Mikroaineistojen tilastolliset tietosuojamenetelmät henkilötilastoissa

Janika Konnu

Tilastotieteen pro gradu -tutkielma

Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos
8. marraskuuta 2006

Tiivistelmä

Janika Konnu: *Mikroaineistojen tilastolliset tietosuojamenetelmät henkilötilastoissa*
Tilastotieteen pro gradu -tutkielma, Jyväskylän yliopisto. 8. marraskuuta, 2006.
Sivuja 167, liitteitä 3.

Tietosuojalla tarkoitetaan yleensä aineistojen suojaamisen lähtökohtana olevia lakeja, eettisiä normeja ja tietosuojamenetelmiä. Puhuttaessa tilastollisesta tietosuojasta tarkoitetaan nimenomaan tietojen suojaamiseen käytettäviä menetelmiä. Tämän tutkielman tarkoituksena on esitellä tietosuojamenetelmiä ja niiden käyttöä. Tutkielma kirjoitettiin tarkoituksellisesti suomeksi, jotta tilastollisista tietosuojamenetelmistä olisi käytettävissä ajantasaista äidinkielistä materiaalia.

Tutkielman alussa esitellään Suomen lainsäädäntöä ja Tilastokeskuksen eettisiä normeja, jotka vaikuttavan aineistojen luovuttamiseen. Muu tutkielma keskittyy tilastollisten tietosuojamenetelmien esittelyyn ja matemaattiseen määrittelyyn. Ensimmäisenä esitellään tietojen suojauksen pohjaksi tarvittavat termit paljastumisriski ja informaatiokato. Nämä ovat kaksi vastakkaista näkökulmaa, joiden arvojen optimoinnilla pyritään muodostamaan mahdollisimman hyvin suojattu mutta silti käyttökelpoinen aineisto. Sellaisia paljastumisriskin tai informaatiokadon mittoja, jotka sopisivat kaikkien menetelmien yhteyteen, ei ole vielä kehitetty, joten menetelmän yhteydessä esitellään tarvittaessa myös menetelmän tilanteeseen sopivat mitat.

Tutkielmassa on esitelty perusteet tilastollisten tietosuojamenetelmien käytölle. Tämän jälkeen määritellään joitakin yleisesti käytössä olevia ja uusimpia vasta kehitettyjä tietosuojamenetelmiä. Yleisemmin käytössä olevia menetelmiä esitellään lyhyesti, koska niiden voidaan olettaa olevan useimmille tuttuja. Uudemmat ja vähemmän käytetyt menetelmät, kuten kohinan lisääminen, mikroaggregointi, PRAM- ja MASSC-menetelmä määritellään tarkemmin.

Tutkielman lopuksi tarkastellaan kahden tietosuojamenetelmän toimintaa käytännössä. Mikroaggregointia ja PRAM-menetelmää sovelletaan suomalaisten opettajien muodostaman kokonaisaineiston suojaamiseen. Suojaaminen toteutetaan μ-Argus-ohjelman avulla. Empiirisen tutkimuksen tavoitteena on saada tietoa menetelmien käytettävyydestä. Lisäksi halutaan saada käsitys suojaamiseen sopivien suojausparametrien arvoista. Tutkielman johtopäätöksiä voidaan hyödyntää tilastollisten tietosuojamenetelmien käyttöönotossa ja niiden tarkemman tutkimisen pohjatietoina.

Avainsanat: tietosuoja, tilastollinen tietosuoja, aineiston suojaaminen, tietosuojamenetelmät, mikroaineisto, mikroaggregointi, PRAM-menetelmä, μ-Argus

Abstract

Janika Konnu: *Statistical Disclosure Control Methods for Personal Microdata*
Master's theses in statistics, University of Jyväskylä. November 8, 2006.
Pages 167, Appendices 3.

Statistical Disclosure Control (SDC) is a part of Statistical Database Protection. The former includes mainly the methodological part of protecting data from being abused when the latter definition includes also the legislation and ethics. The theses was prepared in purpose to introduce basic information and current best methods of the SDC. It was written in Finnish to provide useful material about SDC for Statistics Finland.

The theses begins with an introduction to Finnish legislation and ethical norms used in Statistics Finland. The following parts of the theses are focused in the disclosure control methodology. First disclosure risk and its counter-effect information loss were formulated. There are no universal ways to measure either of those and one has to give also method-specific measures for chosen methods.

When basic information about SDC is provided, the methods for disclosure control of microdata are introduced. There are many different kind of methods and the most common and some of the recently researched ones are chosen to be included. Methods like adding noise, microaggregation, PRAM and MASSC are defined more theoretically and in detail. More commonly known methods like sampling, global recoding, top and bottom coding and local supression are introduced with their basic ideas and theoretical considerations.

In the last parts of the theses two of the introduced methods are tested. The data to be protected is a census of the teachers in Finland. Program called μ-Argus was used to produce the safe data. These two SDC methods were studied with a goal of obtaining basic information about the usability of the method and to get some idea of the values of the parameters that produce the best results. The conclusion can be used in the further research of the SDC methods for microdata.

Keywords: Statistical Disclosure Control, Statistical Disclosure Limitation, data protection, privacy protection, protection methods, microdata, microaggregation, PRAM, μ-Argus

Sisällys

1 Johdanto

Tutkielman tarkoituksena on esitellä mikroaineistojen tietosuojamenetelmiä ja tarkastella joidenkin menetelmien käyttöä. Perinteisesti on totuttu kuulemaan tietoturvasta, jolla tarkoitetaan teknisiä keinoja estää tunkeutuminen sisäiseen verkkoon tai tiedostojen varastaminen verkon kautta. Tietoturvan menetelmiin kuuluvat palomuurit, käyttäjätunnukset ja muut vastaavat suojaukset tietojen suojelemiseksi ulkoiselta uhalta. Tietoturva ei kuitenkaan suojaa aineistoa, joka on luovutettava viraston ulkopuolisen henkilön käyttöön. Tähän tarkoitukseen tarvitaan tietosuojausta, joka keskittyy aineistojen suojaamiseen ei-toivottua tunnistamista vastaan. Tilastollisilla tietosuojamenetelmillä tarkoitetaan erilaisia aineiston muokkaustapoja, joilla alkuperäisen aineiston yksiköt saadaan tunnistamattomiksi ilman, että aineisto muuttuu käyttökelvottomaksi. Lain ja etiikan vaatima yksiköiden identiteetin suojeleminen ja mahdollisimman yksityiskohtainen ja oikeellinen aineisto ovat kaksi toisilleen vastakkaista näkemystä, jotka asettavat rajat tietosuojamenetelmien käytölle ja kehittämiselle.

Mikroaineistojen luovutus tutkimuskäyttöön ei ole yksinkertaista, koska erityisesti mikroaineistot ovat taulukkoaineistoja herkempiä tietosuojarikkomuksille. Paitsi laki myös eettiset normit rajoittavat kaikkea tiedonjakelua. Henkilötietolaissa luetellaan arkaluonteiset tiedot, joiden käsittely on ehdottomasti kielletty. Määräysten noudattamiseksi on ennen yksityiskohtaisten tietojen luovuttamista varmistettava, että yksiköiden identiteetti pysyy suojassa. Ongelmaa lisää internetin kautta laajalti saatavilla olevat tiedot, joita voidaan hyödyntää luovutetun aineiston yksiköiden tunnistamisessa. Henkilöllisyyden tai yrityksen paljastuminen on ikävää tiedonantajan kannalta ja heikentää lisäksi muiden tiedonantajien luottamusta tiedonkerääjää kohtaan. Mikäli tiedonantajat eivät voi luottaa siihen, että heidän identiteettinsä ja vastauksensa pysyvät tunnistamattomana, voivat he kieltäytyä kokonaan vastaamasta tai vastata kysymyksiin epärehellisesti. Molemmissa tapauksissa tiedonkerääjä kärsii tiedon laadun heikentymisestä.

Aiheen ajankohtaisuutta korostaa tietojen käsittelyn automatisoituminen ja tekniikan kehittyminen. Koska aineistot tallennetaan nykyisin sähköisesti, ovat ne helposti käytettäviä ja sisältävät yhä enemmän tietoa. Tutkijat haluavat käyttöönsä yhä useammin mikroaineistoja, koska he eivät saa haluamiaan tuloksia pelkkien tilastojen tai taulukoiden kautta. Paljastamista yrittävän henkilön käytössä on tietotekniikan kehityksen mukanaan tuomia menetelmiä, joilla aineistosta pystytään yhä helpommin ja varmemmin tunnistamaan haluttuja yksiköitä. Aineistojen luovuttaminen tutkimuskäyttöön on siis yhä yleisempää ja toisaalta yksikön paljastaminen aineistosta onnistuu yhä suuremmalla todennäköisyydellä. Tilanteeseen vastatakseen on tiedonkerääjän kehitettävä jatkuvasti uusia ja tehokkaampia tietosuojamenetelmiä. Tietosuojauksen tutkimus etenee juuri nyt yhä kiihtyvää tahtia sekä USA:ssa että Euroopassa ja tämä tutkielma antaa yleiskuvan kehityksen tämän hetkiseen tilanteeseen.

Tutkielman sisältö on seuraava: tutkimustehtävät esitellään luvussa 2 ja tietojen julkaisemiseen liittyvää lainsäädäntöä ja eettisiä normeja käsitellään kolmannessa luvussa. Luku 4 pitää sisällään tiedon paljastumisriskin ja suojatun aineiston hyödynnettävyyden tarkastelun. Tietosuojamenetelmiin liittyviä määritelmiä ja menetelmien jaottelua käsitellään luvussa 5. Varsinaisiin tietosuojamenetelmiin tutustutaan luvuissa 6 ja 7, joista ensimmäisessä käsitellään aineiston rajoittamiseen perustuvia menetelmiä ja jälkimmäisessä aineiston muuntamiseen perustuvia menetelmiä. Kah-

deksannessa luvussa annetaan käytännön esimerkki menetelmistä ja yhdeksäs luku sisältää teoreettisen tarkastelun ja käytännön esimerkin perusteella saadut johtopäätökset. Tutkielman ymmärtämiseen voi hyödyntää kymmenennen luvun sanastoa, jossa annetaan suomenkielisten termien selitykset ja termien englanninkieliset vastineet kirjallisuuteen tutustumisen helpottamiseksi.

2 Tutkimustehtävä

Tutkielmassa esitellään ja vertaillaan uusimpia, laajimmin tutkittuja ja käytössä olevia mikroaineistojen tietosuojamenetelmiä. Menetelmät jaotellaan niiden vaikutustavan mukaan aineistoa rajoittaviin ja aineistoa muokkaaviin menetelmiin. Menetelmien esittelyn yhteydessä annetaan yleisten aineistoon kohdistuvien vaikutusten lisäksi tietoa menetelmien matemaattisesta perustasta. Menetelmien vaikutusta aineistoon tarkastellaan lähinnä frekvenssien muutosten ja esimerkkinä olevan mallin avulla. Paljastumisriskin ja informaatiokadon mittoja esitellään omassa kappaleessaan yleisesti, mutta myös menetelmien yhteydessä esitellään kyseiseen menetelmään sopivia mittoja. Näitä voi hyödyntää suojattujen aineistojen käytettävyyden ja suojauksen tason arvioimisessa.

Osaa tutkielmassa esitellyistä menetelmistä sovelletaan käytäntöön. Sovellusta varten on käytössä aineisto, joka suojataan kokeilumielessä. Kyseinen aineisto on kokonaisaineisto Suomen lukioiden opettajista. Koko aineisto on mukana suojauksessa, vaikka Tilastokeskuksesta ei todellisuudessa luovuteta kokonaisaineistoja. Aineisto sisältää kuitenkin niin vähän muuttujia, ettei aineisto olisi sellaisenaan käytettävissä tutkimukseen. Aineiston avulla onkin tarkoitus saada muutamia yksinkertaisia tuloksia testattavien tietosuojamenetelmien toimivuudesta. Sovelluksen tarkoituksena on testata menetelmien käyttöä käytännössä ja niiden soveltuvuutta erityisesti Tilastokeskuksen aineistojen suojaamiseen. Sovelluksen yhteydessä tarkastellaan menetelmien ominaisuuksia ja niiden eroja ja havaintoja hyödynnetään menetelmiä vertailtaessa.

Tietosuojamenetelmiä vertaillaan ensin niiden teoreettisen perustan pohjalta. Tarkastelussa on erityisesti menetelmien vaikutus aineiston käytettävyyteen sekä paljastumisriskiin. Vaikka varsinaisia mitan arvoja ei kummallekaan ominaisuudelle lasketa, saadaan suojauksen tasosta käsitys tarkastelemalla suojatun ja alkuperäisen aineiston eroja. Samalla tarkastellaan menetelmän käyttöön vaadittavia taustatietoja ja suojaamiselle optimaalisia asetuksia.

Tavoitteena on, että tutkielman sisältämän tiedon ja kokemusten pohjalta voidaan jatkaa mikroaineistojen tietosuojamenetelmien tutkimista ja perehdyttää aineistojen tutkimuskäyttöön luovuttamisen kanssa tekemisissä olevia tietosuojaperiaatteisiin ja suojausmenetelmiin. Lisäksi saadaan käsitystä siitä, ovatko Tilastokeskuksesta luovutettavien mikroaineistojen suojaamiseen käytetyt menetelmät parhaimpia mahdollisia vai tulisiko tilannetta tutkia tarkemmin.

3 Lainsäädännöllinen sekä tilasto- ja tutkimuseettinen perusta

Tilastokeskuksen toimintaa säätelee Euroopan Yhteisöjen (myöhemmin EY) ja Suomen lainsäädännön lisäksi tilastotoimen eettiset ohjeet, kuten esimerkiksi Tilastokeskuksen omat ammattieettiset periaatteet. Lakien tavoitteena on taata yhteiskunnan tiedonsaanti ja toisaalta turvata henkilöiden yksityisyydensuoja sekä elinkeinonharjoittajien liike- ja ammattisalaisuuden suoja. Tilastokeskus keskittyy omilla eettisillä periaatteillaan (Tilastokeskus, 2002a) varmistamaan tiedonantajien luottamuksen. Tilastokeskuksen eettiset periaatteet perustuvat YK:n virallisten tilastojen periaatteisiin ja Kansainvälisen tilastoinstituutin ammattieettisiin suosituksiin. Luottamus Tilastokeskuksen toimintaan tietojen julkaisijana ja varsinkin tietojen kerääjänä ja käsittelijänä on tärkeää, jotta tiedonantajat pystyvät huoletta luovuttamaan tietonsa Tilastokeskuksen käyttöön. Mikäli luottamus horjuu, on todennäköisempää, ettei kerättäviä tietoja saada tai annetut tiedot ovat virheellisiä ja siten epäluotettavia.

3.1 EU-lainsäädäntö

Suomen tilastotoimintaa säätelee paitsi kansallinen lainsäädäntö myös Euroopan Yhteisöjen säädökset. Tärkeimmät periaatteet tilastojen laatimista ja käsittelyä varten sisältyvät Neuvoston asetukseen (EY) N:o 322/97. Asetuksen asettamisen perusteena on yhteisön tilastojen tuottamiseen tarvittavien tietojen suojaaminen. Lisäksi Euroopan laajuisella asetuksella pyritään luomaan samat periaatteet tietojen suojaamiselle koko yhteisön alueella. Asetuksessa korostetaan erityisesti luottamusta tietojen keräämistä ajatellen, jolloin tavoitteena on yhteisymmärrys tiedonantajien kanssa. Vastaavasti korostetaan myös luottamusta tiedon julkaisemisen kannalta, jolloin tavoitteena on yhteiskunnan luottamus julkaistuja tietoja ja niiden perusteella tehtyjä päätöksiä kohtaan.

Asetuksen tärkeimmät säädökset tietosuojan kannalta sisältyvät kolmanteen lukuun: Periaatteet. Kymmenennessä artiklassa luetellaan: ”Yhteisön tilastoja hallitsevat periaatteet ovat puolueettomuus, luotettavuus, relevanssi, kustannustehokkuus, tilastosalaisuus ja avoimuus, joilla taataan sekä deontologiselta että ammatilliselta kannalta mahdollisimman hyvä laatu”. Näistä periaatteista kukin on määritelty tarkemminkin, mutta olennaisinta tietosuojamenetelmien kannalta on tilastosalaisuus, jolla ”tarkoitetaan suoraan tilastollisiin tarkoituksiin tai välillisesti hallinnollisista tai muista lähteistä saatujen yksittäisiin tilastoyksiköihin liittyvien tietojen suojaamista tietosuojarikkomuksia vastaan. Se edellyttää, että estetään saatujen tietojen käyttö muihin kuin tilastollisiin tarkoituksiin sekä tietojen laiton paljastaminen”.

Asetuksen viidennen luvun ”Tilastosalaisuus” 13. artiklan 1. momentissa määrätään, että ”kansallisten viranomaisten ja yhteisön viranomaisen yhteisön tilastojen tuottamiseen käyttämiä tietoja on pidettävä luottamuksellisina, jos niistä voi suoraan tai välillisesti tunnistaa tilastoyksiköt ja ne näin paljastavat yksilötietoja. Tilastoyksikön tunnistettavuutta määritettäessä on otettava huomioon kaikki keinot, joita kolmas osapuoli saattaa kohtuudella käyttää tunnistaakseen kyseisen tilastoyksikön.” Lisäksi 15. artikla määrää, että ”yksinomaan yhteisön tilastojen tuottamiseen saatua luottamuksellista tietoa voivat kansalliset viranomaiset ja yhteisön viranomainen käyttää ainoastaan tilastotarkoituksiin, jolleivät tietojen luovuttajat ole yksiselitteisesti antaneet suostumustaan sen käyttämiseen muihin tarkoituksiin.” Kuitenkin edellises-

sä (14.) artiklassa sanotaan: ”sellaisia luottamuksellisia tietoja, joita ei voi suoraan tunnistaa, voidaan luovuttaa kansallisten viranomaisten kesken sekä kansallisten viranomaisten ja yhteisön viranomaisen kesken siinä laajuudessa, kuin luovuttaminen on välttämätöntä yhteisön erityisten tilastojen tuottamiseksi. Tietojen luovuttamiseen edelleen on oltava tiedot keränneen kansallisen viranomaisen nimenomainen lupa.”

Näiden säädösten perusteella päädytään siihen, että tilastotarkoituksiin kerätyn tiedon luovuttaminen tunnistetietoineen on kielletty paitsi julkisuuteen myös toisille viranomaisille. Lisäksi luovutettavien tietojen tunnistamattomuutta tulee tarkastella paitsi identifioivien muuttujien poistamisen, myös julkaistavien tietojen yhdistettävyyden kannalta. Toisin sanoen on oltava tarkka siitä, onko julkaistavan aineiston tietoja mahdollista hyödyntää muiden aineistojen avulla. Säädöksien nojalla Tilastokeskus on siis velvollinen varmistamaan, ettei sen julkaisemista tiedoista voida julkisiin tai itse kerättyihin tietoihin yhdistämällä tunnistaa tietojen kohteita suoraan tai välillisesti.

3.2 Suomen lainsäädäntö

Kansallisessa lainsäädännössä on useita tilastotointa koskevia lakeja, jotka vaikuttavat osaltaan tilastoalan tietosuojakysymyksiin. Tärkeimpänä näistä on luonnollisesti tilastolaki (280/2004), joka määrittää koko kansallisen tilastotoimen säädökset. Tilastokeskuksessa on käytössä Tilastolain soveltamisohje (Tilastokeskus, 2005a), jossa lain tarkoitus on kirjattu selkeästi esiin. Erityisesti tietosuojakysymyksissä tärkeiksi nousevat henkilötietolaki (523/1999) sekä laki viranomaisten toiminnan julkisuudesta (621/1999). Seuraavassa käsitellään näiden lakien sisältöä ja vaikutusta aineistojen luovuttamiseen.

Laki Tilastokeskuksesta (48/1992) määrittelee Tilastokeskuksen ensisijaiseksi tehtäväksi ”...laatia yhteiskuntaoloja koskevia tilastoja ja selvityksiä sekä huolehtia valtion tilastotoimen yleisestä kehittämisestä yhteistyössä muiden valtion viranomaisten kanssa.” Laissa määrätään, että Tilastokeskuksen on huolehdittava sekä kansallisesta että kansainvälisestä tiedontarpeesta sekä toimittava muiden mahdollisesti määrättävien tehtävien mukaan.

Ensisijaisesti Tilastokeskus on kerännyt tietoja kansallista päätöksentekoa varten. Tilastokeskuksen aineiston hankinnasta määrätään tilastolaissa (4 ja 5§): tiedonantajilta ei tule kerätä muuta kuin ”tilastojen laatimisen kannalta välttämättömät tiedot” ja nekin ”taloudellisesti ja niin, että siitä aiheutuu tiedonantajille mahdollisimman vähän haittaa ja kustannuksia”. Tämä ehto toteutuu parhaiten, jos noudatetaan lain 4§:n ohjetta ”...ensi sijassa käyttää hyväksi julkishallinnon tehtävien hoitamisessa kertyneitä sekä elinkeinon- ja ammatinharjoittajien, yhteisöjen ja säätiöiden tavanomaisen toiminnan seurauksena syntyneitä tietoja.”

Tiedonantajien yksityisyyden suojaamiseksi tilastolain 5§ määrää: ”Tiedot tulee kerätä ja tallettaa ilman tunnistetietoja aina, kun se tilastojen laatimisen kannalta on mahdollista. Tunnistetietoja voidaan kerätä ja tallettaa ainoastaan silloin, kun se on välttämätöntä tietoaineistojen yhdistämiseksi tai kun se on muutoin välttämätöntä yhteiskuntaolojen kehityspiirteitä kuvaavien luotettavien ja vertailukelpoisten tilastojen tuottamiseksi.” Toisin sanoen yksilön henkilökohtaisia tietoja tulee lain mukaan suojella jo tiedonkeruusta asti ja jo edellä mainittiin, ettei epäolennaisia tietoja saa

edes kerätä. Kuitenkin käytännössä suurin osa Tilastokeskuksen aineistosta tallennetaan tunnistetietojen kanssa, koska tietoja tullaan tilastointia tai tutkimusta varten yhdistämään. Itse "tilastot tulee laatia niin, etteivät niistä ole suoraan tai välillisesti tunnistettavissa ne, joita tilastot koskevat, ellei tunnistamista koskeva tieto ole tämän lain mukaan julkinen" (tilastolaki, 11§).

Tilastokeskuksen toimintaa tarkastellessa tulee lisäksi ottaa huomioon Henkilötietolain (523/1999) 11 §, joka yksikäsitteisesti kieltää tiettyjen arkaluonteisten tietojen käsittelyn:

"Arkaluonteisten henkilötietojen käsittely on kielletty. Arkaluonteisina tietoina pidetään henkilötietoja, jotka kuvaavat tai on tarkoitettu kuvaamaan:
1) rotua tai etnistä alkuperää;
2) henkilön yhteiskunnallista, poliittista tai uskonnollista vakaumusta tai ammattiliittoon kuulumista;
3) rikollista tekoa, rangaistusta tai muuta rikoksen seuraamusta;
4) henkilön terveydentilaa, sairautta tai vammaisuutta taikka häneen kohdistettuja hoitotoimenpiteitä tai niihin verrattavia toimia;
5) henkilön seksuaalista suuntautumista tai käyttäytymistä; taikka
6) henkilön sosiaalihuollon tarvetta tai hänen saamiaan sosiaalihuollon palveluja, tukitoimia ja muita sosiaalihuollon etuuksia."

Näinkin yksiselitteiseen kieltoon on säädetty poikkeuksia, joiden mukaan arkaluonteisten tietojen käsittely sallitaan muun muassa "... historiallista tai tieteellistä tutkimusta taikka tilastointia varten" (Henkilötietolaki 12§ 6). Tilastokeskuksella on lupa käsitellä ja tallentaa tätä arkaluonteista materiaalia. Vaikka samassa kohdassa mainitaan myös tieteellinen tutkimus, ei tämä tarkoita sitä, että Tilastokeskuksella olisi automaattisesti oikeus luovuttaa tietojaan tieteellistä tutkimusta varten. Laissa viranomaisten toiminnan julkisuudesta (281/2004 §24 16), jossa luetellaan lain mukaan salassa pidettäviä asiakirjoja määrätään nimittäin: "Salassa pidettäviä viranomaisen asiakirjoja ovat, jollei erikseen toisin säädetä:... tilastoviranomaiselle tilastojen laatimista varten annetut asiakirjat samoin kuin asiakirjat, jotka on vapaaehtoisesti annettu viranomaiselle tutkimusta tai tilastointia varten". Näistä edelleen poikkeuksina ovat Tilastokeskuksen hallussa olevat valtion ja kunnallisten viranomaisten toimintaa ja julkisten palvelujen tuottamista kuvaavat sekä tilastolain 18 §:ssä tarkoitetut yrityksiä ja yhteisöjä koskevat tiedot, kuten toimipaikat ja muut sellaiset.

Kuten edellä olleesta voi päätellä, Tilastokeskus saa käyttöönsä valtavat määrät tietoa, joka toisaalta kiinnostaisi tutkijoita, mutta toisaalta on sellaista, ettei tiedonantajilla ole ollut valtuuksia kieltää sen keräämistä. Tutkijat haluavat käyttöönsä yhä tarkempaa yksikkötasoista tietoa ja se on Tilastokeskuksessa valmiiksi kerättynä. Tästä syystä Suomen lainsäädäntöön on sisällytetty tarkat ohjeet siitä, miten ja millaisessa muodossa tilastotarkoituksiin kerättyä tietoa saa käsitellä tai luovuttaa kolmannelle osapuolelle. Lain mukaan Tilastokeskus voi luovuttaa tietoja tutkijoille lain määräämin edellytyksin.

Tiedonantovelvollisuuden määrittelyn yhteydessä mainitaan useasti "tilastojen laatimisen kannalta välttämättömät tiedot" (Tilastolaki, 5. luku). Tälläkin viitataan siihen, ettei Tilastokeskuksella ole oikeutta hankkia tilastojen laatimisen yhteydessä ylimääräisiä tietoja. Myöskään turvallisuuden tai maanpuolustuksen kannalta salassa pidettäviä tietoja ei saa kerätä. Kerättäessä tietoja tilastotarkoituksiin tulee muistaa,

että ”tilastoja laativan viranomaisen on tietoja kerättäessä kirjallisesti selostettava tiedonantajille tietojen käyttötarkoitus, tilastojen laatimisessa käytettävät menettelytavat, tiedonantovelvollisuus perusteluineen tai se, että tietojen antaminen on vapaaehtoista, tiedonantajan oikeudet, tietojen suojaamisjärjestelyt ja säilytysajat sekä muut tarpeelliset tietojen antamiseen vaikuttavat seikat”. (Tilastolaki, 9§) Näiden tietojen ilmoittamisesta seuraa, ettei tietoja voi välttämättä luovuttaa eteenpäin, mikäli niiden luovuttamista koskevista kysymyksistä ei ole asianmukaisesti informoitu tiedonantajaa tai hänen edustajaansa. Käytännössä Tilastokeskus pyrkii informoimaan aineiston useammista käyttötarkoituksista siten, että tiedonkeruun tuloksena saavien aineistojen luovuttaminen tutkimuskäyttöön joudutaan vain harvoin kieltämään. Kuitenkin aineiston luovuttamisen ongelmat voivat olla seurausta tiedonkeruuvaiheesta ja siksi tiedonkeruu onkin suunniteltava tarkasti.

Laissa on ohjeet myös tiedon säilyttämiseen ja käsittelyyn. Tilastotarkoituksia varten kerättyjen tietojen käsittelyn kaikissa vaiheissa on varmistettava, ”ettei kenenkään yksityisyys tai liike- tai ammattisalaisuus vaarannu. Tietoja on käsiteltävä hyvää tilastotapaa noudattaen ja tilastoalalla yleensä sovellettavien kansainvälisten suositusten ja menettelytapojen mukaisesti”(tilastolaki 10§). Tämän lisäksi tietoja käsitellessä on huolehdittava asianmukaisesta tietojen säilyttämisestä. Se, mihin suojauksella viitataan on käsitelty yksityiskohtaisesti Asetuksessa viranomaisten toiminnan julkisuudesta ja hyvästä tiedonhallintatavasta (1030/1999). Asetuksen 3 §:ssä kerrotaan tarkasti huolehdittavista yksityiskohdista, kuten siitä millaisessa paikassa aineistoja saa säilyttää:
” 3) tietoaineistoja käyttävät, muuttavat ja muutoin käsittelevät vain ne, joiden tehtäviin asian käsittely kuuluu ja että käyttöoikeudet rajataan muutoinkin asianmukaisesti ja käyttöä valvotaan riittävästi;
4) tietoja aineistosta luovuttavat vain ne, joiden tehtäväksi siitä huolehtiminen kuuluu;
5) tietoverkoissa siirrettävä tieto salataan tarpeen mukaan.”
Lisäksi on huolehdittava aineistojen luovuttamista koskevasta päätäntävallasta. Tilastokeskuksen tutkimuskäyttöön anottuihin yksikkötason aineistoihin käyttöluvan myöntää aineistoja hallinnoivan tilastoyksikön tilastojohtaja. Vaikeammissa tapauksissa asiat käsitellään ennen päätöstä tilastoeettisessä lautakunnassa. Mikäli aineisto on anottu käyttöön ulkomaille, tekee päätöksen pääjohtaja tilastoeettisen lautakunnan esityksen pohjalta.

Olennaisin ero luonnollisten henkilöiden kannalta Tilastokeskuksen ja muiden virastojen tietojen suhteen on, ettei Tilastokeskuksen tai muunkaan tilastoviranomaisen tarvitse antaa tietoja edes henkilöä itseään koskevasta asiakirjasta (henkilötietolaki, 27§ 3. kohta). Myöskään muut viranomaiset eivät voi tilastotarkoituksiin kerättyä tietoa hyödyntää, ellei tieto ole alunperin kyseisen viranomaisen luovuttamaa. Tilastotarkoituksiin kerättyyn aineistoon ei voi soveltaa muita kuin erityisesti sitä varten säädettyjä säädöksiä. Tarkemmin tilastotiedon erityisasemasta kertoo: ”Näitä tietoja ei saa luovuttaa käytettäväksi tutkinnassa, valvonnassa, oikeudenkäynnissä, hallinnollisessa päätöksenteossa tai muussa vastaavassa henkilöä, yritystä, yhteisöä tai säätiötä koskevan asian käsittelyssä.” (tilastolaki, 12–13§) Tämä on hyvin erikoinen tilanne sinällään, ettei edes oikeudenkäyntiä varten voida luovuttaa tietoja, jotka on tilastojen laadintaa varten saatu. Lain rajoite perustuu siihen, että tilastointia varten saatavan aineiston on oltava mahdollisimman luotettavaa ja näillä rajoitteilla pyritään varmistamaan tiedonantajien luottamus tilastoja laativia viranomaisia kohtaan.

Mihin Tilastokeskuksen käytössä olevia aineistoja sitten voi luovuttaa? Kysymykseen vastaa tilastolain 13 §:n toinen momentti: ”Tilastoviranomainen voi luovuttaa tilastotarkoituksiin keräämiään salassa pidettäviä tietoja yhteiskuntaoloja koskevia tieteellisiä tutkimuksia ja tilastollisia selvityksiä varten.” Edelleen tulee muistaa se, mitä jo tiedon keruun ja tallennuksen suhteen määrättiin eli ”Henkilötietolaissa tarkoitettuja henkilötietoja ja muiden tilastoyksiköiden tunnistetietoja ei kuitenkaan saa luovuttaa. Välttämättömät tunnistetiedot voidaan kuitenkin luovuttaa toiselle tilastoviranomaiselle sen toimialaan kuuluvien tilastojen laatimista varten” (tilastolaki, 13§).

Henkilötiedoista ”tiedot iästä, sukupuolesta, koulutuksesta ja ammatista” voidaan luovuttaa tieteellistä tai tilastollista tutkimusta varten ”edellyttäen, että tietojen saajalla on henkilötietolain mukainen oikeus käsitellä näitä tietoja” (Tilastolaki, 19§). Henkilötietolain 13§:n mukaan ”Henkilötunnusta saa käsitellä rekisteröidyn yksiselitteisesti antamalla suostumuksella tai, jos käsittelystä säädetään laissa. Lisäksi henkilötunnusta saa käsitellä, jos rekisteröidyn yksiselitteinen yksilöiminen on tärkeää esimerkiksi ”...historiallista tai tieteellistä tutkimusta taikka tilastointia varten” tai jos kaikkien rekisteröityjen suostumusta ei ole mahdollista hankkia. Tällaisten henkilötietojen käyttö tutkimustarkoituksiin vaatii myös asianmukaisen tutkimussuunnitelman, tutkimuksen vastuullisen johtajan sekä takuun siitä, että tietojen paljastuminen ulkopuolisille estetään. Ehdot sisältävät vastuun rekisterin hävittämisestä tai arkistoimisesta kun henkilötietoja ei ole enää välttämätöntä säilyttää. Henkilötietolaissa otetaan huomioon myös tietojen ikä ja laatu sekä tietojen kohteen suostumus yms. Mikäli yksityisyyden suojaaminen on ilmeisen tarpeetonta näiden syiden takia, ei siihen kiinnitetä turhaan huomiota.

Toisin sanoen henkilöitä koskevien tietojen käyttö tutkimustarkoituksiin on mahdollista jopa tunnistetietoineen, mutta tällöin aineistolta tai sen vastaanottajalta vaaditaan monia ominaisuuksia. Yleisemmin pätee se edellä mainittu näkökulma, ettei Tilastokeskus luovuta henkilötietoja tunnistetietoineen.

Luvan tilastotarkoituksiin kerättyjen tietojen luovuttamiseen antaa tilastoviranomainen eli tässä tapauksessa Tilastokeskus. Jopa silloin, kun tiedot on erikseen kerätty jonkun viranomaisen toimesta ja luovutettu sitten tilastoja varten, luovuttaminen tapahtuu tilastoviranomaisen päätöksellä, mutta vain sillä ehdolla, ”ettei tiedon antaminen loukkaa niitä etuja, joiden suojaksi salassapitovelvollisuus on säädetty”. Mikäli tiedot on kerätty tiedonantajan erillisellä suostumuksella, ”lupaa ei saa antaa vastoin suostumuksessa tiedon käytölle ja luovutukselle asetettuja ehtoja”. (Laki viranomaisten toiminnan julkisuudesta, 28§)

Osan tietosuojariskiä muodostavat myös ulkopuolisten palvelujen tarjoajat. Tilastolain 22§ antaa selvän säännön tähänkin tilanteeseen: ”Tilastoja laativa viranomainen voi antaa toimeksiantona suoritettavaksi tilastotuotannon eri vaiheisiin liittyviä osatehtäviä tai tukitehtäviä. Toimeksiantojen käyttöä ja ehtoja harkittaessa ja valvontaa järjestettäessä on erityisesti otettava huomioon tietosuojanäkökohdat.” Yleisesti tilanteita, joissa ulkopuolisia palveluja nimenomaan tietoaineistojen käsittelyyn tarvitaan, pyritään välttämään juuri hankalan valvonnan takia.

Henkilötietolain nojalla ”henkilötietoja voidaan siirtää Euroopan unionin jäsenvaltioiden alueen tai Euroopan talousalueen ulkopuolelle ainoastaan, jos kyseisessä maassa taataan tietosuojan riittävä taso.” Lisäksi ”tietosuojan tason riittävyys on arvioita-

va ottaen huomioon tietojen luonne, suunnitellun käsittelyn tarkoitus ja kestoaika, alkuperämaa ja lopullinen kohde, asianomaisessa maassa voimassa olevat yleiset ja alakohtaiset oikeussäännöt sekä käytännesäännöt ja noudatettavat turvatoimet."

Tilastotarkoituksiin kerättyjen aineistojen suojaamista ja luovutusta koskevissa artikkeleissa mainitaan usein rangaistuksen pelote. Artikkelien kirjoittajat uskovat, että varsinkin tutkijoille maineenmenetys ja muu mahdollinen rangaistus toimii riittävänä suojauksena. He eivät pidä todennäköisenä, että kukaan tutkija haluaisi edes korvausta vastaan yrittää paljastusta seuraamusten pelossa. Suomen lainsäädännössä säädetään paljastamiseen liittyvistä rangaistuksista useassa laissa. Rangaistus riippuu myös siitä, onko kyseessä virkamies, joka rikkoo virkasalaisuutta vai onko kyseessä muu tunkeutuja.

Virkamiesasema on rikoslaissa (39/1889) erityisesti korostettu ja siksi virkasalaisuuden rikkomisesta määrätään muita henkilöitä ankarammat rangaistukset (lain 40 luku 5 §). Virkamiehen paljastamiseen johtaneen teon osoituttua tahalliseksi voidaan virkamies tuomita jopa vankeusrangaistukseen ja viralta pantavaksi. Rangaistusta päätettäessä tahattomuus on lieventävä asia.

Jos joku muu kuin viranomainen rikkoo tilastosalaisuutta määräytyy rangaistus eri rikosnimikkeiden mukaan riippuen siitä, millainen rikkomus on kyseessä (Tilastolaki, 12 ja 13§). Tällaisia rikosnimikkeitä on salassapitovelvollisuuden rikkominen, joka viittaa tehtävien kautta saatujen tietojen paljastamiseen ja rangaistuksena voi olla jopa vankeutta ja tämän lievempi muoto, tilastosalaisuusrikkomus, jossa teon seuraukset ovat vähäisiä ja rangaistuksena on korkeintaan sakkoa. Henkilörekisteririkos on henkilötietojen käsittelyä perusteettomasti tai aiheuttaen rekisteröidylle haittaa. Tästäkin rikoksesta voidaan tuomita vankeuteen. Lievemmän rikkeen tai laiminlyönnin seurauksena henkilö voidaan tuomita henkilörekisteririkkomuksesta sakkoon.

3.3 Eettiset näkökohdat

Ennen käyttöluvan myöntämistä mihinkään luottamukselliseen aineistoon, tulee ottaa huomioon Tilastokeskuksen tietoturvaperiaatteissa määritelty koko käsittelyn läpi jatkuva luottamuksellisuuden turvaaminen. Tämä "ei saa kuitenkaan estää tai merkittävästi vaikeuttaa Tilastokeskuksen oman henkilökunnan tai sen sidosryhmien asianmukaista toimintaa. Tietoturvallisuuden keskeisenä tavoitteena tulisi olla resurssien ja palvelujen helppo ja tehokas käyttö myönnettyjen käyttöoikeuksien puitteissa" (Tilastokeskus, 2002b).

Kansainvälinen tilastoinstituutti (myöhemmin ISI) on julkaissut vuonna 1985 ammattieettisen julistuksen, joka ohjaa myös Tilastokeskuksen työtä. Tilastokeskuksen oma eettinen ohjeisto: Toimi oikein tilastoalalla – Tilastokeskuksen ammattieettinen opas (Käsikirjoja 30), pohjautuu näihin periaatteisiin.

Tilastokeskuksen toiminnan eettisyyttä tarkastellessa voidaan ajatella yhden ISI:n julistuksesta (1989) seuraavan kohdan selittävän tilanteen kaikkein selvimmin. "Tilastollinen aineisto ei ole keskittynyt yksittäisiin identiteetteihin vaan aineisto kerätään, jotta saadaan vastauksia kysymyksiin "Kuinka monta?" tai "Kuinka suuri osa?" eikä suinkaan kysymykseen "Kuka?". Identiteetit ja rekisterit mukana olevista tutkimuskohteista tulisi siksi pitää luottamuksellisina olipa luottamuksellisuus sitten erikseen luvattu tai ei."

Tilastokeskuksessa käsitellään suuressa määrin luottamuksellisia tietoja. Luottamuksellisuus on määritelty myös lain perusteella (esimerkiksi henkilötietolaki). Siitä huolimatta, että tilastojen tuottamiseen käytettävät tiedot ovat luottamuksellisia, eivät ne menetelmät, joilla julkistettu tieto on tuotettu saisi jäädä salaisuudeksi vaan tieto menetelmistä tulisi olla yleisön saatavilla. Mikäli yleisö ei luota julkaistuihin tilastoihin, ei niiden julkaisemisella ole merkitystä. Tilastot onkin laadittava niin, että yksikön tiedot eivät käy tilastoista ilmi, mutta tilastojen tiedot vastaavat todellisuutta ja niiden selittävyys on kerrottu rehellisesti.

Jotkut vastaajat eivät halua tietoja, jotka on kerätty tiettyyn tarkasti määrättyyn tarkoitukseen, käytettävän muihin tarkoituksiin. Kuitenkin Tilastokeskus kerää tietoja vain ja ainoastaan tilastollisiin tarkoituksiin, jonka lisäksi yksikön identiteetti suojataan kaikessa käytössä. Tilastokeskus saa kerättyjen tiedonantojen lisäksi tietoja hallinnollisista rekistereistä, jotka sisältävät usein arkaluonteisiakin tietoja. Näitä tai muita lainsäädännöllisellä velvoitteella tai suostumuksella kerättyjä arkaluonteisia tietoja on käsiteltävä erityisen huolellisesti. Yksiköiden identiteetin suojaaminen on yksi tärkeimmistä velvollisuuksista.

Aina tietoja julkaistaessa on määritettävä informaatiolle sellainen taso, jolla aineiston julkaisu tai muu jakelu ei mahdollista yhdenkään tutkimuskohteen identiteetin paljastumista suoraan tai päättelemällä. Tämä periaate koskee myös mikroaineistojen luovuttamista tutkimuskäyttöön. Absoluuttista suojausta luottamuksellisuuden rikkomista vastaan ei kuitenkaan ole olemassakaan. Useilla eri menetelmillä voidaan kuitenkin pienentää tällaisten rikkomusten todennäköisyyttä. Tavallisin ja potentiaalisesti varmin tapa on aineiston anonymointi. Muita mikroaineistojen suojaamiseen kehittyjä menetelmiä esitellään seuraavissa luvuissa.

3.4 Tilastokeskuksen tietosuojamenettelyt

Tilastokeskuksen aineistojen käyttöä valvotaan tarkasti käyttöoikeuksien avulla. Monimutkaisimmat aineistopyynnöt käsitellään tilastoeettisessä lautakunnassa, joka arvioi pyydetyn aineiston tietosuojariskejä ja päättää sen mukaan sallitaanko aineiston luovuttaminen tutkijalle vai voidaanko aineisto antaa tutkijan käyttöön Tilastokeskuksen tutkimuslaboratoriossa. Tietenkin käyttölupahakemuksia tulee myös sellaisiin aineistoihin, joiden antaminen tutkijan käyttöön on kokonaan evättävä.

Tilastokeskuksen aineistot kiinnostavat tutkijoita ja joka vuosi noin 200 käyttölupahakemusta saapuu Tilastokeskukselle. Tilastoeettinen lautakunta käsittelee uusista käyttölupahakemuksista hankalimmat ja tekee esityksensä kaikista ulkomaille pyydetyistä aineistoista pääjohtajan päätettäväksi. Käyttölupahakemuksista yli 25 vuosittain on vaatinut tarkempaa tarkastelua tilastoeettisessä lautakunnassa. Suurin osa näistä on ollut aineistojen luovutuksia ulkomaille. Tutkimuslaboratoriossa hankkeiden määrä on ollut noin kaksikymmentä viime vuosina. Kaiken kaikkiaan Tilastokeskuksesta luovutetaan tutkimuskäyttöön suunnilleen 130 aineistoa vuosittain. Kiinnostuksen erityisesti mikroaineistoja kohtaan voidaan olettaa olevan kasvussa, koska muutos on jo nähtävissä monien muiden maiden tilastovirastoissa.

Tutkimuslaboratoriossa käytettävistä henkilöitä koskevista aineistoista poistetaan kaikki suoraan tai välillisesti tunnistamiseen johtavat tiedot. Lisäksi muut aineistot muokataan muotoon, joiden pohjalta yritystä tai muuta tietojen kohdetta ei voida suoraan tunnistaa. Tilastokeskuksen ulkopuolelle toimitettavat aineistot suojataan yleen-

sä tarkemmin, vaikka laki on sama molemmille käyttötavoille. Myös aineistojen palauttaminen tai tuhoaminen käyttöajan jälkeen otettu huomioon jo lupaa myönnettäessä.

Tilastokeskuksessa on käytössä yleisiä ohjeita tilastojen, taulukkomuotoisten aineistojen ja yksikköaineistojen tietosuojauksesta. Taulukkomuotoisen aineiston suojaamiseen on annettu kaksi erilaista ohjetta, joista toinen on tarkoitettu henkilöaineistoille ja toinen yritysaineistoille. Ohjeiden tulkinnasta seuraa, etteivät tietosuojauksen periaatteet välttämättä ole täysin yhtenevät kaikkialla Tilastokeskuksessa. Mikroaineistojen luovuttaminen vaatii oman suojauksensa, mutta suojaus riippuu aina aineistosta ja sen tulevasta käytöstä, joten mikroaineistojen suojaamisessa käytetään lähtökohtana lain antamaa ohjenuoraa, jonka mukaan mikään yksiköistä ei saa olla suoraan tai välillisesti tunnistettavissa. Käytännössä aineistoista poistetaan suoraan identifioimiseen riittävät muuttujat ja aineisto karkeistetaan epäsuoran identifioitumisen estämiseksi. Nämä periaatteet ja aineiston käyttäjän velvollisuudet on kirjattu ohjeeseen käyttölupien myöntämisestä Tilastokeskuksen yksikkötason aineistoon (Tilastokeskus, 2005b).

4 Paljastumisriski ja tiedon hyödynnettävyys

Ennen menetelmien esittelyä on syytä pohtia aineiston suojauksen tasoa. Koska varsinaista suojauksen tason mittaa ei ole olemassa, on tarkasteltava paljastumisriskissä olevien yksiköiden lukumäärää ja riskin todennäköisyyttä. Tarkoitusta varten määritellään ensin paljastumisriski ja esitellään sen mittaamiseen sopivia menetelmiä. Lisäksi aineiston suojaamisen johdosta sen hyödyllisyys yleensä kärsii ja siksi on perusteltua tarkastella esiteltäviä menetelmiä myös tiedon hyödynnettävyyden kannalta. Määritellään siis informaatiokato ja esitellään myös sen mittaamiseen kehitettyjä menetelmiä. Molempia määritelmiä voidaan käyttää paitsi aineiston luovuttamiseen liittyvien päätösten perustana myös eri menetelmien välisen vertailun apuna.

Mikroaineiston suojaamisessa tutkitaan toisaalta paljastumisriskiä, toisaalta aineiston hyödyllisyyttä. Näiden kahden näkökulman toisena ääripäänä on aineiston julkaisematta jättäminen tai aineiston koodaaminen salakielelle, jolloin aineistosta ei ole mitään hyötyä käyttäjälle ellei hänellä ole mahdollisuutta purkaa salausta. Toisessa ääripäässä aineistolle ei tehdä mitään muutoksia eli luovutetaan tutkijalle alkuperäinen aineisto, jolloin aineistosta saadaan varmasti oikeat johtopäätökset, mutta mikään yksikkö ei välttämättä ole suojassa. Suojaamisen tavoitteena on löytää tasapaino näiden kahden välille. Tavoite on siis riittävän matala paljastumisriski ja samalla riittävän matala informaation häviäminen.

Mikroaineistojen suojaamisessa ideaalisena tavoitteena olisi siis se, että aineistosta ei voisi tunnistaa yhtään yksikköä edes välillisesti ja että aineiston tilastolliset ominaisuudet olisivat identtisiä alkuperäisen aineiston kanssa, jolloin kaikkien mahdollisten analyysien tulokset olisivat täsmälleen samat sekä suojattua että alkuperäistä aineistoa käyttävälle. Todellisuus on kuitenkin se, että yksikkökohtaisten arvojen muuttaminen aiheuttaa joka tapauksessa vähintäänkin jotain muutoksia aineistosta tehtäviin analyyseihin.

Suojausmenetelmien keskinäistä vertailemista vaikeuttaa erityisesti se, että suojausmenetelmät säilyttävät usein osan aineiston ominaisuuksista ja loput ominaisuuksista muuttuvat. Siksi menetelmiä ei pystytä laittamaan yksikäsitteiseen paremmuusjärjestykseen vaan niitä on helpompi arvioida tilannekohtaisesti. Suojausmenetelmien arvioimiseen ei siis ole olemassa tiettyä standardoitua järjestelmää ja jatkossa tarkastelussa tuodaankin esiin suojausmenetelmien hyviä ja huonoja puolia sekä tilanteita, joissa menetelmät toimivat erityisen hyvin ja jolloin ne pettävät. Vertailun tueksi annetaan myös suuntaa-antavaa tietoa menetelmän vaikutuksesta paljastumisriskiin ja informaation määrään.

4.1 Paljastumisriski

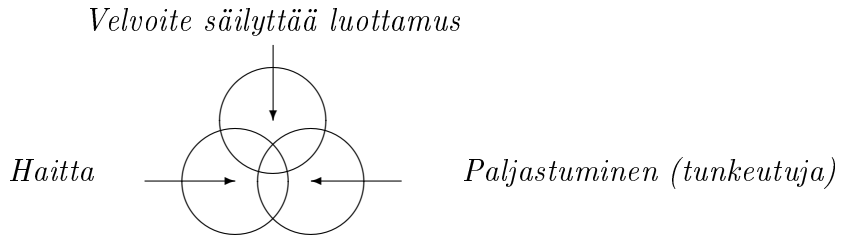
Tässä tutkimuksessa käsitellään mikroaineistoja eli aineistoja, joissa on taulukkomuodossa yksiköihin liittyviä arvoja useille eri muuttujille. Muuttujat voidaan paljastumisriskiä tutkittaessa jakaa avainmuuttujiin eli sellaisiin identifioiviin muuttujiin, joiden avulla paljastamista yrittävä henkilö eli tunkeutuja voi identifioida yksiköt ja tiedon kannalta herkkiin muuttujiin, joiden tietoja tämä haluaa hyödyntää mikäli tunnistaminen onnistuu.

Tarkempi jako muuttujille on seuraava:

- Identifioivat eli avainmuuttujat ovat muuttujia, joiden avulla henkilöt voidaan erottaa toisistaan. Identifioivia muuttujia on kahta eri tyyppiä.
  - Suoraan identifioivia ovat muuttujat, jotka identifioivat yksikön yksikäsitteisesti. Näitä ovat esimerkiksi henkilötunnus tai passin numero.
  - Epäsuoraan identifioivat muuttujat ovat taas sellaisia, jotka identifioivat yksiköt pienellä epävarmuudella. Esimerkkejä näistä ovat nimi, ammatti, asuinalue tai sukupuoli.

  Yleensä epäsuoraan identifioivat muuttujat ovat sellaisenaan hyödyttömiä, mutta käytettäessä useampia niistä yhdessä, identifioituu yksikkö varmasti. Epäsuorien muuttujien arvojen yhdistelmiä, jotka riittävät yksikön identifioimiseen kutsutaan yksikön profiiliksi.
- Lisäksi on herkkiä muuttujia, jotka ovat luottamuksellisia, kuten esimerkiksi palkka, uskonto, etninen tausta tai terveydentila
- ja muuttujia, jotka eivät ole mitään edellä olevista eli sellaisia, jotka eivät vaikuta identifioitumiseen eikä niiden paljastuminen haittaa yksikköä.

Paljastumisriskiä tarkastellessa on hyvä aluksi miettiä, mikä on yksikön kannalta ongelmallista paljastumista. Esimerkiksi se, että tunkeutuja saa selville jonkun tietyn vastaajan olleen mukana tutkimuksessa ei välttämättä vielä haittaa ketään. Jos kuitenkin samalla paljastuu vastaajan mielipiteitä tai ominaisuuksia, on kyseessä ongelmallinen paljastuminen. Ruotsin Statistiska Centralbyrån julkaisun "Statistisk röjandekontroll" (2001) mukaan siitä, mikä on arkaluonteista, on monia näkemyksiä. Lisäksi näkemykset vaihtelevat paitsi maittain myös ajan kuluessa ja lisäksi tilastontuottajilla on omat näkemyksensä ja tiedonantajilla omansa. Tästä huomataan, ettei arkaluonteisista asioista ohjeistaminen ole helppoa. Tilastolain mukaisesti kaikki tilastotarkoituksiin saatu tieto on salassa pidettävää. Paljastumista ei ole pelkästään identifioituminen vaan henkilökohtaisia asioita paljastuu myös silloin, kun tietyn homogeenisen ryhmän ominaisuuksia saadaan selville. Esimerkiksi tietyn työpaikan samaa palkkaa saavien työntekijöiden palkan keskiarvo paljastaa suoraan näiden työntekijöiden palkat.



Kuva 1: Luottamuksellisuuden, paljastumisen ja siitä seuraavien haittojen väliset yhteydet (Fienberg, 2000)

Kuvassa 1 korostetaan näkökulmaa siitä, miten tiedonantajille turvattava luottamuksellisuus, mahdollinen paljastuminen ja siitä seuraavat haitat ovat suhteessa toinen toisiinsa. Kuva 1 antaa aiheen korostaa sitä, että mahdollisesta paljastumisestakin seuraa haittaa yksikölle vain osassa tapauksista. Kuviosta havaitaan myös olemassa olevat haitat, jotka seuraavat paljastumisesta siitä huolimatta, ettei tiedonantajalle

luvattua luottamuksellisuutta ole rikottu. Käytännössä tämä tarkoittaa sitä, että tunkeutuja on mielestään identifioinut aineiston pohjalta jonkun, joka ei todellisuudessa ole edes mukana aineistossa. Hän on tehnyt virheen, mutta voi silti aiheuttaa haittaa tälle väärin tunnistamalleen yksikölle. Vastaavasti pelkästä luottamuksellisuuden rikkomisesta voi seurata vahinkoa ilman paljastumistakin esimerkiksi niin, että aineistoon kuuluva henkilö tiedottaa julkisesti olleensa mukana kyseisessä tutkimuksessa. Tällöin hänen tietojensa suojaaminen vaatii aineiston suojaajalta erityistä vaivannäköä ja samalla hän aiheuttaa muiden samankaltaisia ominaisuuksia omaavien yksilöiden paljastumisriskin kasvun. Tilastoviranomaisen on tällaisessa tilanteessa aloitettava suojaus uudestaan tai ryhdyttävä muihin toimenpiteisiin.

Paljastuminen voidaan luokitella eri luokkiin riippuen siitä, mitä ja miten paljastuu. Identifioituminen tarkoittaa yksikön tunnistamista ja sen seurauksena kaikki aineistossa olevat tiedot kyseisestä yksiköstä paljastuvat. Toinen paljastumistyyppi on ominaisuuksien paljastuminen ilman identifioitumista eli ryhmän yhtenevän ominaisuuden kuten palkan paljastuminen vaikka kukaan ryhmästä ei varsinaisesti tule identifioituneeksi. Lisäksi paljastumisriskille on tietosuojamenetelmien historian aikana esitetty useita muita eri määritelmiä. Paljastumiseksi voidaan tulkita myös tilanne, jossa julkaistavaa aineistoa hyväksikäyttäen tunkeutuja saa jollekin jo aiemmin tutkimalleen herkän muuttujan estimaatille tarkemman arvon.

Mikroaineiston suojaamista tarkastellessa on luonnollista lähteä olettamuksesta, että aineistosta poistetaan kaikki suoraan identifioivat muuttujat sekä sellaiset epäsuoraan identifioivat muuttujat, joiden voidaan olettaa aiheuttavan melkein suoran identifioitumisen (esimerkiksi nimi). Suojaamisen tutkiminen jatkuu siis siitä, ettei jäljellä ole enää itsestäänselviä identifioitumismahdollisuuksia.

Paljastumisriskin määritelmää ja mittaamista varten merkitään jatkossa alkuperäistä aineistoa $n \times p$ -matriisilla X (yksiköt $\times$ muuttujat) ja suojattua aineistoa matriisilla Z. Suojaamisen tavoite on saada matriisin Z yksiköiden paljastumisriski mahdollisimman alhaiseksi ja toisaalta aineiston käyttäjän tulisi saada matriisin Z avulla analyyseistään liki samat tulokset kuin hän saisi matriisin X avulla.

Yksikkökohtainen riski

Paljastumisriskin mittaamiseen voidaan käyttää yksikkökohtaista riskiä. Yksikkökohtainen riski identifioitumiselle määritellään tarkasteltavan aineiston yksikön todennäköisyytenä tulla tunnistetuksi oikein eli tunnistetuksi perusjoukon vastaavaksi yksiköksi. Henkilötilastoja tarkasteltaessa yksikkökohtainen riski määritellään keskittyen lähinnä aineiston ainutlaatuisen tai harvoin esiintyvän profiilin eli samojen identifioivien muuttujien arvojen omaavien riskeihin. Mikäli aineiston ainutlaatuista profiilia vastaa vain muutama yksikkö perusjoukossa, on kyseisen yksikön paljastumisriski korkea. Jos perusjoukossa on useita saman profiilin omaavia yksiköitä, paljastumisriski on pieni. Kun kullekin yksikölle on saatu määritettyä oma paljastumisriskinsä, on näistä helppo valita suojausta vaativat yksiköt.

Yksikkökohtaisen riskin tarkastelu helpottaa suojaamista tiedon hyödyllisyyden näkökulmasta, koska kaikkia yksiköitä ei ole välttämätöntä suojata ja siten luovutettava aineisto on paremmin käytettävissä. Valitettavasti tällaisesta tiedon suojaamisesta saattaa seurata analyyseihin harhaa. Yleensä ainutlaatuiset tapaukset vaikuttavat suuresti keskeisimpiin tunnuslukuihin ja mikäli niiden arvoja joko merkitään puuttuvaksi tiedoksi tai muuten muunnetaan, voi seurauksena olla yllättävän suuri harha.

Tyypillisin esimerkki on yritysaineistoista, jossa helposti tunnistettavien suuryritysten liikevaihdon tai muun sellaisen tiedon korvaaminen puuttuvan tiedon merkinnällä aiheuttaisi suuren harhan talouden lukuihin.

Paljastumisriskiä voidaan mitata yksiköiden yhdistämisellä, jossa valitun mallin perusteella määritetään miten moni suojatun aineiston yksiköistä voidaan yhdistää ulkopuolisen aineiston yksiköihin oikein. Erilaisia yhdistämistapoja on useita ja seuraavassa esitellään kaksi hyvin realistista tapaa yhdistää yksiköitä. Tarkempaan yksikkökohtaisen paljastumisriskin mittaamiseen käytetään kuitenkin epärealistista "worst case scenario" -tilannetta, joka esitellään myöhemmin. Pahin mahdollinen tilanne valitaan laskemisen yksinkertaistamiseksi.

Etäisyyteen perustuvaa yksiköiden yhdistämistä tarkastellessa oletetaan, että sekä alkuperäisessä että suojatussa aineistossa on p muuttujaa. Määritellään suojatun aineiston yksikkö vastaamaan lähintä alkuperäisen aineiston yksikköä ja määritellään "lähin" lyhimmän p-ulotteisen Euklidisen etäisyyden mukaan. Oletetaan lisäksi, että tunkeutuja on saanut käyttöönsä aineiston, joka sisältää vain i alkuperäisen aineiston muuttujista. Hän yrittää näiden i avainmuuttujan perusteella yhdistää alkuperäisen ja suojatun aineiston yksiköitä toisiinsa. Yhdistämisessä lasketaan seuraavaksi alkuperäisen ja suojatun aineiston yksiköiden väliset i-ulotteiset Euklidiset etäisyydet käyttäen vain i avainmuuttujan standardoituja arvoja. Standardoinnilla estetään mitta-asteikoista seuraavat ongelmat. Suojatun aineiston yksikkö merkitään "oikein yhdistetyksi", mikäli sen lähin i-ulotteinen yksikkö on oikea eli lähin myös käytettäessä p-ulotteista etäisyyttä.

Toinen erittäin hyödyllinen paljastumisriskin mitta saadaan tutkimalla pelkkiä aineiston ainutlaatuisia yksiköitä. Kun yksikön tiedetään olevan otosaineiston ainutlaatuinen ja käytetään tätä tietoa lisäinformaationa, voidaan laskea todennäköisyys $\mathbb{P}(PU|SU)$. Tämä on todennäköisyys sille, että yksikkö on ainutlaatuinen myös perusjoukossa ehdolla, että se on ainutlaatuinen kyseisessä otosaineistossa. Toisin ajateltuna kyseessä on todennäköisyys sille, että perusjoukossa on yksikkö, jolla on samat identifioivien muuttujien arvot kuin otosaineiston yksiköllä. Tätä tilannetta voidaan kutsua todennäköisyyteen perustuvaksi yksiköiden yhdistämiseksi.

Identifioituminen

Tarkastellaan paljastumista keskittyen ensin pelkkään identifioitumiseen. Oletetaan, että identifioituminen toteutuu yhdistämällä otoksen ja perusjoukon yksikkö toisiinsa oikein ja yhdistäminen on tapahtunut käyttämällä pelkkiä avainmuuttujia. Tarkastellaan yksikkökohtaista riskiä lähinnä avuksi arvojen peittämiseen. Merkitään paljastamista yrittävän henkilön kohteena olevan aineiston yksikköä j ja hänen käytössään olevan rekisterin yksikköä j^*. Nämä yksiköt ovat ne, jotka tunkeutuja haluaa saada yhdistetyksi toisiinsa käyttämiensä avainmuuttujien avulla. Vertailun tuloksena on identifioituminen, jos rekisteristä oleva yksikkö j^* valitaan vastaamaan yksikköä j ja vastaavuus on oikea eli rekisterin yksikkö j^* on todellakin sama kuin aineiston yksikkö j.

Määritellään seuraavaksi paljastumisskenaario ja otetaan tarkastelun näkökulmaksi "worst case scenario" (Hundepool et al. 2005) eli tilanne jossa kaikki pahimmat paljastumisuhat toteutuvat.

1. Perusjoukon P otos s ollaan julkaisemassa ja otospainot ovat käytettävissä.
2. Tunkeutujan käytössä oleva tiedot sisältävät koko perusjoukon P, jonka seurauksena jokaista $j \in s$ vastaava perusjoukon P yksikkö j^* on aina tunkeutujan käytettävissä.
3. Tunkeutujan käytössä oleva tietoarkisto sisältää yksilökohtaiset identifioivat muuttujat sekä otokseen kuuluvien luokiteltujen avainmuuttujien arvot.
4. Tunkeutuja yrittää yhdistää otoksen yksikön j perusjoukon rekisterissä olevaan yksikköön j^* vertaamalla avainmuuttujien arvoja näiden kahden aineiston välillä.
5. Tunkeutujalla ei ole käytössään mitään muuta informaatiota kuin se, mikä sisältyy hänen tietorekisteriinsä.
6. Identifioituminen tapahtuu, kun otoksen yksikön j ja perusjoukkoon perustuvan rekisterin yksikön j^* välille muodostetaan yhteys ja yksikkö j^* on todellisuudessa se perusjoukon yksikkö, josta otoksen yksikön tiedot johdettiin. Tämä tarkoittaa siis sitä, että muodostetun yhteyden on oltava oikea ennen kuin varsinainen paljastuminen tapahtuu.
7. Tunkeutuja yrittää yhdistää jokaisen otoksen yksikön johonkin perusjoukon rekisterin yksikköön.
8. Avainmuuttujien arvot ovat samat oikean yhdistämisen tapauksessa eli käytössä olevissa tiedostoissa ei ole virheitä, puuttuvia arvoja tai ajan aiheuttamia muutoksia avainmuuttujien osalta.

Merkitään äärellisestä N yksikön perusjoukosta P valittua satunnaisotosta s ja sen kokoa n. Yleiselle perusjoukon yksikölle j merkitään todennäköisyyttä tulla valituksi otokseen merkinnällä $1/w_j$. Tarkastellaan ensin avainmuuttujien ristiintaulukkoa. Avainmuuttujien kombinaatio k on ristiintaulukon k:s solu. Profiilien joukko $1, \ldots, k, \ldots, K$ määrittää sekä perusjoukon että otoksen osituksen soluihin. Tarkastelemalla yksikön $j \in s$ avainmuuttujien arvoja, saadaan yksikkö luokitelluksi oikeaan soluun. Merkitään sen solun indeksiä, johon yksikkö $j \in s$ tulee profiilinsa perusteella luokitelluksi, merkinnällä $k(j)$. Tavallisesti usean otoksen yksikön voidaan olettaa saavan saman avainmuuttujien profiilin.

Merkitään edelleen profiilin k omaavien suojatun aineiston yksiköiden lukumäärää f_k ja vastaavan perusjoukon yksiköiden lukumäärää F_k. F_k on tuntematon kaikille k. Avainmuuttujista riippuen kombinaatioiden yhteismäärä K voi olla hyvinkin suuri. Suojatussa aineistossa on ainoastaan osajoukko kaikista profiileista ja vain sellaiset perusjoukon profiilit, joille $f_k > 0$, ovat paljastumisriskin estimoimisen kannalta mielenkiintoisia.

Identifioitumisen tapauksessa yksikkökohtainen riski otoksen yksikön j paljastumiselle määritellään identifioitumisen todennäköisyytenä eli

$$\rho_j = \mathbb{P}(j \text{ yhdistetty oikein yksikköön } j^* | s, P).$$

Todennäköisyys, että yksikkö $j \in s$ on yhdistetty oikein yksikköön $j^* \in P$, on nolla, jos tunkeutuja ei yhdistä mitään yksiköitä. Tästä johtuen voidaan edellistä ehdollistaa tapahtumalla L_j: "tunkeutuja yrittää identifioida yksikön $j \in s$" ja kirjoittaa

$$\rho_j = \mathbb{P}(j \text{ yhdistetty oikein yksikköön } j^* | s, P, L_j)\mathbb{P}(L_j),$$

missä $\mathbb{P}(L_j)$ on todennäköisyys sille, että tunkeutuja yrittää yhdistää yksikön j johonkin perusjoukon P yksikköön.

Tarkastellaan identifioimisyrityksiä edellä esitellyn paljastumisskenaarion vallitessa. Valitaan pessimistinen lähestymistapa ja oletetaan, että paljastamista yrittävä henkilö yrittää identifioida kaikki aineiston yksiköt, jolloin $\mathbb{P}(L_j) = 1$ kaikille $j \in s$. Lisäksi oletetaan, että kaikille suojatun aineiston avainmuuttujien profiileille on täydellinen pari perusjoukossa. Toisin sanoen muuttujien arvoissa ei ole puutteita tai virheitä. Tämä oletus kasvattaa yksikön j paljastumistodennäköisyyttä. Näistä oletuksista johtuen käytettävän paljastumisskenaarion antama riski r_j ei ole ainakaan pienempi kuin edellä esitelty riski ρ_j, joten

$$\rho_j \leq r_j = \mathbb{P}(j \text{ yhdistetty oikein yksikköön } j^* | s, P, \text{worst case scenario}).$$

Todelliset arvot ρ_j ovat pienempiä kuin käytettävät estimaatit r_j. Tämän takia on perusteltua mitata paljastumisriskiä mitalla r_j.

Yksikkökohtaisen riskin estimoinnista

Paljastumisriski perustuu avainmuuttujien arvojen ristiintaulukoitujen solujen arvoihin. Tästä seuraa, että riski paljastua on sama kaikille samaan soluun kuuluville yksiköille. Merkitään siis riskiä mille tahansa yksikölle j, joka kuuluu soluun k merkinnällä r_k aiemman yksikkökohtaisen merkinnän r_j sijaan. Tarkastellaan edelleen avainmuuttujien profiilia k, missä $k = 1, \ldots, K$. Perusjoukossa on F_k yksikköä, joilla on kyseinen profiili. Näistä aineistoon on valikoitunut f_k yksikköä, eikä yksiköiden välillä ei voida tehdä eroa ilman uuden informaation saamista. Mikäli perusjoukon frekvenssi olisi tunnettu, olisi paljastumisriski helppo asettaa luvuksi $1/F_k$. Koska frekvenssit eivät välttämättä ole tiedossa, eikä niitä saada aina lasketuksikaan, määritellään riski ilman perusjoukon frekvenssejä.

Määritellään yksikkökohtainen riski Benedetti et al. (1998) artikkelin mukaisesti identifioimistodennäköisyyden estimaatin ylärajaksi. Käytetään bayesiläistä lähestymistapaa ja tarkastellaan jakaumaa perusjoukon frekvenssit ehdolla otoksen frekvenssit. Tämän jälkeen yksikkökohtainen riski saadaan $1/F_k$ posteriorin keskiarvona tarkasteltaessa jakaumaa $F_k|f_k$:

$$r_j = \mathbb{E}\left(\frac{1}{F_k}|f_k\right) = \sum_{h \geq f_k} \frac{1}{h}\mathbb{P}(F_k = h|f_k).$$

Tarkastellaan seuraavaa superpopulaatiomallia $F_k|f_k$ todennäköisyysmassafunktion laskemiseksi

$$\pi_k \sim [\pi_k] \propto 1/\pi_k \text{ riippumattomat}, k = 1, \ldots, K$$
$$F_k|\pi_k \sim \text{Poisson}(N\pi_k) \text{ riippumattomat}, F_k = 0, 1, \ldots$$
$$f_k|F_k \sim \text{binomial}(F_k, p_k) \text{ riippumattomat}, f_k = 0, 1, \ldots, F_k.$$

Mallin vallitessa $F_k|f_k$ posteriorijakauma on negatiivinen binomijakauma, jossa onnistumistodennäköisyys on p_k ja onnistumisten lukumäärä on f_k. $F_k|f_k$ saa tiheysfunktiokseen negatiivisen binomijakauman, jossa lasketaan kokeilujen määrää ennen

j:nnettä onnistumista onnistumisen todennäköisyyden ollessa p_k. Tällöin todennäköisyysmassafunktio on

$$\mathbb{P}[F_k = h | f_k = l] = \binom{h-1}{l-1} p_k^l (1-p_k)^{h-l}, \ h \geq l.$$

Käytettäessä edellä ollutta negatiivista binomijakaumaa, paljastumisriski voidaan kirjoittaa muotoon

$$r_k = \mathbb{E}(F_k^{-1} | f_k) = \int_0^\infty \left(\frac{p_k e^{-t}}{1 - q_k e^{-t}} \right)^{f_k} dt, \tag{4.1}$$

missä $q_k = 1 - p_k$.

Toisaalta paljastumisriski voidaan kirjoittaa muuttujanvaihtoa $y = (1 - q_k e^{-t})^{-1}$ käyttäen

$$\begin{aligned} r_k &= \left(\frac{p_k}{q_k} \right)^{f_k} \int_1^{1/p_k} \frac{1}{y} (y-1)^{f_k - 1} \, dy \\ &= \left(\frac{p_k}{q_k} \right)^{f_k} \left\{ \sum_{j=0}^{f_k - 2} (-1)^l \binom{f_k - 1}{l} \frac{{p_k}^{l+1-f_k} - 1}{f_k - l - 1} + (-1)^{f_k} \log(p_k) \right\}, \end{aligned} \tag{4.2}$$

joka pätee, kun $f_k > 1$. Vaihtoehtoinen muuttujanvaihto on $y = e^{-t}$ (μ-Argus manual, 2005), jolloin riskille pätee

$$r_k = {p_k}^{f_k} \int_0^1 t^{f_k - 1} (1 - t q_k)^{-f_k} \, dt.$$

Paljastumisriski voidaan kirjoittaa myös hypergeometrisen funktion avulla seuraavasti:

$$r_k = \frac{{p_k}^{f_k}}{f_k} {}_2F_1(f_k, f_k; f_k + 1; q_k), \tag{4.3}$$

missä

$${}_2F_1(a, b; c; d) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 t^{b-1} (1 - t^{c-b-1})(1 - td)^{-a} \, dt$$

on Gaussin hypergeometrisen sarjan integraaliesitys.

Yksikkökohtaisen paljastumisriskin estimaatti saadaan estimoimalla p_k jostakin edellä olleista riskin esityksistä. Suurimman uskottavuuden estimaatti p_k:lle, kun f_k on tunnettu, saadaan binomimallin vallitessa seuraavasti:

$$\hat{p}_k = \frac{f_k}{F_k}. \tag{4.4}$$

Koska arvoa F_k ei pystytä havaitsemaan, voidaan estimaatin (4.4) sijaan käyttää arvoa

$$\hat{p}_k = \frac{f_k}{\sum_{j:k(j)=k} w_j},$$

missä $\sum_{j:k(j)=k} w_j$ on otosasetelmaan perustuva, mahdollisesti kalibroitu estimaatti arvolle F_k.

Kaavan (4.2) käyttö riskin estimoimisessa johtaa usein epästabiileihin $\hat{p}_k$ estimaatteihin arvojen 0 tai 1 lähellä. Viimeinen mitta (4.3) ei kärsi tästä ongelmasta. Tästä syystä suositeltavin estimaattori yksikkökohtaiselle riskille on

$$\hat{r}_k = \frac{\hat{p}_k^{f_k}}{f_k} {}_2F_1(f_k, f_k; f_k + 1; 1 - p_k).$$

Negatiivinen binomijakauma on määritelty, kun $0 < p_k < 1$. Käytännössä estimaatit voivat silti saada vaihteluvälin reunan arvoja. Koska arvo $\hat{p}_k = 0$ vastaa arvoa $f_k = 0$, sitä ei tulla käsittelemään. Toisaalta, jos $\hat{p}_k = 1$, on ${}_2F_1(f_k, f_k; f_k + 1; 1 - \hat{p}_k) = 1$ ja yksikkökohtainen riski on $1/f_k$.

Jos parametrien f_k ja $1 - \hat{p}_k$ arvot ovat suuria, on hypergeometrisen funktion numeerinen arvioiminen laskennallisesti vaativaa. Siksi kohtuullistenkin solun arvojen tilanteessa on mahdollista käyttää approksimaatiota tarkkojen arvojen sijaan. Approksimaatiot perustuvat hypergeometrisen funktion ${}_2F_1(f_k, f_k; f_k + 1; q_k)$ sarjakehitelmään. Koska sarja hajaantuu vain kun $f_k < 0$, ei hajaantuminen tule kyseeseen paljastumisriskiä tarkasteltaessa. Sarja konvergoi varmasti silloin, kun $f_k > 1$. Riskin mitta tunnetaan myös, kun $f_k = 1$ ja se on $-log(p_k)\frac{p_k}{1-p_k}$. Myös otoksen frekvensseille kaksi ja kolme, mitan tarkat arvot ovat tunnettuja:

$$f_k = 2: \; r_k = \frac{p_k}{{q_k}^2}(p_k log p_k + q_k) \text{ ja}$$

$$f_k = 3: \; r_k = \frac{p_k}{2{q_k}^3}(q_k(3q_k - 2) - 2{p_k}^2 log p_k).$$

Ehdotettava approksimaatio hyödyntää hypergeometrisen funktion ominaisuutta

$${}_2F_1(f_k, f_k; f_k + 1; q_k) = (1 - q_k)^{1-f_k} {}_2F_1(1, 1; f_k + 1; q_k)$$

ja sarjakehitelmää

$${}_2F_1(a, b; c; d) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)} \sum_{n=0}^{\infty} \frac{\Gamma(a+n)\Gamma(b+n)}{\Gamma(c+n)} \frac{d^n}{n!}.$$

Näiden kaavojen avulla saadaan esimerkiksi esitys

$$r_k = \frac{p_k}{f_k} \left\{ 1 + \frac{q_k}{f_k + 1} + \frac{2{q_k}^2}{(f_k + 1)(f_k + 2)} + \frac{3! {q_k}^3}{(f_k + 1)(f_k + 2)(f_k + 3)} + O({f_k}^{-4}) \right\}.$$

Esitystä voidaan muokata haluttaessa ottaen mukaan lisää termejä, jolloin virhetermin vaikutus pienenee tai suuremmille solun arvoille osa termeistä voidaan jättää poiskin. Jos aineiston tilanteesta ei ole varmuutta tai frekvenssit vaihtelevat, suositellaan käytettäväksi seitsemää ensimmäistä termiä. Koska hypergeometrisen funktion käyttö voi olla laskennallisesti vaativaa ja toisaalta edellä olleet approksimaatiot ovat käytettävissä, kannattaa riski estimoida käyttäen $\hat{p}_k$ estimaatteja tässä kaavassa.

Koko aineiston riskin estimoinnista

Yksikkökohtainen paljastumisriski tarkastelee ainoastaan yksikkötason riskiä, mutta aineiston hallinnan kannalta on syytä määritellä koko aineiston paljastumisriski. Määritellään se ensin aineiston odotettavissa olevaksi määräksi identifioituneita yksiköitä. Tämä mitta riippuu aineistossa olevasta yksiköiden määrästä, mutta toinen määriteltävä mitta identifioitumisaste on riippumaton lukumäärästä n.

Määritellään dikotominen satunnaismuuttuja Φ, jonka arvo 1 tarkoittaa oikeaa identifioitumista ja arvo 0 on identifioituminen, joka ei ole oikea. Määritellään edelleen jokaiselle aineiston yksikölle kyseinen muuttuja Φ_j ja oletetaan arvo 1 saatavan korkeintaan todennäköisyydellä r_j. Jatkossa ajatellaan yksinkertaisuuden vuoksi, että todennäköisyys arvolle 1 on tasan r_j. Satunnaismuuttujat Φ eivät ole *i.i.d.*, mutta riski on vakio kaikkien ristiintaulukon solujen yli. Tästä seuraa, että avainmuuttujien kombinaatiolle k on f_k kappaletta *i.i.d.* satunnaismuuttujia Φ_j, joiden oletetaan saavan arvon 1 identifioitumisen ollessa oikea ja sen todennäköisyyden ollessa vakio r_k. Tätä todennäköisyyttä hyväksikäyttäen voidaan johtaa solukohtainen odotettavissa olevien identifioitumisten lukumäärä, joka vastaa arvoa $f_k r_k$. Koko aineiston odotettavissa olevien identifioitumisten määrälle saadaan siten

$$ER = \sum_{j=1}^{n} \mathbb{E}(\Phi_j) = \sum_{k=1}^{K} f_k r_k.$$

Tätä mittaa voidaan käyttää identifioitumisasteen ξ määritelmässä

$$\xi = \frac{1}{n} ER = \frac{1}{n} \sum_{k=1}^{K} f_k r_k.$$

Mitta ξ on globaalin riskin mitta, joka mittaa paljastumisriskiä koko aineistosta eikä riipu aineiston koosta. Tästä syystä mittaa ξ voidaan käyttää aineiston riskin määrittämiseen sekä vertailtaessa erilaisia aineistoja. Yhtä hyvä globaalin riskin mitta saadaan odotettavissa olevien identifioitumisten prosenttiosuudesta $\Psi = 100 \cdot \xi$ %.

4.2 Informaatiokato

Mikroaineistojen suojauksen tarkoitus ei ole muokata suojattavaa aineistoa käyttökelvottomaksi kryptiseksi aineistoksi vaan suojatun aineiston tulee olla edelleen käyttäjälleen hyödyllinen. On siis mielekästä määritellä mittareita myös muokkaamisen seurauksena häviävälle informaatiolle: informaatiokadolle. Informaatiokadon mittaamiseen on paljastumisriskin tavoin esitetty useita eri menetelmiä. Kaikkien näiden tavoitteena on esittää se ero, joka muokatun ja alkuperäisen aineiston tilastollisten ominaisuuksien välillä on.

Esitellään ensin Yancey et al.(2002) artikkelissa käsiteltyjä menetelmiä. Kun informaatiokatoa lähdetään määrittelemään, on ensimmäinen ongelma määrittää olennaiset tilastolliset ominaisuudet ja tämän jälkeen keksiä tapa laskea näiden ominaisuuksien erot kahden aineiston välillä. Jos merkitään alkuperäistä aineistoa X ja suojattua aineistoa Z ja oletetaan molempien olevan kokoa $n \times p$, voisi eräs kiinnostuksen kohde olla aineiston muutos eli mitta erolle $Z - X$. Ensimmäinen määritelmä olisi tällainen:

$$\frac{1}{pn} \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{|x_{ij} - z_{ij}|}{|x_{ij}|}, \tag{4.5}$$

mutta ongelmaksi muodostuisi alkuperäisen aineiston arvot $x_{ij} = 0$. Nimittäjä voidaan korvata jollain vakiolla kun $x_{ij} = 0$, mutta vakion valinnasta riippuen kaavan (4.5) arvot voivat vaihdella suuresti. Erityisesti ongelmia on sellaisessa aineistossa, jossa useat arvot ovat nollia. Lisäksi ehdotuksen (4.5) antamat informaatiokadon mitat voivat olla jopa monta suuruusluokkaa suurempia kuin muiden mittojen, joten tämän vertailu muihin mittoihin on liki mahdotonta. Kun mittaa muokataan, saadaan kaava

$$IL1 = \frac{1}{pn} \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{|x_{ij} - z_{ij}|}{\frac{1}{2}(|x_{ij}| + |z_{ij}|)},$$

jonka arvoille pätee $IL1 \leq 2$. Suuruusluokkaan liittyvä ongelma on siis jo hallinnassa. Huomataan, että tässä kaavassa nollalla jakaminen on epätodennäköisempää kuin edellisessä, mutta ei kuitenkaan mahdotonta. Tämän ongelman lisäksi molemmat edellä esitellyistä informaatiokadon mitoista on skaalattu aineistojen yksittäisillä arvoilla, eivätkä ne tästä syystä ole kovin stabiileja. Nollaa liki olevien arvojen suojaamisesta seuraa muutoksia, joiden vaikutus on suuri myös informaatiokadon mittaan.

Tarkastellaan suojattavaa aineistoa riippumattomina otoksina jostakin jakaumasta. Kun kaikki summan arvot skaalataan muuttujalle yhteisellä arvolla, saadaan arvojen muutoksen mitasta stabiili. Esimerkiksi jos X ja Z ovat riippumattomia satunnaismuuttujia, joiden kummankin keskiarvona on μ ja varianssina σ^2, on satunnaismuuttujan $Y = X - Z$ keskiarvo 0 ja varianssi $2\sigma^2$. Tämän perusteella yleinen skaalaus muuttujalle Y olisi sen keskihajonta $\sqrt{2\sigma^2}$. Tässä tapauksessa keskihajontaa voidaan estimoida käyttämällä otoskeskihajontaa S. Ehdotetun muutoksen jälkeen informaatiokadon mitta on

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

Mitan arvoja laskettaessa muuttujan saamille arvoille on käytetty yhtä ja samaa skaalausta. Nimittäjä ei saa arvoa nolla, elleivät jonkun muuttujan saamat arvot ole samoja läpi koko aineiston.

Yancey et al.(2002) esittävät kahta vaihtoehtoa suojatun aineiston informaatiokadon tarkasteluun. Toinen suoritelluista vaihtoehdoista on jättää koko informaatiokadon

käsite tarkastelematta, koska aineiston suojaamisen tarkoituksena on suojata aineiston yksiköitä ja pyrkiä samalla säilyttämään aineiston tilastolliset ominaisuudet. Jos näitä tavoitteita pitää ensisijaisina, ei ole mielekästä tarkastella suojatun aineiston arvojen muuttumista suhteessa alkuperäisen aineiston arvoihin. Mikäli taas informaatiokatoa halutaan jollakin tavoin mitata, on suositeltavaa mitata sitä käyttämällä yleistä ja joustavaa skaalausta, kuten mitassa $IL1s$.

Domingo-Ferrer ja Torra (2001) esittävät aineiston tilastollisten ominaisuuksien pysyvyyden mittaamiseen kehitettyjä mittoja. Otoksen keskiarvojen vaihtelua mittaa

$$IL2 = \frac{1}{p} \sum_{j=1}^{p} \frac{|\overline{x}_j - \overline{y}_j|}{|\overline{x}_j|}.$$

Periaatteessa tämäkin mitta voi kärsiä nollasta tai suhteellisen pienistä arvoista nimittäjässä. Siksi mitta on tarkoitettu erityisesti aineistoille, joiden muuttujien arvot ovat ei-negatiivisia. Otoksen kovarianssimatriisin muutoksia voidaan mitata tunnusluvulla

$$IL3 = \frac{2}{p(p+1)} \sum_{j=1}^{p} \sum_{k=1}^{j} \frac{|Cov(X)_{jk} - Cov(Y)_{jk}|}{|Cov(X)_{jk}|}$$

ja otoksen varianssien muutosten aiheuttamaa informaatiokatoa mittaa tunnusluku

$$IL4 = \frac{1}{p} \sum_{j=1}^{p} \frac{|Cov(X)_{jj} - Cov(Y)_{jj}|}{|Cov(X)_{jj}|}.$$

Korrelaatiomatriisien erolle mitta saadaan tunnusluvusta

$$IL5 = \frac{2}{p(p-1)} \sum_{j=1}^{p} \sum_{k=1}^{j} |Cor(X)_{jk} - Cor(Y)_{jk}|.$$

Varsinaista koko aineiston informaatiokatoa kuvaavaa mittaa varten edellä esiteltyjä tunnuslukuja yhdistetään eri asioita korostaen. Koska on vaikeaa määritellä yhtä yleistä informaatiokatoa eli päättää, mitkä muutoksista vaikuttavat kaikkein eniten aineiston informaation määrään, käytetään arvojen laskemisessa edellä esiteltyjen tunnuslukujen aritmeettisia keskiarvoja. Mikäli aineiston tulevasta käytöstä tiedetään tarkemmin, voidaan tiettyjä tunnuslukuja korostaa painojen avulla, mutta yleisessä käytössä painottamaton keskiarvo toimii parhaiten. Aiemmin havaittiin, että mitta $IL1$ on ongelmallinen, koska sen nimittäjäksi tulee helposti nolla tai sitä liki oleva arvo. Tämän lisäksi kyseinen tunnusluku mittaa numeroarvojen muutoksia, jotka eivät yleensä ole merkityksellisiä yleisten tilastollisten ominaisuuksien muutosten rinnalla. Näillä perusteluin yksi vaihtoehto aineiston informaatiokadon mitaksi on

$$s0 = \frac{IL2 + IL3 + IL4 + IL5}{4}.$$

Jos taas yksittäisten arvojen muutosta mittaava tunnusluku halutaan ottaa mukaan mittaamisessa, saadaan

$$s1 = \frac{IL1 + IL2 + IL3 + IL4 + IL5}{5}.$$

Yksi vaihtoehto tarkastelulle on jättää pois joku mitoista $IL3$, $IL4$ tai $IL5$ eli kovarianssien, varianssien tai korrelaatioiden eron mitta. Jos näistä kolmesta tunnetaan kaksi, voidaan kolmas laskea niiden perusteella. Koska kovarianssien eroa mittaava tunnusluku $IL3$ kärsii samasta ongelmasta kuin alkuperäinen $IL1$, missä pienimpien arvojen vaikutus tunnusluvun arvoon on suurin, voidaan näistä kolmesta valita juuri kyseinen tunnusluku jätettäväksi tarkastelun ulkopuolelle. Tällöin saadaan

$$s2 = \frac{IL1s + IL2 + IL4 + IL5}{4}.$$

Paljastumisriskin ja informaatiokadon mittojen käyttö menetelmien vertailussa

Menetelmien vertaileminen aineiston paljastumisriskin ja informaatiokadon mittojen avulla on ollut vaikeaa, koska yleensä tiedon suojaamista on mitattu todennäköisyydellä eli sen arvo 0 ja 1 välillä, kun taas informaatiokadon mitta on ollut jatkuva rajoittamaton luku. Jos informaatiokadon mitasta tiedetään, että se on rajoitettu ja raja on riittävän pieni, voidaan menetelmiä vertailla esimerkiksi paljastumisriskin ja informaatiokadon mittojen keskiarvon avulla.

Esimerkiksi Domingo-Ferrer ja Torra (2001) ovat hyödyntäneet mittoja menetelmien vertailussa. He ovat laskeneet vertailuihinsa informaatiokadon mittaamiseen kullekin ominaisuudelle kolme mittaa, joista vain osaa hyödynnetään tulosten laskemisessa. Tutkittavia ominaisuuksia ovat edellä esiteltyyn mittaan $s0$ kuuluvat arvot, mutta jokaiselle näistä on mitattu keskineliövirhe, keskipoikkeama ja absoluuttinen keskipoikkeama. Näistä muiden mittojen erolle käytetään keskipoikkeamaa, mutta korrelaatiomatriisin kohdalla eroa mitataan absoluuttisella keskineliöpoikkeamalla eli mitalla $IL5$. Paljastumisriskin mittaamisessa artikkelissa on käytetty molempia edellä esiteltyjä mittoja ja näiden lisäksi vielä tarkasteltavan arvon paljastumista likimain eli tilannetta, jossa arvolle saadaan tarkka väli, jolla se sijaitsee.

Domingo-Ferrer ja Torra sanovat itsekin, että tulokset ovat vain suuntaa-antavia, vaikka he ovat käyttäneet tarkastelussaan kaikkia näitä mittoja. Mittojen antamia tuloksia ei siis pidä uskoa sokeasti vaikka useita mahdollisia tilastollisia ominaisuuksia olisikin niiden avulla mitattu. On parempi luottaa tietosuojamenetelmistä saamiinsa kokemuksiin, kun pelkkien tunnuslukujen arvoihin.

5 Menetelmien jaottelua

Tilastollisten aineistojen tietosuojamenetelmien kehitys alkoi 1970-luvulla (Domingo-Ferrer, 2002). Kehityksen käynnisti hallinnollisten tilastojen laatimisen muuttuminen käsin tekemisestä tietokoneavusteiseksi. Muutoksen myötä tilastoja on alettu laatia yhä enemmän ja samalla yksityisyyden suoja on tullut yhä tärkeämmäksi. Tätä kautta tietosuojamenetelmien parantaminen on tullut keskeiseksi tutkimusalueeksi.

Tietosuojamenetelmistä puhuttaessa on parasta aloittaa tosiasioista, jotka ennen tietojen julkaisemista on hyväksyttävä. Jos halutaan puhua paljastumisen estämisestä, on ainoa ja siten paras menetelmä helppo valita: jättää aineisto julkaisematta. Mikäli jotain aineistoa julkaistaan, on aina olemassa mahdollisuus myös arkaluonteisen tiedon paljastumiseen. Voi olla, että todennäköisyys paljastumiseen on lähes olematon, eikä paljastumista tapahdu, mutta julkaistu aineisto ei koskaan ole täysin paljastumisriskiä vailla. Tietosuojamenetelmien tarkoituksena on kuitenkin pyrkiä minimoimaan tuo paljastumisriski. Samalla yleensä aiheutuu epävarmuutta mahdolliseen tunnistamiseen eli vaikka paljastamista yrittävä henkilö olisikin mielestään paljastanut jonkun yksikön, ei hän voi olla täysin varma kyseisen yksikön identiteetistä.

Mikroaineiston suojaamiseen liittyvät menetelmät voidaan ensisijaisesti jakaa aineiston rajoittamiseen ja aineiston muuntamiseen perustuviin menetelmiin. Menetelmien erona on nimensä mukaisesti tapa, millä aineisto muuttuu. Rajoittamiseen perustuvat menetelmät eivät millään tavoin muuta julkaistavan aineiston arvoja, mutta osa julkaistavan aineiston yksiköistä tai arvoista on poistettu tai merkitty puuttuvaksi tiedoksi. Aineiston rajoittamiseen perustuvia menetelmiä ovat esimerkiksi otanta ja arvojen peittäminen, missä herkkä arvo jätetään kertomatta. Aineiston muuntamiseen perustuvia menetelmiä, joissa aineistoa muokataan tavalla tai toisella, on runsaasti. Tyypillisimmät esimerkit ovat kohinan lisääminen, luokkien vähentäminen tai niiden laajentaminen ylä- tai alareunoista. Menetelmiä voidaan myös yhdistää käyttämällä esimerkiksi otantaa ja lisäksi muuntaa otokseen valikoituneiden yksiköiden arvoja.

Matriisimenetelmiksi kutsutaan sellaisia menetelmiä, joiden vaikutus aineistoon voidaan suorittaa matriisilaskutoimituksin suojattavan aineiston muodostamalle matriisille (Polettini et al. 2002). Näissä menetelmissä on tietty funktio alkuperäisen datan ja suojattujen arvojen välillä. Funktio voi olla stokastinenkin. Suojaaminen voidaan toteuttaa joko muokkaamalla alkuperäistä aineistoa X, jolloin siitä saadaan suojattu aineisto Z. Vaihtoehtoisesti voidaan luoda täysin synteettinen aineisto Z, jonka luominen toteutetaan säilyttäen haluttu osa alkuperäisen aineiston X tilastollisista ominaisuuksista. Palataan synteettisen aineiston tarkasteluun vasta luvussa 7.8. Aineiston muokkaaminen voidaan esittää matriisimenetelmien osalta matemaattisella mallilla, jossa suojattu aineisto muodostuu matriisituloista alkuperäisen matriisin vasemmalta ja oikealta puolen sekä mahdollisesta kohinan lisäyksestä. Toisin sanoen malli on

$$X -> AXB + C = Z,$$

missä A on jokin $a \times n$ -matriisi, jonka operaatiot vaikuttavat yksiköihin, ja B on jokin $p \times b$ -matriisi, joka vaikuttaa muuttujiin ja C on matriisi, joka lisää häiriötä eli kohinaa.

Tähän tyyppiin kuuluu monia eri menetelmiä. Esimerkiksi

- kohinan lisääminen
- otanta, joka poistaa rivejä matriisista X
- arvojen peittäminen
- simuloidun aineiston lisääminen alkuperäiseen aineistoon eli käytännössä rivien lisääminen matriisiin X
- arvojen vaihtaminen yksiköiden välillä.

Menetelmiä on mahdollista jakaa myös aineiston tyypin kautta. Menetelmiä voidaan tarkastella jakamalla ne jatkuvien muuttujien ja luokiteltujen muuttujien tietosuojamenetelmiin. Tässä tapauksessa jatkuviksi muuttujiksi luokitellaan kaikki sellaiset muuttujat, jotka saavat numeerisia arvoja ja joilla on voimassa aritmeettisia laskutoimituksia. Jatkuvien muuttujien tapauksessa on hyvin todennäköistä, että aineistossa on useita ainutlaatuisia tapauksia ja siten suojaamaton aineisto on erityisen herkkä paljastumiselle. Onneksi kuitenkin näiden muuttujien suojaaminen onnistuu melkein helpoimmin, koska aritmetiikkaa voidaan hyödyntää suojauksessa. Luokitellut muuttujat ovat niitä, joille muuttujanarvot ovat äärelliseltä väliltä, eikä tavalliset aritmeettiset laskutoimitukset arvojen välillä ole järkeviä. Luokitelluille muuttujille suojaus on haastavampaa kuin jatkuville. Suojausta helpottaa tietysti se, että ainutlaatuisia tapauksia ei alunperinkään liene kovin monia. Mikäli luokiteltu muuttuja on järjestysasteikollinen voidaan suojauksessa hyödyntää minimi- ja maksimiarvoja. Jos taas kyseessä on luokitteluasteikollinen muuttuja, ei arvoille voi tehdä muuta kuin parittaisia vertailuja.

Muuttujan tyyppi ja tietosuojamenetelmät

Seuraavassa tarkastellaan eri menetelmien sopivuutta jatkuvalle tai luokitellulle muuttujille. Sopivuudet on merkitty taulukoihin, joista ensimmäisenä on aineiston rajoittamiseen perustuvat menetelmät ja sitten aineiston muokkaamiseen perustuvat menetelmät. Taulukot pohjautuvat Group Crisesin raporttien vastaaviin (Group Crises 2004a, 2004b).

Taulukko 1: Aineiston rajoittamiseen perustuvat menetelmät

Menetelmä	Jatkuva muuttuja	Luokiteltu muuttuja
Otanta		X
Arvojen peittäminen		X

Taulukossa 1 oletetaan, että esitelty suojausmenetelmä on ainoa aineiston suojaamiseen käytettävä menetelmä, jolloin jatkuvan muuttujan arvot voivat helposti johtaa tunnistamiseen.

Taulukko 2: Aineiston muokkaamiseen perustuvat menetelmät

Menetelmä	Jatkuva muuttuja	Luokiteltu muuttuja
Kohinan lisääminen	X	(X)
Sullivanin menetelmä	X	X
Arvojen vaihtaminen yksiköiden välillä	X	X
Mikroaggregointi	X	(X)
Luokkien vähentäminen	X	X
Luokkien laajentaminen mitta-asteikon ylä- tai alarajoilla	X	X
PRAM		X
MASSC		X

Tässä tutkielmassa menetelmät jaotellaan aineiston rajoittamiseen ja aineiston muokkaamiseen perustuviin menetelmiin. Kuitenkin matriisimenetelmiin kuuluvien menetelmien yhteydessä tullaan käyttämään samoja merkintöjä kuin jaottelutavan esittelyn yhteydessä. Jaottelun vaikeutena on ollut se, että eri lähteet tulkitsevat hieman eri näkökulmista samoja menetelmiä ja jakavat ne siten erilailla. Varsinkin luokittelun muuttaminen on nähty toisaalta aineiston rajoittamisena ja toisaalta aineiston muuttamisena. Tässä tutkielmassa käytetään jälkimmäistä näkökulmaa.

6 Aineiston rajoittamiseen perustuvat menetelmät

Aineiston rajoittamiseen perustuvissa menetelmissä on se hyvä puoli, ettei menetelmä voi mitenkään muuttaa luovutettavaa aineistoa. Tietenkin esimerkiksi otannassa vaarana on se, että otos ei vastaakaan koko perusjoukon tilannetta. Silti tutkija voi luottaa siihen, että otannan jälkeen kunkin yksikön muuttujanarvot ovat aitoja. Samoin on myös silloin, mikäli tiedon suojaamiseen käytetään arvojen peittämistä. Tämän menetelmän ongelmaksi muodostuu kuitenkin luovutettavan tiedon rajallisuus. Luovutettu aineisto ei ehkä olekaan riittävän informatiivinen tutkimusta varten.

Aineiston rajoittamiseen perustuvia menetelmiä ovat Group Crisesin (2004c) mukaan muun muassa otanta, arvojen peittäminen sekä luokituksen muuttaminen katsottuna yksityiskohtien vähentämisen näkökulmasta. Tässä tutkielmassa käsitellään kuitenkin luokituksen muuttamiseen perustuvia menetelmiä vasta seitsemännessä luvussa. Tämä siksi, että menetelmä on useissa muissa lähteissä tulkittu aineiston muuttamiseen perustuvaksi.

6.1 Otanta

Otantaa käytetään tietosuojamenetelmänä usein muiden menetelmien ohessa, sillä se on varma tapa pienentää paljastumisriskiä muuttamatta mitään aineiston arvoista. Tarkastellaan kuitenkin ensin otantaa ainoana aineiston suojaukseen käytettävänä menetelmänä. Otannan käytöstä muuhun kuin tietosuojaan on olemassa paljon materiaalia ja koska otanta tietosuojamenetelmänä ei eroa kattavan otoksen valinnasta, otantaan voi perehtyä tarkemmin kirjallisuuteen tutustumalla.

Luonnollisesti pelkän otannan käyttö jatkuvia muuttujia sisältävän aineiston suojaamisessa ei yleensä riitä vähentämään aineiston paljastumisriskiä. Jotkin yleiset muuttujat, kuten ikä, hyvin samanikäisiä ihmisiä sisältävässä aineistossa, ovat toki poikkeus. Yleensä tilanne on kuitenkin se, että jatkuvan muuttujan tietylle arvolle on harvoin kahtakaan yksikköä. Siksi on perusteltua olettaa kahta eri aineistoa tarkastellessa samojen arvojen sattuessa kohdalle kyseessä olevan yksi ja sama yksikkö. Tämän yleisen esimerkin valossa, on perusteltua tarkastella pelkän otannan vaikutusta paljastumisriskiin vain luokiteltujen muuttujien osalta.

Otannassa paljastumisriskin muutosta helppo arvioida, koska tiedetään alkuperäisen aineiston ja otokseen tulevien tapausten lukumäärät ja selvästikin otoksesta poisjääneet yksiköt ovat suojattuja. Otokseen tulevien yksiköiden paljastumisriskin määrittäminen onkin sitten hankalampaa.

Kun paljastusta yrittävä henkilö saa otannalla suojatun aineiston muiden käytössään olevien aineistojen lisäksi, hän käyttää hyödykseen avainmuuttujia (henkilöaineistossa esimerkiksi sukupuoli, ikä ja maantieteelliset muuttujat) päätelläkseen voiko kyseinen yksikkö olla mukana otosaineistossa. Seuraavaksi Fienbergin artikkelin (2000) mukaan hänen on tarkastettava, onko kyseinen yksikkö ainutlaatuinen otosaineistossa ja pohdittava sitten, onko tämä mahdollisesti ainutlaatuinen myös perusjoukossa. Mikäli tunkeutuja arvioi tilanteen väärin, voi siitä seurata väärä yhdistäminen eli otosaineiston tiedot eivät koskekaan haluttua yksikköä, vaikka hän luulee niin. Paljastumisriskiä arvioitaessa tulee siis arvioida kuinka suuren osan yhdistämisistä paljastamista yrittävä henkilö onnistuu saamaan oikein.

6.2 Arvojen peittäminen

Arvojen peittämisen tavoitteena on poistaa aineistosta ainutlaatuisia tapauksia. Mikäli aineiston muuttujien joukossa on jatkuvia muuttujia, yleensä aineistossa melkein kaikki yksiköt ovat ainutlaatuisia. Koska tähän tilanteeseen ainoa ratkaisu arvojen peittämismenetelmällä olisi peittää liki kaikki jatkuvan muuttujan arvot, lienee järkevämpää tulkita tämä menetelmä luokitelluille muuttujille sopivaksi menetelmäksi.

Arvojen peittämistä käytettäessä on tavoitteena saada ainutlaatuisten yksiköiden joukko pienenemään. Arvoja peitetään siten, että ainutlaatuinen yksikkö saadaan puuttuvan arvon avulla osaksi jotain tai joitain joukkoja, jossa on jo useampia yksiköitä. Pelkkien ainutlaatuisten yksiköiden arvojen peittäminen johtaa helposti siihen, että tunkeutuja voi päätellä kyseisen yksikön olleen ainutlaatuinen ja hyödyntää tietoa paljastamisyrityksessään. Tästä syystä peittämistä on lisättävä muihinkin yksiköihin tai tiedon puuttumisen syy on jätettävä erittelemättä.

Paljon erilaisia henkilöitä sisältävän aineiston suojaaminen arvojen peittämisellä johtaa aineiston informaation heikkenemiseen. Ainutlaatuiset yksiköt ovat myös usein sellaisia, jotka vaikuttavat suuresti tunnuslukuihin. Siksi pelkän peittämisen avulla voidaan saada aikaiseksi aineisto, jota käytettäessä aineistolla saatavat tulokset voivat erota suurestikin alkuperäisen aineiston perusteella saatavista tuloksista. Suositeltavaa olisikin peittämisen jälkeen paikata aineisto esimerkiksi imputoinnin avulla.

Seuraavassa on esimerkki kuvitteellisesta, muutamia henkilöitä sisältävästä aineistosta, jonka suojaamiseen käytetään peittämistä. Aineisto ei sinällään vaadi suojaamista, koska mitään arkaluonteista tietoa ei aineiston yhteydessä paljastu. Suojaamiseen saadaan peruste, kun ajatellaan, että aineisto on osa suurempaa aineistoa ja esitellyt muuttujat ovat aineiston henkilöitä kuvaavia muuttujia. Tässä tapauksessa kaikki aineiston yksiköt ovat aineistossaan ainutlaatuisia, joten peittämistä suoritetaan arvioimalla yksiköiden ainutlaatuisuutta perusjoukossa. Esimerkiksi aineiston ensimmäinen yksikkö: nuori mies, joka on lähihoitajana pienellä paikkakunnalla, on melko varmasti perusjoukossa ainutlaatuinen.

Taulukko 3: Arvojen peittäminen, alkuperäinen aineisto

Havainto	Ikä	Sukupuoli	Ammatti	Kotikunta
1	19	Mies	Lähihoitaja	Ruovesi
2	25	Nainen	Muusikko	Helsinki
3	38	Mies	Osastosihteeri	Helsinki
4	45	Nainen	Toimitusjohtaja	Muurame
5	33	Mies	Kokki	Mikkeli
6	28	Nainen	Matemaatikko	Parikkala
7	37	Nainen	Myyjä	Jyväskylä
8	46	Nainen	Opettaja	Kirkkonummi
9	29	Mies	Kirvesmies	Tampere

Taulukko 4: Arvojen peittäminen, suojattu aineisto

Havainto	Ikä	Sukupuoli	Ammatti	Kotikunta
1	19	Mies	Lähihoitaja	*
2	25	*	Muusikko	Helsinki
3	*	Mies	Osastosihteeri	Helsinki
4	45	Nainen	Toimitusjohtaja	*
5	33	Mies	Kokki	Mikkeli
6	28	Nainen	*	Parikkala
7	37	Nainen	Myyjä	Jyväskylä
8	*	Nainen	Opettaja	Kirkkonummi
9	29	Mies	Kirvesmies	Tampere

Suojatun aineiston perusteella ei pitäisi enää olla riskiä edes sisäiseen paljastamiseen eli esimerkiksi paikkakuntalaisten ei pitäisi saada selville kuka kyseinen henkilö on. Suojattu aineisto on kuitenkin turhan suojattu monien asioiden tutkimiseen, esimerkiksi paikkakuntakohtaisia tai ikäryhmittäisiä eroja on vaikea tarkastella, jos aineistossa on paljon puuttuvia tietoja kyseisten muuttujien kohdalla.

7 Aineiston muuntamiseen perustuvat menetelmät

Aineiston muuntamiseen perustuvat menetelmät ovat hyvin käyttökelpoisia. Näiden menetelmien etu on tiedon varma suojaus erityisesti silloin, jos luovutettavat tiedot eivät sisällä juurikaan alkuperäisiä vastauksia vaan ovat mitä suurimmassa määrin muokattuja. Aineiston muuntamiseen perustuvissa menetelmissä on vaikeutena suorittaa suojaus niin, ettei informaatiota katoa liikaa ja toisaalta harhan kasvu saadaan minimoiduksi.

Aineiston muuntamiseen perustuvat menetelmät voidaan jakaa usealla eri tavalla, joista yksi on jako systemaattisiin ja satunnaisiin aineiston muuntamismenetelmiin. Tässä jaossa systemaattisia aineiston muuntamismenetelmiä ovat esimerkiksi mikroaggregointi ja deterministinen pyöristäminen, joissa muuntaminen tapahtuu kaikille soluille samalla tavalla. Satunnaisiin aineiston muuntamismenetelmiin kuuluvat taas arvojen vaihtaminen yksiköiden välillä ja kohinan lisääminen, joissa muuntamiseen liittyy todennäköisyyksiä ja siten muuntaminen eroaa solusta toiseen (Fienberg, 2000).

Suurin osa aineiston muuntamiseen perustuvista menetelmistä on matriisimuotoista suojaamista, jota käsiteltiin luvussa 4.2. Muistutuksen vuoksi tämä tarkoitti merkinnällisesti, että alkuperäisestä aineistosta X, on saatu suojattu aineisto Z seuraavasti

$$Z = AXB + C,$$

missä A on yksiköitä muuttava suojaus, B muuttujia muuttava suojaus ja C arvoja siirtävä suojaus eli kohinan lisääminen.

7.1 Kohinan lisääminen

Erityisesti jatkuvia muuttujia voi suojata lisäämällä muuttujan arvoon kohinaa. Luokitelluille muuttujille kohinan lisääminen soveltuu joissakin tapauksissa. Kohinan lisääminen on erityisen sopiva menetelmä jatkuvan muuttujan tapauksessa, koska menetelmä ei vaadi oletuksia aineiston määrittelyvälille, joka voi olla ääretönkin. Yksinkertaisimmillaan käytetään valkoista kohinaa, jolle pätee $\varepsilon_j \sim N(0, \sigma^2_{\varepsilon_j})$ ja lisäksi $Cov(\varepsilon_t, \varepsilon_r) = 0$, kun $t \neq r$. Toisin sanoen lisättävä kohina on jatkuvaa ja sen aiheuttama muutos aineiston arvoihin muuttaa epätodennäköiseksi sen, että mistään muusta aineistosta löytyisi yksikön muuttujanarvolle täydellistä vastaavuutta. Tietenkin kohinan lisäämisestä riippuen likimääräinen arvo tai väli, jolle arvo kuuluu on mahdollista päätellä.

Matemaattisesti kohinan lisäämistä voidaan tarkastella seuraavasti (Group Crises 2004d): valkoisen kohinan tapauksessa oletetaan, että x_j on alkuperäisen aineiston j:s sarake eli kyseisen muuttujan saamat arvot. Suojatussa aineistossa Z sarake saa arvot

$$z_j = x_j + \varepsilon_j,$$

missä ε_j on valkoista kohinaa.

Yleisesti kohinalla suojatessa ε_j:n varianssit ovat verrannollisia alkuperäisten muuttujien variansseihin. Eli jos σ^2_j on alkuperäinen varianssi aineistossa X_j, tulee uusi varianssi asettaa $\sigma^2_{\varepsilon_j} = \alpha\sigma^2_j$. Aineiston ollessa p-ulotteinen voidaan valkoisen kohinan avulla suojaaminen kirjoittaa muodollisesti

$$Z = X + \varepsilon,$$

missä $X \sim (\mu, \Sigma)$, $\varepsilon \sim N(0, \Sigma_\varepsilon)$ ja $\Sigma_\varepsilon = \alpha \cdot diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$ kaikilla $\alpha > 0$.

Korreloimattoman kohinan käytössä on se etu, että se säilyttää alkuperäisen aineiston keskiarvot ja kovarianssit:

$$\mathbb{E}(Z) = \mathbb{E}(X) + \mathbb{E}(\varepsilon) = \mathbb{E}(X) = \mu$$

$$Cov(Z_k, Z_l) = Cov(X_k, X_l) \quad \text{kaikilla } k \neq l.$$

Toisaalta valkoisen kohinan käytössä menetetään alkuperäisen aineiston varianssien ja korrelaatiokertoimien arvot:

$$Var(Z_j) = Var(X_j) + \alpha Var(X_j) = (1+\alpha)Var(X_j),$$

$$\rho_{Z_k,Z_l} = \frac{Cov(Z_k, Z_l)}{\sqrt{Var(X_k)Var(X_l)}} = \frac{1}{1+\alpha}\rho_{X_k,X_l} \quad \text{kaikilla } k \neq l.$$

Mikäli suojaaja katsoo tärkeämmäksi etteivät varianssit muutu, voidaan kohina lisätä korreloituneena. Toisin sanoen lisättävä kohina $\varepsilon \sim N(0, \Sigma_\varepsilon)$, missä $\Sigma_\varepsilon = \alpha\Sigma$ eli kohinan kovarianssimatriisi on verrannollinen alkuperäisen aineiston kovarianssimatriisiin. Tämän menetelmän kautta suojatun aineiston kovarianssimatriisille saadaan

$$\Sigma_Z = \Sigma + \alpha\Sigma = (1+\alpha)\Sigma$$

ja edelleen korrelaatiokertoimille

$$\rho_{Z_k,Z_l} = \frac{1+\alpha}{1+\alpha}\frac{Cov(X_k, X_l)}{\sqrt{Var(X_k)Var(X_l)}} = \rho_{X_k,X_l}.$$

Valitettavasti tämän menetelmän lopputuloksena ovat harhaiset estimaatit, mikä käy ilmi kovarianssimatriiseille pätevistä yhtälöistä. Tästä huolimatta suojatun aineiston käsittelijä voi estimoida alkuperäisen kovarianssimatriisin hyvin tarkkaan mikäli hänellä on käytössään arvo α.

Kohinan lisääminen korreloituneena johtaa suojatun aineiston parempaan analyyttisyyteen kuin korreloimattoman kohinan lisääminen. Aineiston avulla pystytään estimoimaan useita alkuperäisen aineiston kanssa yhtäpitäviä tilastollisia tuloksia, kunhan α arvo tunnetaan. Tällainen yksinkertainen kohinan lisääminen ei kuitenkaan ole paljoakaan käytetty menetelmä, koska sen antama suoja on erittäin alhainen.

Mikäli kohinan avulla suojaamista kuitenkin halutaan käyttää, saavutetaan parempi suoja, kun mukaan otetaan kohinan lisäksi lineaarinen muunnos. Lineaarisen muunnoksen lisääminen onnistuu vain jatkuville muuttujille.

Käytännössä suojausmenetelmä toteutetaan kahdessa osassa. Ensin lisätään alkuperäiseen aineistoon korreloimaton kohina kuten aiemmin on esitelty ja saadaan aineisto, jota merkitään

$$Z_j = X_j + \varepsilon_j, \quad \text{kun} \quad j = 1, 2, \ldots, p. \tag{7.1}$$

Tässä tapauksessa lisättäväksi kohinaksi voidaan valita muukin kuin normaalisti jakautunut. Yleensä valitaan sellainen kohina, joka noudattaa samaa jakaumaa kuin alkuperäinen aineisto. Seuraavassa vaiheessa muunnetaan edelleen aineistoa Z_j

$$G_j = cZ_j + d_j. \tag{7.2}$$

Kohinan lisäys (7.1) ja muunnos (7.2) voidaan kirjoittaa matriisimuodossa:

$$Z = X + \varepsilon$$

$$G = cZ + D = c(X + \varepsilon) + D, \tag{7.3}$$

missä $X \sim (\mu, \Sigma), \varepsilon \sim (0, \alpha\Sigma), G \sim (\mu, \Sigma)$ ja D on matriisi, jossa j:s sarake sisältää vakion d_j kaikilla riveillä. Muunnoksen (7.3) jakaumista ei voi olla varma, koska lineaarinen muunnos muuntaa viimeistään G:n jakauman erilaiseksi (normaalisti jakautuneet ovat toki poikkeus). Parametrien c ja d_j arvot saadaan seuraavien rajoitteiden avulla $\mathbb{E}(g_j) = \mathbb{E}(x_j)$ ja $Var(g_j) = Var(x_j)$ kaikilla $j = 1, 2, \ldots, p$. Itse asiassa nämä rajoitteet yhtälöparina johtavat yhtälöön

$$d_j = (1 - c)\mathbb{E}(x_j). \tag{7.4}$$

Kyseinen lineaarinen muunnos riippuukin vain yhdestä parametrista c. Tarkempi parametrin määrittäminen riippuu muunnoksen valinnasta. Yhtälö (7.4) johtaa kahteen vaihtoehtoon:

$$g_{j,1} = cz_j + (1 - c)\overline{x}_j$$
$$g_{j,2} = cz_j + (1 - c)\overline{z}_j.$$

Jos näistä muunnoksista valitaan jälkimmäinen, saadaan parametrille c yhtälö:

$$c = \sqrt{\frac{n - 1 - \alpha}{(n - 1)(1 + \alpha)}},$$

missä käytetään rajoitetta $\sigma_X = \sigma_G$, joka tietenkin johtaa rajoitteeseen $\Sigma_X = \Sigma_G$. Jos taas rajoitteista valitaan ensimmäinen, parametrin c arvolle saadaan:

$$c = \sqrt{\frac{n - 1}{n(1 + \alpha) - 1}},$$

joka on asymptoottisesti sama edellisen ratkaisun kanssa. Lisäksi pienille otoksille nämä kaksi ratkaisua eivät juurikaan eroa. Suurille otoksille parametrin c arvolle voidaan käyttää raja-arvoa:

$$\lim_{n \to \infty} c = \frac{1}{\sqrt{1 + \alpha}}.$$

Parametrin c arvot ovat nollan ja yhden välillä, koska $\alpha > 0$.

Tämänkin menetelmän avulla suojatusta aineistosta saadaan valkoisen kohinan tapaan samat odotusarvot ja kovarianssit kuin alkuperäisessä aineistossa on. Valitettavasti harhaa syntyy kuitenkin regressioanalyysissa sekä otoksiin kohdistuvissa analyyseissä vaikka muuten suojattu aineisto on analyyttisesti suhteellisen hyvä. Harha voidaan arvioida c:n arvon avulla ja mikäli niin halutaan, voidaan suojatun aineiston vastaanottajalle antaa tieto käyttöönsä. Arvon c paljastamisesta on valitettavasti seurauksena myös se, että suojatun aineiston suojaus voidaan helposti palauttaa valkoisen kohinan tilanteeseen ja haettu vahvempi suoja menetetään.

Viimeinen kohinan lisäämiseen perustuva menetelmä, josta tässä vielä mainitaan vahvistaa suojausta kohinan lisäyksen jälkeen vielä epälineaarisella muunnoksella. Tätä menetelmää kutsutaan algoritmin kehittäjän mukaan Sullivanin menetelmäksi ja siihen perehdytään tarkemmin seuraavassa luvussa.

7.2 Sullivanin menetelmä

Sullivanin menetelmä perustuu valkoisen kohinan lisäämiseen epälineaariseen muunnoksen kautta. Edellisessä luvussa on esitelty kohinan lisäämiseen perustuvat tietosuojamenetelmät ja siitä syystä tässä luvussa aloitetaan suoraan Sullivanin menetelmän esittelystä perehtymättä tarkemmin kohinan lisäämiseen. Luvun tarkastelu on seuraavaa algoritmia lukuun ottamatta Brandin artikkelista (2002).

Sullivanin menetelmä sopii sekä jatkuville että luokitelluille muuttujille ja säilyttää jakaumat toisin kuin lineaariseen muunnokseen perustuva menetelmä. Matemaattisesti Sullivanin menetelmä on erittäin raskas ja vaatimansa laskentatehon takia sen käyttöönotto on tullut mahdolliseksi vasta tietokoneiden tehojen kasvun myötä. Menetelmän monimutkaisuudesta johtuen esittely on syytä aloittaa menetelmän vaiheista muodostetun algoritmin esittelyllä (Group Crises, 2004d).

1. Lasketaan empiiriset jakaumafunktiot jokaiselle alkuperäisen aineiston muuttujalle.
2. Tasoitetaan jakaumafunktiot ennen niiden muuta käyttöä.
3. Muunnetaan tasoitetut empiiriset jakaumafunktiot yhtenäiseksi satunnaismuuttujaksi ja edelleen standardoidun normaalijakauman satunnaismuuttujaksi.
4. Lisätään kohina standardin normaalijakauman muuttujaan.
5. Muunnetaan saatu muuttuja takaisin jakaumafunktion arvoiksi.
6. Muunnetaan arvot edelleen takaisin alkuperäiselle asteikolleen.

Algoritmin edetessä käytetään etäisyyskriteeriä varmistamaan, ettei suojatun muuttujan ja sitä vastaavan standardoidun muuttujan etäisyys toisistaan ole kahden pienimmän etäisyyden joukossa. Tällä varmistetaan menetelmän tehokkuutta suojaamisessa.

Ensimmäisen ja toisen vaiheen muunnokset eroavat toisistaan riippuen siitä, onko muuttuja jatkuva vai luokiteltu. Jatkuvalle muuttujalle X empiirinen jakauma lasketaan ensimmäisessä vaiheessa. Toisessa vaiheessa laskettu jakauma muunnetaan tasoitetuksi jakaumaksi. Tämä tapahtuu laskemalla arvojen X_i keskiarvot $x_i = (X_i + X_{i+1})/2$, kaikille $i = 0, 1, \ldots, m$ ja lisäksi $x_0 = 2X_1 - X_2$ ja $x_{m+1} = 2X_m - X_{m-1}$. Näitä keskiarvoja käytetään tasoitetun empiirisen jakauman laskemiseen seuraavasti:

$$\overline{F}_x(z) = \hat{F}(x_{i-1}) + \frac{\hat{F}(x_i) - \hat{F}(x_{i-1})}{x_i - x_{i-1}}(z - x_{i-1}), \text{ kun } z \in (x_{i-1}, x_i],$$

missä x_i tarkoittaa laskettuja keskiarvoja ja $\hat{F}(x_i)$ empiirisen jakauman arvoja näillä keskiarvoilla. Yhtälön arvot lasketaan kaikille X:n arvoille ja merkitään $p_i = \overline{F}_x(X_i)$. Saadut arvot muunnetaan standardoituun normaalijakaumaan käyttämällä kertymäfunktiomenetelmää:

$$Z_i = \Phi^{-1}(p_i).$$

Nämä muunnokset ovat standardoidun normaalijakauman muuttujia, koska $\hat{F}(X_{ij}) \simeq \Phi(X_{ij})$. Muunnettujen muuttujien korrelaatiot ovat liki identtiset alkuperäisten muuttujien korrelaatioiden kanssa, kun kaikki muuttujat ovat normaalisia. Jos taas havaitut muuttujanarvot eivät noudata normaalijakaumaa, muunnettujen muuttujien korrelaatiot eroavat huomattavasti alkuperäisistä. Suuruusluokka riippuu empiirisen jakauman standardoitujen arvojen ja standardin normaalijakauman arvojen välisistä eroista.

Luokitellun muuttujan muuntaminen aloitetaan k luokkaa omaavan muuttujan jakamisella $k-1$ Bernoullin muuttujaksi. Tämän jälkeen lasketaan ehdollinen kovarianssimatriisi ($m_{dd.cc}$) apumuuttujille, jotka on johdettu jatkuvista muuttujista:

$$m_{dd.cc} = m_{dd} - {m_{cd}}^T m_{cc}^{-1} m_{cd},$$

missä m_{cc} tarkoittaa jatkuvien muuttujien kovarianssimatriisia, m_{dd} binääristen muuttujien kovarianssimatriisia ja m_{cd} kovarianssimatriisia, joka saadaan jatkuvien ja binääristen muuttujien välille. Kyseisissä merkinnöissä c ("continuos") viittaa jatkuviin muuttujiin ja d ("discrete") vastaavasti luokiteltuihin muuttujiin. Saadun matriisin avulla muodostetaan standardoitujen normaalisten satunnaislukujen F_{dc} matriisi, jossa sarakevektorille f_{dct}:

$$f_{dct} = {m_{cd}}^T m_{cc}^{-1} L_{cc}^2 Z_{ct} + m_{dd.cc}^{\frac{1}{2}} e_{d.ct},$$

missä Z_{ct} tarkoittaa havainnosta t muunnettujen jatkuvien muuttujien vektoria, $L_{cc}^2 = diag(m_{cc})$ on diagonaalimatriisi, jossa otoksen varianssit ovat diagonaalin alkioina ja $e_{d.ct}$ on standardoitujen normaalisten satunnaislukujen vektori. Huomataan, että vektorin f_{dct} arvot noudattavat likimain normaalijakaumaa odotusarvolla 0 ja kovarianssimatriisilla m_{dd}. Lisäksi (Z_{ct}, f_{dct}) on likimain normaalinen ja sillä on melkein sama korrelaatiomatriisi kuin alkuperäisellä aineistolla.

Vaikka f_{dct}:n korrelaatiot ovat liki samat kuin alkuperäisten Bernoullin muuttujien, eivät uudet arvot riipu alkuperäisistä. Jotta alkuperäisten Bernoullin muuttujien ja f_{dct}:n arvojen välille saadaan haluttu yhteys, on tehtävä lisää muunnoksia. Jakauman arvoille kaikilla $j = 1, \ldots, p_d$ asetetaan

$$g_{dct,j} = \Phi(f_{dct,j}).$$

Saadut arvot $g_{dct,j}$ ovat yleisten satunnaismuuttujien realisaatiot, eivätkä nekään vielä riipu Bernoullin muuttujista. Tästä johtuen satunnaismuuttuja $h_{dt,j}$ muodostetaan siten, että se riippuu sekä arvoista $g_{dct,j}$ että alkuperäisistä muuttujista $(x_{dt,j})$:

$$h_{dt,j} = \begin{cases} g_{dct,j}(1 - p_{oj}) & \text{jos } x_{dt,j} = 0, \\ 1 - p_{oj}(1 - g_{dct,j}) & \text{jos } x_{dt,j} = 1, \end{cases} j = 1, \ldots, p_d,$$

missä p_{oj} on j:nnen Bernoullin muuttujan keskiarvo. Seuraavaksi määritetään normaaliset järjestysluvut $R_{d1,j}, \ldots, R_{dn,j}$ aloittamalla pienimmästä muuttujan $h_{dt,j}$ saamasta arvosta:

$$\tilde{R}_{dt,j} = \frac{R_{dt,j} - 0,5}{n}.$$

Näin saadut arvot muunnetaan standardoidun normaalijakauman muuttujiksi kertymäfunktion avulla:

$$Z_{dt,j} = \Phi^{-1}(\tilde{R}_{dt,j}).$$

Yhdistämällä nämä arvot jatkuvien muuttujien kanssa on saatu standardin normaalijakauman muuttujanarvojen vektori jokaiselle havainnolle ts. $Z_t = ({Z_{ct}}^T, {Z_{dt}}^T)$.

Algoritmin neljännessä vaiheessa muunnetut muuttujat Z_t suojataan lisäämällä kohinaa. Tämä vaihe on pääpiirteiltään täsmälleen sama kuin luvun 7.1. valkoisen kohinan lisääminen. Olkoot Z muunnetun standardin normaalijakauman matriisi, jonka rivivektoreita merkitään Z_t ja olkoot U^* kohinan matriisi, jolle pätee $U^* \sim N(0, \alpha M_{ZZ})$. Näillä merkinnöin suojattujen muuttujien matriisin Z^* voi kirjoittaa:

$$Z^* = Z + U^* = Z + \sqrt{\alpha} U T_{ZZ}, \text{ kun } \alpha > 0,$$

missä $U \sim N(0, I_{p \times p})$ ja T_{ZZ} tarkoittaa muunnetun aineiston korrelaatiomatriisin (P_{ZZ}) tulontekijää eli matriisia, jolle ${T_{ZZ}}^T T_{ZZ} = P_{ZZ}$. Koska matriisien Z ja U^* osat ovat normaalisti jakautuneita, ovat myös suojatut arvot Z_t^* normaalisia: $Z^* \sim N(0, M_{Z^*Z^*})$, missä $M_{Z^*Z^*} = M_{ZZ} + \alpha M_{ZZ} = (1 + \alpha) M_{ZZ}$.

Kaikille suojatuille yksiköille lasketaan vektoriin m_i suojatun ja alkuperäisen havainnon välinen etäisyys, joka saadaan alkioittain

$$m_{it} = d_{it} \Sigma_d^{-1} d_{it}^T,$$

missä d_{it} on havaintojen i ja t välisien erojen vektori $d_{it} = z_i - z_t^*$, ja jonka kovarianssimatriisi on $\Sigma_d = \alpha M_{ZZ}$. Sopiva kriteeri etäisyydelle on, ettei mikään etäisyys m_{ii} saa olla kumpikaan pienimmistä etäisyyksistä vektorissa m_i. Mikäli etäisyys on toinen näistä pienimmistä, toistetaan suojaamista eli kohinan lisäämistä kunnes ehto täyttyy.

Algoritmin neljännessä ja viidennessä vaiheessa suojattu yksikkö palautetaan takaisin alkuperäiseen mittakaavaan. Normalisoidut järjestysnumerot D_j^* lasketaan kullekin muuttujalle Z_j^*, $j = 1, \ldots, p$. Tätä varten on muodostettava järjestysnumeroiden vektori R_j^* muuttujien Z_j^* alkiolle laskevassa järjestyksessä ja jakamalla havaintojen lukumäärällä n.

Tämä "empiirinen jakauma" on muunnettava edelleen takaisinmuunnosta varten, koska sen alkioiden arvot riippuvat suoraan otoskoosta. Virhetermit standardoidaan muuntamalla arvot ensin normaalijakaumaan

$$u_{tj}^+ = \frac{u_{tj}^*}{\sqrt{\frac{1}{n-1} \sum_{t=1}^n {u_{tj}^*}^2}}$$

ja lisäämällä sitten järjestyksen vaikutus

$$D_{tj}^* = \frac{R_{tj}^* - \phi(u_{tj}^+)}{n} = \frac{R_{tj}^*}{n} - \eta_{tj}, \ t = 1, \ldots, n \ ,$$

missä $\phi(\cdot)$ on funktio, joka muuntaa u_{tj}^{+} arvot nollan ja ykkösen väliin eli saadaan $0 \leq \eta_{tj} \leq \frac{1}{n}$. Näin on saatu D_{tj}^{*} rajoitetuksi välille

$$\frac{R_{tj}^{*}}{n} \leq D_{tj}^{*} \leq \frac{R_{tj}^{*} - 1}{n}.$$

Sopiva valinta funktioksi $\phi(\cdot)$ on standardin normaalijakauman tiheysfunktio, koska u_{tj}^{+} on normaalisti jakautunut ja tämä valinta johtaa siihen, että η_{tj} on samoin jakautunut.

Lopullinen takaisinmuunnos riippuu jälleen siitä, onko muuttuja jatkuva vai luokiteltu. Jatkuville muuttujille käytetään tasoitetun empiirisen jakauman muunnosta:

$$X_{ctj}^{*} = x_{i-1,j} + \frac{D_{tj}^{*} - \hat{F}(x_{i-1,j})}{\hat{F}(x_{i,j}) - \hat{F}(x_{i-1,j})}(x_{i,j} - x_{i-1,j}), \text{ missä } D_{tj}^{*} \in (\hat{F}(x_{i-1,j}), \hat{F}(x_{i,j})].$$

Binäärisille muuttujille muunnos suoritetaan kaavalla:

$$X_{dtj}^{*} = \begin{cases} 0 & \text{jos } D_{tj}^{*} \in (0, 1 - p_{oj}), \\ 1 & \text{jos } D_{tj}^{*} \in [1 - p_{oj}, 1), \end{cases} \text{ kun } t = 1, \ldots, n \text{ ja } j = 1, \ldots, p_d.$$

Tällainen takaisinmuunnos varmistaa jakaumien likimääräisen säilymisen ja siten myös otoskeskiarvot ja -varianssit ovat likimain samat kuin alkuperäisellä aineistolla. Kuitenkin korrelaatioihin tulee muutoksia useiden numeeristen muutosten sekä normaalijakauman käytöstä alkuperältään ei-normaalisten muuttujien muokkaamiseen. Myös korrelaatiot X:n ja X^{*}:n välillä eroavat muuttujasta toiseen, mikä tarkoittaa, että kohinan suhteellinen määrä vaihtelee. Näiden ongelmien korjaamiseksi ehdotetaan kahta eri iteraatiota (Brand, 2002). Ensinnäkin ristikorrelaatiot muuttujien ja näiden suojattujen vastineiden välillä säädetään robustiin keskiarvoon. Minimoidaan korrelaatiomatriisien arvojen väliset erot. Koska $\mathbb{E}(X_j) = \mathbb{E}(X_j^{*})$, perustuu ristikorrelaatioiden säätäminen ehdolle:

$$X_{tj}^{*} = \rho_{X_j X_j^{*}} + \eta_{tj},$$

missä $\rho_{X_j X_j^{*}}$ tarkoittaa X_j ja X_j^{*} välistä korrelaatiota ja η_{tj} on aineistosta X_j riippumaton virhetermi, jolle $\eta_j \sim (0, \sigma_{\eta_j}^2)$, $Cov(X_j \eta_j) = 0$ ja $\sigma_{\eta_j}^2 = \sigma_{X_j}^2(1 - \rho_{X_j X_j^{*}}^2)$.

Tästä saadaan korrelaatiolle

$$\rho_{X_j X_j^{*}}^2 = 1 - \frac{\sigma_{\eta_j}^2}{\sigma_{X_j}^2}.$$

Voidaan olettaa, että alkuperäisten ja suojattujen muuttujanarvojen välinen korrelaatio on positiivinen, jolloin $\sigma_{\eta_j}^2 < \sigma_{X_j}^2$. Korrelaatio siis kasvaa, kun $\sigma_{\eta_j}^2$ vähenee. Korrelaatiota $\rho_{X_j X_j^{*}}^2$ vastaavien korrelaatioiden säätämiseen voidaan hyödyntää virhetermien muokkaamista.

Kohdekorrelaation ollessa robusti keskiarvo korrelaatiosta $\overline{\rho}$, saadaan

$$\overline{\rho} = \frac{\sum_{j=1}^{p} r_{X_j X_j^{*}} - (\max_{1 \leq j \leq p} r_{X_j X_j^{*}} + \min_{1 \leq j \leq p} r_{X_j X_j^{*}})}{p - 2},$$

missä $r_{X_jX_j^*}$ tarkoittaa X_j ja sen suojatun vastineen X_j^* välistä otoskorrelaatiota.

Varianssien vaihtelun määrän määrittämiseen käytetään yksinkertaista lineaarista approksimaatiota. Muunnosmatriisi B_{aa} määritellään seuraavien alkioiden avulla

$$b_{aaij} = \begin{cases} \frac{1-\overline{\rho}}{1-0{,}5(\overline{\rho}+r_{X_jX_j^*})}, & \text{jos } j = i \text{ ja } i, j = 1, \ldots, p \\ 0 & \text{muulloin} \end{cases},$$

ja uudet standardinormaalijakaumaa noudattavat arvot lasketaan

$$Z_t^* = Z_t + u_t^* B_{aa}.$$

Nämä saadut arvot muunnetaan edelleen takaisin alkuperäiselle mitta-asteikolle. Vaihe toteutetaan iteraatioiden avulla ja niitä toistetaan kunnes suojaamisen kautta saadut ristikorrelaatiot eroavat halutuista ristikorrelaatioiden arvoista on annettua kynnysarvoa vähemmän tai kunnes ennalta määrätty määrä iteraatioita on saavutettu.

Koska suojatun ja alkuperäisen aineiston korrelaatiot yleensä eroavat toisistaan, voidaan käyttää vielä toista iteratiivista säätömenetelmää, jotta diagonaalin ulkopuoliset alkiot näissä kahdessa matriisissa saataisiin liki identtisiksi. Perusideana on edelleen käyttää lineaarista muunnosta virhetermien arvioimiseksi. Muunnokset suoritetaan peräkkäin aloittaen muuttujasta, joka saa $\sum_k (\rho_{X_jX_k} - \rho_{X_j^*X_k^*})^2$ suurimman arvon.

Virhetermien muokkaamista varten muodostetaan jonkin valitun alkuperäisen muuttujan (Z_1) ja muiden suojattujen muuttujien (Z_j^*) arvojen perusteella lineaarinen muunnos H_1^*:

$$H_1^* = b_0 Z_1 + \sum_{j=1}^{p} b_j Z_j^* = Z_1^{+T} b,$$

missä $b_0 = 1 - b_1$. Yhtälöpari, joka toteuttaa toivotut ominaisuudet, on

$$\begin{cases} r(G_1^*, X_1) = \kappa \\ r(G_1^*, X_1^*) = r_{X_1,X_1}, \end{cases}$$

missä G_1^* tarkoittaa takaisin muunnettua muuttujaa, joka vastaa muuttujaa H_1^* ja κ on muuttujan X_1 ja suojattujen muuttujien $X_2^*, \ldots, X_p^*$ välisten ristikorrelaatioiden aritmeettinen keskiarvo eli $\kappa = \frac{1}{p-1} \sum_{j=2}^{p} r_{X_1X_j^*}$. Kertoimet saadaan laskettua ratkaisemalla

$$\Sigma_{Z^+} b = \rho^+,$$

missä Σ_{Z^+} tarkoittaa muuttujien $Z^+ = (Z_1, Z_2^*, \ldots, Z_p^*)$ korrelaatiomatriisia ja $\rho^{+T} = (\kappa, r_{X_1,X_2}, \ldots, r_{X_1,X_p})$. Tämän jälkeen H_1^* arvot voidaan laskea lineaarisen muunnoksen kaavaa käyttäen ja muuntaa takaisin alkuperäiselle mitta-asteikolle algoritmin

neljättä ja viidettä vaihetta noudattaen. Approksimaatiota voidaan toistaa iteratiivisesti kunnes korrelaatioiden konvergointikriteeri täyttyy tai ennalta määrätty maksimimäärä iteraatioita on suoritettu.

Edellä ollut menetelmä soveltuu sekä binääristen että jatkuvien muuttujien suojaamiseen. Luokitelluille muuttujille, joilla on enemmän kuin kaksi luokkaa, lopullinen takaisinmuunnos vaatii oman määrittelyn. Olkoon Z_t^* vektori, joka on määritelty seuraavasti:

$$Z_{ti}^* = \begin{cases} X_{dt1}^* & \text{kun } i = 1 \\ (1 - \sum_{j=1}^{i-1} Z_{tj}^*) X_{dti}^* & \text{kun } i = 2, \ldots, k-1 \end{cases}.$$

Jos suojatun muuttujan X_d^* alkiot määritellään $X_{dt}^* = i$, jos $Z_{ti}^* = 1$ kaikille i.

Sullivanin menetelmä yhdistää kohinan lisäämisen ja muunnokset, jotka eivät yleensä ole lineaarisia. Epälineaaristen muunnosten valinta varmistaa, että muuttujakohtaiset jakaumat säilyvät ainakin likimain. Mikäli jakaumissa on vielä liian suuria eroja, voidaan lisäksi käyttää iteratiivisia menetelmiä, joiden tavoitteena on korjata muunnosten ja suojauksen aiheuttamia eroja korrelaatioissa. Korjauksista johtuen kaikkien muuttujien suojauksen taso ei ole sama.

Sullivanin menetelmällä suojatun aineiston otoskeskiarvot ovat alkuperäisen aineiston odotusarvojen harhattomia estimaatteja. Lisäksi suojattujen binääristen muuttujien otosvarianssit ovat alkuperäisten binääristen muuttujien varianssien harhattomia estimaatteja. Kuitenkin jatkuvien muuttujien varianssit kasvavat, koska

$$V(X_{cj}^* | x_{0,j}, x_{1,j}, \ldots, x_{n,j}) = V(X_{cj}) + \sum_{i=1}^{n} \frac{(x_{i,j} - x_{i-1,j})^2}{12n},$$

missä $x_{i,j}$ viittaa määrittelyjoukon rajoihin. Varianssin kasvu on sitä suurempi, mitä suurempi on otoksen koko ja mitä suuremmat ovat erot alkuperäisen aineiston havaittujen arvojen välillä. Tästä tuloksesta saadaan suojattujen muuttujien avulla lasketulleotosvarianssille $\hat{V}(X_{cj}^*) = \frac{1}{n} \sum_{i=1}^{n} (X_{c,ij}^* - \overline{X}_{cj}^*)^2$, joka on alkuperäisen aineiston otosvarianssin harhainen estimaattori. Alkuperäisen aineiston vaihtelu tulee yliestimoiduksi, koska muunnoksista seuraa ylimääräistä vaihtelua luokkien sisälle. Kuitenkin suojatun aineiston otosvarianssi on alkuperäisen aineiston varianssin tarkentuva estimaattori, koska sen lisäystermin arvot lähestyvät nollaa, kun $n \to \infty$.

Algoritmi on sen verran monimutkainen, ettei muiden ominaisuuksien tarkempia tarkastelu ole mielekästä. Pääasiassa tämä johtuu siitä, että virheiden jakaumaa alkuperäisellä mitta-asteikolla on liki mahdotonta määrätä eksplisiittisesti. Yleiset vaikutukset regressioestimaatteihin voidaan määrittää käyttäen apuna muuttujien virhetermien mallintamista.

Yhteenvetona Sullivanin menetelmästä voidaan todeta sen olevan monimutkainen yhdistelmä kohinan lisäämistä ja epälineaarisia muunnoksia. Muuttujakohtaiset jakaumat säilyvät likimain. Kuitenkin jatkuvien muuttujien varianssit kasvavat pienissä otoksissa muunnosten rakenteen takia. Algoritmin iteratiivisten tarkennusten avulla varmistetaan, että alkuperäisten muuttujien korrelaatiorakenne säilyy suojauksessa.

7.3 Arvojen vaihtaminen yksiköiden välillä

Data swapping eli arvojen vaihtaminen yksiköiden välillä on mikroaineistojen suojausmenetelmä, joka esiteltiin alun perin ainoastaan luokiteltuja muuttujia sisältävän aineiston suojaamiseen. Sen ideana on yksinkertaisesti muuntaa aineistoa vaihtamalla muuttujien arvoja eri yksiköiden kesken. Vaihtamisen tulee tapahtua niin, että marginaalisummat pätevät edelleen vaihtamisen jälkeenkin. Vaikka alkuperäinen menetelmä ei saanut suurta kannatusta, oli se lähtökohtana, kun menetelmää kehitettiin edelleen. Tällä kertaa kehitys tapahtui sekä luokiteltuja että jatkuvia muuttujia sisältävää mikroaineistoa koskevaksi. Menetelmän hyödyllisyys kasvoi, kun arvojen vaihtamista ruvettiin rajoittamaan. Uuden version nimeksi tuli rank swapping eli arvojen vaihtaminen yksiköiden välillä järjestykseen perustuen.

Menetelmän perusideassa (Group Crises, 2004e) tietyn muuttujan X_i arvot järjestetään ensin kasvavaan järjestykseen. Muuttujan X_i järjestetyistä arvoista valitaan satunnaisesti tietyltä väliltä kaksi, joiden paikat vaihdetaan. Tietyllä välillä tarkoitetaan sitä, ettei arvoja tule vaihtaa täysin satunnaisesti vaan on järkevämpää valita ennakkoon raja, joka rajoittaa vaihdettavien arvojen "etäisyyden" toisistaan. Raja voidaan valita esimerkiksi niin, ettei vaihdettavien arvojen järjestysluvut saa erota liikaa toisistaan eli järjestyslukujen erotus ei saa olla enempää kuin tietyn prosenttiosuuden verran koko aineiston yksiköiden lukumäärästä. Esimerkiksi aineiston koon ollessa $N = 1500$ prosenttiosuuden 10% valinnasta seuraa, että vaihdettavien yksiköiden järjestyslukujen eron on oltava alle 150. Esitellyn algoritmin mukaiset muutokset suoritetaan toisistaan riippumattomasti jokaisen muuttujan kohdalla.

Seuraavassa esimerkissä tarkastellaan kuvitteellista aineistoa, jossa on tietoja eri ikäisten henkilöiden asuntojen koosta ja vuokran suuruudesta. Aineiston suojaaminen aloitetaan järjestämällä muuttujien arvot suuruusjärjestykseen ennen vaihtamista. Suuruusjärjestyksestä pidetään kiinni vaihtamalla muuttujien arvoja korkeintaan yhden yksikön yli eri yksiköiden välillä. Muutokset toteutetaan muuttuja kerrallaan muista muuttujista riippumattomasti.

Taulukko 5: Arvojen vaihtaminen yksiköiden välillä, alkuperäinen esimerkkiaineisto

Havainto	Ikä	Asunnon koko m^2	Vuokra/kk
1	19	45	570
2	25	23	220
3	28	67	630
4	29	72	780
5	33	78	810
6	37	157	1120
7	38	128	1050
8	45	135	1340
9	46	59	790
10	49	45	450

Taulukko 6: Arvojen vaihtaminen yksiköiden välillä järjestykseen perustuen prosenttiosuus 20%

Havainto	Ikä	Asunnon koko m^2	Vuokra/kk
1	19	45	220
2	29	45	570
3	28	67	630
4	25	59	780
5	37	78	810
6	33	157	1120
7	46	128	790
8	45	135	1340
9	38	72	1050
10	49	23	450

On perusteltua olettaa, että arvojen vaihtamisella järjestykseen perustuen suojatusta aineistosta saadaan vähemmän muuttuneita tilastollisia arvoja kuin alkuperäisellä menetelmällä, jossa arvot vaihdettiin satunnaisesti. Myöhemmät empiiriset tarkastelut, kuten Domingo-Ferrerin ja Torran (2001) tutkimus, ovat osoittaneet, että arvojen vaihtaminen yksiköiden välillä järjestykseen perustuen on käyttökelpoinen menetelmä, kun arvio perustetaan paljastumisriskin ja informaatiokadon määrään.

Esimerkkiaineiston tapauksessa voidaan todeta, ettei aineiston perusteella tehtävät johtopäätökset juurikaan muutu, kun aineiston arvoja on vaihdettu yksiköiden välillä järjestykseen perustuen. Alkuperäisen aineiston avulla saadaan alle 30-vuotiaiden asuntojen keskimääräiseksi kooksi 52 m^2 ja vuokraksi 550 euroa, kun vastaavat arvot 30–39-vuotiaille ovat 121 m^2 ja 993 euroa ja yli 40-vuotiaille 80 m^2 ja 860 euroa. Vastaavasti suojatusta aineistosta saadut arvot asuntojen koolle ovat: alle 30-vuotiaille 54 m^2, 30–39-vuotiaille 102 m^2 ja yli 40-vuotiaille 95 m^2. Tällä suojauksella keskivuokrat ovat pysyneet ennallaan. Kuitenkin keskimääräinen neliövuokra eri kokoisille asunnoille on muuttunut. Esimerkiksi suurilla asunnoilla (yli 100 m^2) keskineliövuokra on muuttunut 8,4 eurosta 7,7 euroon ja vastaavasti keskikokoisten asuntojen (50–99 m^2) keskineliövuokra on noussut 10,9 eurosta 11,8 euroon. Pienten asuntojen osalta muutosta ei ole tapahtunut. Muutokset vaikuttavat nyt osin suuriltakin, mutta aineistoissa, joissa arvojen erot eivät ole näin suuria, muutoksetkin ovat pienempiä. Toisaalta tässäkin aineistossa olisi saatu suurempiakin eroja, jos arvojen vaihtamiset olisi valittu toisin.

7.4 Mikroaggregointi

Mikroaggregointi on eräs erityisesti jatkuvan mikroaineiston suojaamiseen sopiva menetelmä. Vasta viime vuosina menetelmää on kehitelty luokiteltua aineistoa suojaavaksi. Mikroaggregoinnissa on ideana, että alkuperäisen aineiston sijaan julkaistaankin ryhmitelty eli aggregoitu aineisto. Aineistoa muokataan niin, että alkuperäisen aineiston yksiköt ryhmitellään ominaisuuksien mukaan ryhmiin, joissa yksiköiden lukumäärä on tietty ennalta valittu k tai tarvittaessa samaan ryhmään voidaan laittaa

suurempikin määrä yksiköitä. Lisäksi vaatimuksena on, ettei mikään ryhmän yksiköistä dominoi eli vaikuta liikaa ryhmästä julkaistaviin keskiarvoihin. Toisin sanoen julkaistu aineisto kertoo alkuperäisen aineiston yksiköistä koottujen k:n kokoisten ryhmien keskiarvot.

Group Crisesin (2004d) tutkimuksen mukaan mikroaggregointi toteutetaan käytännössä jakamalla alkuperäisen aineiston n yksikköä kunkin johonkin ryhmään (yhteensä g kpl), joiden jokaisen koko on vähintään k. Ryhmät muodostetaan mahdollisimman samankaltaisista yksiköistä. Jokaiselle ryhmälle lasketaan kunkin p muuttujan arvojen keskiarvo, jolla korvataan ryhmän jokaisen yksikön arvot. Nämä muunnetut arvot ovat lopulta ne, jotka on turvallista julkaista.

Muodollisempi tarkastelu lähtee olettamuksesta, että meillä on aineisto, jossa on n yksikköä, joilla on arvot p muuttujalle. Yksikköä voidaan merkitä $X^T = (X_1, \ldots, X_p)$, missä X_i ilmaisee muuttujien arvoja. Näillä ehdoin muodostuu siis g ryhmää, joissa kussakin on $n_i \geq k$ yksikköä ja $\sum_{i=1}^{g} n_i = n$. Merkitään edelleen x_{ij} i: nnen ryhmän j:nnettä yksikköä, jolloin keskiarvovektori, joka sisältää jokaisen muuttujan keskiarvot, i:nnessä ryhmässä merkitään $\overline{x}_i$ ja koko aineiston yksiköiden keskiarvot sisältävää vektoria $\overline{x}$. Näillä merkinnöillä informaatiokadon kannalta optimaalisin jako k-osajoukkoihin saadaan maksimoimalla ryhmän sisäinen yhtenäisyys. Jos ryhmän yksiköt ovat mahdollisimman samanlaisia ja niiden arvot korvataan kyseisen ryhmän keskiarvoilla, menetetään luonnollisesti vähiten tietoa. Neliösumma-kriteeriä on yleisesti käytetty ryhmän sisäisen homogeenisuuden mittana klusteroinnissa. Määritelmä sisäiselle neliösummalle SSE on

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^T (x_{ij} - \overline{x}_i).$$

Mitä pienempi SSE sitä suurempi sisäinen homogeenisyys. Vastaavasti neliöiden yhteissumma määritellään

$$SST = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^T (x_{ij} - \overline{x}).$$

Neliösummien avulla etsitty paras k -ryhmiin jako minimoi SSE-arvon.

Vaikka tilanne näyttääkin yksinkertaiselta, on neliösummia mahdollista lähteä laskemaan useille eri tavoin ryhmitellyille tapauksille. Klassinen tapa toteuttaa ryhmittely on valita kaikki ryhmät mahdollisimman samankokoisiksi eli käytännössä valita muut paitsi ehkä yksi ryhmä kokoa k oleviksi. Edellä esitelty tapa muodostaa ryhmät eri kokoisiksi, mutta lukumäärää k suuremmiksi, on saanut nimekseen *aineisto-orientoitunut mikroaggregointi*. Aineisto-orientoitunut valinta voi olla hyvin selvästi edullisempi informaation säilymisen kannalta, koska aineiston pakottaminen jakautumaan kiinteäksi valittuun ryhmäkokoon saattaa aiheuttaa erittäin epähomogeenisia ryhmiä. Tällaisia tilanteita, joissa yksiköt luonnollisesti jakautuvat muuttujan arvoiltaan hyvin erilaisiin ryhmiin, on sen verran vähän, että kiinteän ryhmä koon avulla saatava laskennallinen etu kääntää tilanteen sen voitoksi. Toisin sanoen kiinteätä ryhmäkokoa kannattaa käyttää aina, mikäli aineistosta ei selvästi nouse valmista ryhmäjakoa.

Jos $p = 1$ eli kyseessä on yhden muuttujan aineisto tai aineisto on jo ryhmitelty muuttuja kerrallaan, on optimaalinen k-ryhmien määrittely olemassa ja mikroaggregointi

voidaan suorittaa. Mikäli taas $p > 1$ on eksaktin optimaalisen ratkaisun löytyminen ongelmalle todettu NP-kovaksi. Toisin sanoen ainoa mahdollisuus on ottaa käyttöön vähemmän eksakteja menetelmiä tai on hyödynnettävä yhden muuttujan tilanteen algoritmeja tähänkin tapaukseen.

Valitettavasti algoritmin soveltaminen monimuuttuja-aineistoon ei johda haluttuihin lopputuloksiin, koska muuttuja kerrallaan tehtävä mikroaggregointi ei tuota riittävän suojattua aineistoa ja lisäksi useille muuttujille sovellettaessa algoritmi on hyvin hidas. Siksi parempi tapa on käyttää algoritmia, jolla saadaan vastaus molempiin ongelmiin ja joka toimii myös käytännössä. Menetelmiä, joilla monimuuttuja-aineistoa voidaan mikroaggregoida:

- Käytetään yhden muuttujan menetelmää monimuuttuja-aineistoon mikroaggregoimalla muuttuja kerrallaan eli kukin muuttuja mikroaggregoidaan vuorotellen ja riippumattomasti. Tätä menetelmää voi nähdä kutsuttavan nimellä muuttujien yksittäinen järjestäminen (eng. individual ranking). Vaikka menetelmä olisikin suhteellisen nopea, on senkin ongelmana korkea paljastumisriski, josta mainittiin jo yhden muuttujan optimaalisen mikroaggregoinnin yhteydessä.

- Monimuuttuja-menetelmin, kuten monimuuttuja-aineiston järjestäminen järjestykseen projektoimalla akselille pääkomponenttianalyysin tai z-pistemäärän avulla tai käyttämällä suoraan projektoimatonta aineistoa. Kun käytetään projektoimatonta aineistoa, voidaan aineiston kaikki muuttujat mikroaggregoida samalla tai ne voidaan mikroaggregoida riippumattomasti kahden, kolmen tai useamman muuttujan ryhmissä. Suojatun aineiston hyödyllisyyden kannalta on parasta valita yhtä aikaa mikroaggregoitavaksi muuttujia, jotka korreloivat keskenään.

Esimerkkinä projektoimattoman monimuuttuja-aineiston kiinteän ryhmäkoon mikroaggregoinnista toimikoot "Maximum Distance to Average Vector" eli MDAV-algoritmi.

MDAV-algoritmi (Group Crises 2004e)

1. Laske koko aineiston keskiarvot $\overline{x}$ eli jokaisen muuttujan keskiarvo yli koko aineiston. Etsi aineistosta keskiarvovektorin arvoista $\overline{x}$ muuttujan arvoiltaan kaikkein eniten eroava yksikkö x_r. Erotus lasketaan käyttäen mittana euklidista etäisyyttä.
2. Etsi äsken löytämästäsi yksiköstä x_r mitattuna kaikkein kaukaisin yksikkö x_s.
3. Lajittele koko aineisto kahteen eri ryhmään edeltävien x_r ja x_s yksiköiden avulla. Valitse ensimmäiseen ryhmään x_r ja $k-1$ muuttujan arvoiltaan sitä lähintä yksikköä ja vastaavasti toiseen ryhmään x_s ja $k-1$ sitä lähintä yksikköä.
4. Mikäli jakamisen jälkeen jää ainakin $3k$ yksikköä, jotka eivät vielä kuulu kumpaankaan edellisessä vaiheessa muodostettuun ryhmään, on aloitettava ensimmäisestä vaiheesta käyttäen aineistona näitä ryhmien ulkopuolelle jääneitä yksiköitä.
5. Mikäli taas jakamisen jälkeen ryhmiin kuulumattomien yksiköiden lukumäärä on välillä $2k$ ja $3k-1$, jatketaan seuraavasti:

(a) Lasketaan keskiarvot $\overline{x}$ vastaavasti kuin ensimmäisessä vaiheessa, mutta nyt käyttäen aineistona vain näitä ryhmien ulkopuolelle jääneitä yksiköitä.
(b) Etsitään jälleen muuttujan arvoiltaan uudesta keskiarvovektorista $\overline{x}$ eroava yksikkö x_r.
(c) Muodostetaan yksi ryhmä yksikön x_r ympärille $k-1$ yksiköstä, jotka ovat muuttujan arvoiltaan sitä lähimpänä.
(d) Lopuista yksiköistä muodostetaan oma ryhmänsä. Ja poistutaan tästä algoritmista.

6. Mikäli kahden ryhmän ulkopuolelle on jäänyt vähemmän kuin $2k$ yksikköä, muodostetaan näistä yksiköitä oma ryhmänsä ja lopetetaan algoritmi.

Tässä esitelty algoritmi voidaan toteuttaa myös riippumattomasti kullekin muuttujien ryhmälle, joka on saatu jakamalla alkuperäinen aineisto (korreloituneiden) muuttujien mukaisiin ryhmiin.

Menetelmän käytännön esimerkkinä tarkastellaan kuvitteellista aineistoa eri ikäisten henkilöiden asuntojen koosta ja vuokran suuruudesta. Aineisto suojataan kahdella eri tavalla toteuttamalla mikroaggregointi ensin muuttuja kerrallaan ja sitten käyttäen ryhmien määrittelyssä MDAV-algoritmiä.

Taulukko 7: Mikroaggregoinnin esimerkkiaineisto

Havainto	Ikä	Asunnon koko m^2	Vuokra/kk
1	19	45	570
2	25	23	220
3	28	67	630
4	29	72	780
5	33	78	810
6	37	157	1120
7	38	128	1050
8	45	135	1340
9	46	59	790

Aineiston suojaaminen on varmempaa, jos ryhmien muodostamiseen käytetään MDAV-algoritmiä. Muuttujanarvojen keskiarvot ovat seuraavat: iälle 33,3 vuotta, asunnon koolle 84,8 m^2 ja asunnon kuukausi vuokralle 812,2 euroa. Euklidisen etäisyyden perusteella kaukaisimmaksi yksiköksi paljastuu havainto 2. Havainnosta 2 kaukaisin yksikkö on havainto 8. Kaksi havaintoa 2 lähimpinä olevaa havaintoa ovat havainnot 1 ja 3. Vastaavasti havaintoa 8 lähimpinä ovat havainnot 6 ja 7. Koska tämän jälkeen jäljellä on enää kolme yksikköä, muodostavat ne viimeisen ryhmän. Näiden ryhmien perusteella muodostettu suojattu aineisto on esitelty seuraavassa taulukossa.

Jos nyt verrataan ensimmäisenä suojattua aineistoa (taulukko 8) jälkimmäisenä suojattuun (taulukko 9), voidaan huomata ero ainakin siinä, ettei jälkimmäisestä aineistosta saada selville jonkun aineiston henkilön olevan vanhimpaan ryhmään kuuluva, mutta silti asuvan pienessä asunnossa. Jo tällaisten tietojen avulla voidaan usein selvittää poikkeuksellisten yksiköiden identiteetti.

Taulukko 8: Mikroaggregoitu muuttuja kerrallaan. Ryhmäkoko on 3 havaintoa.

Havainto	Ikä	Asunnon koko m^2	Vuokra/kk
1	24	43	473
2	24	43	473
3	24	72	473
4	33	72	793
5	33	72	793
6	33	140	1170
7	43	140	1170
8	43	140	1170
9	43	43	793

Taulukko 9: Mikroaggregoitu muodostaen ryhmät MDAV-algoritmin avulla. Ryhmäkoko on 3 havaintoa.

Havainto	Ikä	Asunnon koko m^2	Vuokra/kk
1	24	45	473
2	24	45	473
3	24	45	473
4	36	70	793
5	36	70	793
6	40	140	1170
7	40	140	1170
8	40	140	1170
9	36	70	793

7.5 Luokkien muuttaminen

Eräs tavallisimmista suojausmenetelmistä on luokittelun väljentäminen eli karkeistaminen. Jos aineiston muuttujien arvot ovat turhan yksityiskohtaisia ja siten aiheuttavat helposti paljastumista tai jopa identifioitumista, on yksityiskohtia helppo vähentää muuttamalla muuttujien luokkia laajemmiksi. Yksinkertaisimpia esimerkkejä lienevät iän luokitteleminen esimerkiksi viiden vuoden pituisiin luokkiin tai ammattien luokitteleminen tarkasta ammattinimikkeestä laajempaan.

Muut luokkien muuttamiseen perustuvat menetelmät on jaettu kahteen eri ryhmään, koska ne muodostuvat luonnollisesti eri tavoista toteuttaa muutos. Luokkien vähentämisellä viitataan menetelmään, jossa luokat käydään läpi yli kaikkien muuttujanarvojen. Luokkien laajentaminen ylä- tai alareunoista viittaa suojaamistapaan, jossa muuttujanarvot ovat jakautuneet likimain normaalisti ja ääriarvojen sijoittaminen samoihin luokkiin parantaa näiden pienten frekvenssien luokkiin sijoittuneiden yksiköiden suojausta. Menetelmiä voidaan käyttää myös yhtä aikaa eli mikäli paljastumisriski muuttujan pienille tai suurille arvoille vaikuttaa luokkien vähentämisenkin jälkeen liian suurelta, voidaan luokkia muuttaa pelkästään näiden arvojen osalta.

Luokkien muuttamisen tavoitteena on saada suojattavassa aineistossa olevien ainutlaatuisten tai harvojen arvoyhdistelmien eli profiilien frekvenssit kasvamaan. Yleensä aineistoa suojatessa on käytössä ennalta määrätty kynnysarvo, jota enemmän yksiköitä kullekin identifioivien muuttujien arvojen yhdisteelle on oltava, jotta aineisto voidaan tulkita suojatuksi. Toisin sanoen, jos joku arvojen yhdistelmä on sellainen, että se esiintyy harvemmin kuin kynnysarvoksi valittu luku, on kyseinen aineisto vielä lisäsuojauksen tarpeessa.

Selvennyksen vuoksi otetaan esimerkiksi aineisto, jossa muuttujia ovat sukupuoli, ammatti ja ikä. Tarkastellaan eri profiileita kynnysarvon kannalta. Oletetaan esimerkiksi, että kynnysarvo 5 olisi aineistolle sopiva. Tällöin, jos aineistosta löytyy esimerkiksi vain 4 naispuolista 57-vuotiasta kirurgia, on jokainen 57-vuotias naiskirurgi on tulkittavissa paljastumisvaarassa olevaksi yksiköksi. Vaikka kynnysarvo annetaan usein perusjoukon kokoon suhteutettuna frekvenssinä, on suojaajan käytettävissä yleensä vain otos koko perusjoukosta. Tällöin on järkevää muokata kynnysarvoa otokseen sopivaksi ja käyttää sitä pelkän otoksen suhteen.

Luokkien vähentäminen

Jos aineistossa on liian yksityiskohtaisia tietoja yksiköistä, on mahdollista siirtyä yleisempään luokitteluun. Esimerkiksi Suomessa henkilöaineistossa yksilön asuinpaikkakunnan kertominen on yleensä liian yksityiskohtainen tieto, joten onkin perusteltua siirtyä ainakin seutukuntien tasolle. Tällöin usean eri osoitteen sisältävä paikkakunta liitetään yhteen useamman paikkakunta-tasoisen luokan kanssa ja siten luokkien määrä vähenee ja yksilön paljastumisriski pienenee. Seurauksena luokkien vähentämisestä on tietysti paljastumisriskin pienenemisen ohella myös informaatiokato, joka joskus voi haitata tutkijoita, jotka olisivat kiinnostuneita tarkemmista tiedoista.

Tarkastellaan luokkien vähentämistä seuraavaksi muodollisemmin. Olkoot X_i luokiteltuja muuttujia, joista useita yhdistetään yhteiseen luokkaan muodostamaan uuden muuttujan Z_i. Näille muuttujille pätee $|D(Z_i)| < |D(X_i)|$, missä $|\cdot|$ tarkoittaa normia. Mikäli tässä halutaan miettiä jatkuvia muuttujia, tarkoittaa jatkuvan muuttujan X_i uudelleen luokittelu itse asiassa uuden luokitellun muuttujan Z_i muodostamista. Tosin sanoen mahdollisesti äärettömän määrittelyvälin omaava $D(X_i)$ on

kuvattu äärelliseksi $D(Z_i)$. Kuitenkin ensisijaisesti menetelmä on tarkoitettu luokitellun aineiston muuntamiseen niin, että ainutlaatuisten tapausten määrä aineistossa saadaan vähenemään. Käytettäessä jatkuvaan aineistoon, menetelmä aiheuttaa huomattavan informaatiokadon ja lisäksi jatkuvan muuttujan etuna oleva aritmeettisten laskutoimitusten käyttö menetetään.

Luokkien laajentaminen ylä- tai alarajoilla

Luokkien laajentaminen ylä- tai alarajoilla ei varsinaisesti ole erillinen menetelmä vaan vain luokkien vähentämisen erikoistapaus. Kuitenkin ylä- ja alarajan määritteleminen vaatii joko jatkuvan tai vähintäänkin järjestysasteikollisen luokitellun aineiston ja eroaa siten yleisestä luokkien vähentämisestä. Näillä perustein menetelmä esitellään erikseen.

Yksinkertaistetusti menetelmässä valitaan sopiva arvo, jota suuremmat arvot kootaan samaan luokkaan aineiston yläreunasta ja vastaavasti tiettyä kynnysarvoa pienemmille arvoille annetaan sama luokka aineiston alareunassa.

7.6 PRAM (The Post-RAndomization Method)

The Post-RAndomization Method (myöhemmin PRAM) on aineiston muokkaamiseen perustuva luokiteltujen mikroaineistojen suojausmenetelmä, jonka suojaus tapahtuu todennäköisyyksien perusteella. PRAM-menetelmällä voidaan suojata ainoastaan luokiteltuja muuttujia, koska sen käyttämisessä tarvittaan siirtymämatriisia, jossa on oltava rivi jokaiselle suojattavan muuttujan mahdollisista arvoista. PRAM-menetelmän soveltamisen seurauksena kunkin yksikön muuttujanarvoista yksi tai useampia on voinut muuttua toiseksi arvoksi. Arvojen muutos perustuu siirtymämatriisin todennäköisyyksiin, joten arvojen on mahdollista pysyä myös ennallaan. Tällaisen Markovin matriisin käyttäminen tekee PRAM-menetelmästä monipuolisen. Se saa menetelmään sisältymään kohinan lisäämisen, arvojen peittämisen sekä uudelleen luokittelun. Monipuolisuutensa takia PRAM-menetelmää käytetään erityisesti identifioivien muuttujien suojaamiseen.

PRAM-menetelmän tavoitteena on muuttaa identifioivien muuttujien arvoja siten, että suojatussa aineistossa muuttujanarvot ovat tietyllä todennäköisyydellä vääriä. Menetelmän etuna on, että yksiköiden tunnistaminen suojauksen jälkeen on hankalaa ja jopa tunnistamisen tapahduttua tunkeutuja ei voi olla täysin varma tunnistamisen oikeellisuudesta. Menetelmä voidaankin nähdä eräänlaisena väärinluokittelun muotona, jossa siirtymätodennäköisyydet (eli todennäköisyydet muuttaa arvo toiseksi) määräävät täysin menetelmän todennäköisyysmekanismin. Siirtymätodennäköisyydet muodostavat Markovin matriisin, jota tämän menetelmän yhteydessä voidaan kutsua PRAM-matriisiksi.

PRAM-menetelmä eroaa tavanomaisesta väärinluokittelusta siten, että mekanismi, jolla luokittelu määräytyy, on tunnettu. Menetelmän mekanismin tuntemisen etuna on, että suojatun aineiston avulla voidaan suorittaa tilastollisia päättelyitä liki tavanomaisesti. Suojatun aineiston avulla voidaan estimoida aineiston käyttäjälle tuntemattoman alkuperäisen aineiston tunnuslukuja. Tämä on tietenkin mahdollista vain siinä tapauksessa, että aineiston käyttäjä tuntee aineiston suojauksessa hyödynnetyn muunnoksen eli siirtymätodennäköisyydet. PRAM-menetelmällä suojatun aineiston luovuttamisen yhteydessä onkin tapana luovuttaa myös suojaamisessa käytetty

PRAM-matriisi. Aineistosta pystytään kuitenkin suorittamaan analyysejä väärinluokitteluun tarkoitetuilla menetelmillä ilman tarkan siirtymämatriisin tuntemistakin.

PRAM-menetelmän informaatiokato ja paljastumisriski riippuvat erityisesti Markovin matriisin valinnasta. PRAM-matriiseja on tutkittu paljon ja kummankin ominaisuuden mittaamiseen on ehditty kehittää useitakin menetelmiä. Tämän tutkielman käsittely perustuu Gouweleeuw et al. (1998) teoreettiseen esitykseen ja de Wolfin ja van Gelderin (2004) tutkimukseen.

PRAM-matriisi

Olkoon X_i alkuperäisen luokitellun aineiston muuttuja, joka halutaan suojata PRAM-menetelmällä ja olkoon Z_i vastaava muuttuja suojatussa aineistossa. Oletetaan lisäksi, että X_i ja siten myös Z_i on luokiteltu K luokkaan eli luokat ovat $1, \ldots, K$. Merkitään PRAM-menetelmän siirtymätodennäköisyyksiä

$$p_{kl} = \mathbb{P}(Z_i = l | X_i = k),$$

mikä tarkoittaa todennäköisyyttä, että alkuperäinen arvo $X_i = k$ muuttuu suojaamisessa arvoksi $Z_i = l$ ja pätee kaikille $k, l = 1, \ldots, K$. Käytettäessä näitä merkintöjä $K \times K$ -matriisin muodostamisessa, saadaan Markovin matriisi $P = \{p_{kl}\}$ eli tässä tapauksessa PRAM-matriisi, jota merkitään lyhyesti P. Markovin matriisillehan pätee, että $P\imath = \imath$, missä $\imath$ tarkoittaa K-vektoria, jonka kaikki alkiot ovat ykkösiä. Käytännössä tämä tarkoittaa sitä, että kunkin matriisin rivin todennäköisyyksien summan on oltava yksi. PRAM-matriisin valinnassa voidaan käyttää myös vahvempaa olettamusta eli vaatia että PRAM-matriisin tulee olla kääntyvä. Vaikka se ei olekaan välttämätöntä, helpottaa käänteismatriisin käyttö X_i:n frekvenssijakauman estimointia alkuperäisestä aineistosta ja vastaavasti myös PRAM-menetelmän käytön aiheuttaman varianssin estimointia.

Käytössä olevin merkinnöin suojaaminen PRAM-menetelmällä tarkoittaa, että tietyn yksikön r valitulle arvolle $X_i = k$ saadaan uusi arvo Z_i todennäköisyysjakaumasta $p_{k1}, \ldots, p_{kK}$. Sama menettely käydään läpi jokaisen alkuperäisen aineiston yksikön kohdalla riippumattomasti muista yksiköistä.

Otetaan esimerkki asian selventämiseksi ja oletetaan, että X_i on sukupuoli ja muuttuja saa arvot $X_i = 1$, jos kyseessä on mies ja $X_i = 2$, jos nainen. Muokataan PRAM-menetelmällä aineistoa, jossa on 100 miestä ja 100 naista, ja valitaan todennäköisyyksiksi $p_{11} = p_{22} = 0,9$. Oletuksena on, että myös suojatussa aineistossa on 100 miestä ja 100 naista, mutta näistä miehistä 10 oli alun perin naisia ja vastaavasti naisista 10 oli alun perin miehiä.

Yleisemmin ajateltuna PRAM-menetelmän vaikutus yksiulotteiseen frekvenssitaulukkoon on

$$\mathbb{E}(\mathbf{T}_{Z_i} | X_i) = P^T \mathbf{T}_{X_i},$$

missä $\mathbf{T}_{X_i} = (T_{X_i}(1), \ldots, T_{X_i}(K))^T$ on merkintä frekvenssivektorille alkuperäisessä aineistossa ja vastaavasti $\mathbf{T}_{Z_i}$ suojatussa aineistossa. Alkuperäisen aineiston frekvenssitaulukon ehdollinen harhaton estimaattori saadaan

$$\hat{\mathbf{T}}_{X_i} = (P^{-1})^T \mathbf{T}_{Z_i}. \tag{7.5}$$

Mikäli käytössä on kaksiulotteinen frekvenssitaulukko, voidaan edellä olevat määrittelyt laajentaa tulkitsemalla kyseinen taulukko vektoreina. Vastaava PRAM-matriisi saadaan Kroneckerin tulona yksiulotteisten taulukoiden matriiseista. Vaihtoehtoisesti voidaan käyttää kaksiulotteista frekvenssitaulukkoa alkuperäisestä aineistosta $T_{X_i\eta}$ ja suojatusta aineistosta $T_{Z_iY_i}$ suoraan matriisimerkinnöin:

$$\hat{T}_{X_i\eta} = (P_{Z_i}^{-1})^T T_{Z_iY_i} P_{Y_i}^{-1}.$$

Paljastumisriskin mittaaminen

Perehdytään de Wolfin ja van Gelderin (2004) käyttämään paljastumisriskin mittaan. Heidän menetelmässään käytetään etukäteen valittua kynnysarvoa, mutta kynnysarvon määrittäminen toteutetaan tavallisuudesta poikkeavasti. PRAM-menetelmän yhteydessä perinteisen kynnysarvon käyttö ei ole mielekästä, koska menetelmä perustuu todennäköisyyksiin ja sen seurauksena paljastumisvaarassa olevat yksiköt vaihtelisivat suojauskerrasta toiseen. Tästä syystä paljastumisriskin mittaamista on tutkittava uudella tavalla.

Paljastumisriski ajatellaan todennäköisyydeksi, että tietty arvo k suojatusta aineistosta onkin pysynyt samana ja siten vastaava arvo alkuperäisestä aineistosta on myös k. Bayesin säännön avulla paljastumisriskin R_{PRAM} voi laskea seuraavasti:

$$R_{PRAM}(k) = \mathbb{P}(X_i = k | Z_i = k) = \frac{\mathbb{P}(Z_i = k | X_i = k)\mathbb{P}(X_i = k)}{\sum_{l=1}^{K} \mathbb{P}(Z_i = k | X_i = l)\mathbb{P}(X_i = l)}.$$

Oletetaan, että PRAM-menetelmää käytetään muuttujien X_i arvokombinaatioiden suojaamiseen, jolloin edellä ollein merkinnöin estimaatiksi saadaan

$$\hat{R}_{PRAM}(k) = \frac{p_{kk} T_{X_i}(k)}{\sum_{l=1}^{K} p_{lk} T_{X_i}(l)}.$$

Todennäköisyydelle $\mathbb{P}(X_i = k)$ käytetään estimaattina arvoa $T_{X_i}(k)/n$, missä n on alkuperäisen mikroaineiston koko.

Tarkastellaan tavallisen kynnysarvon ja PRAM-menetelmän perusteella suojattujen yksiköiden paljastumisriskejä. Tavallisen kynnysarvon tapauksessa suojatuiksi tulkitaan yksiköt, joiden muuttujien arvojen yhdistelmiä on aineistossa vähintään kynnysarvon d osoittama määrä. Tällöinkin suojattu yksikkö voidaan yhdistää ainakin d yksikköön perusjoukossa ja mikäli yhdistämisen oletetaan tapahtuvan satunnaisesti on todennäköisyys pienempi tai yhtä suuri kuin $1/d$. Toisin sanoen kynnysarvoa käytettäessä paljastumisriski on korkeintaan $1/d$. Arvioidaan myös PRAM-menetelmän paljastumisriskiä kynnysarvon kautta. de Wolf ja van Gelder esittävät suojatun yksikön määrittelyksi

$$\hat{R}_{PRAM}(k) \leq \frac{T_{X_i}(k)}{d},$$

missä d on kynnysarvo, joka on määritelty alkuperäiselle mikroaineistolle. Koska $T_{X_i}(k)$ on estimaatti alkuperäisen aineiston muuttujan X_i arvon k frekvenssille, on oikean yhdistämisen kokonaisriski alkuperäisestä aineistosta satunnaisesti valitun ja jonkun $T_{X_i}(k)$ yksikön välillä $1/d$. Perinteisen kynnysarvo-ajattelun mukaan suojattu yksikkö tulee olemaan suojattu myös PRAM-menetelmän paljastumisriskin mukaan.

Lisäksi suojaamattomien yksiköiden määrä PRAM-menetelmässä riippuu ainoastaan alkuperäisistä frekvensseistä ja valitusta PRAM-matriisista eli määrä on riippumaton suojauksen toteutuksesta.

Informaatiokadon mittaaminen

Informaatiokadon mittaamiseen de Wolf ja van Gelder esittävät kolme eri tapaa. Ensimmäinen entropiaan perustuva mitta määritellään

$$\mathrm{EBIL}(P, Z) = -\sum_{r \in Z} \sum_{k=1}^{K} \mathbb{P}(X_i = k | Z_i = l_r) \log \mathbb{P}(X_i = k | Z_i = l_r),$$

missä Z on suojattu mikroaineisto ja l_r on yksikön r arvo muuttujalle Z_i aineistossa Z. Todennäköisyydet $\mathbb{P}(X_i = k | Z_i = l_r)$ ovat järjestykseltään päinvastaisia kuin PRAM-matriisin P määrittelyssä. Jatkossa näitä todennäköisyyksiä merkitään $p_{lk}^{\leftarrow}$. Toinen hyvin samankaltainen mitta määritellään

$$\mathrm{IL}(P, X, Z) = -\sum_{r \in Z} \log \mathbb{P}(X_i = k_r | Z_i = l_r),$$

missä k_r ja l_r ovat yksikön r arvot muuttujalle X_i alkuperäisessä aineistossa X ja vastaavasti muuttujalle Z_i suojatussa aineistossa Z. Suurin ero näillä kahdella mitalla EBIL ja IL on se, että jälkimmäinen käyttää informaatiokadon laskemiseen sekä alkuperäistä että suojattua aineistoa toisin kuin ensimmäinen.

Mikäli halutaan käyttää edellä esitetty lyhempää esitystä, voidaan mitat kirjoittaa muotoon

$$\mathrm{EBIL}(P, Z) = -\sum_{l=1}^{K} \sum_{k=1}^{K} T_{Z_i}(l) p_{lk}^{\leftarrow} \log p_{lk}^{\leftarrow}$$

ja

$$\mathrm{IL}(P, X, Z) = -\sum_{l=1}^{K} \sum_{k=1}^{K} T_{X_i, Z_i}(k, l) \log p_{lk}^{\leftarrow},$$

missä $T_{X_i,Z_i}(k, l)$ tarkoittaa sellaisten yksiköiden määrää, joille $X_i = k$ alkuperäisessä aineistossa X ja $Z_i = l$ suojatussa aineistossa Z. Verrattaessa näitä kahta määrittelyä, voi todeta, etteivät ne poikkea paljoakaan toisistaan, kun lukumäärä$T_{X_i,Z_i}(k, l)$ on lähellä arvoa $T_{Z_i}(l)p_{lk}^{\leftarrow}$. Ehto toteutuu kun yksiköitä on riittävän paljon suhteessa luokkien K lukumäärään.

Jos käytetään samankaltaisia oletuksia kuin paljastumisriskin johtamisessa PRAM-menetelmälle, voidaan todennäköisyydet $p_{lk}^{\leftarrow}$ estimoida seuraavasti:

$$\hat{p}_{lk}^{\leftarrow} = \frac{p_{kl} T_{X_i}(k)}{\sum_{m=1}^{K} p_{ml} T_{X_i}(m)}.$$

Informaatiokatoa voidaan mitata myös tarkastelemalla aineistojen frekvenssien eroja. Jokaisen suojattavan aineiston tarkastelu alkaa yleensä frekvenssitaulukoinnilla vähintään identifioivien muuttujien osalta. PRAM-menetelmän käyttö vaikuttaa taulukoihin ainakin silloin, kun jokin suojattavista muuttujista on mukana taulukossa.

Tästä syystä eräs kehitetty informaationkadon mitta perustuu alkuperäisen ja suojatun frekvenssitaulukon eroihin.

Mitattaessa PRAM-menetelmän vaikutusta frekvenssitaulukoihin, on yksi tapa käyttää alkuperäisen taulukon T_{X_i} ja estimoidun taulukon $\hat{T}_{X_i}$ arvojen suhteellisten etäisyyksien mediaania eli

$$\mathrm{RD}_d = Mediaani\left\{\left|\frac{T_{X_i}(k) - \hat{T}_{X_i}(k)}{T_{X_i}(k)}\right|, k = 1, 2, \ldots, K\right\},$$

missä d viittaa frekvenssitaulukon dimensioon ja nyt käsittely rajoittuu tapauksiin, joille $d = 1, 2$. Tämän lisäksi voidaan laskea suurin suhteellinen ero $m\mathrm{RD}_d$. Ero voi olla ääretön, mikäli $T_{X_i}(k) = 0$ ja $\hat{T}_{X_i}(k) \neq 0$. Kokemusten perusteella tällainen tilanne syntyy vain jos $d = 2$ ja jos kummallakin muuttujalla suuri määrä luokkia. Mikäli tutkitaan tilannetta $d = 2$, on perusteltua tarkastella suurinta arvoa yli kaikkien äärellisten suhteellisen etäisyyksien arvojen ja laskea kuinka usein äärettömiä tapauksia esiintyy.

Toinen informaatiokadon määritelmä, joka käyttää frekvenssitaulukkoja, saadaan tarkastelemalla PRAM-menetelmän käytöstä aiheutunutta varianssin kasvua. Tarkastelussa käytetään yksiulotteisen frekvenssitaulukon estimointia eli yhtälössä (7.5) esiintyvää estimaattoria. Estimaattorissa (7.5) esiintyvän taulukon $\hat{T}_{X_i}$ ehdollinen kovarianssimatriisi on

$$\Sigma^{|}_{\hat{\mathbf{T}}_{X_i}} = Var(\hat{\mathbf{T}}_{X_i}|X_i) = Var((P^{-1})^T\mathbf{T}_{Z_i}|X_i) = (P^{-1})^T Var(\mathbf{T}_{Z_i}|X_i)P^{-1}.$$

Gouweleeuw et al. osoittavat, että

$$Var(\mathbf{T}_{Z_i}|X_i) = \sum_{k=1}^{K} T_{X_i}(k)V_k,$$

missä V_k tarkoittaa matriisia, jonka alkiot $V_k(l, j)$ on määritelty

$$V_k(l,j) = \begin{cases} p_{kl}(1 - p_{kl}) & \text{kun } l = j \\ -p_{kl}p_{kj} & \text{kun } l \neq j \end{cases} \qquad l, j = 1, 2, \ldots, K.$$

Jos halutaan yksi arvo informaatiokadolle, on tässä tapauksessa käytettävä yksiulotteisen frekvenssitaulukon eri luokkien saamien varianssikertoimien mediaania. Mikäli näin tehdään, saadaan

$$\mathrm{CV} = Mediaani\left\{\frac{\sqrt{\Sigma^{|}_{\hat{\mathbf{T}}_{X_i}}(k,k)}}{T_{X_i}(k)}, k = 1, 2, \ldots, K\right\}.$$

Lisäksi voidaan laskea suurin varianssikerroin $m\mathrm{CV}$ jokaisen K luokan suhteen. de Wolfin ja van Gelderin tutkimuksien perusteella $T_{X_i}(k) > 0$ kaikille yksiulotteisten muuttujien X_i luokille k, joten kaikki tutkittavat varianssikertoimet ovat äärellisiä.

Mikäli kumpikaan edellisistä informaatiokadon mitoista ei sovellu aineistolle, voidaan vielä kokeilla lineaariseen regressioon perustuvaa mittaa. Tämä on yleisimpiä PRAM-menetelmällä suojattujen mikroaineistojen analysointiin käytettyjä mittoja. Tarkoituksena on sovittaa regressiosuora yhden luokitellun muuttujan muodostamaan aineistoon ja sitten verrata alkuperäisestä ja suojatusta aineistosta estimoituja regressiokertoimia. Tässä esitettävässä menetelmässä valitaan alkuperäiset muuttujien arvot selitettäviksi ja suojatut arvot selittäviksi muuttujiksi. Malli on siis seuraava:

$$Y = \mathbb{E}\left(\sum_{k=1}^{K} \beta_k \delta(k)\right),$$

missä Y on siis selitettävä muuttuja (alkuperäisestä aineistosta) ja $\delta(k)$ indikaattorimuuttuja, joka vastaa k:tta muuttujan X_i PRAM-menetelmällä suojattua luokkaa. Regressiokertoimet $\beta = (\beta_1, \ldots, \beta_K)^T$ estimoidaan alkuperäisen aineiston avulla ts.

$$\beta = [diag(T_{X_i}(1), \ldots, T_{X_i}(K))]^{-1} \mathbf{T}^y_{X_i},$$

missä $\mathbf{T}^{yT}_{X_i} = (T^y_{X_i}(1), \ldots, T^y_{X_i}(K))$ ja $T^y_{X_i}(k) = \sum_{r \in X} Y_r \delta_{X_i,r}(k)$ on summa yli alkuperäisen aineiston kaikkien yksiköiden, joille $X_i = k$. Kun PRAM-menetelmä on muokannut muuttujan X_i arvot, regressiokertoimet β_k voidaan estimoida käyttämällä

$$\tilde{\beta} = [diag(\hat{T}_{X_i}(1), \ldots, \hat{T}_{X_i}(K))]^{-1} (P^{-1})^T \mathbf{T}^y_{Z_i},$$

missä $\hat{\mathbf{T}}_{X_i}$ on lausekkeesta (7.5) ja $\mathbf{T}^y_{Z_i}$ on määritelty täysin vastaavasti suojaamattomalle aineistolle määritellyn $\mathbf{T}^y_{X_i}$ kanssa. Lopulta haluttu informaatiokadon mitta voidaan kirjoittaa muodossa

$$\text{LRD} = Md\left\{\left|\frac{\beta_k - \tilde{\beta}_k}{\beta_k}\right|, k = 1, \ldots, K\right\}.$$

Myös tässä tapauksessa voidaan halutessa laskea suurin suhteellinen erotus mLRD K:n regressiokertoimen yli.

Erilaisia PRAM-matriiseja

PRAM-matriiseja voi valita satunnaisesti, mutta analyysien helpottamiseksi on järkevämpää valita jollain tapaa symmetrisiä matriiseja. PRAM-matriisin tulee kuitenkin aina toteuttaa jo alussa esitellyt Markovin matriisin ominaisuudet. Erilaiset PRAM-matriisit voidaan jakaa laajempiin ryhmiin ominaisuuksiensa perusteella ja seuraavassa perehdytään de Wolfin ja van Gelderin tutkimuksessaan käyttämiin kolmeen eri tyyppiseen matriisiin.

- Nauhamatriisit $nB(p;b)$, missä n on neliömatriisin koko, p diagonaalilla olevien alkioiden arvo ja b nauhan leveys eli niiden alkioiden p_{kl} lukumäärä, joille $|k-l| < b$. Todennäköisyysmassa $(1-p_{kk})$ on jakautuneena tasaisesti kaikille nauhalla oleville diagonaalin ulkopuolisille alkioille. Esimerkkinä nauhamatriisista olkoon $4B(0,8;2)$ eli

$$\begin{pmatrix} 0,8 & 0,2 & 0 & 0 \\ 0,1 & 0,8 & 0,1 & 0 \\ 0 & 0,1 & 0,8 & 0,1 \\ 0 & 0 & 0,2 & 0,8 \end{pmatrix}$$

- Täysiasteinen matriisi, jossa kaikki diagonaalin ulkopuolella olevat alkiot saavat samat arvot. Merkitään tällaista matriisia $nE(p)$, missä n viittaa neliömatriisin kokoon ja p on diagonaalialkioiden saama arvo. Esimerkkinä tästä matriisityypistä olkoon $3E(0,8)$ eli

$$\begin{pmatrix} 0,8 & 0,1 & 0,1 \\ 0,1 & 0,8 & 0,1 \\ 0,1 & 0,1 & 0,8 \end{pmatrix}.$$

- Täysiasteinen matriisi, jossa diagonaalin ulkopuolella olevien alkioiden arvot riippuvat niitä vastaavista alkuperäisen aineiston frekvensseistä. Merkitään tällaista matriisia $nF(p)$, jossa n on neliömatriisin koko ja p diagonaalialkioiden arvo. Diagonaalin ulkopuolisten alkioiden arvot saadaan laskettua de Wolfin ja van Gelderin tutkimusraportin mukaan seuraavan kaavan avulla

$$p_{kl} = \frac{(1-p_{kk})\left(\sum_{i=1}^{K} T_{X_i}(i) - T_{X_i}(k) - T_{X_i}(l)\right)}{(n-2)\left(\sum_{i=1}^{K} T_{X_i}(i) - T_{X_i}(k)\right)}.$$

  Esimerkkinä tällaisesta matriisista olkoon (de Wolf, van Gelder, 2004 s.8) matriisi $3F(0,6)$, missä $\mathbf{T}_{X_i} = (5576, 24, 632)^T$

$$\begin{pmatrix} 0,6000 & 0,3854 & 0,0146 \\ 0,0407 & 0,6000 & 0,3593 \\ 0,0017 & 0,3983 & 0,6000 \end{pmatrix}.$$

Jos tarkastellaan matriisia $1B(1,1)$, huomataan, että itse asiassa $1B(1,1) = 1E(1) = 1F(1)$ ja tätä matriisia voidaan merkitä myös $\mathbb{1}$.

7.7 MASSC

MASSC-menetelmän nimi on lyhenne sanoista Micro Agglomeration, Substitution for perturbation, Subsampling for suppression, optimal sampling weight Calibration. MASSC-menetelmä siis sisältää useita aiemmissa luvuissa esiteltyjen menetelmien ominaisuuksista ja näiden lisäksi se ottaa huomioon tuotettavan aineiston otospainot (Singh ja Dunteman, 2004).

Menetelmän suojaus tapahtuu neljässä vaiheessa:

- aineiston osittaminen riskiositteisiin,
- muuntaminen optimaalisesti todennäköisyyksiin perustuen,
- peittäminen optimaalisesti todennäköisyyksiin perustuvan otannan avulla ja
- otospainojen optimaalinen kalibrointi muuttujien estimaattien oikeellisuuden säilymiseksi.

Menetelmän viimeisen vaiheen eli kalibroinnin jälkeen aineisto on käytettävissä analyyseihin perusohjelmistojen avulla.

MASSC-menetelmän yhteydessä aineistoa ajatellaan perusjoukkona ja paljastumisriskin muodostaa sisäinen paljastusta yrittävä henkilö eli joku, joka tietää kiinnostuksen kohteensa olevan aineistossa mukana. Aiemmin esiteltyjen menetelmien avulla aineistoa pyrittiin suojaamaan sellaista tunkeutujaa vastaa, jolla ei ole aineistossa olevista henkilöistä tarkempaa lisätietoa. MASSC-menetelmän avulla pystytään tarttumaan tätä vaikeampaan ongelmaan eli suojaukseen tunkeutujalta, jolla on enemmän tietoa aineiston sisällöstä. Samalla menetelmä suojaa myös ilman kyseisiä tietoja paljastamista yrittäviltä. Tämän lisäksi menetelmä tuottaa mallista riippumattomat mitat sekä informaatiokadolle että paljastumisriskille. Näiden yhtäaikainen toteuttaminen onnistuu ottamalla survey-menetelmät mukaan tietosuojaukseen.

Aiempien menetelmien yhteydessä on oletettu, ettei tunkeutujalla ole mitään tietoa aineistossa olevista yksiköistä. Tällainen henkilö etsii aineistosta ainutlaatuisia yksiköitä ja yrittää muita aineistoja apuna käyttäen identifioida näitä ainutlaatuisia yksiköitä. Tällöin eri henkilöillä on käytössä erilaisia tietoja identifioivista muuttujista. Tarkasteluissa on lisäksi otettava huomioon, ettei luovutettavassa aineistossa oleva ainutlaatuinen yksikkö ole välttämättä ainutlaatuinen perusjoukossa, joten aineistoihin jääneistä ainutlaatuisista yksiköistä ei suoraan seuraa identifioituminen. Tavanomaisten suojausmenetelmien tapauksessa aineisto on otos, johon sisältyvistä yksiköistä ei ole erityistä tietoa. MASSC-menetelmän tapauksessa aineisto sen sijaan on perusjoukko ja tunkeutujan oletetaan tietävän keitä aineistoon sisältyy. Näiden tietojen perusteella hänen on helpompi päätellä mikä aineiston yksiköistä vastaa kutakin kiinnostuksen kohdetta.

Tavallisen tietosuojauksen heikkoutena on, että aineiston paljastumisriskin tai informaatiokadon arvioimiseksi on muodostettava malli. Koska MASSC-menetelmä käyttää todennäköisyyksiä apuna suojaamisessa, pystyy se optimoimaan paljastumisriskin ja informaatiokadon arvoja suojauksen edetessä ja tuottamaan optimaalisen lopputuloksen suojaukselle. Paljastumisriskin ja informaatiokadon laskeminen ilman malliin liittyviä oletuksia toteutetaan hyödyntämällä tietoa aineiston suojaamiseen käytetystä satunnaissotkennasta sekä suojatun ja alkuperäisen aineiston välisistä eroista.

MASSC-menetelmän esittelyn yhteydessä otetaan huomioon kaksi erityyppistä paljastumisen muotoa. Ensinnäkin tarkastellaan identifioivien muuttujien arvoiltaan ai-

nutlaatuisen yksikön identifioitumista, jonka seurauksena kaikki sen herkkien muuttujien arvot paljastuvat. Toisena paljastumisen muotona on ryhmä sellaisia yksiköitä, joiden kaikki identifioivien muuttujien arvot ovat samat ja kaikille yhteiset herkkien muuttujien arvot paljastuvat. Perinteisessä tietosuojauksessa jälkimmäinen tilanne yleensä ohitetaan, koska sen riski on varsinkin suurissa aineistoissa vähäinen.

Merkitään yksikkökohtaista paljastumisriskiä δ^p, missä yläindeksi p viittaa tiettyyn profiiliin. Profiililla tarkoitetaan yksikön identifioivien muuttujien arvojen kombinaatiota. Kaikilla niillä yksiköillä, joilla on samat identifioivien muuttujien arvot eli sama profiili, on sama paljastumisriski. Saman profiilin omaavien yksiköiden lukumäärä riippuu identifioivien muuttujien luokkien laajuudesta. Eli mitä laajemmat luokat, sitä vähemmän luokkia ja sitä pienempi määrä erilaisia profiileja ja sen seurauksena yksikkökohtaisia paljastumisriskejä. Kuitenkin haluttaessa on mahdollista määrittää globaalipaljastumisriski δ^c. Aineiston yksiköiden määrä vaikuttaa molempiin paljastumisriskeihin, koska se vaikuttaa suoraan siihen, onko yksikkö ainutlaatuinen vai ei.

Luvussa 4.1 paljastumisriskiä määriteltäessä on esitelty mitta, joka on ehdollinen todennäköisyys $\mathbb{P}(PU|SU)$. Tällä mitataan todennäköisyyttä, että otoksen ainutlaatuinen yksikkö SU (sample unique) on myös perusjoukon ainutlaatuinen yksikkö PU (population unique). Tämä mitta ei kuitenkaan ota huomioon todennäköisyyttä ξ_M^p, että yksikkö, joka on kummassakin tapauksessa ainutlaatuinen, on vielä yhdistetty oikeaan perusjoukon yksikköön. MASSC-menetelmän oletusten takia on otettava huomioon todennäköisyys sille, että kiinnostuksen kohde on mukana aineistossa. Kun yhdistetään nämä kaikki todennäköisyydet ulkoisen ja sisäisen paljastamisen huomioon ottavaksi mitaksi $\delta_U^{p(1)}$ saadaan tulo

$$\delta_U^{p(1)} = \xi_U^p \xi_{s_1}^p \xi_M^p. \tag{7.1}$$

Mitassa ξ_U^p tarkoittaa todennäköisyyttä sille, että profiili "p" on ainutlaatuinen perusjoukossa ja $\xi_{s_1}^p$ todennäköisyys, että kyseinen profiili on mukana aineistossa (s_1) ehdolla, että se on ainutlaatuinen. Esitelty mitta $\delta_U^{p(1)}$ on satunnainen, koska sen laskemisessa käytetään satunnaisia otoksia perusjoukosta. Todennäköisyyden ξ_M^p oikealle yhdistämiselle oletetaan kuitenkin mukavuussyistä olevan yksi. Jos tarkastellaan pelkkää sisäistä paljastamista, jolloin aineiston tulkitaan olevan koko perusjoukko, on todennäköisyys $\delta_U^{p(1)}$ jokaiselle ainutlaatuiselle yksikölle yksi. Sisäisen paljastamisen tapauksessa ja silloin, jos todennäköisyys $\delta_U^{p(1)}$ on melko suuri, on aineisto suojattava ennen luovutusta.

Käytännössä kumpikaan, ehdollinen todennäköisyys $\mathbb{P}(PU|SU)$ tai todennäköisyys $\delta_U^{p(1)}$, ei ole paras mahdollinen paljastumisriskin mitta. Jos keskitytään ulkopuoliseen tunkeutujaan, yksikön identifioiminen suojaamattomasta aineistosta tapahtuu valitsemalla kohde joko ainutlaatuisista profiileista tai profiileista, jotka ovat tunkeutujalle ennestään tuttuja. Tilastollisen tietosuojauksen jälkeen kohteen identifioivien muuttujien arvot ovat voineet muuttua ja siksi identifioimisessa on mukana tiettyä epävarmuutta. Tästä syystä paljastumisriski määritellään edellä annetun $\delta_U^{p(1)}$ sijaan määrittäen riskiä todennäköisyyksien avulla.

Käytetään seuraavia todennäköisyyksiä:

- profiili on perusjoukon ainutlaatuinen,
- profiili on mukana aineistossa,
- profiili on aineiston ainutlaatuinen,
- profiili on pysynyt ennallaan läpi muokkaamisen ja peittämisen ja
- profiili on yhdistetty oikein kohteeseensa.

Ulkoisen paljastamisen riski sellaiselle yksikölle, joka vaikuttaisi olevan aineiston ainutlaatuinen ja jonka profiili on p, voidaan määrittää seuraavasti:

$$\begin{aligned} \delta_u^{p(2)} &= \xi_U^p \xi_{s_1}^p \mathbb{P}(A_{1(s_1)}) \mathbb{P}(A_{2.1(s_1)}) \mathbb{P}(A_{3.12(s_1^*)}) \mathbb{P}(A_{4.123(s_2^*)}) \xi_M^p \\ &= \xi_U^p \xi_{s_1}^p \pi_U^p (1-\psi_U^p) \phi_U^p (1-\chi_U^p) \xi_M^p, \end{aligned} \tag{7.2}$$

missä alaindeksi "u" viittaa suojatun aineiston ainutlaatuiselta vaikuttavaan yksikköön ja jo edellä esitelty "U" alkuperäisen aineiston (s_1) ainutlaatuiseen yksikköön. Mitan $A_{1(s_1)}$ tarkoittaa tapausta (todennäköisyyttä merkitään π_U^p), että aineiston s_1 yksikkö on ainutlaatuinen. Tämä tapahtuma on ehdollinen ehdolla, että yksikkö on ainutlaatuinen myös perusjoukossa ja saa siten todennäköisyyden yksi. Tässä vaiheessa koko merkintä saattaa vaikuttaa turhalta, mutta ei vielä jätetä sitä pois, koska sitä tarvitaan jatkossa sisäistä paljastamista tutkittaessa.

Seuraava tapaus mitassa $\delta_u^{p(2)}$ on $A_{2.1(s_1)}$, jonka todennäköisyyttä merkitään $(1-\psi_U^p)$. Se tarkoittaa tapahtumaa, jossa yksikkö ei ole muuttunut ehdolla, että se on ainutlaatuinen muokattujen yksiköiden muodostamassa aineistossa s_1^*. $A_{3.12(s_1^*)}$ on tapaus (todennäköisyyttä merkitään ϕ_U^p), missä yksikkö ei tule peitetyksi eli poistetuksi aineistosta s_1^*. Tapaus voidaan tulkita myös niin, että yksikkö on edelleen mukana peittämisen jälkeisessä aineistossa s_2^* ja yksikkö on tämän lisäksi ainutlaatuinen ja muokkaamaton. Merkinnällä $A_{4.123(s_2^*)}$ tarkoitetaan tapausta, jota vastaava todennäköisyys on $(1-\chi_U^p)$, kun yksikkö aineistosta s_2^* tulee luokitelluksi ainutlaatuiseksi ja se on suojatun aineiston ainutlaatuinen ja selvinnyt muuttumattomana aineiston muunnoksesta ja peittämisestä. Käytännössä viimeinen merkintä tarkoittaa tilannetta, ettei mikään muu yksikkö ole muokkautunut profiililtaan samanlaiseksi aineiston suojauksen edetessä.

Edellä olleita todennäköisyyksiä tarkastellessa on hyvä huomata, että todennäköisyydet on laskettu aineiston suojaajan eikä tunkeutujan näkökulmasta. Tämän takia niissä ei ole mietitty näyttääkö tunkeutujan kiinnostuksen kohde ainutlaatuiselta vai ei. Näillä merkinnöillä paljastumisriskin todennäköisyyttä on kuitenkin helpompi hallita ja toisaalta aineiston suojaajahan ei voi edes tietää, kuka paljastamisen kohteena on. Siten hän ei voi tietää vaikuttaako kohde suojatun aineiston ainutlaatuiselta tapaukselta vai ei. Tunkeutujan näkökulman lisäämisestä seuraisi, että tarkastelun tulisi sisältää myös tilanne, jossa kohde on suojatun aineiston ainutlaatuinen vaikkei se olisi ollut alkuperäisen aineiston ainutlaatuinen.

Esitellyn paljastumisriskin laskemiseen tarvitaan useita todennäköisyyksiä, jotka on joko laskettava tai estimoitava ennen mitan arvon saamista. Vastaavasti kuten aiempi mitta $\delta_U^{p(1)}$, on myös mitta $\delta_U^{p(2)}$ satunnainen. Mitan arvot riippuvat perusjoukosta ja otoksista eli aineistoista s_1, s_1^* ja s_2^*, jotka voidaan valita satunnaisesti. Otannasta tutun asetelmaperusteisen tulkinnan mukaan perusjoukko on kiinteä, joten todennäköisyydetkin ovat kiinteitä. Tästä seuraa, ettei todennäköisyyksien estimointi ole enää mahdollista ilman oletuksia mallista. Kuitenkin sisäisen paljastamisen tapauksessa

kaikki ξ-todennäköisyydet ovat tunnettuja ja arvoltaan ykkösiä. Aineiston suojaaja pystyy laskemaan loput todennäköisyydet, mikäli hän tuntee suojaukseen käytetyn järjestelmän. Todennäköisyydet ϕ ja χ ovat satunnaisia, koska molemmat riippuvat satunnaisesti muokatuista aineistoista.

Tarkastellaan sitten sisäistä paljastamista. Mallista riippumaton yksikkökohtainen paljastumisriski sellaiselle yksikölle, joka näyttää suojatun aineiston ainutlaatuiselta, mutta ei ehkä alunperin ollutkaan ainutlaatuinen, saadaan seuraavasta:

$$\delta_u^p = \pi_U^p(1-\psi_U^p)\phi_U^p(1-\chi_U^p) + \pi_{NU}^p(1-\psi_{NU}^p)\phi_{NU}^p\chi_{NU}^p. \tag{7.3}$$

Kaavassa olevien todennäköisyyksien π_U^p ja π_{NU}^p summa on yksi ja ne vastaavat osuuksia niistä alkioista, joiden profiili on "p" ja jotka ovat, ja vastaavasti eivät ole, alkuperäisessä aineistossa ainutlaatuisia. Todennäköisyyksien summa saadaan siitä, että tunkeutujan kohde, jonka profiili on "p", on muokkautunut satunnaisesti ja on alunperin kuulunut joko ainutlaatuisten tai yleisten yksiköiden luokaan. Lisäksi on oletettu, ettei sisäinen tunkeutuja tiedä kaikkia kohteensa identifioivia muuttujia aivan tarkalleen, joten se joukko yksiköitä, johon kohde kuuluu, voi muodostua useastakin identifioivien muuttujien luokista. Oletuksesta seuraa, että joukkoon voi kuulua sekä ainutlaatuisia että yleisiä yksiköitä. Mikäli profiilin "p" omaavista yksiköistä suurin osa on ainutlaatuisia, ovat nämä kaksi komponenttia painottuneet sopivasti. Todennäköisyyksien laskemiseen tarvittavat alkuperäisen aineiston osuudet ovat aineiston suojaajan laskettavissa. Muut kaavassa (7.3) olevat todennäköisyydet, joiden alaindeksi on "NU", määritellään samoin kuin niitä vastaavat todennäköisyydet alaindeksillä "U". Nämäkin todennäköisyydet ovat tunnettuja, koska suojaamismekanismin satunnaisuus muuntamisessa ja peittämisessä on tunnettu. Paljastumisriskin mitta δ_u^p samallekin aineistolle vaihtelee, koska eri suojauskertojen aineistot s_1^* ja s_2^* eroavat toisistaan. Mitan avulla voidaan laskea estimaatti äärellisen perusjoukon eli aineiston parametrille $E_{\psi\phi}(\delta_u^p)$, joka on mitan δ_u^p keskiarvo yli kaikkien mahdollisten aineistojen s_1^* ja s_2^*, kun todennäköisyydet (ψ, ϕ) ovat vakioita. Toisaalta tämä estimaatti $E_{\psi\phi}(\delta_u^p)$ kelpaa myös yksittäiseksi kiinteäksi riskin mitaksi, joka voidaan approksimoida melko tarkasti simuloimalla useita eri aineistoja s_1^* ja s_2^*.

Väärinluokittelun todennäköisyys χ_{NU}^p on ehdollinen todennäköisyys sille, että tietty yleinen yksikkö selviää muuttumattomana muokkauksesta ja tulee mukaan suojattuun aineistoon ja väärinluokitelluksi ainutlaatuiseksi yksiköksi. Näin voi tapahtua, jos muut alunperin saman profiilin omaavat yksiköt ovat muuttuneet muokkauksessa profiileiltaan erilaisiksi. Väärinluokittelun todennäköisyyksiä ei tunneta, koska aineistoa suojatessa nähdään ainoastaan sillä kertaa toteutunut tilanne väärinluokittelusta. Arvot voidaan approksimoida melko tarkasti simuloimalla aineistoja s_1^* ja s_2^*. Käytännössä todennäköisyydet kannattaa estimoida yhdestä suojauksen läpi käyneestä aineistosta tutkimalla suojauksessa muuttumattomien ja sen jälkeen väärinluokitelluiksi tulleiden yksiköiden osuutta aineistossa. Voidaan olettaa, että jokaisella suojauksesta muuttumatta selvinneellä yksiköllä on yhtä suuri todennäköisyys tulla väärinluokitelluksi. Lisäksi oletetaan, että aineistossa on riittävästi yksiköitä, jotka tulevat väärinluokitelluiksi. Jälkimmäinen oletuksista voi aiheuttaa ongelmia tarkasteluun, joten joitakin profiileja joutuu yhdistämään järkevämmän estimaatin saamiseksi.

Vaikka profiilikohtaiset paljastumisriskin arvot ovatkin hyödyllisiä, on syytä käyttää jotain yksinkertaisempaa paljastumisriskin mittaa ainakin silloin, kun aineisto on laaja. Laajassa aineistossa on yleensä useita eri profiileja ja siten yksittäisten profiilien riskien perusteella on vaikea hallita koko aineistoa. Yhdistämällä yksittäisten profiilien paljastumisriskin mitat voidaan muodostaa globaalimitta δ^c. Globaalin mitan yhteydessä on hyvä pohtia jälleen myös ulkoista paljastumisriskiä. Koska sisäisellä tunkeutujalla on ennakkotietoa aineistosta, voidaan olettaa ulkoisen paljastamisen riskin olevan korkeintaan yhtä suuri kuin edellä esitelty sisäisen paljastamisen riski.

MASSC-menetelmän perusta

Satunnaistamismenetelmän optimaalinen määrittäminen MASSC-menetelmän käytössä on samankaltainen ongelma kuin optimaalisen otosasetelman määrittämien äärelliselle perusjoukolle. Yhteyteen perustuen sopivat ratkaisut voidaan löytää käyttämällä survey-menetelmiä. MASSC-menetelmän ja otannan yhteyttä voidaan perustella seuraavasti:

1. Aineiston luovuttaminen ilman suojausta on verrattavissa kokonaisaineiston käyttöön otoksen sijaan. Paljastuminen aiheuttaa suuret ”kulut” vastaavasti kuin aineiston kerääminen, mutta informaatiokatoa ei tapahdu.
2. Aineiston suojaamista arvoja korvaamalla voidaan verrata otoksen imputointiin. Optimaalista asetelmaa varten arvon korvaamatta jäämisestä seuraavat ”paljastumiskulut” voidaan minimoida käyttäen harhaa rajoitteena.
3. Otoksen valinnassa käytettävä satunnaistamisasetelma pienentää informaatiokatoa ja kuluja. Aineiston suojaamisen kuluilla tarkoitetaan yksikön aineistosta poistamatta jäämisestä seuraavia ”paljastumiskuluja”. Asetelma saadaan optimaaliseksi minimoimalla kuluja ottaen huomioon tarkkuusrajoitteet.
4. Suojaamisessa voidaan lopuksi kalibroida otospainot vastaavasti kuin otannassa. Molemmissa tapauksissa tavoitteena on paremmat estimaatit. Kalibroinnin jälkeen aineistot ovat analysoitavissa perinteisillä otospainot huomioon ottavilla ohjelmistolla.

MASSC-menetelmän etuna on, että paljastumisriskiä ja informaatiokatoa kontrolloidaan koko suojauksen ajan. Kyseessä ovat suojauksen tärkeimmät ominaisuudet, joiden tavoitteet ovat vastakkaiset. Sopivasti valituille arvoille ε ja δ, tavoitteet voidaan esittää muodossa

$$\max_y\{RRMSE(\hat{\theta}_y)\} \leq \varepsilon,$$
$$\max\{E_{\psi\phi}(\delta_u^p), E_{\psi\phi}(\delta_{nu(d)}^p), E_{\psi\phi}(\delta_{nu(t)}^p), E_{\psi\phi}(\delta_{nu(o)}^p)\} \leq \delta,$$

missä $RRMSE$ tarkoittaa suhteellisen keskineliövirheen neliöjuurta alkuperäisestä aineistosta lasketuista muuttujien summista (θ_y) ja maksimi valitaan yli tutkittavien muuttujien (y). Profiilin paljastumisriskin mitoissa δ^p olevat alaindeksit tarkoittavat tilanteita, jolloin profiili on suojatun aineiston ainutlaatuinen tai profiili on yleinen ja esiintyy kahdesti (double), kolmesti (triple) tai useammin (other).

MASSC-menetelmän neljä vaihetta

Vaihe 1 – Mikroryvästys. Ensimmäinen vaihe sisältää riskiluokkien määrittämisen riskin kannalta olennaisten ja vähemmän olennaisten identifioivien muuttujien perusteella. Olennaisia identifioivista muuttujista ovat ne, jotka ovat helposti paljastamista yrittävän henkilön käytettävissä. Vähemmän olennaisiin taas kuuluvat ne identifioivat muuttujat, joita tunkeutuja ei saa kovin helposti käyttöönsä. Vähemmän olennaiset muuttujat voidaan järjestää sen mukaan, miten vaikeaa niiden perusteella on saada tietoa kohteesta. Riskiluokat määritellään seuraavasti: riskiluokka 0 on niiden mikroaineiston yksiköiden ryhmä, jotka ovat ainutlaatuisia kaikki olennaiset identifioivat muuttujat huomioon otettaessa. Riskiluokka 1 muodostuu yksiköistä, jotka ovat ainutlaatuisia, jos järjestyksessä ensimmäinen vähemmän olennaisten identifioivien muuttujien ryhmä siirretään olennaisten ryhmään. Jatkamalla vastaavasti järjestystä eteenpäin, saadaan muodostettua muut riskiluokat. Viimeinen riskiluokista sisältää siten vain sellaiset yksiköt, joiden profiilit ovat yleisiä kaikkien muuttujien suhteen. Tämä luokka voidaan jakaa vielä luokkiin, joissa on kahdesti, kolmesti tai useammin esiintyvät profiilit. Paljastumisriskin kannalta ajateltuna yksikkökohtainen riski on sitä pienempi, mitä suurempi on riskiluokan numero.

Vaihe 2 – Korvaaminen. Tässä vaiheessa valitaan kullekin yksikölle korvaamista varten pari. Luovuttajaksi valitaan yksikkö, jonka muuttujanarvojen etäisyys vastaanottajaan on pienin. Etäisyyden mittaamiseen käytetään identifioivia muuttujia ja olennaisimpia herkkiä muuttujia. Luovuttajaa ei valita suoraan etäisyyden perusteella vaan valinnassa otetaan huomioon myös globaalit ja paikalliset rajoitteet. Rajoitteiden tavoitteena on pitää tutkimuksessa käytettävien muuttujien yleensä moniulotteiset yhteydet pysyvät ennallaan. Tutkittavat muuttujat ovat yleensä identifioivien ja herkkien muuttujien funktioita, esimerkiksi rajoitteen tavoitteena voisi olla säilyttää mahdollinen yhteys tupakoinnin ja alkoholinkäytön välillä tietyn ikäisten nuorten keskuudessa.

Vaihe 2 on yksi tärkeimmistä MASSC-menetelmän vaiheista, koska sen avulla rajoitetaan korvaamisesta seuraavaa harhaa. Lisäksi harhaa rajoitetaan valitsemalla optimaaliset todennäköisyydet korvaamiselle. Tätä varten kukin riskiluokista jaetaan edelleen korvausosajoukkoihin siten, että kunkin osajoukon vaikutus kokonaisharhaan olennaisimpien tutkimusmuuttujien osalta on mahdollisimman pieni. Optimaalisen valinnan saavuttamiseksi tutkitaan osajoukkokohtaisia todennäköisyyksiä, ψ_{hk}, missä h viittaa riskiluokkaan ja k sen osajoukkoon. Optimaaliset todennäköisyydet saadaan minimoimalla korvaamatta jäämisestä seuraavat paljastumiskulut tutkimusmuuttujien keskineliöpoikkeaman ollessa rajoitettu eli

$$\min_{\psi}\{\sum\nolimits_{hk} c(\psi_{hk})N_{hk}(1-\psi_{hk})\} \text{ ehdolla, että } E_{\psi}(\theta_y^* - \theta_y)^2 \leq \alpha\theta_y^2. \tag{7.4}$$

Minimoitavassa $c(\psi_{hk})$ on korvaamatta jääneiden paljastumiskulufunktio, joka on valittu muuttujan ψ_{hk} suhteen vähenevästi funktioksi, N_{hk} on osajoukon koko ja $N_{hk}(1-\psi_{hk})$ on siten odotettavissa oleva korvaamatta jääneiden yksiköiden lukumäärä. Edelleen θ_y on alkuperäisestä aineistosta laskettu tutkittavan muuttujan y alkuperäinen summa perusjoukossa ja θ_y^* on estimoitu summa korvaamisen jälkeen. Arvolla α tarkoitetaan ylärajaa harhan neliölle suhteessa populaatiototaalin neliöön

ja E_ψ tarkoittaa odotusarvoa yli satunnaisen korvausmekanismin. Seuraavan vaiheen eli aliotannan jälkeen ehdollinen keskiarvo $E_{\phi|\psi}(\hat{\theta}_y^*)$ on täsmälleen θ_y^* ja siksi ehdon (7.4) keskineliövirhe voidaan esittää korvaamisen ja aliotannan jälkeen laskettavan estimaattorin $\hat{\theta}_y^*$ ehdollisen harhan keskineliönä eli

$$\begin{aligned} E_\psi(\theta_y^* - \theta_y)^2 &= E_\psi[B_{\phi|\psi}^2(\hat{\theta}_y^*)] \\ &= E_\psi[\theta_y^* - E_\psi(\theta_y^*)]^2 + E_\psi[E_\psi(\theta_y^*) - \theta_y]^2 \\ &= V_\psi(\theta_y^*) + B_\psi^2(\theta_y^*). \end{aligned}$$

Valintatodennäköisyydet ψ voidaan määrittää pysymään välillä 0–0,5. Toisin sanoen kullakin yksiköllä on positiivinen todennäköisyys tulla korvatuksi, mutta missään osajoukossa korvattujen yksiköiden osuus ei ole yli 50 %. Parametrin ψ arvojen kasvaessa nollasta kohti ylärajaa, kasvaa $MSE(\theta_y^*)$ tai vastaavasti ehdollisen harhan keskineliö parametrille θ_y^*, mutta samalla paljastumiskulut vähenevät. Nämä kaksi funktiota siis kulkevat vastakkaisiin suuntiin, mikä on toivottavaa optimointitilanteessa.

Vaihe 3 – aliotanta. Aineisto on jaettu riskiluokkiin ensimmäisessä vaiheessa, mutta nyt riskiluokat jaetaan edelleen osajoukkoihin ryhmittelyalgoritmilla, jolla pyritään mahdollisimman saamaan kunkin osajoukkoon mahdollisimman samankaltaisia yksiköitä. Arvojen vaihtelua mitataan olennaisten tutkimusmuuttujien (y) osalta. Tavoitteena on löytää optimaalinen valintatodennäköisyys ϕ_{hk} kunkin riskiluokan h osajoukolle j. Optimointi toteutetaan minimoimalla otantaan mukaan tulemisesta seuraavia paljastumiskuluja ehdolla, että olennaisten tutkimusmuuttujien varianssirajoitteet ovat voimassa eli

$$\min_\phi\{\Sigma_{hj} c(\phi_{hj}) N_{hk} \phi_{hj}\}, \text{ kun } E_\psi(V_{\phi|\psi}(\hat{\theta}_y^*))^2 \leq \beta\theta_y^2,$$

missä $c(\phi_{hj})$ on paljastumiskulujen funktio, joka on määritetty muuttujan ϕ_{hk} suhteen kasvavaksi funktioksi. $N_{hk}\phi_{hk}$ on mukaan tulevien yksiköiden odotettavissa oleva lukumäärä, $\hat{\theta}_y^*$ on totaalin estimaatti, joka on laskettu korvauksen ja aliotannan jälkeisestä aineistosta. $V_{\phi|\psi}$ on merkintä varianssille, joka lasketaan satunnaisen aliotantamekanismin vaikutuksista korvausvaiheen läpikäyneeseen aineistoon ja β on yläraja varianssille suhteessa populaatiototaalin neliöön.

Kuten korvausmekanismin yhteydessä voidaan myös ϕ-todennäköisyydet määrittää riippumaan profiileista riskiluokkien kautta. Tätä varten profiilien määrittämät luokat jaetaan pienemmiksi osoitteiksi. Jos valintatodennäköisyyksien ϕ_{hj} annetaan vaihdella 0,5 ja 1 välillä, on jokaisella yksiköllä on positiivinen todennäköisyys tulla poistetuksi aineistosta, mutta poistettujen yksiköiden osuus ei kasva yli 50 % missään ositteessa. Tämä rajoite auttaa kontrolloimaan epätasaisen painotuksen vaikutuksia lopullisiin estimaatteihin. Voidaan osoittaa, että varianssin ja paljastumiskulujen funktiot kulkevat vastakkaisiin suuntiin, kun ϕ_{hj} kasvaa, joten optimointi on mahdollista. Optimoinnin jälkeen otoksen valinta voidaan ajatella kaksiasteiseksi otannaksi. Asetelmassa alkuperäinen aineisto tulkitaan ensimmäisen asteen otokseksi, josta otanta suoritetaan sopivia valintatodennäköisyyksiä käyttäen. Kaksiasteisuuden tavoitteena on saada MASSC-menetelmällä suojattu aineisto sellaiseen muotoon, että

perusohjelmistoja voidaan käyttää aineiston analysoimiseen. Mikäli alkuperäinen suojattava aineisto ei ole otos, voidaan käyttää ositettua yksinkertaista satunnaisotantaa tarvittavan otoksen valitsemiseksi.

Vaihe 4 – Kalibrointi. Viimeisessä vaiheessa kalibroidaan aineiston otospainot, jotka saadaan edellisen vaiheen valintatodennäköisyyksien käänteisarvoina. Kalibroinnin tavoitteena on saada tutkimusmuuttujien (y) ja identifioivien muuttujien estimaatit vastaamaan täsmälleen alkuperäisen aineiston perusteella laskettuja arvoja. Tämä on yksi jälkiosittamisen keino ja auttaa vähentämään korvaamisesta johtuvaa harhaa ja otannasta johtuvaa varianssia.

Informaatiokadon mitoista

Aineiston informaatiokadon mittaamisessa toimivat parhaiten perusmenetelmät, kuten harha, varianssi ja keskineliövirhe (MSE). Kuitenkin informaatiokadon laskemiseksi on tunnettava satunnaistamismekamismi, joka kattaa muutokset alkuperäisestä lopulliseen muokattuun aineistoon. MASSC-menetelmässä satunnaismekanismi tunnetaan ja se on määritelty muokattavien ja peitettävien yksiköiden satunnaisella valinnalla.

Informaatiokadon mitalle, joka on suurin tutkimusmuuttujien keskineliövirheen neliöjuurista eli $\max RRMSE$ ja jonka ylärajana on ε, pätee ennen kalibrointia

$$max_y RRMSE(\hat{\theta}_y^*) \leq \varepsilon, \text{ koska}$$

$$MSE(\hat{\theta}_y^*) = E_{\psi\phi}\left(\hat{\theta}_y^* - \theta_y\right)^2 = E_\psi V_{\phi|\psi}(\hat{\theta}_y^*) + E_\psi B_{\phi|\psi}^2(\hat{\theta}_y^*) \leq (\beta + \alpha)\theta_y^2 = \varepsilon^2\theta_y^2. \quad (7.5)$$

Aineiston suojaajalla on riittävästi tietoa ehdollisen harhan neliön eli kaavan (7.5) toisen termin laskemiseksi. Ensimmäisen termin laskemiseen tarvitaan useita versioita korvauksen jälkeen saatavasta aineistosta ja ne saadaan muuttamalla muuttujan ψ arvoa. Tällaisten versioiden tuottaminen on periaatteessa mahdollista, mutta ei kovin käytännöllistä. Vaihtoehtoisesti myös termiä $V_{\phi|\psi}(\hat{\theta}_y^*)$ voidaan käyttää harhattomana estimaattorina. Aineiston käyttäjä, jolla on käytössään MASSC-menetelmällä suojattu aineisto, ei pysty laskemaan harhatermiä vaan hänen on luotettava aineiston luovuttajaan ja uskottava ettei harha ole liian suuri. Vaihtoehtoisesti aineiston suojaaja voi muodostaa yleistetyn harhafunktion mallin, joka on vastaava kuin yleistetyn varianssifunktion malli otosaineistolle. Mallin avulla aineiston käyttäjä voisi arvioida harhan suuruusluokkaa omissa estimaateissaan. Toisaalta aineiston käyttäjän on mahdollista arvioida ehdollisen varianssin arvoa käyttämällä pelkästään suojattua aineistoa. Jos alkuperäinen aineisto on perusjoukko, ei arviointi ole edes vaikeaa. Jos suojattu aineisto on kaksiasteisen otoksen tulos, otoksen varianssin arvioimiseen tarvitaan erikoisohjelmisto. Tällaisessa tapauksessa varianssin estimointia voi yksinkertaistaa olettamalla, että toisen asteen otanta on vain valinta ensimmäisen asteen otosalkioista. Lisäksi oletetaan, että ensimmäinen otos tai alkuperäinen aineisto perustuu monitasoiseen asetelmaan ja että ensisijaiset otosalkiot on valittu palauttaen.

Paljastumisriskin mitoista

MASSC-menetelmää tarkastellessa on muistettava, että aineiston kaikilla yksiköillä, riippumatta siitä ovatko ne vaarassa tulla paljastetuksi vai eivät, on positiivinen todennäköisyys tulla muokatuksi korvaamalla ja/tai peittämällä. Menetelmää kehitettäessä on todettu, että suojaamalla satunnainen määrä yksiköitä, jotka eivät ole vaarassa paljastua, saavutetaan parempi suojauksen taso. Ylimääräisten yksiköiden muokkaaminen aiheuttaa epävarmuutta siihen, millä aineiston yksiköillä paljastumisriski on todellisuudessa suurin. Epävarmuuden lisäämisen seurauksena ei tarvitse muokata niin suurta määrää yksiköitä ja suojauksen vaikutus informaation määrään ei ole suojauksen hyötyyn nähden liian suuri. Mitä suuremman osan aineistosta suojaa, sitä pienemmäksi ainutlaatuisten yksiköiden osuus yleensä pienenee ja vastaavasti yleisten yksiköiden osuus kasvaa. Siten yleiset yksiköt ovat yhä epätodennäköisemmin vaarassa paljastua ja onkin parempi suojata koko alkuperäinen aineisto aina kuin se on mahdollista.

Paljastumisriski riippuu paljastamista yrittävän henkilön tiedoista, joten oletetaan seuraavaksi paljastumisriskien mittoja varten, että kyseinen henkilö tuntee kaikki kohteensa identifioivien muuttujien arvot. Yksinkertaisuuden vuoksi tarkastellaan globaaleja mittoja, vaikka profiilikohtaisten mittojen laskeminen on tapahtuu vastaavasti. Käytetään ainutlaatuisten yksiköiden luokasta indeksiä h ja merkitään yleisten yksiköiden luokkaa h'. Toisin sanoen alaindeksi hjk tarkoittaa k:nnetta otanta-ositetta, joka on osa j:nnettä korvausosajoukkoa ja molemmat osa ainutlaatuisten yksiköiden luokkaa h. Merkitään tästä lähtien osajoukkoa hjk merkinnällä ν ja vastaavasti olkoot ν' osajoukon $h'jk$ merkintä. Pieni kirjain "u" tarkoittaa pseudo-ainutlaatuista. Pseudo-ainutlaatuinen on suojatun aineiston ainutlaatuinen yksikkö, joka on ollut alunperin ainutlaatuinen tai muokkautunut ainutlaatuiseksi. On myös mahdollista, että yleinen yksikkö on päätynyt pseudo-ainutlaatuiseksi muiden saman profiilin yksiköiden muokkauduttua erilaisiksi tai niiden jäätyä pois aineistosta suojauksen yhteydessä. Vastaavasti merkintä "nu" viittaa pseudo-yleisiin yksiköihin.

Näillä merkinnöillä saadaan yksikölle, joka vaikuttaa ainutlaatuiselta suojatussa aineistossa, paljastumistodennäköisyys sisäisen paljastamisen tapauksessa seuraavasta:

$$\delta_u^c = \sum\nolimits_{\nu} \pi_\nu (1 - \psi_\nu^c) \phi_\nu^c (1 - \chi_\nu^c) + \sum\nolimits_{\nu'} \pi_{\nu'} (1 - \psi_{\nu'}^c) \phi_{\nu'}^c \chi_{\nu'}^c,$$

missä aineiston suojaaja olettaa globaalia mittaa varten, että tunkeutujan kohde voi olla peräisin mistä tahansa riskiositteesta. Kaavan todennäköisyydet profiilikohtaiselle mitalle on muutettava ja näitä merkitään yläindeksillä "p". Paljastumistodennäköisyys lasketaan vastaavasti myös yksiköille, jotka ovat suojatussa aineistossa yleisiä. Esimerkiksi parin (d=double) omaavan yksikön paljastumistodennäköisyys on

$$\delta_{nu(d)}^c = \sum\nolimits_{\nu} \pi_\nu^c (1-\psi_\nu^c) \phi_\nu^c \chi_\nu^c \eta_{\nu(d)}^c (1-\zeta_{\nu(d)}^c) + \sum\nolimits_{\nu'} \pi_{\nu'}^c (1-\psi_{\nu'}^c) \phi_{\nu'}^c (1-\chi_{\nu'}^c) \eta_{\nu'(d)}^c (1-\zeta_{\nu'(d)}^c),$$

missä $\eta_{\nu(d)}^c$ tarkoittaa todennäköisyyttä, että osajoukon hjk ainutlaatuinen yksikkö tulee suojauksesta muuttumattomana selvittyään väärinluokitelluksi yleiseksi parin omaavaksi yksiköksi ja $\eta_{\nu'(d)}^c$ on todennäköisyys sille, että muu kuin parin omaava yleinen yksikkö tulee suojauksesta muuttumatta selvittyään väärinluokitelluksi yleiseksi

parin omaavaksi yksiköksi. Merkintä $\zeta^c_{\nu(d)}$ tarkoittaa todennäköisyyttä, että ainutlaatuinen yksikkö ei muutu suojauksessa ja tulee väärinluokitelluksi sellaiseksi yleiseksi parin omaavaksi yksiköksi, jolla kaikki herkkien muuttujien arvot eivät ole samat kuin parillaan. Todennäköisyys $\zeta^c_{\nu'(d)}$ määritellään vastaavasti myös alunperin yleisille yksiköille. Käytännössä todennäköisyyden $\zeta^c_{\nu(d)}$ oletetaan olevan sama alunperin ainutlaatuisille sekä yleisille yksiköille, koska näin saadaan stabiili estimaatti kyseiselle todennäköisyydelle. Paljastumisriskit, $\delta^c_{\nu(t)}$ ("t=triple") ja $\delta^c_{\nu(o)}$ ("o"=other eli neljä tai enemmän samoja profiileja), voidaan mitata täysin vastaavasti kuin edellä esitelty paljastumisriski.

Yksinkertainen esimerkki MASSC-menetelmästä

Seuraava esimerkki on lainattu suoraan Singhin ja Duntemanin kirjoittamasta artikkelista. Esimerkki on yksinkertaistettu, mutta auttaa ymmärtämään, miten MASSC-menetelmä muokkaa aineistoa ja vaikuttaa yksiköiden riskialttiuteen. Olkoon suojattava aineisto taulukon 10 mukainen kymmenen havainnon aineisto. Olennaiset identifioivat muuttujat ovat ikä (neljä ikäluokkaa: 1=12, 2=17, 3=21, 4=25) ja sukupuoli (1=mies, 0=nainen), eikä muita identifioivia muuttujia ole. Herkkä muuttuja on alkoholinkäyttö (1=käyttää, 0=ei käytä). Paljastumisriskin tarkastelussa sisäiseksi paljastajaksi voidaan ajatella aineistossa olevan nuoren äiti. Jos aineisto on hänen käytettävissään, voi hän tunnistaa lapsensa helposti ja saada selville tämän alkoholinkäytön.

Alkuperäistä aineistoa tarkastellessa huomataan, että seitsemän yksikköä kymmenestä on vaarassa paljastua. Näistä kolme on ainutlaatuisia yksiköitä ja neljä parin omaavia yleisiä yksiköitä, joille herkän muuttujan arvo on sama kuin parin. Alkuperäisen aineiston yksiköistä 70 % on siis vaarassa paljastua ennen suojausta. Kunkin yksikön paljastumisriski on 100 %, koska suojaamattoman aineiston paljastamiseen ei liity mitään epävarmuutta. Yksikkö joko paljastuu tai ei.

Muodostetaan esimerkkiä varten kaksi riskiluokkaa: ainutlaatuisille ja muille yksiköille. Tästä saadaan arvot: $\pi_U = \frac{3}{10}$ ja $\pi_{NU} = \frac{7}{10}$. Etsitään seuraavaksi korvaamisvaiheen parit kaikille yksiköille. Aineiston ollessa näin pieni jätetään korvausvaiheen osajoukkojen muodostaminen väliin ja asetetaan korvaamistodennäköisyydet suoraan arvoiksi $\psi_U = \frac{1}{3}$ ainutlaatuisten ja $\psi_{NU} = \frac{2}{7}$ muiden yksiköiden riskiluokalle. Eli kokonaisosuus korvaamiselle on $\frac{3}{10}$= 30 %. Ainutlaatuisten riskiluokan havainnon 2 sukupuoli ja yleisten riskiluokan havainnon 7 ikä ja sukupuoli ja havainnon 10 ikä tulevat korvatuiksi. Osajoukkoja ei muodosteta otantavaihettakaan varten vaan asetetaan valintatodennäköisyydet $\phi_U = \frac{2}{3}$ ainutlaatuisten ja $\phi_{NU} = \frac{6}{7}$ muiden yksiköiden riskiluokille. Otantaosuus on siis yhteensä $\frac{8}{10}$= 80%. Havainto 1 ainutlaatuisten riskiluokasta ja havainto 5 yleisten riskiluokasta valikoituu aineistosta poistettavaksi. Riskiluokkien otospainot ovat siis 3/2 ja 7/6 ja nämä summautuvat riskiluokkien yksiköiden lukumäärällä painotettuina takaisin perusjoukon yksiköiden lukumääräksi 10. Koska ainoa kontrolloitava arvo on tuo lukumäärä 10, ei otospainoja tarvitse kalibroida.

Taulukosta 10 nähdään kunkin yksikön status suojaamisen jälkeen. Ainutlaatuiset yksiköt muuttuivat suojauksen yhteydessä seuraavasti: Ensimmäinen havainto on jäänyt pois otantavaiheessa. Havainto 2 on selvinnyt otannasta, mutta muuttunut korvausvaiheessa ja vaikuttaa nyt ainutlaatuiselta. Havainto 3 selvisi muuttumattomana suojauksesta, mutta tuli väärinluokitelluksi yleiseksi parin omaavaksi yksiköksi. Yleisistä yksiköistä havainto 4 ei muuttunut suojauksessa, mutta näyttää nyt ainutlaatuiselta.

Havainto 5 jäi aineistosta pois, havainto 6 ei muuttunut, mutta sen status muuttui parillisesta ainutlaatuiseksi. Havainnon 7 arvoja korvattiin ja siksi se näyttäisi olevan osa kolmikkoa, vaikka alunperin se oli parin omaava yksikkö. Havainnot 8 ja 9 selvisivät muuttumattomina ja säilyttivät riskistatuksensa vaikkakin yksi alkuperäisen kolmikon yksiköistä vaihtui. Viimeinen havainto muuttui korvausvaiheessa ja sen alkuperäinen status vaihtui kolmikosta parin omaavaan.

Taulukko 10: MASSC-menetelmän esimerkkiaineisto. (Signh ja Dunteman 2004)

	Aineisto ennen suojausta				Korvaamisen jälkeen		Otannan jälkeen
Havainto	Ikä	Sp	Käyttö	Status ennen suojausta	Ikä	Sp	Status suojauksen jälkeen
1	1	1	1	Ainutlaatuinen, vaarassa	1	1	Poistettu otannassa
2	1	0	1	Ainutlaatuinen, vaarassa	1	1	Pseudo-ainutlaatuinen
3	2	1	1	Ainutlaatuinen, vaarassa	2	1	Pseudo-yleinen, pari
4	2	0	1	Yleinen, pari, vaarassa	2	0	Pseudo-ainutlaatuinen
5	2	0	1	Yleinen, pari, vaarassa	2	0	Poistettu otannassa
6	4	0	0	Yleinen, pari, vaarassa	4	0	Pseudo-ainutlaatuinen
7	4	0	0	Yleinen, pari, vaarassa	3	1	Pseudo-yleinen, kolmikko
8	3	1	1	Yleinen, kolmikko, ei vaaraa	3	1	Yleinen, kolmikko
9	3	1	0	Yleinen, kolmikko, ei vaaraa	3	1	Yleinen, kolmikko
10	3	1	1	Yleinen, kolmikko, ei vaaraa	2	1	Pseudo-yleinen, pari

Lasketaan seuraavaksi esimerkin aineistolle eri mittojen arvoja tarkastelemalla alkuperäistä ja suojattua aineistoa. Yksi ainutlaatuisista yksiköistä on selvinnyt suojauksesta muuttumattomana, mutta on tullut väärinluokitelluksi yleiseksi yksiköksi, joten todennäköisyys $\chi_U = \frac{1}{1}$. Vastaavasti $\chi_{NU} = \frac{2}{4}$, koska neljä yleistä yksikköä jäi muuttumatta ja kaksi niistä väärinluokiteltiin ainutlaatuisiksi. Todennäköisyys $\eta_{U(d)} = \frac{1}{1}$, koska ainutlaatuisista yksiköistä, jotka tulivat väärinluokitelluksi yleisiksi, se ainoa muuttui yleiseksi parin omaavaksi. Koska muita suojauksessa muuttumatta jääneitä ainutlaatuisia yksiköitä ei ollut, ovat $\eta_{U(t)} = \frac{0}{1}$ ja $\eta_{U(o)} = \frac{0}{1}$. Yleisistä yksiköistä, jotka eivät muuttuneet suojauksessa ja tulivat luokitelluiksi yleisiksi, molemmat vaikuttavat kolmikon osalta ja siksi todennäköisyydet ovat $\eta_{NU(d)} = \frac{0}{2}$, $\eta_{NU(t)} = \frac{2}{2}$ ja $\eta_{NU(o)} = \frac{0}{2}$. ζ-todennäköisyyksien ratkaisemiseksi oletetaan, että ne ovat samat molemmissa riskiluokissa. Tällöin $\zeta_d = \frac{0}{1}$, koska yleisistä yksiköistä vain yksi on pysynyt muuttumattomana ja tullut oikein luokitelluksi parin omaavaksi ja esimerkissä ei siksi saada tällaisille yksiköille eri arvoja alkoholin käytössä. Samoin perustein ne kaksi yleistä yksikköä, jotka säilyivät muuttumatta ja päätyivät edelleen osaksi kolmikkoa, antavat todennäköisyyden $\zeta_t = \frac{2}{2}$. Näiden yksiköiden alkoholinkäytön arvot eroavat siis toisistaan. Todennäköisyyttä ζ_o ei ole määritelty, koska suojatussa aineistossa on suurimmillaan kolmen yleisen yksikön ryhmä. Saatujen todennäköisyyksien avulla voidaan laskea

$$\delta_u = \left(\frac{3}{10}\right)\left(1-\frac{1}{3}\right)(1-1) + \left(\frac{7}{10}\right)\left(1-\frac{2}{7}\right)\left(\frac{6}{7}\right)\left(\frac{2}{4}\right) = 0 + 0,2143$$

ja

$$\begin{aligned}\delta nu(d) &= \left(\frac{3}{10}\right)\left(1-\frac{1}{3}\right)\left(\frac{2}{3}\right)\left(\frac{1}{1}\right)\left(\frac{1}{1}\right)\left(1-\frac{0}{1}\right) + \left(\frac{7}{10}\right)\left(1-\frac{2}{7}\right)\left(\frac{6}{7}\right)\left(1-\frac{2}{4}\right)\left(\frac{0}{2}\right) \\ &= 0,1333 + 0\end{aligned}$$

$\delta_{nu(t)}$ on nolla, koska $\eta_{nu(t)}$ on nolla ja ζ_t on yksi. Vastaavasti myös $\delta_{nu(o)}$ saa arvokseen nolla. Edellä esiteltyjen paljastumisriskin mittojen perusteella voidaan todeta, ettei suojatun aineiston yksiköistä yhdelläkään ole 100 % mahdollisuutta tulla paljastetuksi. Suurin riski yksittäiselle yksikölle on 21,43 %. Koska aineistossa on kymmenen havaintoa, saadaan paljastumisvaarassa olevien ainutlaatuisilta vaikuttavien yksiköiden lukumäärän odotusarvoksi $10 \times 0,2143$ eli 2,143. Kuitenkaan yksikään tämän esimerkin kolmesta suojatun aineiston ainutlaatuiselta vaikuttavasta yksiköstä ei ole vaarassa paljastua, koska mikään niistä ei ollut alunperin ainutlaatuinen. Aineistossa on kaksi paria yleisiä parin omaavia yksiköitä. Odotettavissa oleva lukumäärä paljastumisvaarassa oleville yksiköille on $10 \times 0,1333$ eli 1,333 ja todellisuudessa täsmälleen yksi havainto, havainto 2, on vaarassa paljastua.

Seuraavaksi tarkastellaan, miten estimoida tutkimusmuuttujan (y) RRMSE, joka tässä tapauksessa on esimerkiksi alkoholin käyttö aineiston miesten keskuudessa. Otospainolla painotettu keskiarvon estimaatti suojatusta aineistosta on
$N^{-1}\hat{\theta}_y^* = 10^{-1}\left(2 \times \frac{3}{2} + 2 \times \frac{7}{6}\right) = 0,53$ ja todellinen arvo alkuperäisestä aineistosta laskettuna on 0,40. Lasketaan ensin keskineliöpoikkeamaa varten ehdollisen harhan keskineliötä

$$E_\psi[B^2_{\phi|\psi}(N^{-1}\hat{\theta}_y^*)] = N^{-2}\left[\sum_h N_h(1-\psi_h)\psi_h S^2_{\nu,h} + \left(\sum_h \left(\sum_i \nu_i\right)\psi_h\right)^2\right],$$

missä $\nu_i = \tilde{y}_i - y_i$.

Esimerkin tapauksessa saadaan

$$10^{-2}[3(1-1/3)(1/3)(1/3) + 7(1-2/7)(2/7)(0) + (-1(1/3) + 0(2/7))^2] = 1/300.$$

Laskettaessa arvoa $S^2_{\nu,h}$ ainutlaatuisten yksiköiden luokassa, huomataan tähän tarvittavien kolmen arvon osalta, että ν saa ainoastaan yhden nollasta poikkeavan arvon. Tämän seurauksena luokan varianssille olisi saatu arvoksi $(1/3)(2/3)$, jos nimittäjä olisi ollut N_h. Koska nimittäjä on nyt $N_h - 1$, saadaan arvoksi 1/3. Seuraavaksi tarvitaan odotettavissa oleva ehdollinen varianssi, jonka arvoa voidaan estimoida ehdollisella varianssilla eli

$$V_{\phi|\psi}\left(N^{-1}\hat{\theta}_y^*\right) = N^{-2}\sum_h N_h(1/\phi_h - 1)S^2_{y^*,\,h},$$

missä $y^* = \tilde{y}_i$ tai y_i riippuen siitä, tuliko yksikkö muuttuneeksi korvausvaiheessa ja missä merkinnät y_i^*, y_i tarkoittavat korvauksessa käytetyn parin luovuttajan ja vastaanottajan saamia muuttujan y arvoja. Esimerkin tapauksesta huomataan, että $S^2_{y^*}$ on 0 ainutlaatuisten ositteessa ja 10/42 yleisten ositteessa. Näiden avulla lasketulle ehdolliselle varianssille saadaan

$$10^{-2}[7(7/6-1)(10/42)] = 10^{-2}(35/126),$$

eli RMSE-estimaatin arvoksi saadaan 0,0782 ja siten RRMSE on 0,195. Muiden tutkimusmuuttujien RRMSE voidaan laskea vastaavasti.

Esimerkkiä tarkastellessa on oletettu, että alkuperäinen kymmenen yksikön aineisto on perusjoukko. Tämän perusteella aineiston suojaaja pystyy laskemaan MASSC-menetelmällä suojatun aineiston paljastumisriskin ja informaatiokadon. Aineiston käyttäjä taas pystyy estimoimaan otannasta seuraavan ehdollisen varianssin, mutta korvaamisvaiheen aiheuttaman harhan laskemiseksi hänen olisi saatava yleistetty harhanfunktio, josta mainittiin aiemmin.

7.8 Simulointi

Rubin (1993) on esittänyt, että identifioitumisen riski voidaan poistaa kokonaan, jos käytetään alkuperäisen aineiston tilalla synteettistä aineistoa. Synteettinen aineisto muodostetaan simuloimalla, jolloin alkuperäisen ja synteettisen aineiston välillä ei ole mitään suoraa yhteyttä. Simulointi toteutetaan käyttäen esimerkiksi Bayesiläisiä menetelmiä tai moni-imputointia.

Ensi näkemältä loistavan menetelmän ongelmaksi muodostuvat ne synteettisen aineiston yksiköt, jotka kuitenkin vastaavat jotakin alkuperäisen aineiston yksikköä (Fienberg, 2000). Vaikka kyseiset yksiköt eivät todellisuudessa ole joutuneet identifioiduiksi, voi tilanne vaikuttaa siltä varsinkin tunnistetun yksikön näkökulmasta. Lisäksi tunkeutuja voi tunnistuksen tehtyään kokea onnistuneensa tavoitteissaan. Tuskin tunnistettu yksikkökään, olipa se sitten henkilö tai yritys, on tyytyväinen selitykseen, ettei hänen todelliset tietonsa ole paljastuneet. Ongelmaa tilanteessa lisää mahdollisuus saada täysin tekaistua informaatiota jostakusta virheellisesti tunnistetusta. Tällöin haitan määrä saattaa olla jopa suurempi kuin oikean tunnistamisen tilanteessa.

Simulointia on kritisoitu myös siksi, että se on suojausmenetelmä, jossa alkuperäisen aineiston ominaisuuksista vain ne, jotka aineiston suojaaja on valinnut säilytettäväksi simuloidessaan, on suojatun aineiston kautta käytössä. Nämä ominaisuudet ovat lisäksi täysin tilastollisia. Esimerkiksi Group Crisesin raportissa Synthetic Microdata Generation for Database Privacy Protection (2004f) kysytään: miksi turhaan simuloida aineistoa, kun kaikki mitä simuloidun aineiston avulla voidaan laskea, voidaan antaa suoraan tunnuslukuinakin. Tällainen asennoituminen simuloimista kohtaan on ollut vähintäänkin hidasteena simulointiin perustuvien menetelmien kehitykselle (Raghunathan et al. 2003).

Simuloituja aineistoja suositellaan käytettäväksi erityisesti silloin, kun muodostetaan julkisia aineistoja. Toisaalta pienten maiden aineistot ovat joskus paljastumisriskiltään liki julkisten aineistojen tasoisia ja simuloinnin käyttäminen Suomessa myös tutkijoille tarkoitetuissa aineistoissa on näin perusteltua. Simuloidun aineiston käytettävyyteen vaikuttaa eniten mallin sovittamisen onnistuminen. Jos aineiston simuloimisessa käytetty malli ei sovi alkuperäiseen aineistoon, eroaa simuloitu aineisto alkuperäisestä huomattavasti.

Moni-imputointi tietosuojamenetelmänä

Perinteinen tapa käyttää moni-imputointia tietosuojamenetelmänä perustuu idealle käyttää alkuperäisestä aineistosta otettua otosta. Otoksesta poisjääneet yksiköt tulkitaan puuttuviksi tiedoiksi ja ne paikataan tavalliseen tapaan moni-imputoinnin avulla. Ensimmäisten tietosuojaukseen tarkoitettujen moni-imputointi -menetelmien ongelmana oli, ettei niillä suojattuja aineistoja voitu käyttää perinteisten tilastoohjelmistojen avulla ja ne vaativat jopa erityisiä analysointimenetelmiä. Uudempien menetelmien tavoitteena on saada moni-imputoitu aineisto muotoon, jossa sitä voi analysoida perusohjelmistoilla tavallisin analyysimenetelmin. Eräs tapa suojata tällainen aineisto on moni-imputoinnin käyttö herkkien arvojen korvaamisessa (Little, 1993). Aineistosta saatuun otokseen valikoituneiden yksiköiden herkät arvot korvataan moni-imputoinnin avulla ja tämän jälkeen muunnettu aineisto voidaan julkaista. Vastaavasti voidaan suojata vain avainmuuttujia tai molempia muuttujia yhtä aikaa.

Esitellään tässä Rubinin (1993) esittämä moni-imputoinnin malli. Oletetaan, että alkuperäisestä aineistosta, jonka kokona on N, poimitaan n alkion otos s. Poimitun otoksen muuttujat jakautuvat taustamuuttujiin x_A, luovutettavissa oleviin muuttujiin x_B ja herkkiin muuttujiin x_C. Taustamuuttujat tunnetaan kaikille alkuperäisen aineiston N yksikölle, mutta muuttujien x_B ja x_C arvot vain otokseen mukaan tulleille n yksiköille. Ensimmäiseksi otoksen s avulla muodostetaan moni-imputoitu N yksikön aineisto. Tämä uusi perusjoukko sisältää kaikki n otoksen s yksikköä sekä M (moni-imputointien lukumäärä, yleensä kolmen ja kymmenen välillä) eri matriisia (x_B, x_C) -datalle otokseen kuulumattomien $N - n$ yksiköiden arvoiksi. Imputoitujen arvojen vaihtelu takaa, että moni-imputoidun aineiston avulla päädytään oikeisiin johtopäätöksiin. Taustamuuttujien x_A antamaa mallia tietojen (x_B, x_C) ennustamiseen käytetään hyväksi arvojen moni-imputoinnissa perusjoukkoon. Mallin valinta ei ole triviaali. Kun moni-imputoitu perusjoukko on muodostettu, poimitaan siitä n' yksikön otos s'. Tämän saadun otoksen koostumus vastaa n' yksikön otosta alkuperäisestä aineistosta. Vastaava moni-imputointi voidaan toteuttaa M kertaa, jolloin saadaan M kopiota (x_B, x_C) arvoille. Tuloksena on M moni-imputoitua synteettistä aineistoa. Jotta alkuperäisten yksiköiden joutuminen luovutettavaan aineistoon saadaan varmasti estettyä, muodostetaan lopullinen aineisto poimimalla otokset jokaisesta M synteettisestä aineistosta, joista kaikista on poistettu n alkuperäistä yksikköä.

Uudempien menetelmien etuna on paitsi perusohjelmistojen toimiminen myös lisäinformaation hyödyntäminen moni-imputoidun aineiston luomisessa. Lisäksi uudempien menetelmien avulla voidaan julkaista useita yksinkertaisella satunnaisotannalla perusjoukosta valittuja aineistoja sen sijaan, että annettaisiin tutkijan käyttöön asetelmaltaan monimutkaisia otoksia. Yksikertaisen satunnaisotoksen analysoimiseen sopivien ohjelmistojen käyttö on kuitenkin laajempaa kuin monimutkaisempiin aineistojen analysointeihin tarkoitettujen ohjelmistojen.

Olkoon alkuperäinen mikroaineisto kokoa n oleva otos kokoa N olevasta äärellisestä perusjoukosta $P = (x_A, x_B)$, missä $x_A = (x_{A_i}, i = 1, 2, \ldots, N)$ ovat taustamuuttujat (sisältäen design- ja hallinnolliset muuttujat) kaikille perusjoukon yksiköille ja $x_B = (x_{B_i}, i = 1, 2, \ldots, N)$ ovat kiinnostuksen kohteena olevat tutkimusmuuttujat, jotka on havaittu vain otoksen yksiköille. Jälkimmäinen voidaan indeksoida uudestaan: $x_{B_{inc}} = (x_{B_i}, i = 1, 2, \ldots, n)$ ovat yksiköt, jotka muodostavat havaittujen yksiköiden osuuden x_B:stä ja $x_{B_{exc}} = (x_{B_i}, i = n + 1, n + 2, \ldots, N)$ ovat ei havaittu osa eli otokseen kuulumattomat yksiköt. Näillä merkinnöin havaittu mikroaineisto on $s = \{x_A = (x_{A_i}, i = 1, 2, \ldots, N), x_{B_{inc}} = (x_{B_i}, i = 1, 2, \ldots, n)\}$. Yksinkertaisuuden vuoksi oletetaan, ettei havaittujen yksiköiden muodostamassa aineistossa olevilla yksiköillä ole puuttuvia tietoja.

Menetelmä koostuu kahdesta vaiheesta, joista ensimmäisen tarkoituksena on tuottaa useita synteettisiä populaatioita, $P^{(l)} = \{({x_A}^{(l)}, {x_B}^{(l)}, l = 1, 2, \ldots, M)\}$. Toisessa vaiheessa muodostetaan otos kustakin synteettisestä aineistosta ja julkaistaan kyseiset otokset. Yleensä otoksen saamiseksi käytetään yksinkertaista satunaisotantaa.

Ensimmäisen vaiheen toteutus siinä tapauksessa, ettei taustamuuttujien x_A julkaisulle ole estettä, on seuraava: Olkoot ${x_A}^{(l)} = x_A$ ja simuloidaan $(x_{B_{exc}}^{(l)}; l = 1, 2, \ldots, M)$ posteriorin jakauman ehdollisesta jakaumasta $\mathbb{P}(x_{B_{exc}i}|x_A, x_{B_{inc}})$. Tämä jakauma on ehdollinen ehdolla havaittu aineisto s ja mallioletukset. Jos kumpaakaan aineistoista x_A ja x_B ei voida julkaista, voidaan koko populaatio generoida superpopulaation odotettavissa olevan posteriorijakauman $\mathbb{P}(x_{A_f}, x_{B_f}|s)$ avulla. Tällöinkin jakauma on

ehdollinen ehdolla havaittu aineisto ja mallioletukset.

Yleensä perusjoukon koko N on liian suuri, jotta M synteettistä aineistoa olisi mahdollista julkaista. Siksi simuloinnin toinen vaihe varmistaa menetelmän käytännöllisyyden. Toisessa vaiheessa kustakin synteettisestä populaatiosta otetaan yksikertaisella satunnaisotannalla kokoa k oleva otos, $Z^{(l)} = (x^{(l)}_{A_i j_l}, x^{(l)}_{B_i j_l}, j_l = 1, 2, \ldots, k)$, missä $l = 1, 2, \ldots, M$ ja x_{A_i} tarkoittaa yksittäistä taustamuuttujaa. Tämän jälkeen julkaistaan M synteettistä otosta $Z = \{Z^{(l)}, l = 1, 2, \ldots, M\}$. Muitakin vaihtoehtoja julkaistaviksi valittaviin otoksiin on olemassa. Esimerkiksi osa taustamuuttujista x_A voi olla täysin julkaistavissa ja niitä voidaan käyttää suoraan koko perusjoukolle. Julkaistavissa olevia muuttujia voidaan liittää otoksiin $Z^{(l)}$. Jos taas muuttujat x_A ovat täysin suojattavia, voidaan niitä käyttää synteettisten aineistojen luomiseen, mutta julkaista sitten ainoastaan $(x^{(l)}_{B_i j_l}, j_l = 1, 2, \ldots, k)$. Joissakin tapauksissa voi olla toivottavaa käyttää synteettisen aineiston luomiseen muutakin kuin yksinkertaista satunnaisotantaa. Tästä on kuitenkin seurauksena aineiston analysoinnin vaikeutuminen.

Näytetään vielä esimerkki tällaisen synteettisen aineiston analysoinnista. Oletetaan, että aineiston käyttäjä haluaa tehdä päätelmiä perusjoukkoon liittyvästä lukumäärästä $Q = Q(x_A, x_B)$, joka voi riippua sekä muuttujista x_A että x_B. Oletetaan, että yksinkertaisella satunnaisotannalla saatua otosta käyttäessään tutkija käyttäisi pisteestimaattia q yhdessä epävarmuuden mitan v kanssa. Estimaatti q voisi olla esimerkiksi suurimman uskottavuuden estimaatti malliparametrille Q ja epävarmuutta v voisi mitata havaitun informaation käänteisarvo. Vastaavasti kyseiset arvot voisivat olla parametrin Q posteriorin keskiarvo ja varianssi tai harhattoman estimaatin arvo ja sen otosvarianssi.

Olkoot $(q^{(l)}, v^{(l)}), l = 1, 2, \ldots, M$ eri synteettisen aineistojen antamat arvot muuttujille q ja v. Tulkitaan joukko $(q^{(l)}, v^{(l)}, l = 1, 2, \ldots, M)$ synteettisten aineistojen Z_{Syn} yhteenvedoksi ja muodostetaan tämän perusteella approksimaatiot $\mathbb{P}(Q|Z_{Syn})$. Yksinkertaisin approksimaatio perustuu normaalijakaumaan ja käyttää estimaattien keskiarvoa

$$\bar{q}_M = \sum_l q^{(l)}/M$$

estimaatin Q posteriorikeskiarvona ja posteriorin varianssin approksimaationa on

$$T_M = (1 + M^{-1})d_M - \bar{v}_M,$$

missä $\bar{v}_M = \sum_l v^{(l)}/M$ ja $d_m = \sum_l (q^{(l)} - \bar{q}_M)^2/(M - 1)$.

Kuten esimerkistä voidaan huomata, ei moni-imputoinnilla muodostettujen synteettisten aineistojen kanssa toiminen eroa juurikaan tavallisen moni-imputoidun aineiston käytöstä. Tietosuojatilanteessa luovutetaan yleensä useita aineistoja.

8 Menetelmien empiirinen sovellus

Empiirisen sovelluksen ensisijaisena tavoitteena on tarkastella tilastollisten tietosuojamenetelmien toimivuutta Tilastokeskukselle tyypillisessä henkilöaineistossa. Käytettävänä aineistona on pätevien opettajien määrän arvioimista varten kerätty kokonaisaineisto Suomen perus- ja keskiasteen opettajista. Suoraan identifioivat muuttujat on poistettu aineistosta, mutta epäsuoraan identifioivia muuttujia on aineistossa niin paljon, että aineistossa olevien yksiköiden paljastumisriski on hyvin korkea. Paljastamista yrittävän henkilön kohteiksi voidaan ajatella erityisesti ne opettajat, jotka eivät ole päteviä hoitamaansa tehtävään. Aineiston herkkiä muuttujia ovat opettajien pätevyystiedot.

8.1 Aineisto

Tutkimuksen aineistona käytetään Opettajatilasto 2005 -aineistoa, joka on Suomen opettajista keväällä 2005 kerätty kokonaisaineisto. Tutkimuksen osalta rajoitutaan peruskoulujen ja lukioiden opettajiin, jolloin N=52069. Koska tutkimuksen tavoite on tarkastella tietosuojamenetelmiä, eikä niinkään analysoida käytettävissä olevaa aineistoa, voidaan aineistoa rajata edelleen pienemmäksi. Lopulliseksi tutkimusaineistoksi valittiin lukion opettajat, joille N=7798. Aineiston rajaaminen lukion ja peruskoulun opettajiin toteutettiin opettajan virkatiedon perusteella.

Kouluista käytettävissä olevia tietoja ovat koulun nimi, sijaintikunta, opetuskieli, oppilaitostyyppi ja oppilaitoksen opiskelijoiden määrä ala- ja yläluokilla. Opettajiin liittyviä tietoja aineistossa ovat opettajan syntymäaika, ikä, työsuhteen luonne, tehtävätyyppi, opettajan opettamat aineet, kelpoisuus opettamaansa tehtävään, mahdollisen epäpätevyyden syyt ja opettajan muut opettajakelpoisuudet. Tarkemmin aineiston sisältämistä muuttujista on tietoja liitteessä 2.

Aineistossa suoraan identifioivia muuttujia ei varsinaisesti ole, koska opettajien henkilötunnukset ja nimet on poistettu. Epäsuoraan identifioivia muuttujia ovat opettajan syntymäaika ja ikä, koulun sijaintikunta sekä opettajan opettamat aineet. Aineiston herkkiä muuttujia ovat opettajan pätevyystiedot eli opettajan kelpoisuuteen ja kelpoisuuden puutteisiin liittyvät muuttujat. Erityisen herkkä on tieto siitä, onko opettaja kelpoinen hoitamaansa tehtävään.

8.2 Käytettävät suojausmenetelmät

Aineiston suojaamisessa käytettävä μ-Argus-ohjelma on alunperin Alankomaiden tilastoviraston mikroaineistojen suojaamiseen kehittämä ohjelma. Ohjelmaa on kehitetty edelleen Computational Aspects of Statistical Confidentiality (myöhemmin CASC) ja a CENtre of EXcellence for Statistical Disclosure Control (myöhemmin CENEX-SDC) projektien yhteydessä. Molemmat projektit ovat EU:n rahoittamia ja niihin osallistui useita tilastovirastoja ja yliopistoja ympäri Euroopan. Projektien tavoitteena oli tutkia ja kehittää uusia tietosuojamenetelmiä ja edistää niiden soveltamista (CENEX-SDC -kotisivut). Kehitystyö on keskittynyt erityisesti käytännön työkaluihin eli ohjelmistoihin ja niiden jatkokehittämiseen (CASC-kotisivut). Työn tuloksena ovat uudet versiot Argus-ohjelmista, joista toinen, τ-Argus, on tarkoitettu taulukkoaineistojen suojaukseen. Koska kyseessä on yleiseurooppalainen projekti, on yksi

tavoitteista kehittää yhteiset pelisäännöt tietosuojaukselle koko Euroopassa. Vapaasti saatavilla olevat ohjelmat tukevat hyvin tätä tavoitetta. μ-Argus sisältää useita teoriaosuudessa esiteltyjä suojausmenetelmiä:

- luokkien vähentäminen (global recoding),
- arvojen peittäminen (local supression),
- luokkien laajentaminen ylä- tai alareunoista (top and bottom coding),
- PRAM (the Post RAndomisation Method),
- mikroaggregointi (numerical micro aggregation),
- arvojen vaihtaminen yksiköiden välillä järjestykseen perustuen (numerical rank swapping) ja
- Sullivanin menetelmä (Sullivan masking).

Tutkimuksessa ei ole mielekästä tarkastella kaikkia näitä menetelmiä pintapuolisesti ja siksi tutkielmassa on keskitytty erityisesti mikroaggregointiin ja PRAM-menetelmään. Nämä menetelmät tulivat valituiksi, koska ne ovat tällä hetkellä Tilastokeskuksessa käytettävistä menetelmistä poikkeavia ja siksi tulosten kannalta mielenkiintoisimpia. Tulosten perusteella voidaan harkita Tilastokeskuksen tietosuojaamisen tukemista tai käytössä olevien menetelmien korvaamista näillä tutkimuksessa testatuilla.

Aineistoja suojatessa on hyvä huomata, että μ-Argus-ohjelma "nollaa" aiemman muunnoksen, kun aineistolle toteutetaan uusi saman menetelmän muunnos. Muuttuja kerrallaan suoritettava mikroaggregointi saadaan toteutetuksi lisäämällä uusi aggregointi jo aiemmin suojattuun aineistoon. Koska tämä on melko vaivalloista ja jo kirjallisuuden perusteella tiedetään, ettei muuttuja kerrallaan mikroaggregoitu aineisto ole erityisen varmasti suojattu, ei tätä suojaustapaa tutkittu enää empiirisessä tutkimuksessa.

Sullivanin menetelmää voidaan μ-Argus-ohjelmassa käyttää vain, jos aineistosta suojattavaksi valittavat muuttujat sisältävät vähintään kaksi jatkuvaa muuttujaa ja luokiteltujen muuttujien luokkien lukumäärä on vähemmän kuin kaikkien muuttujien lukumäärä yhteensä. Sullivanin menetelmässä on useita parametreja joiden arvot vaikuttavat suojaamiseen. Sullivanin menetelmä olisi ollut hyödyllinen osa empiiristä tutkimusta, koska muiden suojausmenetelmien lopputuloksia olisi voitu verrata kirjallisuudessa varmaksi todetun menetelmän tuloksiin. Koska Sullivanin menetelmä ei sovellu tutkimukseen valitun aineiston suojaamiseen, oli menetelmä jätettävä empiirisen tutkimuksen ulkopuolelle.

Suojaamisen aluksi muuttujille on määritettävä metadata. Metadataan määritellään muuttujan tyyppi eli onko muuttuja numeerinen vai luokiteltu vai onko muuttuja paino, jota on käytetty otoksen valinnassa. Tämän lisäksi annetaan tieto, voidaanko arvoja pyöristää ja liittyykö muuttuja yksiköiden muodostamaan ryhmään (esim. kotitalousmuuttujat). Luokitelluille muuttujille on mahdollista muodostaa tiedosto, jonka perusteella voidaan muuttujan tiedon tarkastelu vaihtaa yksityiskohtaisesta tiedosta yleisempään. Esimerkiksi paikkatiedon kohdalla voitaisiin kuntatason tiedoista siirtyä seutukuntatasolle. Tämän lisäksi muuttujan tärkeyttä identifioitumisessa arvioidaan asteikolla nollasta viiteen ja valitaan muuttujat suojauksen lopuksi tapahtuvaa arvojen peittämistä varten. Mitä pienempi peittämisestä varten valittu paino on, sitä harvemmin kyseisen muuttujan arvoja tullaan peittämään ja vastaavasti suuremman painon omaavien muuttujien arvot tulevat helpommin peitetyiksi. Metadatan määrittelyyn kannattaa panostaa, koska tarkasti määritellyn metadatan avulla

itse suojausprosessi sujuu melko helposti.

Aineiston suojaajan on määritettävä ne muuttujat, joiden avulla epäsuora identifioituminen on mahdollista. Identifioimiseen tarvittavien muuttujien arvot ristiintaulukoidaan ja taulukosta saadaan selville harvinaiset profiilit. Kun ristiintaulukoitavat muuttujat on annettu, ohjelma laskee automaattisesti liian pieniä frekvenssejä sisältävien solujen lukumäärän. Käyttäjän on annettava suurin frekvenssi, jonka solun yksiköt on vielä suojattava. Tilastokeskuksessa käytetään yleisimmin liian pienelle frekvensille arvoa 5 tai pienempi. Haluttaessa voidaan suojaus aloittaa myös paljastumisriskitarkastelulla. Paljastumisriskin määrittämistä varten kullekin aineiston muuttujalle on annettava tarkat tiedot muuttujan vaikutuksesta paljastumisriskiin. Tässä tutkimuksessa käytettiin suojauksen lähtökohtana profiilien frekvenssejä.

Tässä tutkimuksessa varsinainen tavoite ei ollut turvallisen aineiston muodostaminen, joten aineistosta ei suojauksen lopuksi enää peitetty arvoja. Käytännössä suojauksen jälkeen saadut aineistot eivät olisi luovutettavissa tutkijoille. Tutkimuksen kannalta kiinnostavat aineistossa tapahtuneet muutokset ovat näin suojatusta aineistosta tarkasteltavissa. Koska aineistoa ei muokattu valmiiksi, ei aineiston paljastumisriskejä ei ole mielekästä laskea vaan tarkasteluissa keskitytään pelkästään aineistossa tapahtuneisiin muutoksiin.

Mikroaggregoinnin soveltaminen

Aineisto suojattiin mikroaggregoimalla käyttäen suojaamiseen muuttujia: ikä, opettajan virka ja kouluaste, jolla opettaja ensisijaisesti antaa opetusta. Koska tämän hetkisessä μ-Argus-ohjelman versiossa luokiteltujen muuttujien mikroaggregointi ei ole käytettävissä, suoritettiin suojaaminen kaikille muuttujille asettamalla niiden metatietoihin tyypiksi jatkuva muuttuja. Luokiteltujen muuttujien suojaaminen jatkuvana muuttujana aiheuttaa ongelman, kun ohjelma antaa suojauksen kautta sellaisiakin arvoja, joita alkuperäisessä aineistossa ei ollut mukana. Tämä on seurausta siitä, että ohjelmalle on virheellisesti kerrottu muuttujien olevan jatkuvia ja kaikille jatkuvien muuttujien arvoille on järkevä tulkinta.

Luokkien uuden nimeämisen seurauksena suojattujen aineistojen frekvensseihin saatiin muutoksia. Nämä muutokset olivat seurausta mikroaggregointia varten muodostettujen ryhmien koostumuksen muuttumisesta. Kun virkaluokan numero oli aiemmin ollut 4, oli se uuden nimeämisen jälkeen 2. Tämän muutoksen seurauksena vaikuttivat kahden muun muuttujan arvot: ikä ja kouluaste aiempaa enemmän aggregoitavan ryhmän valintaan. Frekvenssien erot johtuivat siis siitä, että ryhmät muodostuivat uuden nimeämisen jälkeen eri yksiköistä kuin alkuperäisiä luokkien nimiä käytettäessä. On kuitenkin hyvä huomata, että käytettäessä MDAV-algoritmiä ryhmien muodostamisessa, on luokkien nimeämisellä eli tässä tapauksessa numeroinnilla vaikutusta siihen, mitkä yksiköt valikoituvat ryhmiin.

Mikroaggregoinnin vaikutusta aineistoon tarkasteltiin muuttamalla ainoaa μ-Argus-ohjelmassa suojauksen tasoon vaikuttavaa parametrin arvoa eli aggregoitavan ryhmän kokoa. Tutkimuksessa ryhmäkokoa vaihdettiin arvosta kaksi alkaen kokoon 11 ja lisäksi testattiin ryhmäkokoa 15. Mikroaggregoinnin suojaavuudesta saatiin käsitystä vertaamalla suojattuja aineistoja alkuperäiseen aineistoon. Lisäksi mikroaggregoitujen aineistojen frekvenssejä verrattiin PRAM-menetelmällä suojattujen aineistojen vastaaviin.

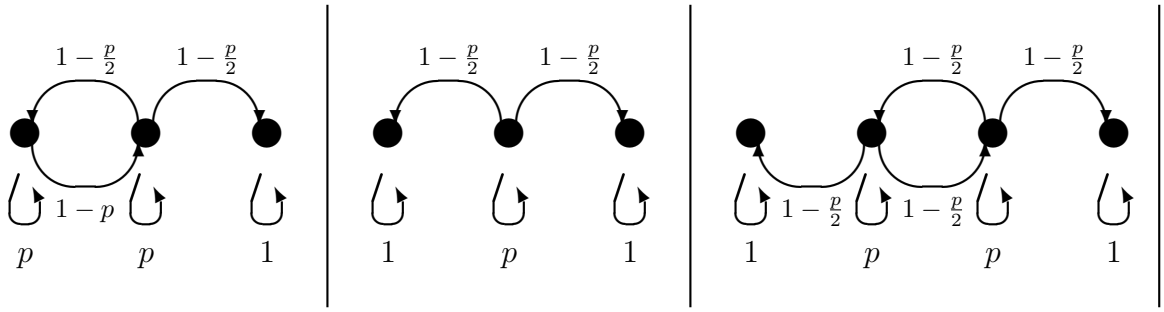
PRAM-menetelmän soveltaminen

PRAM-menetelmällä suojaaminen toteutettiin käyttämällä suojaamiseen lukioaineiston muuttujia: ikä, virka, opettajan opetuksen pääasiallinen kouluaste ja opetettavat aineet eniten opetettavasta kolmanneksi eniten opetettavaan. PRAM-menetelmä toimii ainoastaan luokitelluille muuttujille, joten kaikki muuttujat merkittiin metatietoihin luokitelluiksi, vaikka ikä-muuttujan voisi tulkita myös jatkuvaksi. Koska ikämuuttujalla on kymmeniä eri luokkia, on sen PRAM-matriisi hyvin suuri. Siitä huolimatta ohjelma toimi moitteettomasti suojauksen aikana.

PRAM-menetelmän suojaavuutta voitiin tarkastella kahden eri parametrin arvon muuttamisen seurauksena. PRAM-matriisin diagonaalin arvot voitiin asettaa halutuksi vaihtumistodennäköisyyden 0,01 tarkkuudella. Tämän lisäksi aineiston arvojen vaihtumista voitiin rajoittaa. Raja oli valittava kokonaisluvuista, joiden maksimiarvo riippuu tietenkin suojattavan muuttujan luokkien määrästä. Nämä asetukset ovat muuttujakohtaisia ja siksi suojaus olisi voitu toteuttaa eri PRAM-matriiseja käyttäen kullekin muuttujalle. Lisäksi ohjelmassa olisi voitu asettaa vaihtumistodennäköisyys kullekin luokalle erikseen, mutta tällaisten asetusten tuloksia olisi ollut hyvin vaikea analysoida.

Tutkimuksessa päädyttiin tulosten vertailtavuuden takia pitämään suojausparametrien asetukset samoina kaikille muuttujille. Arvojen vaihtumista ei ensimmäisellä suojauskerralla rajoitettu ollenkaan ja tämän jälkeen kaikkien muuttujien arvot saivat vaihtua korkeintaan kahden luokan verran. Muuttujien arvojen vaihtumiselle asetettu rajoitus valittiin ottamalla huomioon suojattavien muuttujien luokkien lukumäärät. Aste-muuttujalla luokkien lukumäärä on 12. Jos arvojen olisi annettu vaihdella enemmän, ei suojauksessa rajoittamattomasti vaihtuvan ja rajoitetun vaihtumisen välille olisi saatu eroa. Tämän lisäksi muuttujien arvojen vaihtaminen kauempana oleviin arvoihin ei ole sisällöllisesti mielekästä.

Molempien tapausten: vaihtamisen rajoittamisen ja rajoittamatta jättämisen, sisällä tehtiin tarkemmat tarkastelut eri vaihtumistodennäköisyyksien arvoille. Todennäköisyydet lähellä nollaa tarkasteltiin tarkemmin. Muutoksen todennäköisyys sai arvot: 0,03, 0,05, 0,07 ja 0,10. Tätä suuremmille vaihtumistodennäköisyyksille todennäköisyyttä kasvatettiin arvolla 0,05 aina arvoon 0,40 asti.



Kuva 2: PRAM-menetelmässä vaihtumisen rajoittaminen ryhmien sisälle. Rajalla olevan luokan arvo on suojaamisen jälkeen sama todennäköisyydellä yksi.

Tutkimuksen edetessä tuli mielenkiintoiseksi tarkastella tilannetta, jossa aine-muuttujan arvojen vaihtaminen rajoitettiin oman aineryhmän sisälle. Tarkastelussa keskityttiin vain äidinkielen, kielten ja matemaattis-luonnontieteellisten aineiden opettajien frekvensseihin. Suojaus toteutettiin asettamalla muuttujanarvon vaihtumistodennäköisyydet kahden aineryhmän rajalla molemmille rajalla oleville luokille nollaksi. Lisäksi ryhmien sisällä sallittava vaihtuminen oli rajattava korkeintaan yhden luokan päähän, jolloin arvot eivät voineet hypätä edellä mainitun rajan yli. Tarkastelu toteutettiin vaihtumistodennäköisyyksille 0,03, 0,05 ja 0,07.

8.3 Sovellettavat tilastolliset menetelmät

Aineistoa tutkittiin laskemalla erilaisia frekvenssejä ja reunajakaumia sekä sovittamalla opettajan pätevyyttä ennustava malli aineistoon. Kyseiset tarkastelut toteutettiin alkuperäiselle aineistolle sekä suojaamalla saaduille aineistoille. Suojatut aineistot oli muokattu kahdella eri tietosuojamenetelmällä ja suojaaminen oli toteutettu useille parametrien arvoille. Näin tutkimusta varten oli käytettävissä monia aineistoja, joiden avulla frekvenssien sekä mallin estimaattien muutoksia voitiin verrata. Vertailujen tavoitteena oli selvittää suojauksen vaikutusta aineistojen käytettävyyteen. Lisäksi pohdittiin suojausmenetelmien aiheuttamia rajoitteita aineistoissa ja muuttujissa esimerkiksi aineiston kokoon tai muuttujien tyyppiin liittyen.

Aineistoista vertaillaan ensin frekvenssien muutoksia, kun suojatessa siirrytään yhä vahvempaan suojauksen tasoon. Frekvenssien muutosten perusteella pystytään arvioimaan aineiston käytettävyyttä erityisesti tunnuslukujen perusteella tehtäviin analyyseihin. Muuttujat, joiden frekvenssejä laskettiin, olivat: ikä, opettava aine, opettajan virka ja kouluaste, jolla opettaja ensisijaisesti opettaa. Tutkielman osaksi tuotettuihin taulukoihin muuttujien luokkia on yhdistetty tiiviimmän esitysmuodon saamiseksi. Kuitenkin olennaisia muutoksia esitetään myös tarkemmalla luokkatasolla.

Saatujen reunajakaumien muutoksia tarkastellaan verrattuna alkuperäiseen aineistoon, eri parametrien arvoilla suojausmenetelmän sisällä sekä myös suojausmenetelmien välillä. Tämän tarkoituksena on verrata mikroaggregoinnin ja PRAM-menetelmän käytettävyyttä ja toisaalta menetelmän sopivuutta tiettyjä analyysejä varten pyydetyn aineiston tai tietyn tyyppisten muuttujien suojaamiseen.

Frekvenssijakaumien analysoimisen lisäksi aineistoon sovitettiin malli, jolla estimoitiin opettajan pätevyyttä. Mallin vastemuuttujaksi asetettiin tieto siitä, onko opettajan pätevyydessä puutteita vai ei. Muuttujan arvo kertoo, onko opettaja pätevä hoitamaansa tehtävään eli esimerkiksi yläluokilla sijaisuutta tekevä luokanopettaja on epäpätevä vaikka onkin pätevä luokanopettaja. Selittävinä muuttujina mallissa käytetään opettajan ikää, sukupuolta ja opettajan virkaa. Näistä ikää tarkasteltiin jatkuvana muuttujana ja alkuperäinen virka-muuttuja oli mukana aggregoituna kahdeksan luokkaisena virkal-muuttujana. Sukupuoli oli mukana mallissa siitä huolimatta, ettei se ollut mukana suojaamisessa. Sukupuoli-muuttujan saamia parametrin arvoja voikin käyttää verrattaessa tilannetta suojattujen ja suojaamattomien muuttujien välillä.

8.4 Menetelmien soveltaminen frekvenssijakaumille

Molempien suojausmenetelmien tarkastelu aloitettiin tutkimalla suojausparametrien vaikutusta suojattavaan aineistoon. Mikroaggregoinnin yhteydessä keskityttiin valitun ryhmäkoon vaikutukseen suojatun aineiston frekvensseihin. Ryhmien määräytymisessä kaikki suojattavat muuttujat olivat mukana samanaikaisesti eli suojausta ei toteutettu lainkaan muuttuja kerrallaan. Sen sijaan ohjelma etsi ryhmäkoon mukaisen määrän samankaltaisimpia yksiköitä ottaen huomioon kaikkien suojattavien muuttujien arvot ja muodosti näistä aggregoitavan ryhmän.

PRAM-menetelmän osalta kiinnostuksen kohteena oli arvojen vaihtumistodennäköisyyden vaikutus muuttujien frekvensseihin. Lisäksi tarkasteltiin, mikä vaikutus oli arvojen vaihtumisen rajaamisella. Tässä yhteydessä mietittiin myös PRAM-menetelmän soveltuvuutta eri tyyppisten muuttujien suojaamiseen.

Mikroaggregoinnin soveltaminen

Mikroaggregoinnista on käytettävissä μ-Argus-ohjelmassa vasta jatkuvien muuttujien suojaamiseen tarkoitettu versio, joten luokiteltujen muuttujien suojaaminen mikroaggregoimalla ei ole vielä kovin suositeltavaa. Suojaaminen kuitenkin haluttaessa onnistuu, mikäli muuttuja on järjestysasteikollinen tai luokat on määritelty kokonaisluvuin ja kaikki luokat ovat käytössä. Tutkittavan aineiston kohdalta jälkimmäinen ehto poistaa mahdollisuuden soveltaa mikroaggregointia suoraan muun muassa opetettaviin aineisiin. Jos muunlaisia muuttujia halutaan kuitenkin suojata mikroaggregoimalla, voidaan luokat nimetä uudelleen juoksevalla numeroinnilla. Keinotekoista järjestystä käyttäessään on kuitenkin vaarana, että suojauksen lopputulos aiheuttaa virhetulkintoja. Luokkien nimiä muuttaessaan onkin oltava hyvin tarkkana, että uusi luokittelu toimii myös sisällöllisesti.

Tutkielman aineiston muuttujista opettajan ikä, virka ja kouluaste suojattiin mikroaggregoimalla. Ensimmäisellä kerralla aineiston luokittelujen annettiin pysyä ennallaan. Tämän aineiston ongelmaksi muodostuvat opettajan virkatyypit, jollaisia ei lukion opettajissa ole mukana, mutta joiden arvoja saadaan aggregoitavien ryhmien keskiarvona. Vaikka alkuperäisessä aineistossa ei ole ollut erityisopettajia mukana, on suojatussa aineistossa muutamien opettajien virkaluokan koodi muuttunut erityisopettajan koodiksi kuten taulukosta 11 nähdään. Tällainen suojattu aineisto ei kelpaa luovutettavaksi tutkijalle, koska siinä on selvä virhe.

Mikäli taulukon 11 alkuperäisillä luokituksilla saatuja frekvenssejä tarkastellaan yleisemmin, voidaan nähdä, että ryhmäkoko saa kasvaa kuudeksikin ilman, että frekvensseissä tapahtuu edellä mainittu virhe lukuun ottamatta kovin huomattavia muutoksia. Riittävän suuressa aineistossa kuten esimerkkimme aineisto frekvenssien muuttuminen noin viidellä ei vielä aiheuta suuria muutoksia yleisimpiin aineistosta tehtäviin analyyseihin. Tämän tutkielman mikroaggregoitujen aineistojen frekvenssimuutoksia ei kannata tarkastella muuttuja kerrallaan, koska suojaaminen toteutettiin kaikille kolmelle muuttujalle yhtä aikaa. Tämän seurauksena esimerkiksi kouluasteen 01 ainoa yksikkö on voinut muiden muuttujanarvojen perusteella tulla ryhmitellyksi ryhmäkoosta riippuen erilaisiin ryhmiin ja saada tämän perustella hyvinkin paljon ykkösestä eroavan keskiarvon aste-muuttujalle. Sama ilmiö selittää muidenkin pienten frekvenssien näennäisen suurta vaihtumista ryhmäkoosta toiseen.

Taulukko 11: Kolmen muuttujan yhtä aikaisella mikroaggregoinnilla saadun aineiston frekvenssejä. Taulukossa ovat saadut frekvenssit alkuperäisiä ja uusia luokituksia käyttäen. (Alkuperäistä aineistoa on merkitty ryhmäkoolla 1.)

k	Ikä	n		Virka	n		Aste	n	
		alkup.	uusi		alkup.	uusi		alkup.	uusi
1	alle 40	2467		rehtori	432		01	1	
	40–53	2849		luokanopettaja	-		02	10	
	yli 53	2482		erityisopettaja	-		03	343	
				lehtori	5684		04	7	
				opinto-ohjaaja	142		05	6672	
				päätoiminen tuntiop.	1078		06	686	
				sivutoiminen tuntiop.	414		07	22	
				muu virka	48		08	7	
							09	7	
							10	16	
							11	3	
							12	24	
2	alle 40	2468	2468	rehtori	430	428	01	0	0
	40–53	2840	2840	luokanopettaja	-	-	02	6	8
	yli 53	2490	2490	erityisopettaja	4	-	03	346	346
				lehtori	5676	5684	04	6	6
				opinto-ohjaaja	152	142	05	6672	6672
				päätoiminen tuntiop.	1068	1076	06	686	684
				sivutoiminen tuntiop.	420	416	07	22	24
				muu virka	48	52	08	6	8
							09	8	4
							10	16	18
							11	4	4
							12	24	24
6	alle 40	2460	2460	rehtori	420	432	01	0	0
	40–53	2860	2860	luokanopettaja	6	-	02	6	6
	yli 53	2478	2478	erityisopettaja	12	-	03	348	348
				lehtori	5656	5674	04	0	0
				opinto-ohjaaja	156	162	05	6682	6676
				päätoiminen tuntiop.	1086	1080	06	678	684
				sivutoiminen tuntiop.	408	402	07	24	30
				muu virka	54	48	08	12	6
							09	6	6
							10	12	12
							11	6	6
							12	24	24
10	alle 40	2450	2470	rehtori	420	420	01	0	0
	40–53	2878	2858	luokanopettaja	-	-	02	10	10
	yli 53	2470	2470	erityisopettaja	20	-	03	340	340
				lehtori	5658	5658	04	10	10
				opinto-ohjaaja	150	140	05	6678	6668
				päätoiminen tuntiop.	1120	1080	06	680	690
				sivutoiminen tuntiop.	400	380	07	30	30
				muu virka	30	120	08	0	0
							09	10	10
							10	10	10
							11	10	10
							12	20	20

Jos suojaaminen olisi toteutettu muuttuja kerrallaan, olisi ryhmäkoolla kaksi aste-muuttujan ainoa arvon 01 saanut yksikkö muuttunut aste-muuttujan arvolle 02. Tämä johtuu mikroaggregoinnin asetuksesta, jonka mukaan ryhmäkoolla kaksi suojattaessa jokaisen ryhmän on oltava vähintään kahden yksikön muodostama. Tällöin kyseisen ryhmän molempien yksiköiden saama muuttujanarvo on esimerkin tilanteessa joko 01 tai 02 ja näiden keskiarvo: 1,5 pyöristetään ylöspäin. Joissakin tapauksissa tällä mikroaggregoinnin pyöristämistavalla voi olla merkitystä ja siksi se on hyvä pitää mielessä suojausmenetelmää käyttäessään.

Mikroaggregointi tuntuisi toimivan suojattavien muuttujien tapauksessa hyvin, joten on syytä muokata viran luokittelua tarkemman tiedon saamiseksi menetelmän soveltamisesta.

Luokituksen vaihtamisella päästään eroon keskiarvojen aiheuttamista ongelmista. Jo jos verrataan taulukon 11 ryhmäkoolla 2 saamia frekvenssejä, huomataan, ettei alkuperäisestä aineistosta täysin puuttuneita luokanopettajia tai erityisopettajia ilmesty enää aineistoon. Suojatun aineiston oikeellisuus saadaan siis säilymään luokittelun muutoksella.

Vaikka luokittelun muuttaminen vaikuttaa aineistosta saataviin yksittäisiin frekvenssien arvoihin, nähdään edelleen, että frekvenssien arvot säilyvät likimain jopa ryhmäkoolla 6. Ryhmäkoko 10 on jo liian iso, koska useiden pienten frekvenssien, kuten aste-muuttujan luokat 1, 4, 8, 9 ja 11, saama arvo voi olla joko nolla tai kymmenen riippuen ryhmien muodostumisesta. Esimerkiksi arvon 7 saaneen luokan frekvenssin muuttuminen nollaksi voi aiheuttaa aineiston käyttäjälle väärän mielikuvan tutkimuksen tilanteesta tai jopa virheellisiä lopputuloksia.

Tarkastellaan vielä korjatun aineiston frekvenssejä ristiintaulukoitaessa kahta muuttujaa. Taulukko 12 kertoo oletetusti, että opettajan virkaluokalla ja pätevyydellä on yhteys. Liki kaikki aineistossa olevat rehtorit ovat päteviä hoitamaansa tehtävään samoin kuin liki kaikki lehtorin virassa olevat opettajat. Sivutoimisista tuntiopettajista sen sijaan jopa puolet ovat epäpäteviä hoitamaansa virkaan ja päätoimisista tuntiopettajista epäpäteviä on liki viidennes. Näistä tiedoista voidaan tehdä johtopäätös siitä, että mitä pysyvämpi virka opettajalla on, sitä suuremmalla todennäköisyydellä opettaja on pätevä. Tämä on luonnollista, koska opettajan virkaan nimitettäessä pitäisi opettajan olla pätevä. Jos taas opetusta hoidetaan sivutoimisesti, voidaan tehtävään valita helposti myös epäpätevä henkilö.

Opettajan pätevyyden ja iän välillä olevaa korrelaatiota voidaan tarkastella taulukosta 13. Vaikka pätevyys ja ikä korreloivat vahvasti, ei korrelaatio ole yhtä vahvaa kuin opettajan viran ja pätevyyden välillä. Tarkasteltava ikä-muuttuja on taulukossa 13 luokiteltu hyvin laajoihin luokkiin ja tällä voi olla vaikutusta korrelaatioon. Ikä- ja pätevyys-muuttujan ristiintaulukoituja frekvenssejä tarkastellessa näkee, että mikroaggregointi toimii erinomaisesti jatkuvalle muuttujalle. Frekvensseissä ei ole juurikaan eroja vaikka ryhmäkoon annettaisiin kasvaa kymmeneen. Nyt täytyy tietenkin muistaa, että taulukon 13 arvot on saatu aineistosta, joka on yksikkötasolla suojattu ja laajat luokat on määritetty jo suojatuille arvoille.

Taulukosta 12 voidaan tutkia mikroaggregoinnin vaikutusta kaksiulotteisiin reunajakaumiin. Tarkastellaan kahta muuttujaa: virkaluokkaa ja pätevyyttä, joista vain ensimmäinen on ollut mukana suojauksessa. Frekvenssien säilymisestä voidaan todeta, että yksiulotteisten reunajakaumien tarkastelusta saatu johtopäätös mikroaggregoin-

nin toimivuudesta pätee edelleen. Jakauma ei ole taulukossa 12 juurikaan muuttunut ryhmäkokoon kuusi asti ja tarkempi tarkastelu liitteen 1 taulukoista 10–13 osoittaa, ettei mikroaggregoinnilla suojaaminen aiheuta huomattavia muutoksia edes kaksiulotteiseen jakaumaan. Myöskään tutkittavien kahden muuttujan väliseen korrelaatioon suojaamisella ei näyttäisi olevan vaikutusta, koska testien t-arvot pysyvät liki saman suuruisina ryhmäkoon kasvusta huolimatta.

Mikroaggregointi säilyttää erinomaisesti frekvenssien arvot suojauksessa, jos ryhmäkoon annetaan pysyä korkeintaan noin viidessä. Kuitenkin suojauksen tarkoitusta ajatellen tämä voidaan ajatella myös menetelmän huonoksi puoleksi. Jos kiinnostuksen kohteena olevasta yksiköstä tunnetaan muuttujien tarkkoja arvoja tai arvoja suunnilleen, voidaan mikroaggregoinnilla syntyviä keskiarvoja hyödyntää tunnistamisessa. Kuitenkin useamman muuttujan samanaikainen mikroaggregointi vähentää tätä riskiä, kun ryhmät muodostuvat useamman muuttujan arvojen kautta ja siten aina samassa ryhmässä ei ole mukana esimerkiksi vain saman ikäisiä henkilöitä.

Tutkielman luokitellut muuttujat eivät tuo mikroaggregoinnin parhaita puolia esiin, koska käytettävissä oleva sovellus on tarkoitettu jatkuvia muuttujia varten. Kun tarkastellaan frekvenssejä, näyttäisi menetelmä toimivan tällä sovelluksella kelvollisesti myös luokiteltujen muuttujien suojaamisessa. Jatkuvan ikä-muuttujan kohdalla tilanne on erityisen hyvä. Luokitelluilta muuttujilta vaaditaan monia ominaisuuksia tai aineiston suojaajan on nähtävä vaivaa saadakseen muuttujat suojaamiseen sopivaan muotoon.

Taulukko 12: Pätevyyden ja virkaluokkien ristiintaulukointi. Suojattu käyttäen mikroaggregointia. (Alkuperäistä aineistoa on merkitty ryhmäkoolla 1.)

k	Virka	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
	rehtorit	425	7	432
		98,38	1,62	100,00
	lehtorit	5519	165	5684
		97,10	2,90	100,00
	op.-ohjaajat	125	17	142
		88,03	11,97	100,00
1	päätoim. tuntiop.	875	203	1078
		81,17	18,83	100,00
	sivutoim. tuntiop.	208	206	414
		50,24	49,76	100,00
	muu virka	39	9	48
		81,25	18,75	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1422,4648	p<0,0001	
	LR	932,8533	p<0,0001	
	rehtorit	421	7	428
		98,36	1,64	100,00
	lehtorit	5518	166	5684
		97,08	2,92	100,00
	op.-ohjaajat	124	18	142
		87,32	12,68	100,00
2	päätoim. tuntiop.	876	200	1076
		81,41	18,59	100,00
	sivutoim. tuntiop.	212	204	416
		50,96	49,04	100,00
	muu virka	40	12	52
		76,92	23,08	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1392,8053	p<0,0001	
	LR	919,2068	p<0,0001	
	rehtorit	425	7	432
		98,38	1,62	100,00
	lehtorit	5508	166	5674
		97,07	2,93	100,00
	op.-ohjaajat	142	20	162
		87,65	12,35	100,00
6	päätoim. tuntiop.	877	203	1080
		81,20	18,80	100,00
	sivutoim. tuntiop.	200	202	402
		49,75	50,25	100,00
	muu virka	39	9	48
		81,25	18,75	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1414,5238	p<0,0001	
	LR	925,1851	p<0,0001	

Taulukko 13: Pätevyyden tarkastelua. Pätevyyden ja ikäluokkien ristiintaulukointi. Suojattu käyttäen mikroaggregointia. (Alkuperäistä aineistoa on merkitty ryhmäkoolla 1.)

k	Ikä	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
	alle 40 v.	2128	339	2467
		86,26	13,74	100,00
	40–53 v.	2666	183	2849
1		37,07	30,15	36,54
		93,58	6,42	100,00
	yli 53 v.	2397	85	2482
		96,58	3,42	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	195,0332	p<0,0001	
	LR	190,7357	p<0,0001	
	alle 40 v.	2126	342	2468
		86,14	13,86	100,00
	40–53 v.	2662	178	2840
2		93,73	6,27	100,00
	yli 53 v.	2403	87	2490
		96,51	3,49	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	χ^2	199,7616	p<0,0001	
	LR	193,6664	p<0,0001	
	alle 40 v.	2117	343	2460
		86,06	13,94	100,00
	40–53 v.	2682	178	2860
6		93,78	6,22	100,00
	yli 53 v.	2392	86	2478
		96,53	3,47	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	203,9335	p<0,0001	
	LR	197,3288	p<0,0001	
	alle 40 v.	2127	343	2470
		86,11	13,89	100,00
	40–53 v.	2673	185	2858
10		93,53	6,47	100,00
	yli 53 v.	2391	79	2470
		96,80	3,20	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	207,3504	p<0,0001	
	LR	204,5974	p<0,0001	

PRAM-menetelmän soveltaminen

PRAM-menetelmällä suojaaminen toteutettiin sekä arvojen vaihtumista rajoittamatta että rajoittamalla arvojen vaihtuminen korkeintaan kahden arvon päähän. Lisäksi tehtiin tarkempi tarkastelu aine-muuttujan suojaamiselle rajoittamalla suojaaminen aineryhmän sisälle. Vaihtuminen sallittiin esimerkiksi kielten opettajien aineiden välillä, mutta kielen opettajan opetettavan aineen ei annettu vaihtua matemaattiseksi aineeksi. Tämän kokeilun tavoitteena oli selvittää, voiko μ-Argus-ohjelman PRAM-menetelmää soveltaa ryhmiin jakautuneen muuttujan suojaamiseen. Tarkastellaan ensin aine-muuttujan suojaamista kaikilla kolmella eri tavalla ja vasta sen jälkeen tarkemmin näitä tyypillisempiä PRAM-menetelmän tapauksia, joissa suojaaminen tapahtuu koko muuttujalle kerralla.

Luokkien frekvenssit varsinkin aine-muuttujan tapauksessa pyrkivät tasoittumaan heti, kun vaihtumistodennäköisyyttä kasvatettiin. Sama ilmiö oli havaittavissa sekä rajoitettua että rajoittamatonta muutosta käytettäessä. Koska aineistossa on useita sellaisia aineita, joiden opettajia on Suomessa vain muutamia, aiheuttaa tämä jo pienillä vaihtumistodennäköisyyksillä huomattavia muutoksia suojatun aineiston perusteella tehtäviin analyyseihin.

Tasoittuminen on erityisen voimakasta silloin, kun suuren frekvenssin omaava luokka on hyvin pienen frekvenssin omaavan luokan vieressä. Tällöin monet suuren frekvenssin yksiköistä siirtyivät suojaamisessa pienemmän frekvenssin arvolle, eikä päinvastaista siirtymistä juurikaan ollut. Tämän havainnon seurauksena voidaan todeta, että muuttujan luokkien järjestyksellä on suuri merkitys PRAM-menetelmän lopputuloksessa. Mikäli luokat järjestettäisiin suuruusjärjestykseen frekvenssien mukaan, olisi tasoittuminen huomattavasti vähäisempää. Yleensä muuttujien luokat ovat kuitenkin sisällöllisesti mielekkäässä järjestyksessä, joten tasoittumisilmiö on otettava huomioon PRAM-menetelmää käytettäessä.

Reunajakaumaa tasoittava vaikutus oli häiritsevää jo vaihtumistodennäköisyyden ollessa 0,10. Suositeltavaa olisikin asettaa muuttumisen todennäköisyys korkeintaan arvoon 0,07 tai mieluummin jopa niinkin pieneen arvoon kuin 0,05 tai ottaa käyttöön rajoituksia arvojen vaihtumiselle.

Tutkimista jatkettiin rajoittamalla muuttujien arvon vaihtumista. Rajoittamisen tavoitteena oli hillitä arvojen liiallista muuttumista alkuperäisiin arvoihin verrattuna. Aine-muuttujan osalta toiveena oli hillitä myös aineiston laatua haittaavaa reunajakauman tasoittumista. Valitettavasti tavoite ei toteutunut vaikka raja asetettiin niinkin pieneksi kuin kaksi. Käytännössä muuttujanarvojen oli mahdollista muuttua korkeintaan kaksi arvoa pienemmiksi tai kaksi arvoa suuremmiksi. Ongelmaksi muodostuivat aine-muuttujan eri luokkien saamien frekvenssien suuret erot. Vaikka vaihtumista rajoitettiin, kasvoivat suurten frekvenssien vieressä olevat pienet arvot nopeasti paljon suuremmiksi kuin alkuperäisessä aineistossa. Koko aineistoa tarkasteltaessa muutokset olivat kuitenkin hillitympiä kuin ilman rajoja toteutetussa suojaamisessa.

Taulukko 14: PRAM-suojaus ilman vaihdon rajoittamista. Aine on tarkemmissa luokissa. (Alkuperäistä aineistoa on merkitty vaihtumattomuustodennäköisyydellä 1.)

	Vaihtumattomuustodennäköisyys										
Aine	1	0,97	0,95	0,93	0,90	0,85	0,80	0,75	0,70	0,65	0,60
04	7	11	15	18	15	39	35	44	78	66	77
05	777	766	740	731	716	688	660	620	608	562	542
06	66	65	82	74	74	81	89	97	105	98	123
07	1	6	4	20	15	23	34	48	47	58	79
08	3	8	7	15	26	24	39	43	62	69	65
09	2	6	6	16	16	32	30	47	46	67	75
10	1	5	10	10	15	39	35	49	68	52	73
11	18	23	28	32	31	48	53	57	67	77	74
16	672	648	646	632	604	607	579	565	513	483	476
17	67	67	70	77	74	81	84	84	106	112	116
18	2	7	7	12	21	25	29	40	50	56	77
19	881	870	855	828	831	776	732	701	670	628	580
20	354	353	338	335	333	326	295	298	303	278	290
21	203	204	197	199	194	195	196	201	181	178	190
22	56	64	61	64	74	78	78	83	86	94	107
23	53	53	61	67	65	75	76	81	95	96	89

Taulukko 15: PRAM-suojauksessa vaihto rajoitettu kahden yksikön päähän. Aine on tarkemmissa luokissa. (Alkuperäistä aineistoa on merkitty todennäköisyydellä 1.)

	Vaihtumattomuustodennäköisyys										
Aine	1	0,97	0,95	0,93	0,90	0,85	0,80	0,75	0,70	0,65	0,60
04	7	18	24	21	36	44	60	84	85	93	124
05	777	752	734	731	694	655	629	577	521	517	487
06	66	71	72	83	83	110	101	119	133	148	138
07	1	9	21	16	37	40	57	67	107	86	99
08	3	4	3	3	4	5	7	8	7	11	4
09	2	2	2	4	2	3	3	1	4	5	4
10	1	7	7	16	26	23	37	45	62	55	69
11	18	22	25	35	35	40	46	64	70	65	86
16	672	655	649	613	596	574	546	509	465	454	435
17	67	77	76	88	110	120	128	148	161	174	182
18	2	8	17	43	51	67	82	108	141	199	185
19	881	864	855	832	801	765	764	722	672	627	572
20	354	354	354	352	340	362	336	338	345	310	347
21	203	207	210	212	220	229	229	230	232	256	263
22	56	58	58	60	74	68	80	76	97	87	105
23	53	55	52	53	52	56	54	58	53	66	55

Aineiston suojaamisessa PRAM-menetelmällä oli annettu aine-muuttujan arvojen vaihtua rajatta ja rajoitettu vaihtumista kahden luokan päähän. Tämä rajoite oli valittu teknisten syiden takia ja uusi kiinnostus kohdistui nyt sisällölliseen tarkasteluun. Kun aine-muuttujan arvojen vaihtuminen rajattiin sallimalla vaihtuminen vain aineryhmien sisällä, saatiin PRAM-menetelmälle ominaista tasoittumisilmiötä hillityksi.

Valitettavasti ohjelman nykyisessä versiossa olevat suojausparametrien säätömahdollisuudet eivät ole tälle idealle kaikkein parhaimmat. Keinotekoiset rajat saatiin aikaiseksi asettamalla rajalle osuvien aineiden vaihtumistodennäköisyydet nollaan ja rajaamalla vaihtuminen korkeintaan yhden luokan päähän. Rajalla oleviin luokkiin valui muutamia yksiköitä oman ryhmän sisältä viereisestä luokasta, mutta mikään alunperin kyseiseen luokkaan kuulunut yksikkö ei saanut uutta arvoa. Tämä tieto olisi ainakin tunkeutujalle mieleinen ja helpottaisi paljastamista. Siksi suojaamiseen käytetty menetelmä pitäisi kertoa vain yleisesti. PRAM-menetelmällä suojatun aineiston yhteydessä luovutetaan yleensä myös suojaukseen käytetty PRAM-matriisi, mutta tässä tapauksessa sellainen ei tule kysymykseen.

Mikäli vaihtumistodennäköisyyttä kasvatettaisiin suuremmaksi kuin 0,07 (taulukko 16), kasvaisivat ryhmien rajalla olevien luokkien frekvenssit ylisuuriksi ja viereisen luokan frekvenssi taas pienenisi tämän seurauksena enemmän kuin muiden luokkien. PRAM-menetelmän tyypillisempien käyttötapojen yhteydessä on todettu, että vaihtumistodennäköisyyden arvojen pitäisi olla korkeintaan 0,07, joten tässä erikoistapauksessakin voidaan ohjetta pitää lähtökohtana. Taulukon 16 perusteella todennäköisyys 0,05 näyttäisi toimivan frekvenssien kohdalla parhaiten. Koska ryhmien sisälle rajoittuvan PRAM-suojauksen yhteydessä PRAM-matriisin luovuttaminen ei ole mahdollista, ovat vain pienet frekvenssien muutokset sallittavia.

Taulukon 16 frekvenssejä tarkasteltaessa on muistettava, että kahdessa ensimmäisessä suojauksessa olivat mukana myös luokat 12–15. Näissä tapauksissa luokan 11 ja 16 arvoihin vaikuttavat myös nuo ryhmittäisen tarkastelun yhteydessä poisjätetyt luokat. Tietenkin ensimmäisessä suojauksessa, kun arvojen vaihtamista ei ollut rajoitettu mitenkään, vaikuttavat suojauksen jälkeen saatuihin frekvensseihin koko aineiston arvot.

Taulukko 16: PRAM-menetelmä, kun aine-muuttujan arvojen vaihtuminen aineryhmästä toiseen on estetty. Vertailun vuoksi taulukossa on mukana myös aiemmin esiteltyjen suojausten frekvenssit. Taulukon vaakaviivoilla osoitetaan aineryhmän rajan sijainti. (Alkuperäistä aineistoa on merkitty vaihtumattomuustodennäköisyydellä 1.)

Aine	1	0,97			0,95			0,93		
		vaihtamisen rajoittaminen								
		ei	raja 2	ryhmä	ei	raja 2	ryhmä	ei	raja 2	ryhmä
05	777	766	752	781	740	734	779	731	731	780
06	66	65	71	62	82	72	64	74	83	60
07	1	6	9	2	4	21	1	20	16	4
08	3	8	4	2	7	3	3	15	3	2
09	2	6	2	2	6	2	2	16	4	2
10	1	5	7	1	10	7	1	10	16	2
11	18	23	22	18	28	25	18	32	35	18
16	672	648	655	672	646	649	676	632	613	674
17	67	67	77	67	70	76	60	77	88	61
18	2	7	8	14	7	17	23	12	43	40
19	881	870	864	855	855	855	849	828	832	827
20	354	353	354	365	338	354	364	335	352	376
21	203	204	207	204	197	210	205	199	212	197
22	56	64	58	57	61	58	56	64	60	58
23	53	53	55	53	61	52	56	67	53	53
24	15	20	15	17	27	17	12	24	14	17
25	10	14	15	9	20	29	12	18	35	8
26	17	22	34	17	30	30	17	26	35	19
27	1194	1156	1169	1199	1146	1148	1202	1112	1120	1206
28	316	306	310	310	314	315	305	303	323	295
29	151	151	157	153	152	169	155	150	179	160
30	115	118	124	113	113	129	112	120	122	111
31	408	404	401	409	409	396	410	392	386	412

Taulukosta 17 voidaan tarkastella laajempiin luokkiin jaoteltujen arvojen muuttumista sekä rajoittamattoman että arvolla kaksi rajoitetun PRAM-suojauksen seurauksena. Taulukossa 17 muutokset ovat molemmissa tapauksissa vielä hyväksyttävissä, mutta erityisesti muutoksen todennäköisyyden kasvettua arvoon 0,10 alkavat muuttujanarvot jo tasoittua. Jos muutoksen todennäköisyyden annetaan edelleen kasvaa, saadaan myös laajojen luokkien aineistossa näkymään huomattavaa arvojen tasoittumista. Parhaiten ero, jonka arvojen vaihtumisen rajoittaminen saa aikaan, nähdään suurien todennäköisyyksien osalta. Vaikka todennäköisyys arvon vaihtumiselle olisi asetettu arvoksi 0,40, ovat laajojen luokkien saamat frekvenssit vielä hyväksyttävissä. Vastaavalla vaihtumistodennäköisyydellä ilman arvojen vaihtumisen rajoittamista frekvenssit ovat jo kaukana alkuperäisistä.

Luokkien laajentaminen suojauksen jälkeen ei anna syytä aiemmin ehdotetun vaihtumistodennäköisyyden kasvattamiselle. Ensimmäinen ehdotus saatiin aine-muuttujan tarkan luokittelun saamia frekvenssejä tutkimalla. Myös laajojen luokkien yhteydessä voidaan vaihtumistodennäköisyyden arvolla 0,07 saatavat frekvenssi vielä hyväksyä, mutta sitä suuremmat todennäköisyydet aiheuttavat liian suuria muutoksia frekvensseihin.

Tutkitaan vielä kaksiulotteista reunajakaumaa kuten mikroaggregoinnin yhteydessä. PRAM-menetelmällä suojattujen aineistojen avulla muodostettiin kahden muuttujan ristiintaulukointi useille muuttujille. Taulukoinnit toteutettiin sekä rajoittamattomalle että rajoitetulle PRAM-suojaukselle. Nämä taulukot on liitteessä 1. Keskitytään tässä tarkastelemaan kahden suojatun muuttujan yhteisjakauman muutoksia taulukossa 18 ja suojatun ja suojaamattoman muuttujan tilannetta taulukossa 19. Taulukko 19 on esimerkki suojatun ja suojaamattoman muuttujan tilanteesta, kun suojaus on toteutettu virka-muuttujalle rajaamalla arvojen vaihtuminen korkeintaan kahden arvon päähän.

Tarkasteltaessa kahden muuttujan yhteisjakaumaa, ei PRAM-menetelmän tasoittava vaikutus ole enää yhtä selvä kuin yksiulotteisissa jakaumissa (vertaa taulukoita 17 ja 18). Kuitenkin pienten frekvenssien voi huomata edelleenkin kasvavan suurempien kustannuksella. Kuten yksiulotteisessakin tilanteessa kaksiulotteisten frekvenssijakaumien tasoittumista pienentää arvojen vaihtumisen rajoittaminen.

Taulukosta 19 nähdään, että PRAM-menetelmällä suojatun aineiston muuttujien väliset korrelaatiot säilyvät pienillä vaihtumistodennäköisyyksillä hyvin. Kuitenkin suojatun aineiston kaksiulotteisen jakauman frekvenssit näyttävät eroavan alkuperäisistä jo vaihtumistodennäköisyyden arvolla 0,05. Yksittäisessä tapauksessa voisi kyse olla sattumastakin, mutta frekvenssien erot ovat huomattavat jokaisella eri vaihtumistodennäköisyydellä ja vaikka kyseessä on rajoitettu PRAM-suojaus.

Ikä-muuttujan ja pätevyystiedon ristiintaulukoinnissa PRAM-menetelmä suoriutuu paremmin kuin virkaluokan tilanteesta kuten liitteen 1 taulukoista käy ilmi. Tulos on tietenkin parempi, jos muuttujanarvojen vaihtumista rajataan kuin rajoittamattoman suojauksen yhteydessä. Tämän paremman tuloksen syynä voi olla hyvin laajoihin luokkiin luokiteltu muuttuja. Jos tarkastellaan yksiulotteisia ikä-muuttujan reunajakaumia esimerkiksi taulukosta 17, on suuria eroja nähtävissä jo vaihtumistodennäköisyyttä 0,10 käytettäessä. Kaksiulotteisen jakauman arvojen pysyvyys johtuu muutosten tasaisesta jakautumisesta kaikille arvoille, jolloin yksittäisen frekvenssin muutos pysyy pienenä.

Taulukko 17: Tuloksia: PRAM-menetelmällä suojatun aineiston frekvenssejä. (Alkuperäistä aineistoa on merkitty vaihtumattomuustodennäköisyydellä 1.)

Tn	Ikä			Aine1		
		Ilman rajaa	Rajana 2		Ilman rajaa	Rajana 2
	alle 40	2467		äidinkieli	868	
	40–53	2849		kielet	2330	
	yli 53	2482		matemaattiset	2184	
1				historia yms.	1335	
				taide yms.	748	
				muut	333	
	alle 40	2505	2469	äidinkieli	879	867
	40–53	2817	2853	kielet	2322	2342
	yli 53	2476	2476	matemaattiset	2135	2161
0,97				historia yms.	1349	1334
				taide yms.	766	750
				muut	347	344
	alle 40	2516	2472	äidinkieli	898	888
	40–53	2803	2846	kielet	2282	2337
	yli 53	2479	2480	matemaattiset	2077	2130
0,93				historia yms.	1353	1349
				taide yms.	812	748
				muut	376	346
	alle 40	2537	2461	äidinkieli	893	881
	40–53	2798	2857	kielet	2278	2349
	yli 53	2463	2480	matemaattiset	2051	2130
0,90				historia yms.	1406	1334
				taide yms.	780	739
				muut	390	365
	alle 40	2544	2467	äidinkieli	935	876
	40–53	2728	2852	kielet	2272	2374
	yli 53	2526	2479	matemaattiset	1988	2092
0,85				historia yms.	1388	1324
				taide yms.	800	741
				muut	415	391
	alle 40	2685	2465	äidinkieli	1031	887
	40–53	2663	2843	kielet	2149	2476
	yli 53	2450	2490	matemaattiset	1657	1891
0,60				historia yms.	1443	1309
				taide yms.	952	743
				muut	566	492

Taulukko 18: PRAM ilman rajaa: kahden muuttujan ristiintaulukko. Ensimmäiset luvut ovat rajoittamattomasta ja jälkimmäiset kahden arvon päähän rajoitetusta PRAM-suojauksesta. (Alkuperäistä aineistoa on merkitty vaihtumattomuustodennäköisyydellä 1.)

		rajoittamaton			rajoitettu		
Tn	Aine	Ikä			Ikä		
		alle 40v.	41–53v.	yli 53v.	alle 40v.	41–53v.	yli 53v.
	äidinkieli	306	316	246			
	kielet	697	853	780			
1	matemaattiset	654	769	761			
	historia yms.	429	494	412			
	taide yms.	278	282	188			
	muut	103	135	95			
	äidinkieli	312	313	254	304	318	245
	kielet	699	849	774	703	862	777
0,97	matemaattiset	656	733	746	646	763	752
	historia yms.	439	501	409	430	491	413
	taide yms.	286	285	195	277	282	191
	muut	113	136	98	109	137	98
	äidinkieli	291	316	270	305	316	243
	kielet	706	837	796	700	861	786
0,95	matemaattiset	661	728	745	646	758	753
	historia yms.	450	485	410	425	495	409
	taide yms.	289	289	198	276	285	185
	muut	110	148	96	112	140	103
	äidinkieli	319	322	257	314	322	252
	kielet	717	817	748	698	853	786
0,93	matemaattiset	644	704	729	644	746	740
	historia yms.	425	515	413	425	505	419
	taide yms.	302	295	215	277	284	187
	muut	109	150	117	114	136	96
	äidinkieli	322	308	263	305	321	255
	kielet	702	835	741	699	864	786
0,90	matemaattiset	633	720	698	635	745	750
	historia yms.	465	503	438	431	497	406
	taide yms.	281	284	215	279	276	184
	muut	134	148	108	112	154	99

Taulukko 19: Pätevyyden tarkastelua. Pätevyyden ja virkaluokkien ristiintaulukointi. Suojattu PRAM-menetelmällä, jossa arvojen vaihtuminen on rajoitettu korkeintaan kahden päähän. (Alkuperäistä aineistoa on merkitty todennäköisyydellä 1.)

Tn	Virka	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
	rehtorit	425	7	432
		98,38	1,62	100,00
	lehtorit	5519	165	5684
		97,10	2,90	100,00
	op.-ohjaajat	125	17	142
		88,03	11,97	100,00
1	päätoim. tuntiop.	875	203	1078
		81,17	18,83	100,00
	sivutoim. tuntiop.	208	206	414
		50,24	49,76	100,00
	muu virka	39	9	48
		81,25	18,75	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1422,4648	p<0,0001	
	LR	932,8533	p<0,0001	
	rehtorit	427	7	434
		98,39	1,61	100,00
	lehtorit	5420	166	5586
		97,03	2,97	100,00
	op.-ohjaajat	187	19	206
		90,78	9,22	100,00
0,97	päätoim. tuntiop.	910	200	1110
		81,88	18,02	100,00
	sivutoim. tuntiop.	198	201	399
		49,62	50,38	100,00
	muu virka	49	14	63
		77,78	22,22	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1392,4511	p<0,0001	
	LR	905,2113	p<0,0001	
	rehtorit	427	9	434
		98,39	1,61	100,00
	lehtorit	5346	166	5512
		96,99	3,01	100,00
	op.-ohjaajat	216	19	235
		91,91	8,09	100,00
0,95	päätoim. tuntiop.	946	198	1144
		82,69	17,31	100,00
	sivutoim. tuntiop.	202	198	400
		50,50	49,50	100,00
	muu virka	54	19	73
		73,97	26,03	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1346,0828	p<0,0001	
	LR	879,0836	p<0,0001	

Menetelmien eroista

Parhaiten kuvan PRAM-menetelmän ja mikroaggregoinnin eroista frekvenssiaineiston muutoksille saa taulukosta 20. PRAM-menetelmässä pienetkin frekvenssit alkavat kasvaa huomattavasti jo pian vaihtumistodennäköisyyden kasvaessa. Mikroaggregoinnissa vastaavaa kasvua ei huomata. Tällä käyttäjän kannalta eduksi tulkittavalla ominaisuudella on kääntöpuolensa suojaavuuden tasoa ajatellessa. Mikroaggregoidusta aineistosta voidaan vielä suurenkin ryhmäkoon aggregoinnin jälkeen päätellä melko helposti tunnettuja yksiköitä. Jos tunkeutuja tietää tai epäilee esimerkiksi jonkun pienen arvon omaavan yksikön olevan aineistossa, pystyy hän mikroaggregoidusta aineistosta tunnistamaan tämän melkein varmasti. Vaikkei tunnettu arvo olisikaan täsmälleen ennallaan, ei kokoluokka ole muuttunut suojauksen aikana.

PRAM-menetelmän aiheuttamat muutokset luokkien frekvensseihin kasvavat nopeasti vaihtumistodennäköisyyden kasvaessa. Aineisto ei siis sellaisenaan ole käytettävissä tutkimukseen. PRAM-menetelmällä suojatun aineiston kanssa luovutetaan yleensä suojauksessa käytetty PRAM-matriisi, jonka avulla näitä aineistojen välisiä suuriakin eroja voidaan korjata. PRAM-menetelmän käytössä ei siksi kannata asettaa tavoitteeksi frekvenssien tarkkaa säilymistä. Kun PRAM-matriisia hyödynnetään, vaihtuneita arvoja tulisi olla riittävästi aiheuttamassa epävarmuutta tunnistamiseen. Siirtymätodennäköisyyksiä voidaan käyttää arvojen muutoksen päättelemiseen ja melko varmoja tunnistuksia on helppo tehdä, jos vain muutamia muuttujia on ollut suojattavana.

Mikäli taulukkoa 20 tarkastellaan aineiston käyttäjän kannalta, haluaisi hän varmasti mieluiten käyttää mikroaggregoinnilla suojattua aineistoa. PRAM-menetelmällä on frekvenssien tasoittumisen lisäksi rasitteena oikeiden tulosten saamiseksi huomioon otettava PRAM-matriisi. Aineiston suojaajan näkökulmasta mikroaggregoinnilla saatujen frekvenssien pysyvyys taas huolestuttaa, vaikka yksiköiden arvot ovat voineet vaihtua ja vain frekvenssit pysyvät. Mikroaggregoitavan aineiston tapauksessa erityisen riskin muodostavat yksiköt, joiden muuttujanarvojen suurusluokka on sama kaikissa muuttujissa ja siten yksikön voi tunnistaa siitä, että kaikki sen suojatut muuttujanarvot ovat liki ennallaan. Sellaiset yksiköt, joilla on sekä pieniä, että suuria arvoja joutuvat suuremmalla todennäköisyydellä osaksi ryhmään, jonka muuttujanarvojen keskiarvot eroavat enemmän alkuperäisistä.

Kummallakin menetelmällä on siis etunsa ja heikkoutensa. Oikeastaan kahta näin eri tyyppistä menetelmää ei kannattaisi mennä vertailemaan ollenkaan. Näin kuitenkin tehtäessä voisi sanoa, että PRAM-menetelmä hankalammin seurattavine muutoksineen on aineiston suojaavuudelta parempi ja mikroaggregointi taas voittaa suojatun aineiston käytettävyydellä.

Taulukko 20: Aste-muuttujan muunnoksien tarkastelu. (Alkuperäistä aineistoa on merkitty vaihtumattomuustodennäköisyydellä ja ryhmäkoolla 1.)

PRAM ilman rajaa											
Aste	1	0,97	0,95	0,93	0,90	0,85	0,80	0,75	0,70	0,65	0,60
01	1	22	30	59	74	94	142	164	200	246	301
02	10	25	45	61	79	92	139	194	221	241	265
03	343	357	361	359	395	382	430	437	459	455	505
04	7	28	37	57	61	104	140	165	206	272	319
05	6672	6487	6347	6202	5978	5701	5350	5049	4693	4418	4011
06	686	679	678	693	684	697	684	663	648	671	674
07	22	39	61	72	89	150	145	192	228	232	272
08	7	36	33	49	77	108	158	172	250	247	299
09	7	24	44	51	71	101	152	177	229	231	275
10	16	36	64	73	89	119	152	188	233	272	315
11	3	27	43	60	99	110	142	196	201	266	262
12	24	38	55	62	102	140	164	201	230	247	300

PRAM rajana 2											
Aste	1	0,97	0,95	0,93	0,90	0,85	0,80	0,75	0,70	0,65	0,60
01	1	2	3	5	9	18	25	26	31	28	29
02	10	15	12	14	20	23	19	26	41	35	36
03	343	376	417	431	469	547	605	651	735	789	904
04	7	50	88	150	218	287	368	483	559	660	778
05	6672	6475	6377	6182	6047	5717	5382	5114	4748	4485	4083
06	686	711	736	763	767	849	899	946	995	1031	1100
07	22	97	99	175	194	276	408	449	575	654	728
08	7	13	15	27	25	30	40	54	61	68	92
09	7	6	8	8	7	8	8	3	10	10	9
10	16	17	16	17	17	18	16	20	19	17	11
11	3	4	3	4	4	7	4	6	5	4	13
12	24	22	24	22	21	18	24	20	19	17	15

Mikroaggregointi										
Aste	1	2	3	4	5	6	7	8	9	10	11
01	1	0	0	0	0	0	0	0	0	0	0
02	10	8	9	8	10	6	7	8	9	10	11
03	343	346	345	344	345	348	350	344	351	340	341
04	7	6	6	4	5	0	0	8	0	10	11
05	6672	6672	6673	6678	6673	6676	6671	6670	6673	6668	6665
06	686	684	684	684	680	684	693	672	693	690	704
07	22	24	24	20	30	30	21	24	18	30	11
08	7	8	9	12	5	6	7	8	9	0	0
09	7	4	6	4	10	6	7	8	0	10	11
10	16	18	15	16	15	12	14	16	18	10	11
11	3	4	3	4	0	6	0	8	9	10	11
12	24	24	24	24	25	24	28	24	18	20	22

8.5 Menetelmien soveltaminen logistisen mallin yhteydessä

Aineistosta muodostettiin logistinen malli, jossa opettajan pätevyyttä selitettiin usean muuttujan avulla. Logistinen malli on muotoa

$$logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n,$$

missä β_0 on mallin vakiotermi, $\beta_1, \ldots, \beta_n$ ovat estimoitavat parametrit ja $x_1, \ldots, x_n$ ovat mallin selittävät muuttujat. Mallin parametrit estimoitiin suurimman uskottavuuden menetelmällä käyttäen SAS-proseduuria logistic.

Mallin dikotomisena vasteena on muuttuja p, joka kertoo, onko opettajalla puuttuvia pätevyyksiä hoitamaansa tehtävään. Mallinnuksessa mallinnetaan tilannetta, jossa p saa arvon 0 eli puuttuvia pätevyyksiä ei ole. Malliin käytetyt selittävät muuttujat olivat: ikä vuosina, sukupuoli (0=mies, 1=nainen) ja opettajan virkaluokka. Virkaluokat ovat seuraavat:

- 01=rehtori
- 02=(esi-)luokanopettaja
- 03=erityisopettaja
- 04=lehtori
- 05=oppilaanohjaaja/opinto-ohjaaja
- 06=päätoiminen tuntiopettaja
- 07=sivutoiminen tuntiopettaja
- 08=muu virka.

Mallin vertailuluokaksi valittiin sivutoimiset tuntiopettajat (virkaluokka 7), joiden pätevyydessä oli eniten puutteita.

Alkuperäiselle aineistolle saatiin taulukon 21 mukaiset parametrien arvot. Tästä taulukosta näkyy myös se, ettei lukion opettajiin rajoitetussa aineistossa ole lainkaan luokanopettajia tai erityisopettajia eli virkaluokkia 2 ja 3.

Taulukko 21: Mallinnetaan viran vaikutusta pätevyyden arvoon. Alkuperäisen aineiston perusteella saadut estimaatit.

Muuttuja	Estimaatit	s.e.	95% lv.		t-arvo	p-arvo
vakio	0,0348	0,2419	-0,4388	0,5102	0,0207	0,8857
ikä	0,0419	0,0419	0,0324	0,0515	74,0574	<0,0001
sukupuoli	0,3469	0,0982	0,1537	0,5388	12,4816	0,0004
virkaluokka 1	1,8713	0,3303	1,2838	2,5985	32,0911	<0,0001
virkaluokka 4	1,3539	0,1212	1,1091	1,5860	124,8486	<0,0001
virkaluokka 5	-0,0963	0,2365	-0,5420	0,3912	0,1658	0,6839
virkaluokka 6	-0,4467	0,1235	-0,6959	-0,2102	13,0916	0,0003
virkaluokka 8	-0,7627	0,3235	-1,3629	-0,0801	5,5573	0,0184

Alkuperäiseen aineistoon sovitetusta mallista nähdään, että opettajan pätevyyteen vaikuttaa opettajan ikä, sukupuoli ja myös yhtä lukuun ottamatta kaikkien virkaluokkien parametrin estimaatit ovat tilastollisesti merkitseviä. Kuten olettaa saattaa, on rehtorin tai lehtorin virkaluokkaan kuuluminen pätevyyttä parantavia tekijöitä. Vastaavasti virkaluokat päätoiminen tuntiopettaja ja muu virka heikentävät pätevyyttä. Oppilaan tai opinto-ohjaajan virka ei vaikuta pätevyyteen, koska parametrin estimaatti ei ole tilastollisesti merkitsevä. Sukupuolella ja iällä on molemmilla pieni vaikutus pätevyyden arvoon.

Mallia muodostettaessa koettiin mielenkiintoisimmaksi mallintaa pätevyyttä käyttäen selittävänä muuttujana opettajan opettamaa ainetta. Aine-muuttujassa on kuitenkin luokkia, joissa on vain muutama opettaja ja se aiheuttaa ongelmia mallintamisessa. Opettajan virkatyypin käyttäminen selittävänä muuttujana oli toinen sisällöllisesti mielenkiintoinen vaihtoehto, joten se tuli valituksi. Kolmantena mahdollisuutena ollut vaihtoehto: opettajan opetuksen kouluasteen käyttö, ei ollut sisällöllisesti mielenkiintoinen, koska aineisto on lukion opettajista koostuva. Suurin osa opettajista lukioasteen opettajia, eikä aste-muuttuja ole tällöin mielekäs selittävä muuttuja tutkittavaan malliin.

Mikroaggregoinnin soveltaminen

Mikroaggregoinnilla suojaaminen toteutettiin jatkuville muuttujille tarkoitetun mikroaggregoinnin avulla. Ensimmäisen suojauksen yhteydessä, kun luokkien nimeämiseen ei puututtu, saatiin parametrin estimaatteja myös lukioaineistosta puuttuville virkaluokille. Tämä oli tietenkin seurausta siitä, että suojauksessa oli tullut nollasta poikkeavia frekvenssejä sellaisiinkin virka-muuttujan luokkiin, joissa ei alunperin ollut yhtään opettajaa. Mallin sovittaminen ei tällaiseen aineistoon onnistunut kunnolla ja varsinkin pienimmillä suojauksen ryhmäkoilla saatiin erikoisia tuloksia (liitteen 1 taulukot 4, 5 ja 6). Monien estimaattien t-arvot olivat nollia, eikä malli ollut tulkittavissa.

Ongelmasta ei päästy eroon vaihtamalla tarkasteltavaa muuttujaa, vaan sama ongelma olisi kuitenkin tullut vastaan aine- tai aste-muuttujaa käytettäessä. Tästä syystä päädyttiin jo edellä esiteltyyn virka-muuttujan luokkien uudelleen nimeämiseen. Frekvenssiltään nollasta eroavat luokat nimettiin juoksevalla numeroinnilla yhdestä eteenpäin ja analysoinnit suoritettiin uusien luokkien avulla. Tulosten tarkasteluun uudet luokkien nimet kuitenkin palautettiin takaisin alkuperäisiksi, jotta tulosten vertailu PRAM-menetelmän kanssa olisi helpompaa.

Luokkien uudelleen nimeämisen jälkeen suojatusta aineistosta lasketuille mallin estimaateille saatiin mielekkäämpiä tuloksia. Nyt alkuperäisen aineiston kanssa samojen parametrien estimaatit olivat tilastollisesti merkitseviä. Parametrien estimaatit säilyvät mikroaggregoidun aineiston mallinnuksessa liki samoina kuin alkuperäisen aineiston. Arvojen säilyminen myös suurimmilla tutkituilla ryhmäkoilla oli yllättävää. Keskihajonnan arvot pysyvät myös alkuperäisen aineiston kanssa saman suuruisina. Ainoat erot saatiin eri estimaattien t-arvoille, mutta niissäkin erot olivat niin pieniä, että ainoastaan tilastollisen merkitsevyyden taso vaihtui muutamien estimaattien kohdalla. Millekään t-arvolle muutos ei ollut niin suuri, että sellaisesta estimaatista, joka ei alunperin ollut tilastollisesti merkitsevä, olisi suojauksessa saatu tilastollisesti merkitsevä. Myöskään päinvastainen muutosta ei tullut ilmi.

Mikroaggregointia voidaan käyttää aineistolle, jonka perusteella aiotaan tehdä logistinen malli. Kokeilun perusteella suojatun aineiston käyttö ei aiheuta erityisiä eroja estimoinnin tuloksiin ja siten saatua mallia voidaan soveltaa vastaavasti kuin alkuperäisen aineiston perusteella muodostettua.

Taulukko 22: Mallin käyttäytyminen aineistoa suojatessa. Tutkitaan viran vaikutusta pätevyyden arvoon. Suojattu käyttäen mikroaggregointia. (Alkuperäistä aineistoa on merkitty ryhmäkoolla 1.)

Koko	Muuttuja	Estimaatit	s.e.	95% lv.		t-arvo	p-arvo
	vakio	0,0348	0,2419	-0,4388	0,5102	0,0207	0,8857
	ikä	0,0419	0,0419	0,0324	0,0515	74,0574	<0,0001
	sukupuoli	0,3469	0,0982	0,1537	0,5388	12,4816	0,0004
1	virkaluokka 1	1,8713	0,3303	1,2838	2,5985	32,0911	<0,0001
	virkaluokka 4	1,3539	0,1212	1,1091	1,5860	124,8486	<0,0001
	virkaluokka 5	-0,0963	0,2365	-0,5420	0,3912	0,1658	0,6839
	virkaluokka 6	-0,4467	0,1235	-0,6959	-0,2102	13,0916	0,0003
	virkaluokka 8	-0,7627	0,3235	-1,3629	-0,0801	5,5573	0,0184
	vakio	-0,0627	0,2410	-0,5347	0,4104	0,0678	0,7946
	ikä	0,0427	0,0049	0,0333	0,0523	77,6357	<0,0001
	sukupuoli	0,3498	0,0980	0,1570	0,5413	12,1135	0,0004
2	virkaluokka 1	1,9122	0,3289	1,3283	2,6372	33,0001	<0,0001
	virkaluokka 4	1,4054	0,1173	1,1689	1,6307	143,0000	<0,0001
	virkaluokka 5	-0,1310	0,2301	-0,5650	0,3425	0,0017	0,5692
	virkaluokka 6	-0,3704	0,1203	-0,6125	-0,1391	9,0023	0,0021
	virkaluokka 8	-0,9806	0,2908	-1,5285	-0,3787	11,0035	0,0007
	vakio	0,0396	0,2412	-0,4324	0,5136	0,0270	0,8694
	ikä	0,0417	0,0049	0,0322	0,0513	73,0562	<0,0001
	sukupuoli	0,3441	0,0981	0,1510	0,5359	12,2973	0,0005
6	virkaluokka 1	1,8623	0,3302	1,2752	2,5893	31,8109	<0,0001
	virkaluokka 4	1,3506	0,1200	1,1082	1,5804	126,7487	<0,0001
	virkaluokka 5	-0,1754	0,2214	-0,5952	0,2777	0,6272	0,4284
	virkaluokka 6	-0,4465	0,1219	-0,6927	-0,2129	13,4083	0,0003
	virkaluokka 8	-0,6556	0,3233	-1,2551	0,0266	4,1131	0,0426
	vakio	0,0913	0,2391	-0,3768	0,5610	0,1459	0,7025
	ikä	0,0416	0,0049	0,0321	0,0512	72,7097	<0,0001
	sukupuoli	0,3561	0,0976	0,1641	0,5468	13,3175	0,0003
10	virkaluokka 1	1,7914	0,3277	1,2105	2,5145	29,8861	<0,0001
	virkaluokka 4	1,2748	0,1133	1,0462	1,4921	126,6574	<0,0001
	virkaluokka 5	0,0498	0,2498	-0,4146	0,5722	0,0397	0,8421
	virkaluokka 6	-0,5313	0,1159	-0,7649	-0,3089	21,0100	<0,0001
	virkaluokka 8	-0,6154	0,2278	-1,0468	-0,1489	7,3018	0,0069

PRAM-menetelmän soveltaminen

PRAM-menetelmän avulla suojattujen aineistojen mallin estimaatit sekä estimaattien t-arvot alkoivat muuttua suojaustasoa kasvatettaessa. Menetelmää sovellettiin sekä vaihdettavia arvoja rajoittamatta että rajoittamalla vaihtaminen korkeintaan kahden arvon päähän. Saaduissa tuloksissa on huomattavia eroja. Tämä havainto tukee käsitystä, joka saatiin jo frekvenssitaulukoita tarkastellessa. Tuloksista voidaan nähdä, että arvojen vaihtamisen todennäköisyyden on oltava pieni tai vaihtamista on rajoitettava tai on tehtävä näistä molemmat, mikäli tavoitteena on saada aineiston perusteella samoja johtopäätöksiä kuin alkuperäisen aineiston perusteella saataisiin.

Tarkastellaan suojaamista PRAM-menetelmällä ilman vaihtamisen rajoittamista taulukosta 23. Aivan pienillä arvojen vaihtamisen todennäköisyyksillä estimaattien arvot pysyvät alkuperäisen aineiston 95%-luottamusvälillä. Kuitenkin todennäköisyyden kasvettua arvoon 0,10 huomataan jo johtopäätöksiin vaikuttavia muutoksia. Sama tulos saatiin frekvenssitarkasteluissa. Taulukosta 23 nähdään todennäköisyyden ollessa 0,10 vakiotermin arvon muuttuvan tilastollisesti nollasta eroavaksi ja sen saama arvo on liki 0,5. Tällainen ero estimaattien välillä ei mahdu edes 95%-luottamusvälille. Alkuperäisen aineiston perusteella kuitenkin saatiin vakiotermin arvoksi nolla. Vastaavasti muutoksia näkyy myös estimaattien arvoissa vaikkei missään muussa näin suurena.

Mallin parametrien estimoimista tarkasteltiin seuraavaksi aineistosta, joka oli suojattu käyttämällä PRAM-menetelmää rajaamalla arvojen vaihtuminen korkeintaan kahden arvon päähän. Tämän suojauksen tavoitteena oli saada arvot pysymään lähempänä alkuperäistä ja siten saada alkuperäisen aineiston kanssa yhtenevämpiä tuloksia. Kuten frekvenssitarkastelujen yhteydessä havaittiin, frekvenssien muuttuminen oli hillitympää, mutta varsinkin suurien frekvenssien vieressä olleet frekvenssit kasvoivat rajusti. Tästä frekvenssien muutoksesta seuraa muutoksia myös mallin estimaatteihin.

Kuten rajoittamattomassakin PRAM-suojauksessa, myös rajoitetussa estimaattien arvot pysyvät suunnilleen alkuperäisen aineiston perusteella saadun 95% luottamusvälin sisällä kunnes vaihtamisen todennäköisyys kasvaa arvoon 0,10. Tällä todennäköisyydellä lehtorien virkaluokkien väliin jäävät opinto-ohjaajien frekvenssit kasvavat jo niin paljon, että aggregoidun virkaluokan 6 estimaatti muuttuu tilastollisesti nollasta eroavaksi.

PRAM-menetelmällä suojatun aineiston käyttö ilman PRAM-matriisin huomioon ottamista ei ole erityisen suotavaa, mutta tarkastellaan silti mallin estimoinnista saatuja tuloksia. Kun PRAM-matriisia ei ole otettu huomioon, pysyvät mallin parametrien estimaatit hyväksyttävinä molemmissa suojaustavoissa aina vaihtumistodennäköisyyden arvoon 0,07 asti. Tämän jälkeen estimaattien tilastollisessa merkitsevyydessä on eroja verrattuna alkuperäisen aineiston tuloksiin. PRAM-menetelmää voisi siis ajatella käytettävän logistisen mallin muodostamista varten halutun aineiston suojaamiseen pienillä vaihtumistodennäköisyyksillä. PRAM-menetelmän käyttämisestä aiheutuva epävarmuus tunnistamisessa lisää suojaustasoa, vaikka todellinen ero alkuperäiseen aineistoon olisi pienempikin. Mieluiten tällaisessä tilanteessa tulisi kuitenkin käyttää rajoitettua arvojen vaihtumista. Taulukoiden 23 ja 24 arvoissa on sen verran eroa, joka puoltaa rajoitetun vaihtumisen käyttämistä.

Taulukko 23: Mallin käyttäytyminen aineistoa suojatessa. Tutkitaan viran vaikutusta pätevyyden arvoon. Suojattu käyttäen PRAM-menetelmää ilman rajaa. (Alkuperäistä aineistoa on merkitty vaihtumattomuustodennäköisyydellä 1.)

Tn	Muuttuja	Estimaatit	s.e.	95% lv.		t-arvo	p-arvo
	vakio	0,0348	0,2419	-0,4388	0,5102	0,0207	0,8857
	ikä	0,0419	0,0419	0,0324	0,0515	74,0574	<0,0001
	sukupuoli	0,3469	0,0982	0,1537	0,5388	12,4816	0,0004
1	virkaluokka 1	1,8713	0,3303	1,2838	2,5985	32,0911	<0,0001
	virkaluokka 4	1,3539	0,1212	1,1091	1,5860	124,8486	<0,0001
	virkaluokka 5	-0,0963	0,2365	-0,5420	0,3912	0,1658	0,6839
	virkaluokka 6	-0,4467	0,1235	-0,6959	-0,2102	13,0916	0,0003
	virkaluokka 8	-0,7627	0,3235	-1,3629	-0,0801	5,5573	0,0184
	vakio	-0,0543	0,2254	-0,4960	0,3879	0,0581	0,8095
	ikä	0,0440	0,0047	0,0348	0,0534	86,6781	<0,0001
	sukupuoli	0,3515	0,0974	0,1599	0,5418	13,0244	0,0003
0,97	virkaluokka 1	1,3537	0,2391	0,9143	1,8587	32,0547	<0,0001
	virkaluokka 4	1,3401	0,1054	1,1308	1,5447	161,6695	<0,0001
	virkaluokka 5	-0,0876	0,2135	-0,4889	0,3524	0,1685	0,6815
	virkaluokka 6	-0,4184	0,1069	-0,6309	-0,2111	15,3113	<0,0001
	virkaluokka 8	-0,3211	0,2608	-0,8050	0,2255	1,5158	0,2183
	vakio	0,2895	0,2237	-0,1482	0,7291	1,6744	0,1957
	ikä	0,0397	0,0046	0,0308	0,0488	74,7705	<0,0001
	sukupuoli	0,3248	0,0967	0,1344	0,5136	11,2802	0,0008
0,95	virkaluokka 1	1,7134	0,2898	1,1934	2,3430	34,9501	<0,0001
	virkaluokka 4	1,1876	0,1092	0,9692	1,3984	118,3724	<0,0001
	virkaluokka 5	-0,0866	0,2133	-0,4878	0,3527	0,1649	0,6847
	virkaluokka 6	-0,6107	0,1099	-0,8309	-0,3988	30,8591	<0,0001
	virkaluokka 8	-0,2705	0,2502	-0,7353	0,2527	1,1692	0,2796
	vakio	0,1940	0,2197	-0,2362	0,6253	0,7794	0,3773
	ikä	0,0407	0,0046	0,0317	0,0499	77,6875	<0,0001
	sukupuoli	0,3359	0,0966	0,1458	0,5248	12,0843	0,0005
0,93	virkaluokka 1	1,4120	0,2360	0,9798	1,9120	35,7960	<0,0001
	virkaluokka 4	1,2369	0,1016	1,0358	1,4346	148,2935	<0,0001
	virkaluokka 5	-0,1487	0,1947	-0,5158	0,2506	0,5833	0,4450
	virkaluokka 6	-0,5336	0,1031	-0,7380	-0,3334	26,8127	<0,0001
	virkaluokka 8	-0,0463	0,2309	-0,4753	0,4359	0,0401	0,8412
	vakio	0,4952	0,2110	0,0826	0,9098	5,5110	0,0189
	ikä	0,0352	0,0044	0,0267	0,0439	64,2781	<0,0001
	sukupuoli	0,2880	0,0957	0,0997	0,4750	9,0564	0,0026
0,90	virkaluokka 1	1,0765	0,1938	0,7145	1,4782	30,8410	<0,0001
	virkaluokka 4	1,1861	0,1001	0,9882	1,3811	140,5286	<0,0001
	virkaluokka 5	0,0601	0,1952	-0,3066	0,4623	0,0949	0,7580
	virkaluokka 6	-0,5834	0,1007	-0,7830	-0,3877	33,5764	<0,0001
	virkaluokka 8	0,1863	0,2489	-0,2704	0,7133	0,5599	0,4543

Taulukko 24: Mallin käyttäytyminen aineistoa suojatessa. Tutkitaan viran vaikutusta pätevyyden arvoon. Suojattu käyttäen PRAM-menetelmää ja rajana arvoa 2. (Alkuperäistä aineistoa on merkitty vaihtumattomuustodennäköisyydellä 1.)

Tn	Muuttuja	Estimaatit	s.e.	95% lv.		t-arvo	p-arvo
	vakio	0,0348	0,2419	-0,4388	0,5102	0,0207	0,8857
	ikä	0,0419	0,0419	0,0324	0,0515	74,0574	<0,0001
	sukupuoli	0,3469	0,0982	0,1537	0,5388	12,4816	0,0004
1	virkaluokka 1	1,8713	0,3303	1,2838	2,5985	32,0911	<0,0001
	virkaluokka 4	1,3539	0,1212	1,1091	1,5860	124,8486	<0,0001
	virkaluokka 5	-0,0963	0,2365	-0,5420	0,3912	0,1658	0,6839
	virkaluokka 6	-0,4467	0,1235	-0,6959	-0,2102	13,0916	0,0003
	virkaluokka 8	-0,7627	0,3235	-1,3629	-0,0801	5,5573	0,0184
	vakio	-0,0198	0,2363	-0,4826	0,4443	0,0070	0,9333
	ikä	0,0438	0,0049	0,0344	0,0534	81,8421	<0,0001
	sukupuoli	0,3399	0,0980	0,1469	0,5314	12,0179	0,0005
0,97	virkaluokka 1	1,8378	0,3281	1,2559	2,5616	31,3734	<0,0001
	virkaluokka 4	1,3054	0,1148	1,0741	1,5258	129,3657	<0,0001
	virkaluokka 5	0,1557	0,2199	-0,2583	0,6088	0,5015	0,4788
	virkaluokka 6	-0,4155	0,1170	-0,6511	-0,1907	12,6086	0,0004
	virkaluokka 8	-0,9129	0,2699	-1,4227	-0,3571	11,4440	0,0007
	vakio	-0,0456	0,2333	-0,5025	0,4124	0,0381	0,8451
	ikä	0,0446	0,0048	0,0352	0,0540	86,1277	<0,0001
	sukupuoli	0,3298	0,0976	0,1376	0,5206	11,4100	0,0007
0,95	virkaluokka 1	1,8314	0,3270	1,2523	2,5536	31,3723	<0,0001
	virkaluokka 4	1,2920	0,1121	1,0662	1,5076	132,7720	<0,0001
	virkaluokka 5	0,2833	0,2173	-0,1249	0,7319	1,7000	0,1923
	virkaluokka 6	-0,3771	0,1142	-0,6070	-0,1574	10,8931	0,0010
	virkaluokka 8	-1,1036	0,2409	-1,5641	-0,6145	20,9789	<0,0001
	vakio	0,0406	0,2340	-0,4175	0,5002	0,0301	0,8622
	ikä	0,0441	0,0048	0,0347	0,0536	83,7868	<0,0001
	sukupuoli	0,3256	0,0977	0,1332	0,5166	11,0988	0,0009
0,93	virkaluokka 1	1,7672	0,3268	1,1885	2,4890	29,2403	<0,0001
	virkaluokka 4	1,2523	0,1119	1,0268	1,4676	125,1313	<0,0001
	virkaluokka 5	0,2896	0,2118	-0,1089	0,7261	1,8699	0,1715
	virkaluokka 6	-0,4714	0,1129	-0,6990	-0,2546	17,4252	<0,0001
	virkaluokka 8	-0,8344	0,2371	-1,2839	-0,3490	12,3835	0,0004
	vakio	-0,0144	0,2292	-0,4636	0,4351	0,0040	0,9499
	ikä	0,0461	0,0048	0,0369	0,0556	94,0638	<0,0001
	sukupuoli	0,3438	0,0971	0,1529	0,5335	12,5487	0,0004
0,90	virkaluokka 1	1,5749	0,3073	1,0268	2,2475	26,2743	<0,0001
	virkaluokka 4	1,1317	0,1089	0,9132	1,3416	107,9990	<0,0001
	virkaluokka 5	0,7570	0,2167	0,3527	1,2077	12,1984	0,0005
	virkaluokka 6	-0,4689	0,1099	-0,6896	-0,2572	18,1896	<0,0001
	virkaluokka 8	-0,9525	0,2294	-1,3894	-0,4854	17,2359	<0,0001

9 Johtopäätökset

Tässä luvussa tarkastellaan eri tietosuojamenetelmien käytettävyyttä sekä tieteellisestä näkökulmasta että empiirisen sovelluksen kautta saaduista kokemuksista. Tarkoituksena on antaa tutkimuksen aikana saatujen käsitysten perusteella ehdotuksia menetelmien valitaan, kun mikroaineistoa valmistellaan luovuttamista varten.

Ensin käsitellään kirjallisuudesta nousseita näkemyksiä. Usein menetelmillä on omat kannattajaryhmänsä ja tämän seurauksena tietyt artikkelit voivat kehua menetelmää kovastikin, kun taas toisissa menetelmä arvioidaan kehnoksi. Menetelmistä on kuitenkin tullut tutkimuksen aikana tietty käsitys, jonka perusteella arvioita esitetään.

Tutkimisen empiirisen osan perusteella voidaan arvioida ainoastaan kahden menetelmän: mikroaggregoinnin ja PRAM-menetelmän toimivuutta. Näitä menetelmiä tarkastellaan eri tyyppisten muuttujien suojaamista ajatellen.

9.1 Suojausmenetelmien teoreettinen perusta

Eri tietosuojamenetelmillä on omat vahvuutensa ja heikkoutensa. Menetelmien toimivuutta voidaan mitata joko aineiston suojaajan tai aineiston käyttäjän näkökulmasta ja yleensä nämä kaksi näkökulmaa ovat yleensä vastakkaisia. Kuten jo aiemmin on kerrottu, ei mitään menetelmää voi pitää kaikkia muita parempana, koska jokaiselle menetelmälle on olemassa aineisto, jonka suojauksesta se selviää muita paremmin.

Näkökulmia kohinan käytöstä suojaamisessa

Korreloitunutta kohinaa lisäävät menetelmät säilyttävät korrelaatiot vähintään likimain. Sullivanin menetelmän etuna on, että siinä käytetään iteratiivisia menetelmiä säätämään pois pienetkin suojauksessa syntyneet erot korrelaatiomatriiseista. Muuttujakohtaiset jakaumat säilyvät ainoastaan Sullivanin menetelmää käytettäessä. Lisäksi se on ainoa näistä menetelmistä, joka pystyy suojaamaan myös luokiteltuja muuttujia. Jos lineaarisia muunnoksia kohinan lisäämisen ohella käyttävään menetelmään yhdistetään arvojen vaihtamista yksiköiden välillä, saadaan siitä käyttökelpoisempi. Lineaarisia muunnoksia käyttävä menetelmä on kohinan lisäämisen menetelmistä suositeltavin, jos aineistosta on analysoitava osajoukkojen ominaisuuksia. Menetelmiä ei kuitenkaan pystytä laittamaan paremmuusjärjestykseen. Se, mikä menetelmistä kulloinkin sopii parhaiten käytettäväksi, riippuu paljon aineiston käyttäjän analyyseistä ja niiden tavoitteista.

Kirjallisuuden perusteella korreloituneen kohinan käytöllä saadaan riittävän suojattu aineisto, jos aineistossa on yleisesti saatavilla olevia muuttujia vain muutama. Mikäli näitä helposti saatavilla olevia muuttujia on enemmän, ei korreloituneen kohinan käytöllä saavuteta riittävää suojaa. Samoin on tilanne lineaarisen muunnoksen ja valkoisen kohinan lisäämisen tapauksessa. Jos aineiston arvoja kuitenkin kohinan lisäämisen lisäksi vaihdetaan eri yksiköiden välillä, saavutetaan parempi suojauksen taso. Koska Sullivanin menetelmää ei ole käytetty kovin kauan laskentaintensiivisyytensä vuoksi, ei siitä ole paljoa käytännön kokemuksia. Brandin (2002) artikkelin perusteella menetelmä näyttäisi toimivan suhteellisen hyvin. Identifioitumistodennäköisyyksiä ei voida laskea, joten on tyydyttävä toteamaan, että algoritmi näyttäisi suojaavan suurimman osan havainnoista riittävällä tasolla.

Kohinan lisääminen ei ole yleisesti riittävän tehokas suojausmenetelmä. Vaikka suojaus paraneekin siirryttäessä pelkän kohinan lisäämisestä muunnosta sen ohessa käyt-

täviin menetelmiin, jää osa yksiköistä silti riittämättömästi suojatuiksi. Sullivanin menetelmän suojaavuus on suhteellisen tehokasta, mutta sen käyttö vaatii laajaa tuntemusta menetelmästä ja siksi se ei ole sopivin menetelmä yleiseen käyttöön. μ-Argus ohjelman käyttöohjeessa suositellaankin käyttämään Sullivanin menetelmää apuna menetelmien vertailussa ja kehittämisessä.

Näkökulmia mikroaggregoinnista

Mikroaggregointi on alunperin jatkuvien muuttujien suojaamiseen kehitetty tietosuojamenetelmä. Suojaus perustuu alkuperäisten muuttujanarvojen korvaamiseen ryhmän keskiarvolla. Menetelmä vaikeuttaa yksiköiden tunnistamista aineistosta, mutta muista yksiköistä eniten poikkeavat yksiköt on helppo tunnistaa mikroaggregoidusta aineistosta ainakin jos paljastettavan yksikön tiedetään kuuluvan suojattuun aineistoon. Toinen mikroaggregoinnin heikkous on se, että se säilyttää muuttujanarvojen kokoluokan ja tätä tietoa tunkeutuja voi hyödyntää yksiköitä paljastaessaan. Mikroaggregoinnin voidaan siis negatiivisesti ajateltaessa nähdä sotkevan aineiston arvoja ja silti jättää yksiköt paljastumisvaaraan.

Mikäli mikroaggregointia sovelletaan suosituksen mukaisesti muodostamalla ryhmät useamman muuttujan avulla eli suojaamalla useampia muuttujia kerralla, on suojatun aineiston paljastumisriski huomattavasti pienempi. Paljastumisriskin pienenemisestä huolimatta aineiston käytettävyys ei ole erityisesti heikentynyt. Mikroaggregoidun aineiston marginaalisummat ovat samat kuin alkuperäisellä aineistolla. Tarkemmat korrelaatiot usein kärsivät ainakin usean muuttujan yhtäaikaisessa suojaamisessa.

Mikroaggregointia voidaan suositella jatkuvia muuttujia sisältävän aineiston suojaamiseen, kun aineiston yksiköt ovat melko samanlaisia. Poikkeavien yksiköiden suojaaminen mikroaggregoinnilla on liian epävarmaa.

Näkökulmia PRAM-menetelmästä

PRAM-menetelmän suojaavuudesta on kaksi näkökulmaa. Menetelmän kannattajien mielestä menetelmä toimii ja suojaa aineiston riittävästi. Menetelmää kritisoivien mielestä PRAM-menetelmällä suojattu aineisto on muuttunut kelvottomaksi, eikä PRAM-matriisin luovuttaminen aineiston yhteydessä pelasta tilannetta. Lisäksi PRAM-matriisin luovuttaminen nähdään ongelmana siksi, että sen avulla voidaan päätellä suojatun aineiston muutoksia ja siten paljastaa aineistossa olevia yksiköitä.

PRAM-menetelmän suojaustaso riippuu valitusta PRAM-matriisista. Erilaisia matriisityyppejä on kokeiltu ja de Wolfin ja van Gelderin tutkimus osoittaa, että lohkomatriisien avulla on mahdollista saavuttaa parempi suojaustaso kuin yksiulotteisen matriisin avulla. Informaatiokato ei lohkomatriisia käytettäessä kasva yksiulotteisen matriisin tilanteeseen verrattuna. Jos on ennakkoon sovittu suojaustaso, voidaan lohkomatriisien avulla päästä pienempään informaatiokatoon. Näiden tutkijoiden näkemys PRAM-menetelmän kehittämisen suunnasta onkin, että PRAM-menetelmälle saadaan optimaalisempi ratkaisu, kun suojaustaso pidetään vakiona ja informaatiokatoa pyritään minimoimaan. Tämän lisäksi tutkijat uskovat informaatiokadon pienenevän mikäli käytetään PRAM-matriiseja, joiden diagonaalialkioiden annetaan erota toisistaan. Tällaisten matriisien suojaavuuden ja siitä seuraavan informaatiokadon käsitteleminen on tietenkin työläämpää.

Aineiston suojaajan kannalta PRAM-menetelmän etuna on, että siinä on yhdistynyt useita eri menetelmiä ja siten sen tuottama suojaus on monipuolinen. Suurimman

ongelman PRAM-menetelmän käytössä aiheuttaa se, ettei suojattua aineistoa pysty käyttämään ilman PRAM-matriisia. PRAM-menetelmällä suojattu aineisto onkin hankalasti käytettävä verrattuna muilla menetelmillä suojattuihin aineistoihin. Kuitenkin ilman PRAM-matriisin huomioon ottamista aineistosta saatavat tulokset ovat vääriä. Silti PRAM-menetelmä puolustaa paikkaansa menetelmänä, jossa muuttujanarvot voidaan antaa hyvin tarkalla luokituksella.

Vaikka PRAM-menetelmällä suojatun aineiston kanssa luovutetaan suojauksessa käytetty PRAM-matriisi, voi matriisin avulla korkeintaan arvailla vaihtuneita muuttujanarvoja. Tunkeutuja ei voi olla varma tunnistuksestaan vaikka joku yksikkö näyttäisi vastaavan tunkeutujan käytössä olevan toisen aineiston yksikköä tai hänen muuten tuntemaansa yksikköä. Tämä tunnistuksen epävarmuus on PRAM-menetelmän etu verrattuna moneen muuhun menetelmään.

Näkökulmia MASSC-menetelmästä

MASSC-menetelmä on ensimmäinen, joka ottaa huomioon sisäiseksi paljastumiseksi kutsutun paljastamisen ja pyrkii estämään myös tämän. MASSC-menetelmän etuna on sen monipuolisuus: se sisältää arvojen vaihtamista ja otantaa. Koska MASSC-menetelmän toiminta perustuu satunnaisuuteen, on suojaavuus vahvempaa kuin kahden edellä mainitun menetelmän yhdistäminen. Satunnaisuus aiheuttaa suojatun aineiston tutkimiseen niin paljon epävarmuutta, ettei tunnistamisesta voida olla minkään yksikön kohdalla täysin varma. Joissakin tapauksissa suojatun aineiston yksikölle kaikki identifioivien muuttujien arvot voivat olla alkuperäisiä, mutta aineiston käyttäjä tai tunkeutuja ei voi olla varma tunnistamisestaan, koska hän ei tiedä mitkä muuttujien arvoista ovat muuttuneita ja mitkä muuttumattomia. MASSC-menetelmän yhteydessä ongelmaksi voisi nostaa tilanteen, jossa tunkeutuja uskoo tunnistaneensa jonkun aineiston yksiköistä, vaikka onkin ”tunnistanut” väärän yksikön. Tällöin hän kertoo virheellistä tietoa yksiköstä ja voi näin aiheuttaa tälle haittaa.

MASSC-menetelmällä voidaan suojata sekä jatkuvia että luokiteltuja muuttujia. Menetelmän otanta-asetelma perustuu riskiositteisiin, jotka määritellään pelkkien identifioivien muuttujien perusteella. Tällöin tutkimusmuuttujina olevat jatkuvat muuttujanarvot säilyvät suojauksesta ennallaan. Yksi MASSC-menetelmän suurimmista eduista käyttäjän näkökulmasta on, että suojattua aineistoa voidaan analysoida kuten tavallista otosta. Analysoimiseen riittää painot huomioon ottava perusohjelmisto. MASSC-menetelmän identifioivien muuttujien korvausvaihe aiheuttaa luokitteluvirheitä ja siten harhaa regressiomallin parametrien estimaatteihin, jos mallinnetaan herkkien muuttujien riippuvuutta identifioivista muuttujista. Helpommin ratkaistava ongelma on korvauksen aiheuttama harha keskiarvojen ja summien keskineliövirheen estimoinnissa. Se voidaan selvittää yleistetyn harhafunktion mallintamisella vastaavasti kuin jo kehitetyt yleistetyn varianssifunktioiden mallit. Vaihtoehtoisesti voidaan hyödyntää imputoinnin yhteydessä käytettyä varianssin estimointia.

Näkökulmia simuloinnin käytöstä tietosuojamenetelmänä

Alkuperäisen ja simuloidun aineiston välillä ei ole mitään suoraa yhteyttä, joten simuloidun aineiston voidaan ajatella poistavan kaiken paljastumisriskin. Mikäli aineiston käyttäjä tietää käyttävänsä simuloitua aineistoa, ei yksiköiden paljastaminen liene käyttäjän kannalta mielenkiintoista. Kuitenkin tilanteessa, jossa aineiston käyttäjä ei tiedä aineiston olevan simuloitu, voi hän tunnistaa aineistosta yksiköitä ja paljastaa sellaisia tietoja, jotka ovat vääriä. Tällaisen paljastamisen haitallisuutta on vaikea

mitata. Joissakin tapauksissa saadut virheelliset tiedot voivat aiheuttaa jopa enemmän harmia kuin todellisten tietojen paljastuminen. Tämä on yksi syy olla varovainen käyttäessään simulointia tietojen suojaamisessa.

Simuloinnilla on kuitenkin hyvät puolensa. Aineiston käyttäjä saa käyttöönsä laajan ja yksityiskohtaisesti luokitellun aineiston, josta voi tehdä monenlaisia analyysejä. Simuloidun aineiston käytettävyys on aivan toista luokkaa kuin esimerkiksi aiemmin esitellyn PRAM-menetelmällä suojatun aineiston. Tämänkin hyvän puolen kohdalla täytyy olla varovainen, koska simuloimalla saadun aineiston ominaisuudet riippuvat hyvin vahvasti mallintamisen onnistumisesta. Mallia muodostettaessa onkin tiedettävä mitä alkuperäisen aineiston ominaisuuksia lähdetään säilyttämään ja mistä on tarvittaessa vara tinkiä. Jos aineistosta tutkitaan ominaisuuksia tai yhteyksiä, joita ei simuloitaessa ole otettu huomioon, voidaan päätyä täysin virheellisiin johtopäätöksiin.

Mikäli on tarvetta muodostaa yleiseen käyttöön tarkoitettuja aineistoja, on simulointi turvallisin tapa muodostaa sellaisia. Tällaisen aineiston yhteydessä on muistettava ilmoittaa, ettei aineisto sisällä todellisia yksiköitä vaan on simuloitu alkuperäisen aineiston perusteella.

9.2 Suojausmenetelmien käyttö

Tietosuojamenetelmien toimivuutta on testattu aiemmin soveltamalla niitä yleisesti saatavilla oleviin aineistoihin (Domingo-Ferrer ja Torra, 2001). Tutkimuksen tavoitteena oli saada paremmuusjärjestys erilaisten tietosuojamenetelmien välille. Järjestykseen vaikuttavat sekä paljastumisriskien että informaatiokadon mitat. Tutkimus on tehty erikseen jatkuville ja luokitelluille muuttujille. Jatkuvien muuttujien suojaamisessa oli mukana tässä tutkielmassa esitellyistä menetelmistä kohinan lisääminen, mikroaggregointi ja arvojen vaihtaminen yksiköiden välillä. Luokiteltujen muuttujien suojaamista oli kokeiltu laajentamalla luokkia ylärajalla tai alarajalla, laajentamalla luokkia yleisesti ja suojaamalla PRAM-menetelmällä.

Domingo-Ferrerin ja Torran tutkimuksen mukaan jatkuvien muuttujien suojaamisessa toimivat parhaiten arvojen vaihtaminen yksiköiden välillä järjestykseen perustuen ja mikroaggregointi. Arvojen vaihtamisen tapauksessa vaihtamisen etäisyyttä mittaavan arvon oli parasta olla noin 15 prosentin luokkaa. Mikroaggregoinnissa parhaaksi valinnaksi osoittautui kolmen muuttujan samanaikainen suojaaminen. Monet menetelmistä suoriutuivat erinomaisesti informaatiokadon suhteen, mutta samalla aineiston yksiköiden paljastumisriskit kasvoivat hyvin suuriksi. Myös päinvastainen tilanne oli joidenkin menetelmien kohdalla havaittavissa.

Luokiteltujen muuttujien suojaamisessa parhaimmin pärjäsi luokkien laajentaminen ylärajalla. PRAM-menetelmä sen sijaan selvisi hyvin heikosti kokeilusta. PRAM-menetelmän ongelmaksi epäillään tutkimuksen johtopäätöksissä epäillään aineiston muuttujien luokkien lukumäärää. Menetelmä tuntui selviävän kaikkein heikoiten useita luokkia sisältävien muuttujien suojaamisessa.

Tutkielman empiirinen osa eroaa suuresti Domingo-Ferrerin ja Torran tutkimuksesta, koska tässä keskitytään vertailemaan suojattujen aineistojen frekvenssejä ja mallin estimaatteja alkuperäisen aineiston vastaaviin. Domingo-Ferrer ja Torra sen sijaan olivat käyttäneet tutkimuksessaan monia eri mittoja määrittääkseen tarkan järjestyksen menetelmien välille. Tämän tutkielman empiirisen osan tavoitteena oli saada

kokemuksia kahden teoriaosassa esitellyn tietosuojamenetelmän soveltuvuudesta käytännön suojaamiseen. Koska suojaaminen toteutettiin μ-Argus-ohjelmalla, osa suojausmenetelmien käytettävyyden analysoinnista sisältää myös ohjelman käytettävyyden analyysiä. Menetelmien käytettävyyttä tullaan tarkastelemaan yleisen käyttäjän näkökulmasta eli olettaen, ettei käyttäjällä ole erikoistuntemusta käyttämästään tietosuojamenetelmästä.

Mikroaggregoinnin soveltaminen

Mikroaggregointi on tämän hetken μ-Argus-ohjelman rajoitteiden takia käytettävissä vain jatkuvien muuttujien suojaamiseen. Luokiteltuja muuttujia voidaan niin halutessa suojata myös mikroaggregoimalla, mutta suojattavilta luokitelluilta muuttujilta vaaditaan tiettyjä ominaisuuksia. Muuttujien luokkien on oltava merkitty peräkkäisin kokonaisluvuin ja sisällöllisesti järkevien tulosten saamiseksi lukujen on oltava peräkkäisiä.

Empiirisen tutkimuksen alussa mukana olivat opettajien opettavien aineiden muuttujat. Kun jatkuvien muuttujien mikroaggregointia testattiin kyseisille muuttujille, ongelmaksi muodostuivat keskiarvona saatavat luokat, joita alkuperäisessä aineistossa ei ollut. Esimerkiksi lukiossa opetettavien aineiden listaan tulee aineiden keskiarvoina mukaan sellaisia aineita, joita opetetaan vain perusopetuksen ala- tai yläluokilla. Lisäksi siitä, että ”muut aineet” oli koodattu luvuilla 88 tai 99, saadaan opetettavan aineen koodiksi esimerkiksi 74 tai 75, joille ei ole todellista vastinetta opettavien aineiden listassa. Vastaava ilmiö oli voidaan nähdä taulukosta 11, josta huomataan kolmen muuttujan yhtäaikaisesta suojaamisesta seuraavan opettajia sellaisiin virkaluokkiin, joita ei alkuperäisessä aineistossa ollut.

Jos muuttujan luokittelu ei ole järjestysasteikollinen, voidaan luokkien nimet vaihtaa ja suojaaminen suorittaa. Tällöin päästään eroon edellä esitellystä ongelmasta. Luokkien uudelleen nimeämisessä pitää kuitenkin olla varovainen, jos luokiteltu muuttuja ei ole luonnollisesti järjestysasteikollinen. Opetettavien aineiden suojaamista tutkittaessa huomattiin, että yksittäiset tai muutamat opettajat siirtyivät suuria ryhmäkokoja käytettäessä lukuarvoltaan viereiseen luokkaan, vaikka aineen vaihtuminen ei ollut sisällöllisesti järkevää. Erityinen ongelma ilmiöstä tulee tilanteessa, jossa opettajan opettavien aineiden yhdistelmä ei ole järkevä. Tosin opettajilla voi tietenkin olla kummallisia aineyhdistelmiä luonnostaankin.

Mikroaggregointi säilyttää aineiston frekvenssit erittäin hyvin ja siksi mikroaggregoinnilla suojattua aineistoa voi käyttää erityisesti määrällisiin analyyseihin melko huolettomasti. Myös mallin sovituksessa mikroaggregoinnilla suojatun aineiston estimaatit vastasivat hyvin tarkasti alkuperäisen aineiston perusteella saatuja estimaatteja. Pienen epävarmuuden menetelmän toimivuudesta tuo epäilys mikroaggregoinnin suojaavuuden tasosta. Jos mikroaggregoinnin yhdistää otantaan, voidaan menetelmän suojaavuudesta olla varmempia, mutta tietenkin aineiston informaatiokato kasvaa otannan seurauksena.

Yleensä aineistojen suojaamisessa käytetään kynnysarvoa viisi. Mikroaggregoinnissa ryhmäkoon valinta osoittaa pienimmän frekvenssin, joka suojattavien muuttujien luokkien on saatava nollaa lukuun ottamatta. Nyt kun on osoitettu, että mikroaggregointi ryhmäkoolla kuusi antaa vielä riittävän samanlaiset frekvenssit ja mallin estimaatit kuin alkuperäinen aineisto on saanut, voidaan suositella mikroaggregoinnin käyttämistä ryhmäkoolla viisi tai kuusi. Tuota pienempiä ryhmäkokoja käytettäessä

paljastumisriski on liian suuri ja suurempien ryhmäkokojen tilanteessa arvot alkavat erota liikaa alkuperäisen aineiston arvoista. Mikroaggregoinnin käyttö tulisi mahdollisuuksien mukaan rajata vain jatkuvien muuttujien suojaamiseen.

PRAM-menetelmän soveltaminen

PRAM-menetelmän suojausparametrit antavat mahdollisuuden säätää suojausta haluamaansa suuntaan. Tämän voi nähdä menetelmän eduksi, koska optimaalista suojausta on mahdollista etsiä suhteellisen helposti. Toisaalta useamman parametrin käyttäminen saattaa helposti hämmentää suojausta ensimmäisiä kertoja käyttävää. Koska PRAM-menetelmällä suojaamiseen on loputtomasti vaihtoehtoisia PRAM-matriisin määrityksiä, ei tutkielmassa esitellyt kaksi laajempaa ja yksi pieni kokeilu kerro kaikkea menetelmän mahdollisuuksista. Näillä kolmella kokeilulla saadaan kuitenkin ohjeita siitä, millaisista parametrien arvoista kannattaa lähteä liikkeelle.

PRAM-menetelmällä suojattuja aineistoja analysoidessa tuli selväksi, että arvojen vaihtumista kannattaa aina rajata. Kun arvojen annettiin vaihtua rajatta, olivat frekvenssit jo pienillä vaihtumistodennäköisyyksillä täysin alkuperäisistä eroavia. Arvojen vaihtumisen rajoittaminen hillitsi jonkin verran frekvenssien tasoittumista ja antoi siten parempia tuloksia. Molemmissa tapauksissa vaihtumistodennäköisyyden arvo 0,07 oli vielä hyväksyttävissä, vaikkakin rajatun vaihtumisen tilanteessa arvot olivat kyseisellä todennäköisyydellä lähempänä alkuperäisiä.

Frekvenssien eroja tarkastellessa tulee mieleen, onko tavoitteena selvittää PRAM-menetelmän parametrien arvoja, joilla suojatuiden aineistojen yhteydessä PRAM-matriisin luovuttaminen on turhaa. Vaikka PRAM-matriisi luovutettaisiinkin aineiston mukana, tulisi aineistosta saada ilman matriisin käyttämistä edes jotakin käsitystä tutkittavan ilmiön käyttäytymisestä. Siksi frekvenssien tarkastelu ei ole turhaa. Lisäksi frekvenssejä tarkastellessa tulee huolehdituksi, ettei aineiston yksiköiden paljastaminen ole liian helppoa. Esimerkiksi tarkasteltaessa tilannetta aineiston suojaustason kannalta, tarkoittaa 0,10 vaihtumistodennäköisyys sitä, että joka kymmenes arvo voi olla muuttunut. Koska aineiston käyttäjä ei voi tietää mitkä arvoista ovat pysyneet ennallaan ja mitkä muuttuneet, voi tällainen todennäköisyys olla hyvinkin riittävä aineiston suojaamiseen. Mikäli tunkeutuja yrittää tunnistaa yksiköitä ja löytää sopivan yksikön, on hänen vaikea olla varma yhdistämisen oikeellisuudesta. Tällainen sattumanvarainen yhdistäminen on sitä hankalampaa, mitä useampia aineiston muuttujista suojataan.

Empiirisen tutkimuksen perusteella voidaan suositella PRAM-menetelmän käyttämistä rajoittamalla arvojen vaihtuminen sopivalla arvolla ja valita vaihtumistodennäköisyydeksi 0,05–0,07. Lisäksi paljastumisriskin pienentämiseksi on syytä suojata useita luovutettavan aineiston muuttujia.

10 Sanasto

Tietosuojamenetelmiin perehtyessä tulee vastaan useita uusia termejä, joiden merkitys ei heti ole aivan selvä. Alkuunpääsyn helpottamiseksi tähän lukuun on koottu keskeisimpiä termejä selityksineen sekä käännöksineen. Kyseessä oleva tutkimusala on vielä sen verran nuori, etteivät kaikki esitellyt termit ole täysin vakiintuneita. Tästä syystä englanninkielisistä termeistä näkee useita eri versioita kirjallisuuteen tutustuessaan. Onneksi termit muistuttavat yleensä toisiaan ja ainakin asiayhteydestä on helppo ymmärtää mistä on kyse.

Suomenkielisistä termeistä suurin osa on julkaisusta Tilastolliset tietosuojamenetelmät ja niiden käyttö (Hänninen, 1997), mutta osaa termeistä on muokattu ja uusia termejä on täytynyt kääntää tutkielman edetessä. Myös suomenkielinen sanasto tulee muokkautumaan tietosuojausterminologian käytön yleistyessä. Tässä esiteltävä sanasto vastaa tutkielmassa käytettyjä termejä ja toimii apuna tutkielmaa luettaessa.

- *aineiston muuntamiseen perustuvat menetelmät*
  (eng. data perturbative methods)
  Menetelmät, jotka perustuvat aineistossa olevien muuttujanarvojen muokkaamiseen tavalla tai toisella.
- *aineiston rajoittamiseen perustuvat menetelmät*
  (eng. data limiting methods, data non-perturbative methods)
  Menetelmät, jotka perustuvat aineistossa olevien yksiköiden tai muuttujien määrän rajoittamiseen tai muuttujien arvojen peittämiseen.
- *avainmuuttuja (eng. key variable)*
  Sellaisia identifioivia muuttujia, joiden avulla yksikkö voidaan paljastaa epäsuorasti, kutsutaan avainmuuttujiksi. Joskus avainmuuttujilla viitataan yleisesti identifioiviin muuttujiin.
- *epäsuora identifioiminen (eng.)*
  Epäsuora identifioiminen tapahtuu usean muuttujan arvoja tarkastelemalla. Esimerkiksi pienen kunnan asukkaita voidaan helposti tunnistaa jopa otosaineistosta, jos tiedossa on henkilön ikä ja ammatti.
- *herkkä muuttuja (eng. sensitive variable)*
  Herkäksi muuttujaksi kutsutaan sellaista muuttujaa, jonka arvojen paljastuminen aiheuttaa haittaa yksikölle. Esimerkiksi tutkielmassa opettajan pätevyys hoitamaansa tehtävään on herkkä muuttuja.
- *identifioiva muuttuja (eng. identifying variable)*
  Muuttujaa, jonka avulla yksikkö on yhdistettävissä tiettyyn yksikköön, kutsutaan identifioivaksi muuttujaksi. Mikäli kyseessä on yksikäsitteinen yhdistäminen, kutsutaan muuttujaa suoraan identifioivaksi, muulloin puhutaan epäsuoraan identifioivista muuttujista. Ks. epäsuora identifioiminen.
- *informaatiokato (eng. information loss)*
  Aineiston suojaamisesta seuraava aineiston informaation väheneminen.
- *kynnysarvo (eng. threshold)*
  Kynnysarvo on se lukumäärä, jota enemmän yksiköitä kullekin identifioivien muuttujien arvojen yhdisteelle on oltava, jotta aineisto voidaan tulkita suojatuksi.
- *MASSC, Micro Agglomeration, Substitution for perturbation, Subsampling for suppression, optimal sampling weight Calibration*
  Otantateoriaan perustuva tietosuojamenetelmä.

- *paljastumisriski (eng. disclosure risk)*
  Yleisesti paljastumisriskillä tarkoitetaan julkaistavan aineiston yksikön tai yksikön ominaisuuksien riskiä paljastua. Nimitystä käytetään myös puhuttaessa keskimääräisestä aineistoon liittyvästä riskistä yksiköiden paljastumiselle.
- *PRAM, The Post-RAndomization Method*
  Aineistoa muokkaava tietosuojamenetelmä, jonka suojaavuus riippuu todennäköisyyksistä.
- *profiili (eng. profile)*
  Yksikön identifioivien muuttujien saamien arvojen yhdistelmää kutsutaan profiiliksi. Esimerkiksi henkilön ikä, asuinkunta, ammatti ja sukupuoli riittävät usein henkilön tunnistamiseen ja siten suojaamisessa tarkastellaan näiden muuttujien saamien arvojen yhdistelmiä. Tietosuojauksen tavoitteena on tuottaa aineisto, jossa kunkin profiilin frekvenssi on riittävän suuri, ettei tunnistaminen onnistu tai ole enää varmaa.
- *suojausmenetelmä, salausmenetelmä (eng. disclosure method)*
  Menetelmä, jolla alkuperäistä aineistoa muokataan aineiston yksiköiden identiteetin suojaamiseksi.
- *tietosuojaus (eng. inference control, statistical database protection)*
  Yleisnimitys tavoitteelle suojata tiedonantajien yksityisyyttä tietoja julkaistaessa. Tavoitteet sisältävät lait, tilastotuotannnon vaiheet yms.
- *tilastollinen tietosuojaus*
  (eng. statistical disclosure control, statistical disclosure limitation)
  Yleisnimitys menetelmille, jotka tähtäävät julkaistavien tilastojen ja aineistojen yksiköiden identiteetin suojaamiseen. Menetelmät perustuvat useimmiten julkaistavien tietojen rajoittamiseen tai muokkaamiseen.
- *tunkeutuja, paljastamista yrittävä henkilö (eng. intruder, snooper)*
  Henkilö, jonka tavoitteena on tunnistaa aineistossa mukana oleva yksikkö tai muuten murtaa aineiston suojaus esimerkiksi aineiston julkaisijan maineen tahraamiseksi.

Lähteet

[1] *Asetus viranomaisten toiminnan julkisuudesta ja hyvästä tiedonhallintatavasta, 1030/1999.* Helsinki.

[2] Benedetti, R., Franconi, L., Piersimoni, F. 1999. *Per-record risk of disclosure in depend data.*
Julkaisussa: *Statistical Data Protection: Proceedings of the conference (Lisbon, 1998).* Luxembourg: Office for Official Publications of the European Communities.

[3] Brand, Ruth 2002. *Microdata Protection through Noise Addition.*
Julkaisussa: *Inference Control in Statistical Databases*, toim. Domingo-Ferrer, Josep. Berlin/Heidelberg: Springer.

[4] CASC – Computational Aspects of Statistical Confidentiality. Viitattu kevät 2006. <http://http://neon.vb.cbs.nl/casc/>

[5] CENEX-SDC, CENtre of EXcellence for Statistical Disclosure Control. Viitattu kevät 2006. <http://http://neon.vb.cbs.nl/cenex/>

[6] de Wolf, Peter-Paul, van Gelder, Ilan 2004. *An empirical evaluation of PRAM.* Discussion paper 04012. Statistics Netherlands. Voorburg/Heerlen.

[7] Domingo-Ferrer, Joseph 2002. *Advances in Inference Control in Statistical Databases: An Overview.*
Julkaisussa: *Inference Control in Statistical Databases*, toim. Domingo-Ferrer, Josep. Berlin/Heidelberg: Springer.

[8] Domingo-Ferrer, J., Torra, V. 2001. *A Quantitative Comparison of Disclosure Control Methods for Microdata.*
Julkaisussa: *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, toim. Doyle et al. Amsterdam: North-Holland.

[9] Fienberg, S. E. 2000. *Confidentiality and Data Protection Through Disclosure Limitation: Evolving Principles and Technical Advances.* The Philippine Statistician. Vol. 49, s. 1–12.

[10] Gouweleeuw, J., Kooiman, P., Willenborg, L., and de Wolf, P.-P. 1998. *Post Randomisation for Statistical Disclosure Control: Theory and Implementation.* Journal of Official Statistics. Vol. 14, No.4, s. 463–478.

[11] Group Crises 2004a. *Research Reports: Non-Perturbative Methods for Microdata Privacy in Statistical Databases.* Viitattu heinäkuu 2005. <http://vneumann.etse.urv.es/publications/reports/>

[12] Group Crises 2004b. *Research Reports: Perturbative Masking for Microdata Privacy Protection in Statistical Databases.* Viitattu heinäkuu 2005. <http://vneumann.etse.urv.es/publications/reports/>

[13] Group Crises 2004c. *Research Reports: An Introduction to Microdata Protection for Database Privacy.* Viitattu heinäkuu 2005. <http://vneumann.etse.urv.es/publications/reports/>

[14] Group Crises 2004d. *Research Reports: Additive Noise for Microdata Privacy Protection in Statistical Databases.* Viitattu heinäkuu 2005. <http://vneumann.etse.urv.es/publications/reports/>

[15] Group Crises 2004e. *Research Reports: Microaggregation for Privacy Protection in Statistical Databases.* Viitattu heinäkuu 2005. <http://vneumann.etse.urv.es/publications/reports/>

[16] Group Crises 2004f. *Research Reports: Synthetic Microdata Generation for Database Privacy Protection.* Viitattu heinäkuu 2005. <http://vneumann.etse.urv.es/publications/reports/>

[17] *Henkilötietolaki, 523/1999.* Helsinki.

[18] Hundepool, van de Wetering, Ramaswamy, Franconi, Polettini, Capobianchi, de Wolf, Domingo-Ferrer, Torra, Brand, Giessing. 2005. *μ-Argus version 4.0 User's Manual.* Department of Statistical Methods, Statistics Netherlands. Voorburg, The Netherlands. <http://neon.vb.cbs.nl/casc/>

[19] Hänninen, Minna 1997. *Tilastolliset tietosuojamenetelmät ja niiden käyttö.* Katsauksia 1997/3. Tilastokeskus. Helsinki: Edita.

[20] International Statistical Institute 1985. *Declaration on Professional Ethics. Suom.(1989) Kansainvälisen tilastoinstituutin tilastoalan ammattietiikkajulistus.* Tilastokeskus. Helsinki.

[21] *Laki Tilastokeskuksesta, 48/1992.* Helsinki.

[22] *Laki viranomaisten toiminnan julkisuudesta, 621/1999.* Helsinki.

[23] Little, R.J.A. 1993. *Statistical Analysis of Masked Data.* Journal of Official Statistics. Vol. 9, No.2, s. 407–426.

[24] Neuvoston asetus (EY) N:o 322/97, yhteisön tilastoista. Euroopan yhteisöjen virallinen lehti N:o L 52, 22.2.1997 s. 1–7. <http://europa.eu.int/eur-lex>

[25] Polettini, S., Franconi, L., Stander, J. 2002. *Model Based Disclosure Protection.*
Julkaisussa: *Inference Control in Statistical Databases*, toim. Domingo-Ferrer, Josep. Berlin/Heidelberg: Springer.

[26] Raghunathan, T.E., Reiter J.P., Rubin, D.B. 2003. *Multiple Imputation for Statistical Disclosure Limitation.* Journal of Official Statistics. Vol. 19, No.1, s. 1–16.

[27] *Rikoslaki, 39/1889.* Helsinki.

[28] Rubin, D.B. 1993. *Discussion: Statistical Disclosure Limitation.* Journal of Official Statistics. Vol. 9, No.2, s. 461–468.

[29] Singh, A.C., Yu, F., Dunteman, G.H. 2004. *MASSC: A New Data Mask fo Limiting Statistical Information Loss and Disclosure.*
Julkaisussa: *Monographs of official statistics, Work session on statistical data confidentiality (Luxembourg, 2003).* Luxembourg: Office for Official Publications of the European Communities.

[30] Statistiska centralbyrån 2001. *Statistisk röjandekontroll.* Statistiska centralbyrån. Örebro.

[31] Tilastokeskus 2002a. *Toimi oikein tilastoalalla. Tilastokeskuksen ammattieettinen opas.* Tilastokeskus. Käsikirjoja. Helsinki.

[32] Tilastokeskus 2002b. *Tilastokeskuksen tietoturvaperiaatteet 21.10.2002.* TK-46-1269-02. Helsinki.

[33] Tilastokeskus 2005a. *Tilastolain soveltamisohje.* TK-00-198-05. Helsinki.

[34] Tilastokeskus 2005b. *Ohje käyttölupien myöntämisestä Tilastokeskuksen yksikkötason aineistoon 18.2.2005.* TK-00-128-05. Helsinki.

[35] *Tilastolaki, 280/2004.* Helsinki.

[36] Yancey, W.E., Winkler, W.E., Creecy, R.H. 2002. *Disclosure Risk Assessment in Perturbative Microdata Protection.*
Julkaisussa: *Inference Control in Statistical Databases*, toim. Domingo-Ferrer, Josep. Berlin/Heidelberg: Springer.

Liite 1

Taulukko 1: Mikroaggregoinnin vaikutus frekvensseihin. Suojaus mikroaggregoimalla kolme muuttujaa yhdessä. Osa 1. (Alkuperäinen aineisto on merkitty ryhmäkoolla 1.)

Ryhmäkoko	Ikä	n	Virka	n	Aste	n
1	alle 40	2467	rehtori	432	01	1
	40–53	2849	luokanopettaja	-	02	10
	yli 53	2482	erityisopettaja	-	03	343
			lehtori	5684	04	7
			opinto-ohjaaja	142	05	6672
			päätoiminen tuntiop.	1078	06	686
			sivutoiminen tuntiop.	414	07	22
			muu virka	48	08	7
					09	7
					10	16
					11	3
					12	24
2	alle 40	2468	rehtori	428	01	0
	40–53	2840	luokanopettaja	-	02	8
	yli 53	2490	erityisopettaja	-	03	346
			lehtori	5684	04	6
			opinto-ohjaaja	142	05	6672
			päätoiminen tuntiop.	1076	06	684
			sivutoiminen tuntiop.	416	07	24
			muu virka	52	08	8
					09	4
					10	18
					11	4
					12	24
3	alle 40	2472	rehtori	429	01	0
	40–53	2848	luokanopettaja	-	02	9
	yli 53	2478	erityisopettaja	-	03	345
			lehtori	5680	04	6
			opinto-ohjaaja	144	05	6673
			päätoiminen tuntiop.	1068	06	684
			sivutoiminen tuntiop.	408	07	24
			muu virka	69	08	9
					09	6
					10	15
					11	3
					12	24
4	alle 40	2468	rehtori	424	01	0
	40–53	2846	luokanopettaja	-	02	8
	yli 53	2484	erityisopettaja	-	03	344
			lehtori	5670	04	4
			opinto-ohjaaja	152	05	6678
			päätoiminen tuntiop.	1072	06	684
			sivutoiminen tuntiop.	416	07	20
			muu virka	64	08	12
					09	4
					10	16
					11	4
					12	24

Taulukko 2: Mikroaggregoinnin vaikutus frekvensseihin. Suojaus mikroaggregoimalla kolme muuttujaa yhdessä. Osa 2.

Ryhmäkoko	Ikä	n	Virka	n	Aste	n
5	alle 40	2470	rehtori	430	01	0
	40–53	2853	luokanopettaja	-	02	10
	yli 53	2475	erityisopettaja	-	03	345
			lehtori	5693	04	5
			opinto-ohjaaja	150	05	6673
			päätoiminen tuntiop.	1070	06	680
			sivutoiminen tuntiop.	410	07	30
			muu virka	45	08	5
					09	10
					10	15
					11	0
					12	25
6	alle 40	2460	rehtori	432	01	0
	40–53	2860	luokanopettaja	-	02	6
	yli 53	2478	erityisopettaja	-	03	348
			lehtori	5674	04	0
			opinto-ohjaaja	162	05	6676
			päätoiminen tuntiop.	1080	06	684
			sivutoiminen tuntiop.	402	07	30
			muu virka	48	08	6
					09	6
					10	12
					11	6
					12	24
7	alle 40	2457	rehtori	427	01	0
	40–53	2863	luokanopettaja	-	02	7
	yli 53	2478	erityisopettaja	-	03	350
			lehtori	5677	04	0
			opinto-ohjaaja	154	05	6671
			päätoiminen tuntiop.	1071	06	693
			sivutoiminen tuntiop.	399	07	21
			muu virka	70	08	7
					09	7
					10	14
					11	0
					12	28
8	alle 40	2464	rehtori	424	01	0
	40–53	2846	luokanopettaja	-	02	8
	yli 53	2488	erityisopettaja	-	03	344
			lehtori	5662	04	8
			opinto-ohjaaja	152	05	6670
			päätoiminen tuntiop.	1064	06	672
			sivutoiminen tuntiop.	400	07	24
			muu virka	96	08	8
					09	8
					10	16
					11	8
					12	24

Taulukko 3: Mikroaggregoinnin vaikutus frekvensseihin. Suojaus mikroaggregoimalla kolme muuttujaa yhdessä. Osa 3.

Ryhmäkoko	Ikä	n	Virka	n	Aste	n
9	alle 40	2457	rehtori	423	01	0
	40–53	2857	luokanopettaja	-	02	9
	yli 53	2484	erityisopettaja	-	03	351
			lehtori	5674	04	0
			opinto-ohjaaja	126	05	6673
			päätoiminen tuntiop.	1080	06	693
			sivutoiminen tuntiop.	396	07	18
			muu virka	99	08	9
					09	0
					10	18
					11	9
					12	18
10	alle 40	2470	rehtori	420	01	0
	40–53	2858	luokanopettaja	-	02	10
	yli 53	2470	erityisopettaja	-	03	340
			lehtori	5658	04	10
			opinto-ohjaaja	140	05	6668
			päätoiminen tuntiop.	1080	06	690
			sivutoiminen tuntiop.	380	07	30
			muu virka	120	08	0
					09	10
					10	10
					11	10
					12	20
11	alle 40	2453	rehtori	429	01	0
	40–53	2859	luokanopettaja	-	02	11
	yli 53	2486	erityisopettaja	-	03	341
			lehtori	5664	04	11
			opinto-ohjaaja	154	05	6665
			päätoiminen tuntiop.	1067	06	704
			sivutoiminen tuntiop.	385	07	11
			muu virka	99	08	0
					09	11
					10	11
					11	11
					12	22
15	alle 40	2445	rehtori	435	01	0
	40–53	2923	luokanopettaja	-	02	15
	yli 53	2430	erityisopettaja	-	03	345
			lehtori	5668	04	0
			opinto-ohjaaja	150	05	6688
			päätoiminen tuntiop.	1110	06	690
			sivutoiminen tuntiop.	405	07	0
			muu virka	30	08	15
					09	0
					10	15
					11	15
					12	15

Taulukko 4: Mikroaggregoinnin vaikutus frekvensseihin. Suojaus mikroaggregoimalla kolme muuttujaa yhdessä. Aineistossa muuttujan virka nimeäminen kuten alkuperäisessä aineistossa. Osa 1. (Alkuperäistä aineistoa on merkitty ryhmäkoolla 1.)

Ryhmäkoko	Ikä	n	Virka	n	Aste	n
1	alle 40	2467	rehtori	432	01	1
	40–53	2849	luokanopettaja	-	02	10
	yli 53	2482	erityisopettaja	-	03	343
			lehtori	5684	04	7
			opinto-ohjaaja	142	05	6672
			päätoiminen tuntiop.	1078	06	686
			sivutoiminen tuntiop.	414	07	22
			muu virka	48	08	7
					09	7
					10	16
					11	3
					12	24
2	alle 40	2468	rehtori	430	01	0
	40–53	2840	luokanopettaja	-	02	6
	yli 53	2490	erityisopettaja	4	03	346
			lehtori	5676	04	6
			opinto-ohjaaja	152	05	6672
			päätoiminen tuntiop.	1068	06	686
			sivutoiminen tuntiop.	420	07	22
			muu virka	48	08	6
					09	8
					10	16
					11	4
					12	24
3	alle 40	2469	rehtori	429	01	0
	40–53	2857	luokanopettaja	-	02	9
	yli 53	2472	erityisopettaja	9	03	345
			lehtori	5659	04	6
			opinto-ohjaaja	159	05	6673
			päätoiminen tuntiop.	1083	06	684
			sivutoiminen tuntiop.	417	07	24
			muu virka	42	08	9
					09	6
					10	15
					11	3
					12	24
4	alle 40	2460	rehtori	428	01	0
	40–53	2854	luokanopettaja	-	02	8
	yli 53	2484	erityisopettaja	4	03	344
			lehtori	5666	04	8
			opinto-ohjaaja	148	05	6670
			päätoiminen tuntiop.	1080	06	688
			sivutoiminen tuntiop.	416	07	24
			muu virka	56	08	4
					09	8
					10	16
					11	4
					12	24

Taulukko 5: Mikroaggregoinnin vaikutus frekvensseihin. Suojaus mikroaggregoimalla kolme muuttujaa yhdessä. Aineistossa muuttujan virka nimeäminen kuten alkuperäisessä aineistossa. Osa 2.

Ryhmäkoko	Ikä	n	Virka	n	Aste	n
5	alle 40	2470	rehtori	430	01	0
	40–53	2843	luokanopettaja	-	02	10
	yli 53	2485	erityisopettaja	5	03	345
			lehtori	5658	04	5
			opinto-ohjaaja	155	05	6673
			päätoiminen tuntiop.	1100	06	685
			sivutoiminen tuntiop.	410	07	25
			muu virka	40	08	5
					09	10
					10	10
					11	5
					12	25
6	alle 40	2460	rehtori	420	01	0
	40–53	2860	luokanopettaja	6	02	6
	yli 53	2478	erityisopettaja	12	03	348
			lehtori	5656	04	0
			opinto-ohjaaja	156	05	6682
			päätoiminen tuntiop.	1086	06	678
			sivutoiminen tuntiop.	408	07	24
			muu virka	54	08	12
					09	6
					10	12
					11	6
					12	24
7	alle 40	2471	rehtori	420	01	0
	40–53	2863	luokanopettaja	-	02	7
	yli 53	2464	erityisopettaja	21	03	350
			lehtori	5656	04	0
			opinto-ohjaaja	154	05	6678
			päätoiminen tuntiop.	1099	06	679
			sivutoiminen tuntiop.	406	07	28
			muu virka	42	08	7
					09	7
					10	14
					11	7
					12	21
8	alle 40	2440	rehtori	424	01	0
	40–53	2894	luokanopettaja	-	02	8
	yli 53	2464	erityisopettaja	16	03	344
			lehtori	5662	04	8
			opinto-ohjaaja	128	05	6678
			päätoiminen tuntiop.	1112	06	672
			sivutoiminen tuntiop.	416	07	32
			muu virka	40	08	8
					09	8
					10	8
					11	8
					12	24

Taulukko 6: Mikroaggregoinnin vaikutus frekvensseihin. Suojaus mikroaggregoimalla kolme muuttujaa yhdessä. Aineistossa muuttujan virka nimeäminen kuten alkuperäisessä aineistossa. Osa 3.

Ryhmäkoko	Ikä	n	Virka	n	Aste	n
9	alle 40	2457	rehtori	414	01	0
	40–53	2857	luokanopettaja	18	02	9
	yli 53	2484	erityisopettaja	9	03	342
			lehtori	5683	04	9
			opinto-ohjaaja	144	05	6682
			päätoiminen tuntiop.	1098	06	675
			sivutoiminen tuntiop.	396	07	27
			muu virka	36	08	0
					09	9
					10	18
					11	0
					12	27
10	alle 40	2450	rehtori	420	01	0
	40–53	2878	luokanopettaja	-	02	10
	yli 53	2470	erityisopettaja	20	03	340
			lehtori	5658	04	10
			opinto-ohjaaja	150	05	6678
			päätoiminen tuntiop.	1120	06	680
			sivutoiminen tuntiop.	400	07	30
			muu virka	30	08	0
					09	10
					10	10
					11	10
					12	20
11	alle 40	2464	rehtori	429	01	0
	40–53	2881	luokanopettaja	-	02	11
	yli 53	2453	erityisopettaja	11	03	341
			lehtori	5664	04	11
			opinto-ohjaaja	154	05	6665
			päätoiminen tuntiop.	1089	06	693
			sivutoiminen tuntiop.	396	07	22
			muu virka	55	08	0
					09	11
					10	11
					11	11
					12	22
15	alle 40	2475	rehtori	420	01	0
	40–53	2893	luokanopettaja	-	02	15
	yli 53	2430	erityisopettaja	15	03	345
			lehtori	5638	04	0
			opinto-ohjaaja	165	05	6658
			päätoiminen tuntiop.	1095	06	720
			sivutoiminen tuntiop.	435	07	0
			muu virka	30	08	15
					09	0
					10	15
					11	0
					12	30

Taulukko 7: Pätevyyden tarkastelua. Pätevyyden ja ikäluokkien ristiintaulukointi. Suojattu käyttäen mikroaggregointia. Osa 1. (Alkuperäistä aineistoa on merkitty ryhmäkoolla 1.)

k	Ikä	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
1	alle 40 v.	2128	339	2467
		86,26	13,74	100,00
	40–53 v.	2666	183	2849
		37,07	30,15	36,54
		93,58	6,42	100,00
	yli 53 v.	2397	85	2482
		96,58	3,42	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	195,0332	p<0,0001	
	LR	190,7357	p<0,0001	
2	alle 40 v.	2126	342	2468
		86,14	13,86	100,00
	40–53 v.	2662	178	2840
		93,73	6,27	100,00
	yli 53 v.	2403	87	2490
		96,51	3,49	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	χ^2	199,7616	p<0,0001	
	LR	193,6664	p<0,0001	
3	alle 40 v.	2130	342	2472
		86,17	13,83	100,00
	40–53 v.	2671	177	2848
		93,79	6,21	100,00
	yli 53 v.	2390	88	2478
		96,45	3,55	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	197,7091	p<0,0001	
	LR	191,1025	p<0,0001	
4	alle 40 v.	2125	343	2468
		86,10	13,90	100,00
	40–53 v.	2671	175	2846
		93,85	6,15	100,00
	yli 53 v.	2395	89	2484
		96,42	3,58	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	200,1932	p<0,0001	
	LR	192,7479	p<0,0001	

Taulukko 8: Pätevyyden tarkastelua. Pätevyyden ja ikäluokkien ristiintaulukointi. Suojattu käyttäen mikroaggregointia. Osa 2.

Tn	Ikä	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
	alle 40 v.	2129	341	2470
		86,19	13,81	100,00
	40–53 v.	2671	182	2853
5		93,62	6,38	100,00
	yli 53 v.	2391	84	2475
		96,61	3,39	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	χ^2	199,0672	p<0,0001	
	LR	194,5190	p<0,0001	
	alle 40 v.	2117	343	2460
		86,06	13,94	100,00
	40–53 v.	2682	178	2860
6		93,78	6,22	100,00
	yli 53 v.	2392	86	2478
		96,53	3,47	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	203,9335	p<0,0001	
	LR	197,3288	p<0,0001	
	alle 40 v.	2115	342	2457
		86,08	13,92	100,00
	40–53 v.	2682	181	2863
7		93,68	6,32	100,00
	yli 53 v.	2394	84	2478
		96,61	3,39	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	204,0305	p<0,0001	
	LR	198,6061	p<0,0001	
	alle 40 v.	2122	342	2464
		86,12	13,88	100,00
	40–53 v.	2666	180	2846
8		93,68	6,32	100,00
	yli 53 v.	2403	85	2488
		96,58	3,42	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	χ^2	202,1184	p<0,0001	
	LR	196,7408	p<0,0001	

Taulukko 9: Pätevyyden tarkastelua. Pätevyyden ja ikäluokkien ristiintaulukointi. Suojattu käyttäen mikroaggregointia. Osa 3.

Tn	Ikä	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
	alle 40 v.	2116	341	2457
		86,12	13,88	100,00
	40–53 v.	2679	178	2857
9		93,77	6,23	100,00
	yli 53 v.	2396	88	2484
		96,46	3,54	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	199,0038	p<0,0001	
	LR	192,2164	p<0,0001	
	alle 40 v.	2127	343	2470
		86,11	13,89	100,00
	40–53 v.	2673	185	2858
10		93,53	6,47	100,00
	yli 53 v.	2391	79	2470
		96,80	3,20	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	207,3504	p<0,0001	
	LR	204,5974	p<0,0001	
	alle 40 v.	2115	338	2453
		86,22	13,78	100,00
	40–53 v.	2678	181	2859
11		93,67	6,33	100,00
	yli 53 v.	2398	88	2486
		96,46	3,54	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	χ^2	193,6152	p<0,0001	
	LR	187,7638	p<0,0001	
	alle 40 v.	2106	339	2445
		86,13	13,87	100,00
	40–53 v.	2736	187	2923
15		93,60	6,40	100,00
	yli 53 v.	2349	81	2430
		96,67	3,33	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	200,8416	p<0,0001	
	LR	196,4186	p<0,0001	

Taulukko 10: Pätevyyden tarkastelua. Pätevyyden ja virkaluokkien ristiintaulukointi. Suojattu käyttäen mikroaggregointia. Osa 1. (Alkuperäistä aineistoa on merkitty ryhmäkoolla 1.)

Tn	Virka	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
	rehtorit	425	7	432
		98,38	1,62	100,00
	lehtorit	5519	165	5684
		97,10	2,90	100,00
	op.-ohjaajat	125	17	142
		88,03	11,97	100,00
1	päätoim. tuntiop.	875	203	1078
		81,17	18,83	100,00
	sivutoim. tuntiop.	208	206	414
		50,24	49,76	100,00
	muu virka	39	9	48
		81,25	18,75	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1422,4648	p<0,0001	
	LR	932,8533	p<0,0001	
	rehtorit	421	7	428
		98,36	1,64	100,00
	lehtorit	5518	166	5684
		97,08	2,92	100,00
	op.-ohjaajat	124	18	142
		87,32	12,68	100,00
2	päätoim. tuntiop.	876	200	1076
		81,41	18,59	100,00
	sivutoim. tuntiop.	212	204	416
		50,96	49,04	100,00
	muu virka	40	12	52
		76,92	23,08	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1392,8053	p<0,0001	
	LR	919,2068	p<0,0001	
	rehtorit	420	9	429
		97,90	2,10	100,00
	lehtorit	5514	166	5680
		97,08	2,92	100,00
	op.-ohjaajat	129	15	144
		89,58	10,42	100,00
3	päätoim. tuntiop.	869	199	1068
		81,37	18,63	100,00
	sivutoim. tuntiop.	206	202	408
		50,49	49,51	100,00
	muu virka	53	16	69
		76,81	23,19	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1395,2526	p<0,0001	
	LR	913,9562	p<0,0001	

Taulukko 11: Pätevyyden tarkastelua. Pätevyyden ja virkaluokkien ristiintaulukointi. Suojattu käyttäen mikroaggregointia. Osa 2.

Tn	Virka	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
	rehtorit	417	7	424
		98,35	1,65	100,00
	lehtorit	5504	166	5670
		97,07	2,93	100,00
	op.-ohjaajat	138	14	152
		90,79	9,21	100,00
4	päätoim. tuntiop.	872	200	1072
		81,34	18,66	100,00
	sivutoim. tuntiop.	213	203	416
		51,20	48,80	100,00
	muu virka	47	17	64
		73,44	26,56	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1391,7963	p<0,0001	
	LR	918,5203	p<0,0001	
	rehtorit	423	7	430
		98,37	1,63	100,00
	lehtorit	5525	168	5693
		97,05	2,95	100,00
	op.-ohjaajat	133	17	150
		88,67	11,33	100,00
5	päätoim. tuntiop.	866	204	1070
		80,93	19,07	100,00
	sivutoim. tuntiop.	208	202	410
		50,73	49,27	100,00
	muu virka	36	9	45
		80,00	20,00	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1392,6257	p<0,0001	
	LR	916,7628	p<0,0001	
	rehtorit	425	7	432
		98,38	1,62	100,00
	lehtorit	5508	166	5674
		97,07	2,93	100,00
	op.-ohjaajat	142	20	162
		87,65	12,35	100,00
6	päätoim. tuntiop.	877	203	1080
		81,20	18,80	100,00
	sivutoim. tuntiop.	200	202	402
		49,75	50,25	100,00
	muu virka	39	9	48
		81,25	18,75	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1414,5238	p<0,0001	
	LR	925,1851	p<0,0001	

Taulukko 12: Pätevyyden tarkastelua. Pätevyyden ja virkaluokkien ristiintaulukointi. Suojattu käyttäen mikroaggregointia. Osa 3.

Tn	Virka	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
	rehtorit	419	8	427
		98,13	1,87	100,00
	lehtorit	5510	167	5677
		97,06	2,94	100,00
	op.-ohjaajat	138	16	154
		89,61	10,39	100,00
7	päätoim. tuntiop.	868	203	1071
		81,05	18,95	100,00
	sivutoim. tuntiop.	200	199	399
		50,13	49,87	100,00
	muu virka	56	14	70
		80,00	20,00	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1393,1677	p<0,0001	
	LR	912,6710	p<0,0001	
	rehtorit	417	7	424
		98,35	1,65	100,00
	lehtorit	5495	167	5662
		97,05	2,95	100,00
	op.-ohjaajat	136	16	152
		89,47	10,53	100,00
8	päätoim. tuntiop.	870	194	1064
		81,77	18,23	100,00
	sivutoim. tuntiop.	201	199	400
		50,25	49,75	100,00
	muu virka	72	24	96
		75,00	25,00	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1391,0432	p<0,0001	
	LR	912,3280	p<0,0001	
	rehtorit	417	6	423
		98,58	1,42	100,00
	lehtorit	5507	167	5674
		97,06	2,94	100,00
	op.-ohjaajat	112	14	126
		88,89	11,11	100,00
9	päätoim. tuntiop.	873	207	1080
		80,83	19,17	100,00
	sivutoim. tuntiop.	202	194	396
		51,01	48,99	100,00
	muu virka	80	19	99
		80,81	19,19	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1360,6405	p<0,0001	
	LR	906,3643	p<0,0001	

Taulukko 13: Pätevyyden tarkastelua. Pätevyyden ja virkaluokkien ristiintaulukointi. Suojattu käyttäen mikroaggregointia. Osa 4.

Tn	Virka	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
	rehtorit	413	7	420
		98,33	1,67	100,00
	lehtorit	5489	169	5658
		97,01	2,99	100,00
	op.-ohjaajat	126	14	140
		90,00	10,00	100,00
10	päätoim. tuntiop.	872	208	1080
		80,74	19,26	100,00
	sivutoim. tuntiop.	190	190	380
		50,00	50,00	100,00
	muu virka	101	19	120
		84,17	15,83	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1356,6618	p<0,0001	
	LR	893,1480	p<0,0001	
	rehtorit	421	8	429
		98,14	1,86	100,00
	lehtorit	5497	167	5664
		97,05	2,95	100,00
	op.-ohjaajat	139	15	154
		90,26	9,74	100,00
11	päätoim. tuntiop.	860	207	1067
		80,60	19,40	100,00
	sivutoim. tuntiop.	197	188	385
		51,17	48,83	100,00
	muu virka	77	22	99
		77,78	22,22	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1339,2735	p<0,0001	
	LR	892,7383	p<0,0001	
	rehtorit	426	9	435
		97,93	2,07	100,00
	lehtorit	5502	166	5668
		97,07	2,93	100,00
	op.-ohjaajat	133	17	150
		88,67	11,33	100,00
15	päätoim. tuntiop.	898	212	1110
		80,90	19,10	100,00
	sivutoim. tuntiop.	208	197	405
		51,36	48,64	100,00
	muu virka	24	6	30
		80,00	20,00	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1354,6721	p<0,0001	
	LR	898,2604	p<0,0001	

Taulukko 14: PRAM-menetelmän vaikutus frekvenssihin, kun aste on tarkemmissa luokissa. Suojaus ilman vaihdon rajoittamista. (Alkuperäistä aineistoa on merkitty vaihtumattomuustodennäköisyydellä 1.)

	Vaihtumattomuustodennäköisyys										
Aste	1	0,97	0,95	0,93	0,9	0,85	0,8	0,75	0,7	0,65	0,6
01	1	22	30	59	74	94	142	164	200	246	301
02	10	25	45	61	79	92	139	194	221	241	265
03	343	357	361	359	395	382	430	437	459	455	505
04	7	28	37	57	61	104	140	165	206	272	319
05	6672	6487	6347	6202	5978	5701	5350	5049	4693	4418	4011
06	686	679	678	693	684	697	684	663	648	671	674
07	22	39	61	72	89	150	145	192	228	232	272
08	7	36	33	49	77	108	158	172	250	247	299
09	7	24	44	51	71	101	152	177	229	231	275
10	16	36	64	73	89	119	152	188	233	272	315
11	3	27	43	60	99	110	142	196	201	266	262
12	24	38	55	62	102	140	164	201	230	247	300

Taulukko 15: PRAM-menetelmän vaikutus frekvenssihin, kun aste on tarkemmissa luokissa. Suojauksessa vaihto rajoitettu kahden yksikön päähän. (Alkuperäistä aineistoa on merkitty vaihtumattomuustodennäköisyydellä 1.)

	Vaihtumattomuustodennäköisyys										
Aste	1	0,97	0,95	0,93	0,9	0,85	0,8	0,75	0,7	0,65	0,6
01	1	2	3	5	9	18	25	26	31	28	29
02	10	15	12	14	20	23	19	26	41	35	36
03	343	376	417	431	469	547	605	651	735	789	904
04	7	50	88	150	218	1287	368	483	559	660	778
05	6672	6475	6377	6182	6047	5717	5382	5114	4748	4485	4083
06	686	711	736	763	767	849	899	946	995	1031	1100
07	22	97	99	175	194	276	408	449	575	654	728
08	7	13	15	27	25	30	40	54	61	68	92
09	7	6	8	8	7	8	8	3	10	10	9
10	16	17	16	17	17	18	16	20	19	17	11
11	3	4	3	4	4	7	4	6	5	4	13
12	24	22	24	22	21	18	24	20	19	17	15

Taulukko 16: PRAM-menetelmän vaikutus frekvenssihin, kun aine on tarkemmissa luokissa. Suojaus ilman vaihdon rajoittamista. (Alkuperäistä aineistoa on merkitty vaihtumattomuustodennäköisyydellä 1.)

	Vaihtumattomuustodennäköisyys										
Aine	1	0,97	0,95	0,93	0,9	0,85	0,8	0,75	0,7	0,65	0,6
04	7	11	15	18	15	39	35	44	78	66	77
05	777	766	740	731	716	688	660	620	608	562	542
06	66	65	82	74	74	81	89	97	105	98	123
07	1	6	4	20	15	23	34	48	47	58	79
08	3	8	7	15	26	24	39	43	62	69	65
09	2	6	6	16	16	32	30	47	46	67	75
10	1	5	10	10	15	39	35	49	68	52	73
11	18	23	28	32	31	48	53	57	67	77	74
16	672	648	646	632	604	607	579	565	513	483	476
17	67	67	70	77	74	81	84	84	106	112	116
18	2	7	7	12	21	25	29	40	50	56	77
19	881	870	855	828	831	776	732	701	670	628	580
20	354	353	338	335	333	326	295	298	303	278	290
21	203	204	197	199	194	195	196	201	181	178	190
22	56	64	61	64	74	78	78	83	86	94	107
23	53	53	61	67	65	75	76	81	95	96	89
24	15	20	27	24	30	36	50	55	68	81	69
25	10	14	20	18	22	31	32	49	62	78	81
26	17	22	30	26	30	42	60	52	66	64	74
27	1194	1156	1146	1112	1088	1039	1008	927	904	798	781
28	316	306	314	303	311	293	293	286	281	265	273
29	151	151	152	150	152	154	142	171	152	148	152
30	115	118	113	120	113	127	126	135	129	140	144
31	408	404	409	392	387	375	371	353	341	324	307
33	375	365	364	356	369	345	335	320	294	300	271
34	12	15	24	18	29	43	53	39	55	66	76
35	1	6	10	15	22	26	23	46	56	62	75
36	9	15	14	17	31	38	37	51	58	58	68
37	630	613	601	609	586	568	532	528	476	491	440
38	2	9	6	17	16	31	52	53	55	59	68
40	225	222	218	229	218	220	226	213	207	208	208
41	5	9	16	17	28	26	49	58	49	86	75
42	190	189	197	198	196	191	179	177	176	187	193
43	203	203	202	198	196	187	197	193	189	192	201
44	334	330	329	328	313	298	301	284	275	292	277
45	9	15	13	27	20	33	45	42	54	54	73
46	9	13	17	31	26	37	41	58	70	66	70
47	2	9	11	16	17	29	43	51	56	73	60
49	1	7	7	14	12	25	29	44	45	78	78
50	49	58	56	54	64	58	62	78	78	103	76
51	208	211	205	200	207	180	196	193	189	191	192
52	45	52	50	52	60	77	72	86	80	104	98
54	4	5	15	15	22	22	56	54	61	59	79
99	96	105	105	112	129	130	144	144	187	197	206

Taulukko 17: PRAM-menetelmän vaikutus frekvenssihin, kun aine on tarkemmissa luokissa. Suojauksessa vaihto rajoitettu kahden yksikön päähän. (Alkuperäistä aineistoa on merkitty vaihtumattomuustodennäköisyydellä 1.)

	Vaihtumattomuustodennäköisyys										
Aine	1	0,97	0,95	0,93	0,9	0,85	0,8	0,75	0,7	0,65	0,6
04	7	18	24	21	36	44	60	84	85	93	124
05	777	752	734	731	694	655	629	577	521	517	487
06	66	71	72	83	83	110	101	119	133	148	138
07	1	9	21	16	37	40	57	67	107	86	99
08	3	4	3	3	4	5	7	8	7	11	4
09	2	2	2	4	2	3	3	1	4	5	4
10	1	7	7	16	26	23	37	45	62	55	69
11	18	22	25	35	35	40	46	64	70	65	86
16	672	655	649	613	596	574	546	509	465	454	435
17	67	77	76	88	110	120	128	148	161	174	182
18	2	8	17	43	51	67	82	108	141	199	185
19	881	864	855	832	801	765	764	722	672	627	572
20	354	354	354	352	340	362	336	338	345	310	347
21	203	207	210	212	220	229	229	230	232	256	263
22	56	58	58	60	74	68	80	76	97	87	105
23	53	55	52	53	52	56	54	58	53	66	55
24	15	15	17	14	15	20	20	23	17	20	22
25	10	15	29	35	44	47	57	82	117	108	148
26	17	34	30	35	46	66	91	130	139	170	162
27	1194	1169	1148	1120	1085	1049	982	928	866	790	760
28	316	310	315	323	332	322	317	325	316	337	323
29	151	157	169	179	180	202	229	225	258	295	270
30	115	124	129	122	134	141	160	156	180	186	201
31	408	401	396	386	399	378	359	348	351	310	337
33	375	367	355	362	337	341	332	319	286	301	282
34	12	18	23	33	32	38	46	61	68	73	82
35	1	12	17	18	23	37	62	69	75	88	107
36	9	12	19	15	26	31	39	49	49	66	73
37	630	609	600	605	571	535	523	493	470	437	376
38	2	11	11	16	32	40	37	43	66	78	100
40	225	223	222	212	216	232	218	222	215	204	229
41	5	8	12	16	28	23	30	39	44	51	60
42	190	191	187	192	194	190	189	200	187	204	179
43	203	207	201	203	188	191	195	188	186	180	184
44	334	326	323	315	301	304	282	264	273	249	252
45	9	10	21	16	29	32	39	43	49	57	51
46	9	11	10	14	19	13	24	32	28	32	39
47	2	4	2	3	2	2	5	6	4	5	13
49	1	1	2	5	6	9	9	25	19	22	25
50	49	47	55	46	51	53	51	51	56	60	64
51	208	202	198	198	193	177	188	159	152	139	142
52	45	50	40	50	47	50	39	48	52	56	39
54	4	6	11	6	22	22	21	30	29	40	44
99	96	95	97	97	85	92	95	86	91	87	79

Taulukko 18: PRAM-menetelmän vaikutus frekvenssihin. Suojausta on tarkasteltu verraten rajoittamatonta, kahden luokan päähän rajoitettua ja ryhmien sisälle rajoitettua vaihtelua. Osa 1.

Aine	Alkuperäinen	Vaihtumattomuustodennäköisyys 0,97		
		ilman rajaa	rajana 2	ryhmitelty
05	777	766	752	781
06	66	65	71	62
07	1	6	9	2
08	3	8	4	2
09	2	6	2	2
10	1	5	7	1
11	18	23	22	18
16	672	648	655	672
17	67	67	77	67
18	2	7	8	14
19	881	870	864	855
20	354	353	354	365
21	203	204	207	204
22	56	64	58	57
23	53	53	55	53
24	15	20	15	17
25	10	14	15	9
26	17	22	34	17
27	1194	1156	1169	1199
28	316	306	310	310
29	151	151	157	153
30	115	118	124	113
31	408	404	401	409

Taulukko 19: PRAM-menetelmän vaikutus frekvenssihin. Suojausta on tarkasteltu verraten rajoittamatonta, kahden luokan päähän rajoitettua ja ryhmien sisälle rajoitettua vaihtelua. Osa 2.

Aine	Alkuperäinen	Vaihtumattomuustodennäköisyys 0,95		
		ilman rajaa	rajana 2	ryhmitelty
05	777	740	734	779
06	66	82	72	64
07	1	4	21	1
08	3	7	3	3
09	2	6	2	2
10	1	10	7	1
11	18	28	25	18
16	672	646	649	676
17	67	70	76	60
18	2	7	17	23
19	881	855	855	849
20	354	338	354	364
21	203	197	210	205
22	56	61	58	56
23	53	61	52	56
24	15	27	17	12
25	10	20	29	12
26	17	30	30	17
27	1194	1146	1148	1202
28	316	314	315	305
29	151	152	169	155
30	115	113	129	112
31	408	409	396	410

Taulukko 20: PRAM-menetelmän vaikutus frekvenssihin. Suojausta on tarkasteltu verraten rajoittamatonta, kahden luokan päähän rajoitettua ja ryhmien sisälle rajoitettua vaihtelua. Osa 3.

Aine	Alkuperäinen	Vaihtumattomuustodennäköisyys 0,93		
		ilman rajaa	rajana 2	ryhmitelty
05	777	731	731	780
06	66	74	83	60
07	1	20	16	4
08	3	15	3	2
09	2	16	4	2
10	1	10	16	2
11	18	32	35	18
16	672	632	613	674
17	67	77	88	61
18	2	12	43	40
19	881	828	832	827
20	354	335	352	376
21	203	199	212	197
22	56	64	60	58
23	53	67	53	53
24	15	24	14	17
25	10	18	35	8
26	17	26	35	19
27	1194	1112	1120	1206
28	316	303	323	295
29	151	150	179	160
30	115	120	122	111
31	408	392	386	412

Taulukko 21: PRAM-menetelmän vaikutus frekvenssihin. Suojattu sekä rajoittamatta että rajoittamalla vaihtelua. Osa 1. (Alkuperäistä aineistoa on merkitty vaihtumattomuustodennäköisyydellä 1.)

Tn	Ikä			Aine1		
		Ilman rajaa	Raja 2		Ilman rajaa	Rajana 2
	alle 40	2467		äidinkieli	868	
	40–53	2849		kielet	2330	
	yli 53	2482		matemaattiset	2184	
1				historia yms.	1335	
				taide yms.	748	
				muut	333	
	alle 40	2505	2469	äidinkieli	879	867
	40–53	2817	2853	kielet	2322	2342
	yli 53	2476	2476	matemaattiset	2135	2161
0,97				historia yms.	1349	1334
				taide yms.	766	750
				muut	347	344
	alle 40	2507	2464	äidinkieli	877	864
	40–53	2803	2855	kielet	2312	2347
	yli 53	2488	2479	matemaattiset	2134	2157
0,95				historia yms.	1345	1329
				taide yms.	776	746
				muut	354	355
	alle 40	2516	2472	äidinkieli	898	888
	40–53	2803	2846	kielet	2282	2337
	yli 53	2479	2480	matemaattiset	2077	2130
0,93				historia yms.	1353	1349
				taide yms.	812	748
				muut	376	346
	alle 40	2537	2461	äidinkieli	893	881
	40–53	2798	2857	kielet	2278	2349
	yli 53	2463	2480	matemaattiset	2051	2130
0,90				historia yms.	1406	1334
				taide yms.	780	739
				muut	390	365
	alle 40	2544	2467	äidinkieli	935	876
	40–53	2728	2852	kielet	2272	2374
	yli 53	2526	2479	matemaattiset	1988	2092
0,85				historia yms.	1388	1324
				taide yms.	800	741
				muut	415	391

Taulukko 22: PRAM-menetelmän vaikutus frekvenssihin. Suojattu sekä rajoittamatta että rajoittamalla vaihtelua. Osa 2.

Tn	Ikä			Aine1		
		Ilman rajaa	Raja 2		Ilman rajaa	Rajana 2
	alle 40	2562	2462	äidinkieli	940	880
	40–53	2781	2857	kielet	2211	2387
	yli 53	2455	2479	matemaattiset	1940	2047
0,80				historia yms.	1418	1338
				taide yms.	835	743
				muut	454	403
	alle 40	2645	2465	äidinkieli	961	881
	40–53	2712	2862	kielet	2209	2424
	yli 53	2441	2471	matemaattiset	1872	1982
0,75				historia yms.	1448	1322
				taide yms.	849	758
				muut	459	431
	alle 40	2647	2458	äidinkieli	1003	904
	40–53	2669	2857	kielet	2200	2439
	yli 53	2482	2483	matemaattiset	1807	1971
0,70				historia yms.	1402	1303
				taide yms.	865	746
				muut	521	435
	alle 40	2683	2471	äidinkieli	983	887
	40–53	2601	2844	kielet	2148	2471
	yli 53	2514	2483	matemaattiset	1675	1918
0,65				historia yms.	1493	1338
				taide yms.	942	749
				muut	557	435
	alle 40	2685	2465	äidinkieli	1031	887
	40–53	2663	2843	kielet	2149	2476
	yli 53	2450	2490	matemaattiset	1657	1891
0,60				historia yms.	1443	1309
				taide yms.	952	743
				muut	566	492

Taulukko 23: PRAM-menetelmän vaikutus kahden muuttujan ristiintaulukoituihin frekvensseihin. Suojattu rajoittamatta vaihtumista. Osa 1. (Alkuperäistä aineistoa on merkitty vaihtumattomuustodennäköisyydellä 1.)

Tn	Aine	Ikä		
		alle 40 v.	41–53 v.	yli 53 v.
1	äidinkieli	306	316	246
	kielet	697	853	780
	matemaattiset	654	769	761
	historia yms.	429	494	412
	taide yms.	278	282	188
	muut	103	135	95
0,97	äidinkieli	312	313	254
	kielet	699	849	774
	matemaattiset	656	733	746
	historia yms.	439	501	409
	taide yms.	286	285	195
	muut	113	136	98
0,95	äidinkieli	291	316	270
	kielet	706	837	796
	matemaattiset	661	728	745
	historia yms.	450	485	410
	taide yms.	289	289	198
	muut	110	148	96
0,93	äidinkieli	319	322	257
	kielet	717	817	748
	matemaattiset	644	704	729
	historia yms.	425	515	413
	taide yms.	302	295	215
	muut	109	150	117
0,90	äidinkieli	322	308	263
	kielet	702	835	741
	matemaattiset	633	720	698
	historia yms.	465	503	438
	taide yms.	281	284	215
	muut	134	148	108

Taulukko 24: PRAM-menetelmän vaikutus kahden muuttujan ristiintaulukoituihin frekvensseihin. Suojattu rajoittamatta vaihtumista. Osa 2.

Tn	Aine	Ikä		
		alle 40 v.	41–53 v.	yli 53 v.
0,85	äidinkieli	333	323	279
	kielet	708	786	778
	matemaattiset	623	686	679
	historia yms.	469	493	426
	taide yms.	283	288	229
	muut	128	152	135
0,80	äidinkieli	325	355	260
	kielet	704	794	713
	matemaattiset	624	667	649
	historia yms.	468	510	440
	taide yms.	291	292	252
	muut	150	163	141
0,75	äidinkieli	359	330	272
	kielet	741	777	691
	matemaattiset	608	634	630
	historia yms.	501	492	455
	taide yms.	297	314	238
	muut	139	165	155
0,70	äidinkieli	369	339	295
	kielet	727	747	726
	matemaattiset	588	616	603
	historia yms.	476	484	442
	taide yms.	306	308	251
	muut	181	175	165
0,65	äidinkieli	367	320	296
	kielet	708	709	731
	matemaattiset	576	529	570
	historia yms.	524	520	449
	taide yms.	325	320	297
	muut	183	203	171
0,60	äidinkieli	360	360	311
	kielet	748	703	698
	matemaattiset	569	538	550
	historia yms.	512	488	443
	taide yms.	315	361	276
	muut	181	213	172

Taulukko 25: PRAM-menetelmän vaikutus kahden muuttujan ristiintaulukoituihin frekvensseihin. Suojattu rajoittamalla arvojen vaihtuminen korkeintaan kahden arvon päähän. Osa 1. (Alkuperäistä aineistoa on merkitty vaihtumattomuustodennäköisyydellä 1.)

Tn	Aine	Ikä		
		alle 40 v.	41–53 v.	yli 53 v.
1	äidinkieli	306	316	246
	kielet	697	853	780
	matemaattiset	654	769	761
	historia yms.	429	494	412
	taide yms.	278	282	188
	muut	103	135	95
0,97	äidinkieli	304	318	245
	kielet	703	862	777
	matemaattiset	646	763	752
	historia yms.	430	491	413
	taide yms.	277	282	191
	muut	109	137	98
0,95	äidinkieli	305	316	243
	kielet	700	861	786
	matemaattiset	646	758	753
	historia yms.	425	495	409
	taide yms.	276	285	185
	muut	112	140	103
0,93	äidinkieli	314	322	252
	kielet	698	853	786
	matemaattiset	644	746	740
	historia yms.	425	505	419
	taide yms.	277	284	187
	muut	114	136	96
0,90	äidinkieli	305	321	255
	kielet	699	864	786
	matemaattiset	635	745	750
	historia yms.	431	497	406
	taide yms.	279	276	184
	muut	112	154	99

Taulukko 26: PRAM-menetelmän vaikutus kahden muuttujan ristiintaulukoituihin frekvensseihin. Suojattu rajoittamalla arvojen vaihtuminen korkeintaan kahden arvon päähän. Osa 2.

Tn	Aine	Ikä		
		alle 40 v.	41–53 v.	yli 53 v.
0,85	äidinkieli	315	316	246
	kielet	712	853	780
	matemaattiset	654	769	761
	historia yms.	429	494	412
	taide yms.	278	282	188
	muut	103	135	95
0,80	äidinkieli	310	323	247
	kielet	713	870	804
	matemaattiset	601	723	723
	historia yms.	443	492	403
	taide yms.	275	286	182
	muut	120	163	120
0,75	äidinkieli	303	323	255
	kielet	714	891	819
	matemaattiset	601	699	682
	historia yms.	415	497	410
	taide yms.	274	291	193
	muut	158	161	112
0,70	äidinkieli	326	317	261
	kielet	724	883	832
	matemaattiset	580	708	683
	historia yms.	423	482	398
	taide yms.	277	279	190
	muut	128	188	119
0,65	äidinkieli	310	316	261
	kielet	722	907	842
	matemaattiset	581	675	662
	historia yms.	448	491	399
	taide yms.	279	277	193
	muut	131	178	126
0,60	äidinkieli	315	318	254
	kielet	712	907	857
	matemaattiset	576	652	663
	historia yms.	428	487	394
	taide yms.	271	278	194
	muut	163	201	128

Taulukko 27: Pätevyyden tarkastelua. Pätevyyden ja ikäluokkien ristiintaulukointi. Suojattu PRAM-menetelmällä ilman vaihdon rajoittamista. Osa 1. (Alkuperäistä aineistoa on merkitty muuttumattomuus todennäköisyydellä 1.)

Tn	Ikä	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
1	alle 40 v.	2128	339	2467
		86,26	13,74	100,00
	40–53 v.	2666	183	2849
		93,58	6,42	100,00
	yli 53 v.	2397	85	2482
		96,58	3,42	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	195,0332	p<0,0001	
	LR	190,7357	p<0,0001	
0,97	alle 40 v.	2165	340	2505
		86,43	13,57	100,00
	40–53 v.	2635	182	2817
		93,54	6,46	100,00
	yli 53 v.	2391	85	2476
		96,57	3,43	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	χ^2	189,1182	p<0,0001	
	LR	185,9195	p<0,0001	
0,95	alle 40 v.	2175	332	2507
		86,76	13,24	100,00
	40–53 v.	2619	184	2803
		93,44	6,56	100,00
	yli 53 v.	2397	91	2488
		96,34	3,66	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	168,9045	p<0,0001	
	LR	165,9098	p<0,0001	
0,93	alle 40 v.	2181	335	2516
		86,69	13,31	100,00
	40–53 v.	2625	178	2803
		93,65	6,35	100,00
	yli 53 v.	2385	94	2479
		96,21	3,79	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	170,2853	p<0,0001	
	LR	165,3345	p<0,0001	

Taulukko 28: Pätevyyden tarkastelua. Pätevyyden ja ikäluokkien ristiintaulukointi. Suojattu PRAM-menetelmällä ilman vaihdon rajoittamista. Osa 2.

Tn	Ikä	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
	alle 40 v.	2212	325	2537
		87,19	12,81	100,00
	40–53 v.	2622	176	2798
0,90		93,71	6,29	100,00
	yli 53 v.	2357	106	2463
		95,70	4,30	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	χ^2	139,5535	p<0,0001	
	LR	134,1304	p<0,0001	
	alle 40 v.	2225	319	2544
		87,46	12,54	100,00
	40–53 v.	2551	177	2728
0,85		93,51	6,49	100,00
	yli 53 v.	2415	111	2526
		95,61	4,39	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	126,9569	p<0,0001	
	LR	122,7565	p<0,0001	
	alle 40 v.	2241	321	2562
		87,47	12,53	100,00
	40–53 v.	2613	168	2781
0,80		93,96	6,04	100,00
	yli 53 v.	2337	118	2455
		95,19	4,81	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	122,4603	p<0,0001	
	LR	116,2431	p<0,0001	
	alle 40 v.	2329	316	2645
		88,05	11,95	100,00
	40–53 v.	2522	190	2712
0,75		92,99	7,01	100,00
	yli 53 v.	2340	101	2441
		95,86	4,14	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	χ^2	111,3633	p<0,0001	
	LR	111,7773	p<0,0001	

Taulukko 29: Pätevyyden tarkastelua. Pätevyyden ja ikäluokkien ristiintaulukointi. Suojattu PRAM-menetelmällä ilman vaihdon rajoittamista. Osa 3.

Tn	Ikä	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
	alle 40 v.	2333	314	2647
		88,14	11,86	100,00
	40–53 v.	2482	187	2669
0,70		92,99	7,01	100,00
	yli 53 v.	2376	106	2482
		95,73	4,27	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	106,2664	p<0,0001	
	LR	106,2543	p<0,0001	
	alle 40 v.	2387	296	2683
		88,97	11,03	100,00
	40–53 v.	2416	185	2601
0,65		92,89	7,11	100,00
	yli 53 v.	2388	126	2514
		94,99	5,01	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	67,9878	p<0,0001	
	LR	67,4564	p<0,0001	
	alle 40 v.	2389	296	2685
		88,98	11,02	100,00
	40–53 v.	2480	183	2663
0,60		93,13	6,87	100,00
	yli 53 v.	2322	128	2450
		94,78	5,22	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	χ^2	64,7174	p<0,0001	
	LR	63,3974	p<0,0001	

Taulukko 30: Pätevyyden tarkastelua. Pätevyyden ja ikäluokkien ristiintaulukointi. Suojattu PRAM-menetelmällä, jossa arvojen vaihtuminen on rajoitettu korkeintaan kahden päähän. Osa 1. (Alkuperäistä aineistoa on merkitty muuttumattomuus todennäköisyydellä 1.)

Tn	Ikä	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
	alle 40 v.	2128	339	2467
		86,26	13,74	100,00
	40–53 v.	2666	183	2849
1		93,58	6,42	100,00
	yli 53 v.	2397	85	2482
		96,58	3,42	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	195,0332	p<0,0001	
	LR	190,7357	p<0,0001	
	alle 40 v.	2129	340	2469
		86,23	13,77	100,00
	40–53 v.	2670	183	2853
0,97		93,59	6,41	100,00
	yli 53 v.	2392	84	2476
		96,61	3,39	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	χ^2	197,2565	p<0,0001	
	LR	193,0461	p<0,0001	
	alle 40 v.	2126	338	2464
		86,28	13,72	100,00
	40–53 v.	2671	184	2855
0,95		93,56	6,44	100,00
	yli 53 v.	2394	85	2479
		96,37	3,43	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	193,4914	p<0,0001	
	LR	189,3547	p<0,0001	
	alle 40 v.	2133	339	2472
		86,29	13,71	100,00
	40–53 v.	2662	184	2846
0,93		93,53	6,47	100,00
	yli 53 v.	2396	84	2480
		96,61	3,39	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	194,7730	p<0,0001	
	LR	191,1367	p<0,0001	

Taulukko 31: Pätevyyden tarkastelua. Pätevyyden ja ikäluokkien ristiintaulukointi. Suojattu PRAM-menetelmällä, jossa arvojen vaihtuminen on rajoitettu korkeintaan kahden päähän. Osa 2.

Tn	Ikä	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
0,90	alle 40 v.	2122	339	2461
		86,23	13,77	100,00
	40–53 v.	2674	183	2857
		93,59	6,41	100,00
	yli 53 v.	2395	85	2480
		96,57	3,43	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	χ^2	196,1895	p<0,0001	
	LR	191,5906	p<0,0001	
0,85	alle 40 v.	2128	339	2467
		86,26	13,74	100,00
	40–53 v.	2666	186	2852
		93,48	6,52	100,00
	yli 53 v.	2397	82	2479
		96,69	3,31	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	197,5019	p<0,0001	
	LR	194,6586	p<0,0001	
0,80	alle 40 v.	2126	336	2462
		86,35	13,65	100,00
	40–53 v.	2670	187	2857
		93,45	6,55	100,00
	yli 53 v.	2395	84	2479
		96,61	3,39	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	190,7502	p<0,0001	
	LR	187,7288	p<0,0001	
0,75	alle 40 v.	2127	338	2465
		86,29	13,71	100,00
	40–53 v.	2676	186	2862
		93,50	6,50	100,00
	yli 53 v.	2388	83	2471
		96,64	3,36	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	χ^2	194,6645	p<0,0001	
	LR	191,3870	p<0,0001	

Taulukko 32: Pätevyyden tarkastelua. Pätevyyden ja ikäluokkien ristiintaulukointi. Suojattu PRAM-menetelmällä, jossa arvojen vaihtuminen on rajoitettu korkeintaan kahden päähän. Osa 3.

Tn	Ikä	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
	alle 40 v.	2116	342	2458
		86,09	13,91	100,00
	40–53 v.	2677	180	2857
0,70		93,70	6,30	100,00
	yli 53 v.	2398	85	2483
		96,58	3,42	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	203,2035	p<0,0001	
	LR	197,4497	p<0,0001	
	alle 40 v.	2135	336	2471
		86,40	13,60	100,00
	40–53 v.	2656	188	2844
0,65		93,39	6,61	100,00
	yli 53 v.	2400	83	2483
		96,66	3,34	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	190,0390	p<0,0001	
	LR	187,9958	p<0,0001	
	alle 40 v.	2132	333	2465
		86,49	13,51	100,00
	40–53 v.	2655	188	2843
0,60		93,39	6,61	100,00
	yli 53 v.	2404	86	2490
		96,55	3,45	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	χ^2	183,0343	p<0,0001	
	LR	180,4295	p<0,0001	

Taulukko 33: Pätevyyden tarkastelua. Pätevyyden ja virkaluokkien ristiintaulukointi. Suojattu PRAM-menetelmällä ilman vaihdon rajoittamista. Osa 1. (Alkuperäistä aineistoa on merkitty vaihtumattomuustodennäköisyydellä 1.)

Tn	Virka	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
	rehtorit	425	7	432
		98,38	1,62	100,00
	lehtorit	5519	165	5684
		97,10	2,90	100,00
	op.-ohjaajat	125	17	142
		88,03	11,97	100,00
1	päätoim. tuntiop.	875	203	1078
		81,17	18,83	100,00
	sivutoim. tuntiop.	208	206	414
		50,24	49,76	100,00
	muu virka	39	9	48
		81,25	18,75	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1422,4648	p<0,0001	
	LR	932,8533	p<0,0001	
	rehtorit	490	14	504
		97,22	2,78	100,00
	lehtorit	5365	163	5528
		97,05	2,95	100,00
	op.-ohjaajat	147	20	167
		88,02	11,98	100,00
0,97	päätoim. tuntiop.	891	199	1090
		81,74	18,26	100,00
	sivutoim. tuntiop.	215	198	413
		52,06	47,94	100,00
	muu virka	83	13	96
		86,46	13,54	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1300,5852	p<0,0001	
	LR	860,6174	p<0,0001	
	rehtorit	509	9	518
		98,26	1,74	100,00
	lehtorit	5271	162	5433
		97,02	2,98	100,00
	op.-ohjaajat	175	20	195
		89,74	10,26	100,00
0,95	päätoim. tuntiop.	894	205	1099
		81,35	18,65	100,00
	sivutoim. tuntiop.	232	197	429
		54,08	45,92	100,00
	muu virka	110	14	124
		88,71	11,29	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1254,8219	p<0,0001	
	LR	851,0785	p<0,0001	

Taulukko 34: Pätevyyden tarkastelua. Pätevyyden ja virkaluokkien ristiintaulukointi. Suojattu PRAM-menetelmällä ilman vaihdon rajoittamista. Osa 2.

Tn	Virka	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
	rehtorit	568	14	582
		97,59	2,41	100,00
	lehtorit	5190	160	5350
		97,01	2,99	100,00
	op.-ohjaajat	192	24	216
		88,89	11,11	100,00
0,93	päätoim. tuntiop.	873	194	1067
		81,82	18,18	100,00
	sivutoim. tuntiop.	228	199	427
		53,40	46,60	100,00
	muu virka	140	16	156
		89,74	10,26	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1256,5308	p<0,0001	
	LR	838,9814	p<0,0001	
	rehtorit	625	22	647
		96,60	3,40	100,00
	lehtorit	5066	159	5225
		96,96	3,04	100,00
	op.-ohjaajat	224	23	247
		90,69	9,31	100,00
0,90	päätoim. tuntiop.	883	194	1077
		81,99	18,01	100,00
	sivutoim. tuntiop.	243	196	439
		55,35	44,65	100,00
	muu virka	150	13	163
		92,02	7,98	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1169,7898	p<0,0001	
	LR	786,2263	p<0,0001	
	rehtorit	679	35	714
		95,10	4,90	100,00
	lehtorit	4817	161	4978
		96,77	3,23	100,00
	op.-ohjaajat	277	26	303
		91,42	8,58	100,00
0,85	päätoim. tuntiop.	913	184	1097
		82,23	16,77	100,00
	sivutoim. tuntiop.	270	180	450
		60,00	40,00	100,00
	muu virka	235	21	256
		91,80	8,20	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	926,2807	p<0,0001	
	LR	643,3084	p<0,0001	

Taulukko 35: Pätevyyden tarkastelua. Pätevyyden ja virkaluokkien ristiintaulukointi. Suojattu PRAM-menetelmällä ilman vaihdon rajoittamista. Osa 3.

Tn	Virka	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
	rehtorit	819	51	870
		94,14	5,86	100,00
	lehtorit	4562	150	4712
		96,82	3,18	100,00
	op.-ohjaajat	343	35	378
		90,74	9,26	100,00
0,80	päätoim. tuntiop.	931	181	1112
		83,72	16,28	100,00
	sivutoim. tuntiop.	281	166	447
		62,86	37,14	100,00
	muu virka	255	24	279
		91,40	8,60	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	793,0843	p<0,0001	
	LR	572,7200	p<0,0001	
	rehtorit	915	53	968
		94,52	5,48	100,00
	lehtorit	4307	133	4440
		97,00	3,00	100,00
	op.-ohjaajat	393	31	424
		92,69	7,31	100,00
0,75	päätoim. tuntiop.	945	177	1122
		84,22	15,78	100,00
	sivutoim. tuntiop.	314	175	489
		64,21	35,79	100,00
	muu virka	317	38	355
		89,30	10,70	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	787,4065	p<0,0001	
	LR	579,3317	p<0,0001	
	rehtorit	1005	74	1079
		93,14	6,86	100,00
	lehtorit	4036	132	4168
		96,83	3,17	100,00
	op.-ohjaajat	446	43	489
		91,21	8,79	100,00
0,70	päätoim. tuntiop.	984	156	1140
		86,32	13,68	100,00
	sivutoim. tuntiop.	312	161	473
		65,96	34,04	100,00
	muu virka	408	41	449
		90,87	9,13	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	636,3775	p<0,0001	
	LR	471,9289	p<0,0001	

Taulukko 36: Pätevyyden tarkastelua. Pätevyyden ja virkaluokkien ristiintaulukointi. Suojattu PRAM-menetelmällä ilman vaihdon rajoittamista. Osa 4.

Tn	Virka	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
	rehtorit	1116	87	1203
		92,77	7,23	100,00
	lehtorit	3836	134	3970
		96,62	3,38	100,00
	op.-ohjaajat	501	40	541
		92,61	7,39	100,00
0,65	päätoim. tuntiop.	940	156	1096
		85,77	14,23	100,00
	sivutoim. tuntiop.	341	152	493
		69,17	30,83	100,00
	muu virka	457	38	495
		92,32	7,68	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	536,4718	p<0,0001	
	LR	409,2997	p<0,0001	
	rehtorit	1203	88	1291
		93,18	6,82	100,00
	lehtorit	3608	140	3748
		96,26	3,74	100,00
	op.-ohjaajat	574	53	627
		91,55	8,45	100,00
0,60	päätoim. tuntiop.	949	145	1094
		86,75	13,25	100,00
	sivutoim. tuntiop.	338	130	468
		72,22	27,78	100,00
	muu virka	519	51	570
		91,05	8,95	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	394,9709	p<0,0001	
	LR	311,4750	p<0,0001	

Taulukko 37: Pätevyyden tarkastelua. Pätevyyden ja virkaluokkien ristiintaulukointi. Suojattu PRAM-menetelmällä, jossa arvojen vaihtuminen on rajoitettu korkeintaan kahden päähän. Osa 1. (Alkuperäistä aineistoa on merkitty todennäköisyydellä 1.)

Tn	Virka	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
1	rehtorit	425	7	432
		98,38	1,62	100,00
	lehtorit	5519	165	5684
		97,10	2,90	100,00
	op.-ohjaajat	125	17	142
		88,03	11,97	100,00
	päätoim. tuntiop.	875	203	1078
		81,17	18,83	100,00
	sivutoim. tuntiop.	208	206	414
		50,24	49,76	100,00
	muu virka	39	9	48
		81,25	18,75	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1422,4648	p<0,0001	
	LR	932,8533	p<0,0001	
0,97	rehtorit	427	7	434
		98,39	1,61	100,00
	lehtorit	5420	166	5586
		97,03	2,97	100,00
	op.-ohjaajat	187	19	206
		90,78	9,22	100,00
	päätoim. tuntiop.	910	200	1110
		81,88	18,02	100,00
	sivutoim. tuntiop.	198	201	399
		49,62	50,38	100,00
	muu virka	49	14	63
		77,78	22,22	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1392,4511	p<0,0001	
	LR	905,2113	p<0,0001	
0,95	rehtorit	427	9	434
		98,39	1,61	100,00
	lehtorit	5346	166	5512
		96,99	3,01	100,00
	op.-ohjaajat	216	19	235
		91,91	8,09	100,00
	päätoim. tuntiop.	946	198	1144
		82,69	17,31	100,00
	sivutoim. tuntiop.	202	198	400
		50,50	49,50	100,00
	muu virka	54	19	73
		73,97	26,03	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1346,0828	p<0,0001	
	LR	879,0836	p<0,0001	

Taulukko 38: Pätevyyden tarkastelua. Pätevyyden ja virkaluokkien ristiintaulukointi. Suojattu PRAM-menetelmällä, jossa arvojen vaihtuminen on rajoitettu korkeintaan kahden päähän. Osa 2.

Tn	Virka	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
	rehtorit	426	7	433
		98,38	1,62	100,00
	lehtorit	5297	161	5458
		97,05	2,95	100,00
	op.-ohjaajat	250	20	270
		92,59	7,41	100,00
0,93	päätoim. tuntiop.	949	204	1153
		82,31	17,69	100,00
	sivutoim. tuntiop.	197	197	394
		50,00	50,00	100,00
	muu virka	72	18	90
		80,00	20,00	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1355,3429	p<0,0001	
	LR	886,4637	p<0,0001	
	rehtorit	422	8	430
		98,14	1,86	100,00
	lehtorit	5137	168	5305
		96,83	3,17	100,00
	op.-ohjaajat	364	18	382
		95,29	4,71	100,00
0,90	päätoim. tuntiop.	1007	207	1214
		82,95	17,05	100,00
	sivutoim. tuntiop.	191	186	377
		50,66	49,34	100,00
	muu virka	70	20	90
		77,78	22,22	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1261,8161	p<0,0001	
	LR	822,7987	p<0,0001	
	rehtorit	432	8	440
		98,18	1,82	100,00
	lehtorit	4928	161	5089
		96,84	3,16	100,00
	op.-ohjaajat	481	22	503
		95,63	4,37	100,00
0,85	päätoim. tuntiop.	1070	207	1277
		83,79	16,21	100,00
	sivutoim. tuntiop.	189	187	376
		50,27	49,73	100,00
	muu virka	91	22	113
		80,53	19,47	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1250,9146	p<0,0001	
	LR	810,9724	p<0,0001	

Taulukko 39: Pätevyyden tarkastelua. Pätevyyden ja virkaluokkien ristiintaulukointi. Suojattu PRAM-menetelmällä, jossa arvojen vaihtuminen on rajoitettu korkeintaan kahden päähän. Osa 3.

Tn	Virka	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
	rehtorit	427	8	435
		98,16	1,84	100,00
	lehtorit	4756	163	4919
		96,69	3,31	100,00
	op.-ohjaajat	624	28	652
		95,71	4,29	100,00
0,80	päätoim. tuntiop.	1088	198	1286
		84,60	15,40	100,00
	sivutoim. tuntiop.	175	173	348
		50,29	49,71	100,00
	muu virka	121	37	158
		76,58	23,42	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1179,3369	p<0,0001	
	LR	763,6763	p<0,0001	
	rehtorit	433	7	440
		98,41	1,59	100,00
	lehtorit	4627	157	4784
		96,72	3,28	100,00
	op.-ohjaajat	668	29	697
		95,84	4,16	100,00
0,75	päätoim. tuntiop.	1172	212	1384
		84,68	15,32	100,00
	sivutoim. tuntiop.	174	161	335
		51,94	48,06	100,00
	muu virka	117	41	158
		74,05	25,95	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1110,4632	p<0,0001	
	LR	740,1859	p<0,0001	
	rehtorit	433	9	442
		97,96	2,04	100,00
	lehtorit	4349	144	4493
		96,80	3,20	100,00
	op.-ohjaajat	809	35	844
		95,85	4,15	100,00
0,70	päätoim. tuntiop.	1275	209	1484
		85,92	14,08	100,00
	sivutoim. tuntiop.	165	161	326
		50,61	49,39	100,00
	muu virka	160	49	209
		76,56	23,44	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	1106,6336	p<0,0001	
	LR	725,5597	p<0,0001	

Taulukko 40: Pätevyyden tarkastelua. Pätevyyden ja virkaluokkien ristiintaulukointi. Suojattu PRAM-menetelmällä, jossa arvojen vaihtuminen on rajoitettu korkeintaan kahden päähän. Osa 4.

Tn	Virka	Puuttuvia pätevyyksiä		
		ei	on	yhteensä
	rehtorit	431	8	439
		98,18	1,82	100,00
	lehtorit	4252	158	4410
		96,42	3,58	100,00
	op.-ohjaajat	883	30	913
		96,71	3,29	100,00
0,65	päätoim. tuntiop.	1326	214	1540
		86,10	13,90	100,00
	sivutoim. tuntiop.	165	142	307
		53,75	46,25	100,00
	muu virka	134	55	189
		70,90	29,10	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	988,6538	p<0,0001	
	LR	665,5304	p<0,0001	
	rehtorit	435	8	443
		98,19	1,81	100,00
	lehtorit	4069	147	4216
		96,51	3,49	100,00
	op.-ohjaajat	1007	41	1048
		96,09	3,91	100,00
0,60	päätoim. tuntiop.	1357	212	1569
		86,49	13,51	100,00
	sivutoim. tuntiop.	148	126	274
		54,01	45,99	100,00
	muu virka	175	73	248
		70,56	29,44	100,00
	yhteensä	7191	607	7798
		92,22	7,78	100,00
	chi-square	943,1323	p<0,0001	
	LR	641,8620	p<0,0001	

Liite 2

Taulukko 1: Mallin käyttäytyminen aineistoa suojatessa. Tutkitaan viran vaikutusta pätevyyden arvoon. Suojattu käyttäen mikroaggregointia. Osa 1. (Alkuperäistä aineistoa on merkitty ryhmäkoolla 1.)

Koko	Muuttuja	Estimaatit	s.e.	95% lv.		t-arvo	p-arvo
	vakio	0,0348	0,2419	-0,4388	0,5102	0,0207	0,8857
	ikä	0,0419	0,0419	0,0324	0,0515	74,0574	<0,0001
	sukupuoli	0,3469	0,0982	0,1537	0,5388	12,4816	0,0004
1	virkaluokka 1	1,8713	0,3303	1,2838	2,5985	32,0911	<0,0001
	virkaluokka 4	1,3539	0,1212	1,1091	1,5860	124,8486	<0,0001
	virkaluokka 5	-0,0963	0,2365	-0,5420	0,3912	0,1658	0,6839
	virkaluokka 6	-0,4467	0,1235	-0,6959	-0,2102	13,0916	0,0003
	virkaluokka 8	-0,7627	0,3235	-1,3629	-0,0801	5,5573	0,0184
	vakio	-0,0627	0,2410	-0,5347	0,4104	0,0678	0,7946
	ikä	0,0427	0,0049	0,0333	0,0523	77,6357	<0,0001
	sukupuoli	0,3498	0,0980	0,1570	0,5413	12,1135	0,0004
2	virkaluokka 1	1,9122	0,3289	1,3283	2,6372	33,0001	<0,0001
	virkaluokka 4	1,4054	0,1173	1,1689	1,6307	143,0000	<0,0001
	virkaluokka 5	-0,1310	0,2301	-0,5650	0,3425	0,0017	0,5692
	virkaluokka 6	-0,3704	0,1203	-0,6125	-0,1391	9,0023	0,0021
	virkaluokka 8	-0,9806	0,2908	-1,5285	-0,3787	11,0035	0,0007
	vakio	-0,0620	0,2388	-0,5299	0,4066	0,0673	0,7952
	ikä	0,0426	0,0049	0,0331	0,0523	76,3710	<0,0001
	sukupuoli	0,3386	0,0980	0,1457	0,5301	11,9299	0,0006
3	virkaluokka 1	1,6708	0,2924	1,1454	2,3057	32,6528	<0,0001
	virkaluokka 4	1,4163	0,1114	1,1933	1,6313	161,6339	<0,0001
	virkaluokka 5	0,1069	0,2425	-0,3449	0,6123	0,1942	0,6594
	virkaluokka 6	-0,3611	0,1147	-0,5904	-0,1398	9,9140	0,0016
	virkaluokka 8	-0,9998	0,2542	-1,4818	-0,4791	15,4722	<0,0001
	vakio	0,0113	0,2395	-0,4576	0,4815	0,0022	0,9622
	ikä	0,0419	0,0049	0,0324	0,0515	74,2536	<0,0001
	sukupuoli	0,3452	0,0979	0,1525	0,5364	12,4312	0,0004
4	virkaluokka 1	1,8607	0,3284	1,2781	2,5849	32,1102	<0,0001
	virkaluokka 4	1,3679	0,1157	1,1347	1,5901	139,7414	<0,0001
	virkaluokka 5	0,1749	0,2500	-0,2899	0,6978	0,4895	0,4842
	virkaluokka 6	-0,4156	0,1186	-0,6543	-0,1876	12,2731	0,0005
	virkaluokka 8	-1,1244	0,2543	-1,6098	-0,6069	19,5466	<0,0001

Taulukko 2: Mallin käyttäytyminen aineistoa suojatessa. Tutkitaan viran vaikutusta pätevyyden arvoon. Suojattu käyttäen mikroaggregointia. Osa 2.

Koko	Muuttuja	Estimaatit	s.e.	95% lv.		t-arvo	p-arvo
	vakio	0,0178	0,2440	-0,4599	0,4972	0,0053	0,9419
	ikä	0,0418	0,0049	0,0323	0,0515	73,4749	<0,0001
	sukupuoli	0,3463	0,0979	0,1537	0,5376	12,5208	0,0004
5	virkaluokka 1	1,8769	0,3303	1,2894	2,6040	32,2853	<0,0001
	virkaluokka 4	1,3568	0,1210	1,1123	1,5887	125,6418	<0,0001
	virkaluokka 5	-0,0425	0,2358	-0,4865	0,4438	0,0324	0,8571
	virkaluokka 6	-0,4407	0,1240	-0,6908	-0,2031	12,6383	0,0004
	virkaluokka 8	-0,8724	0,3262	-1,4790	-0,1856	7,1527	0,0075
	vakio	0,0396	0,2412	-0,4324	0,5136	0,0270	0,8694
	ikä	0,0417	0,0049	0,0322	0,0513	73,0562	<0,0001
	sukupuoli	0,3441	0,0981	0,1510	0,5359	12,2973	0,0005
6	virkaluokka 1	1,8623	0,3302	1,2752	2,5893	31,8109	<0,0001
	virkaluokka 4	1,3506	0,1200	1,1082	1,5804	126,7487	<0,0001
	virkaluokka 5	-0,1754	0,2214	-0,5952	0,2777	0,6272	0,4284
	virkaluokka 6	-0,4465	0,1219	-0,6927	-0,2129	13,4083	0,0003
	virkaluokka 8	-0,6556	0,3233	-1,2551	0,0266	4,1131	0,0426
	vakio	0,0131	0,2369	-0,4510	0,4783	0,0031	0,9558
	ikä	0,0424	0,0049	0,0329	0,0520	75,3173	<0,0001
	sukupuoli	0,3504	0,0979	0,1577	0,5417	12,8079	0,0003
7	virkaluokka 1	1,7098	0,3091	1,1569	2,3847	30,5951	<0,0001
	virkaluokka 4	1,3392	0,1136	1,1113	1,5578	139,0865	<0,0001
	virkaluokka 5	0,0487	0,2367	-0,3938	0,5403	0,0423	0,8370
	virkaluokka 6	-0,4587	0,1158	-0,6911	-0,2357	15,6827	<0,0001
	virkaluokka 8	-0,7079	0,2653	-1,2072	-0,1596	7,1187	0,0076
	vakio	-0,0756	0,2376	-0,5411	0,3909	0,1011	0,7505
	ikä	0,0437	0,0049	0,0342	0,0534	79,5758	<0,0001
	sukupuoli	0,3544	0,0980	0,1617	0,5458	13,0907	0,0003
8	virkaluokka 1	1,8536	0,3272	1,2740	2,5761	32,0973	<0,0001
	virkaluokka 4	1,3620	0,1115	1,1372	1,5763	149,1579	<0,0001
	virkaluokka 5	0,0407	0,2361	-0,4006	0,5312	0,0298	0,8631
	virkaluokka 6	-0,3706	0,1149	-0,6017	-0,1498	10,4142	0,0013
	virkaluokka 8	-0,9859	0,2166	-1,4018	-0,5488	20,7104	<0,0001

Taulukko 3: Mallin käyttäytyminen aineistoa suojatessa. Tutkitaan viran vaikutusta pätevyyden arvoon. Suojattu käyttäen mikroaggregointia. Osa 3.

Koko	Muuttuja	Estimaatit	s.e.	95% lv.		t-arvo	p-arvo
	vakio	0,0745	0,2384	-0,3920	0,5431	0,0976	0,7548
	ikä	0,0419	0,0049	0,0324	0,0515	73,7112	<0,0001
	sukupuoli	0,3651	0,0976	0,1730	0,5559	13,9797	0,0002
9	virkaluokka 1	1,9553	0,3522	1,3368	2,7420	30,8186	<0,0001
	virkaluokka 4	1,2880	0,1168	1,0509	1,5113	121,5384	<0,0001
	virkaluokka 5	-0,0306	0,2530	-0,5025	0,4970	0,0146	0,9037
	virkaluokka 6	-0,5235	0,1189	-0,7647	-0,2962	19,3793	<0,0001
	virkaluokka 8	-0,7547	0,2332	-1,1984	-0,2790	10,4718	0,0012
	vakio	0,0913	0,2391	-0,3768	0,5610	0,1459	0,7025
	ikä	0,0416	0,0049	0,0321	0,0512	72,7097	<0,0001
	sukupuoli	0,3561	0,0976	0,1641	0,5468	13,3175	0,0003
10	virkaluokka 1	1,7914	0,3277	1,2105	2,5145	29,8861	<0,0001
	virkaluokka 4	1,2748	0,1133	1,0462	1,4921	126,6574	<0,0001
	virkaluokka 5	0,0498	0,2498	-0,4146	0,5722	0,0397	0,8421
	virkaluokka 6	-0,5313	0,1159	-0,7649	-0,3089	21,0100	<0,0001
	virkaluokka 8	-0,6154	0,2278	-1,0468	-0,1489	7,3018	0,0069
	vakio	-0,0183	0,2363	-0,4813	0,4453	0,0060	0,9381
	ikä	0,0426	0,0049	0,0331	0,0523	76,0225	<0,0001
	sukupuoli	0,3633	0,0976	0,1713	0,5539	13,8589	0,0002
11	virkaluokka 1	1,7331	0,3080	1,1830	2,4063	31,6632	<0,0001
	virkaluokka 4	1,3492	0,1102	1,1282	1,5616	149,9270	<0,0001
	virkaluokka 5	0,1128	0,2408	-0,3352	0,6155	0,2195	0,6394
	virkaluokka 6	-0,4721	0,1124	-0,6975	-0,2557	17,6547	<0,0001
	virkaluokka 8	-0,8539	0,2214	-1,2765	-0,4048	14,8806	0,0001
	vakio	-0,0151	0,2436	-0,4918	0,4638	0,0038	0,9507
	ikä	0,0422	0,0049	0,0327	0,0518	75,1455	<0,0001
	sukupuoli	0,3507	0,0976	0,1586	0,5414	12,9058	0,0003
15	virkaluokka 1	1,6436	0,2986	1,1024	2,2871	30,3044	<0,0001
	virkaluokka 4	1,3801	0,1259	1,1249	1,6210	120,0791	<0,0001
	virkaluokka 5	0,0208	0,2379	-0,4282	0,5105	0,0077	0,9302
	virkaluokka 6	-0,4466	0,1272	-0,7041	-0,2034	12,3311	0,0004
	virkaluokka 8	-0,7573	0,3953	-1,4831	0,0936	3,6695	0,0554

Taulukko 4: Mallin käyttäytyminen aineistoa suojatessa. Tutkitaan viran vaikutusta pätevyyden arvoon. Aineistossa muuttujan virka nimeäminen kuten alkuperäisessä aineistossa. Suojattu käyttäen mikroaggregointia. Osa 1. (Alkuperäistä aineistoa on merkitty ryhmäkoolla 1.)

Koko	Muuttuja	Estimaatit	s.e.	95% lv.		t-arvo	p-arvo
	vakio	0,0348	0,2419	-0,4388	0,5102	0,0207	0,8857
	ikä	0,0419	0,0419	0,0324	0,0515	74,0574	<0,0001
	sukupuoli	0,3469	0,0982	0,1537	0,5388	12,4816	0,0004
1	virkaluokka 1	1,8713	0,3303	1,2838	2,5985	32,0911	<0,0001
	virkaluokka 4	1,3539	0,1212	1,1091	1,5860	124,8486	<0,0001
	virkaluokka 5	-0,0963	0,2365	-0,5420	0,3912	0,1658	0,6839
	virkaluokka 6	-0,4467	0,1235	-0,6959	-0,2102	13,0916	0,0003
	virkaluokka 8	-0,7627	0,3235	-1,3629	-0,0801	5,5573	0,0184
	vakio	1,4200	38,0925	-0,5254		0,0014	0,9703
	ikä	0,0423	0,0049	0,0328	0,0519	75,6357	<0,0001
	sukupuoli	0,3415	0,0981	0,1484	0,5332	12,1135	0,0005
2	virkaluokka 1	0,4594	38,0932		2,5957	0,0001	0,9904
	virkaluokka 3	8,5361	228,6	-1,8228		0,0014	0,9702
	virkaluokka 4	-0,0476	38,0919		1,7331	0,0000	0,9990
	virkaluokka 5	-1,5577	38,0924		0,3748	0,0017	0,9674
	virkaluokka 6	-1,8323	38,0919		-0,0474	0,0023	0,9616
	virkaluokka 8	-2,2512	38,0931		-0,1713	0,0035	0,9529
	vakio	1,6339	42,6805	-0,3365		0,0015	0,9695
	ikä	0,0414	0,0049	0,0320	0,0511	72,6261	<0,0001
	sukupuoli	0,3499	0,0979	0,1573	0,5412	12,7791	0,0004
3	virkaluokka 1	0,2857	42,6811		2,4492	0,0000	0,9947
	virkaluokka 3	9,5087	256,1	-0,9294		0,0014	0,9704
	virkaluokka 4	-0,2340	42,6799		1,5676	0,0000	0,9956
	virkaluokka 5	-1,6243	42,6804		0,3404	0,0014	0,9696
	virkaluokka 6	-2,0376	42,6799		-0,2318	0,0023	0,9619
	virkaluokka 8	-2,4294	42,6812		-0,2620	0,0032	0,9546
	vakio	1,4885	38,9960	-0,4538		0,0015	0,9696
	ikä	0,0418	0,0049	0,0323	0,0513	74,3994	<0,0001
	sukupuoli	0,3589	0,0978	0,1665	0,5499	13,4754	0,0002
4	virkaluokka 1	0,4038	38,9967		2,5455	0,0001	0,9917
	virkaluokka 3	8,9898	234,0	-1,4114		0,0015	0,9694
	virkaluokka 4	-0,1197	38,9954		1,6665	0,0000	0,9976
	virkaluokka 5	-1,4755	38,9960		0,4882	0,0014	0,9698
	virkaluokka 6	-1,8934	38,9954		-0,1036	0,0024	0,9613
	virkaluokka 8	-2,5720	38,9963		-0,5513	0,0044	0,9474

Taulukko 5: Mallin käyttäytyminen aineistoa suojatessa. Tutkitaan viran vaikutusta pätevyyden arvoon. Aineistossa muuttujan virka nimeäminen kuten alkuperäisessä aineistossa. Suojattu käyttäen mikroaggregointia. Osa 2.

Koko	Muuttuja	Estimaatit	s.e.	95% lv.		t-arvo	p-arvo
	vakio	1,5063	56,0666	-0,5457		0,0007	0,9786
	ikä	0,0426	0,0049	0,0332	0,0522	76,9600	<0,0001
	sukupuoli	0,3539	0,0980	0,1611	0,5453	13,0443	0,0003
5	virkaluokka 1	0,3566	56,0671		2,6009	0,0000	0,9949
	virkaluokka 3	9,3467	336,4	-1,6351		0,0008	0,9778
	virkaluokka 4	-0,1711	56,0662		1,7154	0,0000	0,9976
	virkaluokka 5	-1,4766	56,0666		0,5879	0,0007	0,9790
	virkaluokka 6	-1,9343	56,0662		-0,0437	0,0012	0,9725
	virkaluokka 8	-2,7108	56,0669		-0,5380	0,0023	0,9614
	vakio	1,4529	46,7173	-0,4120		0,0010	0,9752
	ikä	0,0418	0,0049	0,0324	0,0514	74,2794	<0,0001
	sukupuoli	0,3535	0,0979	0,1609	0,5447	13,0498	0,0003
	virkaluokka 1	0,5815	46,7181		2,7001	0,0002	0,9901
	virkaluokka 2	9,9922	327,0	-1,1154		0,0009	0,9756
6	virkaluokka 3	-1,6075	46,7256		2,0463	0,0012	0,9726
	virkaluokka 4	-0,0823	46,7168		1,6210	0,0000	0,9986
	virkaluokka 5	-1,3186	46,7174		0,5754	0,0008	0,9775
	virkaluokka 6	-1,8770	46,7168		-0,1689	0,0016	0,9680
	virkaluokka 8	-2,3771	46,7176		-0,4211	0,0026	0,9594
	vakio	0,0634	0,2666	-0,4539	0,5949	0,0565	0,8121
	ikä	0,0417	0,0049	0,0322	0,0513	73,5892	<0,0001
	sukupuoli	0,3530	0,0977	0,1608	0,5439	13,0639	0,0003
	virkaluokka 1	1,9888	0,3772	1,3107	2,8190	27,7952	<0,0001
7	virkaluokka 3	-0,4706	0,6483	-1,5603	1,1172	0,5269	0,4679
	virkaluokka 4	1,2984	0,1567	0,9612	1,5883	68,6389	<0,0001
	virkaluokka 5	0,1661	0,2778	-0,3657	0,7358	0,3574	0,5500
	virkaluokka 6	-0,4851	0,1597	-0,8271	-0,1887	9,2214	0,0024
	virkaluokka 8	-0,5720	0,3673	-1,2607	0,1998	2,4242	0,1195
	vakio	0,0914	0,2879	-0,4568	0,6947	0,1007	0,7509
	ikä	0,0421	0,0049	0,0327	0,0517	75,2823	<0,0001
	sukupuoli	0,3578	0,0976	0,1657	0,5486	13,4282	0,0002
	virkaluokka 1	1,7794	0,3692	1,0774	2,5667	23,2246	<0,0001
8	virkaluokka 3	-0,0132	0,8936	-1,4069	2,4759	0,0002	0,9882
	virkaluokka 4	1,2436	0,1887	0,7839	1,5752	43,4111	<0,0001
	virkaluokka 5	0,0297	0,3058	-0,5876	0,6452	0,0094	0,9227
	virkaluokka 6	-0,5127	0,1915	-0,9757	-0,1747	7,1694	0,0074
	virkaluokka 8	-0,5629	0,4213	-1,3589	0,3319	1,7848	0,1816

Taulukko 6: Mallin käyttäytyminen aineistoa suojatessa. Tutkitaan viran vaikutusta pätevyyden arvoon. Aineistossa muuttujan virka nimeäminen kuten alkuperäisessä aineistossa. Suojattu käyttäen mikroaggregointia. Osa 3.

Koko	Muuttuja	Estimaatit	s.e.	95% lv.		t-arvo	p-arvo
	vakio	1,6304	47,4886	-0,2287		0,0012	0,9726
	ikä	0,0404	0,0049	0,0310	0,0500	69,5536	<0,0001
	sukupuoli	0,3532	0,0973	0,1617	0,5434	13,1670	0,0003
	virkaluokka 1	0,4643	47,4894		2,5867	0,0001	0,9922
	virkaluokka 2	10,5978	332,4	-1,4477		0,0010	0,9746
9	virkaluokka 3	-1,5124	47,4970		2,1565	0,0010	0,9746
	virkaluokka 4	-0,2328	47,4881		1,4701	0,0000	0,9961
	virkaluokka 5	-1,3386	47,4888		0,5837	0,0008	0,9775
	virkaluokka 6	-2,0211	47,4881		-0,3150	0,0018	0,9661
	virkaluokka 8	-2,5382	47,4891		-0,5246	0,0029	0,9574
	vakio	0,1658	0,2814	-0,3690	0,7584	0,3472	0,5557
	ikä	0,0407	0,0049	0,0312	0,0503	69,9220	<0,0001
	sukupuoli	0,3602	0,0976	0,1682	0,5509	13,6259	0,0002
	virkaluokka 1	1,9222	0,3914	1,1901	2,7700	24,1193	<0,0001
10	virkaluokka 3	0,4570	0,8875	-0,9160	2,9403	0,2651	0,6066
	virkaluokka 4	1,2169	0,1866	0,7601	1,5440	42,5396	<0,0001
	virkaluokka 5	0,2187	0,3100	-0,4029	0,8474	0,4979	0,4804
	virkaluokka 6	-0,5446	0,1880	-1,0032	-0,2143	8,3920	0,0038
	virkaluokka 8	-1,2908	0,3753	-2,0332	-0,5361	11,8292	0,0006
	vakio	0,1761	0,2752	-0,3459	0,7584	0,4097	0,5221
	ikä	0,0402	0,0049	0,0307	0,0497	68,5931	<0,0001
	sukupuoli	0,3589	0,0974	0,1673	0,5493	13,5767	0,0002
	virkaluokka 1	1,9493	0,3912	1,2177	2,7970	24,8234	<0,0001
11	virkaluokka 3	0,2695	0,9043	-1,1628	2,7685	0,0888	0,7657
	virkaluokka 4	1,2360	0,1855	0,7807	1,5615	44,3837	<0,0001
	virkaluokka 5	0,1800	0,3015	-0,4293	0,7880	0,3563	0,5506
	virkaluokka 6	-0,5450	0,1863	-1,0013	-0,2178	8,5577	0,0034
	virkaluokka 8	-1,1218	0,3084	-1,7508	-0,5143	13,2347	0,0003
	vakio	0,0172	0,2632	-0,4932	0,5425	0,0043	0,9478
	ikä	0,0414	0,0049	0,0320	0,0510	72,5962	<0,0001
	sukupuoli	0,3795	0,0975	0,1877	0,5701	15,1524	<0,0001
	virkaluokka 1	2,2083	0,4092	1,4828	3,1237	29,1280	<0,0001
15	virkaluokka 3	-0,3555	0,6605	-1,4823	1,2465	0,2897	0,5904
	virkaluokka 4	1,3692	0,1601	1,0263	1,6666	73,1020	<0,0001
	virkaluokka 5	0,2639	0,2792	-0,2707	0,8360	0,8933	0,3446
	virkaluokka 6	-0,4770	0,1614	-0,8219	-0,1772	8,7391	0,0031
	virkaluokka 8	-1,2388	0,3603	-1,9334	-0,5053	11,8233	0,0006

Taulukko 7: Mallin käyttäytyminen aineistoa suojatessa. Tutkitaan viran vaikutusta pätevyyden arvoon. Suojattu käyttäen PRAM-menetelmää rajoittamatta vaihtumista. Osa 1. (Alkuperäistä aineistoa on merkitty vaihtumattomuustodennäköisyydellä 1.)

Tn	Muuttuja	Estimaatit	s.e.	95% lv.		t-arvo	p-arvo
	vakio	0,0348	0,2419	-0,4388	0,5102	0,0207	0,8857
	ikä	0,0419	0,0419	0,0324	0,0515	74,0574	<0,0001
	sukupuoli	0,3469	0,0982	0,1537	0,5388	12,4816	0,0004
1	virkaluokka 1	1,8713	0,3303	1,2838	2,5985	32,0911	<0,0001
	virkaluokka 4	1,3539	0,1212	1,1091	1,5860	124,8486	<0,0001
	virkaluokka 5	-0,0963	0,2365	-0,5420	0,3912	0,1658	0,6839
	virkaluokka 6	-0,4467	0,1235	-0,6959	-0,2102	13,0916	0,0003
	virkaluokka 8	-0,7627	0,3235	-1,3629	-0,0801	5,5573	0,0184
	vakio	-0,0543	0,2254	-0,4960	0,3879	0,0581	0,8095
	ikä	0,0440	0,0047	0,0348	0,0534	86,6781	<0,0001
	sukupuoli	0,3515	0,0974	0,1599	0,5418	13,0244	0,0003
0,97	virkaluokka 1	1,3537	0,2391	0,9143	1,8587	32,0547	<0,0001
	virkaluokka 4	1,3401	0,1054	1,1308	1,5447	161,6695	<0,0001
	virkaluokka 5	-0,0876	0,2135	-0,4889	0,3524	0,1685	0,6815
	virkaluokka 6	-0,4184	0,1069	-0,6309	-0,2111	15,3113	<0,0001
	virkaluokka 8	-0,3211	0,2608	-0,8050	0,2255	1,5158	0,2183
	vakio	0,2895	0,2237	-0,1482	0,7291	1,6744	0,1957
	ikä	0,0397	0,0046	0,0308	0,0488	74,7705	<0,0001
	sukupuoli	0,3248	0,0967	0,1344	0,5136	11,2802	0,0008
0,95	virkaluokka 1	1,7134	0,2898	1,1934	2,3430	34,9501	<0,0001
	virkaluokka 4	1,1876	0,1092	0,9692	1,3984	118,3724	<0,0001
	virkaluokka 5	-0,0866	0,2133	-0,4878	0,3527	0,1649	0,6847
	virkaluokka 6	-0,6107	0,1099	-0,8309	-0,3988	30,8591	<0,0001
	virkaluokka 8	-0,2705	0,2502	-0,7353	0,2527	1,1692	0,2796
	vakio	0,1940	0,2197	-0,2362	0,6253	0,7794	0,3773
	ikä	0,0407	0,0046	0,0317	0,0499	77,6875	<0,0001
	sukupuoli	0,3359	0,0966	0,1458	0,5248	12,0843	0,0005
0,93	virkaluokka 1	1,4120	0,2360	0,9798	1,9120	35,7960	<0,0001
	virkaluokka 4	1,2369	0,1016	1,0358	1,4346	148,2935	<0,0001
	virkaluokka 5	-0,1487	0,1947	-0,5158	0,2506	0,5833	0,4450
	virkaluokka 6	-0,5336	0,1031	-0,7380	-0,3334	26,8127	<0,0001
	virkaluokka 8	-0,0463	0,2309	-0,4753	0,4359	0,0401	0,8412

Taulukko 8: Mallin käyttäytyminen aineistoa suojatessa. Tutkitaan viran vaikutusta pätevyyden arvoon. Suojattu käyttäen PRAM-menetelmää rajoittamatta vaihtumista. Osa 2.

Tn	Muuttuja	Estimaatit	s.e.	95% lv.		t-arvo	p-arvo
	vakio	0,4952	0,2110	0,0826	0,9098	5,5110	0,0189
	ikä	0,0352	0,0044	0,0267	0,0439	64,2781	<0,0001
	sukupuoli	0,2880	0,0957	0,0997	0,4750	9,0564	0,0026
0,90	virkaluokka 1	1,0765	0,1938	0,7145	1,4782	30,8410	<0,0001
	virkaluokka 4	1,1861	0,1001	0,9882	1,3811	140,5286	<0,0001
	virkaluokka 5	0,0601	0,1952	-0,3066	0,4623	0,0949	0,7580
	virkaluokka 6	-0,5834	0,1007	-0,7830	-0,3877	33,5764	<0,0001
	virkaluokka 8	0,1863	0,2489	-0,2704	0,7133	0,5599	0,4543
	vakio	0,4399	0,1969	0,0547	0,8267	4,9921	0,0255
	ikä	0,0361	0,0042	0,0280	0,0443	75,5192	<0,0001
	sukupuoli	0,2771	0,0939	0,0923	0,4604	8,7090	0,0032
0,85	virkaluokka 1	0,7110	0,1568	0,4145	1,0311	20,5519	<0,0001
	virkaluokka 4	1,1599	0,0918	0,9798	1,3402	159,4845	<0,0001
	virkaluokka 5	0,1821	0,1804	-0,1564	0,5536	1,0192	0,3127
	virkaluokka 6	-0,4812	0,0936	-0,6651	-0,2979	26,4285	<0,0001
	virkaluokka 8	0,1652	0,1981	-0,2038	0,5766	0,6956	0,4043
	vakio	0,4499	0,1935	0,0714	0,8301	5,4051	0,0201
	ikä	0,0349	0,0041	0,0269	0,0430	72,5968	<0,0001
	sukupuoli	0,3036	0,0929	0,1207	0,4852	10,6721	0,0011
0,80	virkaluokka 1	0,5580	0,1329	0,3046	0,8267	17,6298	<0,0001
	virkaluokka 4	1,2110	0,0896	1,0361	1,3878	182,5758	<0,0001
	virkaluokka 5	0,0930	0,1572	-0,2041	0,4139	0,3500	0,5541
	virkaluokka 6	-0,4438	0,0897	-0,6194	-0,2677	24,5048	<0,0001
	virkaluokka 8	0,1390	0,1851	-0,2067	0,5218	0,5640	0,4527
	vakio	0,5542	0,1876	0,1874	0,9229	8,7286	0,0031
	ikä	0,0344	0,0040	0,0265	0,0423	73,1762	<0,0001
	sukupuoli	0,2603	0,0927	0,0778	0,4414	7,8801	0,0050
0,75	virkaluokka 1	0,6165	0,1287	0,3713	0,8769	22,9394	<0,0001
	virkaluokka 4	1,2215	0,0908	1,0451	1,4014	180,9838	<0,0001
	virkaluokka 5	0,2915	0,1630	-0,0148	0,6264	3,1964	0,0738
	virkaluokka 6	-0,4672	0,0877	-0,6385	-0,2946	28,4145	<0,0001
	virkaluokka 8	-0,0972	0,1517	-0,3844	0,2116	0,4111	0,5214

Taulukko 9: Mallin käyttäytyminen aineistoa suojatessa. Tutkitaan viran vaikutusta pätevyyden arvoon. Suojattu käyttäen PRAM-menetelmää rajoittamatta vaihtumista. Osa 3.

Tn	Muuttuja	Estimaatit	s.e.	95% lv.		t-arvo	p-arvo
	vakio	0,5196	0,1821	0,1634	0,8776	8,1387	0,0043
	ikä	0,0344	0,0039	0,0268	0,0421	78,2345	<0,0001
	sukupuoli	0,2880	0,0918	0,1074	0,4674	9,8417	0,0017
0,70	virkaluokka 1	0,3824	0,1118	0,1680	0,6070	11,6957	0,0006
	virkaluokka 4	1,1895	0,0885	1,0179	1,3652	180,5851	<0,0001
	virkaluokka 5	0,1322	0,1407	-0,1345	0,4186	0,8826	0,3475
	virkaluokka 6	-0,3037	0,0877	-0,4744	-0,1303	11,9879	0,0005
	virkaluokka 8	0,0982	0,1441	-0,1745	0,3917	0,4647	0,4954
	vakio	1,0271	0,1748	0,6860	0,3714	34,5256	<0,0001
	ikä	0,0247	0,0037	0,0175	0,0320	44,3570	<0,0001
	sukupuoli	0,2310	0,0906	0,0527	0,4078	6,5088	0,0107
0,65	virkaluokka 1	0,2942	0,1047	0,0927	0,5037	7,8908	0,0050
	virkaluokka 4	1,0609	0,0882	0,8899	1,2361	144,6406	<0,0001
	virkaluokka 5	0,2633	0,1440	-0,0086	0,5572	3,3448	0,0674
	virkaluokka 6	-0,4238	0,0877	-0,5945	-0,2505	23,3592	<0,0001
	virkaluokka 8	0,1980	0,1475	-0,0802	0,4995	1,8028	0,1794
	vakio	0,8076	0,1729	0,4700	1,1478	21,8229	<0,0001
	ikä	0,0282	0,0037	0,0209	0,0355	57,6181	<0,0001
	sukupuoli	0,3137	0,0897	0,1373	0,4888	12,2470	0,0005
0,60	virkaluokka 1	0,3717	0,1026	0,1747	0,5772	13,1323	0,0003
	virkaluokka 4	0,9899	0,0851	0,8252	1,1589	135,3963	<0,0001
	virkaluokka 5	0,1372	0,1271	-0,1045	0,3948	1,1645	0,2805
	virkaluokka 6	-0,3212	0,0876	-0,4911	-0,1475	13,4396	0,0002
	virkaluokka 8	0,0871	0,1296	-0,1590	0,3499	0,4525	0,5012

Taulukko 10: Mallin käyttäytyminen aineistoa suojatessa. Tutkitaan viran vaikutusta pätevyyden arvoon. Suojattu käyttäen PRAM-menetelmää rajoittaen vaihtuminen korkeintaan kahden arvon päähän. Osa 1. (Alkuperäistä aineistoa on merkitty vaihtumattomuustodennäköisyydellä 1.)

Tn	Muuttuja	Estimaatit	s.e.	95% lv.		t-arvo	p-arvo
	vakio	0,0348	0,2419	-0,4388	0,5102	0,0207	0,8857
	ikä	0,0419	0,0419	0,0324	0,0515	74,0574	<0,0001
	sukupuoli	0,3469	0,0982	0,1537	0,5388	12,4816	0,0004
1	virkaluokka 1	1,8713	0,3303	1,2838	2,5985	32,0911	<0,0001
	virkaluokka 4	1,3539	0,1212	1,1091	1,5860	124,8486	<0,0001
	virkaluokka 5	-0,0963	0,2365	-0,5420	0,3912	0,1658	0,6839
	virkaluokka 6	-0,4467	0,1235	-0,6959	-0,2102	13,0916	0,0003
	virkaluokka 8	-0,7627	0,3235	-1,3629	-0,0801	5,5573	0,0184
	vakio	-0,0198	0,2363	-0,4826	0,4443	0,0070	0,9333
	ikä	0,0438	0,0049	0,0344	0,0534	81,8421	<0,0001
	sukupuoli	0,3399	0,0980	0,1469	0,5314	12,0179	0,0005
0,97	virkaluokka 1	1,8378	0,3281	1,2559	2,5616	31,3734	<0,0001
	virkaluokka 4	1,3054	0,1148	1,0741	1,5258	129,3657	<0,0001
	virkaluokka 5	0,1557	0,2199	-0,2583	0,6088	0,5015	0,4788
	virkaluokka 6	-0,4155	0,1170	-0,6511	-0,1907	12,6086	0,0004
	virkaluokka 8	-0,9129	0,2699	-1,4227	-0,3571	11,4440	0,0007
	vakio	-0,0456	0,2333	-0,5025	0,4124	0,0381	0,8451
	ikä	0,0446	0,0048	0,0352	0,0540	86,1277	<0,0001
	sukupuoli	0,3298	0,0976	0,1376	0,5206	11,4100	0,0007
0,95	virkaluokka 1	1,8314	0,3270	1,2523	2,5536	31,3723	<0,0001
	virkaluokka 4	1,2920	0,1121	1,0662	1,5076	132,7720	<0,0001
	virkaluokka 5	0,2833	0,2173	-0,1249	0,7319	1,7000	0,1923
	virkaluokka 6	-0,3771	0,1142	-0,6070	-0,1574	10,8931	0,0010
	virkaluokka 8	-1,1036	0,2409	-1,5641	-0,6145	20,9789	<0,0001
	vakio	0,0406	0,2340	-0,4175	0,5002	0,0301	0,8622
	ikä	0,0441	0,0048	0,0347	0,0536	83,7868	<0,0001
	sukupuoli	0,3256	0,0977	0,1332	0,5166	11,0988	0,0009
0,93	virkaluokka 1	1,7672	0,3268	1,1885	2,4890	29,2403	<0,0001
	virkaluokka 4	1,2523	0,1119	1,0268	1,4676	125,1313	<0,0001
	virkaluokka 5	0,2896	0,2118	-0,1089	0,7261	1,8699	0,1715
	virkaluokka 6	-0,4714	0,1129	-0,6990	-0,2546	17,4252	<0,0001
	virkaluokka 8	-0,8344	0,2371	-1,2839	-0,3490	12,3835	0,0004

Taulukko 11: Mallin käyttäytyminen aineistoa suojatessa. Tutkitaan viran vaikutusta pätevyyden arvoon. Suojattu käyttäen PRAM-menetelmää rajoittaen vaihtuminen korkeintaan kahden arvon päähän. Osa 2.

Tn	Muuttuja	Estimaatit	s.e.	95% lv.		t-arvo	p-arvo
	vakio	-0,0144	0,2292	-0,4636	0,4351	0,0040	0,9499
	ikä	0,0461	0,0048	0,0369	0,0556	94,0638	<0,0001
	sukupuoli	0,3438	0,0971	0,1529	0,5335	12,5487	0,0004
0,90	virkaluokka 1	1,5749	0,3073	1,0268	2,2475	26,2743	<0,0001
	virkaluokka 4	1,1317	0,1089	0,9132	1,3416	107,9990	<0,0001
	virkaluokka 5	0,7570	0,2167	0,3527	1,2077	12,1984	0,0005
	virkaluokka 6	-0,4689	0,1099	-0,6896	-0,2572	18,1896	<0,0001
	virkaluokka 8	-0,9525	0,2294	-1,3894	-0,4854	17,2359	<0,0001
	vakio	0,0208	0,2261	-0,4222	0,4647	0,0084	0,9268
	ikä	0,0471	0,0047	0,0379	0,0564	98,7449	<0,0001
	sukupuoli	0,3202	0,0973	0,1286	0,5102	10,8312	0,0010
0,85	virkaluokka 1	1,5288	0,3063	0,9832	2,2003	24,9156	<0,0001
	virkaluokka 4	1,0736	0,1073	0,8585	1,2806	100,1828	<0,0001
	virkaluokka 5	0,7628	0,1980	0,3906	1,1710	14,8390	0,0001
	virkaluokka 6	-0,4949	0,1067	-0,7093	-0,2897	21,5328	<0,0001
	virkaluokka 8	-0,7704	0,2156	-1,1806	-0,3314	12,7643	0,0004
	vakio	-0,0747	0,2226	-0,5107	0,3622	0,1127	0,7371
	ikä	0,0482	0,0047	0,0390	0,0575	104,6315	<0,0001
	sukupuoli	0,3455	0,0964	0,1557	0,5339	12,8396	0,0003
0,80	virkaluokka 1	1,5480	0,3049	1,0057	2,2174	25,7677	<0,0001
	virkaluokka 4	1,0665	0,1030	0,8598	1,2654	107,1414	<0,0001
	virkaluokka 5	0,7767	0,1776	0,4402	1,1398	19,1230	<0,0001
	virkaluokka 6	-0,4031	0,1033	-0,6106	-0,2040	15,2209	<0,0001
	virkaluokka 8	-0,9128	0,1766	-1,2538	-0,5594	26,7193	<0,0001
	vakio	-0,0575	0,2232	-0,4947	0,3809	0,0664	0,7966
	ikä	0,0483	0,0047	0,0392	0,0576	106,1418	<0,0001
	sukupuoli	0,3342	0,0959	0,1454	0,5215	12,1443	0,0005
0,75	virkaluokka 1	1,6754	0,3243	1,1032	2,3942	26,6913	<0,0001
	virkaluokka 4	1,0545	0,1057	0,8413	1,2580	99,4788	<0,0001
	virkaluokka 5	0,8224	0,1760	0,4877	1,1811	21,8310	<0,0001
	virkaluokka 6	-0,4140	0,1039	-0,6243	-0,2148	15,8781	<0,0001
	virkaluokka 8	-1,1305	0,1729	-1,4663	-0,7863	42,7507	<0,0001

Taulukko 12: Mallin käyttäytyminen aineistoa suojatessa. Tutkitaan viran vaikutusta pätevyyden arvoon. Suojattu käyttäen PRAM-menetelmää rajoittaen vaihtuminen korkeintaan kahden arvon päähän. Osa 3.

Tn	Muuttuja	Estimaatit	s.e.	95% lv.		t-arvo	p-arvo
	vakio	-0,1512	0,2198	-0,5822	0,2799	0,4731	0,4916
	ikä	0,0500	0,0047	0,0408	0,0593	113,3459	<0,0001
	sukupuoli	0,3722	0,0959	0,1835	0,5596	15,0657	0,0001
0,70	virkaluokka 1	1,4270	0,2879	0,9119	2,0540	24,5646	<0,0001
	virkaluokka 4	1,0831	0,1018	0,8806	1,2810	113,2259	<0,0001
	virkaluokka 5	0,8166	0,1601	0,5116	1,1419	25,9968	<0,0001
	virkaluokka 6	-0,3508	0,0975	-0,5460	-0,1625	12,9459	0,0003
	virkaluokka 8	-0,9134	0,1561	-1,2159	-0,6026	34,2485	<0,0001
	vakio	-0,0837	0,2180	-0,5108	0,3443	0,1474	0,7010
	ikä	0,0487	0,0046	0,0397	0,0578	110,8870	<0,0001
	sukupuoli	0,3706	0,0949	0,1838	0,5561	15,2426	<0,0001
0,65	virkaluokka 1	1,5416	0,3043	1,0008	2,2098	25,6663	<0,0001
	virkaluokka 4	0,9538	0,1023	0,7488	1,1515	86,9503	<0,0001
	virkaluokka 5	1,0568	0,1709	0,7326	1,4056	38,2365	<0,0001
	virkaluokka 6	-0,3463	0,0996	-0,5468	-0,1547	12,0865	0,0005
	virkaluokka 8	-1,3007	0,1560	-1,6051	-0,9922	69,5581	<0,0001
	vakio	-0,1867	0,2150	-0,6080	0,2351	0,7540	0,3852
	ikä	0,0507	0,0046	0,0418	0,0597	122,3797	<0,0001
	sukupuoli	0,3736	0,0948	0,1872	0,5588	15,5481	<0,0001
0,60	virkaluokka 1	1,5554	0,3036	1,0159	2,2221	26,2405	<0,0001
	virkaluokka 4	1,0164	0,1020	0,8124	1,2141	99,3201	<0,0001
	virkaluokka 5	0,8629	0,1510	0,5732	1,1675	32,6446	<0,0001
	virkaluokka 6	-0,3278	0,0976	-0,5244	-0,1399	11,2736	0,0008
	virkaluokka 8	-1,1807	0,1399	-1,4551	-0,9052	71,1913	<0,0001

Liite 3

Taulukko 1: Opettajatiedonkeruu 2005. Muuttujakuvaukset. Osa 1.

Tiedon nimi	Kuvaus
okun	Kuntakoodi
omist	Omistajatyyppi
	1 =yksityinen
	2 =valtio
	3 =kunta
	4 =kuntayhtymä
	5 =Ahvenanmaan maakunta
	9 =muu
okieli	Oppilaitoksen opetuskieli
	1 =suomi
	2 =ruotsi
oltyp	Oppilaitostyyppi
	11 =Peruskoulut
	12 =Peruskouluasteen erityiskoulut
	15 =Lukiot
	19 =Perus- ja lukioasteen koulut
	21 =Ammatilliset oppilaitokset
	22 =Ammatilliset erityisoppilaitokset
	24 =Ammatilliset aikuiskoulutuskeskukset
	61 =Musiikkioppilaitokset
	62 =Liikunnan koulutuskeskukset
	63 =Kansanopistot
	64 =Kansalaisopistot
	65 =Opintokeskukset
	99 =Muut oppilaitokset
tunn	Oppilaitoksen yksilöintinumero TK:n oppilaitosrekisterissä
onimi	Oppilaitoksen nimi
OPPYHT	Oppilaitoksen oppilaat yhteensä
OPP16	Perusopetus, (vuosiluokat 1–6)
OPP79	Perusopetus, (vuosiluokat 7–9)
syntaika	Opettajan syntymäaika
sv	Syntymävuosi
ika	Henkilön ikä vuonna 2005
sp	Sukupuoli
	Väestötietorekisteristä:
svalt	Syntymävaltio
edkans	Aikaisempi kansalaisuus
kiel	Äidinkieli
kansv	Kansalaisuus
askun	Asuinkunta

Taulukko 2: Opettajatiedonkeruu 2005. Muuttujakuvaukset. Osa 2.

Tiedon nimi	Kuvaus
tyosuhd	Työsuhde-muuttuja 1=vakinainen (toistaiseksi määrätty) 2=määräaikainen (ei viransijainen) 3=viransijainen/toimensijainen 4=virkavapaalla 5=osa-aikaeläkkeellä
virka	Tehtävätyyppi perusopetuksessa ja lukiokoulutuksessa 01=Rehtori, perusopetus 02=Rehtori, erityiskoulu 03=Rehtori, nuorten lukiokoulutus 04=Rehtori, perusopetuksen ja lukion yhteinen 05=Rehtori, aikuisten lukiokoulutus 06=Rehtori, muu yhdistelmävirka 07=Esiluokanopettaja 08=Luokanopettaja 09=Erityisluokanopettaja, sopeutumattomien opetus 10=Erityisluokanopettaja, mukautettu opetus 11=Erityisluokanopettaja, vaikeasti kehitysvammaisten opetus 12=Erityisluokanopettaja, kuulovammaisten opetus 13=Erityisluokanopettaja, näkövammaisten opetus 14=Erityisluokanopettaja, liikuntavammaisten opetus 15=Erityisluokanopettaja, dysfasiaopetus 16=Erityisluokanopettaja, autistien opetus 17=Erityisluokanopettaja, muu erityisopetus 18=Osa-aikainen erityisopetus 19=Maahanmuuttajaopettaja (perusopetukseen valmistava opetus) 20=Perusopetuksen lehtori 21=Oppilaanohjaaja 22=Opinto-ohjaaja 23=Yhteinen oppilaanohjaajan ja opinto-ohjaajan tehtävä 24=Lukion lehtori 25=Perusopetuksen ja lukion yhteinen lehtori 26=Lukion ja perusopetuksen yhteinen lehtori 27=Päätoiminen tuntiopettaja perusopetuksessa 28=Päätoiminen tuntiopettaja lukiossa 29=Päätoiminen tuntiopettaja, perusopetuksen ja lukion yhteinen 30=Päätoiminen tuntiopettaja, lukion ja perusopetuksen yhteinen 31=Muu yhdistelmävirka 32=Sivutoiminen tuntiopettaja perusopetuksessa 33=Sivutoiminen tuntiopettaja lukiossa

Taulukko 3: Opettajatiedonkeruu 2005. Muuttujakuvaukset. Osa 3.

Tiedon nimi	Kuvaus
aste	Kouluaste, jolla pääasiallisesti opettaa
	01=Esiopetus
	02=Perusopetus, vuosiluokat 1 – –6
	03=Perusopetus, vuosiluokat 7 – –9
	04=Lisäopetus (10. luokka)
	05=Nuorten lukiokoulutus
	06=Aikuisten lukiokoulutus
	07=Ammatillinen koulutus
	08=Ammattikorkeakoulu
	09=Yliopisto
	10=Vapaa sivistystyö
	11=Taiteen perusopetus
	12=Muu
laste01	Opetusta edellisen lisäksi, mikäli tiedossa
	01=Esiopetus
	02=Perusopetus, vuosiluokat 1 – –6
	03=Perusopetus, vuosiluokat 7 – –9
	04=Lisäopetus (10. luokka)
	05=Nuorten lukiokoulutus
	06=Aikuisten lukiokoulutus
	07=Ammatillinen koulutus
	08=Ammattikorkeakoulu
	09=Yliopisto
	10=Vapaa sivistystyö
	11=Taiteen perusopetus
	12=Muu
	13=Ei opetusta muualla (aiempien lisäksi)
laste02–laste12	Opetusta edellisen lisäksi, mikäli tiedossa
	kuten laste01
aine1	Opettajan eniten opettama aine tai opinnot
	01=Esiopetus
	02=Luokanopetus
	03=Luokkamuotoinen erityisopetus
	04=Maahanmuuttajien perusopetukseen valmistava opetus
	05=Äidinkieli ja kirjallisuus , suomi äidinkielenä
	06=Äidinkieli ja kirjallisuus , ruotsi äidinkielenä
	07=Äidinkieli ja kirjallisuus, saame äidinkielenä
	08=Äidinkieli ja kirjallisuus, romani äidinkielenä
	09=Äidinkieli ja kirjallisuus, viittomakieli äidinkielenä
	10=Äidinkieli ja kirjallisuus, muu oppilaan äidinkieli
	11=Äidinkieli ja kirjallisuus, suomi toisena kielenä
	12=Äidinkieli ja kirjallisuus, ruotsi toisena kielenä
	13=Äidinkieli ja kirjallisuus, suomi saamenkielisille
	14=Äidinkieli ja kirjallisuus, suomi viittomakielisille
	15=Äidinkieli ja kirjallisuus, ruotsi viittomakielisille

Taulukko 4: Opettajatiedonkeruu 2005. Muuttujakuvaukset. Osa 4.

Tiedon nimi	Kuvaus
aine1	Opettajan eniten opettama aine tai opinnot 16=Toinen kotimainen kieli ruotsi (suomenkiel.opetus) 17=Toinen kotimainen kieli suomi (ruotsinkiel.opetus) 18=Saame vieraana kielenä 19=Englanti 20=Saksa 21=Ranska 22=Venäjä 23=Espanja 24=Italia 25=Latina 26=Muu kieli 27=Matematiikka 28=Fysiikka 29=Kemia 30=Maantieto 31=Biologia 32=Ympäristö- ja luonnontieteet 33=Uskonto, evankelis–luterilainen 34=Uskonto, ortodoksinen 35=Muut uskonnot 36=Elämänkatsomustieto 37=Historia 38=Yhteiskuntaoppi 39=Oppilaanohjaus 40=Opinto-ohjaus 41=Osa-aikainen erityisopetus 42=Musiikki 43=Kuvataide 44=Liikunta 45=Terveystieto 46=Käsityö (tekstiilityö) 47=Käsityö (tekninen työ) 48=Käsityö 49=Kotitalous 50=Filosofia 51=Psykologia 52=Tietotekniikka 53=Maa- ja metsätalous, puutarhanhoito 54=Kaupalliset aineet ja konekirjoitus 88=Muut 99=Muut
aine2–aine3	Opettajan toiseksi eniten opettama aine tai opinnot kuten aine1

Taulukko 5: Opettajatiedonkeruu 2005. Muuttujakuvaukset. Osa 5.

Tiedon nimi	Kuvaus
	Kelpoisuustekijät hoidettuun tehtävään perusopetuksessa ja/tai lukiokoulutuksessa
k1	Esiopetuksen opettajan tehtävään
k2	Luokanopettajan tehtävään
k3	Aineenopettajan tehtävään perusopetuksessa
k4	Aineenopettajan tehtävään lukiossa
k5	Erityisluokanopettajan tehtävään
k6	Erityisopettajan tehtävään osa-aikaisessa erityisopetuksessa
k7	Oppilaanohjaajan tehtävään perusopetuksessa
k8	Opinto-ohjaajan tehtävään lukiossa
k9	Muu opettajan kelpoisuus (esim. ammatillisen koulutuksen tai vapaan sivistystyön opettajan kelpoisuus)
	Puuttuvat kelpoisuustekijät hoidettuun tehtävään perusopetuksessa ja/tai lukiokoulutuksessa
p1	Opettajan tehtävään vaadittava tutkinto puuttuu
p2	Muut tehtävään vaadittavat opinnot puuttuvat (esim. erityisopettajan tai opinto-ohjaajan opinnot)
p3	Opettajan pedagogiset opinnot puuttuvat
p4	Luokanopettajan monialaiset opinnot puuttuvat
p5	Hyväksytty oppilaitoksen opetuskielen hallinta puuttuu
p6	Riittävä arvosana opetettavasta aineesta puuttuu
p	1=henkilöllä on jokin puuttuva kelpoisuustekijä 0=henkilöllä ei ole puuttuvaa kelpoisuustekijää
k	1=henkilöllä on jokin kelpoisuustekijä 0=henkilöllä ei ole kelpoisuustekijöitä
kp	Kelpoisuusmerkintä 0=ei merkintää muuttujissa k ja p 1=merkintä jommassakummassa tai molemmissa muuttujissa p ja k
pmuo	Jaottelu peruskoulu ja lukio puolelle koulutusasteen mukaan
pmuo3	Jaottelu peruskoulu ja lukio puolelle oppilaitostyypin mukaan
pmuo2	Jaottelu viran mukaan omiksi ryhmiksi 1=virka peruskoulu (01,02,07,08,09,10,11,12,13,14,15,16,17,18,20,21,27,32) 2=virka lukio (03,05,22,24,28,33) Jos virka on 04,06,19,23,25,26,29,30,31 niin pmuo2=pmuo3
kelpu	Opettajan kelpoisuus hoitamaansa tehtävään viran ja tehtävätyypin mukaan
ainkelp	Aineenopettajakelpoisuus