TARTU ÜLIKOOL

LOODUS- JA TÄPPISTEADUSTE VALDKOND

MATEMAATIKA JA STATISTIKA INSTITUUT

DIANA SOKUROVA

Lokaalne pöördemeetod valikuuringutes

MATEMAATILISE STATISTIKA ERIALA

BAKALAUREUSETÖÖ (9 EAP)

JUHENDAJA: NATALJA LEPIK

TARTU 2018

Lokaalne pöördemeetod valikuuringutes

Bakalaureusetöö

Diana Sokurova

Lühikokkuvõte. Käesolevas bakalaureusetöös antakse ülevaade lokaalsest pöördemeetodist ning võrreldakse seda teiste tuntud valikumeetoditega, rakendades neid reaalsete andmetel. Andmed pärinevad hüpoteetilise küla *StatVillage* andmebaasist. Töö teooriaosas kirjeldatakse lühidalt teisi tuntud valikumeetodeid, täpsem ülevaade antakse lokaalsest pöördemeetodist ja tuuakse näide, kuidas seda kasutada. Praktilises osas rakendatakse Monte-Carlo simulatsiooni, et välja selgitada, milline valikumeetod annab kõige parema tulemuse *StatVillage* andmete korral. Lisaks sellele, lokaalse pöördemeetodiga leitud valimi tasakaalustatust üldkogumi teiste objektide suhtes võrreldakse lihtsa juhusliku valiku abil saadud valimiga.

CERCS teaduseriala: P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Märksõnad: valikuuringud, valikuteooria, statistiline hindamine, pöördemeetod, lokaalne pöördemeetod

The Local Pivotal Method in Survey Sampling

Bachelor's thesis

Diana Sokurova

Abstract. The purpose of this bachelor's thesis is to give an overview of the Local Pivotal Method and compare it with other well-known sampling methods, applied on the real data. Data comes from a hypothetical village "StatVillage". In the theoretical part, a brief overview of common sampling methods and detailed description of the Local Pivotal Method is given, which is followed with example of its application. In the practical part, Monte Carlo simulation is conducted to find out which sampling method gives better estimation for "StatVillage" data. At the end, the spacial balance of sample from the Local Pivotal Method is compared with Simple Random Sample.

CERCS research specialisation: P160 Statistics, operation research, programming, actuarial mathematics

Keywords: survey sampling, sample survey theory, statistical estimation, pivotal method, local pivotal method

Sisukord

Sissejuhatus

Valikuuringute teooria on teadus, mille põhilised eesmärgid on välja töötada selline valimi võtmise strateegia ja hinnangufunktsioon, et hinnangud huvipakkuvatele üldkogumi parameetritele oleksid võimalikult täpsed. Kõige tuntum valikumeetod on lihtne juhuslik valik, kus kõikidel objektidel on võrdne võimalus valimisse sattuda. Kui aga uurijal on olemas taustandmed või muud teadmised tulevase uuringu kohta, siis võib ta neid ära kasutada teistsuguse valikumeetodi läbiviimiseks, mis teatud olukordades annab täpsema hinnangu. Üheks selliseks meetodiks on kihtvalik, mille korral jagatakse üldkogum gruppideks ehk kihtideks mingi tausttunnuse järgi ja igas kihis rakendatakse uurija poolt määratud valikumeetodit. Teiseks on süstemaatiline, mille korral võetakse objekte järjest, ettemääratud ja fikseeritud sammuga. Süstemaatiline ja kihtvalik võtavad arvesse tausttunnuse väärtuseid ehk valimiobjektide asetsemist nende väärtuste suhtes. Kuid ülalkirjeldatud valikumeetodid ei kasuta ära objektide nn ruumilist asukohta. Üheks objektide asukohta arvessevõtvaks meetodiks on lokaalne pöördemeetod (ingl. *The Local Pivotal Method*), mis on aga praktikas vähe levinud. Käesoleva töö eesmärk on uurida seda valikumeetodit põhjalikult ning võrrelda teiste eespool nimetatud meetodiga.

Esimest korda kasutati nimetust *"The Pivotal Method"* valikuuringute teoorias 1998. aastal (Deville ja Tillé, 1998). Artiklis räägiti tagasipanekuta valikumeetoditest, kus valimisse sattumise tõenäosused jaotati kaheks komponendiks. Kõigepealt tutvustati nn lahutamismeetodit (ingl. *The Splitting Method*), mis teatud algotitmi järgi teisendab iteratiivselt objektide valimisse sattumise tõenäosuseid seni, kuni jõutakse nullist ja ühest koosneva tõenäosuste vektorini. Artikli lõpus tutvustati pöördemeetodit, mis võtab juhuslikult kaks objekti ja teisendab nende valimisse sattumise tõenäosuseid, kuni kõikide objektide tõenäosused võrduvad nulli või ühega (Deville ja Tillé, 1998). Seejärel tutvustas Guillaume Chauvet selle modifitseeritud varianti - nn järjestatud pöördemeetodit (Chauvet, 2012). 2012. aastal esitati esimest korda lokaalne pöördemeetod, mille eeliseks on valimi tasakaalustatus teiste objektide suhtes (Grafström, Lundström ja Schelin, 2012).

Töö teoreetilises osas tutvustatakse selliseid levinud valikumeetodeid, nagu lihtne juhuslik valik, süstemaatiline valik, kihtvalik ja selle kaks erijuhtu. Iga valiku puhul kirjeldatakse valiku algoritm, vajalikud esimest ja teist järku kaasamistõenäosuse valemid, kogusumma hinnangu valem, selle teoreetiline dispersioon ja dispersiooni hinnang. Esimest ja teist järku kaasamistõenäosused on olulised dispersiooni hinnangu leidmiseks. Kõik need valemid leitakse valikuuringus tuntud üldise hindamis- ja Sen-Yates-Grundy'i teoreemide kaudu, kus kasutatakse Horovitzi-Thompsoni hinnangut. Lähtuvalt töö eesmärgist võrrelda lokaalset pöördemeetodit teiste tuntud meetoditega, antakse teoreetilise osa lõpus põhjalik ülevaade sellest meetodist.

Praktikas on levinud olukorrad, kus seni tuntud valikumeetodid ei anna häid hinnanguid või neid meetodeid on võimatu rakendada.

Praktilises osas uuritakse kõikide eespool mainitud valikumeetodite hinnangute täpsust, kasutades selleks Monte-Carlo simuleerimist. Töös võrreldakse kõiki valikumeetodeid omavahel, et leida parim valikuviis parameetrite hindamiseks. Andmestikuks on kasutatud hüpoteetilist küla *StatVillage* ning hinnanguteks on valitatud ühe pideva ja ühe diskreetse tunnuse kogusumma.

Töö on koostatud dokumentide ettevalmistussüsteemi LaTeX abil, töö praktilises osas on kasutatud vabavara *R* versiooni 3.4.2 ning järgmisi pakette: *dplyr* (versiooni 0.7.4), *tidyr* (versiooni 0.8.0), *sampling* (versiooni 2.8), *BalancedSampling* (versiooni 1.5.2) ja *ggplot2* (versiooni 2.2.1).

1 Tähistused ja sõnavara

Kõik antud bakalaureusetöös kasutatud tähistused on toodud allpool. Lisaks on välja kirjutatud valikuuringute valdkonna kaks põhiteoreemi.

Üldised

$U = \{1, 2, \ldots, N\}$ - lõplik üldkogum, mis sisaldab objektidele vastavaid järjekorra numbreid
$s \subseteq U$ - lõplik valim
N - üldkogumi maht
n - planeeritud valimi maht
n_s - realiseerunud valimi maht
π_i - objekti i kaasamistõenäosus - tõenäosus, millega objekt i kaasatakse valimisse
π_{ij} -teist järku kaasamistõenäosust - objektide i ja j üheaegne valimisse kaasamistõenaosus
$f = \frac{n}{N}$ - valikusuhe, mis näitab kui suur osa üldkogumist võetakse valimisse
$\overline{Y} = \frac{1}{n}\sum_{i=1}^{N} y_i$ - tunnuse y üldkogumi keskmine
$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ - tunnuse y valimi keskmine
s_y^2, S_y^2 - uuritava tunnuse y dispersioon vastavalt valimis ja üldkogumis
$t = \sum_U y_i$ - tunnuse y kogusumma
$\hat{t}$ - hinnang tunnuse y kogusummale t
$V(\hat{t})$ - hinnangu $\hat{t}$ dispersioon
$\hat{V}(\hat{t})$ - hinnangu $\hat{t}$ dispersiooni hinnang

Süstemaatilise ja kihtvaliku korral

m - valikusamm süstemaatilise valiku korral
c - elementide jääk süstemaatilise valiku korral
H - kihtide koguarv
U_h - kihi h kõigi objektide hulk
$s_h \subset s$ - valim kihis h
N_h - kihi h kõigi objektide maht üldkogumis
n_h - kihi h maht valimis
$f_h = \frac{n_h}{N_h}$ - valikusuhe kihis h
$\overline{Y}_h$ - tunnuse y üldkogumi keskmine kihis h
$\overline{y}_h$ - tunnuse y valimi keskmine kihis h
S_{yh}^2 - uuritava tunnuse y üldkogumi dispersioon kihis h
s_{yh}^2 - uuritava tunnuse y valimi dispersioon kihis h
$t_h = \sum_{U_h} y_i$ - tunnuse y kogusumma kihis h

freim - üldkogumi objektide loend
nihketa hinnang - hinnang parameetrile θ, mille korral $E(\hat{\theta}) = \theta$

Teoreemid

Üldine hindamisteoreem (ÜHT). Üldkogumi kogusumma $t = \sum_{j \in U} y_i$ nihketa hinnang tagasipanekuta valiku korral on

$$\hat{t} = \sum_{i \in s} \frac{y_i}{\pi_i}.$$

Selle hinnangu dispersioon on

$$V(\hat{t}) = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}.$$

Dispersiooni nihketa hinnanguks $\pi_{ij} > 0$ korral on

$$\hat{V}(\hat{t}) = \sum_{i \in s} \sum_{j \in s} \left(1 - \frac{\pi_i \pi_j}{\pi_{ij}}\right) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

(Horvitz ja Thompson, 1952).

Sen-Yates-Grundy teoreem (SYG). Fikseeritud mahuga valiku korral saab hinnangu $\hat{t} = \sum_{i \in s} \frac{y_i}{\pi_i}$ dispersiooni esitada alternatiivsel kujul:

$$V(\hat{t}) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} (\pi_{ij} - \pi_i \pi_j) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

ja eeldusel, et $\pi_{ij} > 0 \, \forall i \neq j \in U$, dispersiooni $V(\hat{t})$ nihketa hinnang on

$$\hat{V}(\hat{t}) = -\frac{1}{2} \sum_{i \in s} \sum_{j \in s, j \neq i} \left(1 - \frac{\pi_i \pi_j}{\pi_{ij}}\right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

(Yates ja Grundy, 1953); (Sen, 1953).

2 Ülevaade tuntud valikumeetoditest

2.1 Lihtne juhuslik valik

Lihtne juhuslik valik on praktikas kõige lihtsam ning teoorias enim uuritud valik. Järgnev peatükk põhineb Imbi Traadi ja Janno Inno õpikul „Tõenäosuslik valikuuring“ (Traat ja Inno, 1997: 90-91). Eristatakse kahte sellist varianti: tagasipanekuta ja tagasipanekuga valik. Esimesel juhul valitud objekt eemaldatakse üldkogumist enne järgmist võtmist, teisel juhul mitte. Antud töös vaadeldakse ainult tagasipanekuta lihtsat juhuslikku valikut.

Algoritm

Lihtsa juhusliku valiku realiseerimiseks on välja töötatud erinevaid algoritme, siin on esitatud järjestusvaliku algoritm.

Olgu üldkogum $U = \{1, 2, \ldots, N\}$. Fikseeritakse valimi maht n. Kõigi n mahuliste valimite arv, mida U-st saab moodustada, on $M = C_N^n$.

1. Igale üldkogumi objektile $i = 1, ..., N$ seada vastavusse juhuslikud arvud ühtlasest jaotusest (üldiselt võib kasutada ükskõik millist pidevat jaotust)

$$u_1, ..., u_N, \quad u_i \sim U(0, 1).$$

2. Järjestada üldkogumi objektid saadud arvude u_i järgi kasvavalt:

$$u_{(i_1)} < u_{(i_2)} < ... < u_{(i_N)}.$$

3. Võtta valimisse esimesed n objekti.

Kaasamistõenäosused

Lihtsa juhusliku valiku korral avaldub objektile i vastav kaasamistõenäosus kujul:

$$\pi_i = \frac{n}{N} \quad \forall i \in U$$

ning teist järku kaasamistõenäosus kujul:

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)} \quad \forall i, j \in U, \quad i \neq j.$$

Kogusumma hinnang ja hinnangu dispersioon

Lihtsa juhusliku valiku korral avaldub kogusumma $t = \sum_U y_i$ nihketa hinnang kujul

$$\hat{t} = N\overline{y},$$

kus $\overline{y}$ on tunnuse y valimi keskmine.
Kogusumma hinnangu dispersioon on

$$V(\hat{t}) = N^2(1-f)S_y^2/n$$

ja dispersiooni hinnang on

$$\hat{V}(\hat{t}) = N^2(1-f)s_y^2/n,$$

kusjuures $f = \frac{n}{N}$ on valikusuhe,

$$S_y^2 = \frac{1}{N-1}\sum_{i\in U}(y_i - \overline{Y})^2$$

on tunnuse y dispersioon üldkogumis ja

$$s_y^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \overline{y})^2$$

on tunnuse y dispersioon valimis.

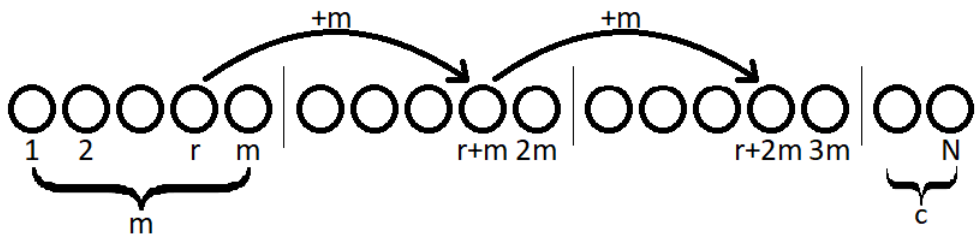
Neid valemeid saab tuletada teoreemidest ÜHT ja SYG ning tuletuskäik on toodud õpikus (Traat ja Inno, 1997: 92-93).

2.2 Süstemaatiline valik

Süstemaatiline valik on valikumeetod, mis põhineb elementide valimisel järjestatud freimist. Kõige levinum süstemaatilise valiku liik on võrdsete tõenäosustega meetod.

Algoritm

Olgu $U = \{1, ..., N\}$ ja valikusamm m on määratud uurija poolt.



Joonis 1: Süstemaatilise valiku näide

1. Esimene element võtta juhuslikult esimese m elemendi hulgast (edaspidi on tähistatud indeksiga r, vt. joonis 1).
2. Iga järgmine valimisse sattunud element on eelmine element pluss samm m.
3. Valimine lõpeb kui jõutakse suurima võimaliku indeksini, mis on üldkogumi mahust N väksem.

Kokku on võimalik saada m erinevat valimit. Iga sellise valimi saamise tõenäosus on $\frac{1}{m}$. Valimimaht n on süstemaatilise valiku korral juhuslik ja määratud sammuga m. Kehtib järgmine seos:

$$N = nm + c, \quad 0 \leq c < m. \tag{1}$$

Seega, realiseerunud valimimaht n_s võib olla kas $n + 1$, kui $r \leq c$, või n, kui $r > c$. (Traat ja Inno, 1997: 111)

Kaasamistõenäosused

Esimest järku kaasamistõenäosus on $\pi_i = \frac{1}{m}$ ja teist järku on

$$\pi_{ij} = \begin{cases} \frac{1}{m}, \text{ kui vahe } i \text{ ja } j \text{ vahel on sammu } m \text{ kordne;} \\ 0, \text{ vastasel juhul.} \end{cases}$$

Vajab märkmist, et paljude objektide korral on $\pi_{ij} = 0$, mis teeb võimatuks $\hat{V}(\hat{t})$ leidmise. (Traat ja Inno, 1997: 112)

Kogusumma hinnang ja hinnangu dispesioon

Kogusumma $t = \sum_U y_i$ nihketa hinnang on

$$\hat{t} = m \sum_s y_i.$$

Kuna kõikide elementidejaoks ei leidu süstemaatilise valiku korral teist järku kaasamistõenäosusi, siis pole võimalik saada dispersiooni nihketa hinnangut. Sel juhul kasutatakse mõnda teist nihkega hinnangut, tavaliselt lihtsa juhusliku valiku hinnangut:

$$\hat{V}(\hat{t}) = N^2(1-f)\frac{s_y^2}{n}.$$

Juhul, kui üldkogum on halvasti järjestatud ja valimis esineb tsüklilisus sammuga m, siis s_y^2 võib osutuda liiga väikeseks ja sel juhul $\hat{V}(\hat{t})$ võib tegelikku dispersiooni alahinnata. (Traat ja Inno, 1997: 113)

2.3 Kihtvalik

Kihtvaliku eesmärk on tagada uuritava tunnuse hinnangu suurem täpsus. Selleks valitakse mõni sobiv tausttunnus, mille järgi jagatakse üldkogum kihtideks. Kihte vaadeldakse üksteisest sõltumatute kogumitena, milles võib üldjuhul rakendada erinevaid valikumeetodeid. Kihid on hästi valitud, kui uuritava tunnuse väärtused on kihtides võimalikult homogeensed. Kihtvalikut kasutatakse sageli ka osakogumite hindamisel, kui soovitatakse leida hea täpsusega hinnanguid osakogumite kaupa. Sel juhul käsitletakse osakogumeid kihtidena. Tausttunnuseks sobib selline tunnus, mis on määratud kõikidel objektidel üldkogumis ja on teada enne uuringu läbiviimist. Taustunuseid võib olla ka rohkem kui üks, siis moodustatakse kihte tausttunnuste ristklassifitseerimise teel.

Olgu lõplik üldkogum $U = 1, ..., N$ jagatud H kihiks $U_1, ..., U_h, ..., U_H$ vastavate mahtudega $N_1, ..., N_h, ...N_H$, kusjuures

$$U = \bigcup_{h=1}^{H} U_h, \quad U_h \cap U_g = \emptyset, \text{ kui } h \neq g,$$

$$N = \sum_{h=1}^{H} N_h.$$

Valiku algoritm ja kaasamistõenäosused sõltuvad disainist, mida rakendatakse igas kihis eraldi. (Traat ja Inno, 1997: 125)

Kogusumma hinnang ja hinnangu dispersioon

Eesmärk on hinnata üldkogumi kogusumma $t = \sum_U y_i$ kihtide kogusummade kaudu:

$$t = \sum_{h=1}^{H} t_h,$$

kus $t_h = \sum_{U_h} y_i$ - uuritava tunnuse kogusumma kihis U_h. Kihtvaliku korral on nihketa hinnang üldkogumi kogusummale t järgmine:

$$\hat{t} = \sum_{h=1}^{H} \hat{t}_h.$$

Hinnangu $\hat{t}$ dispersioon on

$$V(\hat{t}) = \sum_{h=1}^{H} V(\hat{t}_h)$$

ja selle vastav nihketa hinnang

$$\hat{V}(\hat{t}) = \sum_{h=1}^{H} \hat{V}(\hat{t}_h),$$

kus $\hat{V}(\hat{t}_h)$ avaldub sõltuvalt rakendatud valikumeetoditest. (Traat ja Inno, 1997: 126)

2.3.1 Lihtne juhuslik kihtvalik

Lihtne juhuslik kihtvalik on praktikas väga levinud meetod, mille korral igas kihis rakendatakse lihtsat juhuslikku valikut tagasipanekuta. Valimimahud võivad olla kihtides erinevad. Sageli leitakse valimi mahud n_h võrdeliselt kihtide mahtudega N_h.

Algoritm

1. Jagada üldkogumi objektid kihtidesse ühe või mitme abitunnuse põhjal.
2. Igas kihis määrata/leida sobiv valimimaht.
3. Teostada lihtne juhuslik valik igas kihis eraldi (vt. peatükk 2.1).

Kogusumma hinnang ja hinnangu dispersioon

Lihtsa juhusliku valiku korral on kihi sees parameetri t_h hinnang järgmine:

$$\hat{t}_h = N_h \overline{y}_h,$$

kus $\overline{y}_h = \frac{1}{n_h} \sum_{s_h} y_i$ on valimikeskmine kihis U_h.
Lihtsa juhusliku kihtvaliku korral avaldub hinnang kogusummale $t = \sum_U y_i$ kujul

$$\hat{t} = \sum_{h=1}^{H} N_h \overline{y}_h,$$

dispersiooniga

$$V(\hat{t}) = \sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{S_{yh}^2}{n_h}$$

ja dispersiooni nihketa hinnanguga

$$\hat{V}(\hat{t}) = \sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{s_{yh}^2}{n_h}, \tag{2}$$

kus

$$S_{yh}^2 = \frac{1}{N_h - 1} \sum_{U_h} (y_i - \overline{Y}_h)^2,$$

$$s_{yh}^2 = \frac{1}{n_h - 1} \sum_{s_h} (y_i - \overline{y}_h)^2.$$

2.3.2 Süstemaatiline kihtvalik

Süstemaatilise kihtvaliku korral kasutatakse igas kihis lihtsa juhusliku valiku asemel süstemaatilist valikut. Võimaluse korral peaks uuritav tunnus olema igas kihis järjestatud täpsema hinnangu saavutamiseks. Selle jaoks saab kasutada mõnda tausttunnust, mis on uuritava tunnusega tugevalt korreleeritud.

Algoritm

1. Jagada üldkogumi objektid kihtidesse abitunnuse/abitunnuste põhjal.
2. Arvutada valikusamm m_h sõltuvalt igas kihis soovitud valimi mahust (vt. valem (1)).
3. Teostada süstemaatiline valik igas kihis eraldi (vt. peatükk 2.2).

Kogusumma hinnang ja hinnangu dispersioon

Süstemaatilise valiku korral on kihi sees parameetri t_h hinnanguks:

$$\hat{t}_h = m_h \sum_{i \in s_h} y_i,$$

kus m_h on valimis s_h elementide valimise samm.
Süstemaatilise kihtvaliku korral avaldub hinnang kogusummale $t = \sum_U y_i$ kujul

$$\hat{t} = \sum_{h=1}^{H} m_h \sum_{i \in s_h} y_i,$$

ja dispersiooni hinnanguks $\hat{V}(\hat{t})$ on lihtsa juhusliku kihtvaliku dispersiooni hinnang (vt. valem (2)).

3 Ülevaade pöördemeetoditest

3.1 Juhuslik pöördemeetod

Pöördemeetod (nimetatakse ka juhuslikuks pöördemeetodiks) on valikumeetod tagasipanekuta, mida võib rakendada nii võrdsete kui ka ebavõrdsete kaasamistõenäosuste korral. Valikumeetod põhineb objektide kaasamistõenäosuste järjekindlal uuendamisel, kuni kõikide objektide kaasamistõenäosused võrduvad kas 0 või 1. Igal sammul uuendatakse kahe objekti kaasamistõenäosused, kus ainult ühe objekti kaasamistõenäosus muutub võrdseks kas nulli või ühega. Järgnev peatükk põhineb artiklil (Grafström jt, 2012).

Algoritm 1: Juhuslik pöördemeetod

Olgu π_i' - uuendatud kaasamistõenäosus. Objekt i on lõplik (ingl. *finished*), kui $\pi_i' = 0$ või $\pi_i' = 1$. Kui objekt on lõplik, siis teda enam algoritmis ei kasutada ja seega ei ole tal võimalust uuesti valimisse sattuda.

1. Valida juhuslikult kaks objekti i ja j, mille kaasamistõenäosused on vastavalt π_i ja π_j.
2. Muuta vektorit (π_i, π_j) järgmise uuendamisreegli abil:
   (a) Kui $\pi_i + \pi_j < 1$, siis
   $$(\pi_i', \pi_j') = \begin{cases} (0, \pi_i + \pi_j), \text{ tõenäosusega } \frac{\pi_j}{\pi_i + \pi_j} \\ (\pi_i + \pi_j, 0), \text{ tõenäosusega } \frac{\pi_i}{\pi_i + \pi_j}. \end{cases}$$
   (b) Kui $\pi_i + \pi_j \geq 1$, siis
   $$(\pi_i', \pi_j') = \begin{cases} (1, \pi_i + \pi_j - 1), \text{ tõenäosusega } \frac{1-\pi_j}{2-\pi_i-\pi_j} \\ (\pi_i + \pi_j - 1, 1), \text{ tõenäosusega } \frac{1-\pi_i}{2-\pi_i-\pi_j}. \end{cases}$$
3. Alustada jälle sammust 1, kuni kõik objektid on muutunud lõplikuks. (Deville ja Tillé, 1998)

3.2 Lokaalne pöördemeetod

Lokaalne pöördemeetod on loodud selleks, et saavutada tasakaalustatud valikut. Lokaalse pöördemeetodi korral uuendatakse kahe lähima objekti kaasamistõenäosused vastavalt punktis 2 kirjeldatud uuendamisreeglile (vt. algoritm 1). Lähimad objektid on need, mille vaheline kaugus

on kõige väiksem. Kaugust leitakse kaugusfunktsiooni abil. Kõige levinum kaugusfunktsioon on

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2},$$

kus k on valitud tausttunnuste (näiteks koordinaatide) arv, $\mathbf{x} = (x_1, ..., x_k)$ on vektor, mis koosneb k tunnuse väärtusestest objektil x. Objekti y tähistused on analoogilised x-i tähistustega. Lähimate objektide määramiseks on kaks viisi. Esimisel viisil nõutakse, et objektid i ja j oleksid lähimad naabrid teineteisele. Teisel viisel piisab, et j oleks lähim naabel objektile i.

Algoritm 2: Lokaalne pöördemeetod I

1. Valida juhuslikult objekt i.
2. Leida objekt j, mis on lähim naaber objektile i. Kui kahel või rohkemal objektil on kauguse väärtus objektini i võrdne, siis valida nende vahel üks objekt juhuslikult võrdse tõenäosusega.
3. Kui i on samuti lähim naaber objektile j, siis uuendada nende kaasamistõenäosused vastavalt uuendamisreeglile (vt. algoritm 1 punkt 2). Vastasel juhul korrata algoritmi 2 alates punktist 1.
4. Kui kõik objektid on lõplikud, siis valikuprotsess on lõpetatud. Vastasel juhul korrata algoritmi 2 alates punktist 1. (Grafström jt, 2012)

Algoritm 3: Lokaalne pöördemeetod II

1. Valida juhuslikult objekt i.
2. Leida objekt j, mis on lähim naaber objektile i. Kui kahel või rohkemal objektil on kauguse väärtus objektini i võrdne, siis valida nende vahel üks objekt juhuslikult võrdse tõenäosusega.
3. Uuendada objektide i ja j kaasamistõenäosused vastavalt uuendamisreegelile (vt. algoritm 1 punkt 2).
4. Kui kõik objektid on lõplikud, on valikuprotsess lõpetatud. Vastasel juhul korrata algoritmi 3 alates punktist 1. (Grafström jt, 2012)

Kogusumma hinnang ja hinnangu dispersioon

Antud valikumeetodi korral pole eraldi valemit kaasamistõenäosusele, seega kogusumma hindamiseks kasutatakse üldise hindamisteoreemi hinnangut:

$$\hat{t} = \sum_{i \in s} \frac{y_i}{\pi_i}$$

ja selle hinnangu dispersiooni (vt. peatükk 1).

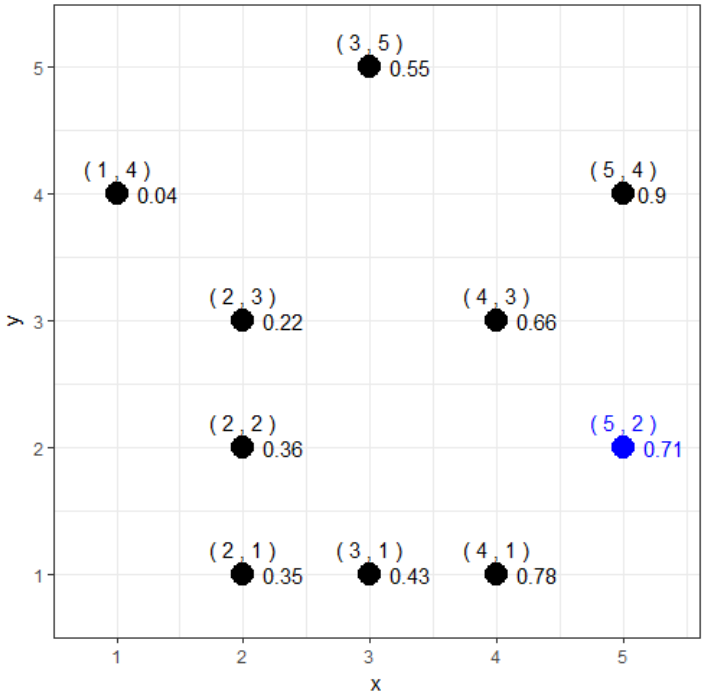
3.2.1 Lokaalse pöördemeetodi I näide

Olgu üldkogumi maht $N = 10$ ja soovitakse saada valimit mahuga $n = 5$, kasutades lokaalset pöördemeetodit I. Iga üldkogumi objekt on punkt ruudustikus 5×5, mille kohta on teada abiinformatsioonina koordinaadid x_i ja y_i ning kaasamistõenäosused $\pi_i, i \in \{1, \ldots, 10\}$. Suurused on kantud tabelisse 3.2.1.

Tabel 1: Näiteandmestik

i	x_i	y_i	π_i
1	1	4	0.04
2	5	2	0.71
3	2	1	0.35
4	2	2	0.36
5	2	3	0.22
6	5	4	0.90
7	3	5	0.55
8	3	1	0.43
9	4	1	0.78
10	4	3	0.66

Objektide valimiseks kasutatakse algoritmi 2, mille põhjal kõigepealt võetakse juhuslikult esimene objekt. Olgu selleks punkt 2 koordinaatidega $(5, 2)$. Näiteandmestik on toodud ka joonisena, kus valitud objekt on sinist värvi.



Joonis 2: Näiteandmestik ruudustikus koos juhuslikult valitud objektiga ja algsete kaasamistõenäosustega

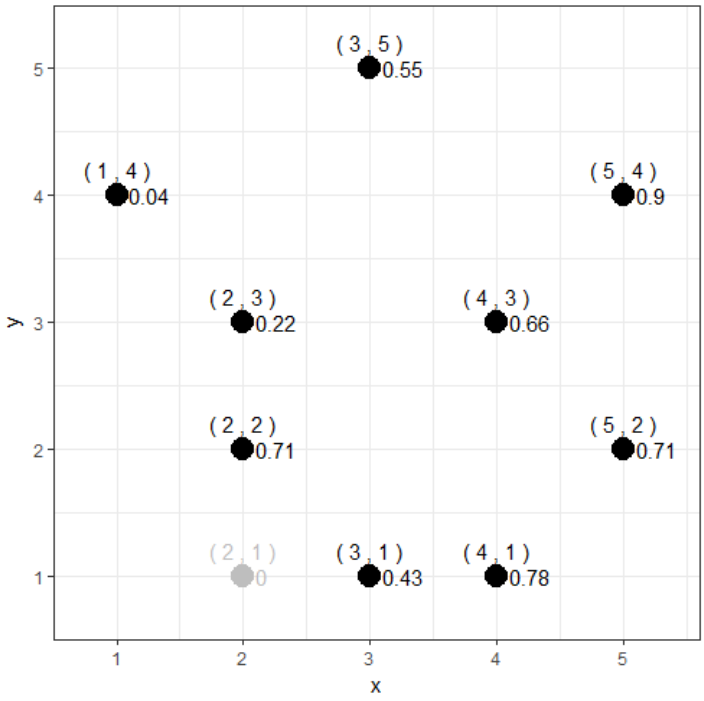
Järgmisena leitakse selle lähim naaber. Antud juhul on kaks lähimat punkti koordinaatidega $(4, 3)$ ja $(4, 1)$, seega lähim naaber valitakse nende vahel juhuslikult. Olgu valitud punkt $(4, 1)$. Seejärel kontrollitakse, kas selle punkti lähim naaber on samuti punkt $(5, 2)$. Joonisel 2 on näha, et punkti $(4, 1)$ lähim naaber on hoopis punkt $(3, 1)$, nende vaheline kaugus võrdub ühega. Siis alustatakse valikuprotsessi uuesti.

Olgu järgmise juhuslikult valitud punkti koordinaadid $(2, 1)$. Selle punkti lähimad naabrid on koordinaatidega $(2, 2)$ ja $(3, 1)$. Nendest valitakse juhuslikult punkt koordinaatidega $(2, 2)$. Selle punkti lähimate naabrite hulgas ongi esimesena valitud punkt koordinaatidega $(2, 1)$ (vt. Lisa 1), seega algoritm jätkub nende kahe punkti kaasamistõenäosuste muutmisega. Vastavalt uuendamisreeglile (vt. algoritm 1 punkt 2):

$$\pi_i + \pi_j = 0.36 + 0.35 = 0.71 < 1$$

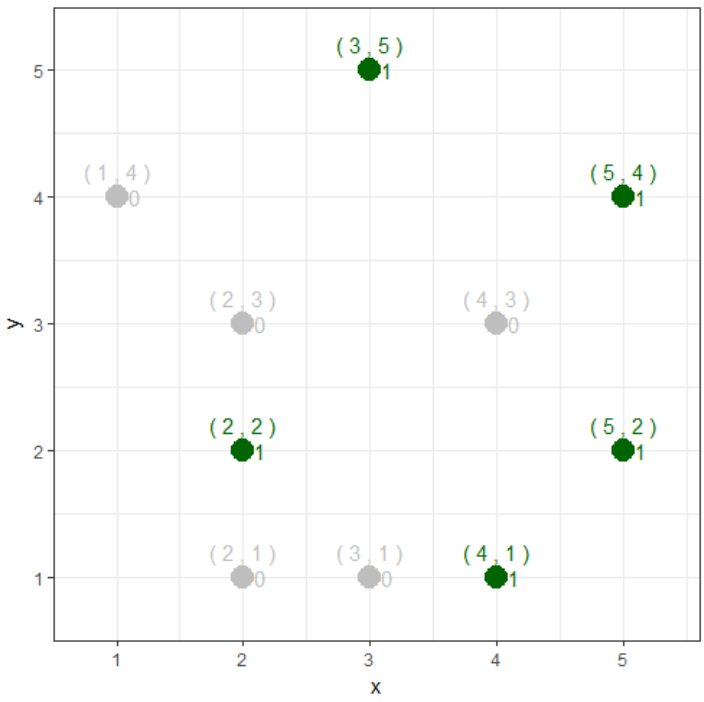
$$
\begin{aligned}
(\pi_i', \pi_j') &= \begin{cases} (0, \pi_i + \pi_j), \text{ tõenäosusega } \frac{\pi_j}{\pi_i + \pi_j} \\ (\pi_i + \pi_j, 0), \text{ tõenäosusega } \frac{\pi_i}{\pi_i + \pi_j} \end{cases} \\
&= \begin{cases} (0, 0.71), \text{ tõenäosusega } \frac{0.36}{0.71} \\ (0.71, 0), \text{ tõenäosusega } \frac{0.35}{0.71} \end{cases} \\
&= \begin{cases} (0, 0.71), \text{ tõenäosusega } 0.5070 \\ (0.71, 0), \text{ tõenäosusega } 0.4930. \end{cases}
\end{aligned}
$$

Antud juhul muudetud kaasamistõenäosused on toodud joonisel 3:



Joonis 3: Näiteandmestiku pärast 2. sammu

Joonisel 3 on näha, et punkt koordinaatidega $(2, 1)$ ei satu valimisse ning järmistel sammudel ei vaadelda seda enam teiste punktide naabrina. Algoritmi jätkatakse, kuni kõik elemendid on lõplikud. Antud andmestiku lokaalse pöördemeetodiga valimini on jõutud 12 sammuga ning lõplik valim on toodud joonisel 4, kus rohelised punktid on kõik lõplikus valimis.



Joonis 4: Näiteandmestikust saadud valim

4 Simuleerimisnäide

Selles peatükis rakendatakse lokaalset pöördemeetodit reaalsete andmete peal ja uuritakse, kui head on selle meetodiga leitud hinnangud võrreldes teiste tuntud meetoditega. Kokku koostatakse selleks 6 valimit järmiste valikumeetodite abil:

1. lihtne juhuslik valik,
2. lihtne juhuslik kihtvalik,
3. süstemaatiline kihtvalik,
4. juhuslik pöördemeetod,
5. lokaalne pöördemeetod I,
6. lokaalne pöördemeetod II.

Töös on vaadeldud kahte erinevat tunnust: üks on pidev ja teine diskreetne tunnus. Otsustusreegliks parima valikumeetodi leidmisel on uuritava tunnuse kogusumma kõige väiksema hinnangu dispersioon $V(\hat{t})$. Teoreetilisi dispersioone hinnatakse Monte-Carlo meetodil võttes üldkogumist korduvaid valimeid

Iga valiku korral võetakse 1000 valimit fikseeritud mahuga. Seejärel arvutatakse iga valimi põhjal hinnagud huvipakkuvale parameetrile $\hat{t}$ nii pideva kui ka diskreetse tunnuse korral. Viimaseks arvutatakse saadud hinnangute Monte-Carlo keskmine ja hinnangute Monte-Carlo standardviga:

$$E_{MC}(\hat{t}) = \frac{1}{1000}\sum_{k=1}^{1000}\hat{t}_k, \tag{3}$$

$$\sqrt{V_{MC}(\hat{t})} = \sqrt{\frac{1}{999}\sum_{k=1}^{1000}(\hat{t}_k, -E_{MC}(\hat{t}))^2}, \tag{4}$$

kus $\hat{t}_k$ on $k.$ valimi põhjal leitud hinnang. Lisaks sellele soovitakse veel kontrollida, kas pöördemeetodite abil hinnangud on nihketa. Selle jaoks võrreldakse saadud hinnangud tegeliku väärtusega.

4.1 Hüpoteetilise küla *StatVillage* kirjeldus

Erinevate valikumeetodite võrdlemiseks kasutatakse külast *StatVillage* pärenevaid andmeid. Järgnev informatsioon küla *StatVillage* kohta pärineb selle küla looja Carl James Schwarz artiklist (Schwarz, 1997). *StatVillage* on hüpoteetiline küla, mille aluseks on tegelikud andmed.

Andmed pärenevad rahvaloendusuuringust, mis toimus 1991. aastal Kanadas.

Küla *StatVillage* on üsna väike ja selle leibkonnad on korrapäraselt paigutatud 128 plokki, kusjuures iga plokk koosneb 8 majast, mis on paigutatud ümber keskse südamiku. Kokku on 1024 maja. Seega moodustavad küla *StatVillage* majad ristkülikukujulise plokkide võrgustiku, kus iga plokk sisaldab kaheksat maja. Igale majale vastab ploki number ja plokisisene majanumber. Järgnevalt on näitena toodud ploki 12 kuju.

Tabel 2: Näide küla *StatVillage* plokk nr.12

1	2	3
4	**12**	5
6	7	8

Küla *StatVillage* iga leibkonna kohta on mõõdetud 48 tunnust. Neid on võimalik jagada viieks rühmaks:

- Demograafilised tunnused - pere suurus ja kooslus vanuseklassi ja soo järgi;
- Sissetulekuid puudutavad tunnused - investeeringud, riiklikud toetused jne;
- Hõivatus tööga;
- Eluaset puudutavad andmed - tüüp, vanus, omanditüüp, väärtus, igakuised elamiskulud jne;
- Andmed kuni kahe perepea kohta (täiskasvanud, kes vastutavad pere heaolu eest) - vanus, sugu, amet, emakeel, haridus, tööalane staatus jne.

Selles asulas on suurema sissetulekuga elanikud koondunud põhjaossa ning vaesemad lõunasse.

Külast *StatVillage* on kolm eri suurusega varianti:

- *Micro village* - 36 plokki;
- *Mini village* - 60 plokki;
- *Maximal village* - 128 plokki.

Käesolevas töös kasutatakse versiooni *Maximal village*.

4.2 Uuritavate tunnuste valik

Praktilise osa eesmärk on anda ülevaade sellest, kuidas hinnatakse pideva ja diskreetse tunnuse kogusummat erinevate valikumeetodite korral. Pidevaks tunnuseks valiti leibkonna sissetulek (*total income of household* = *totinch*). See näitab kogu sissetulekut, mida said kõik 15-aastased ja vanemad isikud leibkonnas kalendriaastal 1990. Kuna hinnatakse tunnuse kogusummat, siis aastane kõikide leibkondade sissetuleku kogusumma saaks liiga suureks arvuks. Seega otsustati hinnata keskmise kuulise sissetuleku kogusummat. Selleks oli tehtud uus tunnus *moninch*, mis näitab leibkonna keskmist kuulist sissetulekut ja arvutatakse järgmiselt:

$$moninch = \frac{totinch}{12}.$$

Diskreetseks tunnuseks valiti leibkonna suurus (*household size* = *hhsize*). Kuna hinnatakse kõikides leibkondades olevate inimeste arvu, siis tulemuseks saadakse hinnang küla *StatVillage* rahvaarvule.

Mõlema tunnuse kogusumma hinnang leitakse kuuest valimist, mis on saadud erinevaid valikumeetodeid kasutades. Nende hulgas on kaks varianti kihtvalikust, mis nõuavad tausttunnust kihtide moodustamiseks. Selle tunnuse abil moodustatud kihid peavad olema uuritava tunnuse suhtes võimalikult homogeensed ning süstemaatilise kithvaliku jaoks peab tausttunnus olema uuritava tunnusega korreleeritud. Autori arvates leibkonna sissetuleku saajate arv (*number of income recipients in household* = *nuirh*) sobib kihtide moodustamiseks, sest on otseselt seotud leibkonna sissetuleku ja leibkonna suurusega. Seda kinnitavad korrelatsioonikordajad, mille väärtused on toodud tabelis 3.

Tabel 3: Korrelatsiooni kordajad

Tunnused	*moninch*	*hhsize*
nuirh	0.49	0.64

Valimi mahuks kihis võetakse üldjuhul proportsionaalne arv kihttunnuse sagedusega üldkogumis. Tabelis 4 on toodud tunnuse *nuirh* väärtused ja nende sagedused.

Tabel 4: Tunnuse *nuirh* sagedustabel

Pole saajat	Üks saaja	2 saajat	3 saajat	4 saajat	5 ja rohkem saajat
8	248	516	146	81	25

Kuna esimese ja viimase väärtuse sagedused on väiksed, otsustati kodeerida tunnust *nuirh* ümber järmiselt:

- „1“ näitab, et leibkonnas on 0 või 1 sissetuleku saajat;
- „2“ näitab, et leibkonnas on kaks saajat;
- „3“ näitab, et leibkonnas on kolm saajat;
- „4“ näitab, et leibkonnas on vähemalt 4 saajat.

Nende väärtuste sagedused ja jaotused üldkogumis on toodud tabelis 5.

Tabel 5: Tunnuse *nuirh* uuendatud sagedused ja jaotused

Sissetuleku saajate arv	0 või 1 saajat	2 saajat	3 saajat	4 ja rohkem saajat
Sagedused	248	516	146	106
Jaotused	25%	50%	15%	10%

Lokaalse pöördemeetodi teostamiseks on kasutanud ploki ja maja numbreid. Kuna kihtvaliku korral on juba eeldatud, et leibkonna sissetuleku saajate arv on teada, siis lisatakse lokaalse pöördemeetodi korral see abitunnuseks.

Kokkuvõttes kasutakse töös viite tunnust: ploki number, maja number, leibkonna keskmine kuuline sissetulek, inimeste arv leibkonnas ja sissetuleku saajate arv leibkonnas.

4.3 Valiku teostamine

Käesoleva töö praktikalises osas soovitakse leida parim valikumeetod tunnuste *moninch* ja *hhsize* kogusummale. Selleks kasutati Monte-Carlo simmulatsiooni kuue erineva valikumeetodi korral. Hinnangu leidmiseks moodustati iga valikumeetodi puhul 1000 valimit ja nende põhjal leiti tunnuste kogusummade hinnangud. Valimi mahuks valiti 300 leibkonda, mis moodustab ligikaudu 30% üldkogumist. Katsed on läbi viidud vabavara *R* abil, mõned koodid tõenäosuslike valikute sooritamiseks on esitatud Sören Mirski bakalaureusetöös (Mirski, 2017).

Kõigepealt sooritati lihtne juhuslik valik, mille puhul piisab ainult üldkogumi mahu ja soovitud

valimi mahu teadmisest. Seejärel sooritati kihtvalikud: lihtne juhuslik kihtvalik ja süstemaatiline kihtvalik, mille korral määratati kihtide mahud. Valimi mahud on leitud võrdeliselt kihtide mahtudega üldkogumis (vt. tabel 6). Kihttunnuseks on kasutatud tunnust *nuirh*.

Tabel 6: Tunnuse *nuirh* sagedused valimis

Sissetuleku saajate arv	0 või 1 saajat	2 saajat	3 saajat	4 ja rohkem saajat	Kokku
Sagedused üldkogumis	248	516	146	106	1024
Kihi osakaal	25%	50%	15%	10%	100%
Sagedused valimis	75	150	45	30	300

Süstemaatilise kihtvaliku korral nõutakse veel kaasamistõenäosusi, milleks otsustati võtta $\frac{1}{m_h}$, kus valikusamm m_h leitakse igas kihis eraldi vastavalt võrrandile 1. Valikusammu m_h on leitud järgmisest valemist: $m_h = \frac{N_h}{n_h}$. Kõikide ülalkirjutatud valikute korral on kasutatud koodid võetud Mirski bakalaureusetööst (Mirski, 2017).

Järgnevalt sooritati pöördemeetodid: juhuslik pöördemeetod, esimene ja teine lokaalne pöördemeetod. Selleks kasutati lisapaketti *BalancedSampling*. Juhusliku pöördemeetodi rakendamiseks kasutati funktsiooni *rpm(prob=p)*, mille argumendiks on kaasamistõenäosuste vektori. Antud katsel võetakse kaasamistõenäosuseks valikusuhe $f = \frac{n}{N}$. Funktsiooni tulemuseks on valikuvektor, mille väärtus on üks, kui objekt sattus valimisse ja null vastasel juhul. Juhusliku pöördemeetodi kasutamise kood, kus andmestik *Yldkogum* on küla *StatVillage* andmestik:

```
install.packages("BalancedSampling") #paketi sisselaadimine
#käivitatakse ainult esimesel korral
library(BalancedSampling) #pöördemeetodi soritamiseks

N = 1024; # üldkogumi maht
n = 300; # valimi maht
p = rep(n/N,N); # kaasamistõenaosuste vektor
```

```
s = rpm(p) #valiku vektor
valimRPM = Yldkogum[s, ] #valim
```

Esimese lokaalse pöördemeetodi jaoks kasutati funktsiooni *lpm1(prob=p, x=maatriks)* ja teise jaoks funktsiooni *lpm2(prob=p,x=maatriks)*. Mõlema funktsiooni puhul võetakse argumendiks *prob* kaasamistõenäosuste vektor, mis on sama juhusliku pöördemeetodi korral ning argumendiks *x* maatriks, mis koosneb abitunuste väärtustest. Abitunnusteks valiti ploki number, maja number ja sissetulekusaajate arv. Mõlema funktsiooni tulemuseks on valikuvektor. Lokaalse pöördemeetodi I ja II kasutamise kood on järgmine:

```
#pakett on sama
library(BalancedSampling) #pöördemeetodi soritamiseks
install.packages("dplyr")
library(dplyr) # töötamine tabeliga

N = 1024; # üldkogumi maht
n = 300; # valikumaht
p = rep(n/N,N); # kaasamistõenäosus
maatriks<- Yldkogum %>%
            select(block,unit,nuirh) %>%
            as.matrix()
#abitunnuste väärtuste maatriks

#Lokaalne pöördemeetod I
s1=lpm1(p,maatriks) #valikuvektor
valimLPM1=Yldkogum[s1, ] #valim

#Lokaalne pöördemeetod II
s2=lpm2(p,maatriks) #valikuvektor
valimLPM2=Yldkogum[s, ] #valim
```

4.4 Tulemused

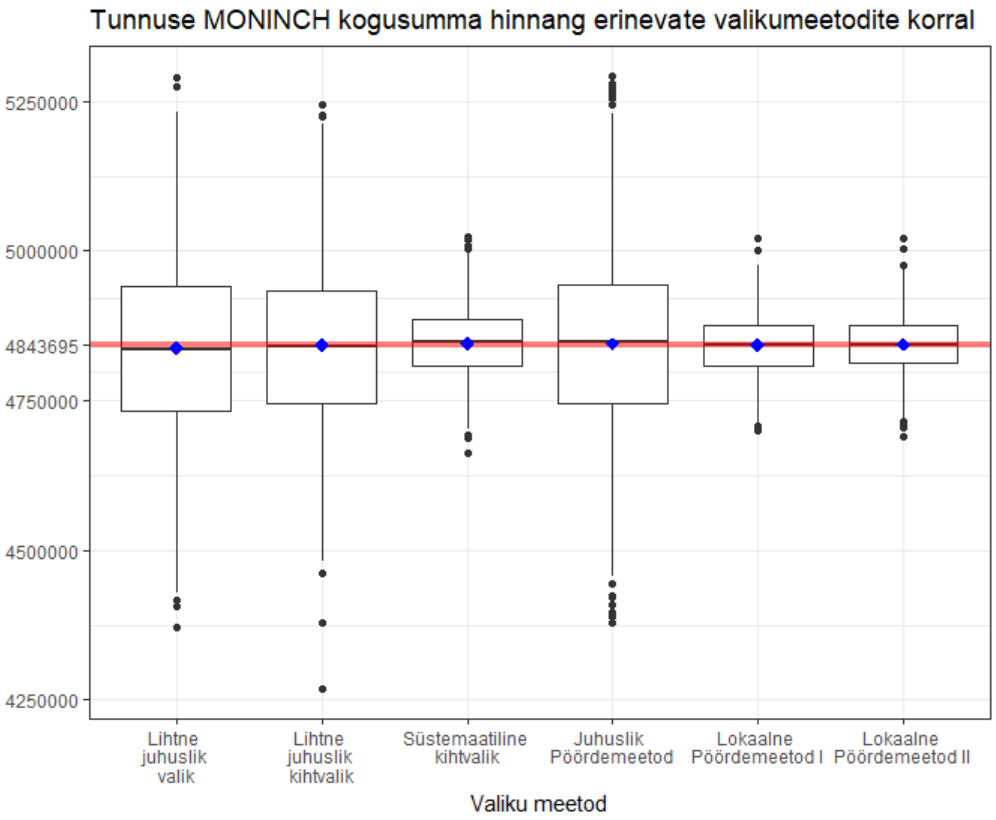
Edaspidi vaadeldakse Monte-Carlo simulatsiooni, kus Monte-Carlo keskmise ja standardvea leidmiseks on kasutatud valemeid (3) ja (4).

Tabel 7: Tunnuse *moninch* hinnangud ja standardsed vead

Tegelik väärtus	4 843 695	
Valikumeetod	$E_{MC}(\hat{t})$	$\sqrt{V_{MC}(\hat{t})}$
Lihtne juhuslik valik	4 838 280	152 718.57
Lihtne juhuslik kihtvalik	4 842 568	139 327.05
Süstemaatiline kihtvalik	4 845 962	58 380.02
Juhuslik pöördemeetod	4 844 994	156 132.98
Lokaalne pöördemeetod I	4 843 181	50 190.02
Lokaalne pöördemeetod II	4 844 111	49 183.08

Tabelis 7 on toodud Monte-Carlo simulatsioonist saadud hinnangud ja nende standardvead kuue erineva valiku korral. Tegelikule väärtusele lähimad hinnangud on saadud mõlemate lokaalsete pöördemeetodite korral (vahe absoluutväärtused on esimese ja teise meetodi korral vastavalt 514 ja 416). Kuna vahe on üsna väike, siis on võimalik järeldada, et mõlemad lokaalse pöördemeetodi abil saadud hinnangud on nihketa. Kõige suurem nihe on tulnud lihtsa juhusliku valiku korral. Võrreldes standardvigasid, on kõige suurem viga saadud juhusliku pöördemeetodi korral, kus hinnangu standardviga on isegi suurem, kui lihtsa juhusliku valiku korral. Lihtne juhuslik kihtvalik annab täpsema hinnangu kui lihtne juhuslik valik, kuid ei kuulu parimate valikumeetodite hulka. Kõige täpsemaid hinnanguid annavad süstemaatiline kihtvalik, lokaalne pöördemeetod I ja lokaalne pöördemeetod II, kus mõlemas lokaalses pöördemeetodis on standrardvead väikese erinevusega.

Järgmisel joonisel on toodud karpdiagrammid, mis kujutavad Monte-Carlo meetodiga leitud hinnanguid iga valikumeetodite korral. Karpdiagrammi horisantaaljooned näitavad kvartiile (sealhulgas mediaani), diagrammi servadel asetsevad maksimaalne ja minimaalne väärtus. Punktidena märgistatakse need väärtused, mis asuvad mediaanist kaugemalt kui poolteist kvariilide vahet. Joonisele on lisatud ka punane joon, mis näitab kõikide leibkondade tegelikku keskmise kuulise sissetuleku kogusummat, mille väärtus on 4 843 695. Sinise rombiga on märgistatud Monte-Carlo meetodi keskväärtus, mille väärtused olid toodud tabelis 7.



Joonis 5: Tunnuse *moninch* karpdiagrammid

Joonisel 5 on näha, et süstemaatilise kihtvaliku, lokaalse pöördemeetodi I ja lokaalse pöördemeetodi II hinnangud on kõige täpsemad ning samal ajal lokaalse pöördemeetodi II korral hinnangute ülemise ja alumise kvartiilide vahe on kõige väiksem. See tähendab, et pool selle meetodiga moodustatud hinnangutest on väga lähedal tegelikule väärtusele.

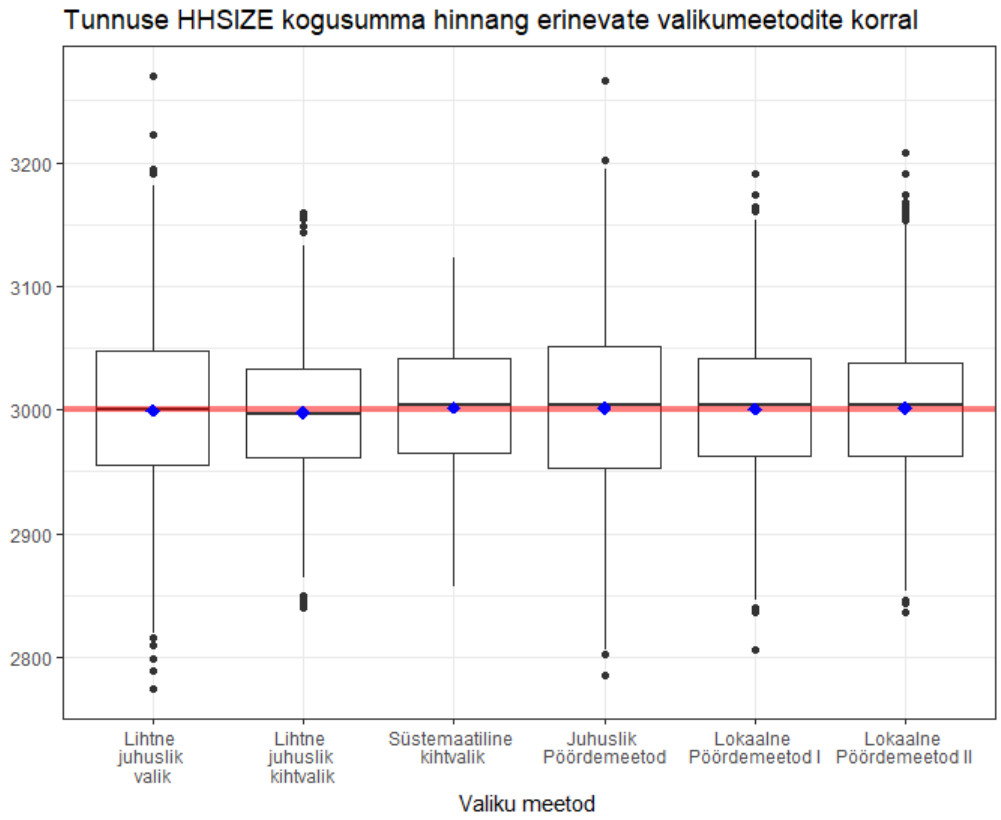
Järgnevalt võetakse vaatluse alla diskreetse tunnuse *nuirh* käitumine erinevate valikumeetodite korral.

Tabel 8: Tunnuse *hhsize* hinnangud ja standardsed vead

Tegelik väärtus	3 000	
Valikumeetod	$E_{MC}(\hat{t})$	$\sqrt{V_{MC}(\hat{t})}$
Lihtne juhuslik valik	2 999.64	73.22
Lihtne juhuslik kihtvalik	2 997.93	54.13
Süstemaatiline kihtvalik	3 002.18	51.15
Juhuslik pöördemeetod	3 001.38	71.17
Lokaalne pöördemeetod I	3 000.8	60.19
Lokaalne pöördemeetod II	3 001.44	58.42

Tabelis 8 toodud hinnangud erinevad tegelikust väärtusest väga vähe. Kõige kaugem hinnang tegelikust väärtusest on saadud süstemaatilise kihtvaliku korral. Samal ajal lihtne juhuslik valik annab kõige suurema standardveaga hinnangu, mis oli oodatav, sest lihtne juhuslik valik ei kasuta ühtegi tausttunnust. Juhusliku pöördemeetodi täpsus ei erine palju lihtsa juhusliku valiku täpsusest. Kõige täpsem hinnang standardvea mõttes on saadud süstemaatilisest kihtvalikust, selle hinnangu standardviga on 51.15. Mõlemad lokaalsed pöördemeetodid ei andnud parimat ega halvimat hinnangut, aga lokaalne pöördemeetod II on veidi täpsem, kui lokaalne pöördemeetod I.

Järgmine karpdiagramm on moodustud sama eeskirjaga nagu joonisel 5



Joonis 6: Tunnuse *hhsize* karpdiagrammid

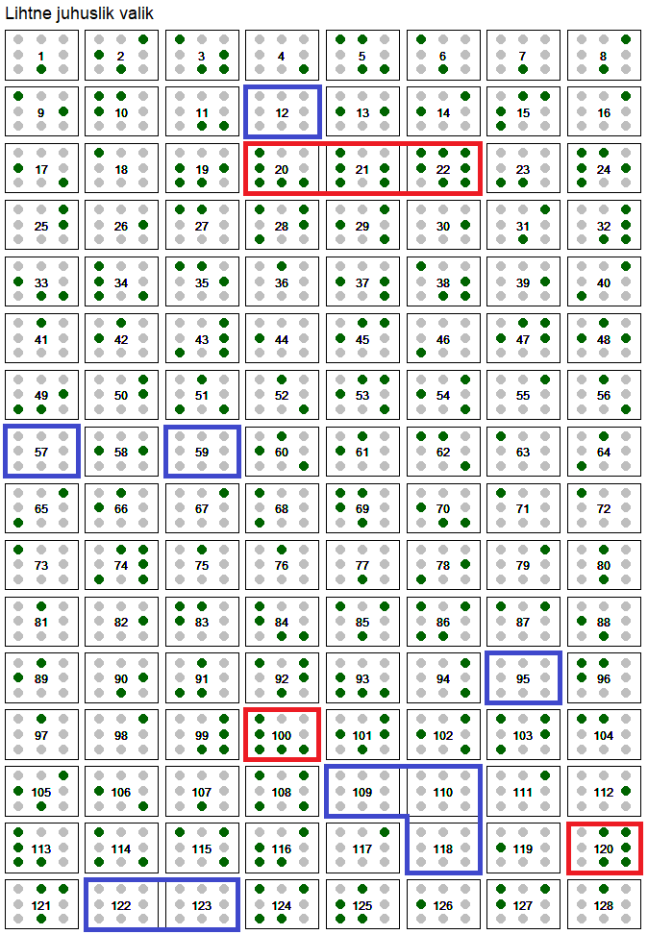
Joonisel 6 on näha, et erinevate valikute jaotused ei erine üksteisest peaaegu üldse . Vastupidiselt ootustele, ei anna mõlemad lokaalsed pöördemeetodid parimat tulemust. Kõige parem hinnang diskreetse tunnuse *hhsize* kogusummale on saadud kihtvalikutest: nii süstemaatilisest kui ka lihtsast juhuslikust valikust. Jääb silma, et süstemaatilise kihtvaliku korral ei ole ühtegi erindit, mis kinnitab, et antud juhul süstemaatiline kihtvalik on parim valikumeetod. Samal ajal lihtsa juhusliku valiku korral ülemise ja alumise kvartiili vahe on väiksem, mis annab hea eelise ka selle kasutamiseks.

Kokkuvõttes, juhusliku pöördemeetodi ehk pöördemeetodi täpsus on sarnane lihtsa juhusliku valiku täpsusega, kuid pöördemeetodi rakendamine on raskem ja aeganõuduvam. Seega juhuslikul pöördemeetodil ei leidu ühtegi eelist. Eelnevast johtuvalt on soovitatav kasutada süstemaatilist kihtvalikut juhul, kui soovitakse minimiseerida aega, või üht lokaalsetest pöördemeetoditest, kui ajakulu pole oluline. Samal ajal lokaalne pöördemeetod II, mis on lihtsam kasutamiseks, annab täpseima hinnangu. Pideva tunnuse korral soovitakse teha valik, kasutades süstemaatilist kihtvalikut juhul, kui eelduseks on aeg, või üht lokaalsest pöördemeetodist, kui eelduseks on täpsus. Diskreetse tunnuse *nuirh* puhul ei anna mõlemad lokaalsed pöördemeetodid häid hinnanguid. Antud juhul on parimad kasutamiseks lihtne juhuslik või süstemaatiline kihtvalik. Kui võrrelda mõlemate lokaalste pöördmeetodite täpsust, siis lokaalne pöördmeetod II annab

täpsema hinnangu kui lokaalne pöördmeetod I.

5 Valimi visualiseerimine

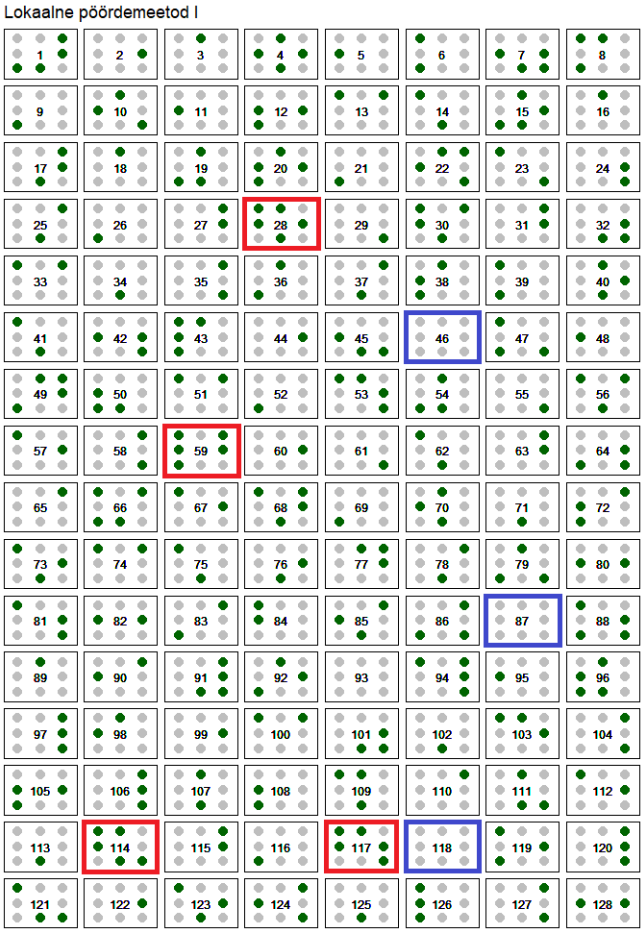
Selleks, et näidata lokaalse pöördemeetodi ruumilise tasakaalu eeliseid, koostati graafik, mis illustreerib küla *StatVillage* paiknemist. Sellel graafikul näitavad ringid maju ruudukujulises plokkides. Tumerohelise värviga on märgistatud majad, mis sattusid valimisse antud valikumeetodi korral. Nende plokkide, mis on eristatud värviga, majade sattumine valimisse on ekstreemne: sinise värviga on need plokid, milles ükski maja ei sattunud valimisse ja punasega need plokid, mille kõigi majade hulgast vähemalt 5 sattusid valimisse. Joonised on koostatud vabavara *R* kasutades. Kood on näitanud lisas 2. Lisas 3 ja 4 on koodi tulemused. Üks joonis vastab valimile, mis on võetud lihtsast juhuslikust valikust ja teine esimesest lokaalsest pöördemeetodist.



Joonis 7: Lihtsa juhusliku valiku valimi visuliseerimine küla *StatVillage* kaardil

Joonisel 7 näeme, et lihtsa juhusliku valiku korral ei pruugi valimisse sattunud objektid jaotuda ühtlaselt üle kogu üldkogumi, üheksast plokist ükski maja ei sattunud valimisse, mille hulgast

5 plokki moodustavad tühjate plokkide kogumeid: plokkid nr. 109, 110 ja 118 on kõrvuti asetsevad tühiplokid ja plokkid nr. 122 ja 123 samuti. Isegi plokid nr 57 ja 59 on üle ühe ploki naabrid. Samal ajal tühiplokk nr 95, tühjaplokkide kolmik ja kaksik asuvad vahetus läheduses. Joonisel 7 esitub 5 plokki, kus vähemalt 5 maja sattus valimisse. Nende hulgas on plokk nr. 22, mille kõigist kaheksast majast 7 maja sattusid valimisse. See asub lähedal plokkidega, kus valimisse sattusid 5 maja igast plokist.



Joonis 8: Lolaalsest pöördemeetodist saadud valimi illustreerimine küla *StatVillage* kaardil

Esimese lokaalse pöördemeetodi korral on ebaühtlus väiksem. Kolmest plokist ei sattunud ühtegi maja valimisse, kuid need plokid on teineteisest kaugel. Analoogiline situatsioon on plokkidega, kus vähemalt 5 maja on võetud valimisse, kokku selliseid plokke on 4 tükki, nad asuvad vähemalt üle 2 ploki. Selles valimis ei leidu plokke, kust oleks valitud rohkem kui 5 maja.

Võrreldes kahte valimit, asetsevad lokaalse pöördemeetodi korral valimisse sattunud objektid ruumis ühtlasemalt ehk ei asetse gruppide kaupa, vaid hajusalt.

Kokkuvõte

Käesoleva bakalaureusetöö eesmärk oli anda teoreetiline ülevaade lokaalsest pöördemeetodist ning võrrelda seda teiste tuntud valikumeetoditega, rakendades kõik valikumeetodeid reaalsetel andmetel. Vaatlusalusteks valikumeetoditeks olid lihtne juhuslik valik, lihtne juhuslik kihtvalik, süstemaatiline kihtvalik, juhuslik pöördemeetod, lokaalne pöördemeetod I ja lokaalne pöördemeetod II.

Teoreetilises osas tuletati esmalt meelde tuntud valikumeetodid: lihtne juhuslik valik, süstemaatiline valik, lihtne juhuslik kihtvalik ja süstemaatiline kihtvalik. Seejärel esitati ülevaade juhuslikust pöördevalikust, lokaalsest pöördevalikust I ja lokaalsest pöördevalikust II. Pöördemeetod ehk juhuslik pöördemeetod põhineb objektide kaasamistõenösuste pideval uuendamisel. Lokaalse pöördemeetodi eeliseks on valimi ruumiline tasakaal, mis on saadud lähimate objektide kaasamistõenäosuste uuendamise käigus. Eristatakse kaht erinevat lokaalset pöördemeetodit: lokaalne pöördemeetod I ja lokaalne pöördemeetod II. Esimese puhul valitakse uuendamiseks objektid, mis on lähimad naabrid teineteisele, teise puhul piisab, et vaid üks objekt oleks lähim naaber teisele. Uue valikumeetodi paremaks mõismiseks toodi väike näide valimi võtmise protsessist, kasutades lokaalset pöördemeetodit I.

Praktilises osas teostati Monte-Carlo simmulatsioon kuue erineva valikumeetodi korral. Kaasatud olid kõik eespool mainitud valikumeetodid. Simmuleerimisel kasutati andmeid, mis pärenevad Kanadas asuvast hüpoteetilisest külast *StatVillage*. Töö käigus selgus, et pideva tunnuse kogusumma hindamisel annavad mõlemad lokaalsed pöördemeetodid täpsemaid hinnanguid, kui teised valikumeetodid. Kusjuures lokaalse pöördemeetodi II hinnang on täpsem. Diskreetse tunnuse puhul ei andnud kumbki lokaaalne pöördemeetod paremaid hinnanguid. Jõuti järeldusele, et pideva tunnuse kogusumma hindamisel parema täpsuse saavutamiseks on soovitatav kasutada lokaalset pöördemeetodit II ning diskreetse tunnuse kogusumma hindamisel süstemaatilist kihtvalikut. Valimi visualiseerimise teel lihtsa juhusliku valiku ja lokaalse pöördemeetodi I võrdlemisel sai kinnitust lokaalse pöördmeetodi I eelis - ruumiline tasakaal.

Autor loodab, et töös esitatud lokaalne pöördemeetod I ja II ning näiteprogrammid tulevad kasuks tulevastele uurijatele.

Viited

Chauvet, G. (2012). On a charaterization of ordered pivotal sampling. *Bernoulli*, *18*(4), 1320–1340.

Deville, J.-C. ja Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, *85*(1), 89–101.

Grafström, A., Lundström, N. L. P. ja Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, *68*, 514–520.

Horvitz, D. G. ja Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663–685.

Mirski, S. (2017). *Tõenäosuslik valikuuring tarkvara r pakettide 'sampling' ja 'survey' abil* (Bakalaureusetöö). Tartu Ülikool.

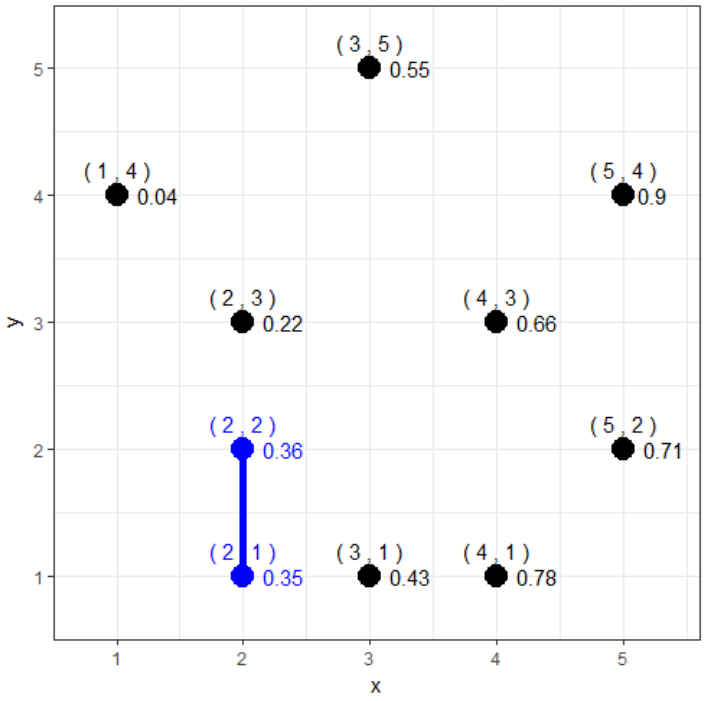
Schwarz, C. (1997). Statvillage: An on-line, www-accessible, hypothetical city based on real data for use in an introductory class in survey sampling. *Journal of Statistics Education*, *5*(2).

Sen, A. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, *5*.

Traat, I. ja Inno, J. (1997). *Tõenäosuslik valikuuring*. Tartu Ülikooli kirjastus.

Yates, F. ja Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, *15*(2), 253–261.

Lisa 1. Näiteandmestik



Lisa 2. *R*-kood *StatVillage*'le vastava kaardi joonistamiseks

Kood kasutab koordinaate x ja y ning valikuindikaatorit, mis on loodud järgmiselt. Koordinaadid x ja y kujutavad maja paiknemist plokis nii nagu näidatud tabelis 9.

Tabel 9: Näide küla *StatVillage* plokk nr.12 koos koordinaatidega

3	1	2	3
2	4	**12**	5
1	6	7	8
y/x	1	2	3

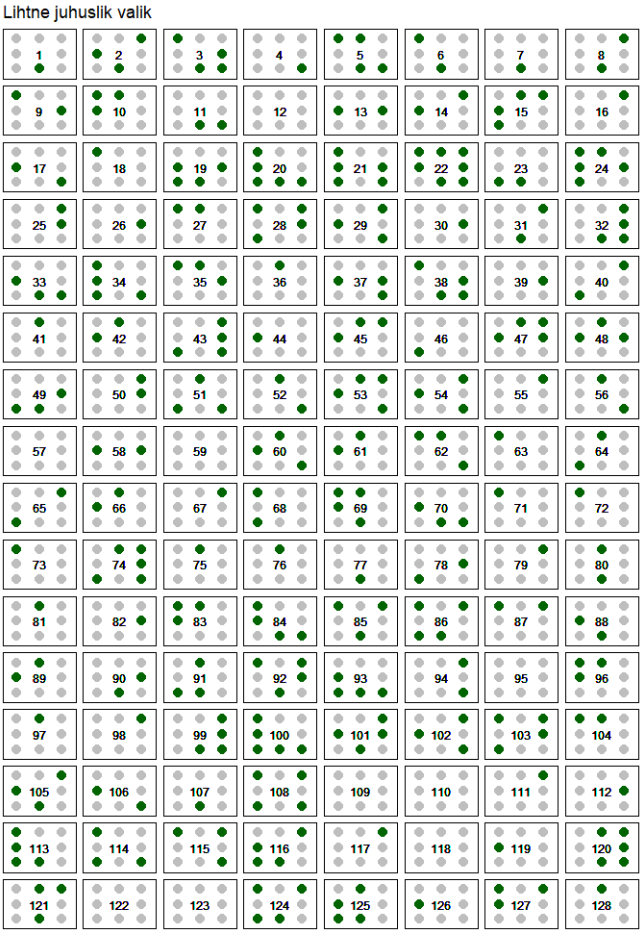
Valikuindikaator näitab, kas objekt (antud juhul maja) on valimis või mitte, tunnuse väärtused vastavalt 1 ja 0. Enne valiku sooritamist on kõikide majade valikuindikaator 0. Funktsioon võtab argumendiks andmestiku ja väljastab nõutava joonise. Andmestikus peavad olema tunnused: *plokk*, *x*, *y*, *valik_ind*.

```
statvilagge<-function(andmestik){
  ggplot(andmestik,aes(x=x,y=y,colour=valik_ind))+
  #võtab x ja y koordinaadid ja määrab vastavalt
  #valikuindikaatori väärtusele värvi
    geom_point(size=3)+ #joonistab punkti  suurusega 3
    facet_wrap(~plokk,ncol=8)+ #eraldab graafiku plokkide kaudu
    theme(strip.background = element_blank(),
          panel.background = element_rect(fill = "white",
          colour = "black"),
          axis.text.x = element_blank(),
          axis.text.y = element_blank(),
          axis.ticks = element_blank(),
          strip.text.x = element_blank(),
          legend.position="none",
          panel.border = element_rect(fill = NA,
            colour = "black",linetype = 1) )+
    labs(x="",y="")+
    #puhastab lisainformatsioonist
    scale_color_manual(breaks = c(0, 1),
                       values=c("grey", "darkgreen"))+
```

```
    #määrab valikuindikaatori väärtusele vastava värvi
    geom_text(data=freim,aes(2,2,label=plokk),
                size=3,color="black")+
    # ploki keskele kirjutab ploki numbri
    scale_x_continuous(limits = c(0.5, 3.5)) +
    scale_y_continuous(limits = c(0.5, 3.5))
    #määrab plokki suurus
}
```

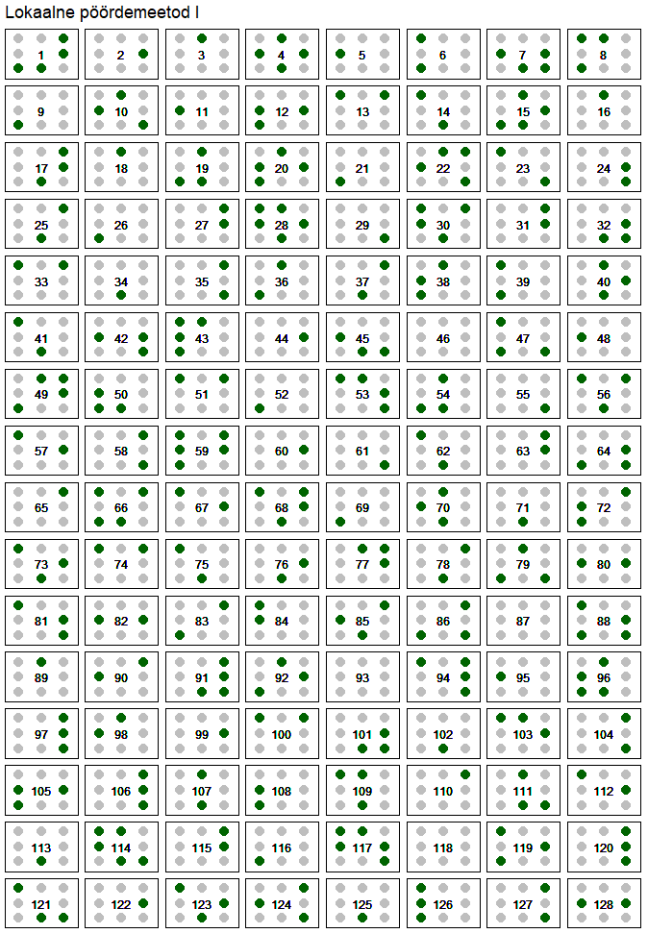
Lisa 3. Kood lihtsa juhusliku valiku teostamiseks ja valimisse sattunud objektide visualiseerimiseks *StatVillage* kaardil

```
library(sampling)
set.seed(2) #määrame seemne, et saada iga kord sama tulemust
s1=srswor(n=300, N=1024)
#lisame saadud väärtused andmestikku valikuindikaatori
#veeru asemele
freim_LJV<-freim %>% mutate(valik_ind=s1)
freim_LJV$valik_ind<-as.factor(freim_LJV$valik_ind)
#Rakendame funktsiooni statvillage
statvilagge(freim_LJV)+
  labs(title="Lihtne juhuslik valik")+ #lisame pealkirja
```



Lisa 4. Kood lokaalse pöördemeetodi I teostamiseks ja valimisse sattunud objektide visualiseerimiseks *StatVillage* kaardil

```
library(BalancedSampling)
N = 1024; n = 300; # üldkogumi ja valimi maht
p = rep(n/N,N); #kaasamistõenäosuse vektor
matriks<- freim %>% select(plokk,maja,nuirh) %>% as.matrix()
#abitunnused maatriksi kujul
set.seed(1201) #määrab seemne
s=lpm1(p,matriks)
freim_lpm1<-freim # kopeerib andmestiku
freim_lpm1$valik_ind[s]<-1
#vastavalt indeksile muudab valikuindikaatori väärtused
#Rakendame funktsiooni statvillage
statvilagge(freim_lpm1)+labs(title="Lokaalne pöördemeetod I")
```



Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Diana Sokurova,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose *Lokaalne pöördemeetod valikuuringutes*, mille juhendaja on Natalja Lepik,

   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, 08.05.2018