Sünteetiliste andmete kasutamisest delikaatsete isikuandmete töötlemisel

Uuno Puus,
Cybernetica AS, Tartu Andmeturbelabor,
uuno.puus@cyber.ee

21. mai 2007. a.

1 Sissejuhatus

Infotehnoloogia järjest laienev kasutusele võtmine tekitab ka uusi probleeme. Üheks selliseks on konfidentsiaalsuse kadumine mitmesuguste delikaatseid isikuandmeid sisaldavate andmebaaside ristkasutusel. Ka Eestis on viimasel ajal käimas elav diskussioon sellel teemal. Ühelt poolt süüdistavad tervishoiuasutuste töötajad Andmekaitse Inspektsiooni selles, et delikaatsete isikuandmete kaitseks kehtestatud seadused jätavad nad ilma võimalusest analüüsida inimeste tervisekäitumist. Selleks olevat ilmtingimata vaja töödelda isikustatud delikaatseid isikuandmeid. Teisalt aga on selge, et kõigi soovijate varustamine isikustatud terviseandmetega viib varem või hiljem inimeste privaatsuse rikkumiseni. Üheks võimaluseks selle probleemi lahendamisel on tegelike delikaatsete andmete asendamine nn sünteetiliste andmetega, mille statistilised omadused (jaotus jms) on küll jäänud samaks, kuid mille puhul on oluliselt raskendatud või isegi võimatu seostada andmebaasis sisalduvaid andmeid konkreetsete isikutega.

Selgub, et sellised meetodid on juba suhteliselt pikka aega olemas. Sünteetiliste andmete genereerimisel lisatakse esialgsetele andmetele nn juhuslik müra või asendatakse delikaatsete andmete andmebaasis sisalduvad andmed esialgsetest andmetest sünteesitud näitajatega. Ühte sellist eksperimenti kirjeldataksegi Jerome P. Reiteri artiklis [1], mis on selle töö aluseks.

2 Andmete varjestamise võimalused

Reaalsete (tegelike) andmete varjestamiseks (delikaatsete isikuandmete varjamiseks nende kaitsmise eesmärgil) on tegelikult mitu võimalust. Üks lihtsamaid on sobilik kasutamiseks arvuliste andmete puhul. Olgu näiteks tarvis uurida palgaandmeid. Selleks, et vältida iga andmebaasis säilitatava isiku konkreetse palganumbri avalikuks tulekut võib palgaandmed salvestada järgmiselt. Andmebaasi kirjutatakse ainult "palk suurem kui 20 000"või "palk väiksem kui 20 000". Sel viisil pole võimalik kodeeritud andmete hulgast enam teada saada konkreetsete inimeste palganumbreid.

Selle idee edasiarendusena võib kogu palgaskaala jaotada samal viisil vahemikeks – näiteks "väiksem kui 8 000", "8 000 kuni 12 000", "12 000 kuni "jne.

Sel viisil andmeid muutes säilib andmetes sisuline informatsioon st palgagruppide lõikes on võimalik andmeid analüüsida, kuid täpseid palganumbreid on neist andmetest keeruline tuletada.

Siiski pole sel viisil võimalik kodeerida kõiki andmeid. Näiteks sel juhul kui tegemist on binaarse ehk kahendtunnusega pole kirjeldatud vahemikeks jaotusest abi. Kui delikaatsete isikuandmete analüüsil on vaja teada, kas tegemist on aidsihaigega või mitte, siis sel juhul tuleb kasutada teisi meetodeid. Üks variant on see, et muudetakse ära vastupidiseks näiteks 10% andmetest. Sel juhul on andmetes esinev viga 10% ehk just nõndapalju, kui seal esineb valesid andmeid, kuid eelisena on saavutatud see, et nende andmete põhjal pole võimalik kindlalt öelda, millistel andmebaasi kantud isikutel on AIDS ja millistel mitte. Hästi valitud müra lisamist konfidentsiaalsust vajavatele andmetele on kirjeldatud artiklis [2].

Näib, et sel viisil andmeid muutes oleme saanud keerulisele probleemile lihtsa lahenduse. Tegelikult pole see nõnda. Müra (10% valesid andmeid) lisamine teeb andmete töötlemise ja analüüsi tunduvalt keerulisemaks. Seetõttu pole võimalik saada rahuldavaid lahendusi sellisel lihtsal viisil. Selleks, et konfidentsiaalsuse saavutamise eesmärgil teadlikult muudetud andmete analüüsil saadud tulemusi saaks usaldada, tuleb andmete muutmiseks kasutada oluliselt keerukamaid meetodeid. Artiklis [1] on välja pakutud ka veidi keerukam, kuid tulemuse mõttes mõistlikum viis.

Kirjeldatud lähenemine koosneb kahest etapist – (1) statistiliste mudelite abil nendesse valimitesse tundmatute väärtuste asemele nn sünteetiliste andmete genereerimine ja (2) saadud andmestikust erinevate valimite välja eraldamine. Järgnevas alajaotuses ongi seda protseduuri täpsemalt kirjeldatud.

3 Sünteetiliste andmete genereerimine

Sünteetiliste andmete genereerimiseks tehakse lähteandmete failis järgmised muudatused. Olgu meil vektor $z = (z_1, \cdots, z_N)$ kus N on objektide arv populatsioonis. Vektor z näitab kas uuringuandmete hulka on valitud vastava (j-nda) objekti andmed (sel juhul $z_j = 1$) või mitte (sel juhul $z_j = 0$).

Olgu veel Y_{obs} $n \times p$ maatriks, mis koosneb nende objektide andmetest, mille puhul $z_j = 1$ ja Y_{nobs} $(N - n) \times p$ uuringusse mitte kaasatud andmete maatriks st nende puhul $z_j = 0$. Kogu andmete maatriks on siis avaldatav kujul $Y = (Y_{obs}, Y_{nobs})$

Olgu X $N \times d$ nn disainimuutujate maatriks, mis määrab objektide hulgas ära mingi klasterduse. See võib teada olla kas või ligilähedaseltki, näiteks kui kasutatakse rahvaloenduse andmeid vms.

Eelpool kirjeldatud andmetest sünteesitakse avalikuks kasutamiseks lubatud andmed järgmisel viisil. Lähteandmeteks on uuringu käigus kogutud andmed (X, Y_{obs}, z). Nendest andmetest lähtuvalt genereeritakse lähteandmete hulka mitte kuuluva $N - n$ objekti jaoks sünteetilised andmed. Sel viisil saadakse kogu populatsiooni hõlmav andmestik. Samuti võib seda sünteetiliste andmete genereerimist rakendada kogu populatsiooni kohta st genereerida sünteetilised andmed ka lähteandmetena kasutatud n objekti kohta.

Järgmisel etapil valitakse juhuslikult kogu populatsiooni sünteetiliste andmete hulgast m valimit, mis antakse avalikult uurijate käsutusse, eeldades, et nende m valimi põhjal pole enam võimalik otseselt teha järeldusi valimisse kuu-

luvate objektide kohta. Samuti eeldatakse, et selline protseduur säilitab andmestikus piisavalt sisulisi seoseid, et need andmed on kasutatavad vajalike uurimuste jaoks.

Kahes järgnevas alajaotuses vaatlemegi, kas eelpool nimetatud eeldused on täidetud st kas sünteetilised andmed säilitavad konfidentsiaalsuse ja on kasutatavad uuringute läbi viimiseks.

4 Konfidentsiaalsuse säilitamine

Konfidentsiaalsus võib saada rikutud kahel viisil – (1) objektide (re)identifitseerimine uuringuandmete põhjal ja (2) objektide kohta tundliku info saamine avalikustatud uuringuandmete põhjal. Esimesel juhul saadakse otseselt teada valimis sisalduvate objektide kohta tundlikku infot, teisel juhul võimaldab andmestik tuletada andmestikus sisalduvate objektide kohta tundlikku informatsiooni.

Sünteetiliste andmete puhul on identifitseerimise risk peaaegu olematu, kuna täielikult sünteetiliste andmete puhul pole nende põhjal kuidagi võimalik teada saada esialgses andmestikus kajastatud väärtusi. Pooleldi sünteetiliste andmete puhul (st juhul kui osa objektide andmeid pärineb lähteandmete hulgast) pole siiski täpselt teada, millised kirjed (objektide andmed) on sünteesitud ja millised mitte. See aga ei luba selle andmestiku põhjal teha järeldusi objektide tegelike andmete kohta.

Siiski mõningatel erijuhtudel on võimalik saada andmestikust tundlikku informatsiooni mõnede objektide kohta tuletamise teel. Näiteks juhul kui vanus on jaotatud eri vanuseklassideks viie aasta kaupa ja mõnes administratiivüksuses on mõnes vanuseklassis väga vähe esindajaid, siis on andmete analüüsimisel võimalik saada nende isikute kohta tundlikku informatsiooni. Selle olukorra tekkimise tõenäosust aitab vähendada sünteetilise müra lisamine andmetele. Selliseid meetodeid on kirjeldatud artiklis [2].

5 Sünteetiliste andmete kasutatavus

Nagu eelmises alajaotuses kirjeldatud, täielikult sünteetiliste andmete puhul on konfidentsiaalsusprobleem enamasti lahendatud. Lisaks konfidentsiaalsuse säilimisele saab sünteetiliste andmete kasutamisel veel esile tuua järgmised kolm eelist. (1) Sünteetilised andmed on populatsiooni hulgast välja valitud juhuvaliku põhimõttel ja seetõttu pole analüüsil vaja muret tunda valimi juhuslikkuse pärast. (2) Kuna andmed on eelnevalt töödeldud, siis sünteetiliste andmete põhjal tehtud järeldused on tõenäoliselt parema kvaliteediga, kui algsete andmete analüüsil saadavad tulemused. (3) Kuna andmed on simuleeritud, siis on võimalik andmestikus säilitada geograafilist asukohta näitavaid identifikaatoreid ja seeläbi on võimalik viia läbi analüüse ka väikestes piirkondades (mis teiste meetodite puhul on raskendatud või pole üldse võimalikud).

Kuid eelpool nimetatud eelistel on ka oma hind. Nimelt sõltub sünteetiliste andmete analüüsil saadud tulemus väga olulisel määral andmete sünteesimiseks kasutatatud võtetest ja meetoditest. Ehk teisisõnu sünteesitud andmete kasutamisel saab uurida vaid seda, mida juba andmete sünteesimise etapil on plaanitud uurida. Seega sõltuvad sünteesitud andmete kasutamisvõimalused olulisel

määral sellest, millisel viisil sünteesitud andmed on saadud. See aga kahandab oluliselt sünteesitud andmete kasutamisvõimalusi.

6 Kokkuvõte

Sünteetiliste genereeritud andmete kasutamine delikaatsete isikuandmete kasutamisel uuringutes võimaldab samaaegselt kasutada delikaatseid isikuandmeid uuringutes ja säilitada delikaatsete isikuandmete konfidentsiaalsust. Siiski pole neid mõlemaid omadusi võimalik ideaalselt rahuldada. Kirjeldatud meetodite praktilisel kasutamisel tuleb alati teha valik konfidentsiaalsuse ja kasutusmugavuse vahel. Kui konfidentsiaalsusnõue on hästi rahuldatud, siis nõuab andmete analüüs sageli harjumatu või eksootilise aparatuuri kasutamist. Samuti võib uuringutulemuste usaldusväärsus sel juhul kannatada.

Teisalt, kui andmed on kergesti analüüsitavad traditsioonilise statistika aparatuuri abil, on nende andmete hulgast võimalik hõlpsamini leida objektide kohta konfidentsiaalset informatsiooni. Ideaalis võiks siiski eelpool kirjeldatud meetodeid kasutada delikaatsete andmete varjestamiseks ning sel viisil lahendada Andmekaitse Inspektsiooni ja tervishoiu- ning ka teiste spetsialistide vahelist vastuolu. Sünteetiliste andmete kasutamisel oleks võimalik delikaatsete isikuandmete töötlemine ilma konfidentsiaalsusnõudeid rikkumata.

Käesoleval ajal on loomisel nn kodeerimskeskus, mis peaks tegelema sama probleemiga st võimaldama delikaatsete isikuandmete andmekaitsega kooskõlas olevat kasutust teadus- ja muudeks uuringuteks. Üks võimalus on lisada selle kodeerimiskeskuse funktsioonide hulka ka delikaatsete isikuandmete varjestamise metoodika välja töötamine ja seejärel ka vajalike (varjestatud) andmete väljastamine uurijatele vajalike uurimuste läbi viimiseks. Loomulikult kuuluks selle keskuse ülesannete hulka sel juhul ka vajalike uurimismetoodikate täiustamine ja uurijate koolitus.

Põhimõtteliselt võimaldaks sünteetiliste andmete tootmine lahendada delikaatsete isikuandmete statistilisel analüüsimisel kasutamise probleemi. Takistuseks on analüütikute vähene kogemus ja sünteetiliste andmete analüüsiks kasutatavate meetodite ebatraditsioonilisus ning mõningatel juhtudel ka komplitseeritus.

Viited

[1] Jerome P. Reiter, Releasing multiply imputed synthetic public use microdata: an illustration and empirical study, Journal of the Royal Statistical Society Series A, Vol. 168, Part 1, pp. 185-205, 2005.

[2] Fuller, W. A., Masking procedures for microdata disclosure limitation, Journal of Official Statistics, Vol. 9, pp. 383-406, 1993.