PATVIRTINTA
Valstybės duomenų agentūros
generalinio direktoriaus
2024 m. sausio 19 d. įsakymu Nr. DĮ-29

STATISTINIO ATSKLEIDIMO KONTROLĖS VADOVAS

Turinys

I SKYRIUS
BENDROSIOS NUOSTATOS

1. Statistinio atskleidimo kontrolės vadovas (toliau – Vadovas) aprašo statistinio atskleidimo kontrolės procesą, statistinio atskleidimo rizikos vertinimo būdus atsižvelgiant į duomenų, įskaitant asmens duomenų, tarp jų ir specialių kategorijų asmens duomenų, statistinių duomenų, administracinių duomenų (toliau apibendrintai vadinama duomenimis) tipus, statistinio atskleidimo kontrolės metodų parinkimą ir taikymą, duomenų panaudojamumo ir informacijos praradimo matus.

2. Vadovo tikslas – užtikrinti statistinio stebėjimo vieneto / asmens duomenų, pakartotinio statistinių duomenų naudojimo konfidencialumą rengiant ir skelbiant statistinę informaciją bei teikiant statistinius duomenis.

3. Vadovu vadovaujasi Valstybės duomenų agentūros (toliau – agentūra) valstybės tarnautojai ir darbuotojai, dirbantys pagal darbo sutartis, kurie atlikdami priskirtas funkcijas tvarko duomenis ar atlieka su duomenų tvarkymu susijusius veiksmus.

4. Agentūros valstybės tarnautojai ir darbuotojai, dirbantys pagal darbo sutartis, ar kitos oficialiąją statistiką tvarkančios įstaigos prireikus konsultuojasi su Metodologijos ir duomenų mokslo grupe (dėl statistinio atskleidimo kontrolės metodų taikymo), Duomenų apsaugos skyriumi (dėl Bendrojo duomenų apsaugos reglamento nustatytų asmens duomenų apsaugos reikalavimų taikymo), Konfidencialių duomenų valdymo komisija, sudaryta Lietuvos statistikos departamento generalinio direktoriaus 2022 m. sausio 28 d. įsakymu Nr. DĮ-39 „Dėl Konfidencialių duomenų valdymo komisijos sudarymo ir jos darbo reglamento patvirtinimo“ (dėl Vadovo taikymo ar probleminių konfidencialių duomenų valdymo atvejų).

II SKYRIUS
PAGRINDINĖS SĄVOKOS IR PAAIŠKINIMAI

5. Vadove vartojamos sąvokos:

5.1. **Asmens duomenys** – bet kokia informacija apie fizinį asmenį, kurio tapatybė nustatyta arba kurio tapatybę galima nustatyti (duomenų subjektas); fiziniu asmeniu, kurio tapatybę galima nustatyti, laikomas asmuo, kurio tapatybę tiesiogiai arba netiesiogiai galima nustatyti visų pirma pagal identifikatorių, kaip antai vardą ir pavardę, asmens identifikavimo numerį, buvimo vietos duomenis ir interneto identifikatorių, arba pagal vieną ar kelis to fizinio asmens fizinės, fiziologinės, genetinės, psichinės, ekonominės, kultūrinės ar socialinės tapatybės požymius.

5.2. **Atsitiktinis (statistinis) triukšmas** – nurodo nepastovumą imtyje, stochastinius sutrikimus regresijos lygtyje ar vertinimo paklaidas. Triukšmas dažniausiai apibrėžiamas kaip atsitiktinis kintamasis.

5.3. **Atskleidimas** – įvykis, kai asmuo, organizacija, įmonė iš paskelbtos statistinės informacijos atpažįsta ar sužino apie konkretų asmenį, organizaciją, įmonę ką nors, kas neturėtų būti sužinoma iš statistinės informacijos. Yra dvi atskleidimo rūšys – tapatumo atskleidimas ir požymio atskleidimas.

5.4. **Duomenų nuasmeninimas** – metodas, kuriuo siekiama panaikinti tapatybės atsekimo galimybę.

5.5. **Duomenų pseudoniminimas** – metodas, kurio metu tiesioginiai identifikatoriai pakeičiami dirbtinai sugeneruotomis reikšmėmis tokiu būdu, kad jų negalima susieti su konkrečiu asmeniu kitaip nei turint raktą, kuris saugiai saugomas atskirai nuo pseudonimintų duomenų.

5.6. **Įsibrovėlis** – duomenų ar informacijos naudotojas, ketinantis atskleisti konkretaus statistinio stebėjimo vieneto duomenis naudodamas paskelbtą statistinę informaciją ir kai kuriais atvejais turimas papildomas žinias.

5.7. **Kvaziidentifikatorius arba raktinis kintamasis** (angl. *quasi-identifikators or key variables*) – duomenų rinkinio kintamasis, kuris sujungtas su kitu (-ais) duomenų rinkinio kintamuoju (-aisiais) (arba kokia nors papildoma, išorine informacija) gali padėti atpažinti statistinį stebėjimo vienetą.

5.8. **Perturbuoti duomenys** – duomenys, kuriems pritaikyta duomenų saugumo technika, pridedanti „triukšmo“ duomenų rinkinyje, tam kad būtų užtikrintas kiekvieno įrašo konfidencialumas.

5.9. **Raktas** (angl. *key*) – vienas kvaziidentifikatorius ar kelių kombinacija.

5.10. **Specialių kategorijų asmens duomenys –** duomenys, atskleidžiantys rasinę ar etninę kilmę, politines pažiūras, religinius ar filosofinius įsitikinimus ar narystę profesinėse sąjungose, taip pat genetinius duomenis, biometrinius duomenis, siekiant konkrečiai nustatyti fizinio asmens tapatybę, sveikatos duomenis ir duomenis apie fizinio asmens lytinį gyvenimą ir lytinę orientaciją.

5.11. **Statistinio atskleidimo kontrolė** – priemonių, kuriomis siekiama sumažinti konfidencialių statistinių duomenų atskleidimo riziką, visuma.

6. Kitos Vadove neapibrėžtos sąvokos suprantamos taip, kaip jos yra apibrėžtos Vadovo III skyriuje išvardytuose teisės aktuose.

III SKYRIUS
REGLAMENTUOJANTYS TEISĖS AKTAI

7. Vadovą reglamentuojantys teisės aktai:

7.1. 2009 m. kovo 11 d. Europos Parlamento ir Tarybos reglamentas (EB) Nr. 223/2009 dėl Europos statistikos, panaikinantis Europos Parlamento ir Tarybos reglamentą (EB, Euratomas) Nr. 1101/2008 dėl konfidencialių statistinių duomenų perdavimo Europos Bendrijų statistikos tarnybai, Tarybos reglamentą (EB) Nr. 322/97 dėl Bendrijos statistikos, Tarybos sprendimą 89/382/EEB, Euratomas, įsteigiantį Europos Bendrijų statistikos programų komitetą, su visais pakeitimais;

7.2. 2016 m. balandžio 27 d. Europos Parlamento ir Tarybos reglamentas (ES) 2016/679 dėl fizinių asmenų apsaugos tvarkant asmens duomenis ir dėl laisvo tokių duomenų judėjimo ir kuriuo panaikinama Direktyva 95/46/EB (Bendrasis duomenų apsaugos reglamentas);

7.3. 2019 m. lapkričio 27 d. Europos Parlamento ir Tarybos reglamentas (ES) 2019/2152 dėl Europos verslo statistikos, kuriuo panaikinama 10 teisės aktų verslo statistikos srityje (toliau – Reglamentas dėl verslo statistikos);

7.4. 2022 m. gegužės 30 d. Europos Parlamento ir Tarybos reglamentas (ES) 2022/868 dėl Europos duomenų valdymo, kuriuo iš dalies keičiamas Reglamentas (ES) 2018/1724 (Duomenų valdymo aktas);

7.5. Lietuvos Respublikos oficialiosios statistikos ir valstybės duomenų valdysenos įstatymas;

7.6. Lietuvos Respublikos asmens duomenų teisinės apsaugos įstatymas.

7.7. Valstybės duomenų agentūros konfidencialumo politikos gairės, patvirtintos agentūros generalinio direktoriaus 2023 m. gegužės 4 d. įsakymu Nr. DĮ-111 „Dėl Valstybės duomenų agentūros konfidencialumo politikos gairių patvirtinimo“;

7.8. kiti teisės aktai, reglamentuojantys oficialiosios statistikos ir duomenų tvarkymą.

IV SKYRIUS
STATISTINIO ATSKLEIDIMO KONTROLĖS PROCESAS

8. Remiantis Bendruoju statistinės veiklos proceso modeliu (angl. *Generic Statistical Business Process Model* (toliau – GSBPM), statistinės informacijos rengimas yra skirstomas į aštuonis procesus (Vadovo 1 priedas).

9. Statistinės informacijos rengimo procesai, kuriuose atliekama statistinio atskleidimo kontrolė: „2. Planavimas“ subprocese „2.5. Duomenų apdorojimo ir analizės metodologijos parengimas“ ir „6. Analizė“ subprocese „6.3. Duomenų atskleidimo kontrolė“.

10. Statistinio atskleidimo kontrolės procesas susideda iš šių etapų:

10.1. statistinio atskleidimo kontrolės poreikio nustatymo;

10.2. duomenų ir kintamųjų analizės;
10.3. statistinio atskleidimo rizikos vertinimo;
10.4. statistinio atskleidimo kontrolės metodų parinkimo ir taikymo;
10.5. prarastos informacijos vertinimo;
10.6. pakartotino statistinio atskleidimo rizikos vertinimo;
10.7. sklaidos.

1 etapas. Statistinio atskleidimo kontrolės poreikio nustatymas

11. Norint nustatyti statistinio atskleidimo kontrolės poreikį, reikia nurodyti Vadovo 2 priedo anketoje statistinio atskleidimo poreikiui nustatyti skelbiamo / viešinamo duomenų rinkinio:

- aprašymą / pavadinimą;
- duomenų šaltinį (-ius);
- naudojimo tikslą: statistikos, administravimo, mokslo tikslams ir pan.;
- statistinio stebėjimo vienetus: gyventojai, privatūs namų ūkiai, juridiniai asmenys (įmonės), šeimos ūkiai ar ūkininkai, būstai, įvykiai ir pan.;
- detalumą: mikroduomenys, agreguoti duomenys ir kt.;
- duomenų rinkinio struktūrą: kintamųjų pavadinimai; kategorijų rinkiniai; klasifikatoriai; geoerdvinė komponentė ir kt.; reikėtų atkreipti dėmesį į tai, kad net jeigu nustatyti statistinio stebėjimo vienetai nėra fiziniai ar juridiniai asmenys, duomenyse gali būti informacijos apie fizinius ir juridinius asmenis. Pavyzdžiui, duomenų rinkinyje, kuriame namai yra pagrindiniai statistinio stebėjimo vienetai, gali būti informacijos apie šiuose namuose gyvenančius asmenis ar jų pajamas ir pan.;
- ar yra specialių kategorijų asmens duomenų ir kokie jie;
- ar kintamieji yra jautrūs, tiesiogiai atpažįstami ir pan.

12. Jeigu duomenų rinkinyje yra konfidencialių duomenų, specialiųjų kategorijų arba jautrių duomenų, pereinama į antrą statistinio atskleidimo kontrolės proceso etapą.

13. Jeigu duomenų rinkinyje nėra konfidencialių duomenų, specialiųjų kategorijų arba jautrių duomenų, pereinama į paskutinį – sklaidos – etapą.

2 etapas. Duomenų ir kintamųjų analizė

14. Antras statistinio atskleidimo kontrolės proceso etapas yra duomenų paruošimas.

14.1. Jeigu pirmame etape buvo nustatyta, kad duomenų rinkinyje yra jautrūs kintamieji ar tiesiogiai atpažįstami kintamieji, t. y. turintys tokių požymių, kaip vardas, asmens kodas, paso numeris, adresas, telefono numeris ar tam tikrais atvejais XY koordinatės, tokie kintamieji turėtų būti arba pašalinti, arba nuasmeninti, arba pseudoniminti ar pan. (pagal naudojimo tikslus).

14.2. Statistinio tyrimo metodikos ir statistiniai formuliarai padeda nustatyti su konfidencialumu susijusią informaciją: identifikatorius, kvaziidentifikatorius, konfidencialius kintamuosius.

14.3. Statistinio tyrimo duomenų šaltiniai ir juos reglamentuojantys teisės aktai padeda nustatyti atskleidimo galimybes ir statistinio atskleidimo kontrolės poreikį:

14.3.1. Jei duomenų rinkinį sudaro administraciniai duomenys ar bet kokio tipo registrai, svarbu žinoti šių duomenų šaltinių viešą prieinamumą.

14.3.2. Jei duomenų rinkinį sudaro ištisinio tyrimo (ar surašymo) mikroduomenys, svarbu užtikrinti duomenų konfidencialumą, taikant atskleidimo kontrolės metodus.

14.4. Jeigu objektas yra imčių tyrimų duomenys, tai informacija apie imties planą, vertinamus kintamuosius ir sritis, imties svorius gali padėti nustatyti su konfidencialumu susijusią informaciją. Pavyzdžiui, verslo statistikos mikroduomenyse išvardyti imties sluoksniai, susiję su dideliais verslo statistiniais stebėjimo vienetais, kelia didesnį atskleidimo rizikos pavojų.

3 etapas. Statistinio atskleidimo rizikos vertinimas

15. Trečiame statistinio atskleidimo kontrolės proceso etape atliekamas statistinio atskleidimo rizikos vertinimas: naudojamos rizikos vertinimo priemonės, padedančios nustatyti, ar duomenų rinkinyje užtikrinamas duomenų konfidencialumas, kad jį būtų galima viešinti. Daugiau informacijos apie statistinio atskleidimo rizikos vertinimą, pateikiama Vadovo V skyriuje „STATISTINIO ATSKLEIDIMO RIZIKOS VERTINIMAS“.

4 etapas. Statistinio atskleidimo kontrolės metodų parinkimas ir taikymas

16. Statistinio atskleidimo kontrolės metodų parinkimas priklauso nuo duomenų apsaugos poreikio (matuojamas pagal statistinio atskleidimo riziką, aprašytą Vadovo V skyriuje), duomenų struktūros ir kintamųjų tipo. Renkantis statistinio atskleidimo kontrolės metodus, reikėtų atsižvelgti ir į jų įtaką duomenims, t. y., ar duomenys, pritaikius pasirinktą metodą, atitiks vartotojų poreikius. Statistinio atskleidimo kontrolės metodų pasirinkimas iš dalies yra bandymų ir klaidų procesas: pritaikius pasirinktą metodą, matuojama atskleidimo rizika, duomenų naudingumas.

5 etapas. Prarastos informacijos vertinimas

17. Pritaikius pasirinktą statistinio atkleidimo kontrolės užtikrinimo metodą, yra vertinama prarastos informacijos apimtis koreguotame duomenų rinkinyje, t. y., kiek koreguotas duomenų rinkinys skiriasi nuo pradinio.

18. Informacijos praradimo matavimo priemonės ir metodai, kurie gali būti taikomi kategoriniams ir tolydiems kintamiesiems, yra aprašyti Vadovo VII skyriuje „DUOMENŲ PANAUDOJAMUMO IR INFORMACIJOS PRARADIMO MATAI“.

6 etapas. Pakartotinas statistinio atskleidimo rizikos vertinimas

19. Pritaikius statistinio atskleidimo kontrolės metodus, pakartotinai įvertinama konfidencialių duomenų atskleidimo rizika taikant pirmame rizikos vertinime pasirinktas rizikos vertinimo priemones. Jei rizika nėra priimtino lygio, kartojamas etapas, kuriame parenkami ir taikomi statistinio atskleidimo kontrolės metodai, taikant skirtingus metodus ir (arba) parametrus.

7 etapas. Sklaida

20. Paskutinis statistinio atskleidimo proceso etapas yra koreguotų duomenų skelbimas / viešinimas.

V SKYRIUS
STATISTINIO ATSKLEIDIMO RIZIKOS VERTINIMAS

PIRMASIS SKIRSNIS
DUOMENŲ ATSKLEIDIMO TIPAI

21. Galimi keli pagrindiniai duomenų atskleidimo tipai:

21.1. ***Tapatybės atskleidimas*** (angl. *identity disclosure*) įvyksta, jei įsibrovėlis susieja žinomą statistinį stebėjimo vienetą su paskelbtų duomenų įrašu. Dažniausiai tai įvyksta, kai įsibrovėlis susieja paskelbtų duomenų įrašą su turima papildoma, išorine informacija, kai paskelbtų duomenų įrašas išsiskiria kurių nors kintamųjų reikšmėmis, kai hiperkubo / gardelės / dažnių lentelės langelio duomenys priklauso vienam arba mažam statistinio stebėjimo vienetų skaičiui (angl. *small counts*), taip pat kai tikėtinas savęs atpažinimas (angl. *self-identification*).

21.2. ***Požymių atskleidimas*** (angl. *attribute disclosure*). Galima išskirti individualių požymių atskleidimą ir grupės požymių atskleidimą:

21.2.1. *Individualių požymių atskleidimas* įvyksta, kai kas nors, turintis tam tikros informacijos apie statistinio stebėjimo vienetą, gali, naudodamas lentelės duomenis ar duomenis iš kitos lentelės, turinčios bendrą požymį, atskleisti anksčiau nežinotas detales. Tai gali įvykti, jei lentelės eilutėje arba stulpelyje, kuriai priklauso langelis su maža reikšme, dominuoja nulinės

reikšmės. Tokiu atveju statistinio stebėjimo vienetas atpažįstamas kai kurių lentelės požymių pagrindu ir atskleidžiamas naujas požymis. Pavyzdys pateiktas *1 lentelėje*.

1 lentelė. Individualių požymių atskleidimo pavyzdys

Pajamos	Vyrai			Moterys			**Iš viso**
	Vedę	Išsituokę	Niekada negyveno santuokoje	Ištekėjusios	Išsituokusios	Niekada negyveno santuokoje	
Aukštos	30	0	2	14	2	0	**48**
Vidutinės	6	7	0	2	2	0	**17**
Mažos	2	0	1	1	0	1	**5**
Iš viso	**38**	**7**	**3**	**17**	**4**	**1**	**70**

Individualūs požymiai atskleidžiami, nes iš *1 lentelės* duomenyse esančių mažų reikšmių galima nustatyti, kad vienintelė niekada negyvenusi santuokoje moteris turi mažas pajamas.

21.2.2. *Grupės požymiams atpažinti* nereikalingas net tapatybės atpažinimas. Grupės požymių atpažinimo atveju galima nustatyti, kad konkrečiomis žinomomis charakteristikomis (amžiaus, lyties, gyvenamosios vietos ir pan.) pasižyminti statistinio stebėjimo vienetų grupė turi tam tikras kitų jautrių kintamųjų reikšmes. Pavyzdžiui, *1 lentelėje* matyti, kad visi išsituokę vyrai turi vidutines pajamas.

21.3. ***Egzistavimo atskleidimas*** (angl. *disclosure of existence*). Laikantis itin griežto supratimo apie požymių atskleidimą, galima išskirti atskirą atvejį, kai pati informacija apie tam tikromis charakteristikomis pasižyminčio statistinio stebėjimo vieneto ar jų grupės egzistavimą laikoma jautria. Pavyzdžiui, kai mažoje seniūnijoje yra pateiktas gyventojų skaičius pagal mažai skaitlias tautybes.

21.4. ***Atskleidimas naudojant skirtumą*** (angl. *disclosure by differencing*) įvyksta, kai paimamas kelių skirtingų lentelių (arba jų dalių) reikšmių skirtumas ir gautoje lentelėje, turinčioje mažas reikšmes, galima pritaikyti anksčiau aprašytus atvejus. Pavyzdžiui, *2 lentelė* ir *3 lentelė* parengtos naudojant tą patį duomenų rinkinį, tačiau skirtingą kintamojo reikšmių grupavimą:

2 lentelė. Gyventojų skaičius, kurių amžius nuo 18 metų

Amžius	<18	18–45	>45
Gyventojų skaičius	5	26	13

3 lentelė. Gyventojų skaičius, kurių amžius nuo 19 metų

Amžius	<19	19–45	>45
Gyventojų skaičius	6	25	13

Galima nesunkiai apskaičiuoti 18 metų gyventojų skaičių pagal gyventojų, priklausančių 18–45 ir 19–45 metų amžiaus grupėms, skaičių skirtumą.

Taip pat atskleidimas naudojant skirtumą dažnai galimas dėl naudojamų skirtingų geografinių kintamųjų, tokių kaip gardelės, NUTS klasifikatoriaus lygiai arba kiti teritoriniai vienetai.

Statistinio atskleidimo kontrolės metodas, nepakeičiantis reikšmių lentelėje, neapsaugos nuo duomenų atskleidimo naudojant skirtumą. Pavyzdžiui, nuo tokio duomenų atskleidimo tipo rizikos apsaugo apvalinimas, o reikšmių slėpimas neapsaugos. Šiuo atveju galimi du pasirinkimai – arba apriboti skelbiamos statistinės informacijos struktūrą, arba įvertinti visus potencialius skirtumus.

21.5. ***Atskleidimas naudojant sujungimą*** (angl. *disclosure by linking tables*). Kai skelbiama daug lentelių, suformuotų vieno statistinių duomenų rinkinio pagrindu, jos gali būti sujungtos naudojant bendrus požymius, taip didinant galimybes atskleisti naudojamus statistinio atskleidimo kontrolės metodus ir tikras nesaugių langelių reikšmes. Kai sujungtos lentelės suformuotos iš to paties

statistinių duomenų rinkinio, nepakanka nagrinėti kiekvienos lentelės apsaugos atskirai. Jei langelis reikalauja apsaugos vienoje lentelėje, jis turi būti apsaugotas visose lentelėse.

ANTRASIS SKIRSNIS
KINTAMŲJŲ KLASIFIKACIJOS

22. Duomenų rinkinio kintamieji turi skirtingą įtaką duomenų atskleidimui, priklausomai nuo to, kokio pobūdžio informacija saugoma tuose kintamuosiuose. Kintamieji gali būti skirstomi:

22.1. ***Identifikavimo kintamieji*** (angl. *identifying variables*) – kintamieji, turintys savyje informacijos, leidžiančios atpažinti statistinio stebėjimo vienetą. Jie dar gali būti skirstomi į dvi grupes:

22.1.1. *Tiesioginiai identifikatoriai* (angl. *direct identifiers*) tiesiogiai ir vienareikšmiškai leidžia identifikuoti statistinio stebėjimo vienetą. Tokie kintamieji dažniausiai būna vardas ir pavardė, asmens kodas, asmens dokumento numeris ir pan. Šie kintamieji visuomet turi būti pašalinti arba užšifruoti viešinamame duomenų rinkinyje. Tačiau vien tik tiesioginių identifikatorių pašalinimas iš duomenų rinkinio dažniausiai neužtikrina pakankamo saugumo.

22.1.2. *Kvaziidentifikatoriai* arba *raktiniai kintamieji* (toliau – kvaziidentifikatoriai). Kvaziidentifikatoriai patys savaime neleidžia atpažinti statistinio stebėjimo vieneto (pvz., vyras / moteris), tačiau kombinacijoje su kitais kintamaisiais gali turėti unikalią reikšmę duomenų rinkinyje (pvz., 18 metų, vyras, susituokęs). Nėra rekomenduojama tiesiog pašalinti kvaziidentifikatorių iš duomenų rinkinio, nes taip gali būti prarasta svarbi informacija. Praktiškai kiekvienas duomenų rinkinio kintamasis gali būti kvaziidentifikatorius.

22.2. ***Neleidžiantys identifikuoti kintamieji*** (angl. *non-identifying variables*) neturi jokios informacijos, leidžiančios atpažinti statistinį stebėjimo vienetą. Taip gali būti dėl to, kad šių kintamųjų nėra jokiame kitame duomenų rinkinyje ar išoriniame duomenų šaltinyje ir jie nėra prieinami įsibrovėliui. Šie kintamieji yra taip pat labai svarbūs statistinio atskleidimo kontrolės procese, nes gali saugoti konfidencialią ar jautrią informaciją.

23. Dar vienas kintamųjų klasifikavimas – skirstymas į **jautrius** (angl. *sensitive*) ir **nejautrius** (angl. *non-sensitive*):

23.1. Jautrių kintamųjų pavyzdžiai – pajamos, religija, politinės pažiūros, su sveikata susijusi informacija ar kita teisės aktais apibrėžta informacija.

23.2. Nejautrūs kintamieji saugo nekonfidencialius duomenis apie statistinio stebėjimo vienetą, pavyzdžiui, gyvenamoji vieta, kaimo / miesto vietovė ir pan. Ši kintamųjų kategorija taip pat yra svarbi statistinio atskleidimo kontrolės procese, nes nejautrūs kintamieji gali būti kvaziidentifikatoriai ir padėti atpažinti statistinio stebėjimo vienetą.

TREČIASIS SKIRSNIS
STATISTINIO ATSKLEIDIMO RIZIKOS VERTINIMAS MIKRODUOMENIMS

Mikroduomenų hierarchinė struktūra

24. Apklausų ir surašymų mikroduomenyse užtikrinamas duomenų konfidencialumas asmens ar vieneto (įmonės, mokyklos, sveikatos priežiūros įstaigos ir t. t.) lygiu. Mikroduomenų failai dažnai turi hierarchinę struktūrą, kai atskiri vienetai priklauso grupėms, pavyzdžiui, asmenys priklauso namų ūkiams. Dažniausia mikroduomenų hierarchinė struktūra yra namų ūkio struktūra namų ūkių apklausų duomenyse, todėl Vadove hierarchinės struktūros duomenų atskleidimo rizika kartais vadinama namų ūkio rizika (angl. *Household risk*). Tačiau šios sąvokos vienodai taikomos ir įmonių duomenims, ir kitiems hierarchinės struktūros duomenims, pavyzdžiui, mokyklos duomenims su mokiniais ir mokytojais arba įmonės duomenims su darbuotojais.

25. Turint hierarchinius duomenis, informacija, surinkta aukštesniu hierarchiniu lygmeniu (pvz., namų ūkio), yra vienoda visiems asmenims, priklausantiems tam aukštesniam hierarchiniam lygmeniui (pvz., namų ūkio pajamos ir būsto išlaidos). Ši hierarchinė struktūra sukuria papildomą informacijos atskleidimo rizikos lygį dėl dviejų priežasčių:

25.1. jei vienas namų ūkio asmuo identifikuojamas, tada namų ūkio struktūra leidžia identifikuoti kitus to paties namų ūkio narius;

25.2. kitų namų ūkio narių kintamųjų reikšmės, būdingos visiems namų ūkio nariams, gali būti panaudotos kitam to paties namų ūkio nariui identifikuoti.

Individuali rizika

26. Informacijos atskleidimo rizikos vertinimas pagrįstas teisingo statistinio stebėjimo vieneto identifikavimo tikimybės vertinimu duomenyse. Kuo retesnė imties kvaziidentifikatorių reikšmių kombinacija, tuo didesnė statistinio stebėjimo vieneto tapatybės atskleidimo rizika. Įsibrovėlis, bandantis susieti statistinio stebėjimo vienetą, kurio imties (duomenų rinkinio) raktų reikšmių derinys duomenyse yra gana retas, su kitu duomenų rinkiniu, kuriame yra tas pats raktas, turės didesnę tikimybę atskleisti tapatybę nei tada, kai tuo pačiu raktu dalijasi didesnis skaičius statistinio stebėjimo vienetų.

27. Pažymėkime duomenų rinkinyje esančių skirtingų raktų skaičių raide *K*. Pažymėkime f_k(*k = 1, ..., K*), – *k-tojo* rakto imties dažnį, t. y. asmenų, kurių kvaziidentifikatorių reikšmės sutampa su *k-tuoju* raktu, skaičių imtyje. Imties dažnis f_k yra toks pat kiekvienam įrašui, besidalijančiam *k-ąjį* raktą. Šių dažnių skaičiavimą iliustruoja toliau pateiktas pavyzdys.

4 lentelė. Duomenų rinkinio pavyzdys, parodantis imties dažnį, populiacijos dažnį

Nr.	Gyvenamoji vieta	Lytis	Išsilavinimas	Užimtumas	Svoris	f_k	F_k
1	Miestas	Moteris	Vidurinis	Dirbantis	180	2	360
2	Miestas	Moteris	Vidurinis	Dirbantis	180	2	360
3	Miestas	Moteris	Pradinis	Neaktyvus	215	1	215
4	Miestas	Vyras	Profesinis	Dirbantis	76	2	152
5	Kaimas	Moteris	Profesinis	Bedarbis	186	1	186
6	Miestas	Vyras	Profesinis	Dirbantis	76	2	152
7	Miestas	Moteris	Pagrindinis	Neaktyvus	180	1	180
8	Miestas	Vyras	Aukštasis	Bedarbis	215	1	215
9	Miestas	Moteris	Vidurinis	Neaktyvus	186	2	262
10	Miestas	Moteris	Vidurinis	Neaktyvus	76	2	262

28. *4 lentelėje* pateikiamos keturių kvaziindentifikatorių – „gyvenamoji vieta“, „lytis“, „išsilavinimas“, „užimtumas“ – kombinacijos dešimčiai statistinio stebėjimo vienetų. Raktų pavyzdžiai yra {„miestas“, „moteris“, „vidurinis“, „dirbantis“} ir {„miestas“, „moteris“, „pradinis“, „neaktyvus“}. Šiame pavyzdyje iš viso yra septyni unikalūs raktai, t. y. *K = 7*. Imties dažnių pavyzdžiai: *k = 1* raktui {„miestas “, „moteris“, „vidurinis“, „dirbantis“} $f_k = 2$, nes šį raktą turi 1-as ir 2-as asmenys. *k = 2* raktui {„miestas“, „moteris“, „pradinis“, „neaktyvus“} $f_k = 1$, nes jis yra vienintelis 3-iam asmeniui.

29. Kuo mažiau kitų statistinio stebėjimo vienetų, su kuriais analizuojamas statistinio stebėjimo vienetas dalijasi savo kvaziidentifikatorių deriniu, tuo didesnė tikimybė, kad statistinio stebėjimo vienetas bus teisingai identifikuotas kitame duomenų rinkinyje, kuriame taip pat yra šie kvaziidentifikatoriai. Net kai tiesioginiai identifikatoriai pašalinami iš duomenų rinkinio, tam statistinio stebėjimo vienetui yra didesnė atskleidimo rizika nei kitiems, darant prielaidą, kad jų imties svoriai yra vienodi. *4 lentelėje* pateikiami *k-tojo* rakto imties dažniai f_k visiems asmenims. Jei $f_k = 1$, tai šis asmuo turi unikalų kvaziidentifikatorių derinį ir vadinamas unikaliu deriniu imtyje (angl. *sample unique*). Duomenų rinkinys 4 lentelėje turi keturis unikalius derinius imtyje.

30. Imties duomenims aktualus *k-tojo* rakto populiacijos dažnis F_k, kuris yra statistinio stebėjimo vienetų skaičius populiacijoje su *k-tuoju* raktu. Populiacijos dažnis nežinomas, jei mikroduomenys yra imties, o ne surašymo. Remiantis tam tikromis prielaidomis, populiacijos dažnio reikšmę galima apskaičiuoti naudojant imties plano svorį w_i kiekvienam imties vienetui *i:*

$$F_k = \sum_{i=1}^{n} w_i^k,$$

čia:
n – imties dydis;
F_k– visų imties vienetų, turinčių tą patį raktą *k*, imties plano svorių suma;

$$w_i^k = \begin{cases} w_i, jeigu\ i\text{-}tasis\ imties\ vienetas\ turi\ k\text{-}tąjį\ raktą \\ 0\ \ kitu\ atveju. \end{cases}$$

31. Populiacijos dažnis F_k yra toks pat visiems įrašams, turintiems raktą *k*. Teisingo identifikavimo rizika yra tikimybė, kad raktas bus susietas su teisingu populiacijos vienetu. Kadangi kiekvienas imties vienetas, turintis raktą *k*, atitinka populiacijos vienetų skaičių F_k, tai teisingo identifikavimo tikimybė yra $1/F\_k$ ir gali būti interpretuojama kaip informacijos atskleidimo rizika. Asmenų, turinčių tą patį raktą, dažniai yra vienodi.

Jei $F_k = 1$, tai raktas *k* yra unikalus derinys imtyje ir populiacijoje, o atskleidimo rizika yra 1. Raktų unikalumas populiacijoje (t. y. raktai, turintys populiacijos dažnį $F_k = 1$) yra svarbus veiksnys, į kurį reikia atsižvelgti vertinant riziką.

32. Individualią atskleidimo riziką labiausiai lemia imties dažniai f_k ir imties plano svoriai w_i (populiacijos dažniai F_k). Tačiau atskleidimo rizikos nustatymas imties duomenims remiantis vien tik dažniais yra laikomas gana konservatyviu, kadangi tokiu atveju nėra tinkamai atsižvelgiama į pritaikytus imčių teorijos metodus ir atskleidimo rizika gali būti pervertinta. Dėl to duomenims gali būti pritaikyti pertekliniai atskleidimo kontrolės metodai. Literatūroje (Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., et al., 2012) yra aprašytas individualios rizikos $r_k$vertinimas *k-tajam* raktui, pritaikant Bajeso statistikos teoriją (angl. *Bayesian approach*) ir panaudojant aposteriorinį F_k skirstinį bei taikant kitus metodus. Vienas iš tokių individualios rizikos r_k vertinimo metodų yra realizuotas R pakete *sdcMicro*.

33. Rizikos matas *j vienetų anonimiškumas* (angl. *j-anonymity*) grindžiamas principu, kad saugiame duomenų rinkinyje statistinio stebėjimo vienetų, turinčių tą patį kategorinių kvaziidentifikatorių reikšmių derinį, skaičius turėtų būti didesnis arba lygus nurodytam slenkščiui *j* ($f_k \geq j, k = 1, \dots, K$). Statistinio stebėjimo vienetas pažeidžia *j* vienetų anonimiškumą, jei kurio nors rakto *k* imties dažnis f_k yra mažesnis už nurodytą slenkstį j ($\exists k = 1, \dots K: f_k < j$). Pavyzdžiui, jei asmuo turi tą patį *k-tąjį* kvaziidentifikatorių derinį kaip ir kiti du imties asmenys (t. y. $f_k = 3$), tai šie asmenys patenkina 3 vienetų anonimiškumą (nes $f_k \geq 3$), bet pažeidžia 4 vienetų anonimiškumą (nes $f_k < 4$).

34. Grįžkime prie *4 lentelės* pavyzdžio. Čia šeši asmenys tenkina 2 vienetų anonimiškumą ir keturi asmenys pažeidžia 2 vienetų anonimiškumą. Asmenys, pažeidžiantys 2 vienetų anonimiškumą, yra unikalūs duomenų rinkinyje kvaziidentifikatorių reikšmių atžvilgiu. Tai reiškia, kad visoms galimoms *K* raktų reikšmėms dažnis $f_k, k = 1, \dots, K,$ lygus vienetui.

35. Rizikos matas *j vienetų anonimiškumas* kiekybiškai gali būti užrašytas kaip skaičius įrašų, pažeidžiančių *j* vienetų anonimiškumą tam tikrai *k-tajai* rakto reikšmei:

$$R_k^j = \sum_{i=1}^{n} I(f_k < j),$$

čia:
I yra indikatorinė funkcija;
n – duomenų rinkinio įrašų skaičius;

$R_k^j - j$ vienetų anonimiškumo rizikos matas *k-tajam* raktui.

36. Parenkant reikiamą *j* vienetų anonimiškumo slenkstį *j* yra svarbu atsižvelgti į imties plano svorius. Kuo didesnis imties plano svoris, tuo vienas statistinio stebėjimo vienetas duomenų rinkinyje reprezentuoja daugiau populiacijos vienetų, dėl to teisingo identifikavimo tikimybė yra mažesnė ir slenkstis gali būti mažesnis.

37. *j* vienetų anonimiškumas dažnai yra būtinas, bet nepakankamas reikalavimas nuasmeninant duomenų rinkinį prieš viešinimą. Dažnai *j* vienetų anonimiškumo slenkstis yra parenkamas tarp 3 ir 5. Jį pasiekti galima naudojant statistinio atskleidimo kontrolės metodus, aprašytus Vadovo VI skyriuje.

38. Atvejis, kai *j* vienetų anonimiškumo tenkinimas yra nepakankama sąlyga norint užtikrinti duomenų rinkinio konfidencialumą, gali būti iliustruotas nagrinėjant ankstesnį pavyzdį, kai prie turimų duomenų pridedamas papildomas neidentifikuojantis, tačiau jautrus kintamas „sveikata“ (*5 lentelė*). Pirmieji du asmenys tenkina 2 vienetų anonimiškumą pagal raktinius kintamuosius „gyvenamoji vieta“, „lytis“, „išsilavinimas“ ir „užimtumas“. Vis dėlto, jei įsibrovėlis žino, kad tam tikras asmuo priklauso imčiai ir turi atitinkamą rakto {„miestas“, „moteris“, „vidurinis“, „dirbantis“} reikšmę, tuomet yra atskleidžiama šio asmens sveikatos būklė („taip“), nes abu asmenys su šiuo raktu turi tą pačią sveikatos kintamojo reikšmę. Atskleidimas negalimas asmenims, turintiems raktą {„miestas“, „vyras“, „profesinis“, „dirbantis“}, nes ketvirto ir šešto asmenų kintamasis „sveikata“ turi skirtingas reikšmes („taip“ ir „ne“).

5 lentelė. l reikšmių skirtingumo kriterijaus pavyzdys

Nr.	Gyvenamoji vieta	Lytis	Išsilavinimas	Užimtumas	Sveikata	Svoris	f_k	F_k	*l* reikšmių skirtingumas
1	Miestas	Moteris	Vidurinis	Dirbantis	Taip	180	2	360	1
2	Miestas	Moteris	Vidurinis	Dirbantis	Taip	180	2	360	1
3	Miestas	Moteris	Pradinis	Neaktyvus	Taip	215	1	215	1
4	Miestas	Vyras	Profesinis	Dirbantis	Taip	76	2	152	2
5	Kaimas	Moteris	Profesinis	Bedarbis	Taip	186	1	186	1
6	Miestas	Vyras	Profesinis	Dirbantis	Ne	76	2	152	2
7	Miestas	Moteris	Pagrindinis	Neaktyvus	Ne	180	1	180	1
8	Miestas	Vyras	Aukštasis	Bedarbis	Taip	215	1	215	1
9	Miestas	Moteris	Vidurinis	Neaktyvus	Ne	186	2	262	2
10	Miestas	Moteris	Vidurinis	Neaktyvus	Taip	76	2	262	2

39. Sakoma, kad duomenų rinkinys tenkina *l* reikšmių skirtingumo kriterijų (angl. *l-diversity*), jei kiekvienam raktui *k* yra bent *l* skirtingų kiekvieno jautraus kintamojo reikšmių. *5 lentelėje* pirmieji du asmenys tenkina tik 1 reikšmės skirtingumo, 4 ir 6 asmenys – 2 reikšmių skirtingumo kriterijų. Reikalingas *l* reikšmių skirtingumo lygis *l* priklauso nuo reikšmių, kurias gali įgyti jautrus kintamasis, skaičiaus. Jei jautrus kintamasis gali įgyti tik dvi reikšmes, tai aukščiausias *l* lygis yra 2. *k-tojo* rakto atžvilgiu unikalus imtyje statistinio stebėjimo vienetas visada tenkins tik 1 reikšmės skirtingumo kriterijų.

40. *l* reikšmių skirtingumo kriterijus yra naudingas, jei duomenų rinkinyje yra kategorinių jautrių kintamųjų, kurie nėra kvaziidentifikatoriai.

Namų ūkio rizika

41. Daugelio socialinių tyrimų duomenys turi hierarchinę struktūrą, kai asmuo priklauso aukštesnio lygio subjektui (pvz., namų ūkiui). Identifikavus vieną namų ūkio narį, galima

identifikuoti ir kitus to namų ūkio narius. Jei atsižvelgsime į namų ūkio struktūrą, asmens identifikavimo rizika r^{hh}, kad bent vienas iš namų ūkio narių bus identifikuotas:

$$r^{hh} = P\left(A_1 \cup A_2 \cup \ldots \cup A_J\right) = 1 - \prod_{j=1}^{J}\left(1 - P\left(A_j\right)\right) = 1 - \prod_{j=1}^{J}\left(1 - r_j'\right),$$

čia:

$A_j, j = 1, \ldots, J$ – įvykis, kad *j*-tasis namų ūkio narys bus identifikuotas;

P(A_j) – tikimybė, kad *j*-tasis namų ūkio narys bus identifikuotas;

r_j' – *j*-tojo namų ūkio nario individuali atskleidimo rizika.

42. Pavyzdžiui, jei namų ūkis turi tris narius su individualiomis atskleidimo rizikomis 0,02, 0,03 ir 0,03, tada namų ūkio rizika yra $1 - (1 - 0{,}02)(1 - 0{,}03)(1 - 0{,}03) = 0{,}078$. Hierarchinė ar namų ūkio rizika negali būti mažesnė už individualią riziką, o namų ūkio rizika visada yra vienoda visiems namų ūkio nariams. Namų ūkio rizika turėtų būti naudojama tais atvejais, kai duomenyse yra hierarchinė struktūra, t. y. kai duomenyse yra pateikta namų ūkio struktūra.

Bendroji rizika

43. Turint individualias arba namų ūkio rizikas, visam duomenų rinkiniui yra skaičiuojama bendroji atskleidimo rizika (angl. *global risk*). Reikia atkreipti dėmesį, kad net ir esant priimtinai bendrajai rizikai gali būti atvejų, kai labai didelės rizikos įrašus kompensuoja keli mažos rizikos įrašai:

43.1. Bendroji duomenų rinkinio rizika yra visų duomenų rinkinio statistinio stebėjimo vienetų rizikų vidurkis:

$$R^g = \frac{1}{n}\sum_{i=1}^{n} r_i' = \frac{1}{n}\sum_{k=1}^{K} f_k r_k,$$

čia:

R^g – bendroji rizika duomenų rinkiniui;

n – imties (duomenų rinkinio) dydis;

K – galimų rakto reikšmių skaičius;

r_i' – *i*-tojo asmens imtyje (duomenų rinkinyje) individuali atskleidimo rizika;

r_k – statistinio stebėjimo vieneto, turinčio *k*-tąjį raktą, atskleidimo rizika;

f_k – *k-tojo* rakto imties dažnis.

Bendroji rizika kinta intervale [0; 1], kuo rizika arčiau 1, tuo tikimybė identifikuoti statistinio stebėjimo vienetą yra didesnė, o kuo arčiau 0 – mažesnė. Bendroji rizika parodo, kokią dalį duomenų rinkinio vienetų įsibrovėlis galėtų identifikuoti. Pavyzdžiui, bendroji rizika $R^g = 0{,}015$ parodo, kad tikėtina identifikuoti 1,5 proc. duomenų rinkinio vienetų.

43.2. Bendrąją riziką taip galima išreikšti kaip galimų identifikuoti duomenų rinkinio statistinio stebėjimo vienetų skaičių, padauginant dydį R^g iš duomenų rinkinio dydžio *n*. Pavyzdžiui, jei turime 500 asmenų duomenų rinkinį (imtį) ir $R^g = 0{,}015$, tai reiškia, kad tikėtina identifikuoti vidutiniškai $0{,}015 \cdot 500 = 7{,}5$ statistinio stebėjimo vienetus šiame duomenų rinkinyje.

KETVIRTASIS SKIRSNIS
STATISTINIO ATSKLEIDIMO RIZIKOS VERTINIMAS AGREGUOTIEMS DUOMENIMS

Nesaugių langelių nustatymo taisyklės

44. Skelbiant agreguotų reikšmių lentelę, įvertinama kiekvieno jos langelio atskleidimo rizika ir nustatomi nesaugūs langeliai. Nesaugiems langeliams nustatyti agreguotų reikšmių lentelėse dažniausiai naudojamos šios taisyklės:

44.1. **Mažiausio dažnio taisyklė.** Pagal šią taisyklę langelio reikšmė laikoma nesaugia, jeigu langelyje atsispindi mažiau kaip *n* statistinio stebėjimo vienetų duomenys. Gali būti taikoma, kai pakanka užkirsti kelią tiksliam statistinio stebėjimo vieneto duomenų atskleidimui. Verslo statistikoje dažniausiai naudojama mažiausio dažnio taisyklė su parametru *n = 3*. Kitose statistikos srityse naudojamas mažiausio dažnio parametras gali būti įvairus, atsižvelgiant į informacijos jautrumą ir kitas savybes.

44.2. **Koncentracijos taisyklės.** Jos gali būti taikomos tais atvejais, kai yra manoma, kad įsibrovėlis gali identifikuoti statistinio stebėjimo vienetus su didžiausiomis dalimis langelyje. Pavyzdžiui, verslo statistinių duomenų atveju, paprastai laikoma, kad įsibrovėliai gali būti statistinio stebėjimo vieneto konkurentai arba trečios šalys, kurios yra gerai informuotos apie situaciją ekonomikos srityje, kuriai priklauso atitinkamas langelis. Šiai taisyklių klasei priskiriamos šios taisyklės:

44.2.1. ***(n; k) dominavimo taisyklė.*** Pagal šią taisyklę langelio reikšmė laikoma nesaugia, jeigu suminė *n* didžiausių statistinio stebėjimo vienetų dalis viršija *k %* langelio reikšmės, t. y.

$$\frac{x_1 + \cdots + x_n}{t_x} \cdot 100 > k,$$

čia:

$x_1 \geq x_2 \geq \ldots \geq x_N$ – tam tikro požymio surūšiuotos statistinio stebėjimo vienetų dalys langelyje ($n < N$), kurio reikšmė t_x

t_x– langelio reikšmė, gaunama:

$$t_x = \sum_{i \leq N} x_i.$$

Taisyklė taikoma, kai tam tikram rodikliui keliami itin griežti konfidencialumo reikalavimai. Dažnai naudojama $(n; k)$ dominavimo taisyklė su parametrais $(2; 85)$ ir $(1; 70)$

44.2.2. ***p % taisyklė.*** Pagal šią taisyklę langelio reikšmė laikoma nesaugia, jeigu iš jos atėmus dviejų didžiausių statistinio stebėjimo vienetų dalis, likusi suma yra mažesnė už didžiausios dalies *p %*, t. y.

$$\frac{\hat{t}_x - (x_1 + x_2)}{x_1} \cdot 100 < p.$$

Siekiant pakeisti *(n; k)* dominavimo taisyklę į *p %* taisyklę, parametras *p* gali būti išskaičiuotas iš sąryšio

$$p = 100 \cdot \frac{100 - k}{k}.$$

Naudojama $(2; 85)$ dominavimo taisyklė gali būti pakeista *p %* taisykle su parametru $p = 17{,}65$.

45. $(n; k)$ dominavimo taisyklės ir *p %* taisyklės parametrai turi būti laikomi konfidencialiais.

46. **Tarptautinės prekybos prekėmis statistikoje taikomas pasyvaus konfidencialumo principas.** Tarptautinės prekybos prekėmis statistinė informacija yra labai detali (Kombinuotoje nomenklatūroje yra daugiau nei 9 000 prekių kodų), todėl įprastų konfidencialumo taisyklių taikymas drastiškai sumažintų statistinės informacijos apimtį. Dėl šios priežasties, kitaip nei kitose statistikos srityse, tarptautinės prekybos prekėmis statistikoje yra taikomas pasyvaus konfidencialumo principas, apibrėžtas Reglamento dėl verslo statistikos 19 straipsnyje. Pagal pasyvaus konfidencialumo principą, tik jei to paprašo duomenis pateikęs statistinio stebėjimo vienetas, oficialiąją statistiką tvarkanti institucija turi nuspręsti ar statistiniai rezultatai, pagal kuriuos galima nustatyti minėtą statistinio stebėjimo vienetą, turi būti skleidžiami arba keičiami tokiu būdu, kad jų sklaida nepakenktų statistinių duomenų konfidencialumui. Paprastai sprendimas grindžiamas bendromis nustatytomis konfidencialumo taisyklėmis. Duomenų teikėjui nepateikus prašymo, konfidencialių duomenų atskleidimo galimybė nėra tikrinama ir statistinė informacija yra skleidžiama. Rengiant tarptautinės

prekybos prekėmis statistiką pagal verslo charakteristikas taikomas aktyvaus konfidencialumo principas, tai yra įprastos nustatytos konfidencialumo taisyklės.

Antrinis konfidencialumas

47. Antrinis konfidencialumas yra nekonfidencialių duomenų laikymas konfidencialiais, siekiant išvengti konfidencialių duomenų atskleidimo, nepaliekant įsibrovėliui galimybės juos nustatyti.

48. Tarkime, *6 lentelėje* turime gyventojų skaičių vietovėse pagal lytį. Vietovėje B yra tik vienas vyras todėl ši statistinė informacija gali atskleisti konfidencialius duomenis ir nėra skelbiama. Tačiau paslėpus šią informaciją ir nepritaikius antrinio konfidencialumo ji galėtų būti lengvai išskaičiuojama ($6 - 5 = 1$ arba $171 - 150 - 20 = 1$).

6 lentelė. Gyventojų skaičius vietovėse pagal lytį

Vietovės pavadinimas	Lytis		
	Iš viso	Moterys	Vyrai
Iš viso	**32**	**155**	**171**
A	**250**	100	150
B	**6**	5	1
C	**70**	50	20

Todėl, norint paslėpti šiuos konfidencialius duomenis, būtina paslėpti ne tik konfidencialų langelį, bet taip pat dar mažiausiai 3 langelius taip, kaip pavaizduota *7 lentelėje*.

7 lentelė. Gyventojų skaičius vietovėse pagal lytį (pritaikius pirminį ir antrinį konfidencialumą)

Vietovės pavadinimas	Lytis		
	Iš viso	Moterys	Vyrai
Iš viso	**326**	**155**	**171**
A	**250**	100	150
B	**6**	X	X
C	**70**	X	X

49. Taikant antrinį konfidencialumą, svarbu atkreipti dėmesį į informacijos praradimą. Konkrečiu atveju būtina paslėpti B miesto moterų skaičių, tačiau galima pasirinkti kurio miesto vyrus ir moteris norime paslėpti: C arba A. *8 lentelėje* pateiktas pavyzdys, kaip galima sugrupuoti konfidencialius duomenis, kad būtų išvengta tiek pirminio, tiek antrinio konfidencialumo. Šiuo atveju svarbu priimti sprendimą, kuri informacija yra svarbesnė: bendras gyventojų skaičius pagal vietoves ar gyventojų skaičius pagal lytį.

8 lentelė. Gyventojų skaičius vietovėse pagal lytį (pritaikius pirminį ir antrinį konfidencialumą)

Vietovės pavadinimas	Lytis		
	Iš viso	Moterys	Vyrai
Iš viso	**326**	**155**	**171**
A	**250**	100	150
B ir C	**76**	55	21

50. Taikant antrinį konfidencialumą svarbus suderinamumas:

50.1. *laiko atžvilgiu* – paėmus skirtingų laikotarpių to paties tyrimo statistinę informaciją neturėtų likti galimybės išskaičiuoti konfidencialių duomenų. Pavyzdžiui, kai skelbiamas gyventojų skaičiaus augimas. Tarkime, B mieste gyventojų skaičiaus augimas, palyginti su praėjusiu laikotarpiu, sudaro 300 proc. (dabar 6 gyventojai, o praėjusį laikotarpį buvo 2 gyventojai). Praėjusį laikotarpį ši informacija nebuvo skelbiama, nes neatitiko konfidencialumo reikalavimų, todėl svarbu

B miesto bendro gyventojų skaičiaus neskelbti ir ataskaitiniu laikotarpiu pritaikant šiam langeliui antrinį konfidencialumą.

50.2. *su kitomis to paties tyrimo lentelėmis* – ta pati informacija gali būti skelbiama skirtingose lentelėse, todėl svarbu, pritaikius antrinį konfidencialumą vienoje lentelėje, taip pat paslėpti langelius su ta pačia informacija kitose lentelėse;

50.3. *kitų tyrimų paskelbta analogiška informacija* – kai antrinis konfidencialumas yra pritaikomas vieno tyrimo statistinei informacijai, yra svarbu įsitikinti, kad analogiška informacija nėra prieinama skelbiant kito tyrimo statistinę informaciją. Pavyzdžiui, jei pritaikytas antrinis konfidencialumas taip, kaip pateikta *7 lentelėje* arba *8 lentelėje*, bet skelbiant kito tyrimo rezultatus yra paskelbiamas, tarkime, C miesto vyrų ir moterų skaičius, tada konfidencialius duomenis galima išskaičiuoti.

VI SKYRIUS
STATISTINIO ATSKLEIDIMO KONTROLĖS METODAI

PIRMASIS SKIRSNIS
STATISTINIO ATSKLEIDIMO KONTROLĖS METODŲ KLASIFIKAVIMAS

51. Statistinio atskleidimo kontrolės metodus (taikomus tolydiems ir diskretiems kintamiesiems) galima suskirstyti į dvi kategorijas:

51.1. Nekeičiantys duomenų rinkinio reikšmių metodai (angl. *non-perturbative methods*). Tokie metodai sumažina duomenų rinkinio detalumą, paslepiant (užmaskuojant) kai kurias duomenų rinkinio reikšmes, keičiant grupavimo ar klasifikavimo reikšmes, bet nekeičiant pačių duomenų reikšmių. Šie metodai padeda sumažinti atskleidimo riziką, tačiau kartu gali vesti prie didelio informacijos kiekio praradimo, taip pat gali pakeisti pačią duomenų struktūrą. Nekeičiantys duomenų reikšmių metodai gali būti nuosekliai taikomi kelioms lentelėms / duomenų rinkiniams, taip išlaikant jų adityvumą, tačiau labai sunku išlaikyti vientisumą tarp daugelio tarpusavyje susietų lentelių ar duomenų rinkinių (pvz., tarp skirtingų Europos Sąjungos šalių duomenų rinkinių, iš to paties duomenų rinkinio parengtų skirtingų lentelių).

51.2. Duomenų rinkinio reikšmes keičiantys metodai (angl. *perturbative methods*). Šie metodai nežymiai pakeičia pačias duomenų rinkinio reikšmes. Tokie metodai dažniausiai nulemia mažesnį duomenų praradimą, dėl to vartotojui duomenų rinkinys tampa naudingesnis. Dauguma duomenų rinkinio reikšmes keičiančių metodų nekeičia duomenų rinkinio struktūros, todėl išlaikomas geresnis palyginamumas tarp skirtingų duomenų rinkinių. Pagrindinis šių metodų reikalavimas yra, kad duomenų pakeitimas nedarytų didelės įtakos pačių duomenų kokybei ir reikšmingai jų neiškreiptų. Duomenų rinkinio reikšmes keičiantys metodai gali būti taikomi mikroduomenims (angl. *pre-tabular perturbation*) arba agreguotiems duomenims (angl. *post-tabular perturbation*). Mikroduomenims taikomi metodai pakeičia mikroduomenis taip, kad iš šių modifikuotų mikroduomenų sudarytos lentelės būtų laikomos pakankamai saugiomis. Agreguotiems duomenims taikomi metodai keičia agreguotų duomenų reikšmes nekeičiant mikroduomenų rinkinio reikšmių.

9 lentelė. Statistinio atskleidimo metodų palyginimas

Nekeičiantys duomenų rinkinio reikšmių metodai	Duomenų rinkinio reikšmes keičiantys metodai
Duomenys ne pakeičiami, bet perkoduojami, paslepiami ir pan.	Duomenys yra nežymiai pakeičiami.
Galimas didelis informacijos praradimas.	Informacijos praradimas kontroliuojamas parametrais.
Gali pasikeisti duomenų rinkinio struktūra.	Duomenų rinkinio struktūra yra išsaugoma.
Nuoseklumas gali būti išsaugomas, bet kai turima daug tarpusavyje susietų ar iš to paties	Nuoseklumas ir (ar) adityvumas gali būti prarastas.

duomenų rinkinio parengtų lentelių, nuoseklumą ir adityvumą išsaugoti sunku.	
Galima išsaugoti duomenų rinkinių adityvumą.	Metodai lankstesni nei nekeičiantys duomenų rinkinio reikšmių metodai.
Atsiranda trūkstamos reikšmės laiko eilutėje.	Poveikis laiko eilutėms nereikšmingas.

52. Kitas galimas statistinio atskleidimo metodų klasifikavimas – į tikimybinius (angl. *probabilistic*) ir deterministinius (angl. *deterministic*):

52.1. Tikimybiniai metodai priklauso nuo tikimybėmis arba atsitiktinių skaičių generavimu paremtų mechanizmų. Kiekvieną kartą taikant tikimybinį metodą gaunamas skirtingas rezultatas. Dėl šios priežasties rekomenduojama naudoti fiksuotą atsitiktiniams skaičiams generuoti naudojamą pradinę reikšmę (angl. *seed*), kad būtų galima gauti pakartojamus rezultatus.

52.2. Deterministiniai metodai pagrįsti tam tikru žinomu algoritmu ir leidžia gauti vienodus rezultatus, jeigu metodas pritaikomas tam pačiam duomenų rinkiniui su tais pačiais parametrais.

ANTRASIS SKIRSNIS
NEKEIČIANTYS DUOMENŲ RINKINIO REIKŠMIŲ METODAI

Globalus perkodavimas

53. Globalus perkodavimas yra deterministinis, nekeičiantis duomenų rinkinio reikšmių metodas. Taikant globalųjį perkodavimą, kategoriniam kintamajam kelios kategorijos yra sujungiamos į vieną, o tolydaus kintamojo reikšmės suskirstomos į intervalus.

54. *10 lentelėje* pateiktas pavyzdys, kaip pasikeičia dažniai kiekvienai kvaziidentifikatorių reikšmių kombinacijai, pritaikius globalųjį perkodavimą. Čia stulpelis f_k nurodo dažnį kiekvienai kvaziidentifikatorių *Regionas*, *Lytis* ir *Religija* kombinacijai.

10 lentelė. Dažnių pasikeitimai

ID	Prieš perkodavimą				Po perkodavimo			
	Regionas	**Lytis**	**Religija**	f_k	**Regionas**	**Lytis**	**Religija**	f_k
1	Regionas 1	Moteris	Katalikas	1	Šiaurės	Moteris	Katalikas	3
2	Regionas 2	Moteris	Katalikas	2	Šiaurės	Moteris	Katalikas	3
3	Regionas 2	Moteris	Katalikas	2	Šiaurės	Moteris	Katalikas	3
4	Regionas 3	Moteris	Protestantas	2	Centrinis	Moteris	Protestantas	2
5	Regionas 3	Vyras	Protestantas	1	Centrinis	Vyras	Protestantas	2
6	Regionas 3	Moteris	Protestantas	2	Centrinis	Moteris	Protestantas	2
7	Regionas 3	Vyras	Protestantas	2	Centrinis	Vyras	Protestantas	2
8	Regionas 4	Vyras	Musulmonas	2	Pietų	Vyras	Musulmonas	3
9	Regionas 4	Vyras	Musulmonas	2	Pietų	Vyras	Musulmonas	3
10	Regionas 5	Vyras	Musulmonas	1	Pietų	Vyras	Musulmonas	3

10 lentelėje matyti, kad prieš perkodavimą yra 3 statistinio stebėjimo vienetai, kurių raktinių kintamųjų reikšmės yra unikalios. Sugrupavus atskirus regionus į tris stambesnes kategorijas – *Šiaurės, Centrinis* ir *Pietų*, randame mažiausiai 2 statistinio stebėjimo vienetus kiekvienai raktinių kintamųjų reikšmių kombinacijai.

55. Taikant globalųjį perkodavimą reikia nuspręsti, kokio dydžio turėtų būti naujosios grupės ir kokios reikšmės turėtų būti sujungtos į vieną grupę. Naujos grupės turėtų būti parinktos atsižvelgiant į vartotojų poreikius ir minimizuojant perkodavimo metu prarandamos informacijos kiekį.

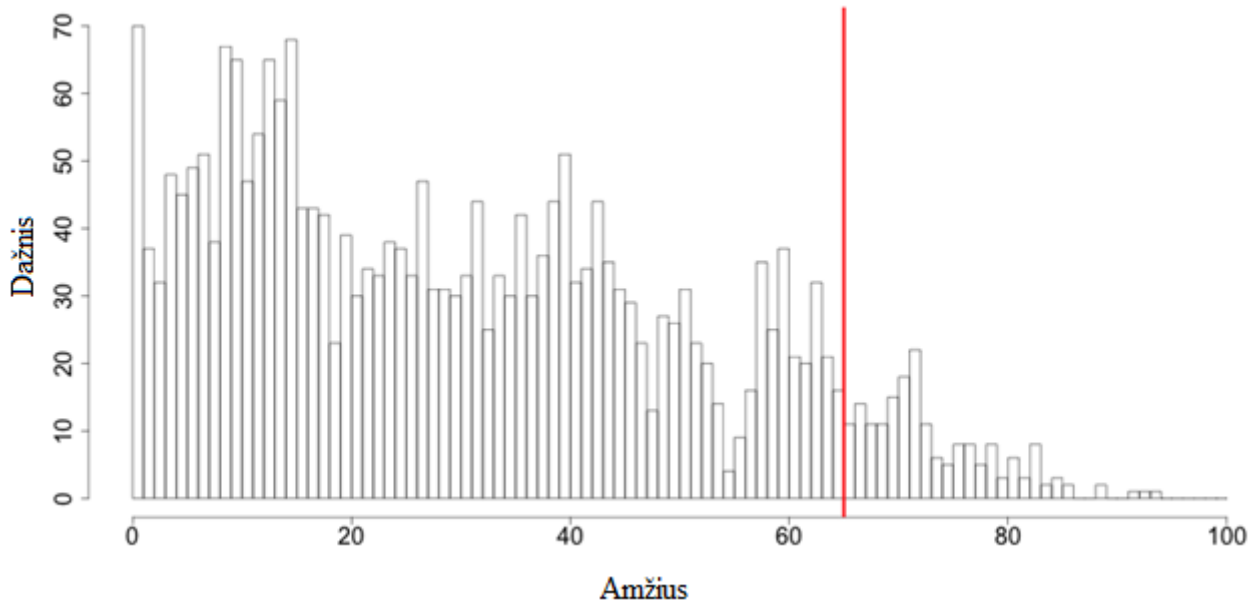
Viršaus ir apačios perkodavimas

56. Šis metodas yra panašus į globalųjį perkodavimą, tačiau vietoje visų reikšmių perkodavimo, jis atliekamas tik viršutinėms ir (arba) apatinėms skirstinio arba kategorinio kintamojo kategorijų reikšmėms. Viršaus ir apačios perkodavimo metodas gali būti taikomas tik tolydiems dydžiams arba kategoriniams kintamiesiems, kuriuos galima suranguoti. Metodas tinkamiausias tokiems duomenims, kurių didžioji dalis susitelkusi skirstinio viduryje su nedideliu kiekiu išsiskiriančių reikšmių. Pavyzdžiui, tokie kintamieji yra amžius ir pajamos: dažniausiai tik kelios reikšmės viršija tam tikrą ribą. Kuo mažiau stebinių kategorijoje, tuo didesnė atskleidimo rizika. Todėl viena iš išeičių tokiu atveju yra sugrupuoti visas reikšmes skirstinio galuose į vieną kategoriją. Taip yra sumažinama atskleidimo rizika, tuo pačiu metu nedarant jokios įtakos kitoms skirstinio reikšmėms.

57. Norint priimti sprendimą, kokią ribą pasirinkti ir kurias reikšmes pergrupuoti, reikėtų:

57.1. peržiūrėti visą skirstinį ir identifikuoti, kuriame taške dažniai tampa mažesni už nustatytą stebinių skaičių, surasti išskirtis. Skirstinyje (*1 pav.*) raudona vertikali linija nurodo siūlomą ribą amžiaus kintamojo viršutinėms reikšmėms perkoduoti;

57.2. išanalizuoti norimų skelbti duomenų panaudojimo tikslą. Pavyzdžiui, jeigu duomenys skirti analizuoti 15–64 metų amžiaus žmonių užimtumą, tai perkodavimas neturėtų būti taikomas amžiaus grupėse nuo 15 iki 64 metų. *1 pav.* pavaizduotame skirstinyje reikėtų perkoduoti visas reikšmes, viršijančias 64 metus, į vieną grupę.



1 pav. Duomenų pasiskirstymas pagal amžių

Apvalinimas

58. Apvalinimas yra panašus į grupavimą, tačiau taikomas tolydiems kintamiesiems. Apvalinimas padeda išvengti atvejų, kai duomenys tiksliai sutampa su išorinio šaltinio duomenimis ir dėl to didėja rizika atpažinti statistinio stebėjimo vienetus. Apvalinimas gali būti naudojamas ir duomenų detalumui sumažinti, pavyzdžiui, pašalinant trupmeninę skaičiaus dalį arba suapvalinant reikšmes iki artimiausios dešimties.

Lokalaus reikšmių slėpimo metodas

59. Lokalaus reikšmių slėpimo metodas gali būti taikomas duomenims jau pritaikius kurį nors iš perkodavimo metodų. Perkodavimas padeda sumažinti reikšmių, kurias reikia paslėpti, skaičių. Lokalus reikšmių slėpimas reiškia, kad tam tikros kintamojo reikšmės pakeičiamos tuščia

reikšme. Reikia pabrėžti, kad taikant šį metodą nėra paslepiamos visos tam tikro kintamojo reikšmės visiems statistinio stebėjimo vienetams, kaip būtų daroma tiesiog pašalinant tiesioginį identifikatorių (pvz., *vardas*). Lokalaus reikšmių slėpimo metodas iliustruojamas *11 lentelėje*.

11 lentelė. Lokalus reikšmių slėpimas

	Prieš lokalųjį reikšmių slėpimą			**Po lokalaus reikšmių slėpimo**		
ID	**Lytis**	**Regionas**	**Išsilavinimas**	**Lytis**	**Regionas**	**Išsilavinimas**
1	moteris	kaimas	aukštesnysis	moteris	kaimas	NA / trūkstama reikšmė
2	vyras	kaimas	aukštesnysis	vyras	kaimas	aukštesnysis
3	vyras	kaimas	aukštesnysis	vyras	kaimas	aukštesnysis
4	vyras	kaimas	aukštesnysis	vyras	kaimas	aukštesnysis
5	moteris	kaimas	vidurinis	moteris	kaimas	vidurinis
6	moteris	kaimas	vidurinis	moteris	kaimas	vidurinis
7	moteris	kaimas	vidurinis	moteris	kaimas	vidurinis

60. Nagrinėkime septynių statistinio stebėjimo vienetų duomenis su trimis kvaziidentifikatoriais: lytimi, regionu ir išsilavinimu. Matome, kad jų kombinacija „moteris“, „kaimas“ ir „aukštesnysis“ yra nesaugi, kadangi ji yra vienetinė turimoje imtyje. Šiuo atveju galima slėpti arba lytį, arba išsilavinimą, kadangi abiem atvejais liks bent 3 statistinio stebėjimo vienetai su vienodomis likusiomis kintamųjų reikšmėmis. Pasirinkimas, kurio kintamojo reikšmes slėpti, turėtų priklausyti nuo to, kuri informacija yra svarbesnė vartotojui ir kuriuo atveju bus minimizuotas bendras paslėptų reikšmių skaičius. Šiame pavyzdyje laikoma, kad informacija apie lytį yra svarbesnė, todėl paslėpta išsilavinimo reikšmė.

61. Kadangi tolydūs kintamieji turi labai daug unikalių reikšmių (pvz., pajamos), *j* vienetų anonimiškumas ir lokalus reikšmių slėpimas nėra tinkami tolydiems kintamiesiems ir daug skirtingų kategorijų turintiems kategoriniams kintamiesiems. Tokiais atvejais pirmiau galėtų būti taikomas perkodavimas, siekiant sumažinti galimų kategorijų skaičių.

TREČIASIS SKIRSNIS
DUOMENŲ RINKINIO REIKŠMES KEIČIANTYS METODAI

Įrašų keitimas

62. Įrašų keitimas yra mikroduomenims taikomas metodas. Mikroduomenyse parenkamos tam tikros poros įrašų (asmenų, namų ūkių ir pan.). Suporuotų įrašų kai kurių kintamųjų reikšmės sutampa, o nesutampančios sukeičiamos tarpusavyje. Sukeičiami kintamieji dažnai būna geografiniai. Šis kintamųjų sukeitimas įneša mikroduomenims neapibrėžtumo, todėl įsibrovėlio išvada apie tam tikrą asmenį / namų ūkį gali būti neteisinga.

63. Įrašų sukeitimas gali būti atsitiktinis arba tikslinis. Atsitiktinio įrašo sukeitimo atveju asmenys / namų ūkiai, kurie bus keičiami, yra atrenkami atsitiktinai su vienoda tikimybe, o tikslinio įrašų sukeitimo atveju yra atrenkami įrašai, turintys didelę atskleidimo tikimybę ir kiekvienam tokiam įrašui parenkama pora.

64. Įrašų keitimo metodas taikomas mikroduomenims tam, kad gautume pakeistą agreguotą lentelę / hiperkubą, kuri skiriasi nuo pradinės.

65. Pavyzdžiui, turime gyventojų skaičių detaliu teritoriniu lygmeniu. Laikoma, kad tų langelių, kuriuose yra mažiau nei 10 gyventojų, reikšmės yra konfidencialios. Tada konkrečiam namų ūkiui iš konfidencialios teritorijos X parenkamas kitas namų ūkis teritorijoje Y. Tada namų ūkiui iš X teritorijos įrašoma Y teritorija, o suporuotam namų ūkiui iš Y teritorijos įrašoma X teritorija paliekant visus kitus tų asmenų duomenis nepakeistus.

Atsitiktinis triukšmas

Atsitiktinio triukšmo moduliai

66. Atsitiktinis triukšmas yra agreguotiems duomenims taikomas metodas, kurį apibūdina tikimybinis skirstinys ir kuriam parenkamas šis skirstinys. Vadove aprašomas atsitiktinio triukšmo metodo variantas yra parengtas Australijos statistikos tarnybos ir pagrįstas vadinamaisiais langelio raktais (angl. *cell keys*). Tai leidžia užtikrinti, kad atsitiktinis triukšmas, kuris yra pritaikomas tam tikram langeliui, visada bus vienodas, netgi jei šis langelis yra skirtingose lentelėse.

67. Atsitiktinio triukšmo įgyvendinimas gali apimti tris modulius:

67.1. langelio rakto modulis (angl. *cell key module*);

67.2. modulis atsitiktiniam triukšmui nustatyti pagal langelio raktą ir triukšmo pasiskirstymo parametrų matricą;

67.3. adityvumo atstatymo modulis.

68. Pirmasis modulis užtikrina duomenų keitimo nuoseklumą (angl. *consistency*), antrasis nustato statistines duomenų keitimo savybes. Trečiasis modulis užtikrina adityvumą, tačiau „sugadina“ nuoseklumą (t. y. padaro taip, kad toje pačioje lentelėje susumavus visas reikšmes gaunama tiksli suma, tačiau tas pats langelis skirtingose lentelėse gali turėti skirtingą reikšmę).

Langelio rakto modulis

69. Langelių raktai turėtų būti gaunami pasinaudojus tolydžiuoju skirstiniu, apibrėžto tam tikru sveikuoju skaičiumi (pvz., nuo 1 iki 100). Procesas, apibrėžiantis langelių raktus, turi būti nuoseklus, t. y. turi būti užtikrinta, kad tas pats langelis visada turės tą patį raktą net ir skirtingose lentelėse. Atsitiktinio triukšmo dydžiui nustatyti kiekviename langelyje naudojamas to langelio raktas ir dažnis. Šis žingsnis yra deterministinis ir gali būti įgyvendinamas taip, kad triukšmo skirstiniai beveik tiksliai atitiktų iš anksto nustatytus skirstinius, kurie turi būti nurodyti kaip metodo parametrai. Proceso atsitiktinumas atsiranda tik nustatant langelių raktus:

69.1. Kiekvienam mikroduomenų rinkinio įrašui priskiriamas atsitiktinis skaičius (vadinamasis įrašo raktas). Įrašo raktas yra pasiskirstęs pagal tolydųjį skirstinį su reikšmėmis intervale [1; 100];

69.2. Skaičiuojant agreguotus duomenis, t. y. skaičiuojant įrašus, kurie turi tam tikras reikšmių kombinacijas (t. y. priklausantys vienam skelbiamos informacijos langeliui), susumuojamos įrašo rakto reikšmės. Įrašo rakto reikšmių sumos rezultatas yra langelio raktas, kuris irgi įgyja reikšmes nuo 1 iki 100. Langelių raktų reikšmės taip pat turės tolydųjį skirstinį su reikšmėmis intervale [1; 100].

Atsitiktinio triukšmo parametrų nustatymas

70. Atsitiktinio triukšmo įtaka gali būti lengvai ir lanksčiai valdoma pasirenkant atitinkamus parametrus apibrėžiančius tikimybinį skirstinį. Net atsitiktinio apvalinimo metodą galima laikyti atsitiktiniu triukšmu su specifiniu triukšmo skirstiniu. Žemiau išvardytos savybės yra būtinos ir (arba) kontroliuojamos parenkant parametrus:

70.1. triukšmo tikimybė / nepaslinktumo savybė (angl. *Noise expectation / Unbiasedness property*);

70.2. triukšmo dispersija;

70.3. tam tikri dažniai (pvz., 1 ar 2) neturi atsirasti duomenyse (t. y. kad nebūtų skelbiama 1 ar 2 statistinio stebėjimo vienetų informacija);

70.4. nulinės reikšmės niekada nekeičiamos (t. y. jeigu pradiniuose duomenyse kažkuriame langelyje yra nulis, tai ir tuose pačiuose langeliuose pritaikius konfidencialumo užtikrinimo metodus turi likti nulis).

71. Norint pritaikyti atsitiktinio triukšmo metodą konkretiems duomenims, ištestuojami skirtingus atsitiktinio triukšmo parametrus turintys variantai ir pasirenkamas geriausiai turimus duomenis tenkinantis variantas.

Adityvumas

72. Nuoseklių langelių raktų naudojimas užtikrina galutinio perturbuoto duomenų rinkinio nuoseklumą. Atsitiktinio triukšmo reikšmė konkrečioje eilėje visada bus ta pati, net kai tas langelis yra kitoje lentelėje. Tačiau kiekvieno langelio perturbacija vyksta nepriklausomai. Dėl to susumavus perturbuotos lentelės duomenis nebus gauta tiksli suma (pvz., šalyje yra lygiai 3 000 000 gyventojų, tačiau susumavę gyventojų skaičių visose savivaldybėse gausime iš viso 2 999 900 gyventojų). Siekiant to išvengti yra naudojamas papildomas adityvus modulis. Pritaikius adityvumo modulį perturbuotų rezultatų lentelės reikšmės gali būti sudėtos gaunant „beveik“ tą pačią reikšmę kaip ir „iš viso“. Tačiau pritaikius adityvumo modulį atsiranda „nereikšmingas“ nenuoseklumas tarp skirtingų lentelių, t. y. tas pat langelis skirtingose lentelėse gali turėti „nereikšmingai“ skirtingas reikšmes.

PRAM metodas

73. PRAM metodas yra duomenų rinkinio reikšmes keičiantis metodas, skirtas kategoriniams kintamiesiems. Metodo esmė yra perklasifikuoti kai kurias vieno ar kelių kategorinių kintamųjų reikšmes taip, kad įsibrovėlis galėtų identifikuoti statistinio stebėjimo vienetą, tačiau su teigiama tikimybe identifikuojamas ne tas asmuo. Tai reiškia, kad įsibrovėlis gali susieti keletą duomenų rinkinio įrašų su turima papildoma informacija, tačiau negali būti užtikrintas, kad susiejimas įvykdytas teisingai.

74. PRAM metodas apibrėžiamas naudojant perėjimų matricą *P*. Tarkime, yra *k* galimų tam tikro kategorinio kintamojo kategorijų / klasių. Matrica P apibrėžia perėjimo tikimybes, t. y. tikimybes, kad kintamojo reikšmė išliks nepakitusi arba bus pakeista į kurią nors iš $k - 1$ kitų galimų reikšmių. Vienam kintamajam, turinčiam *k* kategorijų, *P* matrica yra dydžio $k \times k$.

75. Pavyzdžiui, nagrinėkime kintamąjį *Regionas*, kuris gali įgyti tris skirtingas reikšmes: „didmiestis“, „miestas“ ir „kaimas“. Tokiam kintamajam apibrėžiama 3×3 dydžio matrica *P*:

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0{,}05 & 0{,}8 & 0{,}15 \\ 0{,}05 & 0{,}15 & 0{,}8 \end{bmatrix}$$

76. Matricos reikšmė pozicijoje (1, 1) reiškia tikimybę, kad pirmoji kintamojo reikšmė išliks nepakitusi. Reikšmė pozicijoje (1, 2) – pirmoji kintamojo reikšmė bus pakeista į antrąją ir t. t.

77. Vienas iš būdų užtikrinti, kad kategorinio kintamojo reikšmių skirstinys išliktų nepakitęs, pritaikius duomenims PRAM metodą, yra parinkti perėjimų matricą taip, kad reikšmių vidurkis prieš metodo taikymą ir po jo išliktų nepakitęs. Toks metodas vadinamas invariantišku PRAM metodu ir yra įdiegtas *R* paketo *sdcMicro* funkcijoje *pram()*.

78. *12 lentelėje* pateikta, kaip gali skirtis rezultatai, naudojant paprastą anksčiau aprašytą perėjimų matricą ir panaudojus invariantišką PRAM metodą.

12 lentelė. PRAM metodų rezultatų palyginimas

Reikšmė	Reikšmių skaičius duomenyse prieš PRAM	Reikšmių skaičius duomenyse naudojant PRAM	Reikšmių skaičius duomenyse naudojant invariantišką PRAM
didmiestis	5,000	5,052	4,998
miestas	500	457	499
kaimas	400	391	403

79. Matome, kad naudojant paprastą PRAM metodą rezultatas atspindi reikšmių vidurkį, apskaičiuotą naudojant anksčiau apibrėžtą perėjimų matricą *P* (nedidelis nuokrypis nuo vidurkio atsiranda dėl atsitiktinumo generuojant tikimybes kiekvienam įrašui). Tuo tarpu invariantiško PRAM metodo atveju perėjimo matrica buvo parinkta taip, kad vidurkis sutaptų su pradiniu reikšmių pasiskirstymu.

80. Naudojant invariantišką PRAM metodą, galima pasirinkti norimą bendrą pakeitimų duomenyse skaičių (t. y. kiek iš viso reikšmių bus pakeista). Šis parametras atsispindės sudarant perėjimų matricą.

81. PRAM metodas ypač naudingas, kai turima daug kintamųjų, ir kitų metodų, tokių kaip lokalaus slėpimo ar perkodavimo, naudojimas gali lemti didelį informacijos praradimą. Atskleidimo rizikos ir duomenų panaudojamumo vertinimas pritaikius PRAM metodą yra būtinas.

82. Tam, kad duomenų analitikai galėtų padaryti teisingas statistines išvadas, reikia žinoti apie patį metodą ir perėjimo matricą. Perėjimo matrica, kartu su pradine atsitiktinių skaičių generavimui naudojama reikšme gali leisti atkurti pradines duomenų rinkinio reikšmes. Todėl ypač svarbu skelbiant duomenų rinkinį ir panaudotą perėjimų matricą, kartu neskelbti pradinės atsitiktiniams skaičiams generuoti naudojamos reikšmės.

83. PRAM metodo trūkumas – galimas netikėtinų reikšmių sugeneravimas (pvz., 64 metų moksleivis). Dėl to pritaikius metodą būtina atidžiai peržiūrėti sugeneruotas reikšmes. Taip pat negalimų reikšmių generavimą galima užkardyti perėjimų matricoje pasirinkus atitinkamą tikimybę 0.

Mikroagregavimas

84. Mikroagregavimas labiausiai tinka tolydiems kintamiesiems, tačiau kai kuriais atvejais gali būti pritaikomas ir kategoriniams kintamiesiems. Mikroagregavimas naudingas, kai turima apibrėžta konfidencialumo taisyklė (pvz., *j* vienetų anonimiškumo), leidžianti skelbti tik tokį duomenų rinkinį, kuriame kiekviena tam tikrų kintamųjų reikšmių kombinacija būdinga ne mažiau kaip *k* statistinio stebėjimo vienetų.

85. Norint taikyti mikroagregavimą, pirmiausia sudaromos nedidelės statistinio stebėjimo vienetų grupės, kurios yra homogeniškos tam tikrų pasirinktų kintamųjų atžvilgiu (pvz., turi panašias pajamas ar amžių). Tuomet visiems vienai grupei priklausantiems statistinio stebėjimo vienetams tų kintamųjų reikšmės pakeičiamos viena bendra reikšme, pavyzdžiui, grupės vidurkiu.

86. Yra keletas mikroagregavimo būdų, kurie skiriasi:

86.1. kaip apibrėžiamos homogeniškos grupės;

86.2. algoritmais, kuriais nustatomas statistinio stebėjimo vienetų priklausymas vienai iš homogeniškų grupių;

86.3. bendrõs reikšmẽs, kuria pakeičiamos atskirų statistinio stebėjimo vienetų reikšmės, nustatymu.

87. Praktikoje mikroagregavimo metodas veikia tuo geriau, kuo labiau homogeniškos yra statistinio stebėjimo vienetų grupės.

88. Mikroagregavimą galima pritaikyti naudojant R paketo *sdcMicro* funkciją *microaggregation()*. Pagal nutylėjimą, funkcija parenka 3 homogeniškas grupes ir pritaiko grupės vidurkį, tačiau parametrais galima rinktis kitokį gupių skaičių ar kitą bendrą reikšmę, pavyzdžiui, grupės medianą. Didelėse grupėse, kuriose yra didesnis heterogeniškumas, rekomenduojama rinktis medianą vietoje vidurkio. Kadangi didelėse grupėse dažniau pasitaiko išsiskiriančių reikšmių, medianos pasirinkimas gali apsaugoti nuo to, kad mikroagregavimo metu visiems grupės nariams bus priskirta išsiskirianti reikšmė.

89. *13 lentelėje* iliustruojama, kaip skiriasi mikroagregavimo rezultatas, kai vienu atveju naudojamas vidurkis, kitu – mediana.

13 lentelė. Mikroagregavimo rezultatų palyginimas

ID	Grupė	Pajamos	Mikroagregavimas vidurkiu	Mikroagregavimas mediana
1	1	2,300	2,245	2,300
2	2	2,434	3,608	2,434
3	1	2,123	2,245	2,300
4	1	2,312	2,245	2,300
5	2	6,045	3,608	2,434

6	2	2,345	3,608	2,434

90. Jeigu mikroagregavimas naudojamas kategoriniam kintamajam, tuomet grupės mediana taikoma siekiant pakeisti tos grupės narių kintamojo reikšmes.

91. Reikia atkreipti dėmesį, kad taikant mikroagregavimą keliems skirtingiems kintamiesiems atskirai ir nesirenkant jokių kitų papildomų statistinio atskleidimo kontrolės metodų, išauga statistinio atskleidimo rizika. Tokiu atveju rekomenduojama taikyti daugiamatį mikroagregavimą (angl. *multivariate microaggregation*).

92. Daugiamačio mikroagregavimo metodui pritaikyti pirmiausia pagal kelis skirtingus kintamuosius sukuriamos homogeniškos grupės. Grupės sudaromos remiantis daugiamate atstumo tarp statistinio stebėjimo vienetų funkcija. Tuomet visų kintamųjų reikšmės visiems vienos grupės nariams pakeičiamos bendromis reikšmėmis. Pavyzdys pateikiamas *14 lentelėje*.

14 lentelė. Mikroagregavimas

ID	Grupė	Prieš mikroagregavimą			Po mikroagregavimo		
		Pajamos	Išlaidos	Turtas	Pajamos	Išlaidos	Turtas
1	1	2300	1714	5,3	2285,7	1846,3	6,3
2	1	2434	1947	7,4	2285,7	1846,3	6,3
3	1	2123	1878	6,3	2285,7	1846,3	6,3
4	2	2312	1950	8,0	3567,3	2814,0	8,3
5	2	6045	4569	9,2	3567,3	2814,0	8,3
6	2	2345	1923	7,8	3567,3	2814,0	8,3

Matome, kad grupavimas pagal pajamas, išlaidas ir turtą statistinio stebėjimo vienetus suskirsto į kitokias grupes negu grupuojant tik pagal pajamas.

93. Yra keletas skirtingų daugiamačio mikroagregavimo metodų, kurie skiriasi algoritmu, naudojamu sukurti homogeniškas statistinio stebėjimo vienetų grupes.

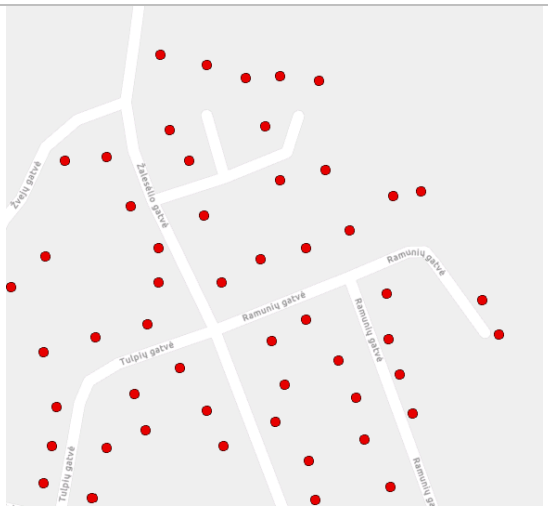
94. Norint taikyti mikroagregavimą keliems kintamiesiems, pirmiausia rekomenduojama nagrinėti kintamųjų koreliacinę ar kovariacinę matricą. Jeigu ne visi kintamieji stipriai koreliuoja tarpusavyje, bet matome kelias kintamųjų grupes, kurių viduje koreliacija yra stipri, mažiau informacijos prarasime taikydami mikroagregavimą atskirai kelioms kintamųjų grupėms. Apskritai mažiau informacijos prarasime taikydami daugiamatį mikroagregavimą, jeigu kintamieji stipriai koreliuoja tarpusavyje.

KETVIRTASIS SKIRSNIS
STATISTINIO ATSKLEIDIMO KONTROLĖS METODŲ TAIKYMAS GEOGRAFINIŲ INFORMACINIŲ SISTEMŲ DUOMENIMS

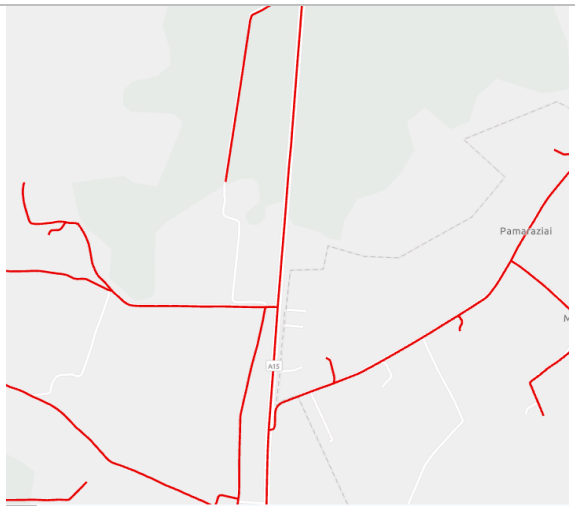
95. Geografinių informacinių sistemų (toliau – GIS) duomenys yra fiziškai arba virtualiai egzistuojantys objektai žemės paviršiuje.

96. Kiekvienas objektas turi vizualinę ir atributinę tarpusavyje susieta išraišką. GIS duomenys skirstomi į du tipus: vektoriniai ir rastriniai:

96.1. *Vektoriniai duomenys*, kurie gali būti: taškai (*2 pav.*), linijiniai (*3 pav.*), poligoniniai (*4 pav.*).



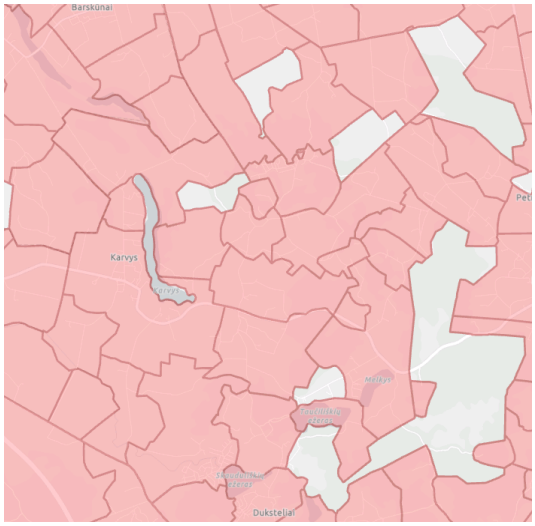
2 pav. Vektoriniai duomenys – taškai



3 pav. Vektoriniai duomenys – linijos

Taškai: apibrėžti XY koordinačių pora. Vaizduojami objektai, kurie neturi plotinės išraiškos, priklausančios nuo mastelio. Statistikoje dažniausiai naudojami adresai, tai yra vienas iš detalesnių statistinių duomenų agregavimo vienetų.

Linijiniai: objektai, kurie priklausomai nuo mastelio neturi plotinės išraiškos, bet turi pradžios, tárpines ir pabaigos koordinates – dažniausiai vaizduojamos infrastruktūrinių objektų ašinės linijos (pvz., kelių, upių, elektros linijų ir t. t.)



4 pav. Vektoriniai duomenys – plotai

Poligoniniai: uždaros linijos suformuoja plotus – t. y. tokie fiziniai objektai kaip ežerai, pastatai, pasėliai ir tokie virtualūs objektai kaip administraciniai vienetai, kaimai ir miestai, gardelės, pašto kodai. Bet koks administracinis-teritorinis vienetas gali būti naudojamas statistiniams duomenims agreguoti.

96.2. *Rastrinis* duomenų tipas (*5 pav.*) gali būti suprantamas kaip paprasčiausias skaitmeninis paveikslėlis. Šiame duomenų tipe informacija saugoma langeliuose (angl. *cell, pixel*). Langeliai yra organizuoti eilutėmis ir stulpeliais ir kiekviena ląstelė saugo vieną reikšmę. Dažniausiai langeliuose saugoma informacija apie spalvą, gali būti koduojamos reikšmės ir apie kitus reiškinius: žemės dangos tipą, žemės paviršiaus temperatūrą ir pan.

5 pav. Rastriniai duomenys

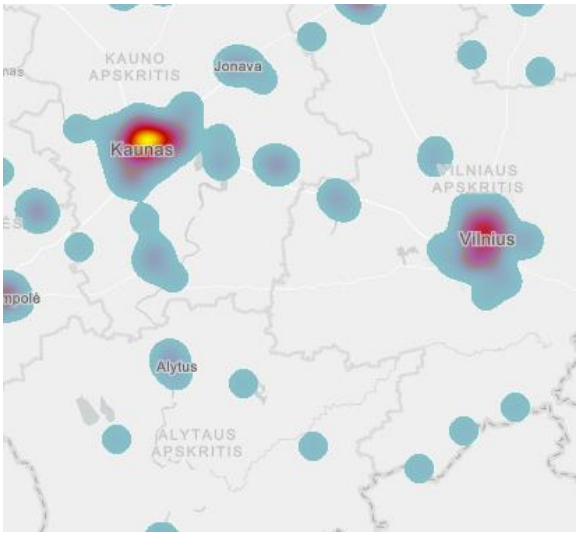
97. Aukščiau aprašyti GIS duomenys nurodo objektus žemės paviršiuje ir patys nėra konfidencialūs (išskyrus atvejus pasienio ruožuose) ir neturi jokios asmeninės informacijos. Bet jungiant su statistika, kuri suteikia informaciją apie asmenį, gali tapti asmens duomenimis, kuriems galios bendros asmens duomenų nuasmeninimo taisyklės. Taip pat reikia atsižvelgti į agreguojamų statistinių duomenų kintamųjų kontekstą ir skaičių, pvz., turint pašto kodo teritorijoje informaciją apie asmenis, jų amžių ir pajamas skelbti asmenis pagal amžiaus grupes ir bendras pajamas arba pajamų rėžius. Tačiau, kai erdvinė informacija ruošiama / skelbiama sprendimams priimti, turėtų būti siekiama maksimalaus išsamumo lygio, kurį galima suderinti su asmenų privatumo apsauga.

98. Alternatyvus būdas – nuasmeninti erdvines teritorijas ir pagal galimybes keisti (stambinti) jų ribas, t. y. išimti identifikavimo kodus ir pavadinimus, pakeičiant juos alternatyviais. Toks sprendimas gali būti tinkamas mokslininkams, bet bus netinkamas viešam naudojimui, nes vartotojai nori sužinoti informaciją pagal tikrus teritorinius vienetus.

99. Pavyzdžiui, siekiant sumažinti asmenų identifikavimo riziką, galima naudoti skirtingo dydžio teritorinius vienetus agreguotiems statistiniams duomenims teikti, t. y. tankiai apgyvendintose teritorijose (miestuose) naudoti 100 m^2 gardelę, rečiau (kaimuose) apgyvendintose teritorijose – 1 km^2 gardelę. Jei to nepakanka, galima apsvarstyti duomenų reikšmes keičiančių metodų taikymą pateikiamiems agreguotiems statistiniams duomenims pagal teritorijas. Pavyzdžiui, teritorijose, kuriose pateikta informacija netenkina konfidencialumo taisyklių, įrašyti kintamojo vidutinę reikšmę, apskaičiuotą iš šalia esančių teritorijų, arba sukeisti skirtingų teritorijų reikšmes, taip pat dažnai naudojama praktika kintamiesiems suteikti intervalines reikšmes (pvz., amžiaus grupės, pajamų intervalai).

100. Kai kuriais atvejais, kai reikia pateikti tik apžvalgą, t. y. užtenka tik vizualinės pateikties, galima naudoti karštų taškų žemėlapį (angl. *heat map)* (*6 pav.*), kuris neleidžia atskleisti išsamios informacijos apie tam tikrą vietą ar asmenį.

101. Jei nėra asmens duomenų atskleidimo rizikos arba ji yra minimali, geografinė informacija turėtų suteikti kuo daugiau informacijos, kad vartotojai galėtų pamatyti padėtį / situaciją jų dominančioje teritorijoje (pvz., nusikalstamumus jų rajone gali padėti bendruomenėms atkreipti dėmesį į besiformuojančią problemą ir laiku susisiekti su atitinkamomis įstaigomis (pvz., policija).



6 pav. Karštų taškų žemėlapis

PENKTASIS SKIRSNIS
STATISTINIO ATSKLEIDIMO KONTROLĖS METODŲ TAIKYMO IMČIŲ TYRIMUOSE PAVYZDŽIAI

102. **Statistinio atskleidimo kontrolės metodų taikymas socialinės statistikos imčių tyrimų mikroduomenims:**

102.1. *Duomenų nuasmeninimas*. Pašalinami visi tiesioginiai identifikatoriai, įskaitant vardą, asmens kodą, adresą, pašto kodą ir kt. Taip pat reikėtų pašalinti ir gimimo datą – paprastai rekomenduojama pašalinti visus panašius kintamuosius, tokius kaip gimimo metai ir mėnuo. Pašalinus tiesioginius identifikatorius, duomenyse vis tiek gali būti unikalių ir retų kintamųjų derinių, leidžiančių identifikuoti šiuos požymius turinčius asmenis / namų ūkius.

15 lentelėje pateikiamas kai kurių pagrindinių kintamųjų sąrašas ir galimas kontrolės metodas. Tai yra pasiūlymai, o ne taisyklės ir sąrašas nėra baigtinis. Pasirinktas informacijos atskleidimo kontrolės metodas turėtų būti tinkamas apklausai ir imties dydžiui. Visada reikėtų atsižvelgti į vartotojų reikalavimus, jei kintamasis reikalingas žemesniam lygiui, nei rekomenduojama, tada kitas kintamasis turėtų būti apsaugotas aukštesniu lygiu. Pavyzdžiui, jei reikalingas tikslus amžius, o ne sugrupuotas, tada atlyginimo ir pajamų kintamieji galėtų būti sugrupuoti stambiau. Imties dydis gali būti skirstomas į:

102.1.1. mažą – jei imtyje yra mažiau nei 1 % populiacijos, daugumos pagrindinių kintamųjų nebūtina apsaugoti. Tačiau visada bus tokių, kaip adresas, kuriems reikalinga kontrolė;

102.1.2. vidutinį – jei imties dydis yra nuo 1 % iki 3 % populiacijos, tikėtina, kad reikės apsaugoti kelis pagrindinius kintamuosius;

102.1.3. didelį – jei imties dydis yra didesnis nei 3 % populiacijos, gali prireikti papildomos apsaugos. Gali tekti pašalinti kai kuriuos įrašus arba lokaliai apsaugoti konkrečius pagrindinius kintamuosius.

15 lentelė. Pagrindinių kintamųjų apsaugos pavyzdžiai, kad duomenys nebūtų atskleisti

Kintamieji	Siūlomas metodas
Geografiniai – gyvenamoji vieta, darbo vieta ir kt.	Žemiausias geografinis lygmuo yra adresas arba XY koordinatė. Jei toks lygmuo neapsaugo asmens duomenų nuo atskleidimo galima stambinti geografinį lygmenį (pvz., agreguoti į gardeles nuo 100 m iki 1 km, gyvenamąsias vietoves / seniūnijas) arba mažinti kitų kintamųjų detalumą.

Kintamieji	Siūlomas metodas
Amžius	Mažose imtyse – paprastai metai gali būti pateikti, vidutinėse ir didelėse – galėtų būti suskirstyta į grupes, tarkime, 5 metų.
Namų ūkio dydis	Žiūrėti į 103.3. *Didelių namų ūkių duomenų tvarkymas*
Gimimo šalis / pilietybė	Apsvarstyti, ar reikia tokio išsamumo lygio. Gali būti priimtina suskirstyti šiuos kintamuosius, pvz., LT, ES, kita.
Profesija / veiklos sritis – pagrindinis darbas, papildomas darbas, ankstesnis darbas ir kt.	Apsvarstyti, ar reikia įtraukti 4 skaitmenų lygį pagal Lietuvos profesijų klasifikatorius (LPK). Rekomenduojama, kad, jei įmonės veiklos sritis pateikta 4 skaitmenų lygiu, profesija galėtų būti tik 3 skaitmenų lygiu ir t. t.
Atlyginimas – bendras ir grynasis, metinis, savaitinis, valandinis ir kt. Premijos ir kt.	Labai dideli atlyginimai ir premijos paprastai apsaugomi perkoduojant didžiausias reikšmes (angl. *top-coding*), savaitės ir valandos įkainiai taip pat.
Pajamos – namų ūkio pajamos, bendrosios ir grynosios ir kt.	Paprastai suapvalinama iki artimiausio 1 000 eurų. Labai didelės vertės galėtų būti užkoduotos, panašiai kaip atlyginimai.

102.2. *Tęstinės apklausos* (angl. *longitudinal*), tokios kaip gyventojų užimtumo (angl. *LFS*) arba pajamų ir gyvenimo sąlygų (angl. *EU-SILC*). Tokio tipo apklausoms, kai susiejamos nuoseklios bangos, gali tekti taikyti kitus papildomus apribojimus, pavyzdžiui, stambesnis demografinių kintamųjų, tokių kaip šeiminė padėtis ir vaikų skaičius, agregavimas, arba modifikuoti duomenis perkoduojant amžiaus, šeimyninės padėties, socialinius ir ekonominius kintamuosius.

102.3. *Didelių namų ūkių duomenų tvarkymas*. Kai tyrimas atliekamas namų ūkio lygiu, mikroduomenys yra hierarchiniai, o identifikavimo raktus gali sudaryti tokie kintamieji kaip namų ūkio amžiaus ir lyties struktūra bei asmenų tarpusavio santykiai, padidėja atskleidimo rizika. Todėl siūloma pašalinti iš duomenų rinkinio namų ūkio įrašus, jei namų ūkį sudaro 10 ar daugiau narių.

103. **Statistinio atskleidimo kontrolės metodų taikymas verslo statistikos imčių tyrimų mikroduomenims.** Šio tipo mikroduomenų atskleidimo rizika yra didesnė nei socialinių (gyventojų) statistinių tyrimų mikroduomenų. Taip yra todėl, kad verslo statistikos tyrimų duomenų dispersija yra didesnė nei socialinių (gyventojų) statistinių tyrimų duomenų. Daug informacijos apie įmones, kuri gali būti panaudota jas identifikuojant paviešintame mikroduomenų rinkinyje, yra skelbiama viešai. *16 lentelėje* pateikiami kai kurie verslo statistikos tyrimų mikroduomenų apsaugos metodai.

16 lentelė. Verslo mikroduomenų apsaugos metodai

Metodas	Apibrėžimas	Pavyzdys
Globalus perkodavimas	Perkodavimas taikomas visam duomenų rinkiniui.	Darbuotojų skaičiaus grupės 250–499, 500–999, 1000 ir daugiau sujungiamos į vieną grupę 250+.
Viršaus / apačios perkodavimas	Visoms reikšmėms, esančios virš arba žemiau parinktos ribos priskiriama ta reikšmė.	Visos apyvartos, viršijančios 500 000 EUR, yra lygios 500 000 EUR.
Mikroagregavimas	Įrašai yra sugrupuojami ir tos grupės suvestinis rodiklis priskiriamas visai grupei.	Įrašai yra surūšiuojami pagal apyvartą didėjimo tvarka ir kiekvienai grupei, pavyzdžiui, po 3 įrašus tikroji apyvarta pakeičiama

Metodas	Apibrėžimas	Pavyzdys
		grupės vidurkiu.

VII SKYRIUS
DUOMENŲ PANAUDOJAMUMO IR INFORMACIJOS PRARADIMO MATAI

104. Taikant bet kurį statistinio atskleidimo kontrolės metodą, kyla tam tikros informacijos praradimo rizika. Statistinio atskleidimo kontrolės metodus rekomenduojame parinkti taip, kad informacijos praradimas būtų minimalus kartu maksimizuojant duomenų panaudojamumą, užtikrinant duomenų konfidencialumą. Duomenų panaudojamumas (angl. *data utility*) šiame kontekste reiškia nuasmeninto duomenų rinkinio tinkamumą statistinei analizei, kurią atlieka galutinis vartotojas. Norint maksimizuoti duomenų panaudojamumą, reikia įsivertinti duomenų panaudojamumą prieš nuasmeninimą ir po jo. Informacijos praradimo matas šiuo atveju yra atvirkštinis dydis duomenų panaudojamumui: kuo didesnis duomenų panaudojamumas po nuasmeninimo, tuo mažesnis informacijos praradimas.

105. **Kategoriniams kintamiesiems taikomi matai**:

105.1. ***Trūkstamų reikšmių skaičius***. Dažniausiai trūkstamos reikšmės atsiranda taikant lokalaus slėpimo metodą. Reikia atkreipti dėmesį, kad trūkstamos reikšmės duomenų rinkinyje gali atsirasti ir dėl kitų priežasčių, pavyzdžiui, dėl neatsakymo. Todėl reikėtų lyginti, kiek trūkstamų reikšmių buvo pradiniame rinkinyje ir kiek jų yra pritaikius statistinio atskleidimo kontrolės metodą.

105.2. ***Pakeistų reikšmių skaičius.*** Skaičiuojama analogiškai kaip ir trūkstamų reikšmių atveju. Lokalaus slėpimo ar kitu metodu paslėptas reikšmes galima laikyti kaip pakeistas į trūkstamą reikšmę ir taip pat įtraukti į šį matą.

105.3. ***Ryšių lentelių palyginimas.*** Reikėtų palyginti pasiskirstymą tarp nuasmeninamo duomenų rinkinio kintamųjų porų. Siekiant išlaikyti duomenų rinkinio tinkamumą statistinei analizei, ryšių lentelės neturėtų smarkiai pasikeisti pritaikius statistinio atskleidimo kontrolės metodus.

106. **Tolydiems kintamiesiems taikomi matai:**

106.1. ***Vidurkis, kovariacijos, koreliacijos koeficientai***. Pagrindinės duomenų charakteristikos neturėtų reikšmingai pasikeisti pritaikius statistinio atskleidimo kontrolės metodus. Galima taikyti ir įvairias kitas charakteristikas, jas skaičiuoti atskiriems kintamiesiems arba nagrinėti daugiamačius skirstinius. Koreliacijos koeficientas ir jo pasikeitimai yra ypač svarbūs, analizuojant duomenų rinkinio tinkamumą regresinei analizei.

106.2. ***Informacijos praradimo matas IL1s.*** Šis matas skirtas atstumui tarp pradinio duomenų rinkinio X ir galutinio duomenų rinkinio Z tolydžių kintamųjų įvertinti. Matas apibrėžiamas taip:

$$IL1s = \frac{1}{pn}\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{|x_{ij} - z_{ij}|}{\sqrt{2}S_j},$$

čia:

p – tolydžių kintamųjų skaičius duomenų rinkinyje;

n – duomenų rinkinio įrašų skaičius;

x_{ij} ir z_{ij} – j-tojo kintamojo reikšmės i-tajam įrašui, atitinkamai originaliame duomenų rinkinyje ir galutiniame;

S_j – j-tojo kintamojo standartinis nuokrypis originaliame duomenų rinkinyje.

Nors funkcijos pavadinimas susietas su duomenų panaudojamumu, ji grąžina informacijos praradimo matą. Jis naudingas lyginant kelis skirtingus metodus. Kuo šio mato reikšmė (t. y. atstumas tarp originalių ir pakeistų duomenų) didesnė, tuo mažesnis duomenų rinkinio panaudojamumas, tačiau tuo pat metu ir mažesnė atskleidimo rizika. Mažesnė mato reikšmė reiškia, kad duomenys, pritaikius nuasmeninimo metodus, nedaug skiriasi nuo originalių.

106.3. ***Tikrinės reikšmės (*angl. *eigenvalues)*.** Taip pat naudojamos skirtingiems metodams palyginti. Šis metodas parodo tikrinių reikšmių skirtumą tarp originalių duomenų rinkinio ir

nuasmenintų duomenų rinkinio. Kuo šis skirtumas didesnis, tuo didesnių pakeitimų duomenų rinkinyje buvo atlikta ir tuo daugiau informacijos prarasta.

107. **Gini koeficientas**. Tai vienas iš dispersijos matų. Jis skaičiuojamas originaliam ir nuasmenintam duomenų rinkiniui ir analizuojamas Gini koeficiento reikšmės pasikeitimas.

108. **Regresija**. Ji gali būti naudinga ne tik palyginant koreliacijų ir kovariacijų matricas, bet ir analizuojant, ar nepasikeitė duomenų rinkinio struktūra pritaikius statistinio atskleidimo kontrolės metodus. Nagrinėjant regresinės lygties parametrus, galima lyginti sąryšius tarp kelių netolydžių kintamųjų. Žinant skelbiamų duomenų tikslą ir sritį, pritaikius regresiją galima analizuoti koeficientų ir pasikliovimo intervalų pokyčius.

109. **Duomenų atvaizdavimas.** Įvairūs grafiniai duomenų vizualizavimo būdai padeda įvertinti pasikeitimus duomenų struktūroje, taikant nuasmeninimo metodus:

109.1. *Histograma arba tankio funkcijos grafikas* parodo galimus pasikeitimus duomenų skirstinyje. Statistinio atskleidimo metodų taikymas turėtų reikšmingai nedaryti įtakos skirstiniui.

109.2. *Stačiakampės diagramos* (angl. *boxplot*) padeda pastebėti pokyčius išsibarsčius tolydžių kintamųjų reikšmėms ir išskirtis.

109.3. *Mozaikinės diagramos* (angl. *mosaic plot*) parodo kategorinio kintamojo reikšmių pasiskirstymo pokyčius, ypač kai lyginami keli skirtingi metodai ar metodo parametrai.

VIII SKYRIUS
NAUDOTOS LITERATŪROS SĄRAŠAS

110. Vadovas parengtas naudojant šiuos literatūros šaltinius:

Antal, L., Enderle, T. and Giessing, S. . (19/05/2017). *Statistical disclosure control methods for harmonised protection of census data* (T. SGA Harmonised protection of census data in the ESS Work package 3).

Benschop, T. and Welch, M. (n.d.) . (be datos). Statistical Disclosure Control for Microdata: A Practice Guide. Nuskaityta iš https://sdcpractice.readthedocs.io/en/latest/

Benschop, T., Machingauta, C. and Welch, M. . (2021). Statistical Disclosure Control: a Practice Guide.

Eurostat. (2021). *European business statistics compilers' manual for international trade in goods statistics, 2021 edition* (2021 leid.).

Government Statistical Service. (October 2014). *GSS/GSR Disclosure Control Guidance for Microdata Produced from Social Surveys, GSS/GSR Disclosure Control Guidance for Microdata Produced from Social Surveys.*

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., Nordholt, E. S., Seri, G., Wolf, P.-P. (2010). Handbook on Statistical Disclosure Control. Nuskaityta iš https://ec.europa.eu/eurostat/cros/system/files/SDC_Handbook.pdf

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., et al. (2012). *Statistical Disclosure Control.* Chichester, UK: John Wiley & Sons Ltd.

IX SKYRIUS
PRIEDAI

111. Vadovo 1 priedas – Statistinės informacijos rengimo proceso schema.

112. Vadovo 2 priedas – Statistinio atskleidimo kontrolės poreikio nustatymo anketa.

113. Vadovo 3 priedas – Rizikų apskaičiavimo ir statistinio atskleidimo kontrolės metodų taikymo pavyzdžiai naudojant *R* paketą.

X SKYRIUS
BAIGIAMOSIOS NUOSTATOS

114. Pasikeitus Vadove nurodytiems teisės aktams, taikomos galiojančios šių teisės aktų redakcijų nuostatos.

_____