VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
MATEMATINĖS STATISTIKOS KATEDRA

Rasa Beržvinskaitė

RANGUOTŲ AIBIŲ ĖMIMO ĮVERTINIŲ TAIKYMAS ŽEMĖS ŪKIO SRITIES DUOMENIMS

Magistro baigiamasis darbas

Vilnius, 2017

Darbo vadovas:

doc. dr. Viktor Skorniakov

Recenzentas:

doc. dr. Rūta Levulienė

Registracijos Nr.: ....

Darbo gynimo data: 2017 - 01 - 19

Turinys

Sutrumpinimai

RSS (angl. *ranked set sampling*) — ranguotų aibių ėmimas;
SRS (angl. *simple random sampling*) — paprastasis atsitiktinis ėmimas;
SP — standartinė produkcija.

1. Ranguotų aibių ėmimas

1.1. Įvadas

Ranguotų aibių ėmimo (toliau RSS — nuo angl. ranked set sampling) metodą pirmasis pasiūlė G. A. McIntyre 1952 m. kaip efektyvų būdą įvertinti vidutinį pašarinių žolių derlių. RSS metodas yra alternatyva paprasto atsitiktinio ėmimo (toliau — SRS nuo angl. simple random sampling) metodui, kai tiriamus objektus galime nesunkiai išrikiuoti pagal dydį, naudodamiesi turima papildoma informacija, tiesiogiai neišmatuodami tiriamo kintamojo reikšmės. RSS tikslas yra iš populiacijos atrinkti elementus taip, kad jie geriau apimtų visą tiriamo kintamojo įgyjamų reikšmių aibę ir tokiu būdu tiksliau atspindėtų tiriamąją populiaciją nei to paties dydžio imtis, išrinkta SRS metodu ([1]).

Ranguotų aibių ėmimas daugiausia taikomas populiacijos vidurkiui, sumai, proporcijai vertinti ekologijos, biologijos, žemės ūkio ir aplinkosaugos, medicinos ir kitose srityse (esant begalinei ir baigtinei populiacijoms), kuriose patį tyrimo kintamąjį išmatuoti yra sudėtinga ir brangu, tačiau populiacijos elementų rangavimas panaudojant papildomą informaciją yra lengvai įgyvendinamas. Taip pat yra rašoma apie RSS panaudojimą konstruojant analogą Wilcoxon-Mann-Whitney testui homogeniškumo hipotezei tikrinti, kokybės kontrolės kortoms konstruoti. Glausta RSS metodo ir jo modifikacijų bei pritaikymo įvairiose srityse apžvalga pateikta A. I. Al-Omari ir C. N. Bouza straipsnyje [4].

Šiame darbe bandžiau pritaikyti RSS metodą žemės ūkio srities duomenims ir palyginti populiacijos parametrų įverčius su paprastosios atsitiktinės imties įverčiais.

Pirmajame skyriuje trumpai apžvelgiama su RSS metodu susijusi teorinė medžiaga: 1.2 skyrelyje aprašoma RSS imties išrinkimo procedūra begalinėms populiacijoms bei baigtinėms populiacijoms su grąžinimu, 1.3 skyrelyje — imties išrinkimo procedūros modifikacijos, kai populiacija yra baigtinė ir ėmimas yra be grąžinimo. 1.4 skyrelyje aptariami RSS metodo privalumai lyginant su SRS metodu; galiausiai 1.5 skyrelyje aprašomas populiacijos vidurkio ir sumos vertinimas subalansuoto imties plano atveju. Antrajame skyriuje apžvelgiami modeliavimui naudoti duomenys ir gauti rezultatai. Toliau pateikiamos išvados.

1.2. Bendra imties išrinkimo procedūra

Darbe [1] aprašoma standartinė RSS imties išrinkimo procedūra begalinėms populiacijoms arba baigtinėms populiacijoms su grąžinimu. Žemiau aptariami pagrindiniai žingsniai. Tarkime, RSS metodu norime iš populiacijos išrinkti k dydžio imtį. Į imtį patekusiems objektams matuosime kintamojo X reikšmes. Atliekame tokius žingsnius:

1. Išrenkame pradinę k dydžio SRS imtį, vadinamą aibe (angl. *set*). Pažymėkime X_{11}, $X_{12},\dots,X_{1k}$ atitinkamų aibės elementų kintamojo X reikšmes. Išrikiuojame elementus pagal dydį. Tam galime elementus tarpusavyje lyginti vizualiai, remtis ekspertine nuomone, ankstesnių tyrimų rezultatais arba su tyrimo kintamuoju gerai koreliuojančiu papildomu kintamuoju, tačiau negalime matuoti tikrųjų tiriamo kintamojo reikšmių $X_{1j}, j=\overline{1,k}$. Surikiuotų pagal dydį elementų kintamojo X reikšmes žymėkime $X_{1[1]},X_{1[2]},\dots,X_{1[k]}$. Į galutinę imtį atrenkamas elementas, turintis mažiausią rangą, t.

y. $X_{1[1]}$. Jam yra išmatuojama ši tiriamo kintamojo X reikšmė, o dydis $X_{1[1]}$ vadinamas pirmąja išankstinio vertinimo pozicine statistika [1] (angl. *first judgement order statistic*). Likusieji $k-1$ elementai mums nebesvarbūs, jiems kintamojo X reikšmė nematuojama.
2. Nepriklausomai nuo pirmosios imties, išrenkama antra k dydžio SRS imtis. Pažymėkime $X_{21}, X_{22}, \ldots, X_{2k}$ — atitinkamų aibės elementų kintamojo X reikšmes. Kaip ir prieš tai, naudodami pagalbinį kintamąjį ar kito tipo informaciją, išrikiuojame elementus pagal dydį. Surikiuotų pagal dydį elementų kintamojo X reikšmes žymėkime $X_{2[1]}, X_{2[2]}, \ldots, X_{2[k]}$. Į galutinę imtį atrenkamas elementas, turintis antrą mažiausią rangą, t. y. $X_{2[2]}$. Jam taip pat yra išmatuojama tiriamo kintamojo X reikšmė ir dydis $X_{2[2]}$ vadinamas antrąja išankstinio vertinimo pozicine statistika (angl. *second judgement order statistic*).
3. Šis procesas kartojamas k kartų, kol į galutinę imtį įtraukiame, mūsų nuomone, didžiausiąjį iš atrinktų k-os imties elementų; jo matavimą žymime $X_{k[k]}$.

Visų k aibių elementus galima surašyti į tokią matricą:

$$\begin{pmatrix} (X_{1[1]}) & X_{1[2]} & \ldots & X_{1[k]} \\ X_{2[1]} & (X_{2[2]}) & \ldots & X_{2[k]} \\ \ldots & & & \\ X_{k[1]} & X_{k[2]} & \ldots & (X_{k[k]}) \end{pmatrix} \tag{1}$$

Į galutinę imtį pakliūna įstrižainės elementai $X_{1[1]}, X_{2[2]}, ..., X_{k[k]}$ (pažymėti skliausteliais), kuriems yra pamatuojamos tikrosios kintamojo X reikšmės.

1-3 žingsniais aprašytas procesas vadinamas ciklu (angl. *cycle*), o dydis k – aibės dydžiu (angl. *set size*). Norint užbaigti vieną ciklą, iš viso reikia panaudoti k^2 populiacijos elementų, kad atskirai suranguotume k nepriklausomų paprastųjų atsitiktinių k dydžio imčių. Norėdami RSS metodu gauti $n = km$ dydžio imtį, atliekame m nepriklausomų ciklų.

Naudojant aukščiau aprašytą procedūrą, gaunama subalansuota ranguotų aibių imtis (angl. *balanced ranked set sample*). Čia terminas „subalansuota“ reiškia, kad į galutinę imtį atrinkome po vienodą skaičių elementų, atitinkančių kiekvieną iš rangų $1, \ldots, k$. Alternatyva subalansuotai imčiai yra nesubalansuota ranguotų aibių imtis (angl. *unbalanced ranked set sample*), kuomet vienų rangų kintamuosius į imtį įtraukiame dažniau nei kitų.

Tolimesniuose skyreliuose naudosime tokius pozicinių statistikų žymėjimus:

- $X_{[i]j}$ — i-oji išankstinio vertinimo pozicinė statistika j-ajame cikle, $i = \overline{1,k}$, $j = \overline{1,m}$;
- $X_{[i]}, i = \overline{1,n}$ — i-oji išankstinio vertinimo pozicinė statistika, kai atliekamas tik vienas ciklas;
- $X_{(i)}, i = \overline{1,n}$ — i-oji pozicinė satistika, išrinkus n dydžio imtį SRS metodu.

[1] laužtiniai skliaustai naudojami vietoje lenktinių skliaustų, nes matavimas $X_{1[1]}$ nebūtinai yra iš tikrųjų mažiausias tarp turimų k elementų, nors mes jį ir įvertinome kaip mažiausią

1.3. Imties išrinkimas iš baigtinės populiacijos

Įvairūs autoriai nagrinėja skirtingas imties išrinkimo iš baigtinės populiacijos technikas pagal tai, kurie elementai imties išrinkimo metu yra grąžinami į populiaciją.

Tarkime, turime N dydžio populiaciją, iš kurios norime išrinkti $n = km$ dydžio imtį. Atliekame m ciklų, kiekviename iš jų išrenkame k elementų.

Galima išskirti kelis pagrindinius imties išrinkimo planus ([2], [3]):

1. Į kiekvieną k dydžio aibę elementai įtraukiami taikant SRS be grąžinimo. Tačiau atlikus rangavimą visi šios aibės elementai, įskaitant pasirinktą elementą, kuriam bus atliekami tikslūs matavimai, grąžinami į populiaciją dar prieš kitos aibės išrinkimą. Tokiu būdu galutinėje $n = km$ dydžio imtyje galimi pasikartojantys elementai. Kai imties dydis yra didelis palyginti su populiacijos dydžiu, šie pasikartojimai gali lemti mažesnį RSS efektyvumą. Šis imties išrinkimo planas kartais vadinamas nulinio lygio (*level-0*) planu.
2. Į kiekvieną k dydžio aibę elementai įtraukiami taikant SRS be grąžinimo. Išrinkus pirmąjį elementą į galutinę imtį, likusieji $k-1$ elementai grąžinami į populiaciją. Antrąją k dydžio aibę renkame iš $N-1$ populiacijos elementų. Išrinkus antrąjį elementą į galutinę imtį, likusieji $k-1$ elementai vėl grąžinami į populiaciją. Trečiąją k dydžio aibę renkame jau iš $N-2$ populiacijos elementų ir t. t. Galutinėje imtyje visi elementai yra skirtingi. Šis imties išrinkimo planas vadinamas pirmojo lygio (*level-1*) planu. Planą galima taikyti, kai $k \leq \frac{N+1}{m+1}$.
3. Į populiaciją negrąžinami nė vienas iš k aibės elementų, nepriklausomai nuo to, ar elementas įtrauktas į galutinę imtį, ar ne. Galutinėje imtyje visi elementai yra skirtingi. Šis imties išrinkimo planas vadinamas antrojo lygio (*level-2*) planu. Planą galima taikyti, kai $mk^2 \leq N$.

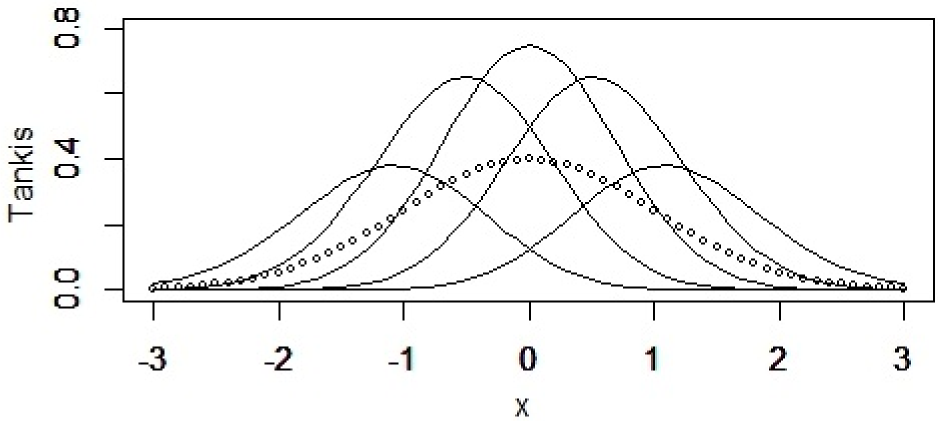
Visi trys planai turi panašias savybes, kai populiacija yra didelė, bet skiriasi, kai populiacija maža. Antrojo lygio planas turi stipresnę neigiamą koreliaciją tarp populiacijos elementų priklausymo imčiai indikatorių ir paprastai yra efektyvesnis nei nulinio ir pirmojo lygio planai.

1.4. RSS metodo pranašumai lyginant su SRS ėmimu

Lyginant SRS ir RSS metodus, svarbus yra pozicinių statistikų ir jų savybių vaidmuo. Toliau paminėsime keletą aspektų, aprašytų D. A. Wolfe ([1]) bei G. P. Patil et al. ([6]) straipsniuose.

Tarkime, turime dvi vienodo dydžio atsitiktines imtis, išrinktas SRS ir RSS metodais iš begalinės populiacijos. SRS imties atveju, visi stebiniai yra tarpusavyje nepriklausomi ir vienodai pasiskirstę, t. y. kiekvienas imties elementas atvaizduoja tipinę tiriamos populiacijos reikšmę. Tuo tarpu subalansuotos RSS imties atveju, stebiniai yra tarpusavyje nepriklausomi, tačiau nėra vienodai pasiskirstę ir jau nereprezentuoja tipinių tiriamos populiacijos reikšmių. Vis dėlto, kiekviena išankstinio vertinimo pozicinė statistika atspindi skirtingas tarpusavyje beveik nesikertančias tiriamos populiacijos dalis. Tokiu būdu RSS metodas geriau nei SRS užtikrina, kad būtų atspindėta visų populiacijos elementų galimų reikšmių aibė.

Iliustracijai panagrinėkime paprastąją atsitiktinę 5 elemetų imtį $X_1, \ldots, X_5$, kur $X_i \sim N(0,1), i = \overline{1,5}$. Tegul $X_{(1)}, \ldots, X_{(5)}$ žymi atitinkamas pozicines statistikas. Grafike pavaizduokime standartinio normalaus skirstinio tankį ir visų penkių pozicinių statistikų marginalius tankius. Jeigu naudosime to paties dydžio RSS imtį, tai, esant idealiam elementų surangavimui, kiekvienas imties stebinys atitiks tarpusavyje nepriklausomas standartinio normalaus skirstinio pozicines statistikas, kurių tankiai grafike pavaizduoti ištisine linija (1 pav.).



1 pav. Standartinio normalaus atsitiktinio dydžio tankis (taškuota linija) ir penkių pozicinių statistikų tankiai (ištisinės linijos)

Tarkime, turime n dydžio paprastąją atsitiktinę imtį $X_1, \ldots, X_n$, išrinktą iš begalinės populiacijos, $X_{(1)}, \ldots, X_{(n)}$ — atitinkamos pozicinės statistikos. Populiacijos vidurkio μ įvertinys yra

$$\hat{\mu}_{SRS} = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{2}$$

Įvertinį galima užrašyti ir naudojant pozicines statistikas:

$$\hat{\mu}_{SRS} = \frac{1}{n} \sum_{i=1}^{n} X_{(i)}. \tag{3}$$

Taip pat iš tos pačios begalinės populiacijos išrinkime vieno ciklo RSS n dydžio imtį $X_{[1]}, \ldots, X_{[n]}$, kur $X_{[i]}$ yra i-oji išankstinio vertinimo pozicinė statistika i-ojoje aibėje. Nors $X_{[i]}$ turi tą patį marginalų skirstinį kaip ir $X_{(i)}$, vis dėlto $X_{[i]}$ yra tarpusavyje nepriklausomi, tuo tarpu $X_{(i)}$ — teigiamai koreliuoti atsitiktiniai dydžiai. RSS metodu gautas populiacijos vidurkio μ įvertinys yra

$$\hat{\mu}_{RSS} = \frac{1}{n} \sum_{i=1}^{n} X_{[i]}. \tag{4}$$

Lygindami (3) ir (4) įvertinius galime pastebėti, kad $\hat{\mu}_{SRS}$ turi didesnę dispersiją dėl teigiamos koreliacijos tarp dydžių $X_{(i)}$ ir tuo nusileidžia įvertiniui $\hat{\mu}_{RSS}$.

Kai populiacija yra baigtinė, atsitiktiniai dydžiai X_i (2) lygtyje jau nebėra tarpusavyje nepriklausomi, bet turi neigiamą koreliaciją, kuri sumažina įvertinio dispersiją. Panašiai ir RSS atveju: dydžiai $X_{[i]}$ (4) lygtyje yra neigiamai koreliuoti ir ši koreliacija taip pat sumažina įvertinio dispersiją lyginant su įvertiniu, kai populiacija begalinė.

1.5. Populiacijos vidurkio ir sumos baigtinėse populiacijose vertinimas, esant subalansuotam imties planui

Tarkime, turime baigtinę N dydžio populiaciją; X yra tyrimo kintamasis, kurio reikšmės apibrėžtos visiems populiacijos elementams. Mus dominantys populiacijos parametrai – vidurkis $\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$ ir suma $t = \sum_{i=1}^{N} X_i = N\mu$.

Žinome, kad SRS imties populiacijos parametrų μ ir t nepaslinkti įvertiniai yra:

$$\hat{\mu}_{SRS} = \frac{1}{n}\sum_{i=1}^{n} X_i; \tag{5}$$

$$\hat{t}_{SRS} = \frac{N}{n}\sum_{i=1}^{n} X_i. \tag{6}$$

Šių įvertinių dispersijos turi pavidalus:

$$D\hat{\mu}_{SRS} = \left(1 - \frac{n}{N}\right)\frac{S^2}{n}; \tag{7}$$

$$D\hat{t}_{SRS} = N^2\left(1 - \frac{n}{N}\right)\frac{S^2}{n}. \tag{8}$$

Čia $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \mu)^2$.

Tegul RSS metodu renkama $n = km$ dydžio imtis ir priklausomai nuo pasirinkto imties išrinkimo metodo (*level-0* – *level-2*) kai kurie elementai į populiaciją nėra grąžinami. Naudosime 1.2 skyrelyje aprašytus žymėjimus.

Populiacijos vidurkio μ ir sumos t įvertiniai esant subalansuotam imties planui yra

$$\hat{\mu}_{RSS} = \sum_{j=1}^{m}\sum_{i=1}^{k}\frac{X_{[i]j}}{km}; \tag{9}$$

$$\hat{t}_{RSS} = N\sum_{j=1}^{m}\sum_{i=1}^{k}\frac{X_{[i]j}}{km} = N\hat{\mu}_{RSS}. \tag{10}$$

Tarkime, rangavimas idealus ir taikomas *level-0* imties planas. Įrodysime keletą teiginių apie įvertinių (9) ir (10) nepaslinktumą ir dispersijų pavidalą (remiantis [2], [1]).

1 teiginys. Įvertiniai (9) ir (10) yra nepaslinkti populiacijos vidurkio ir sumos įvertiniai.

Įrodymas. Iš pradžių nagrinėkime atvejį, kai atliekamas tik vienas ciklas, t. y. $m = 1$ ir $n = k$. Kadangi išankstinio vertinimo pozicinės statistikos $X_{[i]}$ yra pasiskirsčiusios kaip ir to paties dydžio SRS imties pozicinės statistikos $X_{(i)}$, tai

$$E\hat{\mu}_{RSS} = E\left(\frac{1}{k}\sum_{i=1}^{k} X_{[i]}\right) = \frac{1}{k}\sum_{i=1}^{k} EX_{[i]} = \frac{1}{k}\sum_{i=1}^{k} EX_{(i)} = E\left(\frac{1}{k}\sum_{i=1}^{k} X_{(i)}\right) = E\hat{\mu}_{SRS} = \mu. \tag{11}$$

Kai atliekame daugiau negu vieną ciklą, remiantis (11) ir ciklų nepriklausomumu, gauname:

$$E\hat{\mu}_{RSS} = E\left(\frac{1}{km}\sum_{j=1}^{m}\sum_{i=1}^{k} X_{[i]j}\right) = \frac{1}{m}\sum_{j=1}^{m} E\left(\frac{1}{k}\sum_{i=1}^{k} X_{[i]j}\right) = \frac{1}{m}\sum_{j=1}^{m}\mu = \mu. \tag{12}$$

Abiem atvejais

$$E\hat{t}_{RSS} = NE\hat{\mu}_{RSS} = N\mu = t. \tag{13}$$

□

2 teiginys. Kai $m = 1$, tai populiacijos vidurkio μ ir sumos t įvertinių (9) ir (10) dispersijos apibrėžiamos toliau užrašytomis formulėmis.

$$D(\hat{\mu}_{RSS}) = \frac{1}{k}\left(\left(1 - \frac{1}{N}\right)S^2 - \frac{1}{k}\sum_{i=1}^{k}(EX_{(i)} - \mu)^2\right); \tag{14}$$

$$D(\hat{t}_{RSS}) = N^2 D(\hat{\mu}_{RSS}) = \frac{N^2}{k}\left(\left(1 - \frac{1}{N}\right)S^2 - \frac{1}{k}\sum_{i=1}^{k}(EX_{(i)} - \mu)^2\right). \tag{15}$$

Įrodymas. Pasinaudosime gerai žinoma lygybe:

$$(N-1)S^2 = \sum_{i=1}^{N}(X_i - \mu)^2 = \sum_{i=1}^{N} X_i^2 - N\mu^2. \tag{16}$$

Tuomet turėsime:

$$\begin{aligned}
D(\hat{\mu}_{RSS}) = D\left(\sum_{i=1}^{k}\frac{X_{[i]}}{k}\right) = \frac{1}{k^2}\sum_{i=1}^{k} D(X_{[i]}) = \\
= \frac{1}{k^2}\sum_{i=1}^{k}\left(EX_{[i]}^2 - (EX_{[i]})^2\right) = \frac{1}{k^2}\left(\sum_{i=1}^{k} EX_{[i]}^2 - \sum_{i=1}^{k}(EX_{[i]})^2\right) = \\
= \frac{1}{k^2}\left(E\left(\sum_{i=1}^{k} X_{(i)}^2\right) - \sum_{i=1}^{k}(EX_{(i)})^2\right) = \frac{1}{k^2}\left(E\left(\sum_{i=1}^{k} X_i^2\right) - \sum_{i=1}^{k}(EX_{(i)})^2\right) = \\
= \frac{1}{k^2}\left(\sum_{i=1}^{k}\frac{1}{N}\sum_{l=1}^{N} X_l^2 - \sum_{i=1}^{k}(EX_{(i)})^2\right) = \frac{1}{k}\left(\frac{1}{N}\sum_{l=1}^{N} X_l^2 - \frac{1}{k}\sum_{i=1}^{k}(EX_{(i)})^2\right) = \\
= \frac{1}{k}\left(\frac{1}{N}(N-1)S^2 + \mu^2 - \frac{1}{k}\sum_{i=1}^{k}(EX_{(i)})^2\right) = \frac{1}{k}\left(\left(1 - \frac{1}{N}\right)S^2 - \frac{1}{k}\sum_{i=1}^{k}(EX_{(i)} - \mu)^2\right)
\end{aligned} \tag{17}$$

(15) lygybė gaunama pasinaudojus dispersijos savybėmis.

□

3 teiginys. Jei $m > 1$, tai populiacijos vidurkio μ ir sumos t įvertinių (9) ir (10) dispersijos yra atitinkamai

$$D(\hat{\mu}_{RSS}) = \frac{1}{mk}\left(\left(1-\frac{1}{N}\right)S^2 - \frac{1}{k}\sum_{i=1}^{k}(EX_{(i)}-\mu)^2\right); \tag{18}$$

$$D(\hat{t}_{RSS)}) = N^2 D(\hat{\mu}_{RSS}) = \frac{N^2}{mk}\left(\left(1-\frac{1}{N}\right)S^2 - \frac{1}{k}\sum_{i=1}^{k}(EX_{(i)}-\mu)^2\right). \tag{19}$$

Įrodymas. Pasinaudodami 2 teiginio rezultatais ir tuo, kad visi atrinkti į imtį elementai yra grąžinami į populiaciją, gauname:

$$\begin{aligned} D(\hat{\mu}_{RSS}) &= D\left(\sum_{j=1}^{m}\sum_{i=1}^{k}\frac{X_{[i]j}}{km}\right) = \frac{1}{m^2}\sum_{j=1}^{m} D\left(\sum_{i=1}^{k}\frac{X_{[i]j}}{k}\right) = \\ &= \frac{1}{m^2}\sum_{j=1}^{m}\frac{1}{k}\left(\left(1-\frac{1}{N}\right)S^2 - \frac{1}{k}\sum_{i=1}^{k}(EX_{(i)}-\mu)^2\right) = \\ &= \frac{1}{mk}\left(\left(1-\frac{1}{N}\right)S^2 - \frac{1}{k}\sum_{i=1}^{k}(EX_{(i)}-\mu)^2\right) \end{aligned}$$

(19) vėl gaunama pasinaudojus dispersijos savybėmis.

□

1 pastaba. T. R. Dell ir J. R. Clutter ([7]) parodė, kad įvertinys (9) yra nepaslinktas nepriklausomai nuo to, ar rangavimas idealus, ar ne.

2 pastaba. Kai taikomas imties planas *level-1* arba *level-2*, tai ciklai tampa priklausomi vienas nuo kito ir dispersijų išraiškos įgyja sudėtingesnį pavidalą, todėl jų nenagrinėsime (plačiau dispersijos išraišką baigtinių populiacijų atveju nagrinėjo G. P. Patil et al [6] straipsnyje). Tačiau jau *level-0* plano atveju galime pastebėti, kad populiacijos vidurkio įvertinio dispersija neviršija SRS įvertinio dispersijos:

$$\begin{aligned} D(\hat{\mu}_{RSS}) &= \frac{1}{mk}\left(\left(1-\frac{1}{N}\right)S^2 - \frac{1}{k}\sum_{i=1}^{k}(EX_{(i)}-\mu)^2\right) = \\ &= \frac{1}{n}\left(1-\frac{1}{N}\right)S^2 - \frac{1}{mk^2}\sum_{i=1}^{k}(EX_{(i)}-\mu)^2 = \\ &= D(\hat{\mu}_{SRS}) - \frac{1}{mk^2}\sum_{i=1}^{k}(EX_{(i)}-\mu)^2 \le D(\hat{\mu}_{SRS}) \end{aligned} \tag{20}$$

Lygybė galima tik tada, kai $EX_{(i)} = \mu$, t. y. kai rangavimas yra visiškai atsitiktinis.

2. RSS metodo taikymas žemės ūkio srities duomenims

2.1. Duomenys

Naudojau 2015 m. Lietuvos statistikos departamento atlikto Žemės ūkio augalų pasėlių, derliaus ir derlingumo statistinio tyrimo duomenis. Kadangi duomenys konfidencialūs, jie nėra pridedami prie šio darbo.

Statistikos departamento vykdomo tyrimo tikslas — įvertinti žemės ūkio augalų pasėlių plotus, gautą derlių ir derlingumą ataskaitiniais metais. Rodikliai skaičiuojami šalies, apskričių bei savivaldybių lygmeniu. Tyrime naudojama paprastoji atsitiktinė sluoksninė imtis. Populiacija į sluoksnius skaidoma pagal ūkio standartinę produkciją [2]. Imties dydis į sluoksnius paskirstomas pagal Neimano optimalųjį paskirstymo principą. Toliau gauti dydžiai skirstomi savivaldybėmis pagal proporcingąjį imties dydžio paskirstymą į sluoksnius ([5]).

Simuliacijoms naudojau tyrimo metu gautus pirminius duomenis, surinktus iš 6476 ūkininkų ir šeimos ūkių. Juos laikiau baigtine populiacija.

Tyrimo kintamieji — pasėtas ir apsodintas plotas (D1), nuimtas plotas (D2), nuimto derliaus svoris tonomis prieš valymą ir džiovinimą (D3), nuimto derliaus svoris tonomis po valymo ir džiovinimo (D4). Šių kintamųjų reikšmės skaičiuojamos svarbiausiems augalams ir jų grupėms [3], pateikiamoms 1 lentelėje .

Tikrosios populiacijos parametrų (sumų) reikšmės pateikiamos 2 lentelėje.

Panašius į mano turimus duomenis ir RSS metodo bei jo modifikacijų pranašumus prieš SRS metodą nagrinėjo J. C. Stroka ([8]) bei C. E. Husbey et al ([9]). Abiejuose darbuose naudojami tie patys duomenys apie Ohajo valstijoje auginamus kukurūzus (bendras ūkio plotas, kukurūzais apsodintas plotas, nuimtas plotas, nuimtas kukurūzų derlius, nuimto derliaus natūrinis logaritmas bei derlingumas), gauti iš 6346 ūkių. Tyrimo kintamuoju ir pagalbiniu rangavimo kintamuoju buvo renkamos įvairios nurodytų kintamųjų poros, tokiu būdu analizuojant RSS įverčio standartinio nuokrypio sumažėjimą, lyginant su SRS įverčio paklaida, esant įvairaus stiprumo koreliacijai tarp tyrimo ir pagalbinio kintamojo. Abiejuose darbuose pasiekti teigiami rezultatai, parodę, kad net ir esant labai silpnai koreliacijai (apytiksliai 0,1) RSS metodas, nors ir nežymiai, bet pagerina įverčio tikslumą.

Statistikos departamento atliekamo tyrimo kintamieji yra analogiški Ohajo tyrimo kintamiesiems, tačiau yra keletas pagrindinių skirtumų.

Statistikos departamentas tiria įvairias augalų kultūras auginančius ūkius, be to vienos

[2] Standartinė produkcija (toliau — SP) — ekonominiam ūkio dydžiui nustatyti naudojamas rodiklis. Jis rodo kiekvieno žemės ūkio statistinio rodiklio produkcijos vertę atitinkamame regione vidutinėmis sąlygomis. Vertė apskaičiuojama dauginant produkcijos vienetą iš supirkimo kainos. Pridėtinės vertės mokestis, produktų mokesčiai ir tiesioginės išmokos neįskaičiuojami.

[3] Ne kiekvienam augalui ar augalų grupei skaičiuojamos visų kintamųjų D1 − D4 reikšmės: D4 skaičiuojamas tik augalams 140 − 165, o augalams 101 ir 200 skaičiuojamos tik kintamojo D1 reikšmės. 1011 ir 2001 augalų grupės turi tik kintamuosius D2 ir D4

Kodas	Pavadinimas
101	Kultūrinės ganyklos (5 metų ir senesnės, iš viso)
1011	Kultūrinės ganyklos (5 metų ir senesnės, šienui)
140	Vasariniai kviečiai
164	Žieminiai rapsai
165	Vasariniai rapsai
177	Bulvės
181	Lauko daržovės
185	Cukriniai runkeliai
190	Kukurūzai silosui, žaliajam pašarui
200	Daugiametės žolės iki 5 metų (iš viso)
2001	Daugiametės žolės iki 5 metų (šienui)
211	Prekinių šiltnamių daržovės
230	Sėklavaisiai
231	Kaulavaisiai
238	Uogos

1 lentelė. Tyrime naudojami augalai ir jų grupės

Kodas	D1	D2	D3	D4
101	63506,17	-	-	-
140	253385,38	253256,57	1537534,5	1478450,94
164	59496,46	59320,63	215512,85	203858,67
165	20376,93	20212,25	41875,89	39384,75
177	4246,43	4225,39	94561,98	-
181	2691,5	2662,84	56344,15	-
185	8389,77	8362,28	416080,33	-
190	6305,89	6290,97	153765,99	-
200	45545,37	-	-	-
211	53,26	39,36	5174,34	-
230	1947,85	1498,39	10006,23	-
231	131,75	60,63	38,38	-
238	6311,54	4576,1	4630,53	-
1011	-	31749,9	75575,46	-
2001	-	23643,2	63995,23	-

2 lentelė. Tikrosios populiacijos sumų reikšmės

augalų rūšys auginamos dažniau nei kitos. Analizuojant kiekvieną augalą ar jų grupę atskirai, galima parinkti pagalbinį kitamąjį (standartinę produkciją arba bendrą ūkio plotą), kuris pakankamai stipriai koreliuotų su tyrimo kintamuoju, ir todėl tikėtis, kad bus pasiekti panašūs rezultatai kaip minėtuose darbuose. Tačiau iš tikrųjų mano turimoje populiacijoje kiekvienas ūkis augina tik kelias augalų kultūras ir duomenyse yra daug nulinių reikšmių, dėl kurių koreliacijos koeficientai sumažėja ir kai kurie tampa neigiami. Todėl mane domino, ar esamų sąlygų pakanka, kad galima būtų pagerinti bent dalies įverčių tikslumą, rangavimo kintamuoju imant SP, kuri, kaip minėta, šiuo metu taikoma populiacijos skirstymui į sluoksnius.

3 lentelėje pateikiamas ūkių, auginančių atitinkamus augalus ar jų grupes, skaičius.

Kodas	101	140	164	165	177	181	185	190	200	211	230
Ūkių skaičius	3779	4493	1393	852	2760	1405	229	340	3312	96	1029
Kodas	231	238	1011	2001							
Ūkių skaičius	277	611	2456	2140							

3 lentelė. Ūkių, auginančių tam tikrus augalus ir jų grupes, skaičius

Šiame darbe naudotiems duomenims būdinga dešininė asimetrija ir jie nėra pasiskirstę pagal normalųjį dėsnį, todėl taikiau Spirmeno koreliacijos koeficiento įvertį. Pilnas koreliacijų matricas galima rasti prieduose esančiose 1 – 15 lentelėse. Taip pat, naudodama reikšmingumo lygmenį $\alpha = 0,05$, visoms kintamųjų poroms tikrinau hipotezę :

$$\begin{aligned} H_0 &: \text{ koreliacijos koeficientas lygus 0;} \\ H_1 &: \text{ koreliacijos koeficientas nelygus 0.} \end{aligned} \tag{21}$$

Rodiklius sugrupavau į kelias grupes pagal koreliacijos tarp kintamųjų D1–D4 ir SP stiprumą. 4 ir 5 lentelėse pateiktos šios grupės, kai koreliacija skaičiuojama tik iš atitinkamą kultūrą auginančių ūkių duomenų bei iš pilno duomenų rinkinio.

Į grupę „Ryšio nėra“ patenka rodikliai, kuriems hipotezė (21) nėra atmetama.

Grupės nr.	Koreliacijos stiprumas	Rodiklių kodai
1	Stipri teigiama koreliacija (0,7–1)	140, 164, 185
2	Vidutinė teigiama koreliacija (0,5–0,7)	165
3	Silpna teigiama koreliacija (0,2–0,5)	177, 181, 190, 211
4	Labai silpna teigiama koreliacija (0–0,2)	101, 200, 230, 238,1011, 2001
5	Ryšio nėra	231

4 lentelė. Rodiklių grupės, naudojant dalį duomenų

Grupės nr.	Koreliacijos stiprumas	Rodiklių kodai
1	Stipri teigiama koreliacija (0,7–1)	140
2	Vidutinė teigiama koreliacija (0,5–0,7)	164
3	Silpna teigiama koreliacija (0,2–0,5)	165, 185
4	Labai silpna teigiama koreliacija (0–0,2)	190
5	Ryšio nėra	200, 211, 230 (D2, D3), 231 (D2, D3)
6	Labai silpna neigiama koreliacija (0– -0,2)	101, 177, 181, 230 (D1), 231 (D1), 238,1011, 2001

5 lentelė. Rodiklių grupės, naudojant pilną duomenų rinkinį

2.2. RSS ir SRS metodų palyginimas taikant juos turimiems duomenims

Iš turimos populiacijos SRS ir RSS metodais rinkau vienodo dydžio ($n = 100$) atsitiktines imtis. Naudodama RSS metodą taikiau *level-1* ir *level-2* subalansuotus imčių planus ir parametrų reikšmių rinkinius, nurodytus 6 lentelėje.

Ciklų skaičius m	Aibės dydis k
1	100
2	50
4	25
5	20
10	10

6 lentelė. Naudotų RSS metodo parametrų reikšmės

Kaip rangavimo kintamąjį naudojau standartinę produkciją. Jei keli ūkiai turėjo vienodą SP, tai didesniu laikiau tą, kuris turi didesnį bendrą žemės ūkio naudmenų plotą.

Kiekvienai simuliacijai įvertinau kintamųjų D1 – D4 populiacijos sumas kiekvienam augalui ar augalų grupei (iš viso 42 populiacijos parametrų reikšmes) naudodama formules (10) bei (6) ir atlikau po 25000 iteracijų. Gautus rezultatus vertinau lygindama tokias įverčių charakteristikas:

- populiacijos sumų įverčių vidurkius: $\bar{\hat{X}} = \frac{1}{25000}\sum_{i=1}^{25000}\hat{X}_i$;
- šaknį iš vidutinės kvadratinės paklaidos: $RMSE = \sqrt{\frac{1}{25000}\sum_{i=1}^{25000}(\hat{X}_i - X)^2}$
- RSS ir SRS metodais gautų RMSE santykį: $RMSE_{sant.} = \frac{RMSE_{RSS}}{RMSE_{SRS}}100\%$
- vidutinę absoliutinę įverčio paklaidą: $\bar{e} = \frac{1}{25000}\sum_{i=1}^{25000}|\hat{X}_i - X|$
- absoliutinę paklaidą: $e = |\bar{\hat{X}} - X|$

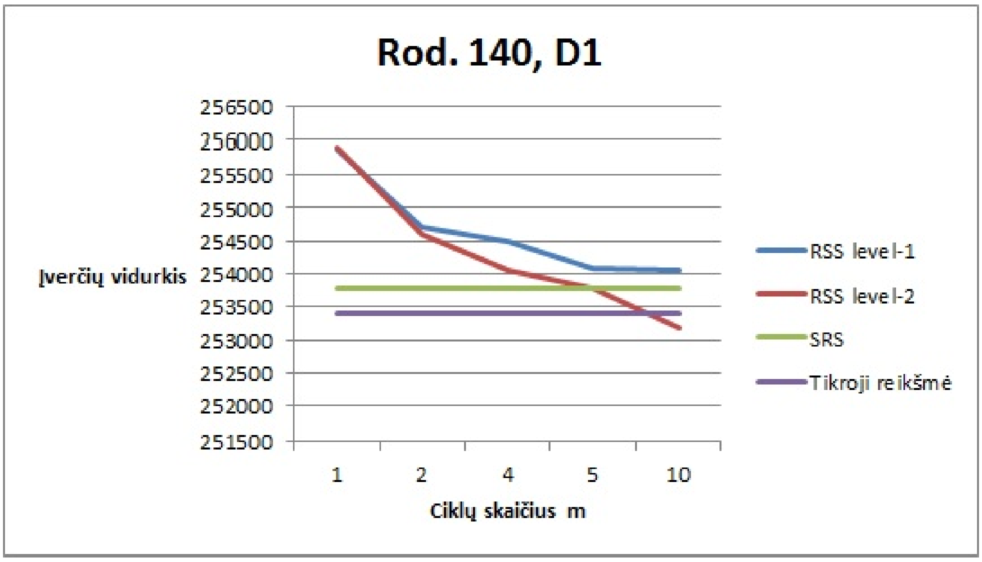
Čia X – tikroji populiacijos parametro reikšmė, o X_i — jos įvertis, gautas i-oje iteracijoje.

$RMSE$, vidutinę absoliutinę paklaidą $\bar{e}$ ir paklaidą e apskaičiavau atitinkamais kintamųjų D1 – D4 matavimo vienetais ir procentine išraiška (kokią tikrųjų populiacijos sumų dalį sudaro paklaida). Gautus rezultatus visiems kintamiesiems D1 – D4 visoms augalų grupėms galima rasti prie darbo pridedamoje Excel laikmenoje rezultatu_palyginimas.xlsx. Analizei naudoti grafikai ir paklaidos, apibendrintos grupėms pagal koreliacijos stiprumą, pateikti laikmenoje grafikai.xlsx.

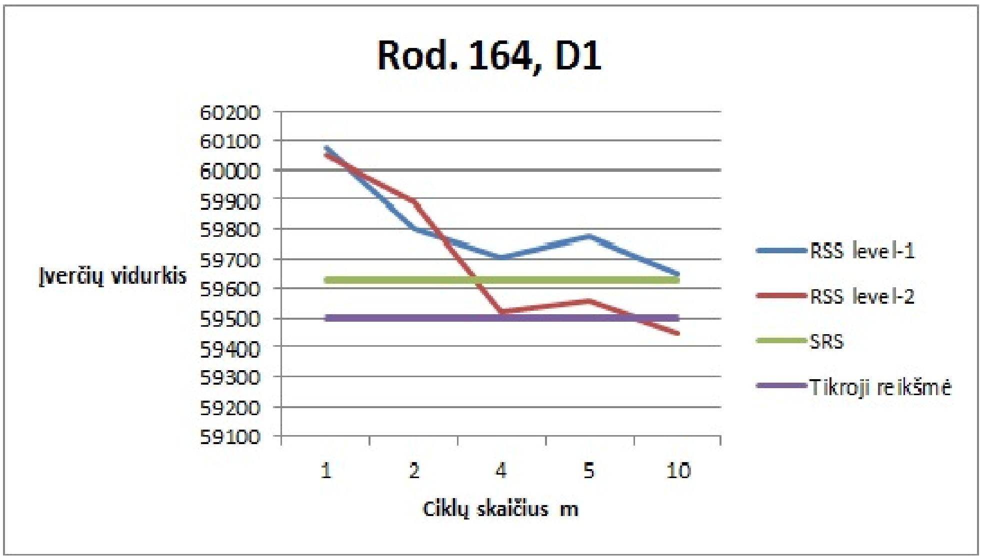
Imties dydžio pasirinkimui jokių teorinių skaičiavimų netaikiau. Tiek bendras imties dydis, skirtas vertinti populiacijos parametro sumą ($n = 100$), tiek simuliacijų skaičius, naudotas Monte–Karlo įvertinių konstravimui (25000 iteracijų kiekvienam parametrų rinkiniui) atrodė „iš akies“ priimtini, kad būtų įmanoma daryti pakankamai tikslias preliminarias išvadas apie lyginamus metodus nagrinėjamoje situacijoje.

Kadangi buvo nagrinėta daug augalų grupių ir kintamųjų, o rezultatai daug kur panašūs, toliau apžvelgsiu kai kuriuos gautus rezultatus tik pavieniams kintamiesiems ar augalų grupėms. Taip pat nagrinėsiu paklaidas, suvidurkintas visiems augalams ir visiems kintamiesiems D1 – D4, patenkantiems į tą pačią grupę pagal koreliacijos stiprumą (pagal 5 lentelę).

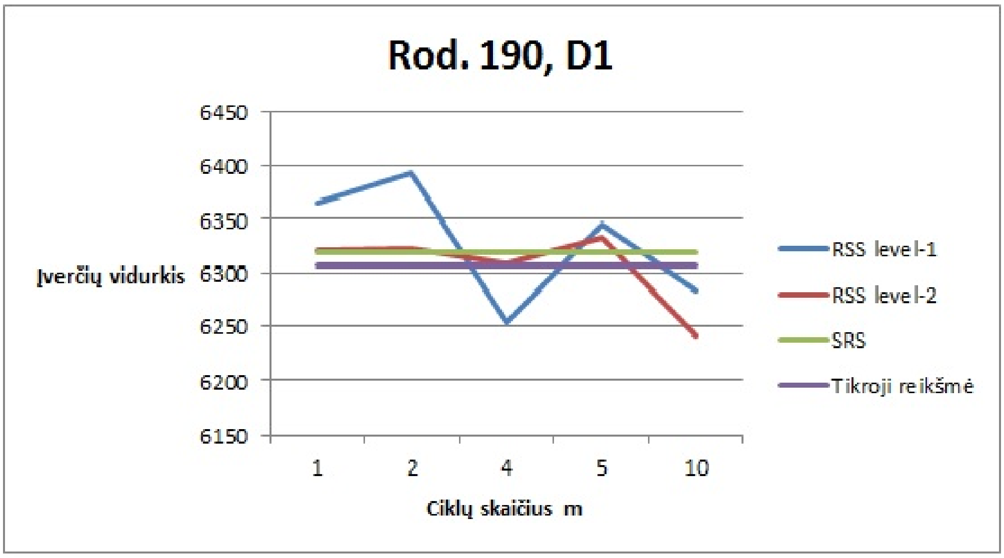
Pirmiausia lyginau populiacijos sumų įverčių vidurkius $\bar{\hat{X}}$, gautus taikant skirtingus imčių planus. Visais atvejais gautas vidurkis yra panašus į tikrąją populiacijos sumą, be to, ir SRS, ir RSS metodų atveju daliai kintamųjų gautas įvertis buvo mažesnis už tikrąją sumą. Nematyti jokio aiškaus dėsningumo, kad kuriuo nors metodu tendencingai būtų gaunami didesni ar mažesni už tikrąsias reikšmes įverčiai. Vis dėlto, esant stipresnei teigiamai koreliacijai tarp tyrimo kintamojo ir SP galima pastebėti, kad abiejų lygių RSS metodais gautas įvertis yra didesnis, kai naudojamas vienas ciklas, ir mažėja, kai ciklų skaičius didinamas. Kai koreliacija yra silpna arba jokio ryšio tarp tyrimo ir pagalbinio rangavimo kintamojo nėra, ciklų skaičius neturi aiškios įtakos įverčio didumui. Šį dėsningumą iliustruoja 2 - 5 grafikai.



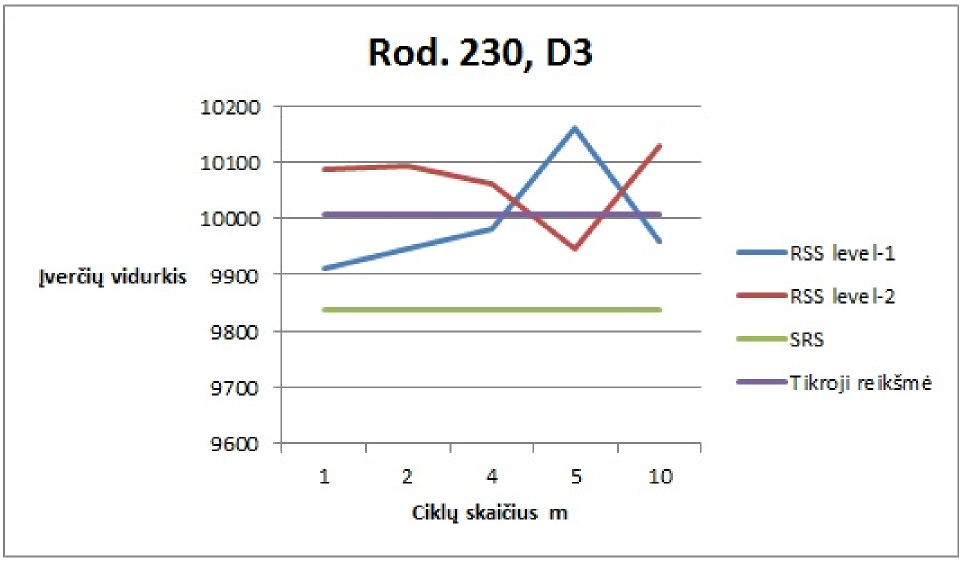
2 pav. Iverčių vidurkis, stipri teigiama koreliacija



3 pav. Iverčių vidurkis, vidutinė teigiama koreliacija



4 pav. Iverčių vidurkis, labai silpna teigiama koreliacija



5 pav. Iverčių vidurkis, ryšio nėra

Toliau nagrinėjau paklaidą *e*. 7 lentelėje parodyta, keliems kiekvienu metodu gautiems įverčiams paklaida absoliutiniu didumu yra iki 1 proc., 1–2 proc., 2–3 proc. ir viršija 4 proc.

	SRS	RSS level-1					RSS level-2				
Ciklų skaičius m:		1	2	4	5	10	1	2	4	5	10
Paklaida, proc.											
< 1	33	26	31	39	33	36	31	36	35	38	33
1–2	7	12	8	3	8	4	8	3	7	1	9
2–3	0	4	0	0	1	2	2	2	0	0	0
3–4	2	0	3	0	0	0	0	1	0	0	0
>4	0	0	0	0	0	0	1	0	0	3	0

7 lentelė. Įverčių, kurie turi atitinkamą paklaidą, skaičius

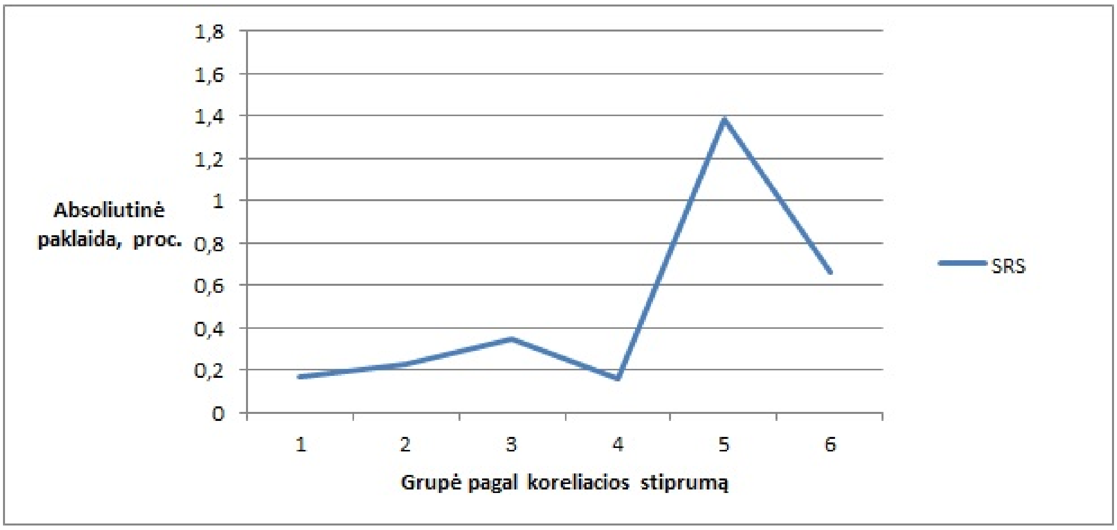
Matome, kad nė vienas metodas neišsiskiria iš kitų pagal savo paklaidos dydį. Didžiausią paklaidą (>3 proc.) turi augalų grupių 211 ir 231 įverčiai. Didžiausia gauta paklaida — 4,26 proc. (RSS *level-2* metodas, $m = 5, k = 20$, rod. 231, kintamasis D1).

Taip pat palyginau, kiek įverčių, gautų RSS metodu, turėjo mažesnę absoliutinę paklaidą nei SRS metodu gauti įverčiai (8 lentelė). Naudojant parametrų reikšmes ($m = 1, k = 100$) ir ($m = 2, k = 50$), abiejų lygių imties plano atveju RSS rezultatai paklaidos e atžvilgiu yra prastesni nei SRS imties plano. Kai ciklų skaičius $m > 2$, RSS ir SRS rezultatai apylygiai. Iš viso 124 atvejais iš 210 RSS *level-2* įverčiai turi mažesnę paklaidą nei RSS *level-1* įverčiai.

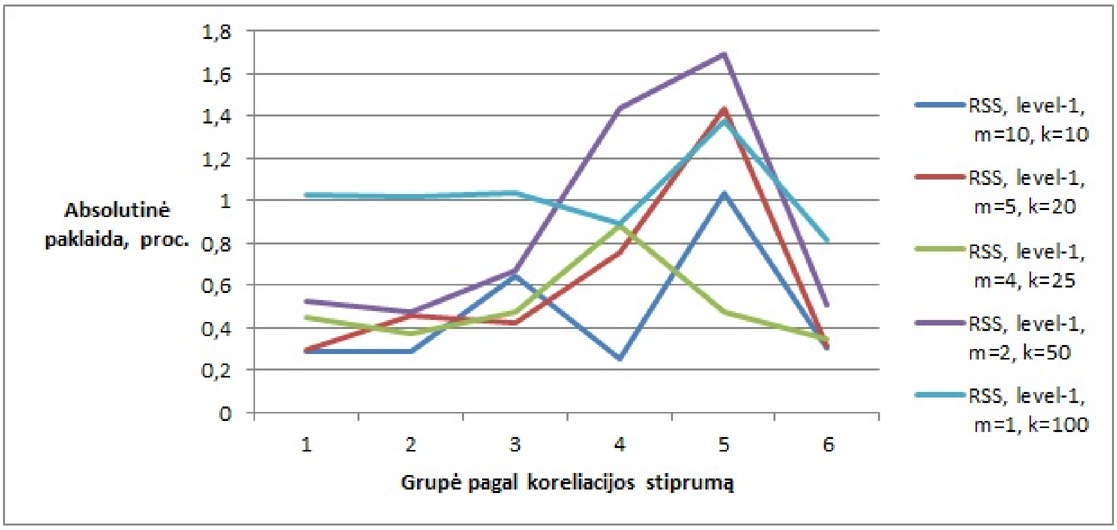
	RSS level-1					RSS level-2				
Ciklų skaičius m:	1	2	4	5	10	1	2	4	5	10
Įverčių skaičius:	11	13	20	21	20	17	16	22	23	24

8 lentelė. Įverčių, kurių paklaida e mažesnė už SRS paklaidą, skaičius

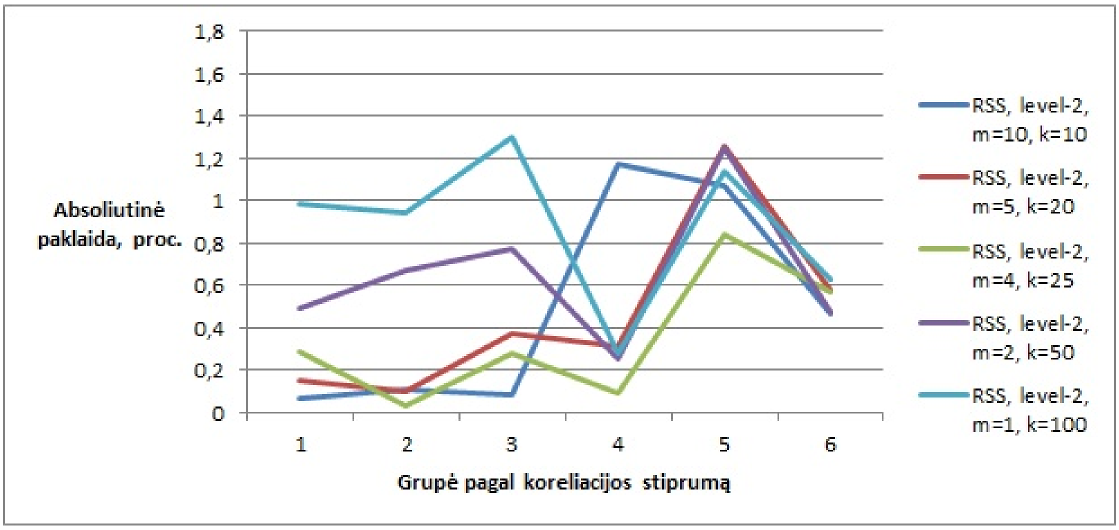
Norėdama palyginti, kaip skiriasi įverčių absoliutinės paklaidos, esant skirtingo stiprumo koreliacijai tarp kintamųjų ir naudojant skirtingus metodus, nubrėžiau absoliutinės paklaidos e procentine išraiška grafikus (pav. 6 – 8). Čia ir toliau grupės numeriai, atidėti horizontalioje ašyje, žymi rodiklių grupes 5 lentelėje.



6 pav. Absoliutinė paklaida, proc. SRS metodas



7 pav. Absoliutinė paklaida, proc. RSS 1-ojo lygio metodas



8 pav. Absoliutinė paklaida, proc. RSS 2-ojo lygio metodas

Nė vienoje grupėje SRS metodu gautų įverčių paklaidų vidurkis neviršija 1,39 proc. Esant įvairaus stiprumo teigiamai koreliacijai, paklaidos yra labai panašios. Jos išauga grupėse, kuriose tiesinio ryšio tarp tyrimo ir pagalbinių kintamųjų nėra arba yra labai silpna neigiama koreliacija. Čia reikėtų atkreipti dėmesį, kad šis paklaidos padidėjimas niekaip nesusijęs su koreliacija tarp tyrimo kintamojo ir SP, nes SRS atveju jokia papildoma informacija nėra naudojama. Šiose grupėse paklaidos galimai išauga dėl to, kad kai kuriuos į grupę patenkančius augalus augina labai mažai ūkių. Abiejų lygių RSS metodais gautų įverčių paklaidų kaitos kreivė yra panaši į SRS paklaidų. 1-ojo lygio RSS įverčių paklaidų vidurkis neviršija 1,69 proc., o 2-ojo lygio RSS — 1,31 proc. 1-ojo lygio RSS atveju išsiskiria įverčiai, kai naudotas vienas ciklas, o 2-ojo lygio atveju — kai naudotas 1 ir 2 ciklai: šių įverčių paklaidos esant stipriai, vidutinei ir silpnai teigiamai koreliacijai viršija paklaidas, kai naudotas didesnis ciklų skaičius. 22 atvejais iš 30 RSS *level-2* įverčiai grupėse pagal koreliacijos stiprumą turi mažesnę paklaidą nei RSS *level-1*. Be to, esant silpnai neigiamai koreliacijai arba kai ryšio nėra, *level-2* įverčių paklaidos tarpusavyje skiriasi mažiau nei *level-1* imties plano paklaidos.

Toliau nagrinėjau vidutinę absoliutinę paklaidą $\bar{e}$ grupėse pagal koreliacijos stiprumą (9 lentelė).

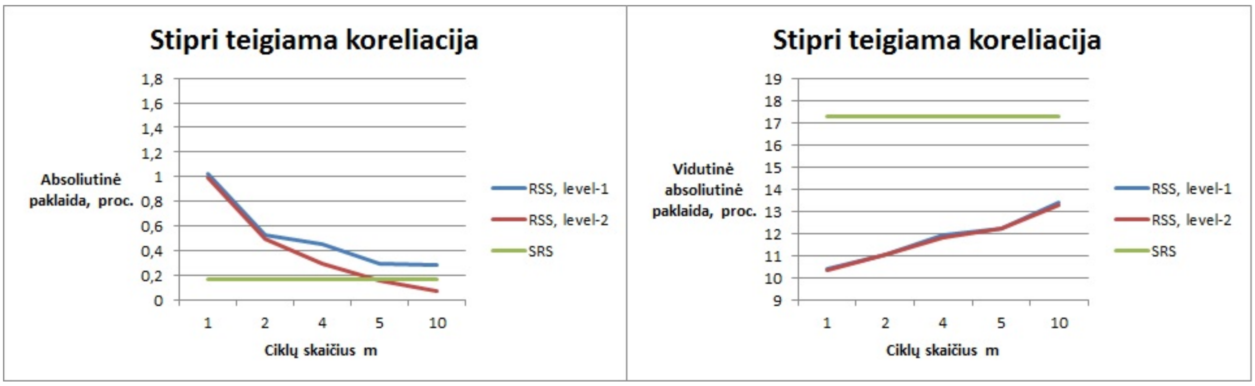
Nr.	SRS	RSS, level-1, m=1	RSS, level-1, m=2	RSS, level-1, m=4	RSS, level-1, m=5	RSS, level-1, m=10	RSS, level-2, m=1
1	17,262	10,384	11,056	11,912	12,199	13,413	10,346
2	26,732	20,629	21,062	21,941	22,278	23,428	20,532
3	47,909	45,173	45,772	46,203	46,152	46,353	45,656
4	61,679	62,179	61,697	61,141	61,684	61,599	61,449
5	112,443	112,785	114,265	113,069	113,522	113,081	113,046
6	64,124	64,208	64,176	63,875	64,178	64,075	64,05

Nr.	RSS, level-2, m=2	RSS, level-2, m=4	RSS, level-2, m=5	RSS, level-2, m=10
1	11,021	11,792	12,221	13,3
2	21,101	21,878	22,041	23,207
3	45,802	45,717	46,019	46,248
4	61,848	61,568	61,772	61,03
5	113,174	113,283	112,145	112,619
6	64,14	64,024	63,9	64,393

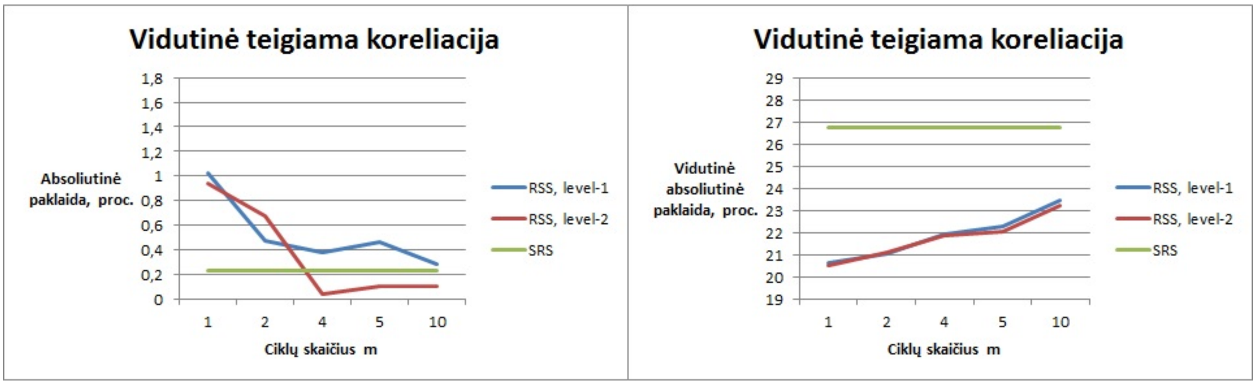
9 lentelė. Vidutinė absoliutinė paklaida grupėse pagal koreliacijos stiprumą

Grupėse, kuriose koreliacija didesnė nei 0,5, RSS metodo paklaidos truputį mažesnės už SRS metodo: pirmoje grupėje skirtumas yra maždaug 4–7 proc. punktai, antroje grupėje — 2–6 proc. punktai. Kitose grupėse jokių reikšmingų skirtumų tarp SRS ir RSS metodų nematyti. Lyginant RSS metodą skirtingų parametrų m ir k reikšmėms, galima pastebėti, kad paklaidos truputį skiriasi, kai koreliacija stipri ir vidutinė (kai $m = 1$, paklaidos mažiausios, o kai $m = 10$ — didžiausios), ir beveik nesiskiria, kai koreliacija yra silpnesnė.

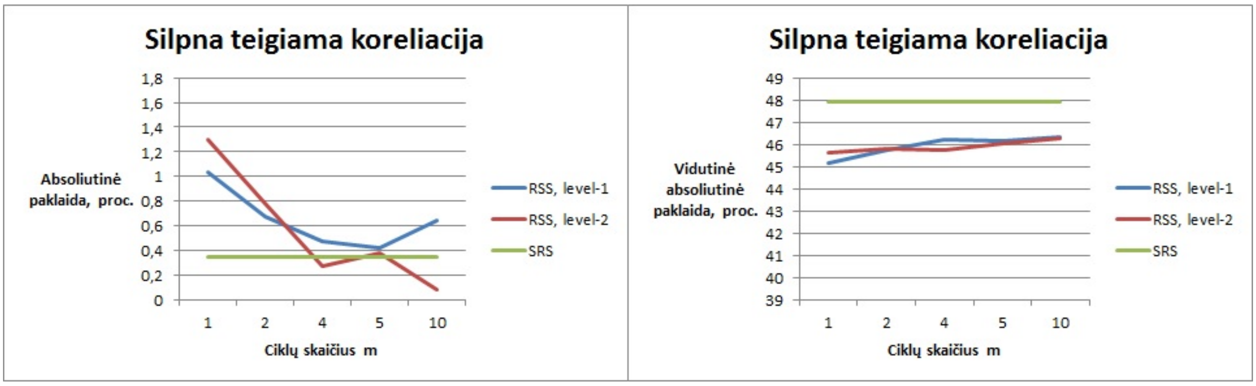
Vidutinę absoliutinę paklaidą $\bar{e}$ kiekvienoje grupėje pagal koreliacijos stiprumą palyginau su absoliutine paklaida e (9 – 14 pav.).



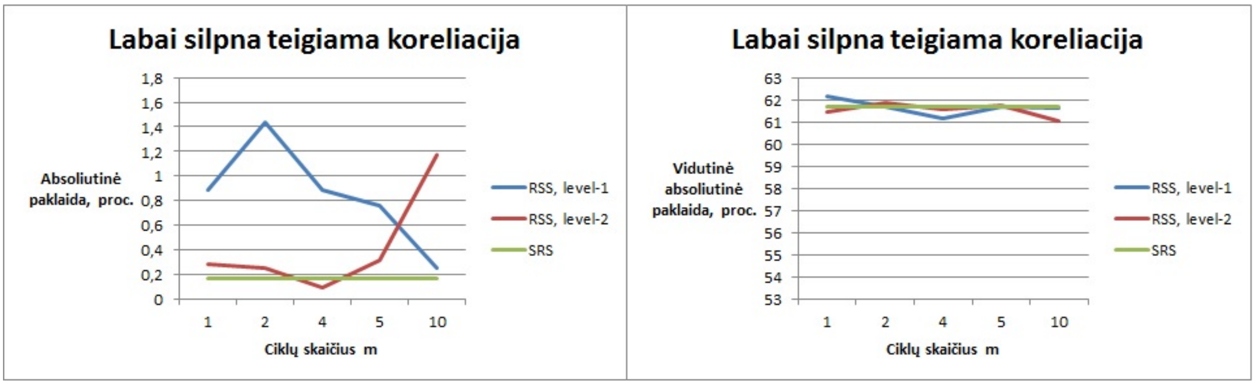
9 pav. Absoliutinė ir vidutinė absoliutinė paklaidos, proc. Stipri teigiama koreliacija



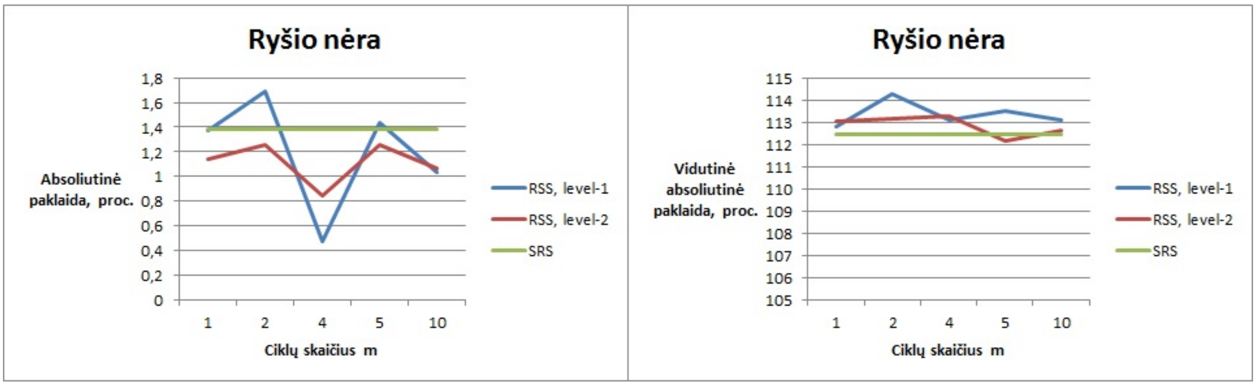
10 pav. Absoliutinė ir vidutinė absoliutinė paklaidos, proc. Vidutinė teigiama koreliacija



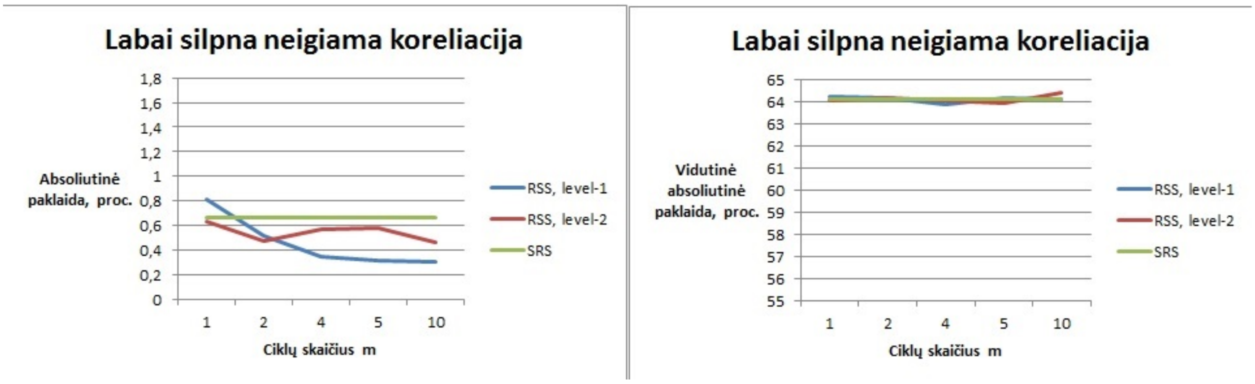
11 pav. Absoliutinė ir vidutinė absoliutinė paklaidos, proc. Silpna teigiama koreliacija



12 pav. Absoliutinė ir vidutinė absoliutinė paklaidos, proc. Labai silpna teigiama koreliacija



13 pav. Absoliutinė ir vidutinė absoliutinė paklaidos, proc. Ryšio nėra

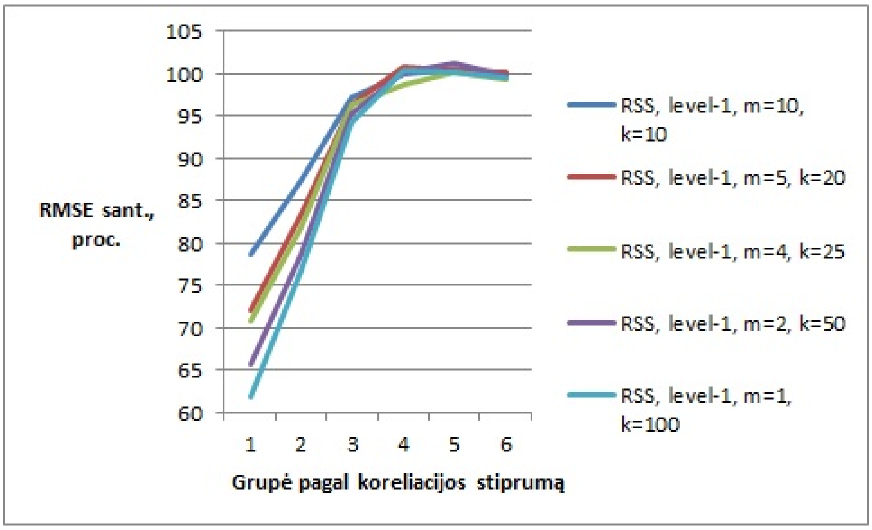


14 pav. Absoliutinė ir vidutinė absoliutinė paklaidos, proc. Labai silpna neigiama koreliacija

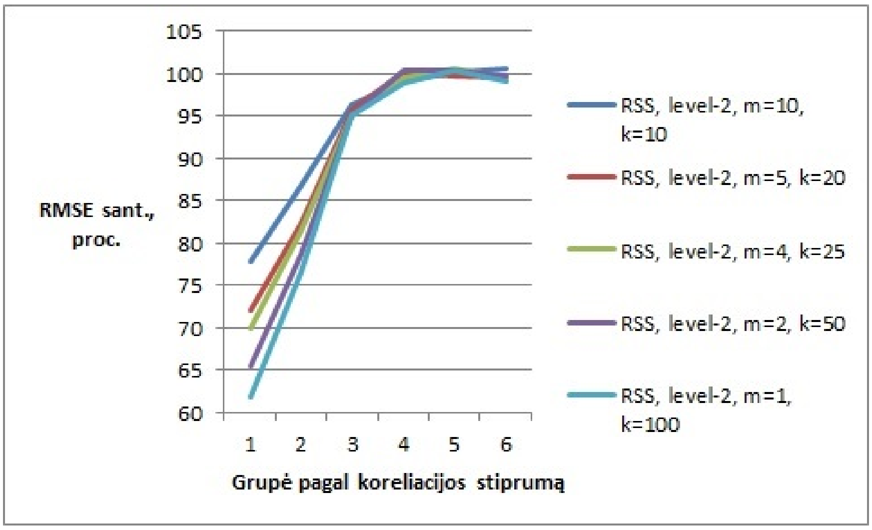
Stiprios ir vidutinės teigiamos koreliacijos grupėse (9 – 10 pav.) RSS metodo (abiejų lygių) vidutinė absoliutinė paklaida $\bar{e}$ gana ženkliai mažesnė už SRS metodo paklaidą. Skirtumas tarp RSS ir SRS paklaidų didžiausias, kai naudojamas 1 ciklas, ir jis mažėja didinant ciklų skaičių (nuo beveik 7 iki maždaug 3,8 procentinių punktų, esant stipriai koreliacijai, ir nuo apytiksliai 6,1 procentinio punkto iki 3,3, esant vidutinei koreliacijai). Absoliutinės paklaidos e atžvilgiu situacija priešinga: RSS paklaida didžiausia naudojant parametrus ($m = 1, k = 100$) ir mažėja didinant parametro m reikšmę. Pirmojo lygio RSS paklaida šiose grupėse visada viršija SRS paklaidą. Antrojo lygio RSS paklaida beveik visoms parametrų m ir k reikšmėms yra mažesnė už pirmojo lygio RSS paklaidą, be to, kai $m \geq 5$ ir $m \geq 4$, atitinkamai pirmoje ir antroje grupėse pagal koreliacijos stiprumą RSS paklaida tampa, nors ir nežymiai, bet mažesnė už SRS paklaidą. Silpnos teigiamos koreliacijos grupėje (11 pav.) RSS paklaida $\bar{e}$ vis dar mažesnė už SRS paklaidą, bet šis skirtumas dar mažesnis nei anksčiau aprašytoje grupėje. Taip pat iki maždaug 1 proc. punkto sumažėja skirtumas tarp RSS paklaidos, kai $m = 1$ ir $m = 10$. Absoliutinių paklaidų e išsidėstymas šioje grupėje nedaug skiriasi nuo paklaidų vidutinio stiprumo koreliacijos grupėje. Esant labai silpnai teigiamai ir neigiamai koreliacijai (12 ir 14 pav.) vidutinė absoliutinė paklaida $\bar{e}$ beveik sutampa abiejų lygių RSS ir SRS metodams, o kai ryšio tarp rangavimo ir tyrimo kintamųjų nėra (13 pav.), RSS abiejų lygių įverčiai paklaidos $\bar{e}$ atžvilgiu yra šiek tiek prastesni nei SRS įverčiai. Tuo tarpu absoliutinė paklaida e paskutinėse dviejose grupėse yra mažesnė RSS įverčiams.

Apibendrinant galima sakyti, kad paklaidos $\bar{e}$ atžvilgiu RSS metodas yra pranašesnis, kai koreliacija stipri, o šis pranašumas mažėja, kai koreliacija darosi silpnesnė. Paklaidos e atžvilgiu esant stipriai koreliacijai geriau pasirodo SRS metodas, o esant silpnai neigiamai koreliacijai ir kai ryšio nėra — RSS metodas. Vis dėlto, reikia atkreipti dėmesį į tai, kad absoliutinės paklaidos e skalė labai skiriasi nuo vidutinės absoliutinės paklaidos $\bar{e}$. Todėl nagrinėjant paklaidą e vieno ar kito metodo pranašumas prieš kitą pasireiškia tik dešimtosiomis procento dalimis.

Kadangi *RMSE* rezultatai nelabai skiriasi nuo vidutinės absoliutinės paklaidos $\bar{e}$, jų čia neapžvelgsiu ir pereisiu prie *RMSE* santykio tarp RSS ir SRS metodų (15 – 16 pav.).



15 pav. RSS *level-1* ir SRS metodų *RMSE* santykis



16 pav. RSS *level-2* ir SRS metodų *RMSE* santykis

Didelių skirtumų tarp 1-ojo ir 2-ojo lygio RSS įverčių nematyti. RSS ėmimo įverčių $\boldsymbol{RMSE}$ yra mažesnė nei SRS metodo $\boldsymbol{RMSE}$ pirmose trijose grupėse, t. y. kai koreliacija didesnė nei 0,2. Kai koreliacija stipri, abiejų lygių RSS ėmimo atvejais didžiausias skirtumas tarp RSS ir SRS paklaidų yra tada, kai naudojamas tik vienas ciklas ($\boldsymbol{RMSE_{RSS}}$ sudaro beveik 62 proc. nuo SRS metodo $\boldsymbol{RMSE}$) ir šis skirtumas palaipsniui mažėja, kai ciklų skaičius didinamas iki 10 ($\boldsymbol{RMSE_{RSS}}$ sudaro apie 78 proc. nuo $\boldsymbol{RMSE_{SRS}}$). Kai koreliacija vidutiniškai stipri, santykinė $\boldsymbol{RMSE}$ svyruoja nuo maždaug 76,5 proc., kai $\boldsymbol{m} = 1$, iki maždaug 87 proc., kai $\boldsymbol{m} = 10$. Trečioje grupėje pagal koreliacijos stiprumą, didinant ciklų skaičių, santykinė $\boldsymbol{RMSE}$ kinta nuo 94,16 proc. iki 97,2 proc., kai naudojamas *level-1* imties planas, ir nuo 94,98 proc. iki 96,4 proc., kai naudojamas *level-2* planas. Kai koreliacija dar silpnesnė, su visomis parametrų $\boldsymbol{m}$ ir $\boldsymbol{k}$ reikšmėmis ir abiejų lygių planais RSS ir SRS paklaidų santykis yra apie 99–101 proc.

Išvados

Nagrinėjant absoliutinę paklaidą e galima pastebėti, kad tiek RSS, tiek SRS metodais gautos reikšmės yra mažos: kai $m \leq 2$, RSS įverčių paklaida yra kiek didesnė nei SRS, o likusiais atvejais gana panaši. Remiantis didžiųjų skaičių dėsniu,

$$\bar{\hat{X}} \approx E\hat{X} = \text{tikroji populiacijos suma.}$$

Taigi ši paklaida iš esmės rodo Monte–Karlo metodo paklaidą vertinant įvertinių vidurkį ir nėra labai svarbi praktiniu požiūriu.

Remiantis tuo pačiu didžiųjų skaičių dėsniu,

$$\begin{aligned} RMSE &\approx \sqrt{E(\hat{X} - \text{tikroji populiacijos suma})^2}, \\ RMSE_{sant.} &\approx \sqrt{\frac{E(\hat{X}_{RSS} - \text{tikroji populiacijos suma})^2}{E(\hat{X}_{SRS} - \text{tikroji populiacijos suma})^2}}, \\ \bar{e} &\approx E|\bar{X} - \text{tikroji populiacijos suma}|. \end{aligned}$$

Kadangi simuliacijoms naudotas gana didelis iteracijų skaičius — po 25000 kiekvienam parametrų rinkiniui — tai dešinėje pusėje esantys teoriniai vidurkiai įvertinti gana tiksliai[4]. Todėl darbe atlikti skaičiavimai leidžia teigti, kad RSS metodas gana reikšmingai pagerina įvertinių tikslumą nagrinėtomis sąlygomis.

[4]analiziškai vertinimo paklaida nevertinta

Santrauka

Beržvinskaitė R. Ranguotų aibių ėmimo įvertinių taikymas žemės ūkio srities duomenims: Statistikos magistro baigiamasis darbas / vadovas doc. dr. V. Skorniakov; Vilniaus universitetas, Matematikos ir informatikos fakultetas, Matematinės statistikos katedra. – Vilnius, 2017. – 26 p.

Šiame darbe trumpai pristatyti ranguotų aibių ėmimo įvertiniai ir teorinis jų pranašumas prieš paprastojo atsitiktinio ėmimo įvertinius. RSS metodas pritaikytas praktiškai žemės ūkio srities duomenims. Simuliacijai naudoti pirmojo ir antrojo lygių RSS imčių planai ir penki metodo parametrų reikšmių rinkiniai, atlikta 25000 iteracijų. Gauti rezultatai palyginti su paprastosios atsitiktinės imties rezultatais įvairių rūšių paklaidų atžvilgiu. Padaryta išvada, kad esant pakankamai stipriai koreliacijai tarp tyrimo ir pagalbinio rangavimo kintamųjų bei tam tikroms metodo parametrų reikšmėms RSS metodu gautas įvertis yra tikslesnis nei SRS įvertis. Antrojo lygio RSS įverčiai kai kuriais atvejais yra truputį tikslesni nei pirmojo lygio įverčiai.

Summary

Beržvinskaitė R. Application of Ranked Set Sampling to Agricultural Data: Master's Work in Statistics / supervisor associate professor V. Skorniakov; Department of Mathematical Statistics, Faculty of Mathematics and Informatics, Vilnius University. – Vilnius, 2017. – 26 p.

Ranked set sampling estimators and their theoretical advantage against simple random sample estimators were shortly introduced in this work. RSS method was applied to agricultural data. There were *level-1* and *level-2* sample designs and five different sets of method parameters used to simulation. 25000 iterations were performed. The results gained with RSS method were compared to the results of simple random sampling with respect to different types of errors. It was concluded that RSS estimator was more precise than SRS estimator when the correlation between the variable of interest and ranking variable were strong enough and parameters of the method were well-chosen. RSS *level-2* estimators were slightly more precise than *level-1* estimators in several cases.

Literatūra

[1] D. A. Wolfe *Ranked Set Sampling: Its Relevance and Impact on Statistical Inference,* ISRN Probability and Statistics, 2012, Volume 2012, 32 p.

[2] B. Kowalczyk *Ranked Set Sampling and Its Applications in Finite Population Studies,* Statistics in Transition, December 2004 Vol. 6, No. 7, pp.1031-1046

[3] O. Ozturk *Estimation of Population Mean and Total in a Finite Population Setting Using Multiple Auxiliary Variables,* Journal of Agricultural, Biological and Environmental Statistics, Vol. 19, No. 2, pp. 161-184

[4] A. I. Al-Omari, C. N. Bouza *Review of Ranked Set Sampling: Modifications and Applications,* Revista Investigacion Operacional, 2014, Vol. 35, No. 3, pp. 215-240

[5] Žemės ūkio augalų pasėlių, derliaus ir derlingumo statistinio tyrimo metodika, 2012 m. https://osp.stat.gov.lt/documents/10180/550594/ZU_augalu_paseliu_metodika_2012.pdf

[6] G. P. Patil, A. K. Sinha, C. Taillie *Finite Population Corrections for Ranked Set Sampling,* Ann. Inst. Statist. Math., 1995, Vol. 47, No. 4, pp. 621-636

[7] T. R. Dell, J. L. Clutter *Ranked set sampling theory with order statistics background*, Biometrics, 1972, Vol. 28, pp. 545-555

[8] C. J. Stroka *Extending RSS to survey methodology,* dissertation, The Ohio State University, 2008

[9] C. E. Husby, E. A. Stasny, D. A. Wolfe *An application of Ranked Set Sampling for Mean and Median Estimation Using USDA Crop Production Data*, Journal of Agricultural, Biological and Environmental Statistics, 2005, Vol. 10, No. 3, pp. 354-373

Priedai

1. Koreliacijos koeficientų įverčiai

n=3779	SP	D1	D2	D3	D4
SP	1.00000000	0.06639462	-	-	-
D1	0.06639462	1.00000000	-	-	-
D2	-	-	-	-	-
D3	-	-	-	-	-
D4	-	-	-	-	-
n=6476	SP	D1	D2	D3	D4
SP	1.000000	-0.14909	-	-	-
D1	-0.14909	1.000000	-	-	-
D2	-	-	-	-	-
D3	-	-	-	-	-
D4	-	-	-	-	-

1 lentelė. Spirmeno koreliacijos koeficientai rodikliui 101

n=4493	SP	D1	D2	D3	D4
SP	1.0000000	0.8758932	0.8756131	0.8548861	0.8548698
D1	0.8758932	1.0000000	0.9999091	0.9789234	0.9788858
D2	0.8756131	0.9999091	1.0000000	0.9790704	0.9790337
D3	0.8548861	0.9789234	0.9790704	1.0000000	0.9999419
D4	0.8548698	0.9788858	0.9790337	0.9999419	1.0000000
n=6476	SP	D1	D2	D3	D4
SP	1.00000	0.76081	0.76053	0.75140	0.75139
D1	0.76081	1.00000	0.99989	0.99268	0.99267
D2	0.76053	0.99989	1.00000	0.99280	0.99279
D3	0.75140	0.99268	0.99280	1.00000	0.99998
D4	0.75139	0.99267	0.99279	0.99998	1.00000

2 lentelė. Spirmeno koreliacijos koeficientai rodikliui 140

n=1393	SP	D1	D2	D3	D4
SP	1.0000000	0.7798317	0.7746024	0.7407572	0.7400876
D1	0.7798317	1.0000000	0.9947062	0.9620954	0.9619875
D2	0.7746024	0.9947062	1.0000000	0.9676972	0.9675883
D3	0.7407572	0.9620954	0.9676972	1.0000000	0.9998635
D4	0.7400876	0.9619875	0.9675883	0.9998635	1.0000000
n=6476	SP	D1	D2	D3	D4
SP	1.00000	0.53047	0.52851	0.52731	0.52729
D1	0.53047	1.00000	0.99572	0.99509	0.99509
D2	0.52851	0.99572	1.00000	0.99937	0.99937
D3	0.52731	0.99509	0.99937	1.00000	1.00000
D4	0.52729	0.99509	0.99937	1.00000	1.00000

3 lentelė. Spirmeno koreliacijos koeficientai rodikliui 164

n=852	SP	D1	D2	D3	D4
SP	1.0000000	0.6476533	0.6420609	0.6142982	0.6135135
D1	0.6476533	1.0000000	0.9885237	0.8981292	0.8976317
D2	0.6420609	0.9885237	1.0000000	0.9116608	0.9111672
D3	0.6142982	0.8981292	0.9116608	1.0000000	0.9997242
D4	0.6135135	0.8976317	0.9111672	0.9997242	1.0000000
n=6476	SP	D1	D2	D3	D4
SP	1.00000	0.32253	0.32203	0.32137	0.32136
D1	0.32253	1.00000	0.99415	0.99355	0.99354
D2	0.32203	0.99415	1.00000	0.99941	0.99941
D3	0.32137	0.99355	0.99941	1.00000	1.00000
D4	0.32136	0.99354	0.99941	1.00000	1.00000

4 lentelė. Spirmeno koreliacijos koeficientai rodikliui 165

n=2760	SP	D1	D2	D3	D4
SP	1.0000000	0.2318245	0.2263875	0.2303615	-
D1	0.2318245	1.0000000	0.9899788	0.8543246	-
D2	0.2263875	0.9899788	1.0000000	0.8650044	-
D3	0.2303615	0.8543246	0.8650044	1.0000000	-
D4	-	-	-	-	-
n=6476	SP	D1	D2	D3	D4
SP	1.00000	-0.13540	-0.13700	-0.13658	-
D1	-0.13540	1.00000	0.99691	0.98397	-
D2	-0.13700	0.99691	1.00000	0.98712	-
D3	-0.13658	0.98397	0.98712	1.00000	-
D4	-	-	-	-	-

5 lentelė. Spirmeno koreliacijos koeficientai rodikliui 177

n=1405	SP	D1	D2	D3	D4
SP	1.0000000	0.2765071	0.2676105	0.2668055	-
D1	0.2765071	1.0000000	0.9849236	0.8593742	-
D2	0.2676105	0.9849236	1.0000000	0.8751032	-
D3	0.2668055	0.8593742	0.8751032	1.0000000	-
D4	-	-	-	-	-
n=6476	SP	D1	D2	D3	D4
SP	1.00000	-0.09698	-0.09849	-0.09853	-
D1	-0.09698	1.00000	0.99395	0.8593742	-
D2	-0.09849	0.99642	1.00000	0.99755	-
D3	-0.09853	0.99395	0.99755	1.00000	-
D4	-	-	-	-	-

6 lentelė. Spirmeno koreliacijos koeficientai rodikliui 181

n=229	SP	D1	D2	D3	D4
SP	1.0000000	0.8186860	0.8096836	0.7669102	-
D1	0.8186860	1.0000000	0.9942344	0.9324750	-
D2	0.8096836	0.9942344	1.0000000	0.9392563	-
D3	0.7669102	0.9324750	0.9392563	1.0000000	-
D4	-	-	-	-	-
n=6476	SP	D1	D2	D3	D4
SP	1.00000	0.21472	0.21409	0.21400	-
D1	0.21472	1.00000	0.99343	0.99340	-
D2	0.21409	0.99343	1.00000	0.99997	-
D3	0.21400	0.99340	0.99997	1.00000	-
D4	-	-	-	-	-

7 lentelė. Spirmeno koreliacijos koeficientai rodikliui 185

n=340	SP	D1	D2	D3	D4
SP	1.0000000	0.4833379	0.4857763	0.4249157	-
D1	0.4833379	1.0000000	0.9984562	0.8263718	-
D2	0.4857763	0.9984562	1.0000000	0.8306820	-
D3	0.4249157	0.8263718	0.8306820	1.0000000	-
D4	-	-	-	-	-
n=6476	SP	D1	D2	D3	D4
SP	1.00000	0.09520	0.09522	0.09484	-
D1	0.09520	1.00000	1.00000	0.99983	-
D2	0.09522	1.00000	1.00000	0.99984	-
D3	0.09484	0.99983	0.99984	1.00000	-
D4	-	-	-	-	-

8 lentelė. Spirmeno koreliacijos koeficientai rodikliui 190

n=3312	SP	D1	D2	D3	D4
SP	1.0000000	0.2129261	-	-	-
D1	0.2129261	1.0000000	-	-	-
D2	-	-	-	-	-
D3	-	-	-	-	-
D4	-	-	-	-	-
n=6476	SP	D1	D2	D3	D4
SP	1.00000	0.01875	-	-	-
D1	0.01875	1.00000	-	-	-
D2	-	-	-	-	-
D3	-	-	-	-	-
D4	-	-	-	-	-

9 lentelė. Spirmeno koreliacijos koeficientai rodikliui 200

n=96	SP	D1	D2	D3	D4
SP	1.0000000	0.5513324	0.3530434	0.3789247	-
D1	0.5513324	1.0000000	0.6119610	0.4921679	-
D2	0.3530434	0.6119610	1.0000000	0.8940253	-
D3	0.3789247	0.4921679	0.8940253	1.0000000	-
D4	-	-	-	-	-
n=6476	SP	D1	D2	D3	D4
SP	1.00000	0.01931	0.01751	0.01754	-
D1	0.01931	1.00000	0.91706	0.91705	-
D2	0.01751	0.91706	1.00000	0.99999	-
D3	0.01754	0.91705	0.99999	1.00000	-
D4	-	-	-	-	-

10 lentelė. Spirmeno koreliacijos koeficientai rodikliui 211

n=1029	SP	D1	D2	D3	D4
SP	1.00000000	0.1227394	0.09694985	0.09999165	-
D1	0.12273945	1.0000000	0.66284665	0.65135451	-
D2	0.09694985	0.6628466	1.00000000	0.98491836	-
D3	0.09999165	0.6513545	0.98491836	1.00000000	-
D4	-	-	-	-	-
n=6476	SP	D1	D2	D3	D4
SP	1.0000	-0.11989	-0.01984	-0.01882	-
D1	-0.11989	1.000	0.46915	0.46422	-
D2	-0.01984	0.46915	1.0000	0.98973	-
D3	-0.01882	0.46422	0.98973	1.0000	-
D4	-	-	-	-	-

11 lentelė. Spirmeno koreliacijos koeficientai rodikliui 230

n=277	SP	D1	D2	D3	D4
SP	1.000000000	-0.06021223	-0.002049642	-0.004429483	-
D1	-0.060212227	1.00000000	0.373136763	0.371116299	-
D2	-0.002049642	0.37313676	1.000000000	0.926607257	-
D3	-0.004429483	0.37111630	0.926607257	1.000000000	-
D4	-	-	-	-	-
n=6476	SP	D1	D2	D3	D4
SP	1.00000	-0.06491	-0.01667	-0.01704	-
D1	-0.06491	1.00000	0.22915	0.22911	-
D2	-0.01667	0.22915	1.00000	0.92849	-
D3	-0.01704	0.22911	0.92849	1.00000	-
D4	-	-	-	-	-

12 lentelė. Spirmeno koreliacijos koeficientai rodikliui 231

n=611	SP	D1	D2	D3	D4
SP	1.0000000	0.1731736	0.2013580	0.1956208	-
D1	0.1731736	1.0000000	0.7852313	0.7690285	-
D2	0.2013580	0.7852313	1.0000000	0.9722652	-
D3	0.1956208	0.7690285	0.9722652	1.0000000	-
D4	-	-	-	-	-
n=6476	SP	D1	D2	D3	D4
SP	1.00000	-0.10680	-0.04117	-0.04059	-
D1	-0.10680	1.00000	0.64153	0.63764	-
D2	-0.04117	0.64153	1.00000	0.98928	-
D3	-0.04059	0.63764	0.98928	1.00000	-
D4	-	-	-	-	-

13 lentelė. Spirmeno koreliacijos koeficientai rodikliui 238

n=2456	SP	D1	D2	D3	D4
SP	1.00000000	-	0.08233377	0.08245976	-
D1	-	-	-	-	-
D2	0.08233377	-	1.00000000	0.95068006	-
D3	0.08245976	-	0.95068006	1.00000000	-
D4	-	-	-	-	-
n=6476	SP	D1	D2	D3	D4
SP	1.000000	-	-0.17037	-0.17006	-
D1	-	-	-	-	-
D2	-0.17037	-	1.000000	0.99646	-
D3	-0.17006	-	0.99646	1.000000	-
D4	-	-	-	-	-

14 lentelė. Spirmeno koreliacijos koeficientai rodikliui 1011

n=2140	SP	D1	D2	D3	D4
SP	1.00000000	-	0.2163002	0.2072667	-
D1	-	-	-	-	-
D2	0.2163002	-	1.00000000	0.9538555	-
D3	0.2072667	-	0.9538555	1.00000000	-
D4	-	-	-	-	-
n=6476	SP	D1	D2	D3	D4
SP	1.000000	-	-0.04075	-0.04184	-
D1	-	-	-	-	-
D2	-0.04075	-	1.000000	0.99762	-
D3	-0.04184	-	0.99762	1.000000	-
D4	-	-	-	-	-

15 lentelė. Spirmeno koreliacijos koeficientai rodikliui 2001

2. R programomos kodas

```
########################################
###     KORELIACIJOS KOEFICIENTU IVERCIAI     ###
duomenys<-pop_duom
duomenys[,2]<-as.factor(duomenys[,2])

#Koreliacijos koeficientu iverciai
(koreliacijos<-by(duomenys[,c(3,5, 6, 7, 8)],duomenys[,2], function(x){cor(x, method="spearman")} ))

#H: koreliacija=0
library(Hmisc)
(koreliacijos2<-by(duomenys[,c(3,5, 6, 7, 8)],duomenys[,2], function(x){rcorr(as.matrix(x),
                                                                              type="spearman")} ))

#dazniu lentele
(dazniai<-table(duomenys$rod_id))

########################################
###      VIENOS IMTIES ISRINKIMAS SRS METODU      ###

#pop_sar - lentele, kurioje yra ukiu sarasas su papildomu kintamuoju,
#pagal kuri ranguosime;
#pop_duom - lentele, kurioje yra tyrimo duomenys: visi tiriami rodikliai, d1-d4;
#imt_sar - sudaromas atrinktu i galutine imti ukiu sarasas;
#imt_duom - i imti atrinktu ukiu duomenys: visi tiriami rodikliai, d1-d4;
#m =0;
#k =0;
#n - imties dydis;
#rankvar -tuščias;


SRS<-function(pop_sar, pop_duom, n, m=0, k=0, rankvar1=NULL, rankvar2=NULL )
{
  N<-nrow(pop_sar)

  imt_sar<-pop_sar[sample(1:N, n, replace=FALSE),]
  imt_duom<-merge(imt_sar, pop_duom, all.x=TRUE)

  rm(imt_sar)

  return(imt_duom)
}

###########################################
###     VIENOS IMTIES ISRINKIMAS RSS level-1 METODU   ###

#pop_sar - lentele, kurioje yra ukiu sarasas su papildomu kintamuoju,
#pagal kuri ranguosime;
#pop_duom - lentele, kurioje yra tyrimo duomenys: visi tiriami rodikliai, d1-d4;
```

```
#imt_sar - sudaromas atrinktu i galutine imti ukiu sarasas;
#imt_duom - i imti atrinktu ukiu duomenys: visi tiriami rodikliai, d1-d4;
#m - vieno ciklo imties dydis;
#k - ciklu skaicius;
#n - galutines imties dydis;
#rankvar - kintamasis, pagal kuri ranguojami ukiai, turi buti lenteleje pop_sar;

RSS.level1<- function(pop_sar, pop_duom, n, m, k, rankvar1, rankvar2=NULL )
{
 likusi_pop<-pop_sar

 for (i in 1:k)
 {
  for (j in 1:m)
  {
   N<-nrow(likusi_pop)
   imtis<-likusi_pop[sample(1:N, m, replace=FALSE),]
   if(is.null(rankvar2))
   {
    imtis<-imtis[order(imtis[,rankvar1]),]

   }
   else
   {
    imtis<-imtis[order(imtis[,rankvar1], imtis[,rankvar2]),]
   }
   imtis<-imtis[j,]
   imtis$k<-i
   imtis$m<-j

   if (i==1 & j==1)
   {
    imt_sar<-imtis
   }
   else
   {
    imt_sar<-rbind(imt_sar, imtis)
   }
   library(sqldf)
   pag<-sqldf("select * from likusi_pop where U_ID not in
(select U_ID from imt_sar)")
   likusi_pop<-pag
   rm(pag)
  }
 }

 imt_duom<-merge(imt_sar, pop_duom, all.x=TRUE)

 rm(imt_sar)
 rm(imtis)
```

```
  rm(likusi_pop)

  return(imt_duom)
}

#########################################
###      VIENOS IMTIES ISRINKIMAS RSS level-2 METODU  ###

#pop_sar - lentele, kurioje yra ukiu sarasas su papildomu kintamuoju, pagal
# kuri ranguosime;
#pop_duom - lentele, kurioje yra tyrimo duomenys: visi tiriami rodikliai, d1-d4;
#imt_sar - sudaromas atrinktu i galutine imti ukiu sarasas;
#imt_duom - i imti atrinktu ukiu duomenys: visi tiriami rodikliai, d1-d4;
#m - vieno ciklo imties dydis;
#k - ciklu skaicius;
#n - galutines imties dydis;
#rankvar - kintamasis, pagal kuri ranguojami ukiai, turi buti lenteleje pop_sar;

RSS.level2<- function(pop_sar, pop_duom, n, m, k, rankvar1, rankvar2=NULL )
{
  likusi_pop<-pop_sar

  for (i in 1:k)
  {
    for (j in 1:m)
    {
      N<-nrow(likusi_pop)
      imtis<-likusi_pop[sample(1:N, m, replace=FALSE),]
      if(is.null(rankvar2))
      {
        imtis<-imtis[order(imtis[,rankvar1]),]

      }
      else
      {
        imtis<-imtis[order(imtis[,rankvar1], imtis[,rankvar2]),]
      }
      imtis<-imtis[j,]
      imtis$k<-i
      imtis$m<-j

      if (i==1 & j==1)
      {
        imt_sar<-imtis
      }
      else
      {
        imt_sar<-rbind(imt_sar, imtis)
      }
      library(sqldf)
```

```
    pag<-sqldf("select * from likusi_pop where U_ID not in (select U_ID from imtis)")
    likusi_pop<-pag
    rm(pag)
   }#end for(j)
  }#end for(i)

  imt_duom<-merge(imt_sar, pop_duom, all.x=TRUE)

  rm(imt_sar)
  rm(imtis)
  rm(likusi_pop)

  return(imt_duom)
}



##############################################
###    POPULIACIJOS SUMU D1-D4 IVERCIAI PASIRINKTAM RODIKLIUI  ###

#imt_duom -lentele, kurioje yra i imti patekusiu ukiu duomenys;
#rod - norimo rodiklio kodas;
#Npop - populiacijos dydis;
#n - imties dydis (n=k*m RSS atveju);

iverciai<-function(imt_duom, Npop, n)
{
  rod_id<-c(140, 181, 177, 230, 231, 238, 164, 165, 185, 190, 211, 101,
 200, 1011, 2001)
  pag<-function(rod)
  {
  duom<-imt_duom[which(imt_duom$rod_id==rod),]
  ivert<-as.data.frame(t(apply(duom[,c("D1", "D2", "D3", "D4")], 2, sum)))
  rm(duom)
  ivert$D1_ivert<-ivert$D1*Npop/n
  ivert$D2_ivert<-ivert$D2*Npop/n
  ivert$D3_ivert<-ivert$D3*Npop/n
  ivert$D4_ivert<-ivert$D4*Npop/n
  ivert$rod_id=rod
  ivert<-ivert[,c("D1_ivert", "D2_ivert", "D3_ivert", "D4_ivert", "rod_id")]
  return(ivert)
  }
  library(foreach)
  ivert<- foreach(rod = rod_id, .combine='rbind') %do%
  {
  pag(rod)
  }

  return(ivert)
}
```

```
##############################################
### 25000 KARTU ATLIEKAMAS IMTIES ISRINKIMAS IR VERTINIMAS ###

#iter - kiek kartu pasirinktu metodu rinksime imti ir skaiciuosime iverciu skaitines
#charakteristikas
#metodas - imties isrinkimo metodas, {"SRS", "RSS.level1", "RSS.level2"}
#m - vieno ciklo imties dydis;
#k - ciklu skaicius;
#n - galutines imties dydis

Monte_Karlas3<-function(pop_sar, pop_duom, iter, metodas, k=0, m=0, n=0,
rankvar1, rankvar2=NULL)
{
 if(exists("rezultatai")) #jei lentele "rezultatai" jau yra, tai ja istriname
 {
  rm(rezultatai)
 }

 #Nustatome darbini kataloga
 setwd('C:/Users/Rasa/Documents/...')

 #-----------------------------------------------------------------------------------------------------#
 #Pagalbine funkcija, kuri kreipiasi i imties isrinkimo funkcija, iverciu skaiciavimo funkcija ir grazina
 #vienos iteracijos metu gautus parametru ivercius
 pag<-function(pop_sar, pop_duom, iter, metodas, k, m, n, rankvar1, rankvar2)
 {
   #nurodome programa, kurioje aprasytos naudojamos imties isrinkimo ir iverciu
   #skaiciavimo funkcijos
   source("Funkcijos.R")

   #pasirinktu metodu isrenkame imti
   if (metodas=="SRS")
   {
    imties_duomenys<-SRS(pop_sar, pop_duom, n, rankvar1, rankvar2)
   }
   else if(metodas=="RSS.level1")
   {
    imties_duomenys<-RSS.level1(pop_sar, pop_duom, n, m, k, rankvar1, rankvar2)
   }
   else if (metodas=="RSS.level2")
   {
    imties_duomenys<-RSS.level2(pop_sar, pop_duom, n, m, k, rankvar1, rankvar2)
   }

   #kiekvienam rodikliui skaiciuojame ivercius
   iverciai.MK<-iverciai(imties_duomenys, 6476, n)
   iverciai.MK$iter_nr<-p

   #pasaliname imties, gautos sioje iteracijoje, duomenis ir pereiname prie kitos iteracijos
```

```
    rm(imties_duomenys)

    #graziname sioje iteracijoje gautus rezultatus
    return(iverciai.MK)
  }

  #-------------------------------------------------------------------------------------------------------#

  #Paralelizacijai naudosime biblioteka doParallel
  library(doParallel)
  cluster = makeCluster(4, type = "SOCK")
  registerDoParallel(cluster)
  library(foreach)

  #pagalbine funkcija pag iskvieciame "iter" kartu, iteracijos atliekamos paraleliai
  ivert.MK<-foreach(p =1:iter, .combine=rbind) %dopar%
  {
    pag(pop_sar, pop_duom, iter, metodas, k, m, n, rankvar1, rankvar2)

  }#end for(p)

  #uzdarome sukurta klasteri
  stopCluster(cluster)

  #Grazinami visu iteraciju rezultatai
  return(ivert.MK)
}

#SRS:
MK_SRS<-Monte_Karlas3(pop, pop_duom, 25000, "SRS",  n=100, rankvar1="ZEM_SO")

#RSS, level-1, 1x100:
MK_RSS1_1X100<-Monte_Karlas3(pop, pop_duom, 25000, "RSS.level1",  1,100, 100,
"ZEM_SO", "ZEM_PLOT")

#RSS, level-1, 2x50:
MK_RSS1_2X50<-Monte_Karlas3(pop, pop_duom, 25000, "RSS.level1",  2, 50, 100,
"ZEM_SO", "ZEM_PLOT")

#RSS, level-1, 4x25:
MK_RSS1_4X25<-Monte_Karlas3(pop, pop_duom, 25000, "RSS.level1",  4,25, 100,
"ZEM_SO", "ZEM_PLOT")

#RSS, level-1, 5x20:
MK_RSS1_5X20<-Monte_Karlas3(pop, pop_duom, 25000, "RSS.level1",  5, 20, 100,
"ZEM_SO", "ZEM_PLOT")

#RSS, level-1, 10x10:
MK_RSS1_10X10<-Monte_Karlas3(pop, pop_duom, 25000, "RSS.level1",  10, 10, 100,
"ZEM_SO", "ZEM_PLOT")
```

```
#RSS, level-2, 1x100
MK_RSS2_1X100<-Monte_Karlas3(pop, pop_duom, 25000, "RSS.level2",  1, 100, 100,
"ZEM_SO", "ZEM_PLOT")

#RSS, level-2, 2x50:
MK_RSS2_2X50<-Monte_Karlas3(pop, pop_duom, 25000, "RSS.level2",  2, 50, 100,
"ZEM_SO", "ZEM_PLOT")

#RSS, level-2, 4x25:
MK_RSS2_4X25<-Monte_Karlas3(pop, pop_duom, 25000, "RSS.level2",  4, 25, 100,
"ZEM_SO", "ZEM_PLOT")

#RSS, level-2, 5x20:
MK_RSS2_5X20<-Monte_Karlas3(pop, pop_duom, 25000, "RSS.level2",  5, 20, 100,
"ZEM_SO", "ZEM_PLOT")

#RSS, level-2, 10x10:
MK_RSS2_10X10<-Monte_Karlas3(pop, pop_duom, 25000, "RSS.level2",  10, 10, 100,
"ZEM_SO", "ZEM_PLOT")

#########################
###  PAKLAIDU SKAICIAVIMAS    ###

#Tikrosios populiacijos sumu reiksmes
pasetas_plotas<-setNames(aggregate(pop_duom$D1, by=list(pop_duom$rod_id),
FUN=sum), c("rod_id", "D1_pop"))
nuimtas_plotas<-setNames(aggregate(pop_duom$D2, by=list(pop_duom$rod_id),
FUN=sum), c("rod_id", "D2_pop"))
prad_derlius<-setNames(aggregate(pop_duom$D3, by=list(pop_duom$rod_id),
FUN=sum), c("rod_id", "D3_pop"))
derlius_po_valymo<-setNames(aggregate(pop_duom$D4, by=list(pop_duom$rod_id),
 FUN=sum), c("rod_id", "D4_pop"))

#duom - lentele su visu iteraciju iverciais
paklaidos<-function(duom)
{
  library(foreach)
  #prijungiame kiekvieno rodiklio tikrasias populiacijos sumas
  duom<-Reduce(function(x, y) merge(x, y, by="rod_id"), list(duom, pasetas_plotas,
nuimtas_plotas, prad_derlius, derlius_po_valymo ))

 #kiekvienai iteracijai skaiciuojame absoliutini skirtuma tarp ivercio ir tikrosios
#populaicijos sumos
 duom$abs_D1<-abs(duom$D1_ivert-duom$D1_pop)
 duom$abs_D2<-abs(duom$D2_ivert-duom$D2_pop)
 duom$abs_D3<-abs(duom$D3_ivert-duom$D3_pop)
 duom$abs_D4<-abs(duom$D4_ivert-duom$D4_pop)

 #lenteliu sujungimas, reikalinga naudojant cikla "foreach"
```

```
 func<-function(x, y){merge(x, y)}

 #iverciu vidurkis (is visu iteraciju)
 vidurkiai<-foreach(x=c("D1_ivert", "D2_ivert", "D3_ivert", "D4_ivert"),
 .combine='func') %do%
  {
    setNames(aggregate(duom[,x], by=list(duom$rod_id), FUN=mean),
c("rod_id", paste("vid",x, sep=".")))
  }

 #sujungiame visas paklaidu lenteles pagal rod_id ir prijungiame tikrasias populiacijos sumas
 duom2<-Reduce(function(x, y) merge(x, y, by="rod_id"), list(vidurkiai, vid_paklaida,
                                                     pasetas_plotas, nuimtas_plotas, prad_derlius,
derlius_po_valymo ))

  return(duom2)

}

pakl_SRS<-paklaidos(MK_SRS)
pakl_RSS1_10x10<-paklaidos(MK_RSS1_10x10)
pakl_RSS1_5x20<-paklaidos(MK_RSS1_5x20)
pakl_RSS1_4x25<-paklaidos(MK_RSS1_4x25)
pakl_RSS1_1x100<-paklaidos(MK_RSS1_1x100)
pakl_RSS1_2x50<-paklaidos(MK_RSS1_2x50)
pakl_RSS2_5x20<-paklaidos(MK_RSS2_5x20)
pakl_RSS2_4x25<-paklaidos(MK_RSS2_4x25)
pakl_RSS2_1x100<-paklaidos(MK_RSS2_1x100)
pakl_RSS2_10x10<-paklaidos(MK_RSS2_10x10)
pakl_RSS2_2x50<-paklaidos(MK_RSS2_2x50)


#RMSE
RMSE<-function(duomenys){

func<-function(x, y){rbind(x, y)}

pag1<-function(duomenys, x) {
 duom<-duomenys[which(duomenys$rod_id==x),]
 rmse1<-sqrt(sum((duom$D1_ivert-pasetas_plotas[
                                        which(pasetas_plotas$rod_id==x),2])^2)/25000)
 rmse2<-sqrt(sum((duom$D2_ivert-nuimtas_plotas[
                                       which(nuimtas_plotas$rod_id==x),2])^2)/25000)
 rmse3<-sqrt(sum((duom$D3_ivert-prad_derlius[
                                      which(prad_derlius$rod_id==x),2])^2)/25000)
 rmse4<-sqrt(sum((duom$D4_ivert-derlius_po_valymo[
                                     which(derlius_po_valymo$rod_id==x),2])^2)/25000)
 rod_id<-x
 rmse1<-cbind(rod_id, rmse1, rmse2, rmse3, rmse4)
 return(rmse1)
```

```
}

pag2<-foreach(x=c(101,140,164,165,177,181,185,190,200,211,230,231,238,1011,2001),
           .combine='func') %do%
            {
                rmse<-pag1(duomenys, x)
             }

 return(pag2)
}

RMSE_SRS<-RMSE(MK_SRS)

RMSE_RSS1_10x10<-RMSE(MK_RSS1_10x10)
RMSE_RSS1_5x20<-RMSE(MK_RSS1_5x20)
RMSE_RSS1_4x25<-RMSE(MK_RSS1_4x25)
RMSE_RSS1_2x50<-RMSE(MK_RSS1_2x50)
RMSE_RSS1_1x100<-RMSE(MK_RSS1_1x100)

RMSE_RSS2_10x10<-RMSE(MK_RSS2_10x10)
RMSE_RSS2_5x20<-RMSE(MK_RSS2_5x20)
RMSE_RSS2_4x25<-RMSE(MK_RSS2_4x25)
RMSE_RSS2_2x50<-RMSE(MK_RSS2_2x50)
RMSE_RSS2_1x100<-RMSE(MK_RSS2_1x100)
```