

TARTU ÜLIKOOL
LOODUS- JA TÄPPISTEADUSTE VALDKOND
MATEMAATIKA JA STATISTIKA INSTITUUT

Elis Sams
**4MOST vaatlusprogrammi tunnuste
tihedusfunktsioonide taastamine
vaatlusandmetelt**
Matemaatiline statistika
Bakalaureusetöö (9 EAP)

Juhendajad: Prof Elmo Tempel
MSc Anne Selart

TARTU 2022

**4MOST VAATLUSPROGRAMMI TUNNUSTE
TIHEDUSFUNKTSIOONIDE TAASTAMINE
VAATLUSANDMETELT**

Bakalaureusetöö

Elis Sams

Lühikokkuvõte

4MOST (4-metre Multi-Object Spectroscopic Telescope) on multifibriline spektroskoopiline uuringuseade, mis on võimeline vaatlama tuhandeid astronoomilisi objekte samaaegselt. 4MOST-i abil plaanitakse vaadelda suur osa lõunataevast selle esimeste tööaastate jooksul.

Vaatlusandmete statistilises analüüsis on oluline teada huvipakkuvate tunnuste jaotust, kuid vaatluste käigus ei vaadelda enamasti kõiki sihtkataloogis olevaid objekte. Seetõttu ei kirjelda vaatlusandmete pealt saadud pideva tunnuse tihedusfunktsiooni hinnang kõigi kataloogis olevate objektide vastava tunnuse tihedust, vaid ainult konkreetsete vaatlusandmete oma. Veelgi enam, vaatlus-simulatsioonidest tuletatud objekti edukalt vaatlemise tõenäosused on nihkega. Bakalaureusetöö eesmärgiks on taastada valitud tunnuste tihedusfunktsioonid simuleeritud vaatlusandmetelt. See saavutatakse korrigerides esmalt objektide edukalt vaatlemise tõenäosused, misjärel taastatakse valitud tunnuste tihedusfunktsioonid kaalutud tuumameetodi abil kasutades korrigeeritud tõenäosuseid.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: Astronoomilised vaatlused, mitteparameetriline statistika, statistiline hindamine, valikumeetod, tõenäosusjaotused.

RECONSTRUCTING DENSITY FUNCTIONS FROM THE 4MOST SURVEY OBSERVATIONAL DATA

Bachelor thesis

Elis Sams

Abstract

4MOST (4-metre Multi-Object Spectroscopic Telescope) is a multi-object spectroscopic survey facility capable of observing thousands of astronomical objects simultaneously. 4MOST is to survey a significant fraction of the southern sky in its first years of operation.

In the statistical analysis of survey output data, it is valuable to know the distributions of the variables under investigation. However, not all objects in the target catalogue are successfully observed during observations. This means that the density function of a continuous variable estimated solely from observational data produces a biased estimate. Furthermore, the probabilities of observing an object successfully deduced from simulations, appear to be biased. The aim of this bachelor's thesis is to reconstruct selected variables' unbiased density function estimates from observational data. This is achieved by first correcting the biased successful observation probabilities and then reconstructing the variables' density functions using weighted kernel density estimation with the corrected probabilities.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics.

Key Words: Astronomical observations, non-parametric statistics, statistical estimation, sampling method, probability distributions.

Sisukord

Sissejuhatus	4
1 4MOST vaatlusprogramm	6
2 4MOST vaatluskataloog	8
2.1 Astronoomiliste terminite lühiseletused	8
2.2 Vaatluste simulatsioonid	10
2.3 Kasutatavate tunnuste kirjeldus	11
3 Matemaatiline taust	15
3.1 Horwitz-Thompsoni kogusumma hinnang	15
3.2 Tuumameetod tihedusfunktsiooni hindamiseks	16
3.2.1 Kaalutud tuumameetod	19
3.3 Nadaraya-Watsoni tuumaregressiooni hinnang	20
4 Probleemi püstitus	22
5 Metoodika	24
5.1 Kaasamistõenäosuste korrigeerimine	24
5.2 Taastatud tihedusfunktsiooni täpsuse hindamine	29
6 Tulemused	31
Kokkuvõte	35
Kasutatud allikad	37
Lisa 1. Taastatud tihedusfunktsioonid	39

Sissejuhatus

Astronoomilistel vaatlustel on tänapäevani oluline osa vaadeldava universumi objektide omaduste, tekke ja evolutsiooni kirjeldamisel. Astronoomiliste objektide vaatlemiseks kasutatakse teleskoopi – uuringuseadet, mille abil kogutakse huvipakkuvatelt objektidelt kiirgust. Kiirgusspekter annab teavet objekti füüsikaliste ja keemiliste omaduste kohta, mille kaudu on võimalik kinnitada või ümber lükata tehtud oletusi ja teooriaid ja püstitada uusi küsimusi. Astronoomiliste objektide hulka kuuluvad teiste seas näiteks tähed, supernoovad, galaktikad ja kvasarid.

Viimastel aastatel on planeeritud ja ehitatud mitmeid multifibriilisi spektrograafilisi uuringuseadmeid, mis on võimelised vaatlema palju enam astronoomilisi objekte korraga, kui varasemad uuringuseadmed. Veel enam, uued seadmed vaatlevad objekte palju täpsemalt, andes väljundiks märkimisväärselt rohkem informatsiooni kui kunagi varem. Üheks selliseks uuringuseadmeks on 4MOST, mille taust ja spetsiifika on aluseks käesolevale tööle.

Võime paljusid objektide samaaegselt vaadelda toob endaga kaasa keerulisi vaatlusstrateegilisi probleeme. Ühe vaatlusuuringu objektid ei täida enam kõiki valgust koguvaid fiibreid, mistõttu ühendatakse kõikide planeeritud vaatlusuuringute objektid ja vaadeldakse neid korraga, nii et ühel vaatlemisel oleksid kõik fiibrid täidetud. Vaatluste optimeerimiseks arendatakse välja algoritmid, mis määravad, millised objektid vaatlusesse valitakse ja millisel viisil neid vaadeldakse. Päris vaatlusi jäljendatakse simulatsioonide abil, et selle varal välja töötada vaatlusalgoritmid ja teha ettevalmistusi vaatlusandmete töötlemiseks.

Vaatluste käigus ei vaadelda edukalt kõiki planeeritavaid objekte. See tähendab, et pärast vaatlusi ei ole kõikide algselt kataloogis olevate objektide kohta vaatlusandmeid, vaid ainult mingil valimil ehk nendel objektidel, mis said edukalt vaadeldud. Seetõttu vaatlusandmete pealt hinnatud tunnuste tihedusfunktsioonid ei kirjelda kõigi kataloogis olevate objektide vastava tunnuse jaotust, vaid ainult vaatlusandmete valimi oma. Selle põhjuseks on see, et vaatlemata jäävad objektid ei jää

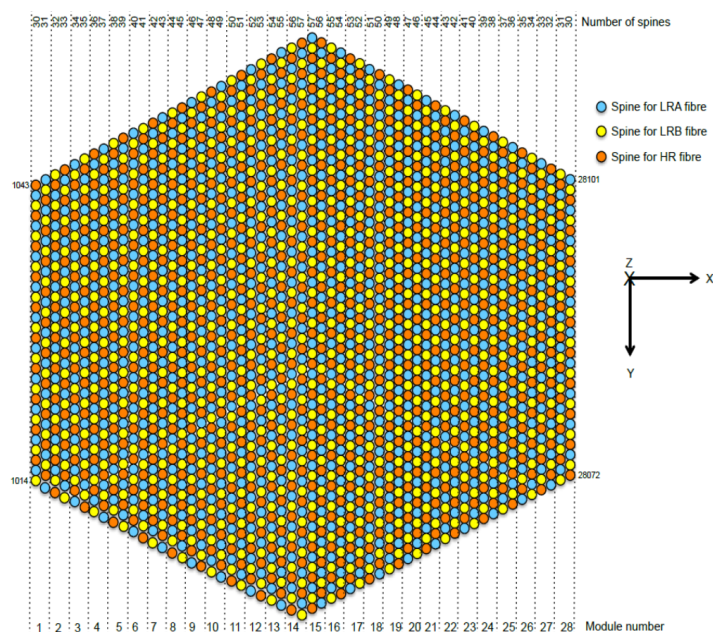
vaatlemata juhuslikult, mistõttu ei ole tiheduse hinnang vaadeldud objektide pealt kõigi kataloogis olevate objektide suhtes korrektne. Bakalaureusetöö eesmärgiks on taastada ühe vaatlusuuringu kataloogi simuleeritud vaatlusvalimi pealt tähesuuruse ja punanihke tihedusfunktsioonid.

Töö jaguneb kuueks osaks. Esimeses kahes osas tutvustatakse 4MOST vaatlusprogrammi, vaatlusandmete simuleerimisloogikat ja vaatluskataloogi. Kolmandas osas tehakse ülevaade lahenduse matemaatilisest taustast. Neljandana kirjeldatakse detailselt probleemi ja viiendana metoodikat selle lahendamiseks. Kuuendana esitatakse tulemused ja esitatakse taastatud tihedusehinnangud tähesuurusele ja punanihkele.

Autor tänab juhendajaid Elmo Templit ja Anne Selartit suunamise, tähelepanelike märkuste ja pühendatud aja eest.

1 4MOST vaatlusprogramm

4MOST (4-metre Multi-Object Spectroscopic Telescope) on laia vaateväljaga optiline multi-fiibriline spektroskoopiline uuringuseade, mis paigaldatakse ESO (European Southern Observatory) teleskoobile VISTA ja mida kasutatakse ESO Paranal Observatooriumis Tšiilis. 4MOST uuringuseadme abil plaanitakse vaadelda suur osa lõunataevast selle esimese 5 tööaasta jooksul. ([4MOST Science Overview n.d](#))



Joonis 1: 4MOST vaatevälja 2436 optilise fiibri asetuse fokaaltasandil ja nende kaardistus kolmele spektrograafile (Allikas: [4MOST Subsystems n.d](#)).

4MOST-i uuringuseadmel on lai 4,2 ruutkaarekraadine vaateväli koos 2436 optilise fiibriga, nendest 812 edastavad kõrge lahutusvõimega spektrograafide ja 1624 kahele madala lahutusvõimega spektrograafide (De Jong *et al.*, 2019). Kuusnurkset vaatevälja kujutab Joonis 9, kus on näha fiibrite asetuse ning millisele spektrograafidele fiber andmeid edastab. Optiliste fiibrite asukohad fokaaltasandil on konfigureeritavad fikseeritud raadiuses, et valgust koguda vaadelda plaanitavatelt astronoomilistelt objektidelt (Guiglion *et al.*, 2019). Astronoomiline objekt võib olla nii

galaktika, täht, kvasar kui ka supernoova.

4MOST-i esimese 5 aasta vaatlusprogrammi kaasati vaatlusuuringud läbi valiku-
protsessi. 70% vaatlusajast jagati 4MOST konsortiumi insituutide uuringutele ning
ülejäänu jagati avalikele uuringutele. Omavahel konkureerivatest konsortiumi vaat-
lusuuringutest valiti vaatlusprogrammi 10, mis hõlmavad väga erisuguseid sihtob-
jekte nii Linnutee galaktikas kui ka väljaspool. Vaatlusprogrammi esitatud uuringu
mahu hindamiseks esitati lisaks ka täpne sihtobjektide kataloog selliste hinnatud
parameetritega nagu objekti koordinaadid taevas, objekti heledus ning vajalik sum-
maarne säriaeg. (De Jong *et al.*, 2019)

4MOST uuringuseade võimaldab vaadelda tuhandeid objekte samaaegselt, mis tä-
hendab, et üksik vaatlusuuring ei täida kõiki fiibreid. Seetõttu vaadeldakse mitmete
uuringute objekte üheaegselt, mis tõstatab vajaduse efektiivse vaatlusstrateegia jä-
rele. (Tempel *et al.*, 2020b) Esimene vaatlusstrateegia optimeerimise väljakutse
oli leida teleskoobi suuna ja vaatlusaja muster ehk plaadistumuster (ingl *tiling
pattern*) sõltuvalt vaatlusi mõjutavatest faktoritest, mille leidmise algoritmi on
kirjeldatud artiklis Tempel *et al.*, 2020a. Teine väljakutse oli leida igale fiibrile
astronoomiline objekt, mida vaadelda, eeldades, et plaadistumuster on eelnevalt
fikseeritud. Fiiber-objekt paaride määramise algoritmi tutvustatakse artiklis Tem-
pel *et al.*, 2020b. Mainitud algoritmid on käesolevas töös olulised, sest need on
kasutatavas andmestikus aluseks kolmele valikutunnusele - nende tunnuste simu-
leerimisprotsessi tutvustatakse alapeatükis 2.2.

4MOST avaldab suurt mõju mitmele kaasaegse astrofüüsika võtmevaldkonnale
([4MOST Science Overview n.d](#)). 4MOST on praegu tootmise, integratsiooni ja
katsefaasis ning teadustegevus algab 2024. aasta esimeses kvartalis ([4MOST Over-
view n.d](#)).

2 4MOST vaatluskataloog

Lõputöös kasutatakse vaatluskataloogi 1 914 489 astronoomilise objektiga ning etteantud 36-st tunnusest peamiselt 7-t. Kataloog on pärit ühest vaatlusuuringust 10-ne seast, mis on valitud 4MOST vaatlusprogrammi ja mille objektideks on kvasarid. Kuigi vaatlusuuringud on oma eesmärkide ja kirjeldavate tunnuste poolest erinevad, piisab probleemi lahendamiseks ühest näitekataloogist, sest objektide vaatlemise simuleerimisloogika on sama. Simuleerimist on vaja selleks, et mängida läbi objektide valik vaatluseks kõikide objektide kataloogist ning vaatlusprotsessi. Alapeatükis 2.2 tehakse ülevaade vaatluste simulatsiooni tööprotsessist, mida rakendatakse töös kasutatava kataloogi peal, aga mida on üldiselt võimalik rakendada ka suvalise 4MOST kataloogi peal. Alapeatükis 2.3 kirjeldatakse kataloogist kasutatavaid tunnuseid.

2.1 Astronoomiliste terminite lühiseletused

Alapeatükis selgitatakse lühidalt töös kasutatavaid astronoomilisi termineid. Ülevaade terminitest on kirjutatud Tenjes ja Reivelt, 2020 põhjal.

Objekti absoluutne heledus L on on ajaühikus astronoomiliselt objektilt kiiratud koguerenergia ehk energia kiirgamise kiirus kogu lainepikkuste vahemikus. Objekti vaadeldes mõõdetakse aga näivat heledust l , mis on tähe poolt tekitatud ja Maa peal registreeritud kiirgusvoog ning sõltub objekti kaugusest. Objekti absoluutne heledus on leitav näiva heleduse ja objekti kauguse kaudu.

Absoluutne ja näiv **tähesuurus** on vastavalt objekti absoluutne ja näiv heledus logaritmilisel skaalal. Kui kahe objekti näivad heledused on vastavalt l_1 ja l_2 , siis nende heleduste erinevus tähesuurustes on $m_1 - m_2 = -2,5 \log \frac{l_1}{l_2}$, kus m_1 ja m_2 on kahe objekti näivad tähesuurused. Absoluutne tähesuurus, mis on kasutusel ka käesolevas töös, on 10 parseki ehk ligikaudu 32,62 valgusaasta kaugusel asuva objekti näiv tähesuurus. Absoluutsed tähesuurused on ligikaudu vahemikus -10

kuni +15, kus väiksema väärtusega tähesuurus märgib heledamaid tähti ja suurema väärtusega tähesuurus nõrgemaid tähti.

Punanihe on astronoomilise objekti elektromagnetkiirguse spektri nihkumine pikema lainepikkuse poole. Punanihke abil on võimalik määratleda objekti kaugus.

Ekvaatorilised koordinaadid on tähistaeva pöörlemisest ja vaateleja asukohast sõltumatud astronoomilise objekti asukohta määravad koordinaadid. Ekvaatoriliste koordinaatide esimeseks koordinaadiks on otsetõus, mis on nurk piki taevaekvaatorit, täpsemalt nurk taevaekvaatorile projekteeritud objekti suuna ja nullsuuna vahel. Nullsuunaks on võetud Päikese asukoht kevadisel pööripäeval. Teiseks koordinaadiks on objekti nurkkaugus taevaekvaatori tasandist kas lõuna -või põhjapooluse suunas. Seda nurkkaugust nimetatakse käändeks, taevaekvaatoril on kääne 0 kraadi. Seega objekti ekvaatorilised koordinaadid on (otsetõus, kääne) või vastavate tähistustega (RA, Dec).

Kvasarid on kaugel asuvad aktiivsete galatikate sarnased väga võimsad ja heledad objektid, mille keskmes asub ülimassiivne must auk. Kvasaritele kui ka aktiivsetele galaktikatele on iseloomulik kiire muutlikkus ja erinev kiirguse iseloom ülejäänud galaktikatest.

2.2 Vaatluste simulatsioonid

Vaatluste simulatsioon tähendab vaatlusprotsessi läbimängimist, mis koosneb astronoomiliste objektide vaatlusesse valimisest kogu kataloogist ja vaatlusest endast, andes väljundiks, millised objektid said kõikide vaadeldute hulgast edukalt vaadeldud. Edukalt vaatlemine tähendab nõutud signaali-müra suhte saavutamist.

Esimene osa vaatlusrealisatsiooni simuleerimisloogikast toetub optimaalsele plaadistumustri leidmise algoritmile ehk reeglitele, kuidas teleskoopi üle vaadeldava taevaosa suunata. Ühe „plaadi“ all mõeldakse teleskoobi vaatevälja orientatsiooni ja keskpunkti. Mitu plaati samade ekvaatoriliste koordinaatide ja vaatevälja pöördenurgaga moodustavad ühe vaatlusploki. Optimaalne plaadistumuster on modelleeritud märgitud punkti protsessina, mille väljundiks on iga vaatlusploki ekvaatorilised koordinaadid, vaatevälja pöördenurk, taeva valgusolud vaatlemise ajal ja säriaeg (objektilt valguse kogumise aeg). (Tempel *et al.*, 2020a)

Teine osa simulatsiooniprotsessist on vaadeldavate objektide ja teleskoobi fiibrite paaride moodustamine. Ülevaade vaatluskataloogis olevatele astronoomilistele objektidele fiibrite määramise algoritmist on kirjutatud Tempel *et al.*, 2020b põhjal.

Optimaalse plaadistumustri leidmine koos eelnevalt mainitud parameetritega enne fiiber-objekt paaride määramist on oluline, sest paaride määramisel võetakse arvesse ka kõiki saadavaid vaatevälju (plaate) ning see mõjutab fiiber-objekt paaride tõenäosuseid. Kui optimaalne plaadistus on leitud ja fikseeritud, määratakse vastav fiiber-objekt konfiguratsioon ehk tõenäosused igale fiiber-objekt paarile, et antud objekt vaadeldakse antud fiibri abil. Tõenäosused määratakse selliselt, et ühtede vaatlusuuringute objekte ei eelistata üle teiste, samamoodi ei anta eelist ka lühema säriajaga objektidele pikema säriajaga objektide ees. Oluline on märkida, et igas vaatlusuuringu kataloogis ei ole vaja vaadelda kõiki objekte, vaid igal vaatlusuuringul on määratud edukalt vaadeldud objektide suhte osakaal. Seetõttu on oluline küsimus, kuidas teadusliku väljundi maksimeerimise ja iga vaatlusuuringu edukalt vaadeldud objektide osakaalu saavutamise eesmärgil valida vaatlemiseks

õiged sihtobjektid.

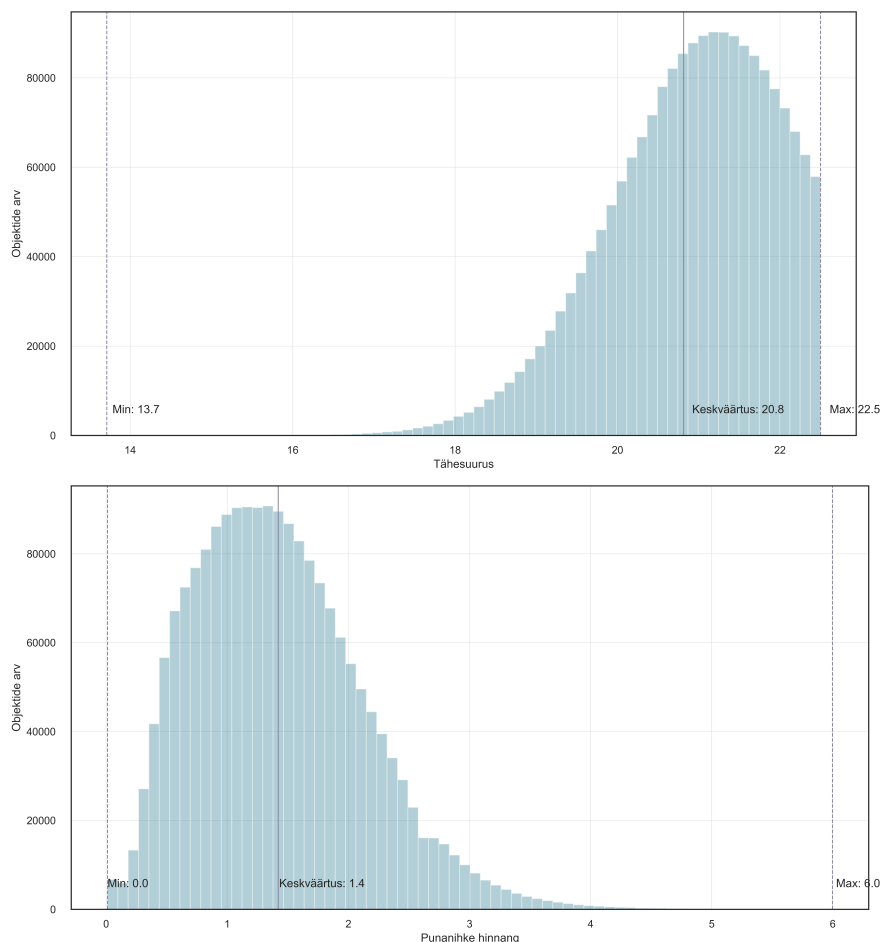
Tõenäosuslik lähenemine on kasulik, sest määratud tõenäosuste abil on lihtne arvutada sihtobjektide selektsioonifunktsiooni ehk eeskirja, millised kindla vaatlusuringu sihtobjektid vaatlusesse kaasatakse. Lisaks on hõlbus vaatlusi leitud tõenäosuste pealt simuleerida ehk leida, millised objektid vaatlusesse kaasatakse ja seejärel ka, millised objektid edukalt vaadeldakse. Tehes palju vaatlussimulatsioone saab hinnata, millised on üksikute objektide vaatlemise tõenäosused ja edukalt vaatlemise tõenäosused tingimusel, et nad on vaadeldud. Töös kasutatavad objektide kaasamistõenäosused ja edukalt vaatlemise tõenäosused on arvutatud ligikaudu 200 simulatsiooni pealt ning neid tunnuseid kirjeldatakse detailselt järgmises alapeatükis.

2.3 Kasutatavate tunnuste kirjeldus

Alapeatükis antakse töös kasutatavate tunnuseliikide ja üksikute tunnuste ülevaade. Tunnuseliikide hulka kuuluvad esiteks astronoomilise objekti ehk praegusel juhul kvasari omadusi kirjeldavad tunnused ja teiseks valikutunnused.

Astronoomilise objekti omadusi kirjeldavate tunnuste hulgas on olulisimad **tähesuurus** ehk heledus ning **punanihke hinnang**. Mõlemad tunnused on pidevad, Joonis 2 kujutab mõlema tunnuse jaotuseid koos keskmise, minimaalse ja maksimaalse väärtusega. Punanihke hinnangu tunnusele viidatakse edaspidi kui lihtsalt punanihkele. Objekti asukoha ekvaatorilistes koordinaatides määravad tunnused **RA** ehk otsetõus ja **Dec** ehk kääne. Oluline on märkida, et tähesuuruse ning koordinaatide tegelikud väärtused on iga kataloogi objekti jaoks teada, samas kui punanihe on hinnatud suurus ning selle suuruse tegelik väärtus selgub alles peale vaatlusi.

Lisaks astronoomilise objekti omadusi kirjeldavatele tunnustele on kataloogis 3 tunnust, mis kirjeldavad objekti edukalt vaatlemist ehk sisuliselt valimit: vaadeldud objekti täielikkus, objekti vaatlemise tõenäosus ning objekti edukalt vaatlemise



Joonis 2: Kvasari tähesuuruse ja punanihke histogrammid kogu kataloogi andmete põhjal. Vertikaalsed jooned vasakult paremale märgistavad vastavalt minimaalselt, keskmist ja maksimaalset väärtust.

tõenäosus tingimusel, et objekt on vaadeldud. Olgu kataloogi objektide arv N ning $i \in \{1, \dots, N\}$.

Vaadeldud objekti täielikkus on pidev arvuline tunnus vahemikus $(0 \dots 18,74)$, ning kirjeldab ühe simuleeritud vaatluse käigus saavutatud i -nda objekti täielikkust. Olemuselt on tunnus suhe

$$frac_i = \frac{(\text{vaadeldud säriaeg})_i}{(\text{nõutud säriaeg})_i},$$

kus $frac_i$ on i -nda objekti täielikkus, vaadeldud säriaeg on vaatluse käigus saavutatud säriaeg ning nõutud säriaeg on enne vaatlusi paika pandud säriaeg. Täielikkuse väärtus 1,0 ja üle tähendab, et objekt on täielikult ehk edukalt vaadeldud, sest nõutud säriaeg on saavutatud. Kui täielikkus on alla 1,0, tähendab see objekti aläsäritatust ning seega ka mitte-täielikkust. Kuigi tunnus on pidev arvuline tunnus, saab sellest lihtsalt tuletada valikuvektori

$$r_i = \begin{cases} 1, & frac_i \geq 1,0 \\ 0, & \text{muidu.} \end{cases}$$

Selguse huvides on oluline märkida, et objekti täielikult vaatlemine ja edukalt vaatlemine on sünonüümsed.

Objekti vaatlemise tõenäosuse hinnang ehk $\hat{P}(\text{vaadeldud})$ on tõenäosus, et objekt on vaadeldud. Ühe vaatlusuringu teaduslike eesmärkide saavutamiseks ei ole nõutud kõikide objektide vaatlemine, vaid mingi ettemääratud edukalt vaadeldud objektide osakaalu saavutamine kõikidest kataloogi objektidest. Seetõttu on vajalik tõenäosus, et objekt on üldse fiiber-objekt algoritmi poolt vaatluseks valitud. See ongi objekti vaatlemise tõenäosuse hinnang. Siinkohal on oluline teha vahet vaadeldud objektil ja edukalt vaadeldud objektil. Objekti vaatlemise tõenäosus ei ole sama, mis objekti edukalt vaatlemise tõenäosus. Objekti vaatlemise tõenäosuse hinnang on leitud ligikaudu 200 $frac$ tunnuse simulatsiooni pealt

$$\hat{P}(\text{vaadeldud})_i = \frac{\sum_{j=1}^s I(frac_{ji} > 0)}{s}, \quad i \in 1, \dots, N,$$

kus s on simulatsioonide arv ja $frac_{ji}$ on j -nda simulatsiooni i -da objekti täielikkus.

Objekti edukalt vaatlemise tõenäosuse hinnang tingimusel, et objekt on vaadeldud ehk $\hat{P}(\text{edukalt vaadeldud}|\text{vaadeldud})$. Objekti edukalt vaatlemine tähendab valitud objekti jaoks määratud signaali-müra suhte saavutamist (Walcher *et al.*, 2019) ning sellele toetudes on määratud vajalik säriaeg. Tõenäosuse hinnang

on leitud samuti *frac* tunnuse simulatsioonide pealt.

$$\hat{P}(\text{edukalt vaadeldud}|\text{vaadeldud}) = \frac{\sum_{j=1}^s I(\text{frac}_{ji} \geq 1)}{\sum_{j=1}^s I(\text{frac}_{ji} > 0)}, \quad i \in 1, \dots, N,$$

kus s on simulatsioonide arv ja frac_{ji} on j -nda simulatsiooni i -nda objekti täielikkus. Edukalt vaadeldud objektide osakaal on võetud nende objektide hulgast, millel on registreeritud mingisugune nullist suurem säriaeg, seega on tegemist tingliku tõenäosuse hinnanguga.

Viimase kahe tõenäosuse abil on võimalik leida objekti edukalt vaatlemise tõenäosuse hinnang

$$\begin{aligned} \hat{\pi}_i &= \hat{P}(\text{edukalt vaadeldud} \cap \text{vaadeldud})_i = \\ &= \hat{P}(\text{edukalt vaadeldud}|\text{vaadeldud})_i \cdot \hat{P}(\text{vaadeldud})_i. \end{aligned} \quad (1)$$

3 Matemaatiline taust

Matemaatilise tausta peatükis tutvustatakse töö põhiprobleemi lahendamiseks vajalikke teadmisi kasutatavatest meetoditest ning esitatakse nende matemaatilised definitsioonid.

3.1 Horwitz-Thompsoni kogusumma hinnang

Ülevaade valikuuringute teooria mõistetest on kirjutatud Traat ja Inno, 1997 põhjal.

Olgu vaatluse all N objektist koosnev üldkogum ning olgu tunnuse y väärtused üldkogumi objektidel $y_1, \dots, y_N$.

Üldkogumi kogusummaks t nimetatakse tunnuse y kõikide väärtuste summat

$$t := \sum_{i=1}^N y_i. \quad (2)$$

Valikudisain on tõenäosusjaotus $p(s)$ kõigi antud valiku korral võimalike valimite hulgal $S = \{s_1, \dots, s_M\}$. Iga valim s_k $k \in \{1, \dots, M\}$ realiseerub tõenäosusega $p(s_k)$. Antud valikudisaini $p(s)$ üldkogumi objekti i kaasamistõenäosuseks π_i nimetatakse tõenäosust, millega see objekt kaasatakse valimisse.

Kui üldkogumist on võetud mingi valikudisaini järgi valim valimimahu ehk objektide arvuga n , siis oluliseks muutub tunnuse y väärtustel $y_1, \dots, y_n$ defineeritud statistiku võimalikult täpne hindamine. See tähendab sellise valimi pealt leitud statistiku väärtuse leidmist, mis ühtiks kõikide üldkogumi väärtuste pealt leitud tegeliku statistiku väärtusega. Kui hinnang on nihketa, tähendab see, et $E(\hat{\theta}) = \theta$, kus $\hat{\theta}$ on valimi pealt hinnatud parameeter ja θ on tegelik parameeter. Järgnev Horwitz-Thompsoni kogusumma hinnang on nihketa hinnang üldkogumi kogusummale.

Kogusumma (2) Horwitz-Thompsoni hinnanguks nimetatakse summat

$$\hat{t} = \sum_{i=1}^n \frac{y_i}{\pi_i}. \quad (3)$$

Kaasamistõenäosuse pöördväärtust

$$\omega_i = \frac{1}{\pi_i} \quad (4)$$

nimetatakse ka laiendusteguriks, mis näitab, mitmele üldkogumi objektile väärtus y_i laiendatakse või teisiti öeldes, mitut objekti väärtus y_i üldkogumis esindab. Töös viidatakse laiendustegurile ka kui objekti kaalule. Kaalude abil on võimalik leida hinnang üldkogumi objektide arvule

$$\hat{N} = \sum_{i=1}^n \omega_i. \quad (5)$$

3.2 Tuumameetod tihedusfunktsiooni hindamiseks

Ülevaade ühemõõtmelisest tuumameetodist on kirjutatud Hollander, Wolfe ja Chicken, 2014 põhjal.

Üks lihtne võimalus tunnuse tihedusfunktsiooni hindamiseks ja esitamiseks on histogramm, nagu on seda ka eelmistes peatükkides graafiliselt tähesuuruse ja punanihke puhul tehtud. Histogrammil on oma miinused - tiheduse esitus sõltub tulpade arvust, lisaks on tiheduse hinnang tulpade tõttu hüplik ja tulba sisse jäävatel tunnuse väärtustel on konstantselt sama tiheduse hinnang. Nende probleemide lahendamiseks saab tiheduse hindamiseks kasutada tuumameetodit.

Lõplikust arvust punktidest saab juhusliku suuruse X tihedusfunktsiooni $f(x)$ hinnata tuumameetodil järgmiselt

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - X_i}{h}\right), \quad (6)$$

kus n on valimi punktide arv, h on akna laius, X_i on juhusliku suuruse valimi i -s väärtus ning $K(x)$ tuumafunktsioon. Tuumafunktsioon on keskväärtusega 0 ning lõpliku dispersiooniga juhusliku suuruse tihedusfunktsioon. Täiendavalt nõutakse, et tuumafunktsioon oleks sümmeetriline, $K(x) = K(-x)$.

Akna laius $h > 0$ on vabalt määratav pidev arvuline parameeter ja sellest sõltub hinnangufunktsiooni siledus. Suurema h korral on hinnangufunktsioon siledam ehk üksikute punktide mõju hinnangufunktsioonile on väiksem ning väiksema h korral on vastupidi. Analüütiliselt soovitakse akna laiust h määrata nii, et keskmine integreeritud ruutviga

$$MISE = E \left(\int \left(f(x) - \hat{f}(x) \right)^2 dx \right) \quad (7)$$

oleks minimaalne, kus $f(x)$ on tegelik tihedusfunktsioon ja $\hat{f}(x)$ on hinnatud tihedusfunktsioon.

Üldiselt kasutatakse optimaalse h leidmisel ristvalideerimist. Kui tuumafunktsioonina kasutatakse normaaljaotust ja lähendatakse samuti normaaljaotust või normaaljaotusele sarnast jaotust, siis akna laiuse valiku rusikareegliks võib pidada eeskirja

$$h = 1,06n^{1/5} \cdot \min \left\{ \hat{\sigma}, \frac{IQR}{1,34} \right\}, \quad (8)$$

kus $\hat{\sigma}$ on valimist hinnatud standardhälve ja IQR on kvartiilhaare.

Teine võimalus on kasutada automaatset akna laiuse valikut, mis põhineb sisendiks antavatel punktidel. Sel juhul ei ole h väärtus konstantne ja tiheduse tuumahinnang on

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_i} K\left(\frac{x - X_i}{h_i}\right). \quad (9)$$

Kui juhusliku suuruse X valimi i -nda väärtuse läheduses on tihedalt punkte, on kasulik võtta h väärtuseks väiksem arv, sest selles punktis on oodatud punktide rohkuse tõttu täpsem tihedusehinnang. Sarnaselt, kui i -nda väärtuse ümbruses on vähe punkte, on mõttekas võtta h väärtuseks suurem arv, et vältida hinnangufunktsiooni ülesobitamist üksikutele punktidele selles ümbruses.

Üheks automaatse aknalaiuse valiku meetodiks on Silvermani aknalaiuse valik. Esimalt tehakse piloothinnang $\hat{f}_p$ fikseeritud aknalaiuse (8) abil ja seejärel leitakse

$$h_i = \lambda_i h, \quad (10)$$

kus

$$\lambda_i = \left(\frac{\hat{f}_p(X_i)}{\left(\prod_{j=1}^n \hat{f}_p(X_j)\right)^{1/n}} \right)^{-\alpha}.$$

Valemis toodud α on valitav tundlikkuse parameeter, mille soovituslikuks väärtuseks on üldiselt 0,5.

Ülevaade mitmemõõtmelisest tuumameetodist on kirjutatud García-Portugués, 2022 põhjal.

Tuumameetodiga saab hinnata ka mitmemõõtmelise juhusliku suuruse $\mathbf{X} \in \mathbb{R}^p$ tihedusfunktsiooni. Olgu juhuslik valim $\mathbf{X}_1, \dots, \mathbf{X}_n$. Tuumameetodil hinnatud tihedusfunktsioon $\hat{f}(\mathbf{x}) \in \mathbb{R}$ avaldub kujul

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n K_H(\mathbf{x} - \mathbf{X}_i), \quad (11)$$

kus $K_H(\mathbf{x} - \mathbf{X}_i) := |H|^{-1/2} K(|H|^{-1/2}(\mathbf{x} - \mathbf{X}_i))$. Siin $K(\mathbf{x})$ on p -mõõtmeline tuumafunktsioon, mis on sümmeetriline ja unimodaalne punktis $\mathbf{0}$ ja H on $p \times p$ sümmeetriline (st $H = H^T$) ning positiivselt määratud maatriks. Eeldades, et H

on reaalarvuline ja sümmeetriline, on H positiivselt määratud parajasti siis, kui kõik tema omaväärtused on positiivsed.

Kuna täieliku aknalaiuse maatriksi puhul on vaja leida kordades rohkem aknalaiuse parameetreid, kui ühemõõtmelisel juhul, siis osutub see arvutuslikult keeruliseks. Seetõttu on tehtud lihtsustus ja aknalaiuse maatriksiks H on võetud $H = \text{diag}(h_1^2, \dots, h_p^2)$. Sel juhul saab (11) kirjutada kujul

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n K_{h_1}(x_1 - X_{i1}) \times \dots \times K_{h_p}(x_p - X_{ip}), \quad (12)$$

kus $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$ on juhusliku valimi i -s liige ja $\mathbf{h} = (h_1, \dots, h_p)^T$ on aknaalaiuse vektor.

3.2.1 Kaalutud tuumameetod

Olukorras, kus andmed pole täielikud, tuleb tavalise tuumameetodiga hinnatud tunnuse tihedusfunktsioon nihkega. Kui valimis on puuduvad väärtused, siis saadud hinnang kirjeldab vaid seda konkreetset valimit, mitte üldkogumit. Puuduvate andmete korral saab rakendada sarnast ideed nagu Horwitz-Thompsoni hinnang kogusummale (3) – kaalutud tuumameetodit, mida käsitleb Dubnicka, 2009. Siinkohal on idee kohandatud koguvalimi ja puuduvate andmete pealt üldkogumi ja valimi peale, nii et valimisse mittejäävaid objekte vaadeldakse kui „puuduvaid“ andmeid.

Olgu $\mathbf{r}$ binaarne kaasamisindikaatori vektor pikkusega N

$$r_i = \begin{cases} 1, & \text{kui objekt } i \text{ on valimis} \\ 0, & \text{muidu,} \end{cases} \quad (13)$$

kus N on üldkogumi objektide arv ja r_i on vektori $\mathbf{r}$ elemendid. Olgu tähistuse

lühiduse mõttes

$$K_h(x - X_i) := \frac{1}{h} K\left(\frac{x - X_i}{h}\right). \quad (14)$$

Parandatud tiheduse hinnang puuduvate andmete korral avaldub järgmiselt (Dubnicka, 2009):

$$\hat{f}^{HT}(y) = \frac{1}{N} \sum_{i=1}^N \frac{r_i}{\pi_i} K_h(x - X_i).$$

Kuna $r_i = 0$, kui objekt ei kuulu valimisse, saab valemi valimi ja üldkogumi objektide arvu hinnangu (5) kaudu esitada kujul

$$\hat{f}^{HT}(y) = \frac{1}{\hat{N}} \sum_{i=1}^n \omega_i K_h(x - X_i). \quad (15)$$

Võrreldes tavalise tuumameetodi valemiga (6), on kaalutud tuumameetodi valemi (15) summasse lisatud iga valimi objekti kohta kaal ω_i . Saab mõelda, et tavalise tuumameetodi valemis on iga objekt kaaluga 1. Üldkogumi objektide arvu hinnangut on otstarbekas kasutada tegeliku N asemel seetõttu, et ei tekiks probleeme hinnangufunktsiooni korrektse normaliseerimisega. Normaliseerimise all mõeldakse hinnangu korrutamist sellise arvuga, mille puhul tuleks tihedusfunktsiooni alune pindala kokku täpselt 1,0. Valimi juhuslikkuse tõttu ei pruugi alati kehtida, et $1/N \sum_{i=1}^n \omega_i = 1$, kuid $1/\hat{N} \sum_{i=1}^n \omega_i = 1$ kehtib alati.

3.3 Nadaraya-Watsoni tuumaregressiooni hinnang

Tuumaregressioon on edasiarendus tuumameetodil tihedusfunktsiooni hindamisest üldise funktsiooni hindamisele (Hollander, Wolfe ja Chicken, 2014). Tuumaregressiooni hinnangut saab kasutada, et leida seos uuritava tunnuse Y ja seletava tunnuse X vahel. Hinnangut saab vaadata kui juhusliku suuruse Y tinglikku keskväärtuse hinnangut (García-Portugués, 2022):

$$\hat{E}(Y|X) = g(X),$$

kus g on mingi funktsioon, mida soovitakse tuumaregressiooniga leida.

Järgnevalt kasutatakse eelmises alapeatükis defineeritud tähistuse lühendust (14).

Olgu i -nda üldkogumi objekti jaoks defineeritud abitunnuste vektor $\mathbf{x}_i$ k abitunnusega, mis on seotud uuritava tunnusega

$$\mathbf{x}_i = (x_{1i}, \dots, x_{ki}).$$

Nadaraya-Watsoni tuumaregressiooni hinnang on

$$\hat{g}(\mathbf{x}_i) = \hat{E}(Y|\mathbf{X} = \mathbf{x}_i) = \frac{\sum_{j=1}^n y_j K_H(\mathbf{x}_i - \mathbf{x}_j)}{\sum_{j=1}^n K_H(\mathbf{x}_i - \mathbf{x}_j)} \quad (16)$$

$$= \frac{\sum_{j=1}^n y_j K_{h_1}(x_{i1} - x_{j1}) \times \dots \times K_{h_k}(x_{ik} - x_{jk})}{\sum_{j=1}^n K_{h_1}(x_{i1} - x_{j1}) \times \dots \times K_{h_k}(x_{ik} - x_{jk})}, \quad (17)$$

kus $\mathbf{x}_i = (x_{i1}, \dots, x_{ik})^T$ ja $\mathbf{x}_j = (x_{j1}, \dots, x_{jk})^T$ on abitunnuste vektorid ja y_j on uuritava tunnuse j -s väärtus. Siinkohal on oluline tähele panna, et K_H on mitmemõõtmeline tuumafunktsioon diagonaalse aknalaiuse maatriksiga H , mida saab esitada ühemõõtmeliste tuumafunktsioonide korrutisena (12). (Hollander, Wolfe ja Chicken, 2014)

4 Probleemi püstitus

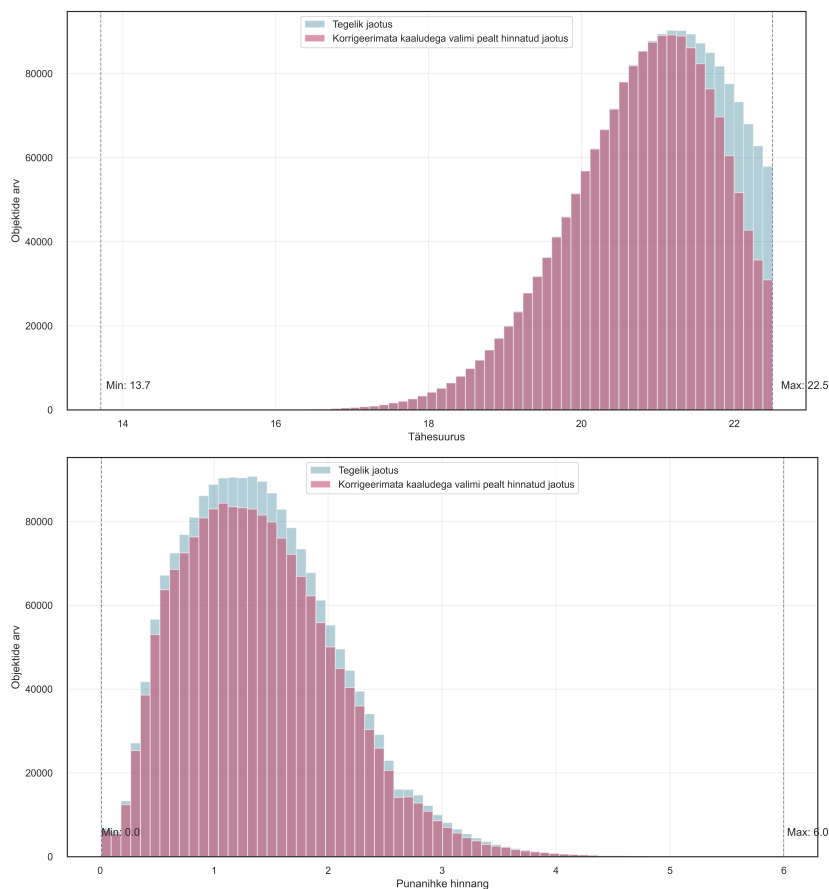
Sadu vaatlusuuringuid simuleerides ning simulatsioonide pealt objekti edukalt vaatlemise tõenäosused leides võiksid need tõenäosused lähendada objekti vaatlusvalimisse kaasamise tõenäosuseid. Õiged kaasamistõenäosused on olulised, et leida iga objekti kaal ehk mitut objekti konkreetne valimiobjekt kogu kataloogis esindab. Õi-gete kaaludega saab konstrueerida suvalise pideva tunnuse kaalutud histogrammi või kaalutud tihedusfunktsiooni hinnangu, mis taastab tegeliku tihedusfunktsiooni.

Probleemiks on see, et hetkel simulatsioonide kaudu hinnatud kaasamistõenäosused (1) on teatud regioonis nihkega ning kaalud ei summeeru üldkogumi objektide arvuks. Tekkivat nihet on tähesuuruse ja punanihke hinnangu kaudu kujutatud Joonisel 3. Jooniselt on näha, et korrigeerimata algsel kujul kaalud ei taasta tunnuse jaotust kataloogis, tähesuuruse puhul tekib nihe suurte väärtuste korral ning punanihke hinnangu puhul keskvärtuse ümbruses.

Nihe võib tekkida mitmel põhjusel. Andmestikus on ligikaudu 7% andmeid, kus edukalt vaatlemise tõenäosus (1) on 0, mis tähendab, et neid objekte ei vaadeldagi või vaadeldakse väga harva. See võib tuleneda sellest, et

1. objekte on rohkem kui fiibritele mahub ja seega on mõne objekti vaatlusesse valiku tõenäosus nii väike, et objekti ei valitagi kunagi või valitakse väga harva,
2. mõne objekti ettenähtud säriaeg on suurem, kui summaarne säriaeg kõiki-de vaateväljade peale kokku, mis seda objekti hõlmavad, ja sellisel juhul eemaldatakse selline objekt algoritmi poolt juba enne fiiber-objekt paaridele tõenäosuste määramist (Tempel *et al.*, 2020b).

Lõputöö eesmärgiks on esiteks korrigeerida objektide edukalt vaatlemise (kaasamistõenäosuste) hinnangud (1), et need lähendaksid tegelikke kaasamistõenäosuseid ja summeeruksid üldkogumi objektide arvuks. Teiseks taastatakse töös tähesuuruse



Joonis 3: Objekti tähesuuruse ja punanikke histogrammid kogu kataloogi andmete põhjal (sinine) võrdluses kaalutud histogrammiga valimi põhjal (violetne). Kaalud on leitud Horwitz-Thompsoni kogusumma hinnangu laiendustegurina ehk kaaluna $\frac{1}{\hat{\pi}_i}$, kus $\hat{\pi}_i$ on objekti edukalt vaatlemise tõenäosuse hinnang. Kaalu seletus on toodud alapeatükis [3.1](#).

ja punanikke tihedusfunktsioonid kasutades korrigeeritud kaale ja hinnatakse kaasamistõenäosuste korrigeerimismeetodite täpsust.

5 Metoodika

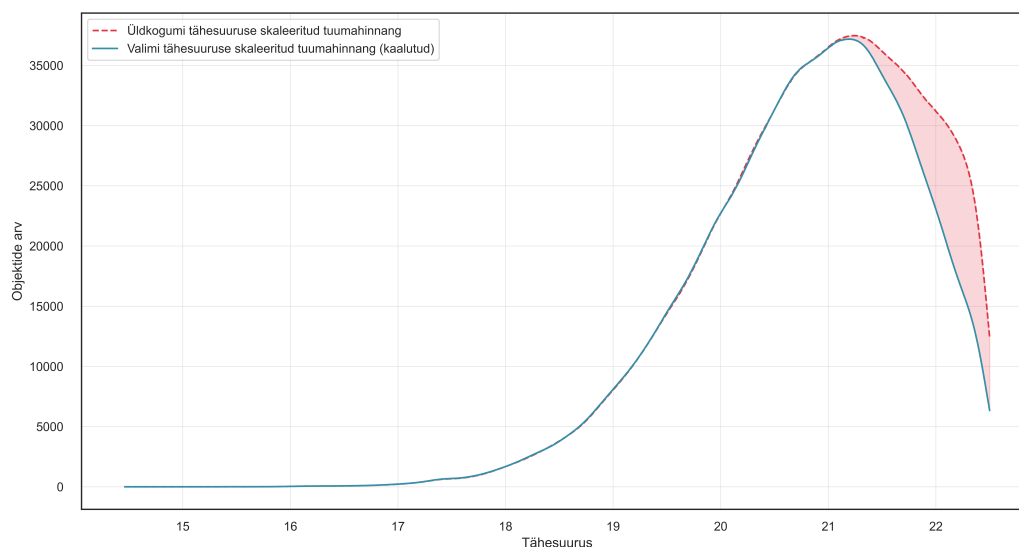
Metoodika peatükis kirjeldatakse probleemi lahendamise teoreetilist protsessi kahes osas. Esimeses osas tehakse ülevaade kaasamistõenäosuste korrigeerimise meetoditest ja teises osas taastatud tihedusfunktsiooni hinnangute täpsuse hindamise meetodist. Kaasamistõenäosuste korrigeerimise all mõeldakse kas olemasolevate tõenäosuste kohandamist või täiesti uute kaasamistõenäosuste leidmist nii, et need lähendaksid tõeliseid kaasamistõenäosuseid.

5.1 Kaasamistõenäosuste korrigeerimine

Kaasamistõenäosuste nihke korrigeerimiseks on lahendamise käigus jõutud kahe erineva meetodini. Kirjapildi lihtsustamiseks viidatakse kogu kataloogi objektidele kui üldkogumile ja kataloogis vaadeldud objektidele kui valimile. Edaspidi mõeldakse skaleerimise all funktsiooni mingi konstandiga läbi korrutamist.

Esimese meetodiga püütakse leida juba olemasolevatele objekti edukalt vaatlemise tõenäosustele (1) korrigeerimisfunktsioon, mis sõltuks tähesuurusest. Olgu konstrueeritud tavaline kaalumata tihedusfunktsiooni hinnang (6) tähesuurusele üldkogumis ja algsete kaaludega (leitud olemasolevate kaasamistõenäosuste (1) kaudu) kaalutud tihedusfunktsiooni hinnang (15) tähesuurusele valimis. Tähistatagu neid hinnanguid vastavalt $\hat{f}_u(mag)$ ja $\hat{f}_v(mag)$.

Olgu $\hat{f}_u(mag)$ skaleeritud punktide arvuga N ja olgu $\hat{f}_v(mag)$ skaleeritud valimi kaalude summaga, mis annab hinnangu $\hat{N}$. Seega skaleeritud hinnangud on $N \cdot \hat{f}_u(mag)$ ja $\hat{N} \cdot \hat{f}_v(mag)$. Skaleering on vajalik selleks, et teha üldkogumi ja valimi tihedusfunktsioonide hinnangud võrreldavaks. Algsed kaalud on nihkega, mistõttu hindavad need üldkogumi mahtu väiksemaks, kui see tegelikult on. Algseid kaale kasutatakse kaalutud tuumahinnangu normaliseerimises (15). Seetõttu ei joundu normaliseeritud tuumahinnangud (6) ja (15) väiksemate tähesuuruse väärtuste korral nii, nagu on näha histogrammil Joonisel 3. Pärast skaleerimist on see probleem

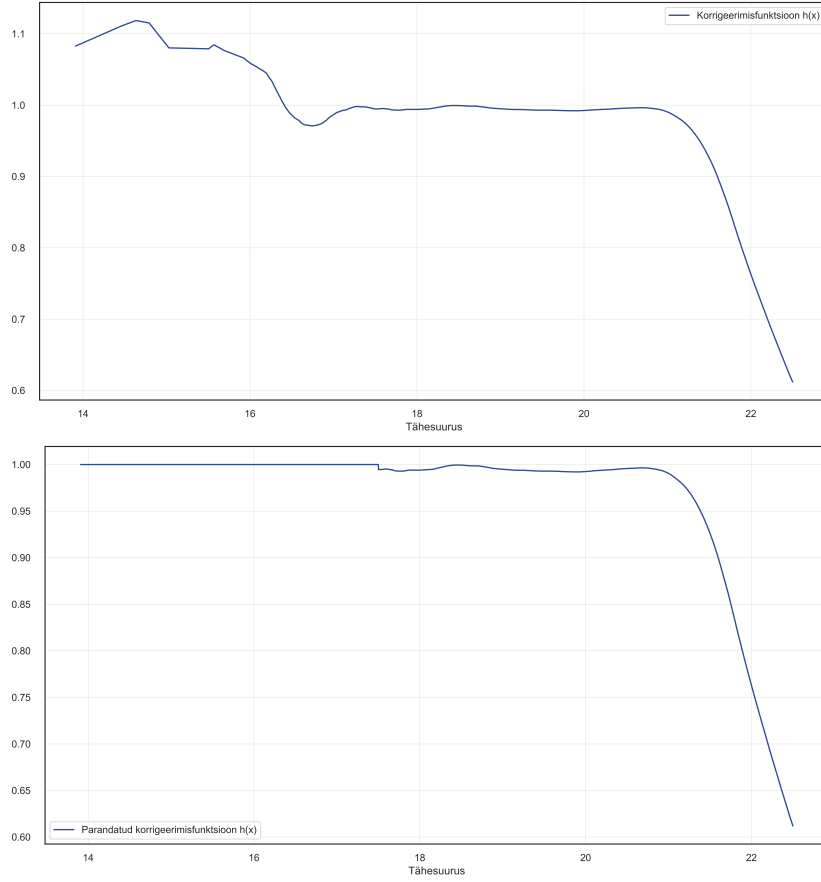


Joonis 4: Tuumameetodil leitud tähesuuruse tihedusfunktsiooni hinnangud üldkogumis (katkendlik punane joon) ja kaalutud valimis (pidev sinine joon). Hinnangud on leitud 100 000 objekti suurusest valimist, mis on võetud juhuslikult tagasipane-kuta kogu kataloogist.

lahendatud. Skaleeritud tuumahinnangud on näha tähesuuruse korral Joonisel 4.

Kaalude korrigeerimise idee seisneb sellise funktsiooni leidmises, mis näitab, mitu korda valimi pealt hinnatud kogu objektide arv erineb tegelikust kogu objektide arvust üldkogumis mingi tähesuuruse m väärtuse puhul. Joonisel 4 on see erinevus näha helepunase alana ehk see osa üldkogumi tihedusfunktsiooni hinnangust, mida kaalutud valimi oma ei kata. Teisisõnu näitab erinevus seda, kui palju rohkematele või vähematele objektidele üldkogumis peaks tähesuuruse m väärtus laienema. Seega saab leida iga objekti laiendustegurile ω_i (kaalule) korrigeeriva väärtuse selle tähesuuruse väärtuse järgi. Kuna objekti laiendustegur ehk kaal on kaasamistõenäosuse pöördväärtus, saab probleemi üle viia kaasamistõenäosuste korrigeerimisele ehk tuleb leida korrigeerivad väärtused kaasamistõenäosustele.

Korrigeerimisfunktsiooniks on $h(mag)$, mis sõltub tähesuurusest ja on olemuselt kahe skaleeritud hinnangufunktsiooni suhe



Joonis 5: Korrigeerimisfunktsioon $h(mag)$ (18) (ülemine graafik) ja parandatud korrigeerimisfunktsioon $h^{par}(mag)$ (19) (alumine graafik).

$$h(mag) = \frac{\hat{N} \cdot \hat{f}_v(mag)}{N \cdot \hat{f}_u(mag)}. \quad (18)$$

Korrigeeritud kaasamistõenäosused saab, kui korrutada algne tõenäosus (1) läbi korrigeerimiskordajaga

$$\hat{\pi}_i^{korr} = \hat{\pi}_i \cdot h(mag_i),$$

kus mag_i on i -nda objekti tähesuuruse väärtus.

Korrigeerimisfunktsioon $h(mag)$ on toodud Joonisel 5 ülemises osas, kus on näha võnkumist suhte väärtuse 1,0 ümbruses. See võnkumine tasandub tähesuuruse

väärtuse $\sim 16,5$ ümbruses. Selline võnkumine väiksemate väärtuste puhul võib tulla sellest, et selles piirkonnas on vähe objekte ja seetõttu tuumahinnang tuleb ebatäpsem. Võnkumise tasandamiseks on võimalik korrigeerimisfunktsiooni väärtuseks omistada 1,0, kui tähesuurus on väiksem, kui mingi arv tähesuuruse vahemikus, praegusel juhul on selleks valitud 17,5. Parandusega korrigeerimisfunktsioon avaldub järgmiselt:

$$h^{par}(mag) = \begin{cases} 1, & x \leq 17,5 \\ h(mag), & \text{muidu.} \end{cases} \quad (19)$$

Selline lahendus ei ole üldistav, vaid sõltub konkreetse kataloogi tähesuuruse jaotusest.

Teine meetod põhineb Nadaraya-Watsoni tuumaregressioonil. Olgu uuritav tunnus valikuvektor $\mathbf{r}$ elementidega r_i

$$r_i = \begin{cases} 1, & \text{kui objekt } i \text{ on vaadeldud} \\ 0, & \text{muidu.} \end{cases}$$

Valikuvektori iga liige on Bernoulli jaotusega juhuslik suurus $r_i \sim B(\pi_i)$, kus π_i on tõenäosus, et $r_i = 1$. Bernoulli jaotusega juhusliku suuruse keskväärtus on $E(r_i) = \pi_i$, mis tähendab, et kaasamistõenäosuseid (1) saab hinnata valikuvektori keskväärtuse kaudu $\hat{E}(r_i) = \hat{\pi}_i$.

Nadaraya-Watsoni tuumaregressioon hindab tinglikku keskväärtust $\hat{E}(Y|\mathbf{X} = \mathbf{x})$ (16), kus $\mathbf{x}$ on abitunnuse vektor. Kui uuritavaks tunnuseks on valikuvektor $\mathbf{r}$, siis

$$\hat{E}(r_i|\mathbf{X} = \mathbf{x}_i) = \hat{\pi}_i.$$

Oluline erinevus esimese korrigeerimismeetodiga on see, et Nadaraya-Watsoni tuumaregressiooni kaudu kaasamistõenäosuste hindamine ei kasuta *frac* tunnuse mit-

mesaja läbimängimise tulemusel leitud kaasamistõenäosuste hinnanguid (1). Selle asemel hinnatakse kaasamistõenäosused uuesti abitunnuste abil, kus uuritavaks tunnuseks on üheainsa *frac* simulatsiooni pealt leitud valikuvektor $\mathbf{r}$ (13).

Esimene mõte kaasamistõenäosuste hindamiseks binaarse valikuvektori pealt on ilmselt logistiline regressioon, kuid praeguses olukorras tuleb välja, et seda ei ole kõige mõistlikum kasutada. Nadaraya-Watsoni tuumaregressiooni eelis logistilise regressiooni ees on see, et Nadaraya-Watsoni hinnang ei pea olema monotoonne. Logistiline regressioon hindab uuritava tunnuse keskväärtust sigmoidfunktsioonina, mis on teatavasti kas kasvav või kahanev funktsioon. Praeguste andmete ja konteksti puhul ei ole monotoonsuse eeldus loogiline. Näiteks, kui võtta abitunnuseks tähesuurus, siis suurema tähesuuruse väärtuse korral ei pruugi objekti kaasamistõenäosus alati olla tingimata kas suurem või väiksem kui väiksema tähesuuruse väärtuse korral. Ruutliikme juurde lisamine selle probleemi lahendamiseks samuti ei paranda olukorda.

Kui korrigeeritud kaasamistõenäosused on leitud kas esimese või teise meetodiga, saab taastada valimi pealt kaalutud tihedusfunktsiooni hinnangu (15) valitud pidevale tunnusele. Käesolevas töös taastatakse korrigeeritud kaasamistõenäosustega vaadeldud objektide valimi pealt tähesuuruse ja punanihke tihedusfunktsioonid. Tunnuste tihedusfunktsiooni arvutamisel on kasutatud standardse normaaljaotuse tuumafunktsiooni $N(\mu = 0, \sigma = 1)$, mille tihedusfunktsioon on kujul

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \exp \frac{-z^2}{2}$$

Kuna punanihke ja tähesuuruse tihedusfunktsioonid üldkogumis on unimodaalsed ja sarnanevad kujult normaaljaotusega, kasutatakse Silvermani aknalaiuse määramise meetodit.

5.2 Taastatud tihedusfunktsiooni täpsuse hindamine

Korrigeeritud kaasamistõenäosuste täpsust saab silma järgi hinnata valimit ja üldkogumit võrdlevate graafikute abil, kuid erinevate meetodite võrdlemiseks on parem esitada ka arvuline parameeter. Kuna tähesuuruse ja punanihke teoreetilised tihedusfunktsioonid ei ole teada, vaid ainult tuumameetodi hinnang üldkogumi pealt, siis ei saa kasutada integreeritud ruutviga *MISE* (7). Tunnuse tihedusfunktsiooni taastamise meetodite võrdlemiseks on konstrueeritud eraldi näitaja. Olgu pideva juhusliku suuruse X väärtuste lõigus $[X_{min}, X_{max}]$ võetud ühtlaselt k punkti.

$$X_{min} = x_1 < x_2 < \dots < x_k = X_{max}.$$

Olgu üldkogumi pealt hinnatud pideva tunnuse tihedusfunktsioon jällegi $\hat{f}_u(x)$ ja valimi pealt taastatud tunnuse tihedusfunktsioon $\hat{f}_v(x)$. Ruutjuur keskmistatud ruutveast punktides $x_1, \dots, x_k$ on

$$RMSE := \sqrt{\frac{1}{k} \sum_{i=1}^k (\hat{f}_u(x_i) - \hat{f}_v(x_i))^2}. \quad (20)$$

Selline veaparameeter ei anna absoluutses mõttes informatsiooni tiheduse hindamisel tehtud vea kohta, kuna sõltub punktide arvust k . Siiski on seda kasulik rakendada erinevate taastatud tihedusfunktsioonide võrdluses. Veel enam, sellist veaparameetrit on lihtne ja kiire arvutada võrreldes funktsioonialuse integraali leidmisega.

Selleks, et vältida liiga optimistlikku veahinnangut, leitakse mudel (kas korrigeerimisfunktsioon $h(mag)$ või Nadaraya-Watsoni tuumaregressiooni hinnang $g(\mathbf{x})$) treeningandmete pealt ja veahinnang (20) testandmete pealt. Veahinnang võib olla liiga optimistlik, kui leida mudel ja hinnata selle headust sama andmestiku peal, sest mudel kirjeldab ära osa konkreetse valimi eripärast. Treening- ja testandmete kasutamisega saab leevendada valimi eripära mõju veahinnangule.

Treening -ja testandmete võtmise jaoks jagatakse esmalt kogu kataloog juhuslikult pooleks. Üks pool on treeningandmete jaoks ja teine pool testandmete jaoks. Üks treeningandmestik on 10 000 objekti suurune juhusliku tagasipanekuta valiku teel saadud andmestik, mis on võetud treeningandmete poolelt. Samamoodi on võetud 10 000 objekti suurune testandmestik testandmete poolelt. Kui mudel on ühe treeningandmestiku peal leitud ja testandmete pealt veahinnang leitud, salvestatakse see ning tehakse kogu protsess uuesti. Selliseid iteratsioone tehakse iga korrigeerimismeetodiga kokku 300.

Selline valimi suurus vea hindamise jaoks on valitud seetõttu, et tuumameetodil tihedusfunktsioon hindamine on ajaliselt keerukas operatsioon. 10 000 juhuslikult valitud punkti pealt tunnuse tihedusfunktsiooni hindamine annab piisavalt täpse representatsiooni tegelikust tunnuse tihedusfunktsioonist. Lisaks on sellise suurusega valimi pealt tihedusfunktsiooni leidmise aeg mõistlik, et teha iga meetodi korral 300 veahinnangu leidmise iteratsiooni.

Töös võrreldakse kolme erinevat tunnuse tihedusfunktsiooni taastamise meetodi kombinatsiooni.

1. Korrigeerimisfunktsioon $h(mag)$
2. Nadaraya-Watsoni meetod, $\mathbf{x} = (\text{tähesuurus})$,
3. Nadaraya-Watsoni meetod, $\mathbf{x} = (\text{tähesuurus}, RA, Dec)$.

Meetodi headust hinnatakse veahinnangute (20) keskmisega.

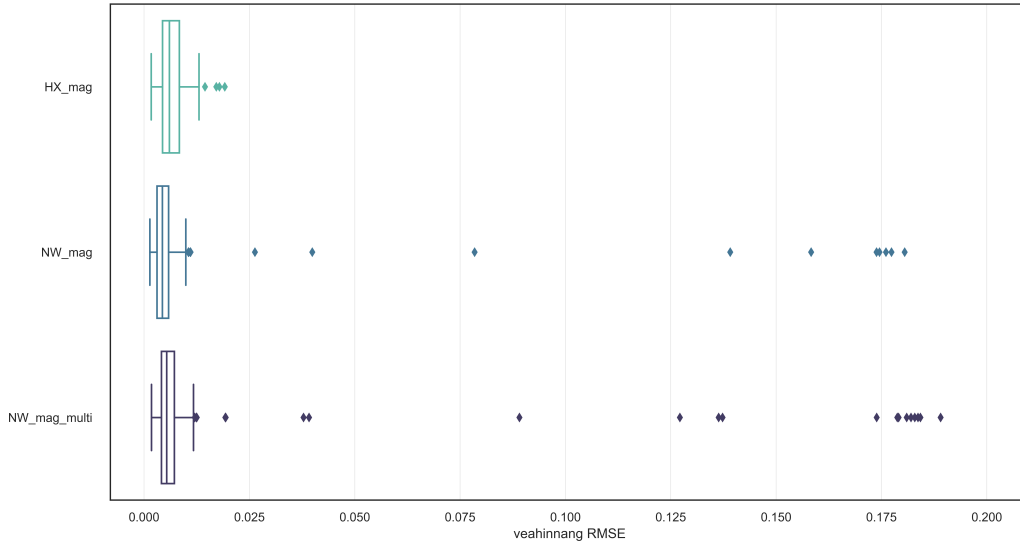
6 Tulemused

Tulemuste peatükis esitatakse eelmise peatüki lõpus loeteluna esitatud kolmel erineval kaasamistõenäosuste korrigeerimisviisil saadud kaalutud tuumahinnangute täpsusehinnangud (20), kus ühtlaselt võetud punktide arv tunnuse teljel on $k = 1000$. Seejärel võrreldakse veahinnanguid omavahel. Täpsuse hinnangud on leitud nii tähesuuruse kui ka punanihke tihedusfunktsiooni taastamisel kasutades metoodika peatükis kirjeldatud treening -ja testandmete meetodit. Lisaks veahinnangute esitamisele taastatakse ja esitatakse graafiliselt kolme korrigeerimisviisiga saadud kaasamistõenäosustega vaatlusandmete pealt kaalutud tuumahinnangud nii tähesuurusele kui ka punanihkele.

Kaasamistõenäosuste korrigeerimise meetodid ja vaatlusandmete pealt kaalutud tuumahinnangu leidmine on implementeeritud Pythonis. Tuumameetodil hindamine on tehtud *KDEpy* lisamooduliga ([KDEpy n.d](#)) ja Nadaraya-Watsoni (NW) tuumaregressioon *statsmodels* lisamooduliga (Perktold, Seabold ja Taylor, 2022).

Taastatud tähesuuruse tihedusfunktsiooni veahinnangud iga korrigeerimismeetodi kohta on kujutatud Joonisel 6. Esiteks on näha, et kõigi kolme meetodi puhul langevad keskmised ühesuguselt 0,01 ümbrusesse. Korrigeerimisfunktsiooni $h(mag)$ puhul on keskmine viga veidi kõrgem kui Nadaraya-Watsoni meetodite puhul. Lisaks kõrgemale keskmisele langevad korrigeerimisfunktsiooni puhul 75% veahinnangutest suuremasse vahemikku, kuid ekstreemseid väärtusi on ka vähem. NW meetodite puhul on huvitav tähele panna, et mitme tunnusega abivektori kasutamise puhul on vigade keskmine veidi kõrgem kui ühe tunnusega abivektori puhul. See võib tuleneda sellest, et ekvaatorilised koordinaadid ei kirjelda hästi objekti kaasamistõenäosuse varieeruvust ja ei paranda tuumaregressiooni mudelit.

Veel on näha, et NW meetodite puhul on mitmeid ekstreemseid väärtusi või erindeid, mis erinevad enamus veahinnangutest. See tuleneb tõenäoliselt sellest, et aknalaius h NW mudelile arvutati ühe korra ristvalideerimise abil enne 300 veahinnangu iteratsiooni leidmist arvutuste kiirendamise eesmärgil. Seetõttu võis ette

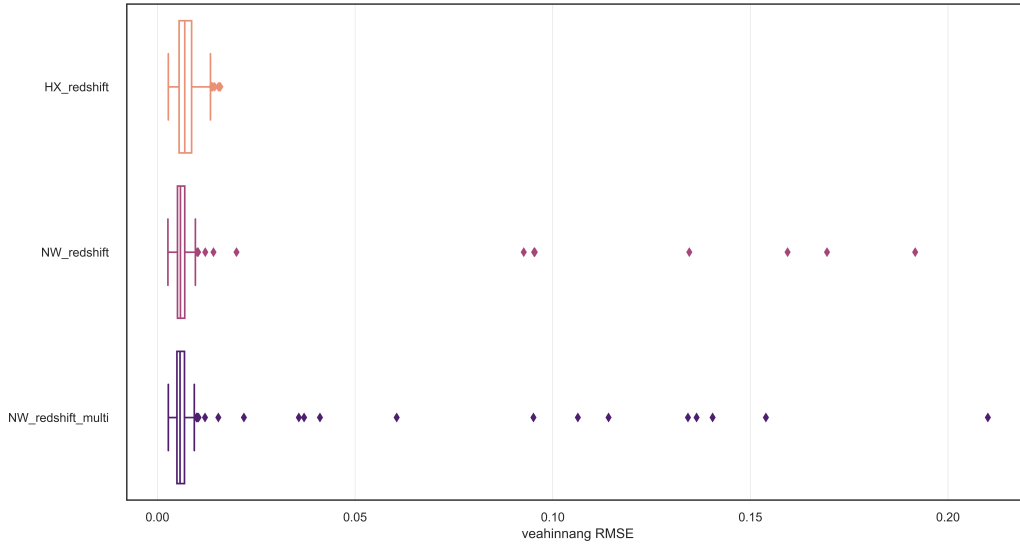


Joonis 6: Karpdiagramm tähesuuruse kaalutud tihedusfunktsiooni hindamisel tekkinud vigade $RMSE$ kohta 300-l juhusliku treening -ja testandmete paari iteratsioonil. Korrigeerimismeetodid on 1) HX_mag – korrigeerimisfunktsiooni meetod (19) 2) NW_mag – Nadaraya-Watsoni meetod ühe abitunnusega $\mathbf{x} =$ (tähesuurus) 3) NW_mag_multi – Nadaraya-Watsoni meetod kolme abitunnusega $\mathbf{x} =$ (tähesuurus, RA , Dec).

juhtuda selliseid juhuslikke treeningandmeid, mille kuju oli veidi teistsugune sellest, mille peal aknalaiust leiti ja seetõttu ka kaasamistõenäosuste hinnangud ei tulnud mitte kõige optimaalsema aknalaiuse valiku tõttu sama head. See asjaolu näitab ka, et NW meetodi headus sõltub väga tugevalt aknalaiuse valikust.

Taastatud punanihke tihedusfunktsiooni veahinnangud iga korrigeerimismeetodi kohta on kujutatud Joonisel 7. Jällegi esineb väga sarnane pilt tähesuuruse tihedusfunktsiooni veahinnangu joonisega. Korrigeerimisfunktsiooni $h(mag)$ puhul on veahinnangute keskmine veidi kõrgem, kui NW meetodite puhul ning NW meetoditel on rohkem erindeid. Siiski tunduvad punanihke korral ühe abitunnuse ja mitme abitunnuse mudelite veahinnangute keskmised ühel joonel paiknevat.

Veahinnangute kaudu on võimalik saada üldine arusaam erinevate korrigeerimismeetodite ja kaalutud tuumahinnangu täpsusest, kuid lisaks pakub huvi ka taastatud tihedusfunktsiooni kuju ja selle sarnanemine üldkogumi pealt hinnatud tihe-



Joonis 7: Karpdiagramm punanihke kaalutud tihedusfunktsiooni hindamisel tekkinud vigade $RMSE$ kohta 300-l juhusliku treening -ja testandmete paari iteratsioonil. Korrigeerimismeetodid on 1) $HX_redshift$ – korrigeerimisfunktsiooni meetod (19) 2) $NW_redshift$ – Nadaraya-Watsoni meetod ühe abitunnusega $\mathbf{x} = (\text{tähesuurus})$ 3) $NW_redshift_multi$ – Nadaraya-Watsoni meetod kolme abitunnusega $\mathbf{x} = \text{tähesuurus}, RA, Dec$.

dusfunktsiooniga. Selleks taastatakse nii tähesuuruse kui punanihke tihedusfunktsioonid kõigi kolme korrigeerimismeetodiga valimi pealt ning võrreldakse üldkogumi hinnatud tihedusfunktsiooni kujuga. Üldkogumiks on võetud iga meetodi puhul 100 000 objekti suurune juhuslik tagasipanekuta valim kogu sihtkataloogist. Joonised leiab Lisa 1 alt.

Nagu taastatud tihedusfunktsioonide graafikutelt näha, lähendavad kõigi kolme meetodi puhul kaalutud tuumameetodil tihedusfunktsioonide hinnangud üldkogumi pealt hinnatud tihedusfunktsiooni suhteliselt hästi. Silma järgi on võimatu otsustada, milline korrigeerimisviis kõige täpsemalt üldkogumi tihedusfunktsiooni taastab. Ainuke erand on mitme abitunnusega NW meetod, kus valimi pealt hinnatud tihedusfunktsioon on tähesuuruse puhul ebatäpsem, mis kajastub ka Joonisel 6. Ülejäänud kahe meetodi puhul on näha, et tähesuuruse puhul järgib taastatud tihedusfunktsioon üldkogumi tihedusfunktsiooni hinnangut veidi paremini, kui

punanihke puhul. See on ilmselgelt nii, kuna kõikides kaasamistõenäosuste korrigimisviisides on tähesuurus seletava tunnuseks sees.

Korrigeerimismeetodite puhul tuleb lisaks täpsusele tähele panna ka mudelite leidmise kiirust arvutuslikult. Tuumameetodil tihedusfunktsiooni hinnangu leidmine naiivsel kujul on arvutuslikult keerukas operatsioon ja vähegi rohkemate andmete puhul võib see ebarealistlikult kaua aega võtta. Arvutuste kiirendamiseks on olemas mitmed võimalusi, näiteks kasutades naiivse valemipõhise lähenemise asemel puudepõhist mudelit ([TreeKDE n.d](#)) nagu käesolevas töös või Fourier' kiirteisendust (FFT). Viimane meetod annab küll väga kiiresti tulemust, kuid pole see eest kõige täpsem ning punkte saab väärtustada vaid ühtlaste vahedega võrgustikul. Mõnikord võib tuumafunktsiooni leidlik valimine normaaljaotuse asemel arvutamist kiirendada, et summa liikmeid tuumahinnangus vähendada. Selline on näiteks *triweight* tuumafunktsioon, mis kujult sarnaneb normaaljaotusele, kuid mille väärtuseks on 0, kui sisendparameetri absoluutväärtus on suurem ühest.

Kokkuvõte

Bakalaureusetöö eesmärgiks oli taastada etteantud vaatlusuuringu kataloogi simuleeritud vaatlusvalimi pealt tähesuuruse ja punanihke tihedusfunktsioonid. See saavutati korregerides esmalt edukalt vaatlemise tõenäosused (kaasamistõenäosused) ning seejärel rakendades kaalutud tuumameetodit vaatlusvalimi peal.

Töös esitati kaasamistõenäosuste korregerimiseks kaks meetodit. Esimene meetod kasutas ära juba varasemalt simulatsioonide pealt arvutatud kaasamistõenäosused ning leidis nendele tähesuurusest sõltuva korregerimisfunktsiooni $h(mag)$. Teise meetodiga hinnati kaasamistõenäosused simuleeritud valikuvektori pealt uuesti kasutades Nadaraya-Watsoni tuumaregressiooni. Nadaraya-Watsoni tuumaregressiooni puhul kasutati kahte abitunnuste varianti: $\mathbf{x} = (\text{tähesuurus})$ ja $\mathbf{x} = (\text{tähesuurus}, RA, Dec)$.

Meetodite täpsus leiti juhuslikult valitud treening -ja testandmetel, kus treeningandmetel leiti kas korregerimisfunktsioon $h(mag)$ või Nadaraya-Watsoni tuumaregressioon ja täpsusparameetri väärtus hinnati testandmetelt. Korregerimisfunktsiooni meetod andis nii tähesuuruse kui punanihke puhul veidi kõrgema, kuid mitte märkimisväärselt kõrgema veahinnangu. Tähesuuruse puhul andis mitme abitunnusega tuumaregressioon samuti veidi kõrgema veahinnangu, kui ühe abitunnusega tuumaregressioon. See tähendab, et mitme abitunnusega tuumaregressioonil ei tundu olevat eeliseid ühe abitunnusega tuumaregressiooni ees ja lisaks võtab see arvutustlikult kauem aega, mistõttu on see ülemäärane.

Lõpuks taastati tähesuuruse ja punanihke tihedusfunktsioonid vaatlusvalimilt kõigil kolmel viisil. Taastatud tihedusfunktsioonid lähendavad üldkogumi vastava tunnuse jaotust hästi, nii et silmaga on raske eristada, millise meetodi puhul taastatud funktsioon üldkogumi tihedusfunktsiooni paremini lähendab.

Mõlemal meetodil, nii korregerimisfunktsioonil ja Nadaraya-Watsoni tuumaregressioonil, on omad eelised ja puudused. Korregerimisfunktsioon on implementee-

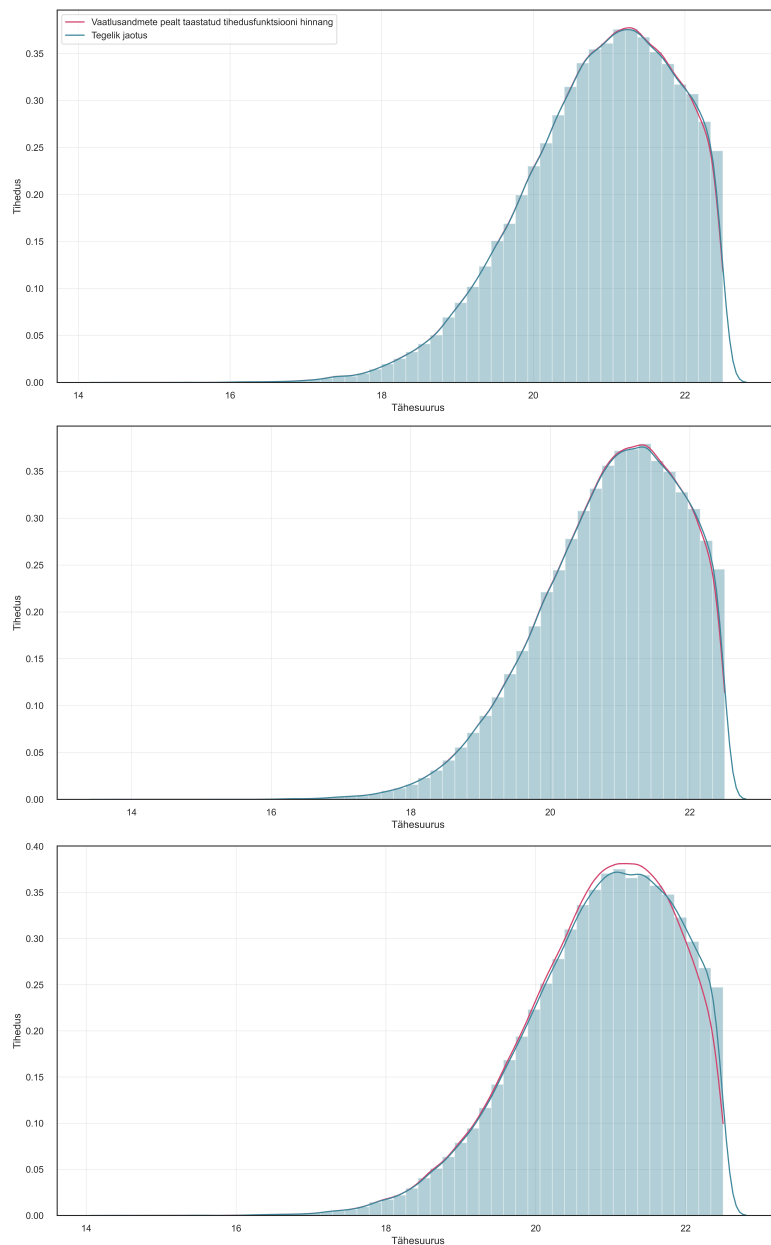
ritav rohkemate sammudega, mistõttu saab seda tõhusamalt arvutuslikus mõttes optimiseerida. Samas ei ole see meetod väga üldistav, see tähendab, et korrigeerimisfunktsiooni $h(mag)$ peab teatud piirkonnas manuaalselt väärtustama andmete põhiselt ebatäpsuste tõttu. Nadaraya-Watsoni meetod seevastu on üldistav, see tähendab, et tulemused tulevad enamasti erinevate andmete puhul sarnaselt head ja manuaalseks väärtustamiseks pole põhjust. Samas võtab töös kasutatavate teekidega optimaalse aknalaiuse leidmine kaua aega ja arvutuste kiirendamist hinnangu leidmise jaoks ei olnud võimalik teha. Meetodi valik sõltub kasutaja võimalustest, kuid täpsuse mõttes ei ole neil kahel meetodil märkimisväärset erinevust.

Kasutatud allikad

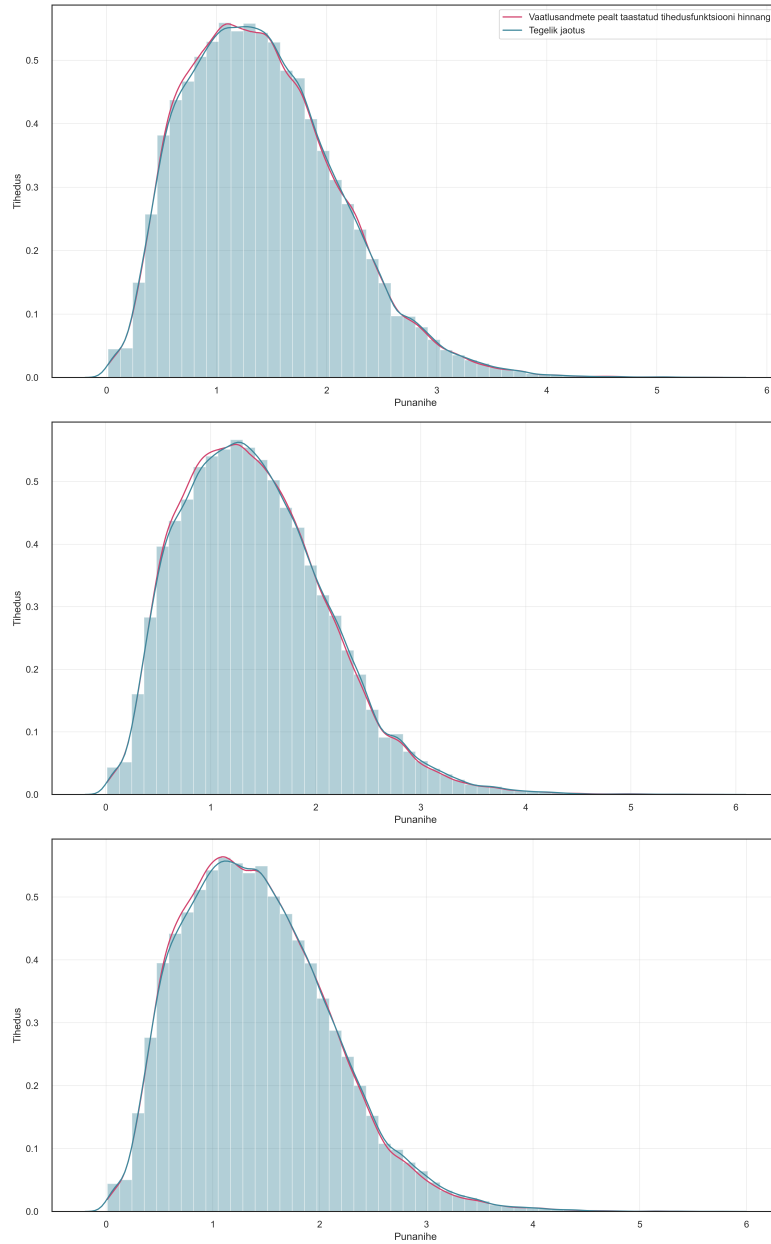
- 4MOST Overview* (n.d). URL: <https://www.4most.eu/cms/> (vaadatud 12.02.2022).
- 4MOST Science Overview* (n.d). URL: <https://www.4most.eu/cms/science/overview/4most.eu> (vaadatud 12.02.2022).
- 4MOST Subsystems* (n.d). URL: <https://www.4most.eu/cms/facility/subsystems/> (vaadatud 12.02.2022).
- De Jong, Roelof S. *et al.* (2019). “4MOST: Project overview and information for the First Call for Proposals”. *The Messenger* vol. 175 pp. 3-11, 9 pages. ISSN: 0722-6691. DOI: [10.18727/0722-6691/5117](https://doi.org/10.18727/0722-6691/5117).
- Dubnicka, Suzanne R. (2009). “Kernel Density Estimation With Missing Data And Auxiliary Variables”. 51(3), lk. 247–270. DOI: [10.1111/j.1467-842X.2009.00541.x](https://doi.org/10.1111/j.1467-842X.2009.00541.x).
- García-Portugués, Eduardo (2022). *Notes for Nonparametric Statistics*. Version 6.5.7. ISBN 978-84-09-29537-1. URL: <https://bookdown.org/egarpor/NP-UC3M/> (vaadatud 19.04.2022).
- Guiglion, Guillaume *et al.* (2019). “4MOST Survey Strategy Plan”. *Published in The Messenger* vol. 175 pp. 17-21, 5 pages. ISSN: 0722-6691. DOI: [10.18727/0722-6691/5119](https://doi.org/10.18727/0722-6691/5119).
- Hollander, Myles, Douglas A. Wolfe ja Eric Chicken (2014). *Nonparametric Statistical Methods*. 3. väljaanne. Hoboken, New Jersey: John Wiley & Sons, Inc. ISBN: 978-0-470-38737-5.
- KDEpy* (n.d). URL: <https://kdepy.readthedocs.io/en/latest/> (vaadatud 12.02.2022).
- Perktold, Josef, Skipper Seabold ja Jonathan Taylor (2022). *KernelReg*. URL: https://www.statsmodels.org/dev/generated/statsmodels.nonparametric.kernel_regression.KernelReg.html (vaadatud 12.02.2022).

- Tempel, E. *et al.* (2020a). “An optimised tiling pattern for multi-object spectroscopic surveys: application to the 4MOST survey”. *Monthly Notices of the Royal Astronomical Society* 497.4, lk. 4626–4643. ISSN: 0035-8711, 1365-2966. DOI: [10.1093/mnras/staa2285](https://doi.org/10.1093/mnras/staa2285).
- (märts 2020b). “Probabilistic fibre-to-target assignment algorithm for multi-object spectroscopic surveys”. *Astronomy & Astrophysics* 635, A101. ISSN: 0004-6361, 1432-0746. DOI: [10.1051/0004-6361/201937228](https://doi.org/10.1051/0004-6361/201937228).
- Tenjes, Peeter ja Kaido Reivelt (2020). *Astronoomia. Õpik füüsikahuvilistele*. URL: <https://opik.fyysika.ee/index.php/book/view/71> (vaadatud 24.04.2022).
- Traat, Imbi ja Jaano Inno (1997). *Tõenäosuslik valikuuring*. Tartu: Tartu Ülikooli Kirjastus. ISBN: 9985-56-226-7.
- TreeKDE* (n.d). URL: <https://kdepy.readthedocs.io/en/latest/API.html#treekde> (vaadatud 12.02.2022).
- Walcher, C. Jakob *et al.* (2019). “4MOST Scientific Operations”. en. *The Messenger vol. 175* pp. 12-16, 5 pages. ISSN: 0722-6691. DOI: [10.18727/0722-6691/5118](https://doi.org/10.18727/0722-6691/5118).

Lisa 1. Taastatud tihedusfunktsioonid



Joonis 8: Vaatlusvalimi pealt taastatud tihedusfunktsioonid tähesuurusele kolmel erineval viisi. Üldkogumiks on võetud 100 000 objekti suurune juhuvalim kogu sihtkataloogist. Ülevalt alla: 1) Korrigeerimisfunktsioon $h(mag)$ 2) Nadaraya-Watsoni meetod ühe abitunnusega $\mathbf{x} = (\text{tähesuurus})$ 3) Nadaraya-Watsoni meetod kolme abitunnusega $\mathbf{x} = (\text{tähesuurus}, RA, Dec)$.



Joonis 9: Vaatlusvalimi pealt taastatud tihedusfunktsioonid punanihkele kolmel erineval viisi. Üldkogumiks on võetud iga meetodi puhul 100 000 objekti suurune juhuvalim kogu sihtkataloogist. Ülevalt alla: 1) Korrigeerimisfunktsioon $h(mag)$ 2) Nadaraya-Watsoni meetod ühe abitunnusega $\mathbf{x} = (\text{tähesuurus})$ 3) Nadaraya-Watsoni meetod kolme abitunnusega $\mathbf{x} = (\text{tähesuurus}, RA, Dec)$.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Elis Sams,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „4MOST vaatlusprogrammi tunnuste tihedusfunktsioonide taastamine vaatlusandmetelt“, mille juhendajad on Elmo Tempel ja Anne Selart, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.
2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commons litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.
3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.
4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Elis Sams

10.05.2022