the purpose of creating different household and family related characteristics.

Keywords: censuses, households, nuclear families, block diagrams, logistic regression.

SISUKORD

Lühendite loetelu

ADS	Aadressiandmete süsteem
ADS_ID	Aadressi identifikaator aadressiandmete süsteemis
ADS_OID	Aadressiobjekti identifikaator aadressiandmete süsteemis
EHIS	Eesti Hariduse Infosüsteem
LAB	Lähteandmebaas
REL	Rahva ja eluruumide loendus
REL11	2011. aasta rahva ja eluruumide loendus
REL leibkond	Küsitluspõhine leibkond; ütluspõhine leibkond; majapidamisüksuse mõiste põhine leibkond
REGREL	Registripõhine rahva ja eluruumide loendus
REGREL leibkond	Registripõhine leibkond; eluruumipõhine leibkond; aadressipõhine leibkond
RR	Rahvastikuregister
SA	Statistikaamet

SISSEJUHATUS

Rahva ja eluruumide loendus on üks andmekogumise meetodeid, mis võimaldab saada teatud ajahetke seisuga andmeid riigi rahvastiku ja eluruumide kohta. Loenduse eesmärk on koguda kõikseid andmeid riigi elanike arvu, paiknemise ja elamistingimuste kohta, samuti inimeste soo, vanuse, haridustaseme, elatusallikate, tööhõive, tegevusalade ja paljude muude näitajate, sealhulgas leibkondade ja perekondade arvu ning koosseisu kohta.

Eestis on seatud eesmärgiks järgmine, 2020/2021 aasta rahva ja eluruumide loendus läbi viia senisest erinevalt, registripõhiselt. See tähendab, et inimeste küsitlemise asemel kasutatakse riiklikes andmekogudes olemas olevaid andmeid. Käesolevas töös keskendutakse loenduse osale, mis käsitleb leibkondade ja perekondade arvu ning koosseisu registripõhist määratlust. Magistritöö ei ole ette nähtud hindama ega kirjeldama registrite valmisolekut nõutud tunnuste moodustamiseks – selle kohta on võimalik lugeda registripõhise rahva ja eluruumide loenduse metoodika väljatöötamise lõpparuandest [1].

Töös eeldatakse, et registrite valmisolek on tagatud ning tulemuseni jõudmiseks otsitakse tehnilisi lahendusi. Kuna kaugeltki mitte kõik, mida rahvusvaheline loendusstandard leibkonna ja perekonna tunnuste kohta ette näeb, ei ole administratiivsest registrist otsesel kujul kättesaadav, siis tuleb leibkonnad ja perekonnad kokku panna kaudse info põhjal. Magistritöö eesmärk on pakkuda välja algoritmid, mida on võimalik testida käesoleva aasta sügisel Statistikaameti poolt läbiviidaval pilootloendusel ning mis sobivusel on hiljem kasutatavad ka registripõhisel rahva ja eluruumide loendusel.

Magistritöö on jagatud kolmeks peatükiks. Esimeses peatükis seletatakse lahti töö kesksed mõisted – leibkond ja perekond. Lähemalt peatutakse vabaabielupartnerite teemal kui perekondade määratlemise peamisel kitsaskohal. Samuti tutvustatakse leibkonna ja perekonnaga seonduvaid kohustuslikke tunnuseid ja nende rahvusvahelisele loendusstandardile vastavat jaotust.

Teises peatükis antakse ülevaade statistilisest metoodikast, mida töös rakendatakse. Tutvustatakse lineaarset regressioonimudelit ja näidatakse, et selle eeldused ei kehti binaarse uuritava tunnuse korral. Seejärel jõutakse logistilise regressioonimudelini. Kirjeldatakse logistilise regressioonimudeli parameetrite hindamist suurima tõepära meetodil, esitatakse

mõned levinumad mudeli sobivuse ja prognoosivõime näitajad ning tutvustatakse lühidalt diagnostiliste testide analüüsimeetodit.

Töö kolmas osa põhineb 2011. aasta rahva ja eluruumide loenduse andmetel. Kuna leibkonna defineerimiseks on kaks erinevat võimalust, millest üks on iseloomulik küsitluspõhisele ning teine registripõhisele loendusele, siis 2011. aasta rahva ja eluruumide loenduse andmete taustal uuritakse lähemalt leibkonna definitsiooni valikust tulenevaid erinevusi kui muutust, mida loendusmeetodi vahetamine endaga paratamatult kaasa toob. Eraldi alapunktides käsitletakse vabaabielupartnerite registripõhist määratlemist, mis on väga oluline ja ühtlasi kõige keerulisem osa perekonnatunnuste moodustamisel. Andmete peal proovitakse vabaabielupartnerite „avastamiseks" kahte erinevat võimalust – ühel juhul rakendatakse logistilist regressioonimudelit ning teisel juhul Soome Statistikaametis kasutavat metoodikat. Kolmandas osas esitatakse kaks plokkskeemi: esimene mitte abielus olevate lasteta partnerite registripõhiseks määratlemiseks ning teine leibkonnaliikmete jagamiseks perekondadesse eeldusel, et kõik partnerlusseosed on varasemalt kindlaks tehtud.

Magistritöö lisades on esitatud kahe logistilise regressioonimudeli SAS *proc logistic* väljatrükid (lisa 4, lisa 5), töös sisalduvate plokkskeemide programmikood (lisa 2) ning leibkonnaga ja perekonnaga seonduvate loendustunnuste jaotus (lisa 1) ja algoritm jaotuseni jõudmiseks (lisa 3). Viimast töös ei käsitleta, kuid see on esitatud magistritöö lisades oma praktilisele väärtuse tõttu pilootloenduse läbiviimisel.

Analüüsi läbiviimiseks on kasutatud tarkvarapaketti *SAS Enterprise Guide*, autori koostatud algoritmid on kirjutatud maatriksarvutuse protseduuri *IML* (*interactive matrix language*) abil ning toodud käesoleva töö lisades. Töö kolmandas peatükis olevad plokkskeemid on tehtud diagrammide koostamise tarkvaraga *MS Visio*. Tabelite tegemiseks on kasutatud nii *SAS*i kui ka tabelarvutus- ja tabeltöötlusprogrammi *MS Excel*, viimasega on tehtud ka illustreerivad joonised. Magistritöö on kirjutatud tekstitöötlusprogrammiga *MS Word*.

Kasutatud kirjanduse loetelu on esitatud allikale viitamise järjekorras, viitamiseks kasutatakse nurksulgi, kus number näitab viite järjekorda kasutatud kirjanduse loetelus.

1. LEIBKOND JA PEREKOND

Esimene peatükk põhineb registripõhise rahva ja eluruumide loenduse metoodika väljatöötamise lõpparuandel [1, lk 97-99] ning Euroopa Nõukogu ja Parlamendi Määrusel EÜ nr 763/2008 rahva ja eluruumide loenduste kohta [2] ja Euroopa Komisjoni Määrusel EÜ nr 1201/2009 [3], millega rakendatakse eelmist määrust seoses andmete ning nende jaotuste tehniliste spetsifikatsioonidega.

1.1. Leibkonna definitsioon. Registripõhise leibkonna erinevus küsitluspõhisest

Leibkond (*household*) on ühest või mitmest isikust koosnev statistiline üksus. Leibkonnad jagunevad tavaleibkondadeks (*private household*), institutsionaalseteks leibkondadeks ja püsielukohata isikute „leibkondadeks“. Rahvusvaheline loendusstandard kohustab leibkonnaliikmete vaheliste seoste määratlemist ja leibkonnaliikmete perekondadesse jagamist vaid tavaleibkondades, institutsionaalsetes leibkondades on see vabatahtlik. Magistritöö raames tegeletakse vaid tavaleibkondadega.

Euroopa Liidu liikmesriigid võivad tavaleibkondade määratlemisel valida kahe erineva variandi vahel.

Esimese variandi rakendamisel lähtutakse majapidamisüksuse mõistest, mille kohaselt moodustavad leibkonna isikud, kes elavad omaette elamuüksuses või selle mingis osas ning varustavad end toidu ja vajaduse korral muu eluks vajalikuga. Leibkonnaliikmed võivad oma sissetulekuid väiksemal või suuremal määral ühiselt kasutada. Isikute leibkonda kuulumise määravad selle variandi puhul kaks tingimust – ühine eluase (sama alaline elukoht) ja ühised majapidamiskulud. Selline leibkonna määratlemine on iseloomulik küsitluspõhisele loendusele, mistõttu edaspidi nimetatakse töös niisuguseid leibkondi küsitluspõhisteks leibkondadeks, ütluspõhisteks leibkondadeks või REL leibkondadeks.

Teise variandi puhul määratletakse leibkond vaid ühe tingimuse, ühise eluaseme põhjal. Leibkonna aadressipõhisel määratlemisel loetakse kõik samas elamuüksuses alaliselt elavad isikud ühte leibkonda, mistõttu langeb asustatud elamuüksuste arv ja neis elavate leibkondade arv kokku. Selle variandi kasutamine on tüüpiline registripõhises loenduses, kus puudub võimalus koguda ja arvestada teavet ühist eluaset jagavate isikute majandamissuhete kohta.

Käesolevas töös nimetatakse niisuguseid leibkondi eluruumipõhisteks leibkondadeks, aadressipõhisteks leibkondadeks või REGREL leibkondadeks.

1.2. Perekonna definitsioon

Tuumperekond (*family nucleus*) on määratud väga kitsas tähenduses – see on statistiline üksus, mis koosneb vähemalt kahest isikust, kes elavad alaliselt samas leibkonnas ning on üksteisega seotud abikaasadena, kooselupartneritena või vanema ja lapsena. Seega kujutab tuumperekond endast leibkonna „allüksust“, milleks võib olla lasteta paar, ühe või enama lapsega paar või ühe või enama lapsega üksikvanem. Laps (poeg/tütar) tähendab tuumperekonna kontekstis lihast, kasu- ja/või lapsendatud last (olenemata vanusest või perekonnaseisust), kes elab ühe või mõlema vanemaga samas leibkonnas ning kellel ei ole samas leibkonnas partnerit ega lapsi. Ajutise eeskoste all olevaid lapsi ja nende eestkostjaid tuumperekonnana ei käsitleta. Seega võib leibkond sisaldada ühte või mitut tuumperekonda, kuid leibkonda võivad kuuluda ka isikud, kes pole omavahel seotud abikaasa/elukaaslase ega lapse ja vanema suhetega. Käesolevas töös nimetatakse tuumperekonda ka lihtsalt perekonnaks.

1.2.1. Vabaabielupartnerid

Perekonna koosseisu registripõhise määramise peamiseks kitsaskohaks on vabaabielu-partnerite (*partners in a consensual union*) teema – küsitluspõhiselt lihtsasti kogutav informatsioon, mis üheski administratiivses registris ei kajastu. Määruse [3] kohaselt peetakse kaht isikut registreerimata kooselus elavateks partneriteks, kui nad

- kuuluvad ühte leibkonda ja
- neil on abielusarnane suhe ja
- nad ei ole üksteisega abielus (ega ela registreeritud kooselus).

Isiku kuuluvust leibkonda eeldusel, et leibkonna moodustavad kõik samas eluruumis elavad isikud, on võimalik registripõhiselt kindlaks teha tänu sellele, et rahvastikuregister sisaldab isikute aadresse. Leibkondade moodustamist lihtsustab asjaolu, et RR on liidestatud aadressiandmete süsteemiga (ADS), mille eesmärk on tagada aadressiobjektide ühene identifitseerimine nii nende asukohas kui ka erinevates andmekogudes ning muuta võrreldavaks erineval ajal ja eri põhimõtetel esitatud koha-aadressid [4]. Samuti on rahvastikuregistris kajastatud isiku ema ja isa isikukood ning seaduslik perekonnaseis ja

abikaasa isikukood. Vastust küsimusele, kuivõrd kahe isiku suhe sarnaneb abielule, ühestki andmekogust mõistagi teada ei saa, seetõttu on oluline vabaabielupartnerite nn kaudne tuvastamine.

Allpool on kirjeldatud, kuidas toimub vabaabielupartnerite registripõhine määratlemine Soome Statistikaametis. Seal on kasutatud registriandmeid kombineeritult küsitluspõhise uuringuga alates 1970. aastast. Kuni 1990. aastani olid vabas kooselus paarid Soome rahvastikuregistrist tuletatavad vaid ühiste laste olemasolul. Et tegelikkusele lähemat pilti saada, on alates 1992. aastast programmi redigeeritud nii, et pärast abikaasade seostamist eluruumipõhistes leibkondades identifitseerib programm vabaabielupartneriteks inimesed, kes elavad samas eluruumis, kellel ei ole abikaasat, kes on vähemalt 18-aastased, erinevast soost ja kelle vanusevahe jääb alla 16 aasta ning kelle puhul on tagatud sugulussidemete puudumine. Need reeglid ei kehti kooselupaaridele, kellel on ühiseid lapsi. Kui isiku jaoks on rohkem kui üks sobiv „partnerikandidaat“ loetakse vabaabielupartneriteks need, kelle vanusevahe on väiksem. Kui ühe isiku jaoks kehtivad toodud kriteeriumid rohkem kui nelja isikuga, siis programm kedagi partneriteks ei klassifitseeri. [5]

Käesoleva töö kolmandas peatükis proovitakse vabaabielupartnerite määratlemiseks kahte võimalikku lähenemist – esimesel juhul püütakse vabaabielupartnerid kindlaks teha logistilisest regressioonimudelist ning teisel juhul naaberriigiga sarnaseid kriteeriume kasutades.

1.3. Kohustuslikud loendustunnused

Leibkonna- ja perekonna koosseisu määravad kuus loendustunnust: *seisund leibkonnas*, *seisund perekonnas*, *tuumperekonna tüüp*, *tuumperekonna suurus*, *tavaleibkonna tüüp*, *tavaleibkonna suurus* [2]. Edasi on kirjeldatud eraldi iga tunnuse kohta Euroopa komisjoni määruses [3] antud rahvusvahelisele loendusstandardile vastavat jaotust ja selle detailsuse astet kui nõutud tulemust töö praktilisele osale.

Loendustunnus ***seisund leibkonnas*** liigitab loendusrahvastiku vastavalt leibkonna- ja peresuhetele kolmel erineva detailsusega tasandil. Esimesel, kõige üldisemal tasandil tehakse vahet tavaleibkonnas ja väljaspool tavaleibkonda elavate isikute vahel. Järgmisel, detailsemal tasandil eristatakse tavaleibkonnas elavate seas tuumperekonda kuuluvad ja tuumperekonda mittekuuluvad isikud. Kolmandal tasandil liigitatakse tuumperekondade liikmed omakorda

abikaasadeks, partneriteks registreeritud kooselus, partneriteks registreerimata kooselus, üksikvanemateks ja lasteks. Soovi korral võivad riigid eristada ka abikaasasid ja partnereid eri- ja samasoolistes kooseludes ning paaride ja üksikvanemate lapsi. Tavaleibkonnas elavate, kuid tuumperekonda mittekuulujate hulgas eristatakse üksielavaid ja mitmeliikmelises leibkonnas elavaid isikud. Viimaste hulgas võivad riigid vajaduse korral välja tuua sugulaste ja (üksnes) mittesugulastega leibkonna moodustavad isikud. Väljaspool tavaleibkonda elavate hulgas tuleb eristada asutusleibkonnas elavaid isikuid ja esmaselt kodutuid (isikud, kes ei ela varjupaigas ning kellel puudub igasugune peavari). Soovi korral võidakse ka asutusleibkondadesse kuuluvad isikud liigitada omakorda tuumperekonda kuulumise ja peresuhete alusel, tehes vahet partnerite, üksikvanemate ja laste vahel.

Loendustunnus ***seisund perekonnas*** kattub tunnuse *seisund leibkonnas* selle osaga, mis liigitab tavaleibkonnas elavaid tuumperekonna liikmeid. Tuumperekondade liikmed jagatakse abikaasadeks, partneriteks registreeritud kooselus, partneriteks registreerimata kooselus, üksikvanemateks ja lasteks. Soovi korral võivad riigid eristada abikaasasid ja partnereid sama- ja erisoolistes kooseludes ning paaride ja üksikvanemate lapsi. Leibkonnas võib olla ka isikuid, kes ei kuulu tuumperekonda, nemad liigitatakse tunnuse *seisund perekonnas* jaotusse „Ei kohaldata“.

Loendustunnus ***tavaleibkonna tüüp*** liigitab isikud vastavalt tavaleibkonna tüübile, milles isik elab. Kõige üldisemal tasandil liigitakse tavaleibkonnad leibkonda kuuluvate tuumperekondade arvu järgi kolme rühma: tavaleibkonnad, mis ei sisalda ühtegi tuumperekonda, ühepereleibkonnad ning kahe- või mitmepereleibkonnad. Esimesse rühma kuuluvate leibkondade hulgas eristatakse järgmisel tasandil ühe- ja mitmeliikmelisi leibkondi. Ühepereleibkondade puhul tehakse järgmisel tasandil vahet abielupaaride, registreeritud kooselus elavate paaride, registreerimata kooselus elavate paaride, üksikisade ja üksikemade leibkondade vahel. Veelgi detailsemal tasandil eristatakse abielu- ja kooselupaaridega ühepereleibkondade seas lasteta paare, alla 25-aastas(t)e lapsega(lastega) ja 25-aastas(t)e või vanema(te) lapsega(lastega) paare. Ka üksikvanemate leibkonnad alaliigitatakse vastavalt sellele, kas noorim leibkonnas elav laps on alla 25-aastane või vanem. Soovi korral võivad riigid eristada abielu- ja kooselupaaridega ühepereleibkondi täiendavalt selle järgi, kas tegemist on eri- või samasooliste paaridega. Kahe- või mitmepereleibkondades elavaid isikuid väiksematesse rühmadesse ei jaotata.

Loendustunnus ***tavaleibkonna suurus*** liigitab isikud vastavalt leibkonnaliikmete arvule. Loendusstandardi kohaselt tuuakse välja 1-, 2-, 3-, 4-, 5-, 6-, 7-, 8-, 9-, 10- ja 11+ liikmelises leibkonnas elavad isikud.

Loendustunnus ***tuumperekonna tüüp*** liigitab isikud vastavalt tuumperekonna tüübile, milles isik elab. Tuumperekondade puhul tehakse kõige üldisemal tasandil vahet abielupaaride, registreeritud kooselus elavate paaride, registreerimata kooselus elavate paaride, üksikisade ja üksikemade perekondade vahel. Detailsemal tasandil eristatakse abielu ja kooselupaaridega perekondade seas lasteta paare, alla 25-aastas(t)e lapsega(lastega) ja 25-aastas(t)e või vanema(te) lapsega(lastega) paare. Ka üksikvanemate perekonnad liigitatakse omakorda vastavalt sellele, kas noorim peres elav laps on alla 25-aastane või vanem. Soovi korral võivad riigid alaliigitada abielu- ja kooselupaaridega perekondi täiendavalt selle järgi, kas tegemist on eri- või samasooliste paaridega.

Loendustunnus ***tuumperekonna suurus*** liigitab isikud vastavalt perekonnaliikmete arvule. Loendusstandardi kohaselt tuuakse välja 1-, 2-, 3-, 4-, 5-, 6-, 7-, 8-, 9-, 10- ja 11+ liikmelises tuumperekonnas elavad isikud.

Käesolevas töös moodustatakse kõik kuus leibkonna ja perekonna koosseisu iseloomustavat loendustunnust vastavalt loendusstandardile, lähtudes sealjuures Eesti seadustest. Kuna Eesti kuulub (veel) nende riikide hulka, kus registreeritud partnerlust pole perekonnaõiguse mõttes kasutusse võetud [6], siis pole antud kooseluliiki, nagu ka samasooliste kooselu, töös käsitletud. Tavaleibkondadesse ja perekondadesse jagatakse vaid Eesti püsielanikud ning jättes kõrvale seadustest tulenevad erandid [6], tehakse seda kõige täpsemal detailsuse astmel, mida loendusstandard ette näeb. Väljaspool tavaleibkonda elavad isikud ei ole käesoleva töö huviorbiidis. Tunnuste *seisund leibkonnas*, *seisund perekonnas*, *tavaleibkonna tüüp*, *tuumperekonna tüüp, tavaleibkonna suurus* ja *tuumperekonna suurus* väärtuste klassifikatsioon on toodud lisas 1. SAS IML programmikood nimetatud tunnuste väärtuste määramiseks on esitatud lisas 3.

2. STATISTILINE METOODIKA

2.1. Lineaarne regressioon

Alapunkt 2.1. põhineb allikal [7, lk 8-10].

Lineaarne regressioonimudel esitatakse maatrikskujul järgmiselt:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

kus $\boldsymbol{Y}$ on (n×1) maatriks n vaadeldava väärtusega, $\boldsymbol{X}$ on (n×p) maatriks teadaolevate elementidega, $\boldsymbol{\beta}$ on mudeli tundmatute parameetrite (p×1) maatriks ning $\boldsymbol{\varepsilon}$ on (n×1) maatriks, mis sisaldab otseselt mittevaadeldavaid juhuslikke vigu. ($\boldsymbol{\beta} = \alpha, \beta_1, \dots, \beta_k$; $p = k$+1).

Lihtsuse mõttes vaadatakse käesolevas alapunktis niisugust mudelit, kus p=1. Klassikalise lineaarse mudeli korral tehakse eeldused:

(a) y on x-i lineaarne funktsioon, millele on liidetud juhuslik viga
$y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1, \dots, n$;

(b) juhuslikud vead on keskväärtusega 0, st x ja ε ei ole korreleeritud
$E(\varepsilon_i) = 0$;

(c) juhuslikud vead on konstantse hajuvusega
$D(\varepsilon_i) = \sigma^2$;

(d) juhuslikud vead on sõltumatud
$COV(\varepsilon_i, \varepsilon_j)$=0, $i \neq j$;

(e) juhuslikud vead on normaaljaotusega
$\varepsilon_i \sim N(0, \sigma^2)$.

Regressioonanalüüsi eesmärgiks on hinnata tundmatute parameetrite vektorit $\boldsymbol{\beta}$. Reeglina kasutatakse selleks vähimruutude meetodit, st mudeli parameetrite väärtused tuleb valida sellised, et erinevused tegelikult mõõdetud sõltuva tunnuse väärtuste ja mudeli järgi prognoositud väärtuste vahel oleksid minimaalsed. Ülal toodud eelduste mittekehtimisel saadakse üldiselt ebaefektiivne mudel – st seosed, mis tegelikkuses kehtivad, võivad mudelis tulla mitteolulised ja vastupidi, tegelikult mitteolulised seosed võivad osutuda mudelis olulisteks. [8]

Kui uuritaval tunnusel on ainult kaks võimalikku väärtust, näiteks 0 ja 1, siis on endiselt põhjendatud arvata, et eeldused (a), (b) ja (d) on täidetud. Sellest, et eeldused (a) ja (b) kehtivad, saab aga lihtsasti järeldada, et (c) ja (e) ei kehti.

Oletame, et (a) kehtib ning y on binaarne tunnus võimalike väärtustega 0 ja 1, siis

$$y_i = 1 \overset{(a)}{\Rightarrow} \varepsilon_i = 1 - \alpha - \beta x_i \,,$$
$$y_i = 0 \overset{(a)}{\Rightarrow} \varepsilon_i = -\alpha - \beta x_i.$$

Järelikult ε_i jaoks on ainult kaks erinevat väärtust, mistõttu on võimatu, et ε_i on normaaljaotusega (vastasel juhul peaks ε_i väärtuste hulk olema lõpmatu ning ülalt ja alt tõkestamata). Seega, kui (a) kehtib ja y on binaarne tunnus, siis (e) ei saa kehtida.

Keskväärtuse definitsiooni põhjal:

$$E(y_i) = 1 \cdot P(y_i = 1) + 0 \cdot P(y_i = 0).$$

Kui p_i on tõenäosus, et y_i võtab väärtuse 1, siis kehtib

$$E(y_i) = p_i. \tag{2}$$

Üldiselt iga 0/1-tunnuse jaoks on selle keskväärtuseks lihtsalt tõenäosus, et tunnus võrdub ühega. Eeldused (a) ja (b) ütlevad aga midagi muud. Võttes võrduse (a) mõlemast poolest keskväärtuse, saame:

$$E(y_i) = E(\alpha + \beta x_i + \varepsilon_i) = E(\alpha) + E(\beta x_i) + E(\varepsilon_i) \,,$$

$$E(y_i) = \alpha + E(\beta x_i). \tag{3}$$

$$\overset{(2),(3)}{\Longrightarrow} \; p_i = \alpha + E(\beta x_i). \tag{4}$$

Seost (4) kutsutakse ka lineaarseks tõenäosusmudeliks (*linear probability model*). Seega y=1 on x-i lineaarne funktsioon. Regressioonikordajatel on siin otsene tähendus: x väärtuse ühe ühikuline muutus toob endaga kaasa β ühikulise muutuse tõenäosuses, et y=1.

Kuna x-i käsitletakse fikseerituna, siis $D(\varepsilon_i) = D(y_i)$. Üldiselt on binaarse tunnuse dispersioon esitatav kujul

$$D(y_i) = p_i \cdot (1 - p_i).$$

Seega saame:

$$D(\varepsilon_i) = p_i \cdot (1 - p_i) = (\alpha + \beta x_i) \cdot (1 - \alpha - \beta x_i).$$

Vea dispersioon on maksimaalne, kui p_i=0,5 ja läheb väiksemaks, kui p_i on nulli või ühe lähedal, mistõttu on eeldus (c) rikutud.

Eelnev on sissejuhatus logistilise regressiooni alapunkti.

2.2. Logistiline regressioon

Alapunkt 2.2. põhineb allikal [7, lk 13-14].

Peamine probleem lineaarse tõenäosusmudeli (4) puhul on see, et lineaarne funktsioon on oma olemuselt tõkestamata, kuid tõenäosused on tõkestatud nulli ja ühega. Tuleb kasutada niisugust seosefunktsiooni, mille rakendamisel ei oleks tõenäosused enam tõkestatud – üheks võimaluseks on logaritmiline tõepärafunktsioon. Esitades tõenäosuse tõepärasuhtena, oleme kõrvaldanud ülemise tõkke ning võttes saadud suhtest logaritmi, oleme kõrvaldanud ka alumise tõkke. Seades saadud tulemuse vastavusse seletavate tunnuste lineaarse kombinatsiooniga, olemegi saanud logistilise regressioonimudeli (*logit* mudeli). Logistiline regressioonimudel k seletatava tunnusega avaldub kujul

$$logit(y_i) = ln\left[\frac{p_i}{1 - p_i}\right] = \alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}, \tag{5}$$

kus $p_i = P(y_i = 1)$. Parameetervektori $\boldsymbol{\beta}$ hindamiseks kasutatakse suurima tõepära meetodit. Võrduse (5) vasak pool kujutab endast logaritmilist tõepärasuhet, millest p_i avaldades saame:

$$p_i = \frac{exp(\alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + exp(\alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}, \tag{6}$$

$$p_i = \frac{1}{1 + exp(-\alpha - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})}. \tag{6'}$$

Sellega on tagatud, et ükskõik, millega $\boldsymbol{\beta}$ ja $\boldsymbol{x}$ asendada, siis p_i jääb alati nulli ja ühe vahele.

2.2.1. Parameetrite hindamine suurima tõepära meetodil

Alapunkt 2.2.1. põhineb allikal [7, lk 36-39].

Olgu meil k statistiliselt sõltumatut tunnust n indiviidi kohta (i=1, …, n). Iga i-nda indiviidi kohta on andmestikus juhusliku suuruse y väärtus y_i, mis võib olla kas 0 või 1, ning seletatavate tunnuste vektor $\boldsymbol{x}_i$=$[1, x_{i1}, \dots, x_{ik}]^T$. Olgu p_i tõenäosus, et y_i=1 ning eeldame, et andmed on genereeritud logistilise mudeli poolt, mille kohaselt

$$p_i = \frac{1}{1 + exp(-\boldsymbol{\beta x_i})}\,. \tag{7}$$

Konstrueerime nüüd tõepärafunktsiooni $L = P\ (y_1,\ y_2, \dots, y_n)$. Kuna eeldame, et vaatlused on sõltumatud, siis saame kõikide y_i jaoks kirjutada tõepärafunktsiooni kujul

$$L = P\ (y_1) \cdot P(y_2) \cdot \dots \cdot P(y_n) = \prod_{i=1}^{n} P\ (y_i). \tag{8}$$

Definitsiooni järgi $P(y_i = 1) = p_i$ ja $P(y_i = 0) = 1 - p_i$. Sellest järeldub, et

$$P\ (y_i) = p_i^{y_i} \cdot (1 - p_i)^{1-y_i}\,. \tag{9}$$

Arvestades (8) ja (9), saame:

$$L = \prod_{i=1}^{n} p_i^{y_i} \cdot (1 - p_i)^{1-y_i} \ = \ \prod_{i=1}^{n} (\frac{p_i}{1 - p_i})^{y_i} \cdot (1 - p_i). \tag{10}$$

Tehniliselt on lihtsam kasutada logaritmilist funktsiooni ning kuna logaritm on kasvav funktsioon, siis mis iganes maksimeerib logaritmi, maksimeerib ka esialgse funktsiooni. Võttes võrduse (10) mõlemalt poolest logaritmi, saame:

$$ln\, L = \sum_{i=1}^{n} y_i \cdot ln\ (\frac{p_i}{1 - p_i}) + \sum_{i=1}^{n} \ln(1 - p_i). \tag{11}$$

Asendades (7) võrrandisse (11), saame:

$$ln\, L = \sum_{i=1}^{n} \boldsymbol{\beta x}_i y_i - \sum_{i=1}^{n} ln\ (1 + \exp\ (\boldsymbol{\beta x}_i)). \tag{12}$$

Sellega oleme viinud logaritmilise tõepärafunktsiooni nii lihtsale kujule kui võimalik. Järgmine samm on leida selline $\boldsymbol{\beta}$, mille korral saavutab logaritmiline tõepärafunktsioon (12) maksimaalse väärtuse. Tõepärafunktsiooni maksimeerimiseks on kõige levinum võimalus võtta sellest tuletis $\boldsymbol{\beta}$ järgi, võrdsustada tuletis nulliga ning avaldada sealt $\boldsymbol{\beta}$.

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \boldsymbol{x}_i y_i - \sum_{i=1}^{n} \boldsymbol{x}_i \left(1 + \exp\left(-\boldsymbol{\beta}\boldsymbol{x}_i\right)\right)^{-1} = \sum_{i=1}^{n} \boldsymbol{x}_i y_i - \sum_{i=1}^{n} \boldsymbol{x}_i \hat{y}_i = 0, \qquad (13)$$

kus

$$\hat{y}_i = \frac{1}{1 + \exp\left(-\boldsymbol{\beta}\boldsymbol{x}_i\right)}.$$

Kuna $\boldsymbol{x}_i$ on vektor, siis (13) kujutab endast tegelikult võrduste süsteemi, mis koosneb $k + 1$ võrdusest, üks iga $\boldsymbol{\beta}$ elemendi jaoks. Võrdus (13) lahendatakse iteratiivselt – meetodeid on erinevaid, kuid erinevus on vaid koondumise kiiruses, tundlikkuses algväärtuste suhtes ja arvutuslikus raskusastmes; tulemus on igal juhul sama. Kõige laialdasemalt kasutatav meetod on Newton-Raphsoni algoritm, mille lahenduskäik on toodud allpool.

Olgu $\boldsymbol{U}(\boldsymbol{\beta})$ logaritmilise tõepärafunktsiooni esimene tuletis $\boldsymbol{\beta}$ suhtes ja $\boldsymbol{I}(\boldsymbol{\beta})$ olgu logaritmilise tõepärafunktsiooni teiste osatuletiste maatriks, siis

$$\boldsymbol{U}(\boldsymbol{\beta}) = \frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \boldsymbol{x}_i y_i \sum_{i=1}^{n} \boldsymbol{x}_i \hat{y}_i = 0, \qquad (14)$$

$$\boldsymbol{I}(\boldsymbol{\beta}) = \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta}} = -\sum_{i=1}^{n} \boldsymbol{x}_i \, \boldsymbol{x}_i{}' \hat{y}_i (1 - \hat{y}_i). \qquad (15)$$

$\boldsymbol{U}(\boldsymbol{\beta})$ nimetatakse ka gradiendiks või skoorifunktsiooniks ning $\boldsymbol{I}(\boldsymbol{\beta})$ Hesse maatriksiks. Newton-Raphsoni algoritm näeb välja järgmine:

$$\boldsymbol{\beta}_{j+1} = \boldsymbol{\beta}_j - \boldsymbol{I}^{-1}(\boldsymbol{\beta}_j)\boldsymbol{U}(\boldsymbol{\beta}_j), \qquad (16)$$

kus $\boldsymbol{I}^{-1}$ on $\boldsymbol{I}$ pöördmaatriks. Praktikas vajame algväärtusi $\boldsymbol{\beta}_0$. Algväärtused asendatakse võrduse (16) paremale poole, misjärel saadakse esimese iteratsioonisammu tulemus $\boldsymbol{\beta}_1$. Saadud väärtused asendatakse tagasi võrrandi (16) paremale poole, esimesed ja teised tuletised arvutatakse uuesti ning saadakse $\boldsymbol{\beta}_2$. Protsessi jätkatakse kuni parameetrite

hinnangute erinevus kahe iteratsioonisammu vahel on väiksem kui teatud etteantud kriteerium.

Kui parameetri hinnangu $\boldsymbol{\beta}_j$ absoluutväärtus on 0,1 või väiksem, siis on vaikimisi koondumise kriteeriumiks $\left|\boldsymbol{\beta}_{j+1} - \boldsymbol{\beta}_j\right| < 0{,}0001$; kui parameetri hinnangu $\boldsymbol{\beta}_j$ absoluutväärtus on suurem kui 0,1, loetakse vaikimisi koondumise kriteeriumiks $\left|\frac{\boldsymbol{\beta}_{j+1} - \boldsymbol{\beta}_j}{\boldsymbol{\beta}_j}\right| < 0{,}001$.

2.2.2. Mudeli sobivuse testid

Alapunkt 2.2.2. põhineb peamiselt allikatel [9] ja [10].

Vastamaks küsimusele „Kas logistiline regressioonimudel sobib andmetega?“ on kaks erinevat lähenemisviisi. Üks võimalus on hinnata mudeli prognoosivõimet, st arvutada statistik, mis mõõdab, kui hästi saab prognoosida sõltuva tunnuse väärtust sõltumatute tunnuste kaudu. Kõige levinum on determinatsioonikordaja (R^2 statistik), mille väärtus varieerub nulli ja ühe vahel ning mille suurem väärtus näitab mudeli paremat prognoosivõimet. R^2 arvutamiseks on mitmeid erinevaid võimalusi ja üheselt ei ole võimalik öelda, milline neist on parim.

Olgu logistilise regressioonimudeli parameetrid hinnatud suurima tõepära meetodil ning olgu $L(0)$ tõepärafunktsioon juhul, kui mudelis ei ole ühtegi hinnatavat parameetrit (ainult vabaliikmega mudel), $L(\hat{\beta})$ hinnatud mudeli tõepärafunktsioon ja n valimimaht. Allpool on esitatud kolm erinevat R^2 statistikut.

McFadden' i R^2:

$$R^2{}_{McF} = 1 - \left(\frac{ln\, L(\hat{\beta})}{ln\, L(0)}\right).$$

Cox and Snell' i R^2:

$$R^2{}_{C\&S} = 1 - \left(\frac{L(0)}{L(\hat{\beta})}\right)^{\frac{2}{n}}.$$

Märkus: $R^2{}_{C\&S}$ ülemine tõke avaldub kujul $1 - (L(0))^{\frac{2}{n}}$, mistõttu on $R^2{}_{C\&S}$ maksimaalne väärtus alati väiksem kui üks. Probleemi lahendamiseks pakkus Nagelkerke välja nn parandatud R^2 statistiku, mille maksimaalne väärtus on 1.

Nagelkerke R^2:

$$R^2{}_N = \frac{1 - \left(\frac{L(0)}{L(\hat{\beta})}\right)^{\frac{2}{n}}}{1 - (L(0))^{\frac{2}{n}}}.$$

Teine võimalus mudeli sobivuse kindlakstegemiseks on kasutada mudeli sobivuse teste (*Goodness of fit* (*GOF*) testid). Allpool esitatakse kolme erineva *GOF* testi teststatistikud (Hälbimus, Pearson'i χ^2 ja Hosmer-Lemeshow'i teststatistik). Järgnevas eeldatakse, et uuritava tunnuse Y võimalikud väärtused on 0 ja 1.

Hälbimus (*Deviance*):

$$D = 2\sum_j O_j ln\frac{O_j}{E_j},$$

kus j tähistab lahtrit kahemõõtmelises tabelis, mille ühes reas on sõltumatute tunnuste kõikvõimalike kombinatsioonide sagedused (profiilid) Y=0 jaoks ja teises reas Y=1 jaoks. O_j on j-nda lahtri vaadeldud sagedus ja E_j mudeli põhjal leitud oodatav sagedus.

Pearson'i χ^2:

$$\chi^2 = \sum_j \frac{(O_j - E_j)^2}{E_j},$$

kus j, O_j ja E_j on sama tähendusega nagu hälbimuse arvutamisel.

Mõlemad statistikud on asümptootiliselt χ^2-jaotusega, mille vabadusastmete arvuks on profiilide arv miinus mudeliga hinnatavate parameetrite arv.

Hälbimus ja Pearson'i χ^2 statistik sobivad hästi hindama diskreetsete tunnustega mudeli sobivust; kui mudelis on ka pidevaid tunnuseid, kasutatakse mudeli sobivuse hindamiseks Hosmer-Lemeshow'i testi. See baseerub χ^2 statistikul ja andmete grupeerimisel olenevalt Y=1 jaoks hinnatud tõenäosusest – tõenäosused järjestatakse kasvavalt ning jagatakse ligikaudu võrdse suurusega G gruppi (vaikimisi G=10).

Esimene grupp moodustatakse 10% valimist, kelle puhul eeldatavad tõenäosused on kõige madalamad, teine grupp järgmisest 10% jne.

Kui mudelis on ka pidevaid tunnuseid, siis võibki igal vaatlusel olla erinev tõenäosus, mistõttu hinnatud tõenäosused võivad varieeruda ka grupisiseselt. Et arvutada oodatav sagedus Y=1 jaoks, võtab Hosmer-Lemeshow'i test arvesse hinnatud tõenäosuste grupikeskmise ja korrutab selle vaatluste arvuga grupis. Sama tehakse Y=0 jaoks ning seejärel arvutatakse igas grupis Pearson'i χ^2 statistik.

Hosmer-Lemeshow'i teststatistik esitatakse kujul:

$$H = \sum_{k=0}^{1} \sum_{g=1}^{\mathrm{G}} \frac{(o_{kg} - e_{kg})^2}{e_{kg}} ,$$

kus o_{0g} ja o_{1g} tähistavad vastavalt Y=0 ja Y=1 tegelikku sagedust g-ndas grupis ning e_{0g} ja e_{1g} vastavaid sagedusi mudelist hinnatuna.

Hosmer ja Lemeshow näitasid simulatsioonide abil, et eeldusel $p + 1 < G$, on teststatistik asümptootiliselt χ^2 jaotusega vabadusastmete arvuga G-2. Samas, viimasel ajal on arutletud selle üle, et testi tulemused sõltuvad tugevalt valitud gruppide arvust G [11].

On oluline teada, et mudeli prognoosivõime ja mudeli headuse näitajad testivad erinevaid asju ja seetõttu on põhjendatud, kui mudelil on näiteks kõrge R^2, kuid *GOF* testi puhul saadakse väike p-väärtus, või vastupidi. *GOF* testid ei näita, kuivõrd mudeli prognoosid vastavad tegelikkusele, vaid nende abil vastatakse küsimusele, kas mudelit keerulisemaks muutes (seosefunktsiooni vahetamine, koosmõjude lisamine), oleks võimalik saada veelgi paremaid prognoose.

2.2.3. Diagnostiliste testide analüüs

Alapunkt 2.2.3. põhineb allikal [12, lk 17-20]. Diagnostiliste testide analüüsimeetodid pärinevad meditsiinist, kus neid kasutatakse sõeluuringutes uurimaks teatud haiguse esinemist. Käesolevas alapunktis on sõnastus viidud näite varale, mis lähtub magistritöös toodud probleemipüstitusest – partnerite määratlemisest. Sellest lähtuvalt defineeritakse ka diagnostilistes testides kasutatavad mõisted *tundlikkus*, *spetsiifilisus*, *positiivne prognoosiväärtus* ja *negatiivne prognoosiväärtus*.

Magistritöös uuritakse logistilist regressioonimudelit, kus sõltuv tunnus *Y* koosneb mehe ja naise seosepaarist. *Y* on binaarne tunnus väärtusega 1, kui mees ja naine on abikaasad või vabaabielupartnerid, ning väärtusega 0 kõikide teiste suhtetüüpide korral. Tabeli 1 read tähistavad andmetest teadaolevat (*observed*) suhtetüüpi uuritavate isikute vahel ning veerud mudelist prognoositud (*expected*) suhtetüüpi.

Tabel 1. Diagnostilised testid

		Y prognoositud (E)	
		Partnerlusseos olemas (+)	Partnerlusseos puudub (-)
Y tegelik (O)	Partnerlusseos olemas (+)	a	b
	Partnerlusseos puudub (-)	c	d

Tundlikkus defineeritakse kui tõenäosus, et isikud määrati mudeli põhjal partneriteks juhul, kui nad ka tegelikult seda on. Teisisõnu, tundlikkus mõõdab, kuivõrd tõenäoline on see, et mudel tuvastab partnerite olemasolu partnerlusseose tegelikul olemasolul. Tundlikkus esitatakse valemiga:

$$P(E^+|O^+) = \frac{a}{a+b}.$$

Spetsiifilisus defineeritakse kui tõenäosus, et isikuid ei määratud mudeli põhjal partneriteks juhul, kui nad ka tegelikult partnerid ei ole. Spetsiifilisus esitatakse valemiga:

$$P(E^-|O^-) = \frac{d}{c+d}.$$

Ideaalkujul peaksid nii tundlikkus kui spetsiifilisus olema kõrged, kuid mõnikord tuleb teha kompromisse; näiteks kõrge tundlikkusega võib kaasneda madal spetsiifilisus, või vastupidi.

Valepositiivne prognoosiväärtus ilmneb, kui mudeli põhjal määratakse isikud partneriteks, kuid reaalselt nad partnerid ei ole. Valepositiivsuse määr esitatakse kujul:

$$P(O^-|E^+) = \frac{c}{a+c}.$$

Valenegatiivne prognoosiväärtus ilmneb, kui mudeli põhjal isikuid partneriteks ei määrata, kuid reaalselt on tegu partneriga. Valenegatiivsuse määr esitatakse kujul:

$$P(O^+|E^-) = \frac{b}{b+d}.$$

Ideaalkujul soovitakse näha nii c kui b väärtust nullina, kuid mudelis, kus andmemaht on suur, on seda üldiselt võimatu teostada.

3. PRAKTILINE TÖÖ

3.1. Ülevaade rahva ja eluruumide loenduse andmetest

Praktilise osa aluseks on võetud 2011. aastal toimunud rahva ja eluruumide loenduse (REL11) anonüümitud andmed. REL11 andmestik koosneb mitmest erinevast lähteandmebaasist (LAB). Allpool antakse ülevaade nendest tunnustest, mida magistritöös kasutatakse.

Tabel 2. Ülevaade kasutatavates andmetest; RELi lähteandmebaas F_ISIK

Lühinimetus	Seletus (võimalikud väärtused)
ISIK_ID	RELi sisene isiku unikaalne identifikaator.
ANONYM_ISIKUKOOD	Anonüümitud isikukood, unikaalne iga isiku jaoks.
SUGU	Isiku sugu (1= mees, 2=naine).
VANUS	Isiku vanus loendusmomendil, täisaastates.
SYNNIAEG	Isiku sünniaeg.
STAATUS	STAATUS võimaldab eristada püsielanikud, ajutiselt riigis viibivad ja riigist püsivalt lahkunud isikud.
LEKO_ID	Leibkonna identifikaator, unikaalne iga leibkonna jaoks.
ELRU_ID	Eluruumi identifikaator, unikaalne iga eluruumi jaoks.
EMA_ID	Isiku ema RELi sisene identifikaator.
ISA_ID	Isiku isa RELi sisene identifikaator.
PERE_ID	Perekonna identifikaator, unikaalne iga perekonna jaoks.
PERES_ROLL	Püsielaniku roll perekonnas(1=abikaasa, 2=vabaabielupartner, 3=üksikvanem, 4=laps). Arvutatud suhtetüüpide järgi.
SEISUND_LEIBKONNAS	Tavaleibkondade leibkonnaliikmete jaotus: 01 abikaasa ilma lasteta perekonnatuumas, 02 abikaasa lastega perekonnatuumas, 03 vabaabielupartner ilma lasteta perekonnatuumas, 04 vabaabielupartner lastega perekonnatuumas, 05 üksikvanem, 06 laps, elab koos mõlema abielus vanemaga, 07 laps, elab koos mõlema vabaabielus vanemaga, 08 laps, elab koos üksikvanemaga, 09 üksi elav 1-liikmelise leibkonna liige, 10 elab koos perekonnatuuma liikmetega, 11 elab koos isikutega, kes ei ole perekonnatuuma liikmed.

Tabel 3. Ülevaade kasutatavates andmetest; RELi lähteandmebaas F_LEIBKONNASUHE

Lühinimetus	Seletus (võimalikud väärtused)
SEOS_ID	Unikaalne identifikaator selles lähteandmebaasis.
LEKO_ID	Leibkonna identifikaator, unikaalne iga leibkonna jaoks.
OSAPOOL1_ID	Iga leibkonna sees on moodustatud selle liikmete vahel kõikvõimalikud paarid. OSAPOOL1_ID ja OSAPOOL2_ID on vastavad RELi sisesed isiku identifikaatorid.
OSAPOOL2_ID	
SEOS	Kahe samas leibkonnas elava püsielaniku vaheline suhe. OSAPOOL1 on OSAPOOLEL2-le (1=abikaasa; 2= elukaaslane; 3= laps (sh lapsendatud); 4=abikaasa või elukaaslase laps; 5=ema/isa (sh lapsendanu); 6=ema/isa abikaasa või elukaaslane; 7= õde/vend (sh poolõde/ poolvend või vanema abikaasa/elukaaslase laps); 8=vanavanem (sh vanavanema abikaasa/elukaaslane); 9=lapselaps (sh abikaasa/ elukaaslase lapselaps); 10=muu sugulane (sh abikaasa/ elukaaslase sugulane); 11=mittesugulane; -2=teadmata).

Tabel 4. Ülevaade kasutatavates andmetest; RELi lähteandmebaas F_LEIBKOND

Lühinimetus	Seletus (võimalikud väärtused)
LEKO_ID	Leibkonna identifikaator, unikaalne iga leibkonna jaoks.
LEKO_LIIK	LEKO_LIIK võimaldab eristada tavaleibkonnad, institutsionaalsed leibkonnad ja püsielukohata isikute "leibkonnad".

Kirjeldatud lähteandmebaaside linkimiseks kasutatakse järgmisi seoseid:

F_ISIK.ISIK_ID=F_LEIBKONNASUHE.OSAPOOL1_ID,
F_ISIK.ISIK_ID=F_LEIBKONNASUHE.OSAPOOL2_ID,
F_ISIK.LEKO_ID=F_LEIBKOND.LEKO_ID=F_LEIBKONNASUHE.LEKO_ID.

3.2. Leibkonna suurus sõltuvalt leibkonna definitsioonist

Registripõhine leibkond koosneb ühest kuni mitmest majapidamisüksuse mõiste põhisest leibkonnast. Eelmisel rahva ja eluruumide loenduse (REL11) käigus loendati 1 279 328 püsielanikku, kes jagunesid 599 832 RELi tavaleibkonnaks. Võrdluseks, kui kõik samas elamuüksuses elavad isikud oleksid ühte leibkonda loetud, siis tekkinuks samadest isikutest

557 095 tavaleibkonda – seda on 42 737 leibkonna võrra vähem. Täpsem võrdlus leibkonna suuruse lõikes on toodud tabelis 5.

Tabel 5. Tavaleibkondade arv leibkonna suuruse lõikes; REL11 andmetel

		Aadressipõhine leibkond (REGREL leibkond)		Majapidamisüksuse mõiste põhine leibkond (REL leibkond)	
		N	%	N	%
Leibkonna suurus	1	191 234	34,33	239 587	39,94
	2	168 097	30,17	173 345	28,9
	3	96 285	17,28	95 129	15,86
	4	65 110	11,69	63 244	10,54
	5	23 723	4,26	20 481	3,41
	6	8000	1,44	5566	0,93
	7	2828	0,51	1622	0,27
	8	1017	0,18	501	0,08
	9	423	0,08	198	0,03
	10	187	0,03	79	0,01
	11+	191	0,03	80	0,01
Kokku		**557 095**	**100**	**599 832**	**100**

Aadressipõhise leibkonna definitsiooni järgi on ühe- ja kaheliikmelisi leibkondi vähem kui majapidamisüksuse mõiste põhise definitsiooni järgi. Suuremate leibkondadega on olukord vastupidine – neid on aadressipõhiselt rohkem. Saadud tulemus on igati ootuspärane.

Kaheliikmeliste ja suuremate leibkondade osakaal on REGREL leibkondade hulgas suurem kui REL leibkondade hulgas. Ligi 40% kõigist REL leibkondadest on üheliikmelised, samas kui aadressipõhistest leibkondadest on üheliikmelisi alla 35%. Edasi pakub huvi, kas on võimalik välja selgitada, mida leibkonna definitsiooni muutus endaga täpsemalt kaasa toob. Piltlikult öeldes, kas isikud, kes on loendusel öelnud, et moodustavad leibkonna üksinda, aga on aadressipõhiselt mitmeliikmelise leibkonna liikmed, elavad pigem kahe-kolmekesi või kaheksa-üheksakesi? Kas sellega, et leibkonna definitsiooni muutus toob kaasa leibkonna koosseisu muutuse, on seotud pigem mehed või naised, noored või vanad?

Isikute jaotusest leibkonniti sõltuvalt leibkonna definitsioonist annavad ülevaate tabelid 6 ja 7.

Tabel 6. Isikute jaotus leibkonniti sõltuvalt leibkonna definitsioonist

		Majapidamisüksuse mõiste põhise leibkonna (REL leibkond) suurus																**Kokku**
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	
Aadressipõhise leibkonna (REGREL leibkond) suurus	**1**	191 234	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**191 234**
	2	23 606	312 588	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**336 194**
	3	12 060	16 464	260 331	.	.	.	.	.	.	.	.	.	.	.	.	.	**288 855**
	4	6497	7988	12 603	233 352	.	.	.	.	.	.	.	.	.	.	.	.	**260 440**
	5	3490	4520	6228	9332	95 045	.	.	.	.	.	.	.	.	.	.	.	**118 615**
	6	1478	3100	3099	5588	3595	31 140	.	.	.	.	.	.	.	.	.	.	**48 000**
	7	612	1200	1800	2252	2060	1260	10 612	.	.	.	.	.	.	.	.	.	**19 796**
	8	259	442	750	1304	730	522	441	3688	.	.	.	.	.	.	.	.	**8136**
	9	95	166	285	624	540	168	91	200	1638	.	.	.	.	.	.	.	**3807**
	10	56	84	141	284	245	174	70	64	72	680	.	.	.	.	.	.	**1870**
	11	47	60	75	140	95	78	63	24	18	60	363	.	.	.	.	.	**1023**
	12	24	30	27	56	65	18	49	8	18	20	33	264	.	.	.	.	**612**
	13	4	8	12	4	10	12	21	8	9	30	.	12	143	.	.	.	**273**
	14	11	12	12	4	10	6	.	8	9	.	.	.	26	28	.	.	**126**
	15	1	10	3	4	.	.	7	8	.	.	.	12	.	.	.	.	**45**
	16	2	4	3	12	.	.	.	.	.	.	11	.	.	.	.	32	**64**
	17	5	6	9	8	10	.	.	.	18	.	.	.	.	14	15	.	**85**
	20	1	6	3	12	.	18	.	.	.	.	.	.	.	.	.	.	**40**
	21	21	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**21**
	33	30	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	**33**
	59	54	2	3	.	.	.	.	.	.	.	.	.	.	.	.	.	**59**
Kokku		**239 587**	**346 690**	**285 387**	**252 976**	**102 405**	**33 396**	**11 354**	**4008**	**1782**	**790**	**407**	**288**	**169**	**42**	**15**	**32**	**1 279 328**

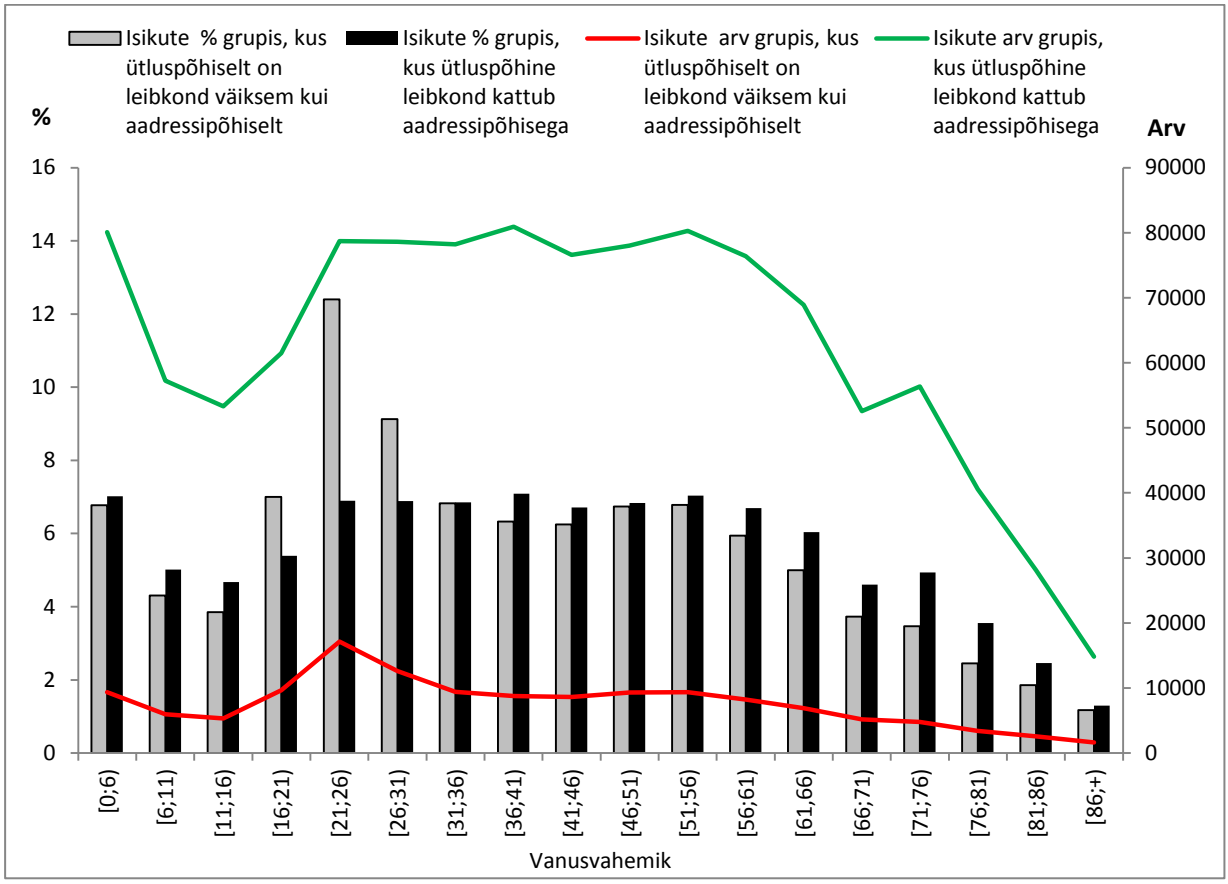
Tabel 7. Isikute jaotus leibkonniti sõltuvalt leibkonna definitsioonist; osakaalud

Aadressipõhise leibkonna (REGREL leibkond) suurus	Majapidamisüksuse mõiste põhise leibkonna (REL leibkond) suurus															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
1	79,82	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2	9,85	90,16	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	5,03	4,75	91,22	.	.	.	.	.	.	.	.	.	.	.	.	.
4	2,71	2,3	4,42	92,24	.	.	.	.	.	.	.	.	.	.	.	.
5	1,46	1,3	2,18	3,69	92,81	.	.	.	.	.	.	.	.	.	.	.
6	0,62	0,89	1,09	2,21	3,51	93,24	.	.	.	.	.	.	.	.	.	.
7	0,26	0,35	0,63	0,89	2,01	3,77	93,46	.	.	.	.	.	.	.	.	.
8	0,11	0,13	0,26	0,52	0,71	1,56	3,88	92	.	.	.	.	.	.	.	.
9	0,04	0,05	0,1	0,25	0,53	0,5	0,8	4,99	91,9	.	.	.	.	.	.	.
10	0,02	0,02	0,05	0,11	0,24	0,52	0,62	1,6	4,04	86,1	.	.	.	.	.	.
11	0,02	0,02	0,03	0,06	0,09	0,23	0,55	0,6	1,01	7,59	89,2	.	.	.	.	.
12	0,01	0,01	0,01	0,02	0,06	0,05	0,43	0,2	1,01	2,53	8,11	91,7	.	.	.	.
13	0	0	0	0	0,01	0,04	0,18	0,2	0,51	3,8	.	4,17	84,6	.	.	.
14	0	0	0	0	0,01	0,02	.	0,2	0,51	.	.	.	15,4	66,7	.	.
15	0	0	0	0	.	.	0,06	0,2	.	.	.	4,17	.	.	.	.
16	0	0	0	0	.	.	.	.	.	.	2,7	.	.	.	.	100
17	0	0	0	0	0,01	.	.	.	1,01	.	.	.	.	33,3	100	.
20	0	0	0	0	.	0,05	.	.	.	.	.	.	.	.	.	.
21	0,01	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
33	0,01	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.
59	0,02	0	0	.	.	.	.	.	.	.	.	.	.	.	.	.

Tabeli lugemine: 79,82% püsielanikest, kes elavad üheliikmelises REL leibkonnas, elavad ka üheliikmelises REGREL leibkonnas; 90,16% inimestest, kes elavad kaheliikmelises REL leibkonnas, elavad kaheliikmelises REGREL leibkonnas, 4,75% kolmeliikmelises REGREL leibkonnas jne.

Kui aadressipõhiselt oleks tegu üheliikmelise leibkonnaga, siis on see leibkond üheliikmeline ka RELi mõistes, ent mida suurem on registripõhine leibkond, seda rohkem on variante, kui mitmeks küsitluspõhiseks leibkonnaks see võib jaguneda. Vastamaks küsimusele „Kas ühes elamuüksuses elavad inimesed, kes ütlevad end olevat erineva leibkonna liikmed, paistavad millegi poolest silma?“ võrreldi kaht gruppi inimesi – esimene grupp moodustati neist, kelle puhul leibkonna definitsiooni muutus ei toonud kaasa leibkonna koosseisu muutust (1 141 138 inimest, 89%), ning teine grupp neist, kes loenduse käigus ütlesid või kelle kohta öeldi, et nemad ei moodusta üht leibkonda kõigiga, kes elavad samas elamuüksuses (138 190 inimest).

Meeste ja naiste osakaal nimetatud gruppides on sisuliselt sama, kuid jagades isikud viieaastastesse vanusrühmadesse (joonis 1) selgus, et enim puudutab leibkonna definitsiooni erinevus 21–25-aastaseid noori, järgnevad 26–30 ja 16–20-aastaste vanusrühmad.



Joonis 1. Aadressi- ja ütluspõhise leibkonna kattuvus vanusgrupiti; isikute osakaal ja arv

Selle teadmise taustal kontrolliti, kas 16–30-aastaste puhul võiks tegu olla (üli)õpilastega, kes on kolinud vanematest eraldi (üli)koolile lähemale, näiteks mõne vanema sugulase juurde, või

jagavad eluruumi (üli)koolikaaslastega, kuid majandavad end seejuures eraldi. Selleks võeti Eesti Hariduse Infosüsteemist (EHIS) nimetatud vanusvahemikus isikute andmed nende isikute kohta, kes olid 31.12.2011 seisuga õppurid ja lingiti nad REL11 andmetega. 21–25 aastaste isikute hulgas, kelle leibkonna suurus jäi mõlema leibkonna definitsiooni juures samaks, oli õppureid 37,3%; vaadates sama vanusrühma grupis, kus leibkonna suurus muutus, oli õppureid 43,5%. See teeb õppurite osakaalu erinevuseks nimetatud gruppides veidi üle 6 protsendipunkti. Vanusrühmades 16–20 ja 26–30 oli õppurite osakaal kahes grupis enam-vähem võrdne.

Need on muutused leibkonna koosseisus, millega tuleb paratamatult arvestada, kui minna üle küsitluspõhise loenduse pealt registripõhisele. Reaalselt võib registri kasutamine õige elukoha registreerimata jätmisest tulenevalt erinevust veelgi suurendada.

3.3. Vabaabielupartnerid logistilisest regressioonimudelist

Igast vähemalt kaheliikmelisest REL11 tavaleibkonnast kaasati analüüsi kõikvõimalikud seosepaarid erinevast soost isikute vahel. Uuritavaks tunnuseks võeti binaarne tunnus Y, mille väärtus on 1, kui vaadeldav seosepaar koosneb abikaasadest või vabaabielupartneritest, ja 0 kõigi teiste teadaolevate seosetüüpide puhul (vt tabel 3). Paare, kelle vahel oli seosetüüp teadmata, analüüsi ei kaasatud. Näiteks kui tegu on viieliikmelise leibkonnaga, kus on abielupaar, nende kaks poega ja üks tütar, siis sellest leibkonnast kaasatakse analüüsi kuus seosepaari: ema ja isa, ema ja esimene poeg, ema ja teine poeg, tütar ja esimene poeg, tütar ja teine poeg, tütar ja isa; ema ja isa puhul Y=1, teiste seosepaaride puhul Y=0.

Selle tulemusel saadi 733 733 seosepaarist koosnev andmestik, millele sobitati mitmeid logistilisi regressioonimudeleid. Kõigist seosepaaridest 273 896 (37%) olid partnerlusseosed. Käesolevas alapunktis kirjeldatakse lähemalt ja võrreldakse omavahel kahte võimalikku mudelit (M1 ja M2) partnerite määratlemiseks.

Argumenttunnustena on esimeses mudelis (M1) iga seosepaari kohta kolm pidevat tunnust ja üks kolmeväärtuseline diskreetne tunnus:

— **m_age** (meessoost isiku vanus täisaastates, loendusmomendi seisuga);
— **f_age** (naissoost isiku vanus täisaastates, loendusmomendi seisuga);

— **age_dif** (isikute vanusevahe erinevus, absoluutväärtus)[1];
— **child** (1, kui kummalgi pole last; 2, kui naisel ja/või mehel on laps, aga mitte ühine; 3, kui naisel ja mehel on ühine laps).

Seega mudel M1 hinnatakse kujul:

$$logit(y_i) = \alpha + \beta_1 m\_age + \beta_2 f\_age + \beta_3 age\_dif + \beta_4(child = 2) + \beta_5(child = 3).$$

Teises mudelis (M2) on lisaks eelpool nimetatud tunnustele üks binaarne tunnus:

— **hh_size** (1, kui isikud elavad kaheliikmelises leibkonnas; 0, kui suuremas).

Seega mudel M2 hinnatakse kujul:

$$logit(y_i) = \alpha + \beta_1 m\_age + \beta_2 f\_age + \beta_3 age\_dif + \beta_4(child = 2) + \beta_5(child = 3) + \\ + \beta_6(hh\_size = 1).$$

Tehtud mudelite kohta on lisatud SAS *proc logistic* väljatrükid magistritöö lisades 4 ja 5, edasine arutelu tugineb nendele.

Mudeli M1 parameetervektori $\boldsymbol{\beta} = (\alpha, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ hinnanguks saadi suurima tõepära meetodil $\widehat{\boldsymbol{\beta}}$ =(-0,90; 0,14; -0,02; -0,33; -1,21; 2,67) ja mudeli M2 parameetervektori $\boldsymbol{\beta} = (\alpha, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)$ hinnanguks $\widehat{\boldsymbol{\beta}}$ = (1,86; 0,12; -0,05; -0,33; -0,90; 3,55; 1,86).

Hosmer-Lemeshow'i testi põhjal tuleb mõlema mudeli korral nullhüpotees kummutada (H_0: mudel sobib andmetega). Nagu metoodikat tutvustavas osas kirjeldati, siis testi tulemus näitab, et mudelisse erinevaid tunnuseid lisades ja mudelit tunnuste koosmõjudega täiendades oleks võimalik saada paremaid prognoose kui olemasoleva mudeliga. See aga ei tähenda automaatselt, et praegune mudel prognoosib (süstemaatiliselt) valesti.

Mudelite prognoosivõimet uuriti Nagelkerke parandatud determinatsioonikordajaga (*proc logistic* väljundis *Max-rescaled R-Square*). Esimese mudeli jaoks saadi selle väärtuseks 0,93 ja teise mudeli jaoks 0,95. Mõlema mudeli prognoosivõimet prooviti veelgi suurendada, lisades mudelisse erinevaid koosmõjusid – kuna oluliselt paremaid prognoose ei saadud, siis otsustati jääda lihtsasti interpreteeritavate koosmõjudeta mudelite juurde.

[1] |m_age-f_age| (võib tegelikust vanusevahest veidi erineda)

Mudelite võrdlemise seisukohalt vaadatakse edasi diagnostiliste testide tulemusi. Sobitatud mudelite jaoks valiti klassifitseerimistabelist tõenäosuse otsustuspunktid Youdeni[2] indeksi järgi. Selline lähenemine valiti, kuna antud ülesandpüstituse puhul ei ole põhjust arvata, et valepositiivne prognoosiväärtus on kuidagi halvem kui valenegatiivne, või vastupidi. Mõlemal juhul on võimalik ja vajalik tulemusi mudelijärgselt parandada (selle kohta on tehtud ettepanekud käesoleva alapunkti lõpus).

Youdeni indeksi järgi võeti M1 jaoks otsustuspunkti väärtuseks 0,34 ning M2 jaoks 0,26 – valitud punktide juures arvutatud tundlikkus, spetsiifilisus, valepositiivsuse määr ja valenegatiivsuse määr on toodud tabelis 8.

Tabel 8. Diagnostiliste testide tulemused mudelite M1 ja M2 jaoks

Mudel	Otsustuspunkt (PLEVEL)	Tundlikkus	Spetsiifilisus	Valepositiivsuse määr	Valenegatiivsuse määr
M1	0,34	97,7 %	96,6%	5,5%	1,3%
M2	0,26	98,4%	97,9%	3,5%	1,0%

Mudeli M1 valepositiivsuse määr on 5,5% ja valenegatiivsuse oma 1,3%, mudeli M2 vastavad näitajad on juba oluliselt paremad – 3,5% ja 1,0%. Kuna teine mudel erineb esimesest vaid ühe binaarse tunnuse võrra, siis on selgelt näha, et teadmisest, kas isikud elavad kaheliikmelises või suuremas leibkonnas, on partnerite mudelipõhisel määratlemisel oluline tähtsus. Valenegatiivsuse määr langes 0,3 protsendipunkti võrra ja valepositiivsuse määr koguni 2 protsendipunkti võrra. Sellest, kui palju tehti ühe või teise mudeli tulemusel partnerite määratlemise jaoks õigeid või valesid otsuseid, annavad ülevaate tabelid 9 ja 10.

Tabel 9. Mudeli M1 põhjal prognoositud tulemuste kooskõla tegelikkusega

		Y prognoositud		
		ei ole partnerid	on partnerid	**Kokku**
Y tegelik	ei ole partnerid	444 263	15 574	**459 837**
		60,55	2,12	**62,67**
	on partnerid	6377	267 519	**273 896**
		0,87	36,46	**37,33**
	Kokku	**450 640**	**283 093**	**733 733**
		61,42	**38,58**	**100**

[2] Youdeni $J = \max\{\text{Sensitivity} + \text{Specificity} - 1\}$

Tabel 10. Mudeli M2 põhjal prognoositud tulemuste kooskõla tegelikkusega

		Y prognoositud		
		ei ole partnerid	on partnerid	**Kokku**
Y tegelik	ei ole partnerid	449 974	9863	**459 837**
		61,33	1,34	**62,67**
	on partnerid	4453	269 443	**273 896**
		0,61	36,72	**37,33**
	Kokku	**454 427**	**279 306**	**733 733**
		61,93	**38,07**	**100**

Kuna mudeli sobitamiseks võeti aluseks kõik vastassooliste seosepaarid, siis kindlasti kaasati osaliselt ka järgmisi seoseid: ema ja poeg, tütar ja isa, õde ja vend, vanavanem ja lapselaps. Samuti olid uuritud seosepaaride hulgas abielus isikud. Kuna RRis on nimetatud seosed küllaltki hästi kajastatud või kaudselt tuletatavad, siis mudeli prognooside korrigeerimine näeb ette järgmist:

1. Vaadatakse kõiki seosepaare, mis määrati mudeli poolt partnerlusseosteks ning kui seal hulgas on lähisugulasi, siis prognoosiväärtus muudetakse nulliks.
2. Vaadatakse kõiki seosepaare, mida ei määratud mudeli poolt partnerlusseosteks, kui sealt õnnestub tuvastada abielus isikuid, siis prognoosiväärtus muudetakse üheks.
3. Kontrollitakse, kas mudeli tulemusel on üks isik seotud mitme partneriga. Kui jah, siis võetakse kasutusele teatud kriteeriumid, kas ja mille alusel üht seost teisele eelistada; üks võimalik lähenemine on toodud alapunktis 3.4.

3.4. Vabaabielupartnerid Soome eeskujul

Mõningate eranditega, kuid üldiselt Soomest eeskuju võttes, prooviti REL11 andmetel vabaabielupartnerid kindalaks teha logistilist regressioonimudelit kasutamata. Selleks moodustati RELi lähteandmebaasidest uus baas: analüüsi kaasati kõik tavaleibkondades elavad püsielanikud. Loomaks RRiga võimalikult sarnane „teadmiste hulk“, eeldati seekord, et on teada isiku ema ja isa identifikaator (seosed LABist F_ISIK, tabel 2), et abielus isikud on seotud abikaasa identifikaatoriga (seosed LABist F_LEIBKONNASUHE, tabel 3) ning oletati, et vabaabielupartnerid ei ole teada. Võrdluse raames defineeriti leibkond nii nagu RELis (info LABist F_ISIK.LEKO_ID, tabel 4).

Abielus mitteolevate isikute puhul uuriti, kas ühes leibkonnas on isikuid, kellel on ühine laps; kas laps elab vanematega samas leibkonnas või mitte, ei olnud määrav. Edasi käsitleti ka neid partneritena. Siinkohal tuleb tõdeda, et tänapäeval on erinevates kärgperedes elamine küllaltki tavaline ning teiste seas on olemas näiteks leibkonnad, mille liikmeteks on ema kahe lapsega ja mõlema lapse isa, kes ei ole sama isik – üheks võimaluseks on sel juhul lugeda partneriteks need lapsevanemad, kelle laps on noorem, magistritöös nii ka tehti.

Kui abikaasad ja ühiste laste emad-isad on teada, jääb üle tuvastada ilma ühiste lasteta vabaabielupartnerid. Selleks kontrollitakse iga täisealise isiku ***A*** puhul, kellele eelneva tulemusel ei ole partnerit leitud, kui mitu täisealist isikut ***B***, kellele samuti ei ole eelneva põhjal partnerit leitud, elab isikuga ***A*** samas leibkonnas ja vastab järgmistele kriteeriumidele:

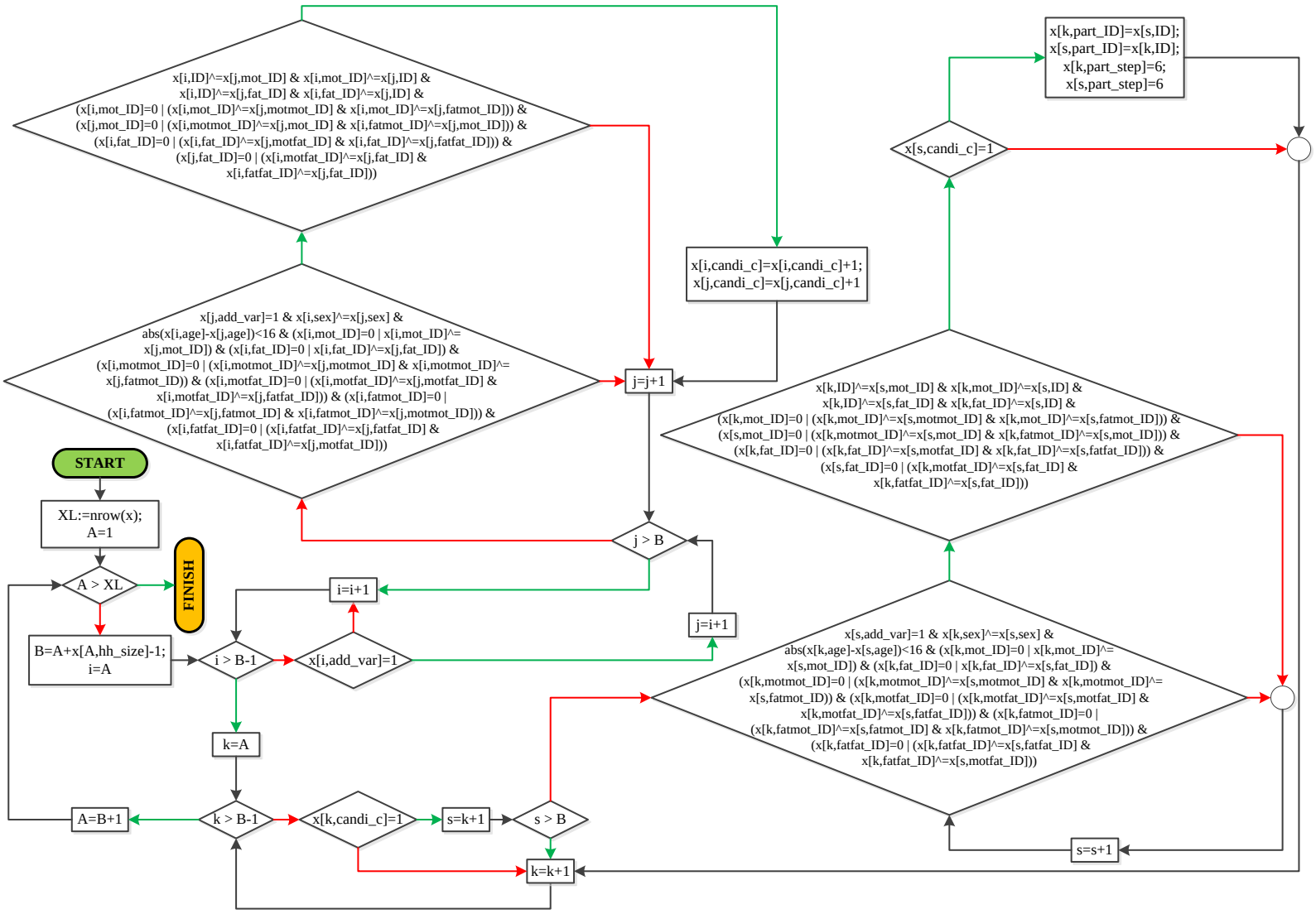
— ***A*** ja ***B*** on erinevast soost;
— ***B*** vanus täisaastates (loendusmomendil) ei erine ***A*** omast üle 16 aasta;
— ***A*** ja ***B*** ei ole lähisugulased (välistatakse olukord, kus paari võivad moodustada vanem ja laps, laps oma onu või tädiga, või onude ja tädide lapsed omavahel).

Kui isiku ***A*** jaoks leitakse ainult üks isik ***B***, kes nimetatud kriteeriumidele vastab ning samuti ei kehti isiku ***B*** jaoks üle ühe niisuguse seose, siis isikud ***A*** ja ***B*** defineeritakse partneriteks. Kui seos ***A*** ja ***B*** vahel ei ole üks-ühene, siis ühtegi partnerlusseost ei defineerita. Meeldetuletuseks: kuni üks neljale seose puhul defineeritakse Soomes partneriteks need, kelle vanusevahe on väiksem; RELi andmete põhjal viis selline lähenemine olukorra pigem tegelikkusest kaugemale.

Magistritöö praktilise ülesandena koostati kahest osast koosnev IML algoritm (programmikood on esitatud lisas 2), mille esimeses osas *partners* püütakse tuvastada lasteta vabaabielupartnerid eeldusel, et abikaasad on teada ja samas leibkonnas elavaid isikuid, kel on ühiseid lapsi, käsitletakse vabaabielupartneritena. Algoritmi teises osas *families* jagatakse leibkonnaliikmed vastavalt definitsioonile perekondadesse, seda osa tutvustatakse punktis 3.5. Algoritmis kasutatavad tähistused on toodud tabelis 11 ja plokkskeem lasteta vabaabielupartnerite määratlemiseks joonisel 2.

Tabel 11. Algoritmis kasutatavad tähistused (vt joonis 2 ja joonis 3)

Lühend	Seletus {veeru nr algoritmis}
A	Leibkonna esimese liikme järjekorranumber (noorim isik leibkonnas)
add_var	1, kui isik on mitteabielus täisealine ning elab vähemalt kaheliikmelises leibkonnas ja tal ei ole sellest leibkonnast kellegagi ühiseid lapsi; 0 muul juhul (moodustatud enne algoritmi) {x[,19]}
age	Vanus täisaastates, loendusmomendil {x[,16]}
B	Leibkonna viimase liikme järjekorranumber (vanim isik leibkonnas)
birth	Isiku sünniaeg (kasutatakse leibkonnaliikmete vanuse järgi järjestamiseks) {x[,18]}
candi_c	„Partnerikandidaatide“ arv (algväärtus 0, isiku kohta leitakse, mitu isikut sellest leibkonnast antud kriteeriumidel talle partneriks sobiks; vabaabielu-partnerid määratakse üksnes siis, kui saadakse üks ühele seos) {x[,20]}
fam_nr	Perekonna number (algoritmis algväärtus 0) {x[,5]}
fat_ID	Isa identifikaator {x[,10]}
fat_hh_nr	Isa leibkonna number {x[,11]}
fatfat_ID	Vanaisa (isa isa) identifikaator {x[,15]}
fatmot_ID	Vanaema (isa ema) identifikaator {x[,14]}
hh_nr	Leibkonna number {x[,2]}
hh_size	Leibkonna suurus {x[,4]}
ID	Isiku identifikaator {x[,1]}
mot_hh_nr	Ema leibkonna number {x[,9]}
mot_ID	Ema identifikaator {x[,8]}
motfat_ID	Vanaisa (ema isa) identifikaator {x[,13]}
motmot_ID	Vanaema (ema ema) identifikaator {x[,12]}
nr	Leibkonnaliikme number vanuselises järjekorras, alustades noorimast {x[,3]}
part_ID	Partneri identifikaator, eelnevalt täidetud, kui samas leibkonna elab abikaasa ja/või ühiste laste ema/isa {x[,6]}
part_step	Tunnus selle kohta, millisel sammul isikule partner leitakse; eelnevalt on täidetud, kui samas leibkonnas elab abikaasa ja/või ühiste laste ema/isa; kui partner leitakse algoritmi sees, siis täidetakse part_step=6 {x[,7]}
sex	Isiku sugu (1 mees, 2 naine) {x[,17]}
XL	Kõigi inimeste arv, ehk ridade arv algandmestikus



Joonis 2. Vabaabielupartnerite määratlemine (Märkus: Tingimuse kontrollist väljub roheline nool, kui vastav tingimus kehtib, ning punane nool, kui ei kehti)

Partnerite algoritmi rakendamise tulemusel saadi tegelikkusele küllalt lähedane olukord, kus 1 279 328 isikust erines partneri (abikaasad ja vabaabielupartnerid) olemasolu või identifikaator ainult 14 288 isiku puhul (ca 1,5 %).

Eesmärgiga avastada võimalikku süstemaatilist kõrvalekallet, uuriti tulemusi lähemalt; st võeti iga leibkonna sees kõikvõimalikud sellised seosepaarid vastassooliste leibkonnaliikmete vahel (info LABist F_LEIBKONNASUHE, tabel 3), kus suhte tüüp oli teada ning võrreldi RELi teadaolevaid partnereid nendega, kes määrati partneriks algoritmi põhjal.

Sellest, kui palju tehti algoritmi põhjal partnerite määratlemisel õigeid või valesid otsuseid, annab ülevaate tabel 12.

Tabel 12. Algoritmi põhjal prognoositud tulemuste kooskõla tegelikkusega

		Y prognoositud		
		ei ole partnerid	on partnerid	**Kokku**
Y tegelik	ei ole partnerid	456 223	3614	**459 837**
		62,18	0,49	**62,67**
	on partnerid	2409	271 487	**273 896**
		0,33	37,00	**37,33**
	Kokku	**458 632**	**275 101**	**733 733**
		62,51	**37,49**	**100**

Seosepaaridest, mis tegelikult partnerlusseosed ei ole, määrati algoritmi tulemusel partnerlusseoseks 1,31% (so valepositiivsuse määr); tegelikest partnerlusseostest jääb algoritmi tulemusel tuvastamata 0,53% (so valenegatiivsuse määr). Seosepaarid võib jagada nelja suuremasse gruppi sõltuvalt sellest, kas osapooled on partnerid RELi andmetel ja/või kas nad määrati partneriteks ühiste laste olemasolu või algoritmis toodud kriteeriumide alusel.

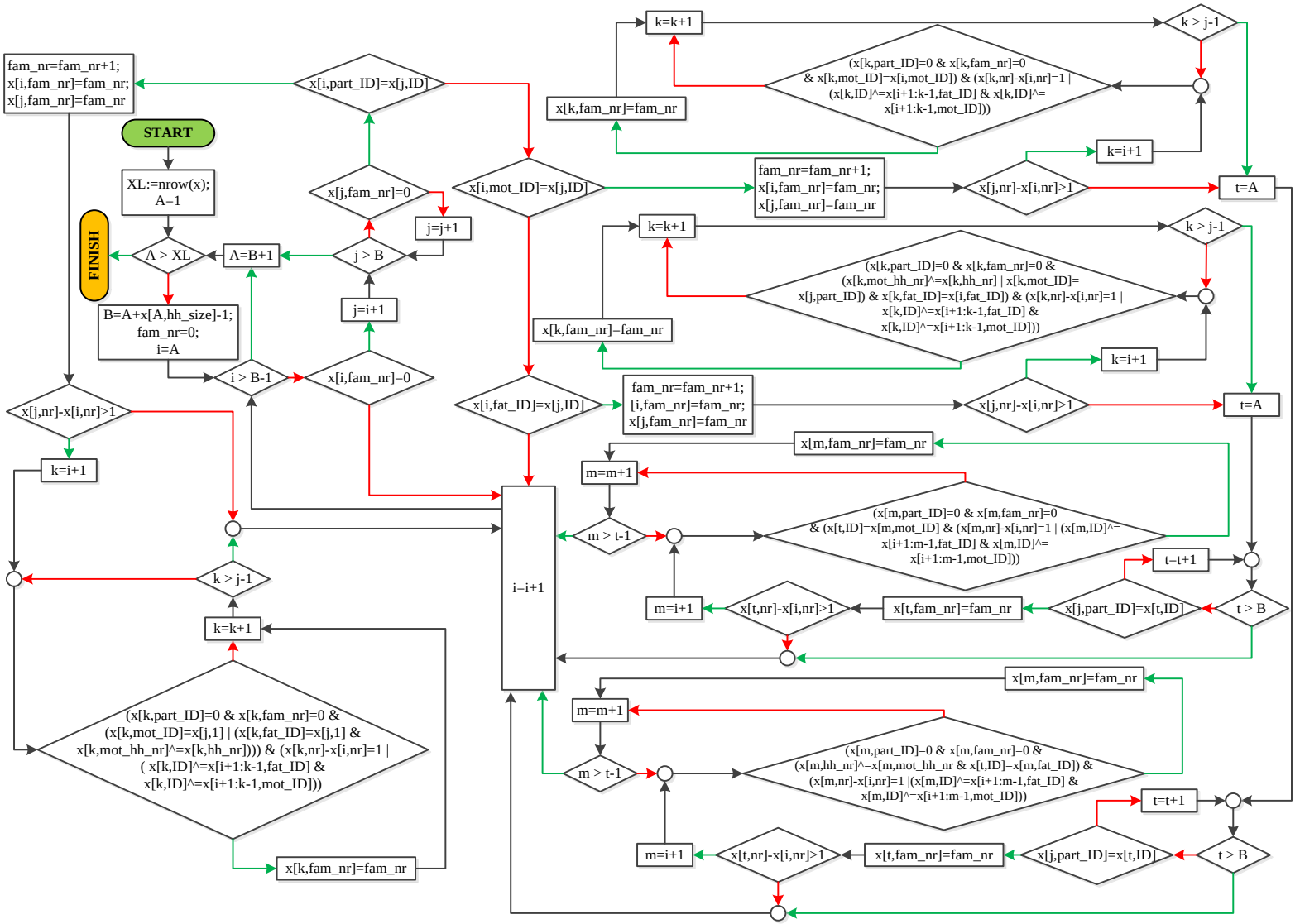
Kuna uuritud seosepaaride hulgas on igas vanuses isikuid, siis on suurim see grupp, mille moodustavad isikud, kes ei ole RELi andmetel partnerid ning keda ei määranud partneriteks ka algoritm (456 223 seosepaari). Valdavalt on tegemist selliste paaridega, kus vähemalt üks osapool on alaealine. Suuruselt järgmine grupp koosneb RELi partneritest, kelle vahel tuvastati partnerlusseos ka algoritmi põhjal või ühiste laste olemasolu tõttu – 271 487 seosepaari. Neist 44 644 defineeriti partnerlusseosteks ühiste laste olemasolu alusel ning 37 840 algoritmis toodud kriteeriumide tulemusel. Lisaks kuulub selles gruppi 189 003 abikaasadevahelist seost (kuna rahvastikuregistris on selline info olemas, siis eeldati ka

algoritmi puhul, et abikaasad on teada). Kokku tehti partnerlusseose olemasolu või selle puudumise kohta õige otsus 727 710 seosepaari puhul (sh abikaasad).

Vale otsus tehti vaid 6923 seosepaari puhul. Algoritmi tulemusel saadi 2409 seosepaari, mis koosnevad RELi andmetel vabaabielupartneritest, kuid kelle vahel algoritm partnerlusseost tuvastada ei suutnud; neist 1375 puhul ületas vanusevahe etteantud kriteeriumit (partnereid otsiti vaid seosepaaride hulgast, kus isikute vanusevahe jäi alla 16 aasta). Seosepaare, mis RELi andmetel ei koosne ei vabaabielupartneritest ega abikaasadest, kuid kaudsete kriteeriumide alusel siiski partnerlusseosteks defineeriti, on 3614 – neist 1285 defineeriti partnerlusseosteks ühiste laste olemasolu tõttu ja 2329 teistel tingimustel.

3.5. Leibkonnaliikmete jagamine perekondadesse

Olles tuvastanud ühes eluruumis elavate isikute vahel vanem/laps seosed ning teinud kindlaks abikaasad ja vabaabielupartnerid, tuleb leibkonnaliikmed tuumperekondadesse jagada. Arvestades kõikvõimalike seoseid ühes eluruumis elavate isikute vahel, koostati algoritm, mille tulemusel on leibkonnaliikmed vastavalt definitsioonile perekondadesse jagatud. Tehniliselt täidab algoritm oma eesmärki olenemata sellest, kuidas tulevikus partnerlusseosed kindlaks tehakse. Algoritmi tähistused on tabelis 11 ning plokkskeem joonisel 3. (IML kood asub lisas 2, osas *families*).



Joonis 3. Leibkonnaliikmete jagamine perekondadesse (Märkus: Tingimuse kontrollist väljub roheline nool, kui vastav tingimus kehtib, ning punane nool, kui ei kehti)

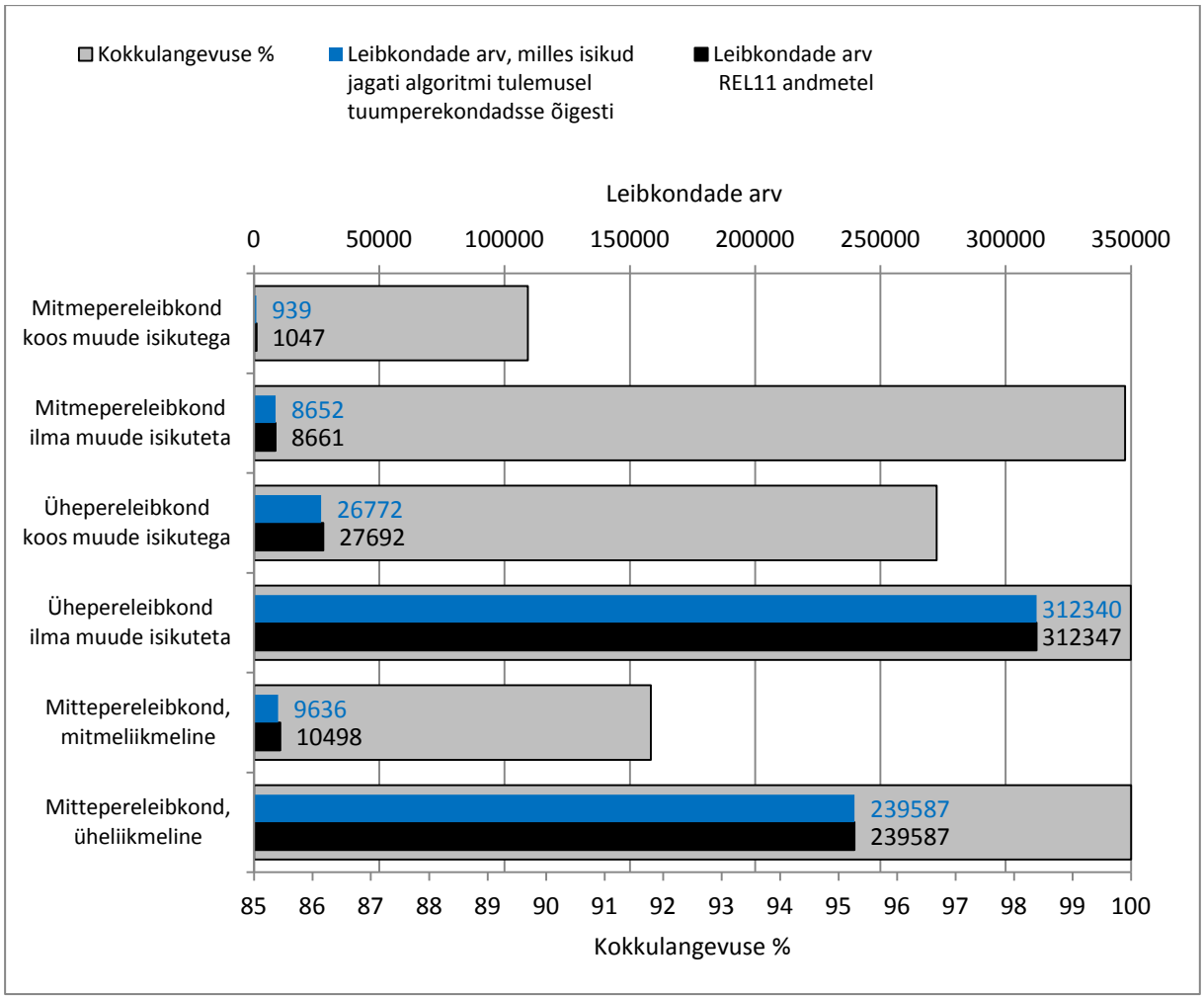
Eeldades, et kõik partnerlusseosed on teada, testiti algoritmi korduvalt REL11 andmete peal. Järgnev võrdlus tuumperekondade kokkulangevuse kohta (leibkonnatüübiti) annab kinnitust, et algoritmist tingituna süstemaatilisi kõrvelekaldeid ei täheldatud.

RELi LABis on 599 832 tavaleibkonda, neist 597 926 jagati isikud algoritmi tulemusel tuumperekondadesse õigesti. Seega oli erinevusi vaid 1906 leibkonna puhul, mis on 0,32% kõigist tavaleibkondadest ja 0,53% kaheliikmelistest või suurematest tavaleibkondadest. Tabelis 13 on esitatud tavaleibkondade üldine jaotus leibkonna tüübi järgi (infoallikas Eesti SA andmebaas [13]).

Tabel 13. Tavaleibkonnad leibkonna tüübi järgi [13]

Tavaleibkonna tüüp	N
Mittepereleibkond, üheliikmeline	239 587
Mittepereleibkond, mitmeliikmeline	10 498
Ühepereleibkond ilma muude isikuteta	312 347
Ühepereleibkond koos muude isikutega	27 692
Mitmepereleibkond ilma muude isikuteta	8661
Mitmepereleibkond koos muude isikutega	1047
Kokku	599 832

Saadud tulemust illustreerib joonis 4, millel on mustaga kujutatud vastavat tüüpi leibkondade tegelik arv ja sinisega nende leibkondade arv, mille liikmed jagunesid algoritmi põhjal tuumperekondadesse samamoodi nagu on kirjeldatud RELi LABis; halliga on kujutatud kokkulangevuse protsent.



Joonis 4. Tuumperekondade moodustamise algoritmi tulemused leibkonnatüübiti

Eraldi tähelepanu pöörati 1906 leibkonnale, mille korral algoritm ei andnud täpselt sama tulemust, mis RELi LABis kirjas. Tehti kindlaks, et kokkulangevuse puudumine ei ole tingitud mitte mõnest algoritmis olevast (süstemaatilisest) veast, vaid muudest põhjustest. Näiteks sellest, et algoritm ei olegi mõeldud töötama leibkondade puhul, milles on samasooliste partnerid.

3.6. Järeldused

Magistritöös testiti eelmise loenduse andmete peal kolme erinevat võimalust vabaabielupartnerite määratlemiseks – tehti kaks logistilist regressioonimudelit (M1 ja M2) ning kirjutati IML kood määratlemaks vabaabielupartnerid selliselt nagu teeb Soome Statistikaamet. Tabel 14 esitab kokkuvõtlikult saadud tulemust.

Tabel 14. Vabaabielupartnerite määratlemine; kokkuvõtlik tulemus

	Andmetest teadaolev arv	Õigesti määratletute arv ja osakaal					
		Mudel 1 (M1)		Mudel 2 (M2)		Algoritm	
Ei ole partnerid	459 837	444 263	96,60%	449 974	97,9%	456 223	99,20%
On partnerid	273 896	267 519	97,70%	269 443	98,4%	271 487	99,10%
Kokku	733 733	711 782	97,0%	719417	98,00%	727 710	99,20%

Selgub, et ka vaid nelja tunnuse põhjal on võimalik suhteliselt täpselt määratleda, kas kaks leibkonnaliiget on partnerid või mitte. Mudel M1, mis võttis arvesse osapoolte vanused, vanusevahe ja (ühiste) laste olemasolu, eksis parima otsustuspunkti juures kokku vaid 3% juhtudest.

Praktiliseks rakenduseks on selline eksimus siiski absoluutarvudes liiga suur, seega tasub arvesse võtta veel mõnda lisatingimust. Mudel M2, milles lisaks eelpool nimetatud tunnustele on binaarne tunnus leibkonna suuruse kohta (kaheliikmeline või suurem), vähendas kogu eksimust 2%-ni ehk 1,5 korda. Samas moodustab ka selline eksimus kontrollimise aluseks võetud reaalse andmehulga puhul 14 316 seosepaari, mille korral mudel ei suuda isikutevahelise suhte staatust (kas tegu on partneritega või mitte) õigesti prognoosida.

Kõige parema tulemuse saavutas vaieldamatult Soome Statistikaameti algoritm, mis eksis vaid 0,8% juhtudest. Suurem täpsus võrreldes mudelitega M1 ja M2 on osaliselt seletatav ka sellega, et algoritm kasutab veelgi rohkem informatsiooni, võttes arvesse ka osapooltevahelisi sugulussidemeid.

Tehtud analüüside põhjal soovitab töö autor vabaabielupartnerite määratlemiseks algoritmipõhist lähenemist, tehes esialgu isikute soo, vanuse, vanusevahe ja sugulussidemete kohta samasugused nõudmised nagu tehakse Soome Statistikaametis. Kui tulevikus osatakse toodud kriteeriumidele paremaid hinnanguid anda, on algoritm lihtsasti ümberhäälestatav.

KOKKUVÕTE

Järgmine rahva ja eluruumide loendus Eestis plaanitakse läbi viia registripõhiselt. Magistritöö raames tegeldi loenduse osaga, mis käsitleb leibkondade ja perekondade registripõhist määratlemist. Magistritöö eesmärk oli pakkuda välja algoritmid, mida on võimalik testida käesoleva aasta sügisel Statistikaameti poolt läbiviidaval pilootloendusel ning mis sobivusel on hiljem kasutatavad ka registripõhisel rahva ja eluruumide loendusel. 2011. aasta rahva ja eluruumide loenduse (REL11) anonüümitud andmetele tuginedes viidi teoreetilise osa taustal läbi mitmed praktilised ülesanded.

Leibkonna defineerimiseks on kaks võimalust, millest üks on iseloomulik küsitluspõhisele loendusele ja teine sobib hästi registripõhisega. Üheks töö eesmärgiks oli uurida, kas ja kui suur erinevus tekib leibkondade koosseisus, kui muuta leibkonna definitsiooni ning kas definitsiooni vahetamisega kaasnev erinevus puudutab võrdselt mehi ja naisi ning erinevas vanuses inimesi. Analüüsist selgus, et tavaleibkondades elavatest püsielanikest 89% moodustaksid täpselt sama leibkonna olenemata sellest, et leibkonnad defineeritakse erinevalt. Meeste ja naiste jaotus osutus enam-vähem võrdseks. Vanusjaotust vaadati viieaastastes gruppides ning selgus, et enim on seotud leibkonna definitsiooni erinevusest tuleneva muutusega leibkonna koosseisus 21–25-aastased noored, järgnevad 26–30 ja 16–20-aastaste vanusgrupid.

Perekonnad otsesel kujul üheski administratiivses registris ei kajastu. Küll aga on rahvastikuregistris isikud seotud oma vanematega ja abielus isikud abikaasaga. Käesoleva töö peamine eesmärk oli leida võimalusi, kuidas niisuguse mittetäieliku info korral leibkonnaliikmed perekondadesse jagada, seejuures osutus kõige keerulisemaks vabaabielupartnerite määratlemine. Erinevaid lähenemisviise tutvustati, analüüsiti ja neist tehti kokkuvõtlikud järeldused (viimased on esitatud alapunktis 3.6.).

Töö praktilise osana on koostatud algoritmid, mis on ette nähtud kasutamiseks ja vajadusel ka täiendamiseks käesoleva aasta sügisel, mil registripõhisele loendusele ülemineku raames viiakse läbi pilootloendus.

KASUTATUD KIRJANDUS

[1] **Puur, A.**, **Sakkeus, L.**, **Aben, S**. TLÜ Eesti Demograafia Instituut ja Ernst & Young Baltic AS. (2013). REGREL metoodika väljatöötamise projekti lõpparuanne. [PDF]. http://www.stat.ee/dokumendid/76831
(06.05.2014)

[2] Euroopa Nõukogu ja Parlamendi Määrus EÜ nr 763/2008. (2008). [PDF]
https://www.stat.ee/dokumendid/30152
(11.03.2014)

[3] Euroopa Komisjoni Määrus EÜ nr 1201/2009.(2009). [PDF].
http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:329:0029:0068:ET:PDF
(11.03.2014)

[4] Aadressiandmete süsteem. (2007). – *Riigiteataja* I, 71, 439.

[5] Quality description, families. (2012). – *Statistics Finland* [WWW]
http://www.stat.fi/til/perh/2012/02/perh_2012_02_2013-11-22_laa_001_en.htmld
(01.04.2014).

[6] Registreeritud partnerlus. (2013). [WWW]
http://europa.eu/youreurope/citizens/family/couple/registered-partners/index_et.htm
(11.03.2014).

[7] **Allison, Paul D.** (1999). Logistic Regression Using the SAS® System: Theory and Application. Cary, NC: SAS Institute Inc.

[8] **Käärik, E.** (2007). Matemaatiline statistika II. – *Loengukonspekt*.

[9] **Allison, Paul D.** (2014). Measures of Fit for Logistic Regression. Statistical Horizons LLC and the University of Pennsylvania. SAS Global Forum. Paper 1485-2014.

[10] The Hosmer-Lemeshow goodness of fit test for logistic regressioon. – *TheStatsGeek.* [WWW]
http://thestatsgeek.com/2014/02/16/the-hosmer-lemeshow-goodness-of-fit-test-for-logistic-regression/
(01.04.2014).

[11] **Allison, Paul D.** (2013). Why I Don't Trust the Hosmer-Lemeshow Test for Logistic Regression. – *Statistical Horizons.* [WWW]
http://www.statisticalhorizons.com/hosmer-lemeshow
(09.04.2014).

[12] **Scheaffer , Richard L.** (1999). University of Florida. Categorical Data Analysis

[13] RL0712: Tavaleibkonnad ja nende liikmed leibkonna tüübi järgi, 31. detsember 2011 – Eesti *Statistikaameti andmebaas*. [WWW].
http://pub.stat.ee
(14.05.2014)

Lisa 1. Kohustuslike tunnuste loendusstandardile vastav jaotus

Allpool toodud jaotuses mõeldakse isiku all Eesti püsielanikku. Sulgudes on toodud koodiline väärtus, mida kasutatakse IML-algoritmis (vt lisa 3).

Seisund leibkonnas *(hh_status_J)*

0. Kokku
1. Tavaleibkonnas elavad isikud
  1.1. Tuumperekonda kuuluvad isikud
    1.1.1. Abielus isikud (1110)
    1.1.2. Partnerid registreerimata kooselus (1120)
    1.1.3. Üksikvanemad (1130)
    1.1.4. Pojad/tütred
      1.1.4.1. Ei ole üksikvanema lapsed (1141)
      1.1.4.2. Üksikvanema lapsed (1142)
  1.2. Isikud, kes ei kuulu tuumperekonda
    1.2.1. Üksinda elavad (1210)
    1.2.2. Ei ela üksinda (1220)
2. Isikud, kes ei ela tavaleibkonnas

Seisund perekonnas *(fam_status_J)*

0. Kokku
1. Partnerid
  1.1. Abielupaari moodustavad isikud (11)
  1.2. Partnerid registreerimata kooselus (12)
2. Üksikvanemad (20)
3. Pojad/tütred
  3.1. Ei ole üksikvanema lapsed (31)
  3.2. Üksikvanema lapsed (32)
4. Ei kohaldata (40)

Tavaleibkonna tüüp *(hh_type_J)*

0. Kokku

1. Leibkonnad, mis ei moodusta pereleibkonda

1.1. Üheliikmelised leibkonnad (110)

1.2. Mitmeliikmelised leibkonnad (120)

2. Ühepereleibkonnad

2.1. Abielupaariga leibkonnad

2.1.1. Abielupaarid, kelle lapsed ei ela kodus (211)

2.1.2. Abielupaarid, kelle vähemalt üks alla 25aastane laps elab kodus (212)

2.1.3. Abielupaarid, kelle noorim kodus elav laps on vähemalt 25aastane (213)

2.2. Registreerimata kooselupaaride leibkonnad

2.2.1. Registreerimata kooselupaarid kodus elavate lasteta (221)

2.2.2. Registreerimata kooselupaar, kelle vähemalt üks alla 25aastane laps elab kodus (222)

2.2.3. Registreerimata kooselupaarid, kelle noorim kodus elav poeg/tütar on vähemalt 25aastane (223)

2.3. Üksikisade leibkonnad

2.3.1. Üksikisade leibkonnad, kelle vähemalt üks alla 25aastane laps elab kodus (231)

2.3.2. Üksikisade leibkonnad, kelle noorim kodus elav poeg/tütar on vähemalt 25aastane (232)

2.4. Üksikemade leibkonnad

2.4.1. Üksikemade leibkonnad, kelle vähemalt üks alla 25aastane laps elab kodus (241)

2.4.2. Üksikemade leibkonnad, kelle noorim kodus elav poeg/tütar on vähemalt 25aastane (242)

3. Kahe- või mitmepereleibkonnad (300)

Tuumperekonna tüüp *(fam_type_J)*

0. Kokku
1. Abielupaaride perekonnad
  1.1. Abielupaaride perekonnad, kelle lapsed ei ela kodus (11)
  1.2. Abielupaar, kelle vähemalt üks alla 25aastane laps elab kodus (12)
  1.3. Abielupaaride perekonnad, kelle noorim kodus elav poeg/tütar on vähemalt 25aastane (13)
2. Registreerimata kooselus paaride perekonnad
  2.1. Registreerimata kooselus elav paar, kelle lapsed ei ela kodus (21)
  2.2. Registreerimata kooselus elav paar, kelle vähemalt üks alla 25aastane laps elab kodus (22)
  2.3. Registreerimata kooselus elav paar, kelle noorim kodus elav poeg/tütar on vähemalt 25aastane (23)
3. Üksikisade perekonnad
  3.1. Üksikisade perekonnad, kelle vähemalt üks alla 25aastane laps elab kodus (31)
  3.2. Üksikisade perekonnad, kelle noorim kodus elav poeg/tütar on vähemalt 25aastane (32)
4. Üksikemade perekonnad
  4.1. Üksikemade perekonnad, kelle vähemalt üks alla 25aastane laps elab kodus (41)
  4.2. Üksikemade perekonnad, kelle noorim kodus elav poeg/tütar on vähemalt 25aastane (42)

Tavaleibkonna suurus *(hh_size_J)*

0. Kokku
1. 1 isik (1)
2. 2 isikut (2)
3. 3 isikut (3)
4. 4 isikut (4)
5. 5 isikut (5)
6. 6 isikut (6)
7. 7 isikut (7)
8. 8 isikut (8)
9. 9 isikut (9)
10. 10 isikut (10)
11. Vähemalt 11 isikut (11)

Tuumperekonna suurus *(fam_size_J)*

0. Kokku
1. 2 isikut (2)
2. 3 isikut (3)
3. isikut (4)
4. 5 isikut (5)
5. 6 isikut (6)
6. 7 isikut (7)
7. 8 isikut (8)
8. 9 isikut (9)
9. 10 isikut (10)
10. Vähemalt 11 isikut (11)

Lisa 2. Algoritm – lasteta vabaabielupartnerite määratlemine; perekondade moodustamine

```
proc iml; start; use tp_maj_v2_temp1;
read all var _all_ into x;
veerunimed={"ID" "hh_nr" "nr" "hh_size" "fam_nr" "part_ID" "part_step" "mot_ID"
"mot_hh_nr" "fat_ID" "fat_hh_nr" "motmot_ID" "motfat_ID" "fatmot_ID" "fatfat_ID"
"age" "sex" "birth" "add_var" "candi_c"}; *tähistusi vt tabel 11;
count=0;
XL=nrow(x);
start_p:
A=1;
do while (A<XL);
B=A+x[A,4]-1;
if count=0 then goto partners;
if count=1 then goto families;
partners: *ühiste lasteta mitteabielus paaride leidmine;
do i=A to B-1;
if x[i,19]=1 then
do j=i+1 to B;
if x[j,19]=1 & x[i,17]^=x[j,17] & abs(x[i,16]-x[j,16])<16
& (x[i,8]=0 | x[i,8]^=x[j,8]) & (x[i,10]=0 | x[i,10]^=x[j,10])
& (x[i,12]=0 | (x[i,12]^=x[j,12] & x[i,12]^=x[j,14]))
& (x[i,13]=0 | (x[i,13]^=x[j,13] & x[i,13]^=x[j,15]))
& (x[i,14]=0 | (x[i,14]^=x[j,14] & x[i,14]^=x[j,12]))
& (x[i,15]=0 | (x[i,15]^=x[j,15] & x[i,15]^=x[j,13])) then do;
if x[i,1]^=x[j,8] & x[i,8]^=x[j,1]
& x[i,1]^=x[j,10] & x[i,10]^=x[j,1]
& (x[i,8]=0 | (x[i,8]^=x[j,12] & x[i,8]^=x[j,14]))
& (x[j,8]=0 | (x[i,12]^=x[j,8] & x[i,14]^=x[j,8]))
& (x[i,10]=0 | (x[i,10]^=x[j,13] & x[i,10]^=x[j,15]))
& (x[j,10]=0 | (x[i,13]^=x[j,10] & x[i,15]^=x[j,10])) then do;
x[i,20]=x[i,20]+1;
x[j,20]=x[j,20]+1;
end; end;
end; *j kinni;
end; *i kinni;
do k=A to B-1;
if x[k,20]=1 then
do s=k+1 to B;
if x[s,19]=1 & x[k,17]^=x[s,17] & abs(x[k,16]-x[s,16])<16
& (x[k,8]=0 | x[k,8]^=x[s,8]) & (x[k,10]=0 | x[k,10]^=x[s,10])
& (x[k,12]=0 | (x[k,12]^=x[s,12] & x[k,12]^=x[s,14]))
& (x[k,13]=0 | (x[k,13]^=x[s,13] & x[k,13]^=x[s,15]))
```

```
& (x[k,14]=0 | (x[k,14]^=x[s,14] & x[k,14]^=x[s,12]))
& (x[k,15]=0 | (x[k,15]^=x[s,15] & x[k,15]^=x[s,13])) then do;
if x[k,1]^=x[s,8] & x[k,8]^=x[s,1]
& x[k,1]^=x[s,10] & x[k,10]^=x[s,1]
& (x[k,8]=0 | (x[k,8]^=x[s,12] & x[k,8]^=x[s,14]))
& (x[s,8]=0 | (x[k,12]^=x[s,8] & x[k,14]^=x[s,8]))
& (x[k,10]=0 | (x[k,10]^=x[s,13] & x[k,10]^=x[s,15]))
& (x[s,10]=0 | (x[k,13]^=x[s,10] & x[k,15]^=x[s,10])) then do;
if x[s,20]=1 then do;
x[k,6]=x[s,1];
x[s,6]=x[k,1];
x[k,7]=6; x[s,7]=6;
end; end; end;
end; *s kinni;
end; *k kinni;
A=B+1; *järgmisesse leibkonda;
if A>=XL then count=1;
if count=1 then goto start_p;
families: *leibkonnaliikmete jagamine tuumperekondadesse (x[,5]);
fam_nr=0;
do i=A to B-1;
if x[i,5]=0 then *vaatame ainult neid, kes tuumperekonda ei kuulu;
do j=i+1 to B;
if x[j,5]=0 then do;
/* I kui isikul on samas leibkonnas partner */
if x[i,6]=x[j,1] then do;
fam_nr=fam_nr+1;
x[i,5]=fam_nr;
x[j,5]=fam_nr;
/* kui sellel partneril on lapsi, kel endal pole partnerit ega last, ja kes pole
üheski teises tuumperekonnas (meessoost partneri puhul ei kaasa last, kelle ema on
samas leibkonnas)*/
if x[j,3]-x[i,3]>1 then
do k=i+1 to j-1;
if x[k,6]=0 & x[k,5]=0
& (x[k,8]=x[j,1] | (x[k,10]=x[j,1] & x[k,9]^=x[k,2])) then do;
if x[k,3]-x[i,3]=1 then x[k,5]=fam_nr;
if x[k,3]-x[i,3]>1 then do;
if x[k,1]^=x[i+1:k-1,10] & x[k,1]^=x[i+1:k-1,8] then x[k,5]=fam_nr;
end; end; end; end;
/* II kui isikul on samas leibkonnas ema */
if x[i,6]=0 then do;
if x[i,8]=x[j,1] then do;
fam_nr=fam_nr+1;
x[i,5]=fam_nr;
x[j,5]=fam_nr;
```

```
/* kui emal on veel lapsi, kel pole samas leibkonnas partnerit ja kes pole üheski
teises tuumperekonnas ja kellel endal pole samas leibkonnas lapsi */
if x[j,3]-x[i,3]>1 then
do k=i+1 to j-1;
if x[k,6]=0 & x[k,5]=0 & x[k,8]=x[i,8] then do;
if x[k,3]-x[i,3]=1 then x[k,5]=fam_nr;
if x[k,3]-x[i,3]>1 then do;
if x[k,1]^=x[i+1:k-1,10] & x[k,1]^=x[i+1:k-1,8] then x[k,5]=fam_nr;
end; end; end;
/* kui lisaks emale on ka ema partner samas leibkonnas (see partner võib olla või
mitte olla jooksva isiku isa) */
do t=A to B;
if x[j,6]=x[t,1] then do;
x[t,5]=fam_nr;
/* kui sellel partneril on veel lapsi, kel pole samas leibkonnas partnerit ega ema
ja kes pole üheski teises tuumperekonnas ja kellel pole endal samas leibkonnas
lapsi */
if x[t,3]-x[i,3]>1 then
do m=i+1 to t-1;
if x[m,6]=0 & x[m,5]=0 & (x[m,2]^=x[m,9] & x[t,1]=x[m,10]) then do;
if x[m,3]-x[i,3]=1 then x[m,5]=fam_nr;
if x[m,3]-x[i,3]>1 then do;
if x[m,1]^=x[i+1:m-1,10] & x[m,1]^=x[i+1:m-1,8] then x[m,5]=fam_nr;
end; end; end; end; end; end; else;
/* III kui isikul on samas leibkonnas isa (ema ei ole samas leibkonnas!) */
if x[i,9]^=x[i,2] then do;
if x[i,10]=x[j,1] then do;
fam_nr=fam_nr+1;
x[i,5]=fam_nr;
x[j,5]=fam_nr;
/* kui isal on veel lapsi, kel pole samas leibkonnas partnerit ega ema (va siis,
kui ema on isa partner) ja kes pole üheski teises tuumperekonnas ja kellel pole
endal selles leibkonnas lapsi. */
if x[j,3]-x[i,3]>1 then
do k=i+1 to j-1;
if x[k,6]=0 & x[k,5]=0
& (x[k,9]^=x[k,2] | x[k,8]=x[j,6]) & x[k,10]=x[i,10] then do;
if x[k,3]-x[i,3]=1 then x[k,5]=fam_nr;
if x[k,3]-x[i,3]>1 then do;
if x[k,1]^=x[i+1:k-1,10] & x[k,1]^=x[i+1:k-1,8] then x[k,5]=fam_nr;
end; end; end;
/* kui lisaks isale on ka isa partner samas leibkonnas */
do t=A to B;
if x[j,6]=x[t,1] then do;
x[t,5]=fam_nr;
```

```
/* kui sellel partneril veel lapsi, kel pole samas leibkonnas partnerit ja kes pole
üheski teises tuumperekonnas ja kellel endal pole selles leibkonnas lapsi */
if x[t,3]-x[i,3]>1 then
do m=i+1 to t-1;
if x[m,6]=0 & x[m,5]=0 & x[t,1]=x[m,8] then do;
if x[m,3]-x[i,3]=1 then x[m,5]=fam_nr;
if x[m,3]-x[i,3]>1 then do;
if x[m,1]^=x[i+1:m-1,10] & x[m,1]^=x[i+1:m-1,8] then x[m,5]=fam_nr;
end; end; end; end; end; end; end; end; end;
end; *j kinni;
end; *i kinni;
A=B+1; *järgmisesse leibkonda;
end;
create work.tp_maj_v2_temp2 from x[colname=veerunimed]; /*tulemuste salvestamine*/
append from x;
finish;
run; quit;
```

Lisa 3. Algoritm – kuus kohustuslikku loendustunnust

```
proc iml; start; use tp_maj_v1_temp4; read all var _all_ into x;
veerunimed={"ID" "hh_nr" "nr" "hh_size" "fam_nr" "part_ID" "part_step" "mot_ID"
"mot_hh_nr" "fat_ID" "fat_hh_nr" "age" "sex" "birth" "nr_insidefam" "fam_size"
"fam_status_J" "type_of_fam_J" "hh_status_J" "type_of_hh_J" "fam_size_J"
"hh_size_J"}; *tähistusi vt tabel 11 ja lisa 1;
XL=nrow(x);
A=1;
do v= A to XL; if x[v,5]=0 then do;
/*kui isik on väljaspool tuumperekonda, siis tunnuseid "fam_status_J" (seisund
tuumperekonnas) ja "fam_type_J" ei moodustata (99)*/
x[v,17]=99; x[v,18]=99;end;end;
/* I: moodustatakse tunnused "fam_status_J" (SEISUNDTUUMPEREKONNAS x[,17]) ja
"fam_type_J" (TUUMPEREKONNA TÜÜP x[,18]) */
do until (A>XL);
B=A+x[A,16]-1; *A on esimene (noorim) liige ja B on viimane (vanim) liige peres;
do i=A to B-1;
if x[i,17]=0 then
do j=i+1 to B;
if x[j,17]=0 then do;
if x[i,6]=x[j,1] & (x[i,7]=1 |x[i,7]=3 | x[i,7]=4) then do; *abikaasa?;
x[i,17]=11; x[j,17]=11;
if x[i,16]=2 then x[A:B,18]=11; *kui perekonnas ei ole lapsi;
if x[i,16]>2 then
do s = i+1 to B;
if x[s,12]<25 then x[A:B,18]=12; *kui perekonnas on alla 25 aastaseid lapsi;
if x[s,18]=0 & x[s,12]>=25 then x[A:B,18]=13; *kui noorim laps vähemalt 25;
if (x[s,18]=12 | x[s,18]=13) & x[s,17]=0 then x[s,17]=41;
end; end;
if x[i,6]=x[j,1] & (x[i,7]=2 |x[i,7]=5 | x[i,7]=6) then do; *vabaabielupartner?;
x[i,17]=12; x[j,17]=12;
if x[i,16]=2 then x[A:B,18]=21; *kui perekonnas ei ole lapsi;
if x[i,16]>2 then
do s = i+1 to B;
if x[s,12]<25 then x[A:B,18]=22; *kui perekonnas on alla 25 aastaseid lapsi;
if x[s,18]=0 & x[s,12]>=25 then x[A:B,18]=23; *kui noorim laps vähemalt 25;
if (x[s,18]=22 | x[s,18]=23) & x[s,17]=0 then x[s,17]=41;
end; end;
if x[i,7]=0 then do; *kui isikul pole ei abikaasat ega vabaabielupartnerit;
if x[i,1]=x[j,10] then do; *üksikisade pered;
x[i,17]=30;x[j,17]=42;
if x[j,12]<25 then x[A:B,18]=31;
if x[j,18]=0 & x[j,12]>=25 then x[A:B,18]=32;
end;
if x[i,1]=x[j,8] then do; *üksikemade pered;
```

```
x[i,17]=30;x[j,17]=42;
if x[j,12]<25 then x[A:B,18]=41;
if x[j,18]=0 & x[j,12]>=25 then x[A:B,18]=42;
end; end; end; end; end;
A=B+1; *järgmisesse tuumperekonda;
end;
/* II: moodustatakse tunnus "hh_status_J" (SEISUND LEIBKONNAS x[,19])*/
do i=1 to XL;
if x[i,5]=0 then do; /*kui isik ei kuulu tuumperekonda*/
if x[i,4]=1 then x[i,19]=1210; *ei kuulu tuumperekonda, elavad üksinda;
if x[i,4]>1 then x[i,19]=1220; *ei kuulu tuumperekonda, elavad mitmekesi;
end;
if x[i,5]^=0 then do; *kui isik kuulub tuumperekonda;
if x[i,17]=11 then x[i,19]=1110;
if x[i,17]=12 then x[i,19]=1120;
if x[i,17]=30 then x[i,19]=1130;
if x[i,17]=41 then x[i,19]=1141;
if x[i,17]=42 then x[i,19]=1142;
end; end;
/* III: moodustatakse tunnus "hh_type_J" (TAVALEIBKONNA TÜÜP x[,20])*/
A=1;
do until (A>XL);
B=A+x[A,4]-1;
do i=A to B;
if x[i,20]=0 then do;
if x[i,5]>1 then x[A:B,20]=300; *mitmepereleibkonnad;
if x[i,5]=1 then do; *ühepereleibkond;
if x[i,18]=11 then x[A:B,20]=211;
if x[i,18]=12 then x[A:B,20]=212;
if x[i,18]=13 then x[A:B,20]=213;
if x[i,18]=21 then x[A:B,20]=221;
if x[i,18]=22 then x[A:B,20]=222;
if x[i,18]=23 then x[A:B,20]=223;
if x[i,18]=31 then x[A:B,20]=231;
if x[i,18]=32 then x[A:B,20]=232;
if x[i,18]=41 then x[A:B,20]=241;
if x[i,18]=42 then x[A:B,20]=242;
end; end;
if x[i,20]=0 & x[i,4]=1 then x[i,20]=110; *ei moodusta pereleibkonda, üksinda;
if x[i,20]=0 & x[i,4]>1 then x[i,20]=120; *ei moodusta pereleibkonda, mitmekesi;
end;
A=B+1; *järgmisesse tuumperekonda;
end;
/* IV: moodustatakse tunnused "fam_size_J" (TUUMPEREKONNA SUURUS x[,21]) ja
"hh_size_J" (TAVALEIBKONNA SUURUS x[,22])*/
do i= 1 to XL;
```

```
if x[i,5]=0 then x[i,21]=0;
if x[i,5]^=0 then do;
if x[i,16]>=11 then x[i,21]=11; else x[i,21]=x[i,16];
end;
if x[i,4]>=11 then x[i,22]=11; else x[i,22]=x[i,4];
end;
/*tulemuste salvestamine*/
create work.tp_maj_v1_temp5 from x[colname=veerunimed];
append from x;
finish;
run; quit;
```

Lisa 4. SAS *proc logistic* väljund M1 kohta

The LOGISTIC Procedure

Model Information	
Data Set	WORK.ANDMED
Response Variable	y
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	733733
Number of Observations Used	733733

Response Profile		
Ordered Value	**y**	**Total Frequency**
1	1	273896
2	0	459837

Probability modeled is y=1.

Class Level Information			
Class	**Value**	**Design Variables**	
child	**1**	-1	-1
	2	1	0
	3	0	1

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterin	**Intercept Only**	**Intercept and Covariates**
AIC	969533.39	132368.96
SC	969544.89	132438.00
-2 Log L	969531.39	132356.96

R-Square	0.6805	Max-rescaled R-Square	0.9281

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	837174.427	5	<.0001
Score	574481.880	5	<.0001
Wald	80659.6658	5	<.0001

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
m_age	1	19082.1331	<.0001
f_age	1	330.9040	<.0001
age_dif	1	48721.9959	<.0001
child	2	17839.4884	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-0.8988	0.0226	1588.6733	<.0001
m_age		1	0.1392	0.00101	19082.1331	<.0001
f_age		1	-0.0151	0.000832	330.9040	<.0001
age_dif		1	-0.3273	0.00148	48721.9959	<.0001
child	2	1	-1.2062	0.0146	6785.7590	<.0001
child	3	1	2.6720	0.0201	17640.6887	<.0001

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
m_age	1.149	1.147	1.152
f_age	0.985	0.983	0.987
age_dif	0.721	0.719	0.723
child 2 vs 1	1.297	1.244	1.351
child 3 vs 1	62.673	58.883	66.706

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	99.4	Somers' D	0.989
Percent Discordant	0.5	Gamma	0.990
Percent Tied	0.1	Tau-a	0.463
Pairs	125947514952	c	0.994

Partition for the Hosmer and Lemeshow Test					
		y = 1		**y = 0**	
Group	**Total**	**Observed**	**Expected**	**Observed**	**Expected**
1	74297	0	0.20	74297	74296.80
2	73497	4	4.03	73493	73492.97
3	73300	9	24.58	73291	73275.42
4	73458	35	108.23	73423	73349.77
5	73349	263	688.99	73086	72660.01
6	73299	3471	6293.54	69828	67005.46
7	73412	53380	48495.38	20032	24916.62
8	73398	72275	72676.69	1123	721.31
9	73573	72877	73466.14	696	106.86
10	72150	71582	72138.12	568	11.88

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
32679.0741	8	<.0001

Classification Table									
	Correct		**Incorrect**		**Percentages**				
Prob Level	**Event**	**Non-Event**	**Event**	**Non-Event**	**Correct**	**Sensi-tivity**	**Speci-ficity**	**False POS**	**False NEG**
0.000	274E3	0	46E4	0	37.3	100.0	0.0	62.7	.
0.020	274E3	364E3	96159	287	86.9	99.9	79.1	26.0	0.1
0.040	273E3	387E3	73001	538	90.0	99.8	84.1	21.1	0.1
0.060	273E3	4E5	59882	800	91.7	99.7	87.0	18.0	0.2

0.080	273E3	41E4	50230	1049	93.0	99.6	89.1	15.5	0.3
0.100	273E3	416E3	43379	1305	93.9	99.5	90.6	13.7	0.3
0.120	272E3	422E3	37917	1573	94.6	99.4	91.8	12.2	0.4
0.140	272E3	425E3	34534	1845	95.0	99.3	92.5	11.3	0.4
0.160	272E3	428E3	31390	2119	95.4	99.2	93.2	10.4	0.5
0.180	271E3	431E3	28862	2504	95.7	99.1	93.7	9.6	0.6
0.200	271E3	433E3	26606	2907	96.0	98.9	94.2	8.9	0.7
0.220	271E3	435E3	24551	3256	96.2	98.8	94.7	8.3	0.7
0.240	27E4	437E3	22555	3641	96.4	98.7	95.1	7.7	0.8
0.260	27E4	439E3	20809	4122	96.6	98.5	95.5	7.2	0.9
0.280	269E3	44E4	19416	4704	96.7	98.3	95.8	6.7	1.1
0.300	269E3	442E3	18199	5175	96.8	98.1	96.0	6.3	1.2
0.320	268E3	443E3	16692	5847	96.9	97.9	96.4	5.9	1.3
0.340	268E3	444E3	15577	6377	97.0	97.7	96.6	5.5	1.4
0.360	267E3	445E3	14679	7135	97.0	97.4	96.8	5.2	1.6
0.380	266E3	446E3	13575	7854	97.1	97.1	97.0	4.9	1.7
0.400	265E3	447E3	12732	8592	97.1	96.9	97.2	4.6	1.9
0.420	264E3	448E3	11880	9621	97.1	96.5	97.4	4.3	2.1
0.440	263E3	449E3	11019	10449	97.1	96.2	97.6	4.0	2.3
0.460	263E3	45E4	10264	11356	97.1	95.9	97.8	3.8	2.5
0.480	262E3	45E4	9714	12337	97.0	95.5	97.9	3.6	2.7
0.500	26E4	451E3	8949	13717	96.9	95.0	98.1	3.3	3.0
0.520	259E3	451E3	8346	14703	96.9	94.6	98.2	3.1	3.2
0.540	258E3	452E3	7912	15763	96.8	94.2	98.3	3.0	3.4
0.560	257E3	452E3	7425	17362	96.6	93.7	98.4	2.8	3.7
0.580	255E3	453E3	7013	18417	96.5	93.3	98.5	2.7	3.9
0.600	254E3	453E3	6701	19534	96.4	92.9	98.5	2.6	4.1
0.620	253E3	453E3	6369	21167	96.2	92.3	98.6	2.5	4.5
0.640	252E3	454E3	6073	22200	96.1	91.9	98.7	2.4	4.7
0.660	251E3	454E3	5839	23322	96.0	91.5	98.7	2.3	4.9
0.680	249E3	454E3	5570	24639	95.9	91.0	98.8	2.2	5.1
0.700	248E3	454E3	5374	25571	95.8	90.7	98.8	2.1	5.3
0.720	247E3	455E3	5146	26835	95.6	90.2	98.9	2.0	5.6
0.740	246E3	455E3	4959	28033	95.5	89.8	98.9	2.0	5.8
0.760	245E3	455E3	4796	28955	95.4	89.4	99.0	1.9	6.0
0.780	244E3	455E3	4606	30058	95.3	89.0	99.0	1.9	6.2
0.800	243E3	455E3	4414	31235	95.1	88.6	99.0	1.8	6.4
0.820	242E3	456E3	4255	32350	95.0	88.2	99.1	1.7	6.6
0.840	24E4	456E3	4089	33603	94.9	87.7	99.1	1.7	6.9
0.860	239E3	456E3	3906	35045	94.7	87.2	99.2	1.6	7.1
0.880	237E3	456E3	3707	36786	94.5	86.6	99.2	1.5	7.5
0.900	235E3	456E3	3507	38667	94.3	85.9	99.2	1.5	7.8
0.920	233E3	457E3	3281	41085	94.0	85.0	99.3	1.4	8.3
0.940	229E3	457E3	2989	44410	93.5	83.8	99.3	1.3	8.9
0.960	224E3	457E3	2646	50075	92.8	81.7	99.4	1.2	9.9
0.980	21E4	458E3	2204	63580	91.0	76.8	99.5	1.0	12.2
1.000	0	46E4	0	274E3	62.7	0.0	100.0	.	37.3

Lisa 5. SAS *proc logistic* väljund M2 kohta

Model Information	
Data Set	WORK.ANDMED
Response Variable	y
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	733733
Number of Observations Used	733733

Response Profile		
Ordered Value	**y**	**Total Frequency**
1	1	273896
2	0	459837

Probability modeled is y=1.

Class Level Information			
Class	**Value**	**Design Variables**	
child	1	-1	-1
	2	1	0
	3	0	1
hh_size	0	-1	
	1	1	

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	969533.39	98408.863
SC	969544.89	98489.404
-2 Log L	969531.39	98394.863

R-Square	0.6949	Max-rescaled R-Square	0.9478

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	871136.525	6	<.0001
Score	586088.395	6	<.0001
Wald	84509.9519	6	<.0001

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
m_age	1	14733.4515	<.0001
f_age	1	3065.2678	<.0001
age_dif	1	46469.8668	<.0001
child	2	29661.1099	<.0001
hh_size	1	24529.9123	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	1.8572	0.0303	3767.3766	<.0001
m_age		1	0.1236	0.00102	14733.4515	<.0001
f_age		1	-0.0504	0.000910	3065.2678	<.0001
age_dif		1	-0.3285	0.00152	46469.8668	<.0001
child	2	1	-0.8971	0.0151	3535.4626	<.0001
child	3	1	3.5469	0.0213	27659.6026	<.0001
hh_size	1	1	1.8557	0.0118	24529.9123	<.0001

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
m_age	1.132	1.129	1.134
f_age	0.951	0.949	0.953
age_dif	0.720	0.718	0.722
child 2 vs 1	5.770	5.495	6.060
child 3 vs 1	491.160	457.598	527.184
hh_size 1 vs 0	40.912	39.055	42.857

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	99.5	Somers' D	0.992
Percent Discordant	0.3	Gamma	0.993
Percent Tied	0.1	Tau-a	0.464
Pairs	125947514952	c	0.996

Partition for the Hosmer and Lemeshow Test					
		y = 1		**y = 0**	
Group	**Total**	**Observed**	**Expected**	**Observed**	**Expected**
1	72403	0	0.13	72403	72402.87
2	73568	3	3.22	73565	73564.78
3	73416	8	22.10	73408	73393.90
4	73406	39	100.48	73367	73305.52
5	73378	161	510.96	73217	72867.04
6	73340	1428	3235.48	71912	70104.52
7	73385	53379	49880.29	20006	23504.71
8	73367	72708	72815.03	659	551.97
9	73518	72720	73383.00	798	135.00
10	73952	73450	73945.22	502	6.78

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
41575.0215	8	<.0001

Classification Table									
	Correct		**Incorrect**		**Percentages**				
Prob Level	**Event**	**Non-Event**	**Event**	**Non-Event**	**Correct**	**Sensi-tivity**	**Speci-ficity**	**False POS**	**False NEG**
0.000	274E3	0	46E4	0	37.3	100.0	0.0	62.7	.
0.020	274E3	378E3	81375	291	88.9	99.9	82.3	22.9	0.1
0.040	273E3	405E3	54427	554	92.5	99.8	88.2	16.6	0.1

0.060	273E3	421E3	39195	831	94.5	99.7	91.5	12.6	0.2
0.080	273E3	43E4	30311	1125	95.7	99.6	93.4	10.0	0.3
0.100	272E3	435E3	24540	1440	96.5	99.5	94.7	8.3	0.3
0.120	272E3	44E4	20331	1782	97.0	99.3	95.6	7.0	0.4
0.140	272E3	442E3	17492	2110	97.3	99.2	96.2	6.0	0.5
0.160	271E3	445E3	15038	2521	97.6	99.1	96.7	5.3	0.6
0.180	271E3	447E3	13291	2942	97.8	98.9	97.1	4.7	0.7
0.200	271E3	448E3	11996	3344	97.9	98.8	97.4	4.2	0.7
0.220	27E4	449E3	11109	3706	98.0	98.6	97.6	3.9	0.8
0.240	27E4	449E3	10389	4136	98.0	98.5	97.7	3.7	0.9
0.260	269E3	45E4	9863	4453	98.0	98.4	97.9	3.5	1.0
0.280	269E3	45E4	9401	4850	98.1	98.2	98.0	3.4	1.1
0.300	269E3	451E3	9071	5117	98.1	98.1	98.0	3.3	1.1
0.320	269E3	451E3	8775	5374	98.1	98.0	98.1	3.2	1.2
0.340	268E3	451E3	8510	5635	98.1	97.9	98.1	3.1	1.2
0.360	268E3	452E3	8257	5950	98.1	97.8	98.2	3.0	1.3
0.380	268E3	452E3	7994	6213	98.1	97.7	98.3	2.9	1.4
0.400	267E3	452E3	7796	6492	98.1	97.6	98.3	2.8	1.4
0.420	267E3	452E3	7551	6730	98.1	97.5	98.4	2.7	1.5
0.440	267E3	452E3	7339	7012	98.0	97.4	98.4	2.7	1.5
0.460	267E3	453E3	7128	7300	98.0	97.3	98.4	2.6	1.6
0.480	266E3	453E3	6941	7602	98.0	97.2	98.5	2.5	1.7
0.500	266E3	453E3	6742	7898	98.0	97.1	98.5	2.5	1.7
0.520	266E3	453E3	6563	8252	98.0	97.0	98.6	2.4	1.8
0.540	265E3	453E3	6392	8623	98.0	96.9	98.6	2.4	1.9
0.560	265E3	454E3	6167	9068	97.9	96.7	98.7	2.3	2.0
0.580	264E3	454E3	5947	9491	97.9	96.5	98.7	2.2	2.0
0.600	264E3	454E3	5754	9952	97.9	96.4	98.7	2.1	2.1
0.620	263E3	454E3	5541	10456	97.8	96.2	98.8	2.1	2.2
0.640	263E3	454E3	5359	10951	97.8	96.0	98.8	2.0	2.4
0.660	262E3	455E3	5156	11474	97.7	95.8	98.9	1.9	2.5
0.680	262E3	455E3	4992	12054	97.7	95.6	98.9	1.9	2.6
0.700	261E3	455E3	4813	12726	97.6	95.4	99.0	1.8	2.7
0.720	26E4	455E3	4639	13449	97.5	95.1	99.0	1.7	2.9
0.740	26E4	455E3	4472	14211	97.5	94.8	99.0	1.7	3.0
0.760	259E3	456E3	4314	14956	97.4	94.5	99.1	1.6	3.2
0.780	258E3	456E3	4137	15928	97.3	94.2	99.1	1.6	3.4
0.800	257E3	456E3	3993	17040	97.1	93.8	99.1	1.5	3.6
0.820	256E3	456E3	3814	18308	97.0	93.3	99.2	1.5	3.9
0.840	254E3	456E3	3643	19854	96.8	92.8	99.2	1.4	4.2
0.860	252E3	456E3	3447	21704	96.6	92.1	99.3	1.3	4.5
0.880	25E4	457E3	3224	24026	96.3	91.2	99.3	1.3	5.0
0.900	247E3	457E3	2999	27018	95.9	90.1	99.3	1.2	5.6
0.920	243E3	457E3	2755	31231	95.4	88.6	99.4	1.1	6.4
0.940	237E3	457E3	2531	36827	94.6	86.6	99.4	1.1	7.5
0.960	231E3	458E3	2286	43102	93.8	84.3	99.5	1.0	8.6
0.980	22E4	458E3	1987	53510	92.4	80.5	99.6	0.9	10.5
1.000	0	46E4	0	274E3	62.7	0.0	100.0	.	37.3

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Kairiin Kütt

Sünnikuupäev: 29.03.1989

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

Leibkonnad ja perekonnad registripõhises rahva ja eluruumide loenduses,

mille juhendaja on Mare Vähi,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **19.05.2014**