Duomenų sintetinimo metodų palyginimas

Simona Stankevičiūtė, Rūta Levulienė

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Taikomosios matematikos institutas, Naugarduko g. 24, Vilnius, Lietuva
simona.stankeviciute@mif.stud.vu.lt

Santrauka. Tyrime palyginti du sintetinių duomenų generavimo metodai – grandininių lygčių metodas (angl. *Fully Conditional Specification*, FCS) ir apibendrintieji adityvieji padėties, mastelio ir formos parametrų modeliai (angl. *Generalized Additive Models for Location, Scale and Shape*, GAMLSS), taikant juos paslaugų sektoriaus įmonių mėnesinių pajamų duomenims. Rezultatai parodė, kad FCS geriau atkuria originalių duomenų pasiskirstymą, tačiau GAMLSS pasižymi mažesne reikšmės atskleidimo rizika ir gali būti laikomas saugesniu konfidencialumo užtikrinimo požiūriu.

Raktiniai žodžiai: sintetiniai duomenys, duomenų konfidencialumas, duomenų atskleidimo rizika, FCS, GAMLSS.

1 Įvadas

Šiuolaikinėje duomenimis grįstoje ekonomikoje organizacijos vis dažniau susiduria su būtinybe suderinti duomenų prieinamumą su konfidencialumo užtikrinimu. Didėjant duomenų analitikos poreikiui ir kartu griežtėjant duomenų apsaugos reikalavimams, sintetiniai duomenys tampa vis svarbesne alternatyva tradiciniams duomenų atskleidimo kontrolės metodams [1]. Sintetiniai duomenys laikomi tinkama alternatyva realiems duomenims todėl, kad jie generuojami remiantis statistiniais modeliais, kurie atkuria esmines pradinio duomenų rinkinio struktūrines savybes – skirstinius, priklausomybes tarp kintamųjų ir variaciją. Nors individualios stebėjimų reikšmės nėra tikros, tinkamai sukurti sintetiniai duomenys leidžia atlikti daugumą statistinių analizių ir gauti panašias išvadas kaip naudojant originalius duomenis.

Tarptautinėje praktikoje sintetiniai duomenys jau plačiai taikomi oficialiojoje statistikoje – administraciniams, medicininiams ir apklausų duomenų rinkiniams. Tuo tarpu Lietuvoje sintetinių duomenų taikymas dar yra ankstyvoje stadijoje ir daugiausia nagrinėjamas kaip perspektyvus metodas konfidencialumui užtikrinti.

DOI: https://doi.org/10.15388/LMITT.2026.27

Sintetiniams duomenims generuoti taikomi įvairūs metodai, tarp kurių dažnai naudojami statistiniais modeliais pagrįsti metodai. Vienas iš jų – grandininių lygčių metodas (angl. *Fully Conditional Specification*, FCS), kuris yra tinkamas mikroduomenų sintezei. Šis metodas naudojamas empiriniuose tyrimuose, pavyzdžiui, kategorinių duomenų sintezei 2021 m. Liuksemburgo gyventojų surašymo duomenų poaibyje [2].

Kitas oficialios statistikos institucijose naudojamas metodas – apibendrintieji adityvieji padėties, mastelio ir formos parametrų modeliai (angl. *Generalized Additive Models for Location, Scale and Shape*, GAMLSS). Šie modeliai leidžia išlaikyti svarbiausias duomenų statistines savybes ir priklausomybes tarp kintamųjų. Pavyzdžiui, šis metodas buvo taikytas Jungtinės Karalystės Nacionalinės statistikos tarnybos generuojant konfidencialų administracinių duomenų rinkinį [3].

Šio straipsnio tikslas – įvertinti ir palyginti FCS ir GAMLSS metodų taikymo galimybes sintetinant paslaugų sektoriaus įmonių duomenis. Toliau straipsnio antrajame skyriuje aprašomi tyrime naudoti duomenys ir jų paruošimo procedūros. Trečiajame skyriuje pristatomi taikyti sintetinių duomenų generavimo metodai bei pateikiami jų taikymo rezultatai, apimantys sintetinių ir originalių duomenų statistinį ir vizualinį palyginimą bei duomenų atskleidimo rizikos vertinimą. Ketvirtajame skyriuje pateikiamos tyrimo išvados.

2 Duomenys

Tyrime naudojami paslaugų sektoriaus juridinių asmenų ekonominės veiklos duomenys, surinkti vykdant Valstybės duomenų agentūros mėnesinį statistinį tyrimą [5]. Analizei pasirinktas duomenų poaibis, apimantis informaciją apie įmonių ekonominę veiklą, jų veiklos klasifikaciją, teisinę formą ir darbuotojų skaičių.

Pagrindinis tyrimo dėmesys skiriamas ekonominės veiklos rezultatą apibūdinančiam rodikliui – mėnesinėms pajamoms, kurios laikomos jautria informacija ir todėl buvo sintetinamos. Duomenų rinkinyje buvo ir keletas neigiamų mėnesinių pajamų reikšmių. Tokios reikšmės gali būti susijusios su nuostolinga veikla ar skolomis. Todėl neigiamos pajamos šiuo atveju interpretuojamos kaip nuostolio išraiškos.

Prieš analizę duomenys buvo paruošti: pašalinti įrašai su trūkstamomis reikšmėmis pasirinktų kintamųjų aibėje. Kategoriniai kintamieji paversti faktoriais, kad būtų tinkami naudoti statistiniuose modeliuose. Po duomenų paruošimo analizėje naudoti 3732 įrašai.

3 Metodai ir rezultatai

Tyrime duomenų sintetinimui buvo taikyti du skirtingo tipo metodai – neparametrinis ir parametrinis. Neparametriniam sintetinių duomenų generavimui buvo taikytas FCS (angl. *Fully Conditional Specification*) metodas, naudojant klasifikavimo ir regresijos medžių modelį (angl. *Classification and Regression Trees*, CART). FCS metodas grindžiamas nuosekliu kiekvieno kintamojo modeliavimu, sąlygojant jį kitais duomenų rinkinio kintamaisiais, o šis procesas kartojamas iteratyviai, kol pasiekiamas stabilumas.

CART modelis yra neparametrinis metodas, kuris nereikalauja prielaidų apie duomenų pasiskirstymą ir leidžia modeliuoti sudėtingus netiesinius ryšius tarp kintamųjų. Jis sudaro sprendimų medį, rekursyviai skaidydamas duomenis į homogeniškas grupes pagal kovariantes. Kiekviename paskutiniame medžio mazge susidaro stebėjimų grupė su panašiomis kovariančių reikšmėmis, o sintetinės reikšmės generuojamos atsitiktinai imant jas iš to mazgo empirinio pasiskirstymo. Tokiu būdu išlaikomi ryšiai tarp kintamųjų, tačiau neatkuriami konkretūs originalūs įrašai.

Parametriniam sintetinių duomenų generavimui buvo taikytas GAMLSS (angl. Generalized Additive Models for Location, Scale and Shape) metodas. Tai lankstus regresinis modeliavimo metodas, leidžiantis modeliuoti ne tik atsako kintamojo vidurkį, bet ir kitus pasiskirstymo parametrus – dispersiją, asimetriją ir ekscesą – kaip funkcijas nuo aiškinamųjų kintamųjų. Kadangi nagrinėjamo rodiklio pasiskirstymas buvo stipriai asimetriškas, prieš modeliavimą buvo išbandytos kelios transformacijos. Nustatyta, kad geriausią skirstinio prisitaikymą užtikrina logaritminė transformacija, kuri gali būti taikoma tik teigiamoms reikšmėms.

Atsižvelgiant į tai, iš analizės buvo pašalintos neigiamos pajamų reikšmės. Reikia pabrėžti, kad tokių stebėjimų buvo labai nedaug – kai kuriais mėnesiais jų visai nepasitaikė, o kituose jų skaičius svyravo nuo 1 iki 8, todėl jų pašalinimas neturėjo reikšmingos įtakos bendram duomenų pasiskirstymui.

Tyrime buvo išbandyti keli galimi pasiskirstymo modeliai, o galutiniam sintetinių reikšmių generavimui pasirinktas ST3 (*skew–t*) tipo pasiskirstymo modelis, leidžiantis modeliuoti asimetriškus ir sunkias uodegas turinčius pasiskirstymus. Modelio parametrai buvo įvertinti naudojant pradinius duomenis, o sintetinių reikšmių generavimas atliktas remiantis įvertintais pasiskirstymo parametrais.

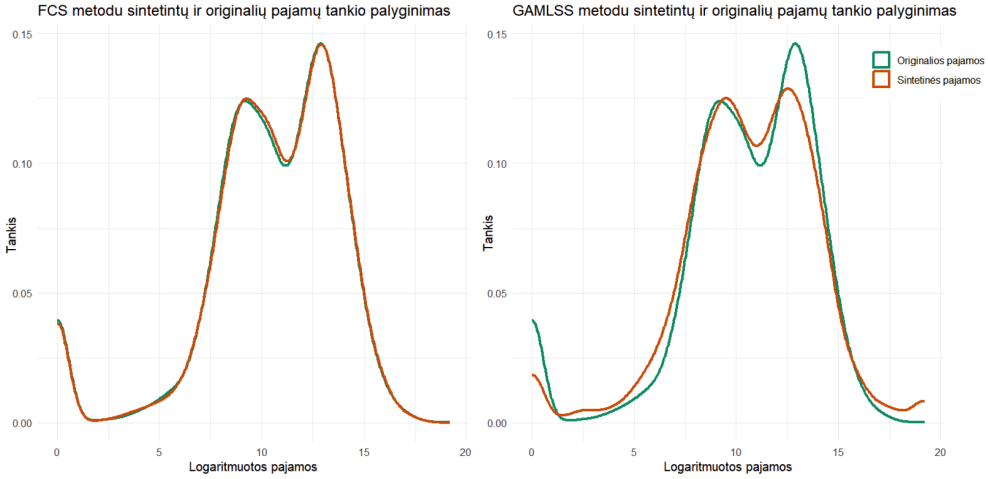
Sugeneravus sintetines reikšmes, buvo atliktas sintetinių ir originalių reikšmių palyginimas. Vertinimui naudoti pagrindiniai statistiniai rodikliai –

kvartiliai, mediana ir vidurkis – bei vizualiai analizuoti pajamų tankio pasiskirstymai. Pasiskirstymų panašumui įvertinti taip pat buvo taikytas Kolmogorovo–Smirnovo testas.

Siekiant įvertinti metodų naudingumą konfidencialumo užtikrinimo požiūriu, atlikta duomenų atskleidimo rizikos analizė. Vertinimas atliktas naudojant tapatybės atskleidimo ir atributo atskleidimo rodiklius.

Tapatybės atskleidimo rizika vertina tikimybę identifikuoti konkretų įrašą pagal kvaziidentifikatorių kombinaciją (repU). Šis rodiklis apskaičiuojamas nustatant, kiek įrašų turi identiškas kvaziidentifikatorių reikšmes, ir vertinant, kiek iš jų yra unikalūs, t.y. lengvai atsekami konkrečiam subjektui. Tuo tarpu atributo atskleidimo rizika vertina galimybę atspėti jautraus atributo reikšmę naudojant sintetinius duomenis (TCAP). TCAP rodiklis apskaičiuojamas vertinant, kokia tikimybė teisingai prognozuoti originalaus duomenų rinkinio įrašo jautraus atributo reikšmę, remiantis sintetinių duomenų sąlyginėmis tikimybėmis, esant toms pačioms kvaziidentifikatorių reikšmėms.

1 paveiksle pateikti logaritmuotų mėnesinių pajamų tankio pasiskirstymų palyginimai. Grafikai rodo, kad FCS metodu generuoti duomenys labai artimai atitinka originalių duomenų pasiskirstymą, o statistinių rodiklių reikšmės skiriasi minimaliai. Tuo tarpu GAMLSS metodu susintetintų duomenų pasiskirstymas vizualiai šiek tiek skiriasi nuo originalių reikšmių, pastebimi nedideli nukrypimai.



1 pav. Sintetinių ir originalių mėnesinių pajamų tankio pasiskirstymų palyginimas.

1 lentelėje pateikti Kolmogorovo–Smirnovo testo rezultatai. Abiem atvejais gautos p reikšmės viršijo pasirinktą reikšmingumo lygį, todėl statistiškai reikšmingo skirtumo tarp sintetinių ir originalių duomenų pasiskirstymų nenustatyta.

1 lentelė. Kolmogorovo–Smirnovo testo rezultatai.

Metodas	p reikšmė
FCS	0.9829
GAMLSS	0.0694

2 lentelėje pateikti tapatybės ir atributo atskleidimo rizikos rodikliai. Abiejų sintetinių duomenų rinkinių atveju tapatybės atskleidimo rodikliai sutampa – repU reikšmė siekia 18,06 %. Tai reiškia, kad maždaug 18 % įrašų yra unikalūs pagal naudojamus kvaziidentifikatorius ir teoriškai galėtų būti identifikuojami, jei būtų prieinama papildoma informacija.

Atributo atskleidimo rodikliai tarp metodų skiriasi. Neparametrinio metodo atveju TCAP rodiklis siekia 13,88 %, o parametrinio metodo atveju – 10,02 %. Šis rodiklis parodo tikimybę, kad naudojant kvaziidentifikatorių kombinaciją galima teisingai priskirti jautraus atributo reikšmę konkrečiam įrašui sintetiniame duomenų rinkinyje.

2 lentelė. Tapatybės atskleidimo ir atributo atskleidimo rizikos rodikliai.

Metodas	repU	TCAP
FCS	18.06 %	13.88 %
GAMLSS	18.06 %	10.02 %

4 Išvados

Tyrimo metu buvo pritaikyti dviejų skirtingų tipų duomenų sintetinimo metodai – neparametrinis FCS ir parametrinis GAMLSS. Rezultatai parodė, kad FCS metodu generuotų duomenų pasiskirstymas labiau atitiko originalių duomenų struktūrą nei GAMLSS metodu sugeneruotų duomenų pasiskirstymas.

Atskleidimo rizikos vertinimas parodė, kad tapatybės atskleidimo rizika abiem metodais iš esmės nesiskyrė, tačiau atributo atskleidimo rodiklis GAMLSS metodo atveju buvo mažesnis. Tai rodo, kad parametrinis metodas šiuo atveju gali būti laikomas saugesniu konfidencialumo užtikrinimo požiūriu, nors FCS metodas geriau išlaiko pradinių duomenų statistines savybes.

Literatūra

[1] Shlomo, N. (2025). Statistical disclosure control. Reference Module in Social Sciences.

[2] Basheer Kalash, C. L. (2025). Approaches to Synthetic Data Generation: Insights from Luxembourg's Census Data. Expert Meeting on Statistical Data Confidentiality.

[3] Jackson, J., Mitra, R., Francis, B., Dove, I. (2022). Using Saturated Count Models for User-Friendly Synthesis of Large Confidential Administrative Databases. Journal of the Royal Statistical Society: Series A (Statistics in Society), 185(4), 1613–1643.

[4] Raab, G. M., Nowok, B., Dibben, C. (2025). Practical privacy metrics for synthetic data.

[5] Valstybės duomenų agentūra. Paslaugų įmonių veiklos statistinio tyrimo metodika. Prieiga per internetą: https://osp.stat.gov.lt/documents/10180/687662/Paslaugu-imoniu-veiklos-metodika.pdf