Anonymisointimenetelmiä henkilötietoa sisältävälle rivitason tiedolle

Johannes Rajala

Pro gradu -tutkielma
Toukokuu 2024

Tarkastajat:
Katariina Perkonoja
Joni Virta

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

Johannes Rajala: Anonymisointimenetelmiä henkilötietoa sisältävälle rivitason tiedolle
Pro gradu -tutkielma, 44 s., 28 liites.
Tilastotiede
Toukokuu 2024

Euroopan Unionin yleinen tietosuoja-asetus ja Suomen laki sosiaali- ja terveystietojen toissijaisesta käytöstä säätelevät henkilötietoa sisältävän tiedon toissijaista käyttöä Suomessa. Jos henkilötieto anonymisoidaan, ei siihen sovelleta enää tietosuoja-asetusta tai toisiolakia; anonyymiä tietoa ei lueta henkilötiedoksi, jolloin sen käyttö on vapaampaa. Anonymisoinnin tarkoituksena on muuttaa tieto muotoon, jossa havaintoyksiköihin ei kohdistu paljastumisen riskiä. Anonymisointi kuitenkin heikentää tiedon käytettävyyttä, eli kykyä tehdä sillä tilastollista päättelyä, joka olisi yhtenevää alkuperäisellä tiedolla tehtyyn päättelyyn.

Tässä tutkielmassa tarkastellaan viittä rivitason tiedon anonymisointimenetelmää: k-anonymiteettiä, l-diversiteettiä, spektraalista kohinaa, spektraalista sarakepermutaatiota ja kryptografista RSA-menetelmää. Menetelmiä tarkastellaan niiden tuottamien aineistojen yksityisyydensuojan, käytettävyyden ja samankaltaisuuden perusteella.

Spektraalinen sarakepermutaatio tuotti yksityisyydensuojaltaan ja samankaltaisuudeltaan parhaat aineistot. Oikeilla parametrivalinnoilla, k-anonymiteetti ja l-diversiteetti tuottivat käytettävyydeltään parhaat aineistot. RSA:lla salattujen aineistojen käytettävyys ja samankaltaisuus olivat huonoja, eikä niiden yksityisyydensuojaa voitu arvioida tutkielman empiirisillä menetelmillä.

Asiasanat: anonymisointi, yksityisyydensuoja, rivitason tieto, henkilötieto, toisiolaki.

Sisällys

1 Johdanto

Kerättävän tiedon (eng. data) ja näistä johdettujen aineistojen (eng. dataset) määrä sekä yksityiskohtaisuus on tietotekniikan kehityksen ansiosta kasvanut ja tämä on vastaavasti mahdollistanut monipuolisemman tieteellisen tutkimuksen tekemisen. Yksityiskohtaiseen aineistoon kuuluminen aiheuttaa kuitenkin suuremman *paljastumisen* tai *tietosuojarikkomuksen* (eng. disclosure) riskin aineiston *havaintoyksiköille* (eng. statistical unit) [1]. Tietosuojarikkomuksessa aineistossa olevasta havaintoyksiköstä on mahdollista oppia jotain, mitä ennen aineiston julkaisua ei tiedetty [1]. Tästä syystä aineistoja tai niiden perusteella saatuja tuloksia julkaistaessa täytyy olla erityisen tarkkana, ettei havaintoyksiköitä altisteta liian suurelle paljastumisen riskille. Paljastumisen välttämiseksi on kehitetty erilaisia *tilastollisen tietosuojan menetelmiä* (eng. statistical disclosure control), jotka pyrkivät eri tilastotieteellisin tavoin suojaamaan aineistoja erilaisilta paljastumisriskeiltä [1].

Suomessa, *henkilötietoa* (määritelmä 1) sisältävän tiedon käyttöä säätelee muun muassa Euroopan Unionin yleinen tietosuoja-asetus (myöhemmin tietosuoja-asetus) [2] ja *toissijaista käyttöä* Suomen laki sosiaali- ja terveystietojen toissijaisesta käytöstä (myöhemmin toisiolaki) [3]. Toissijainen käyttö tarkoittaa henkilötiedon käyttöä jossain muussa kuin tiedon alkuperäisessä käyttötarkoituksessa [3]. Esimerkiksi terveyskeskuksessa potilaiden hoitoa varten kerätyn tiedon käyttäminen tieteelliseen tutkimukseen on toissijaista käyttöä.

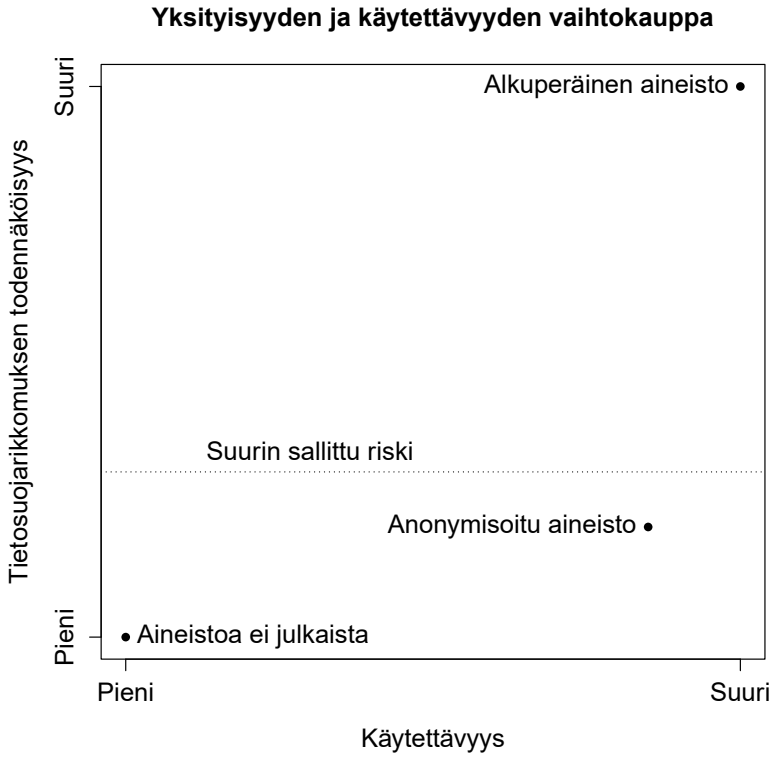
Määritelmä 1. *Henkilötieto*

Henkilötieto on tunnistettuun tai tunnistettavissa olevaan *luonnolliseen henkilöön* liittyvää tietoa. Tunnistettavissa olevana pidetään luonnollista henkilöä, joka voidaan suoraan tai epäsuorasti tunnistaa erityisesti tunnistetietojen, kuten nimen, henkilötunnuksen, sijaintitiedon, verkkotunnistetietojen taikka yhden tai useamman hänelle tunnusomaisen fyysisen, fysiologisen, geneettisen, psyykkisen, taloudellisen, kulttuurillisen tai sosiaalisen tekijän perusteella. [2]

Toisiolaissa sallittuja toissijaisia käyttökohteita ovat muun muassa tilastointi, tieteellinen tutkimus sekä kehittämis- ja innovaatiotoiminta [3, 4]. Toissijaista käyttöä varten voi anoa joko *tietolupaa* tai *tietopyyntöä* Suomen tietosuojaviranomaiselta Findatalta [3]. Tietolupa mahdollistaa henkilötietoa sisältävän tiedon käyttämisen luvan mukaiseen käyttökohteeseen ja tietopyynnöllä voi saada käyttöön tilastomuotoista, luotettavasti anonymisoitua tietoa [3, 4]. Tilastomuotoisella tiedolla tarkoitetaan taulukkomuotoista tietoa (luku 2.1). Toisiokäytössä olevan tiedon perusteella julkaistavien tulosten tulee olla anonyymejä ja tämä voidaan saavuttaa esimerkiksi anonymisoinnilla. Anonymisointi tarkoittaa tiedon saattamista peruuttamattomasti muotoon, josta yksittäiset henkilöt eivät ole tunnistettavissa, eikä tähän riitä pelkkä suorien tunnisteiden, kuten koko nimen tai henkilötunnuksen, poistaminen [5, 6]. Anonymisoitua tietoa ei katsota henkilötiedoksi, joten siihen ei sovelleta enää tietosuoja-asetusta tai tietosuojasäännöksiä [2, 6]. Näin ollen anonymisoitu aineisto voidaan julkaista esimerkiksi osana tieteellistä julkaisua sen jälkeen, kun Findata on varmistunut aineiston anonymiteetistä [3].

Anonymisointi johtaa aina *samankaltaisuuden* (eng. resemblance) muutoksiin ja *käytettävyyden* (eng. utility) heikkenemiseen. Tästä ilmiöstä käytetään termiä *yksityisyyden ja käytettävyyden vaihtokauppa* (eng. privacy-utility tradeoff) [1], jota

on havainnollistettu kuvassa 1. Tässä tutkielmassa samankaltaisuudella tarkoitetaan alkuperäisen ja anonyymin tiedon muuttujien todennäköisyysjakaumien yhtäläisyyttä, käytettävyydellä kykyä tehdä alkuperäisen kaltaisia tilastollisia päätelmiä anonymisoidusta tiedosta ja *yksityisyydellä* tai *yksityisyydensuojalla* (eng. privacy) anonymisoidun tiedon tarjoamaa suojaa erilaisia tietosuojarikkomuksia vastaan.



Kuva 1: Yksityisyyden ja käytettävyyden vaihtokauppa. Tietosuojarikkomuksen riskiä ei synny, jos aineistoa ei julkaista, mutta silloin aineistoa ei voi myöskään käyttää. Jos aineisto sen sijaan julkaistaan käsittelemättä, on sen käytettävyys suuri, mutta niin on myös tietosuojarikkomuksen todennäköisyyskin. Hyvä anonymisointimenetelmä heikentää mahdollisimman vähän aineiston käytettävyyttä ja saavuttaa samanaikaisesti riittävän yksityisyyden tason (suurin sallittu riski).

Tässä tutkielmassa tarkastellaan henkilötietoa sisältävän *rivitason tiedon* (eng. micro data) anonymisointimenetelmiä, niiden teoriaa ja käytännön soveltuvuutta Cohort Study of Mobile Phone Use and Health (COSMOS) -tutkimuksen aineistoon. Tutkielmassa käsitellään viittä anonymisointimenetelmää, joista k-anonymiteetti [5] sekä tästä johdettu l-diversiteetti [7] ovat usein sovellettuja tilastollisia tietosuojamenetelmiä [8], jotka perustuvat aineiston karkeistamiseen. Myös Findata soveltaa k-anonymiteettiin perustuvaa minimifrekvenssiperiaatetta [4]. Lisäksi tutkielmassa tarkastellaan kahta spektraaliseen anonymisaatioon [9] perustuvaa menetelmää, kohinan lisäämistä ja sarakepermutaatiota. Spektraalinen sarakepermutointi säilyttää aineiston muuttujien ensimmäiset ja toiset momentit ainakin likimain ja spektraalinen kohina toimii sille verrokkina [9]. Monet tilastollisen päättelyn menetelmät perustuvat näiden momenttien estimointiin, minkä takia menetelmiä haluttiin tes-

tata käytännössä. RSA-algoritmi sen sijaan on yleisesti sovellettu salausalgoritmi [10] ja sen soveltamisesta tietojen anonymisointiin on saatu lupaavia tuloksia [11].

2 Tietosuojakäsitteitä

Tässä luvussa määritellään tutkielman kannalta keskeiset tietosuojakäsitteet, kuten erilaiset tieto- ja muuttujatyypit, anonyymin ja pseudonyymin tiedon määritelmät, sekä mitä menetelmiä tutkielmassa käytetään anonymisoinnin onnistumisen – eli riittävän tietosuojan saavuttamisen – mittaamiseen.

2.1 Tieto- ja muuttujatyypit

Tutkielmassa keskitytään henkilötietoa sisältävän rivitason tiedon (määritelmä 2) anonymisointimenetelmiin. Henkilötietoa sisältävä aineisto luokitellaan rivitason tiedoksi, kun aineiston jokaisella rivillä on kuvattu yhden henkilön tiedot. Taulukossa 1 on esitetty kuvitteellinen esimerkki tällaisesta aineistosta.

Määritelmä 2. *Rivitason tieto*

Aineistoa M, jossa on n *havaintoyksikköä* (eng. records) ja p *muuttujaa* (eng. variables, attributes), kutsutaan rivitason tiedoksi, kun jokainen havaintoyksikkö $i = 1, \ldots, n$ on esitetty omalla rivillään ja muuttujat $M_j,\ j = 1 \ldots, p$ muodostavat aineiston sarakkeet. Muuttujat voivat olla sekä kvantitatiivisia että kvalitatiivisia. [1]

Taulukko 1: Esimerkki rivitason tiedosta.

Nimi	Sukupuoli	Syntymävuosi	Maakunta	Ansiotulo	Paino (kg)
Matti Meikäläinen	Mies	1975	Uusimaa	33600	85
Markku Mäkinen	Mies	1991	Kymenlaakso	29500	65
Maija Mäkelä	Nainen	1985	Lappi	32400	70

Rivitason tiedon muuttujat voidaan jakaa neljään, osin päällekäiseen kategoriaan [1, 12]:

1. *Suorat tunnisteet* (eng. direct identifier). Suorat tunnisteet ovat muuttujia, jotka suoraan määrittelevät kyseisen henkilön täysin. Tällaisia muuttujia ovat esimerkiksi henkilön koko nimi ja henkilötunnus.

2. *Kvasitunniste* (eng. quasi-identifier) (määritelmä 3). Kvasitunniste on joukko aineiston muuttujia, jonka avulla havaintoyksikkö voidaan tunnistaa populaatioista. Mikä tahansa aineiston muuttuja voi kuulua kvasitunnisteeseen. Esimerkiksi ammatti ja paikkakunta voivat yhdessä muodostaa kvasitunnisteen, sillä nämä voivat joissain tapauksissa auttaa rajaamaan tiedon yksittäiseen havaintoyksikköön.

3. *Luottamukselliset muuttujat* (eng. confidential attributes). Luottamukselliset muuttujat ovat muuttujia, joiden paljastumisesta voi koitua haittaa havaintoyksikölle. Tällaisia muuttujia voivat olla esimerkiksi henkilön vakaumus ja terveystiedot.

4. *Julkiset muuttujat* (eng. non-confidential attributes). Julkiset muuttujat ovat muuttujia, joiden arvot eivät ole luottamuksellisia. Julkisten muuttujien yhdisteet saattavat kuitenkin muodostaa kvasitunnisteen; esimerkiksi Verohallinnolta saatavat nimi, syntymävuosi ja ansiotulot saattavat yksilöidä henkilön.

Muuttujien erottelu on tärkeää, sillä suorien tunnisteiden poistamisen lisäksi monet anonymisointimenetelmät perustuvat kvasitunnisteiden tai luottamuksellisten muuttujien tunnistamiseen ja käsittelemiseen. Tarkastellaan esimerkissä 1 taulukon 1 rivitason tiedon muuttujien jakamista näihin luokkiin.

Esimerkki 1. *Rivitason tiedon muuttujien jako*

Taulukon 1 muuttujat voidaan jakaa edellä mainittuihin luokkiin esimerkiksi seuraavasti: nimi on suora tunniste, paino on luottamuksellinen muuttuja ja kvasitunnisteen muodostavat sukupuoli sekä Verohallinnon julkisista tiedoista saatavat syntymävuosi, maakunta ja ansiotulot.

Määritelmä 3. *Kvasitunniste*

Olkoon P populaatio, $p \in P$ havaintoyksikkö ja Ω muuttuja-avaruus. Olkoon $P_0 \subseteq P$ otos ja $\Omega_0 \subseteq \Omega$ otoksessa havaitut muuttujat. Merkitään muuttujajoukon $Q \subseteq \Omega_0$ kaikkien mahdollisten arvojen joukkoa $D(Q)$. Funktio $f_Q : P \to D(Q)$ kuvaa havaintoyksikön p tämän muuttujajoukolle Q saamiksi arvoiksi $f_Q(p)$. Tällöin alkukuva $f_Q^{-1} : D(Q) \to 2^P$ antaa kaikkien niiden havaintoyksiköiden joukon $f_Q^{-1}(d)$ jotka saavat arvon d muuttujajoukolle Q.

Osajoukkoa $Q \subseteq \Omega_0$ kutsutaan kvasitunnisteeksi jos

$$\exists p \in P_0, \quad \text{jolle} \quad f_Q^{-1}(f_Q(p)) = \{p\},$$

eli jos otoksessa P_0 on havaintoyksikkö, jonka arvot muuttujilla Q yksilöivät tämän havaintoyksikön uniikisti populaatiossa P.

Kvasitunnisteiden määrittelyssä on siis tärkeää tunnistaa otoksen taustalla oleva populaatio ja aineistossa olevien muuttujien suhde oletettuun populaatiojakaumaan. Jos otos esimerkiksi koostuu pelkistä turkulaisista, niin on epätodennäköistä pystyä yksilöimään yhtä henkilöä pelkän sukupuolen ja iän perusteella, jos aineistossa ei ole iältään poikkeavia havaintoja, joiden esiintyminen populaatiossa on harvinaista. Sen sijaan, jos aineistossa on iältään poikkeavia havaintoja, niin ikä voi yksinään muodostaa kvasitunnisteen, jos joku aineiston henkilö on suoraan tunnistettavissa sen avulla populaatiosta. Jos otos on poimittu useilta eri paikkakunnilta, saattaa ikä, sukupuoli ja henkilön asuinpaikkakunta muodostaa kvasitunnisteen, sillä esimerkiksi harvaan asutulla paikkakunnalla jotkin yhdistelmät voivat olla harvinaisempia.

Toinen oleellinen seikka kvasitunnisteiden määrittelyssä on ulkopuolisen tiedon saatavuus sekä tietojen yhdistämisen laillisuus, sillä käytännössä yksilöiden tunnistaminen kvasitunnisteiden avulla tapahtuu yhdistelemällä aineiston tietoa toisiin aineistoihin. Tämä tunnistaminen voi tapahtua jo kerralla, eli kvasitunnisteet löytyvät suoraan jostain toisesta aineistosta, josta löytyy myös havaintoyksiköiden jokin suora tunniste tai sitten tunnistaminen tapahtuu useamman aineiston yhdistelyn seurauksena. Jos aineiston käyttöä ei ole rajoitettu tai ei ole erikseen kielletty sen yhdistämistä muihin ulkopuolisiin aineistoihin, voi seurauksena olla tarkoituksellinen

tai vahingossa tapahtuva havaintoyksiköiden tunnistaminen. Useimmiten aineistojen käyttötarkoitusta ja yhdistelyä on kuitenkin rajoitettu ja tietoturvaloukkaukset tapahtuvat laittomin keinoin. Aineiston tietosuojaa ja kvasitunnisteita määrittäessä olisikin siis hyvä ottaa huomioon myös tällaiset tilanteet noudattaen EU:n yleisen tietosuoja-asetuksen anonymisointia koskevaa kohtuullisuusperiaatetta (määritelmä 4).

Kvasitunnisteeseen valittavien muuttujien määrittelyyn ei siis ole olemassa mitään yleispätevää menetelmää. Lisäksi, mitä enemmän muuttujia kvasitunnisteessa on, sitä enemmän aineiston käytettävyys yleensä kärsii, kun useampaan muuttujaan joudutaan kohdistamaan tilastollisen tietosuojan mukaisia menetelmiä. Toisaalta, kvasitunnisteen virheellinen määrittely saattaa aineiston julkaisun myötä johtaa identiteettirikkomukseen. Huomionarvoista on sekin, että tiedon määrän lisääntyessä myös yhdistettävien aineistojen määrä kasvaa ja siksi kvasitunnisteisiin perustuvat tilastollisen tietosuojan menetelmät eivät tarjoa yhtä hyvää suojaa ajan yli kuin jotkin toisiin periaatteisiin pohjautuvat tietosuojamentelmät [13].

Rivitason tiedon lisäksi on olemassa *taulukkomuotoista tietoa* (eng. tabular data). Taulukkomuotoinen tieto voidaan muodostaa rivitason tiedon perusteella. Seuraavaksi esitellään kaksi taulukkomuotoista tietotyyppiä: frekvenssitaulut ja keskiarvotaulut. Frekvenssitauluissa, jokainen solu vastaa niiden vastaajien lukumäärää, jotka kuuluvat kyseiseen soluun ja keskiarvotauluissa jokainen solun arvo vastaa tietyn muuttujan keskiarvoa toisen muuttujan tasojen kesken (esimerkki 2). [1]

Esimerkki 2. *Taulukkomuotoinen tieto*

Taulukon 1 rivitason tiedosta voidaan muodostaa esimerkiksi seuraavat frekvenssi- ja keskiarvotaulut.

(a) Frekvenssitaulu

Sukupuoli	Lukumäärä
Mies	2
Nainen	1

(b) Keskiarvotaulu

Sukupuoli	Painon keskiarvo (kg)
Mies	75
Nainen	70

2.2 Anonyymi ja pseudonyymi tieto

Anonymisointia varten täytyy luonnollisesti määritellä, mitä anonyymi tieto tarkoittaa. Tässä tutkielmassa anonyymin ja pseudonyymin tiedon (määritelmät 4 ja 5) määrittelyssä sovelletaan sekä tietosuoja-asetusta että Findatan määritelmiä anonymisoinnille ja pseudonymisoinnille [2, 4].

Määritelmä 4. *Anonyymi tieto*

Tietoa kutsutaan anonyymiksi, jos yksittäistä henkilöä ei voida suoraan tai välillisesti tunnistaa, ainoastaan yhtä henkilöä koskevia päätelmiä ei voi tehdä, sekä on mahdotonta tai kohtuuttoman vaikeaa palauttaa tieto muotoon, josta yksittäinen henkilö olisi tunnistettavissa. [2, 4]

Määritelmä 5. *Pseudonyymi tieto*

Tietoa kutsutaan pseudonyymiksi, jos henkilötietoja ei voida yhdistää tiettyyn havaintoyksikköön käyttämättä jotain lisätietoa. Lisätieto voi olla esimerkiksi erillään säilytettävä salausavain tai sanakirja, jolla alkuperäisen aineiston suorat tunnisteet salattiin ja jolla salaus voidaan purkaa. [2, 4]

Anonyymi ja pseudonyymi tieto suojaavat havaintoyksiköitään. Tästä eteenpäin aineiston muokaamista, joka tekee aineistosta anonyymia tietoa kutsutaan anonymisoinniksi. Vastaavasti pseudonyymin aineiston tapauksessa käytetään termiä pseudonymisointi. Anonymisointi eroaa pseudonymisoinnista siinä, ettei pseudonymisointi estä peruuttamattomasti luonnollisen henkilön tunnistamista. Pseudonymisoinnissa suorat tunnisteet säilytetään erillään pseudonymisoidusta tiedosta ja aineistossa käytetään pseudotunnisteita, kuten juoksevaa numerointia tai muuta vastaavaa avainta. Tämä pseudotunniste on yhdistettävissä alkuperäisiin erillään säilytettäviin suoriin tunnisteisiin, jolloin havaintoyksikkö on siis edelleen tunnistettavissa ja aineisto voidaan palauttaa alkuperäiseen muotoon. [4]

Toinen lähestymistapa tuottaa anonyymiä tietoa on tehdä *synteettisiä aineistoja*, joissa alkuperäisellä tunnisteellisella aineistolla opetetaan malli, josta voidaan tuottaa uusia, alkuperäisen kaltaisia havaintoja. Synteettisen aineiston tietosuoja perustuu sovelletun menetelmän kykyyn suojata alkuperäiset havainnot osana tiedon tuottamisprosessia [14]. Tässä tutkielmassa ei käsitellä synteettisiä aineistoja.

Keskeistä on, että tietosuoja-asetusta tai tietosuojasäännöksiä ei sovelleta anonyymiin tietoon, mutta sen sijaan pseudonyymi tieto luetaan edelleen henkilötiedoksi [2, 6].

2.3 Tietosuojarikkomukset

Riittämättömästi anonymisoidun aineiston julkaiseminen saattaa johtaa paljastumiseen tai tietosuojarikkomukseen (määritelmä 6). Tietosuojarikkomukset voidaan jakaa karkeasti seuraaviin luokkiin: *indentiteettirikkomuksiin* [1, 12] (eng. identity disclosure), *ominaisuusrikkomuksiin* [1, 12] (eng. attribute disclosure), *päättelyrikkomuksiin* [9] (eng. inferential / prediction disclosure) ja *osallisuusrikkomuksiin* [12] (eng. membership disclosure).

Määritelmä 6. *Tietosuojarikkomus*

Tietosuojarikkomuksessa tai paljastumisessa aineistossa olevasta havaintoyksiköstä on mahdollista oppia jotain, mitä ennen aineiston julkaisua ei tiedetty. [1]

Identeettirikkomuksessa hyökkääjä voi yhdistää aineiston havaintoyksikön luonnolliseen henkilöön [1, 12]. Jos esimerkiksi pseudonymisointiavaimet vuotavat tai kvasitunnistetta ei ole käsitelty kunnolla, voi identiteettirikkomus tapahtua. Ominaisuusrikkomuksessa hyökkääjä onnistuu yhdistämään havaintoyksikön luottamuksellisen muuttujan arvon luonnolliseen henkilöön [1, 12]. Ominaisuusrikkomus voi tapahtua esimerkiksi, jos aineistoon muodostuu kvasitunnisteen perusteella havaintoyksiköiden osajoukko, jossa kaikilla havaintoyksiköillä on sama arvo jollain luottamuksellisella muuttujalla. Tällöin hyökkääjän ei tarvitse edes yksilöidä kohdettaan voidakseen päätellä tästä jotain uutta. Osallisuusrikkomuksessa hyökkääjä pystyy suurella todennäköisyydellä päättelemään, kuuluuko luonnollinen henkilö aineistoon

vai ei [12]. Osallisuusrikkomuksesta voi koitua haittaa, jos kyseiseen aineistoon kuuluminen itsessään paljastaa jotain luottamuksellista. Päättelyrikkomuksessa hyökkääjä onnistuu päättelemään luonnollisen henkilön muuttujien likimaisen arvojoukon [9]. Anonymisoinnin ja pseudonymisoinnin tarkoituksena on suojata tietosuojarikkomuksia vastaan. Pseudonymisointi suojaa ainoastaan identiteettirikkomusta vastaan ja anonymisoinnin antama suoja vaihtelee käytetystä menetelmästä riippuen.

Tässä tutkielmassa keskitytään identiteetti- ja päättelyrikkomukseen ja luvussa 2.6 esitetään tarkemmin työssä käytetyt mittarit näiden riskin arvioimiseksi.

2.4 Samankaltaisuus

Anonymisointi muuttaa aina aineiston rakennetta. Tämä tarkoittaa, että alkuperäinen aineisto ja sen anonymisoitu versio eivät ole kaikin puolin samankaltaisia. Samankaltaisuudella tarkoitetaan alkuperäisen ja anonymisoidun aineiston muuttujien todennäköisyysjakaumien samankaltaisuutta, jota voidaan mitata useiden erilaisten kvantitatiivisten ja kvalitatiivisten menetelmien avulla.

Tässä tutkielmassa samankaltaisuutta mitataan hajontakuvioita, keskiarvoja, variansseja ja korrelaatioita tarkastelemalla. Nämä tarkastelukohteet valittiin, sillä ne kuvaavat yksiulotteisten jakaumien sijaintia ja hajontaa, sekä muuttujien välisiä riippuvuuksia: asioita, joista tutkimuksessa ollaan yleensä kiinnostuneita. Lisäksi, monet menetelmät perustuvat näiden suureiden estimoimiseen. Hajontakuviot kuvan 2 tapaan esitetään anonymisointimentelmien teorioiden yhteydessä luvussa 3 ja muut tarkastelut tapahtuvat luvussa 4. Hajontakuviot perustuvat tutkielmassa käsiteltävään COSMOS-tutkimuksen [15] aineistoon, joka esitellään tarkemmin luvussa 4.

Aineistojen vertailun mahdollistamiseksi, sekä alkuperäisen että anonymisoitujen aineistojen numeeriset muuttujat normalisoidaan alkuperäisen aineiston muuttujien keskiarvojen ja -hajontojen avulla (määritelmä 7), mikä tekee aineistojen muuttujista yksiköttömiä. Ilman normalisointia, muuttujien mittayksiköt vaikuttaisivat samankaltaisuuden estimaatteihin.

Määritelmä 7. *Aineistojen normalisointi*

Olkoon M alkuperäinen, määritelmän 2 mukainen aineisto ja $\hat{M}$ sen anonymisoitu versio. Olkoon $\mu_j^{(M)}$ ja $\sigma_j^{(M)}$ alkuperäisen aineiston muuttujan M_j keskiarvo ja -hajonta. Nyt kaikilla $j = 1, \ldots, p$ muuttujat M_j ja $\hat{M}_j$ normalisoidaan keskiarvon $\mu_j^{(M)}$ ja keskihajonnan $\sigma_j^{(M)}$ avulla:

$$N_j = \frac{M_j - \mu_j^{(M)}}{\sigma_j^{(M)}}, \text{ ja vastaavasti}$$

$$\hat{N}_j = \frac{\hat{M}_j - \mu_j^{(M)}}{\sigma_j^{(M)}}.$$

Normalisoiduista muuttujista koostuville aineistoille N ja $\hat{N}$ lasketaan muuttujien keskiarvot, varianssit ja korrelaatiot, joita verrataan aineistojen välillä (määritelmä 8).

Määritelmä 8. *Keskiarvojen, varianssien ja korrelaatioiden samankaltaisuus*

Olkoon N normalisoitu alkuperäinen aineisto ja $\hat{N}$ normalisoitu anonymisoitu aineisto. Olkoon $\mu_j^{(N)}$ normalisoidun alkuperäisen aineiston muuttujan N_j keskiarvo ja $\mu_j^{(\hat{N})}$ normalisoidun anonyymin aineiston muuttujan $\hat{N}_j$ keskiarvo. Nyt keskiarvojen samankaltaisuuden S_μ odotusarvo ja keskihajonta lasketaan seuraavasti:

$$\mathrm{E}[S_\mu] = \frac{1}{p}\sum_{j=1}^{p} \mu_j^{(\hat{N})} - \mu_j^{(N)} \text{ ja}$$

$$\mathrm{kh}[S_\mu] = \sqrt{\frac{1}{p}\sum_{j=1}^{p}\left(\mu_j^{(\hat{N})} - \mu_j^{(N)} - \mathrm{E}[S_\mu]\right)^2}.$$

Olkoon $\sigma_j^{2(N)}$ normalisoidun alkuperäisen aineiston muuttujan N_j varianssi ja $\sigma_j^{2(\hat{N})}$ normalisoidun anonyymin aineiston muuttujan $\hat{N}_j$ varianssi. Nyt varianssien samankaltaisuuden S_{σ^2} odotusarvo ja keskihajonta lasketaan seuraavasti:

$$\mathrm{E}[S_{\sigma^2}] = \frac{1}{p}\sum_{j=1}^{p} \sigma_j^{2(\hat{N})} - \sigma_j^{2(N)} \text{ ja}$$

$$\mathrm{kh}[S_{\sigma^2}] = \sqrt{\frac{1}{p}\sum_{j=1}^{p}\left(\sigma_j^{2(\hat{N})} - \sigma_j^{2(N)} - \mathrm{E}[S_{\sigma^2}]\right)^2}.$$

Olkoon $\rho^{(N)}$ normalisoidun alkuperäisen aineiston korrelaatiomatriisi ja $\rho^{(\hat{N})}$ normalisoidun anonyymin aineiston korrelaatiomatriisi. Nyt korrelaatioiden samankaltaisuuden S_ρ odotusarvo ja keskihajonta lasketaan seuraavasti:

$$\mathrm{E}[S_\rho] = \frac{2}{p(p-1)}\sum_{i=1}^{p}\sum_{j>i}^{p} \rho_{ij}^{(\hat{N})} - \rho_{ij}^{(N)} \text{ ja}$$

$$\mathrm{kh}[S_\rho] = \sqrt{\frac{2}{p(p-1)}\sum_{i=1}^{p}\sum_{j>i}^{p}\left(\rho_{ij}^{(\hat{N})} - \rho_{ij}^{(N)} - \mathrm{E}[S_\rho]\right)^2}.$$

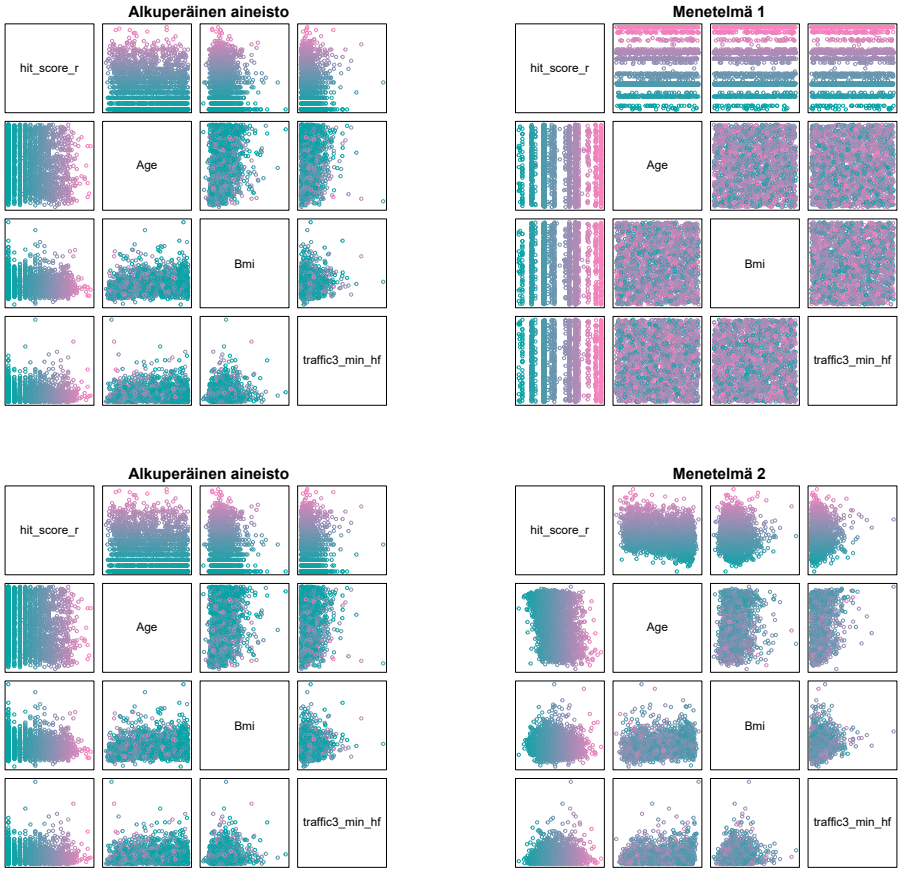
Sekä odotusarvojen että keskihajontojen ollessa lähellä nollaa, on anonymisointi aiheuttanut vain pieniä muutoksia muuttujien odotusarvoihin, variansseihin ja korrelaatioihin. Jos arvot ovat taas suuria, on anonymisointi aiheuttanut isompia muutoksia kyseisiin arvoihin.

Anonymisointimenetelmät voidaan laittaa samankaltaisuudeltaan paremmuusjärjestykseen erotusten odotusarvojen ja keskihajontojen perusteella. Tulokset esitetään *samankaltaisuustaulukoiden* avulla, joita havainnollistetaan taulukossa 3.

Taulukko 3: Anonymisoitujen aineistojen samankaltaisuustaulukko. Aineistosta M on tehty kaksi anonymisoitua versiota $\hat{M}_1$ ja $\hat{M}_2$. Aineiston $\hat{M}_1$ muuttujien odotusarvojen keskierotus alkuperäisen aineiston muuttujien odotusarvoista on 0 ja erotusten keskihajonta on 0, eli muuttujien odotusarvot ovat säilyneet hyvin samankaltaisina. Vastaavasti varianssien keskierotus on 0 ja keskihajonta on 0,05, eli muuttujien varianssit ovat säilyneet keskimäärin samankaltaisina, mutta muuttujakohtaista vaihtelua variansseissa kuitenkin esiintyy. Korrelaatioiden keskierotus alkuperäisen aineiston korrelaatioista on 0,01 ja keskihajonta on myös 0,01, eli korrelaatiot ovat muuttuneet vähän. Aineistolle $\hat{M}_2$ pätee samankaltaiset tulkinnat, mutta poikkeamat alkuperäisestä aineistosta ovat suurempia. Lisäksi, keskiarvojen poikkeama on negatiivinen, eli anonymisointi on keskimäärin laskenut muuttujien arvoja. Taulukkojen perusteella aineisto $\hat{M}_1$ muistuttaa lähemmin alkuperäistä aineistoa kuin $\hat{M}_2$.

Aineisto	E[S_μ]	kh[S_μ]	E[S_{σ^2}]	kh[S_{σ^2}]	E[S_ρ]	kh[S_ρ]
$\hat{M}_1$	0	0	0	0,05	0,01	0,01
$\hat{M}_2$	-0,3	0,15	0,3	0,1	0,1	0,05

Samankaltaisuutta mitataan samankaltaisuustaulukoiden lisäksi hajontakuvioita tarkastelemalla. Tällä tavalla saadaan visuaalisesti tietoa siitä, miten menetelmät ovat muuttaneet aineiston muuttujien välisiä riippuvuuksia. Hajontakuvioiden tulkintaa esitellään kuvassa 2.



Kuva 2: Alkuperäisen ja kahden anonymisoidun aineiston hajontakuviot. Kuvassa vasemmalla esitetään COSMOS-aineiston neljän muuttujan (hit_score_r, age, bmi ja traffic3_min_hf) mukainen hajontakuvio ja oikealla anonymisoitujen aineistojen mukaiset hajontakuviot samoilla muuttujilla. Menetelmä 2 (oikea alakulma) säilyttää aineiston hajontakuvion alkuperäisen kaltaisena paljon paremmin kuin menetelmä 1 (oikea yläkulma), eli muuttujien väliset riippuvuudet säilyvät paremmin ennallaan menetelmässä 2. Vaaleanpunainen väri hajontakuvion pisteellä tarkoittaa suurta arvoa muuttujalla hit_score_r kyseisessä aineistossa ja sininen pientä arvoa.

2.5 Käytettävyys

Anonymisointi vähentää aina aineiston käytettävyyttä, eli kykyä tehdä sillä esimerkiksi tilastollista päättelyä, mallinnusta, luokittelua tai ennustamista. Käytettävyys ja samankaltaisuus ovat osittain päällekkäisiä käsitteitä: korkea samankaltaisuus implikoi yleensä korkeaa käytettävyyttä, mutta päinvastaista päätelmää ei kuitenkaan välttämättä voida tehdä. Esimerkiksi luokittelussa, aineiston jokaisen sarakkeen kertominen jollain vakiolla ei heikentäisi sen käytettävyyttä, vaikka samankaltaisuus muuttuisikin täysin.

Tässä tutkielmassa anonymisoitujen aineistojen käytettävyyttä tarkastellaan lineaarisella ja logistisella regressiolla, sillä COSMOS-tutkimuksessa [15] binääristä vastetta mallinnettiin juuri logistisella regressiolla ja lineaarinen regressio toimii samankaltaisena mallina jatkuvalle vasteelle. Lineaarisen regression malleissa tarkas-

tellaan piste-estimaatteja ja niiden luottamusvälejä sekä selitysastetta (eng. coefficient of determination). Logistisen regression malleissa tarkastellaan selitysasteen sijasta *AUC-pistemäärää* (eng. area under the curve).

Määritellään yleistetty lineaarinen malli kirjan [16] mukaan.

Määritelmä 9. *Yleistetty lineaarinen malli*

Yleistetty lineaarinen malli koostuu kolmesta osasta.

1. Eksponenttiperheeseen kuuluva satunnainen vastemuuttuja $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, jonka komponentit oletetaan toisistaan riippumattomiksi.

2. Lineaarinen ennustin $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$, jossa $\boldsymbol{X}$ on selittävien muuttujien matriisi, jossa on n riviä ja $p+1$ saraketta ja $\boldsymbol{\beta}$ on parametrivektori.

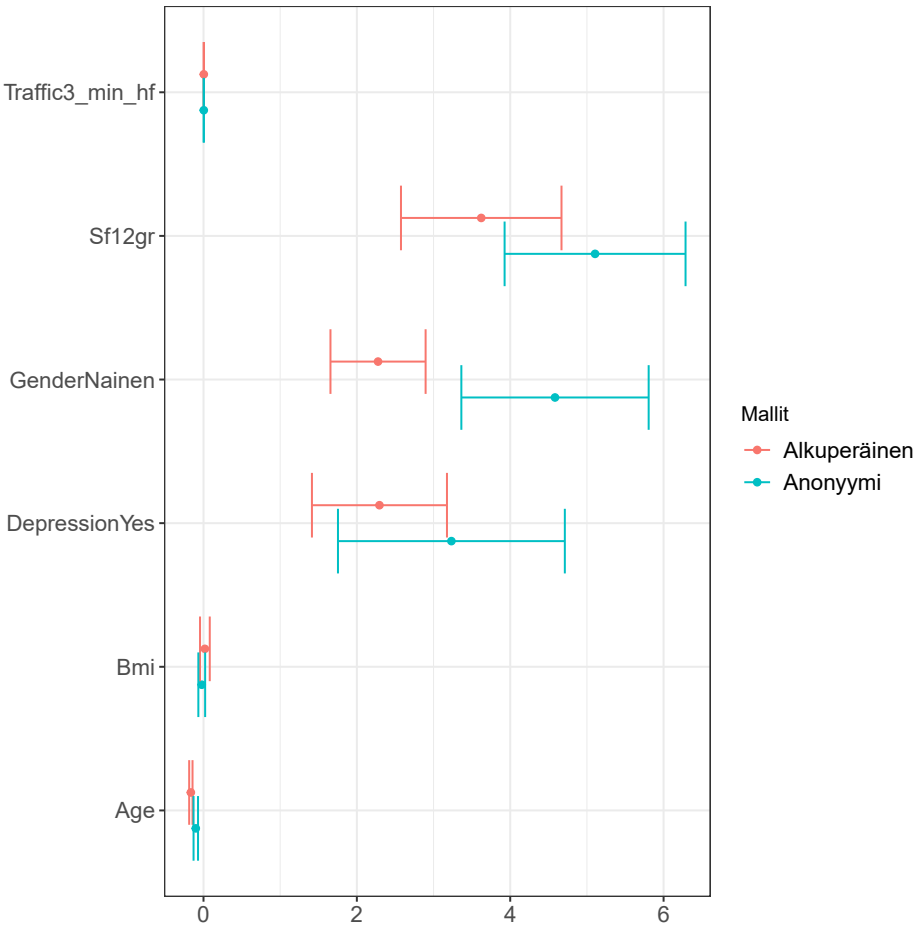
3. Monotoninen ja differentioituva linkkifunktio g, joka yhdistää satunnaisvektorin $\boldsymbol{y}$ arvot lineaariseen ennustimeen.

Olkoon $\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{y}]$. Nyt siis

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = g\left(\mathrm{E}\left[\boldsymbol{y}\right]\right) = \boldsymbol{X}\boldsymbol{\beta}.$$

Kaikilla eksponenttiperheen malleilla on olemassa *luonnollinen parametri*. Normaalijakaumalla – jonka varianssi on tunnettu – luonnollinen parametri on keskiarvo ja binomijakaumalla logaritminen vetosuhde. Linkkifunktiota, joka kuvaa odotusarvon $\boldsymbol{\mu}$ luonnolliseksi parametriksi kutsutaan *kanoniseksi linkkifunktioksi*. Lineaarisessa regressiossa kanoninen linkkifunktio on identiteettifunktio, $g(x) = x$, kun $x \in \mathbb{R}$ ja logistisessa regressiossa logit-muunnos, $g(x) = \log(\frac{x}{1-x})$, kun $x \in (0, 1)$. [16]

Regressiomalleissa ollaan kiinnostuttu tarkastelemaan selittävien muuttujien yhteyttä vastemuuttujaan tarkastelemalla estimoidun mallin parametrivektorin estimaatteja sekä näiden luottamusvälejä. *Luottamusvälikuvaaja* (kuva 3) kuvaa sekä alkuperäiselle että anonymisoidulle aineistolle sovitetun mallin parametrien piste-estimaatit ja niiden luottamusvälit samaan kuvaan ja antaa visuaalisesti tietoa, miten anonymisointi on vaikuttanut aineistojen pohjalta tehtävien regressiomallien piste-estimaatteihin ja luottamusväleihin. Siirtyneet piste-estimaatit ja leveydeltään muuttuneet luottamusvälit kertovat käytettävyyden heikkenemisestä.



Kuva 3: Alkuperäisellä ja anonyymillä aineistolla opetettu lineaarinen regressiomallin parametrien estimaatit sekä näiden 95 %:n luottamusvälit. Anonyymillä aineistolla sukupuolen, hyvinvoinnin ja masennuksen piste-estimaatit eroavat hieman alkuperäisistä ja usean piste-estimaatin luottamusvälit ovat leveämpiä. Anonymisointimenetelmä laskee siis aineiston käytettävyyttä.

Lineaarisen regression tapauksessa estimaattien ja luottamusvälien lisäksi käytettävyyttä mitataan selitysasteella R^2 (määritelmä 10), joka kertoo kuinka hyvin malli ennustaa vastemuuttujan arvoja verrattuna siihen, että jokaisen havainnon ennusteena käytettäisiin vasteiden keskiarvoa. $R^2 = 0$ tarkoittaa, että malli ei ole parempi kuin keskiarvon ennustaminen kaikille havaintoyksiköille. $R^2 = 1$ tarkoittaa, että malli sovittaa havaintoyksiköiden arvot täydellisesti. Selitysasteen lasku kertoo käytettävyyden heikkenemisestä.

Määritelmä 10. *Selitysaste*

$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{\mu}_i)^2}{\sum_i^n (y_i - \bar{y})^2},$$

jossa $\hat{\mu}_i$ on estimoitu odotusarvo havainnolle y_i ja $\bar{y}$ on havaintojen otoskeskiarvo. [16]

Selitysasteiden absoluuttisista muutoksista on vaikea päätellä yksinään, onko

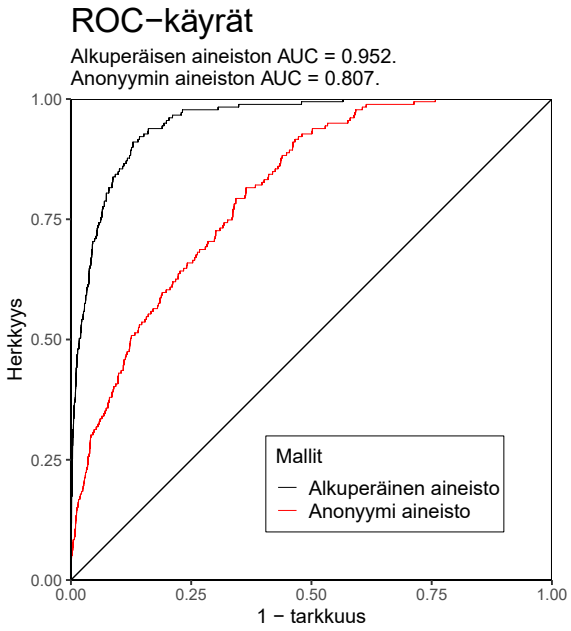
anonymisointi onnistunut vai ei. Sen avulla voidaan kuitenkin asettaa eri anonymisointimenetelmiä paremmuusjärjestykseen, jossa pienin selitysasteen laskeminen tarkoittaa onnistuneinta anonymisointia käytettävyyttä mitattaessa. Taulukossa 4 esitellään *selitysastetaulukko*.

Taulukko 4: Selitysastetaulukko. Ensimmäiselle anonymisoidulle aineistolle ($\hat{M}_1$) sovitetun lineaarisen regressiomallin selitysaste on 0,25 yksikköä pienempi kuin alkuperäiselle aineistolle sovitetulla mallilla. Anonymisointimenetelmä 1 on siis heikentänyt aineiston käytettävyyttä. Toiselle anonymisoidulle aineistolle ($\hat{M}_2$) sovitetun lineaarisen regressiomallin selitysaste on 0,4 yksikköä pienempi, eli anonymisointimenetelmä 2 on heikentänyt aineiston käytettävyyttä vielä enemmän kuin menetelmä 1.

Aineisto	Selitysaste
$\hat{M}$	0,87
$\hat{M}_1$	0,62
$\hat{M}_2$	0,47

Logistisen regression tapauksessa selitysasteen sijasta tarkastellaan AUC-pistemäärää. AUC-pistemäärä on pinta-ala, joka jää *ROC-käyrän* (eng. receiver operating characteristic curve) alle. ROC-käyrä kuvaa luokittelijan herkkyyden $1-$tarkkuuden funktiona. AUC-pistemäärä kertoo kuinka hyvin luokittelija erottaa positiiviset ja negatiiviset tulokset toisistaan. Mitä suurempi AUC-pistemäärä on, sitä paremmin luokittelija pystyy erottamaan luokat toisistaan. AUC-pistemäärä vaihtelee 0 ja 1 välillä, missä 1 tarkoittaa täydellistä erottelukykyä ja 0,5 puolestaan vastaa satunnaista arvausta. AUC-pistemäärä on siis mittari luokittelijan – tässä tapauksessa logistisen regression – suorituskyvyn arvioimiseksi. [16]

Kuvassa 4 havainnollistetaan ROC-käyrien ja AUC-pistemäärien tulkintaa.



Kuva 4: Alkuperäiselle ja anonymisoidulle aineistolle sovitettujen logististen regressiomallien ROC-käyrät. Anonymisoidulle aineistolle sovitetun logistisen regressiomallin AUC-pistemäärä on 0,145 yksikköä pienempi kuin alkuperäiselle aineistolle sovitetulla mallilla, eli anonymisointi heikentää aineiston käytettävyyttä jonkin verran.

2.6 Yksityisyys

Anonymisoinnin tarkoituksena on saavuttaa sellainen yksityisyyden taso, että aineisto voidaan turvallisesti luovuttaa taholle, jolle ei voi antaa pseudonyymiä tai tunnisteellista henkilötietoa. Aineiston yksityisyydellä tarkoitetaan sen kykyä suojata siinä esiintyviä havaintoyksiköitä erilaisilta tietosuojarikkomuksilta (luku 2.3). Aineisto määritellään siis anonyymiksi, kun se on saavuttanut määritelmän 4 mukaisen yksityisyyden tason. Anonymiteetin toteaminen on kuitenkin hankalaa ja vaatii Suomessa aina Findatan tapauskohtaista arviointia. [3]

Tässä tutkielmassa keskitytään aiheen rajaamiseksi vain identiteetti- ja päättelyrikkomuksen riskin estimoimiseen käyttäen mittareina *tunnistusastetta* (eng. reidentification rate) (määritelmä 11) sekä Laskon ja Vinerbon [9] kehittämää *ennuste-etäisyyttä* (eng. prediction distance), *ennuste-epäselvyyttä* (eng. prediction ambiguity) ja *ennuste-epävarmuutta* (eng. prediction uncertainty) (määritelmät 12, 13 ja 14).

Määritelmä 11. *Tunnistusaste*

Tunnistusaste kertoo osuuden anonyymin aineiston riveistä, jotka voidaan yhdistää niitä vastaaviin alkuperäisiin riveihin, kun arvaukseksi valitaan havaintoyksikkö, johon on lyhin matka metriikassa s. Tunnistusaste lasketaan määritelmän 7 mukaan normalisoiduille aineistoille.

Olkoon M_Q alkuperäinen aineisto, jossa on pelkästään kvasitunnisteeseen kuuluvat muuttujat ja $\hat{M}_Q$ sen anonymisoitu versio. Anonyymin aineiston jokainen rivi

$\hat{M}^i{}_Q$, jossa $i \in \{1, \ldots, n\}$, vastaa alkuperäisen aineiston rivin $M^i{}_Q$ havaintoyksikköä. Olkoon s jokin metriikka. Nyt tunnistusaste

$$t(\hat{M}_Q, M_Q, s) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\text{argmin}_k s(\hat{M}^i{}_Q, M^k{}_Q)=i},$$

jossa $\mathbf{1}$ on indikaattorifunktio ja $k \in \{1, \ldots, n\}$.

Tässä tutkielmassa tunnistusasteen metriikkana käytetään euklidista metriikkaa, $s(p, q) = \sqrt{(p_1 - q_1)^2 + \ldots + (p_n - q_n)^2}$. Tunnistusastetta laskettaessa on tärkeää, että kaikki havaintoyksiköt löytyvät samoilta riveiltä molemmissa aineistoissa: suureessa ollaan kiinnostuttu siitä, minimoituuko anonymisoidun havaintoyksikön etäisyys itsensä vai jonkun muun havaintoyksikön kohdalla alkuperäisessä aineistossa. Koska pienintä etäisyyttä laskiessa voi tulla tasapelejä ja nämä tasapelit ratkaistaan arpomalla, lasketaan tunnistusaste 1000 kertaa tästä aiheutuvan satunnaisuuden takia. Tunnistusasteiden keskiarvot esitetään *tunnistusastetaulukoissa*, joita havainnollistetaan taulukossa 5.

Taulukko 5: Alkuperäisen ja anonymisoidun aineiston tunnistusastetaulukko. Alkuperäisen aineiston kaikki havaintoyksiköt voidaan tunnistaa kvasitunnisteen perusteella, sillä tunnistusasteiden keskiarvo on yksi. Anonymisoidussa aineistossa tunnistusasteiden keskiarvo on 0,05, eli viisi prosenttia anonymisoidun aineiston havaintoyksiköistä pystytään keskimäärin yhdistämään niitä vastaaviin alkuperäisen aineiston havaintoyksiköihin kvasitunnisteen perusteella. Mitä pienempi tunnistusaste, sitä paremman suoja anonymisoitu aineisto tarjoaa identiteettirikkomusta vastaan.

Aineisto	Tunnistusaste
M_Q	1
$\hat{M}_Q$	0,05

Tässä tutkielmassa päättelyrikkomuksen riskiä mitataan kolmella eri mittarilla: ennuste-etäisyydellä, -epäselvyydellä ja -epävarmuudella, jotka esitellään tarkemmin määritelmissä 12, 13 ja 14. Nämä mittarit kertovat, kuinka hyvin hyökkääjä onnistuu päättelemään alkuperäisen havaintoyksikön arvoja (etäisyys), kuinka varma hyökkääjä voi olla arvauksensa todenmukaisuudesta (epäselvyys) ja kuinka tärkeää on oikean valinnan tekeminen suunnilleen yhtä hyvien vaihtoehtojen joukosta (epävarmuus). Yleisesti riskimäärittelyssä oletetaan hyökkääjällä olevan jotain lisäinformaatiota, joka tässä tutkielmassa on koko alkuperäinen aineisto. Tämä on epärealistinen oletus, mutta tämän avulla voidaan määritellä *riittävän suojan* kriteeri, joka on riippumaton kvasitunnisteen valinnasta. Kriteerin toteutuessa, keskimääräiseen havaintoyksikköön i kohdistuu pienempi tai yhtä suuri päättelyrikkomuksen riski tämän kuulessa anonymisoituun aineistoon verrattuna tilanteeseen, jossa alkuperäinen aineisto julkaistaisiin ilman havaintoyksikköä i [9]. Jokaiselle määritelmän 7 mukaisesti normalisoidun alkuperäisen aineiston havaintoyksikölle i lasketaan ennuste-etäisyys, -epäselvyys ja -epävarmuus suhteessa sekä normalisoituun anonymisoituun aineistoon $\hat{M}$ että loppuun alkuperäisestä aineistosta M^{-i}. Laskettujen suureiden perusteella estimoidaan kaikkien havaintoyksiköiden ennuste-etäisyyksien,

-epäselvyyksien ja -epävarmuuksien kvantiilifunktiot suhteessa sekä alkuperäiseen että anonymisoituun aineistoon.

Määritelmä 12. *Ennuste-etäisyys*

Ennuste-etäisyys kertoo, kuinka lähelle todellista arvojoukkoa hyökkääjän paras arvaus osuu, ja pieni etäisyys tarkoittaa arvauksen osuneen lähelle todellista arvoa. Olkoon $M_{n\times p}$ alkuperäinen aineisto, $\hat{M}$ sen anonymisoitu versio, M^{-i} alkuperäinen aineisto ilman riviä i ja s jokin metriikka. Ennuste-etäisyys on etäisyys alkuperäisen aineiston rivistä j lähimpään riviin anonyymissä aineistossa (tai lähimpään riviin lopussa alkuperäisessä aineistossa) mitattuna metriikassa s. Toisin sanoen

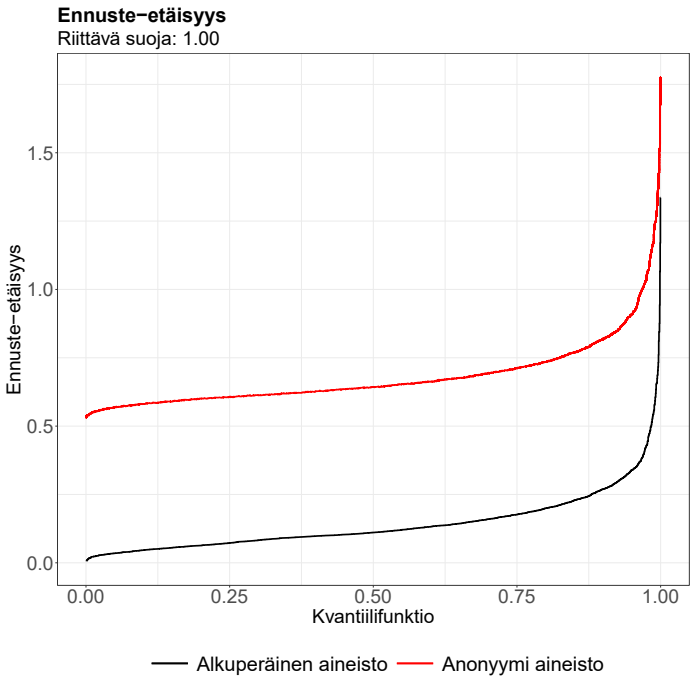
$$d\left(M^i, \hat{M}, s\right) = \min_k s\left(M^i, \hat{M}^k\right) \text{ ja}$$

$$d\left(M^i, M^{-i}, s\right) = \min_{k\neq i} s\left(M^i, M^k\right). \text{ [9]}$$

Tarkastellaan esimerkissä 3 ennuste-etäisyyksien kvantiilifunktioiden kuvaajia ja niiden tulkintaa.

Esimerkki 3. *Ennuste-etäisyys -kuvaaja*

Olkoon M jokin aineisto ja $\hat{M}$ sen anonymisoitu versio. Lasketaan ennuste-etäisyys jokaiselle alkuperäisen aineiston havaintoyksikölle i, muodostetaan etäisyyksien kvantiilifunktiot ja kuvataan kvantiilifunktiot samaan kuvaajaan.



Kuva 5: Ennuste-etäisyyksien kvantiilifunktiot anonyymille ja alkuperäiselle aineistolle. Koska anonymisoidun aineiston ennuste-etäisyyden kvantiilifunktio ei saa missään pienempiä arvoja kuin verrokkijakauma (riittävä suoja = 1), havaintoyksikön kuuluminen anonymisoituun aineistoon ei keskimäärin paranna hyökkääjän mahdollisuutta oppia tästä jotain uutta aineiston perusteella.

Alkuperäisen aineiston kvantiilifunktiot kertovat, kuinka vaikeaa aineistoon M kuulumattomasta havaintoyksiköstä olisi keskimäärin tehdä todenmukaisia päätelmiä, jos aineisto M julkaistaisiin ilman anonymisointia. Anonyymin aineiston kvantiilifuntiot kertovat, kuinka vaikeaa anonymisoituun aineistoon $\hat{M}$ *kuuluvasta* havaintoyksiköstä olisi keskimäärin tehdä todenmukaisia päätelmiä, jos aineisto julkaistaisiin anonymisoituna. Jos anonyymin aineiston mittareiden kvantiilifunktiot eivät saa pienempiä arvoja kuin alkuperäisen aineiston kvantiilifunktiot missään pisteessä, toteutuu riittävän suojan kriteeri. Tällöin anonymisoituun aineistoon kuulumisesta ei keskimäärin koidu liian suurta päättelyrikkomuksen riskiä. Toisin sanoen kriteeri vaatii, että kaikilla kolmella mittarilla $F_M^{-1}(p) \leq F_{\hat{M}}^{-1}(p)\ \forall p \in [0,1]$, jossa p on todennäköisyys ja F_M ja $F_{\hat{M}}$ ovat jonkin mittarin kertymäfunktiot alkuperäiselle ja anonymisoidulle aineistolle. [9]

Määritelmä 13. *Ennuste-epäselvyys*

Ennuste-epäselvyys kertoo, kuinka varma hyökkääjä voi olla siitä, että paras arvaus on oikein. Pieni epäselvyys kertoo, että paras arvaus on huomattavasti uskottavampi kuin k:nneksi paras arvaus. Suuri epäselvyys kertoo, että päätös k:n parhaan havaintoyksikön joukosta ei ole ilmiselvä. Olkoot M, $\hat{M}$, M^{-i} ja s määritelty kuten edellä. Olkoon ympäristöparametri $k \in \{2, \ldots, n\}$ kokonaisluku. Nyt ennuste-epäselvyys

$$c(M^i, \hat{M}, s, k) = \frac{s\left(M^i, \hat{M}^{(1)}\right)}{s\left(M^i, \hat{M}^{(k)}\right)} \text{ ja}$$

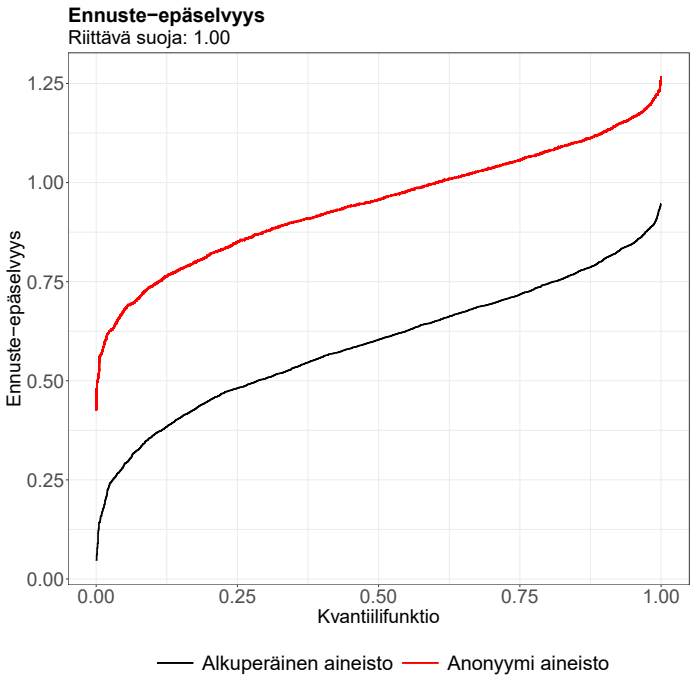
$$c(M^i, M^{-i}, s, k) = \frac{s\left(M^i, M^{-i(1)}\right)}{s\left(M^i, M^{-i(k)}\right)},$$

jossa $\hat{M}^{(l)}$ on i:nnettä alkuperäistä havaintoyksikköä l:nneksi lähin havaintoyksikkö anonymisoidussa aineistossa (tai lopussa alkuperäisessä aineistossa), ilmaisee osamäärän havaintoyksikköä i lähimmän havaintoyksikön etäisyyden ja k:nneksi lähimmän havaintoyksikön etäisyyden välillä. [9]

Tässä tutkielmassa ympäristöparametri $k = 5$, sillä se mukailee Findatan minimifrekvenssiperiaatetta [4] ja Lasko ja Vinterbo [9] suosittelevat ympäristöparametrin kooksi lukua viiden ja kymmenen väliltä. Tarkastellaan esimerkissä 4 ennuste-epävarmuuksien kvantiilifunktioiden kuvaajia ja niiden tulkintaa.

Esimerkki 4. *Ennuste-epäselvyys -kuvaaja*

Olkoon M jokin aineisto ja $\hat{M}$ sen anonymisoitu versio. Lasketaan ennuste-epäselvyys jokaiselle alkuperäisen aineiston havaintoyksikölle i, muodostetaan epäselvyyksien kvantiilifunktiot ja kuvataan kvantiilifunktiot samaan kuvaajaan.



Kuva 6: Ennuste-epäselvyyksien kvantiilifunktiot anonyymille ja alkuperäiselle aineistolle. Koska anonymisoidun aineiston ennuste-epäselvyyden kvantiilifunktio ei saa missään pienempiä arvoja kuin verrokkijakauma (riittävä suoja = 1), ei valinta viiden lähimmän havaintoyksikön välillä ole keskimäärin ilmiselvää, sillä hyökkääjällä on tavallisesti viisi noin yhtä hyvää vaihtoehtoa.

Tässä tutkielmassa riittävän suojan kriteerin toteutumista mitataan laskemalla osuus anonyymien kvantiilifunktioiden pisteistä, jotka ovat suurempia tai yhtä suuria kuin alkuperäiset kvantiilifunktiot näissä pisteissä. Jos näiden pisteiden osuus on 1, toteutuu riittävän suojan kriteeri kyseisellä mittarilla. Jos osuus kyseisellä mittarilla on 0, ei kriteeri toteudu lainkaan. Korkeampi osuus tarkoittaa siis parempaa suojaa päättelyrikkomusta vastaan. Koska riittävän suojan kriteeri perustuu pelkästään aineistoille estimoituihin kvantiilifunktioihin, ei kriteeri ota kantaa yksilötason päättelyrikkomuksen riskeihin tai niiden muutoksiin alkuperäisen ja anonymisoidun aineiston väillä. Tämä tarkoittaa, että yksilön suoja päättelyrikkomusta vastaan saattaa olla huonompi anonymisoidussa aineistossa, vaikka riittävän suojan toteutuessa anonymisoitu aineisto keskimäärin tarjoaakin paremman suojan päättelyrikkomusta vastaan.

Määritelmä 14. *Ennuste-epävarmuus*

Ennuste-epävarmuus kertoo k:n parhaan arvauksen arvojoukon vaihtelusta. Jos vaihtelu on pieni, niin hyökkääjä voi päätellä havaintoyksikön muuttujien arvoja, vaikka paras arvaus ei osuisikaan oikeaan, kunhan myös ennuste-etäisyys on pieni. Jos epävarmuus on suuri, on hyökkääjän tehtävä oikea valinta k:n uskottavimman havaintoyksikön joukosta saadakseen käsityksen havaintoyksikön muuttujien arvoista. Olkoot M, $\hat{M}$, M^{-i}, s ja k määritelty kuten edellä. Olkoon v funktio, joka laskee muuttujien varianssien keskiarvon. Ennuste-epävarmuus

$$u(M^i, \hat{M}, s, k) = v\left(\hat{M}^{(i:k)}\right) \text{ ja}$$

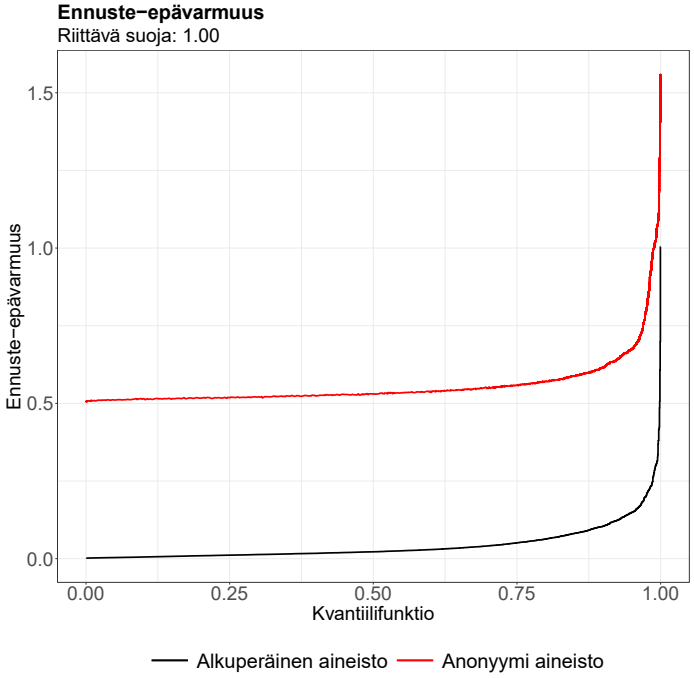
$$u(M^i, M^{-i}, s, k) = v\left(M^{-i(i:k)}\right),$$

missä $\hat{M}^{(i:k)}$ on i:nnettä alkuperäistä havaintoyksikköä lähimmän k:n havaintoyksikön joukko anonymisoidussa aineistossa (tai lopussa alkuperäisessä aineistossa), ilmaisee tämän joukon arvojen keskimääräisen varianssin. [9]

Tarkastellaan esimerkissä 5 ennuste-epäselvyyksien kvantiilifunktioiden kuvaajia ja niiden tulkintaa.

Esimerkki 5. *Ennuste-epävarmuus -kuvaaja*

Olkoon M jokin aineisto ja $\hat{M}$ sen anonymisoitu versio. Lasketaan ennuste-epävarmuus jokaiselle alkuperäisen aineiston havaintoyksikölle i, muodostetaan epävarmuuksien kvantiilifunktiot ja kuvataan kvantiilifunktiot samaan kuvaajaan.



Kuva 7: Ennuste-epävarmuuksien kvantiilifunktiot anonyymille ja alkuperäiselle aineistolle. Koska anonymisoidun aineiston ennuste-epävarmuuden kvantiilifunktio ei saa missään pienempiä arvoja kuin verrokkijakauma (riittävä suoja = 1), on viiden parhaan arvauksen arvojoukossa riittävästi vaihtelua. Hyökkääjä ei siis voi päätellä havaintoyksiköiden muuttujien likimaista arvojoukkoa.

Ennuste-epäselvyys ja -epävarmuus ovat samankaltaisia mittareita, mutta ne mittaavat kuitenkin eri asioita. Korkea epävarmuus ei välttämättä implikoi korkeaa epäselvyyttä, eikä tämä päde myöskään toisin päin. Esimerkiksi tilanteessa, jossa k parasta arvausta ovat arvojoukoltaan hyvin erilaisia (suuri epävarmuus), voi paras arvaus olla huomattavasti uskottavampi kuin k:neeksi paras arvaus (pieni epäselvyys). Päinvastaisessa tilanteessa, jossa hyökkääjällä on k likimain yhtä hyvää arvausta (suuri epäselvyys), voi joukko koostua likimain identtisistä havaintoyksiköistä (pieni epävarmuus). Optimaalisessa anonyymissä aineistossa k parasta arvausta olisivat toisistaan poikkeavia (korkea epävarmuus), mutta kuitenkin yhtä kaukana todellisesta havaintoyksiköstä (korkea epäselvyys).

3 Anonymisointimenetelmät

Tässä luvussa esitellään anonymisointimenetelmät, joita sovelletaan luvussa 4 COSMOS-tutkimuksen aineistoon. k-anonymiteettiä (luku 3.1) sekä l-diversiteettiä (luku 3.2) on sovellettu laajalti kirjallisuudessa [8]. Spektraaliset anonymisointimenetelmät (luku 3.3) pyrkivät säilyttämään tiettyjä aineiston rakenteita, jotka ovat oleellisia tutkielmassa sovellettavien analyysimenetelmien kannalta [9]. RSA-algoritmilla (luku 3.4) on sen sijaan saatu lupaavia tuloksia [11], joten menetelmää haluttiin testata toisenlaiseen aineistoon.

3.1 *k*-anonymiteetti

k-anonymiteetti (määritelmä 15) on Sweeneyn [5] vuonna 2002 kehittämä rivitason tiedon ominaisuus, jonka tarkoituksena on estää identiteettirikkomukset kvasitunnisteen kautta [5]. Aineiston voi jakaa kvasitunnisteen perusteella osajoukkoihin ja näitä osajoukkoja kutsutaan tässä tutkielmassa *ekvivalenssiluokiksi*. Ekvivalenssiluokka on siis aineiston osajoukko, jossa kaikilla havaintoyksiköillä on samat arvot kvasitunnisteeseen kuuluvilla muuttujilla. k-anonymiteetin toteuttavassa aineistossa jokaisessa ekvivalenssiluokassa on vähintään k havaintoyksikköä, eli vaikka hyökkääjä tietäisi havaintoyksikön kvasitunnisteen muuttujien arvot, ei tätä voisi erottaa ekvivalenssiluokan vähintään $k-1$ muusta havaintoyksiköstä kvasitunnisteen perusteella. [5]

Määritelmä 15. *k-anonymiteetti*

Aineisto M toteuttaa k-anonymiteetin, jos kaikille aineiston havaintoyksiköille M^i on olemassa vähintään $k-1$ havaintoyksikköä, joilla $M_Q^i = M_Q^1 = \ldots = M_Q^{k-1}$, jossa M_Q^i on havaintoyksikön i arvojoukko kvasitunnisteen muuttujilla. [5, 7]

Käytännössä k-anonymiteetti saavutetaan tunnistamalla aineistossa olevat kvasitunnisteet ja muodostamalla näiden ekvivalenssiluokat. Mikäli jossain ekvivalenssiluokassa on vähemmän kuin k havaintoyksikköä, tulee kvasitunnisteessa olevia muuttujia käsitellä siten, että ekvivalenssiluokkien kooksi saadaan vähintään k. Tämä voidaan saavuttaa esimerkiksi luokkia yhdistelemällä tai arvoja karkeistamalla. Tässä tutkielmassa numeeristen muuttujien tapauksessa käytetään keskiarvoa ja kategoristen muuttujien tapauksessa moodia, sillä monet menetelmät perustuvat keskiarvojen estimoimiseen ja moodia käyttämällä aineiston arvojoukko ei kasva, mikä helpottaa käytettävyyden tutkimista. k-anonymiteetin voi kuitenkin saavuttaa muillakin tavoilla, kuten esimerkiksi havaintoyksiköitä poistamalla tai muuttujien arvoja piilottamalla. Esimerkissä 6 ja kuvassa 8 havainnollistetaan minkälaisia vaikutuksia k-anonymiteetillä voi olla aineistoon.

Esimerkki 6. *2-anonymiteetti*

Alkuperäinen aineisto (a) ja sen 2-anonyymi versio (b), kun kvasitunnisteeseen valitaan ikä ja sukupuoli. Aineistosta (a) tunnistetaan ekvivalenssiluokat, joita karkeistetaan keskiarvon ja moodin avulla 2-anonymiteetin saavuttamiseksi. Anonymisoidun aineiston (b) arvojoukko on tästä johtuen suppeampi. Päänsäryn arvojoukko ei ole muuttunut kummassakaan aineistosta, koska se ei kuulu kvasitunnisteeseen.

(a) Alkuperäinen aineisto.

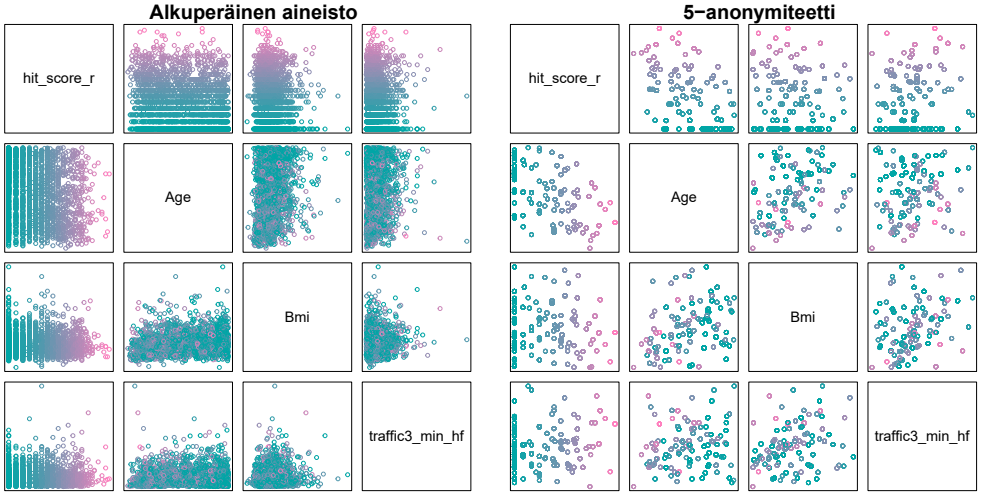
Ikä	Sukupuoli	Päänsärky
18	Mies	0
19	Mies	0
30	Nainen	1
25	Nainen	0
20	Muu	1

(b) 2-anonyymi aineisto.

Ikä	Sukupuoli	Päänsärky
18,5	Mies	0
18,5	Mies	0
25	Nainen	1
25	Nainen	0
25	Nainen	1

Jos yhdestä aineistosta julkaistaan useampi k-anonymisoitu versio, on uudemmissa julkaisuissa luettava kvasitunnisteeseen kaikki jo julkaistut muuttujat. Julkaisut yhdessä eivät välttämättä muuten toteuta k-anonymiteettiä, vaikka ne erikseen sen toteuttaisivatkin. Jos taas jo julkaistuun k-anonyymiin aineistoon halutaan lisätä havaintoyksiköitä, täytyy uudessa julkaisussa lukea kaikki jo julkaistut muuttujat kvasitunnisteeseen. Vaihtoehtoisesti, havaintoyksiköitä voi muokata ja lisätä lähimpiin ekvivalenssiluokkiin jo julkaistussa aineistossa. [5]

Pieni lisäsuoja, joka pätee lähes kaikille anonymisointimenetelmille, on aineiston rivien sekoittaminen ennen julkaisua. Tällöin samaan aineistoon perustuvien julkaisujen havaintoyksiköistä ei voi tehdä helposti päätelmiä pelkästään yhdistelemällä julkaisujen rivejä.



Kuva 8: Alkuperäisen aineiston ja 5-anonymiteetti-aineiston hajontakuvio, kun aineiston kaikki muuttujat valitaan kvasitunnisteeseen. Karkeistamisesta johtuen 5-anonymiteetin toteuttava aineisto on harvempi ja muodoltaan hieman muuttunut, eli muuttujien väliset riippuvuudet eivät pysy täysin ennallaan. Vaaleanpunainen väri tarkoittaa suurta arvoa muuttujalla hit_score_r kyseisessä aineistossa ja sininen pientä arvoa.

k-anonymiteetin ongelmiin lukeutuvat muun muassa luvussa 2.1 esitelty kvasitunnisteen muodostamisen vaikeus ja siitä koituvat ongelmat, mitkä saattavat johtaa identiteettirikkomukseen. Lisäksi, k-anonymiteetti ei lähtökohtaisesti anna mitään suojaa ominaisuusrikkomusta vastaan. Jos jossain anonyymin aineiston ekvi-

valenssiluokassa kaikilla havaintoyksiköillä on sama arvo jollain luottamuksellisella muuttujalla, saa hyökkääjä tietoa havaintoyksiköstä ilman identiteettirikkomusta. Esimerkin 6 2-anonyymissä aineistossa hyökkääjä voisi esimerkiksi päätellä, että aineiston miehillä ei esiinny ollenkaan päänsärkyä. Tämän takia voi olla perusteltua jättää luottamukselliset muuttujat pois kvasitunnisteesta, jolloin ekvivalenssieluokkien yhteydessä olevien luottamuksellisten muuttujien arvot määräytyvät enemmän tai vähemmän sattumalta. k-anonymiteetti sopii näistä syistä parhaiten aineistolle, jossa ei ole ole lainkaan luottamuksellisia muuttujia ja kvasitunnisteeseen valittavat muuttujat ovat perusteltavissa.

3.2 l-diversiteetti

l-diversiteetti on aineiston ominaisuus, jonka Machanavajjhala ym. [7] kehittivät vuonna 2007 k-anonymiteetin ominaisuusrikkomukseen liittyvien heikkouksien korjaamiseen. l-diversiteetin tarkoitus on suojata aineistoa sekä identiteettirikkomuksilta että luottamuksellisten muuttujien ominaisuusrikkomuksilta edellyttämällä, että aineiston jokaisessa ekvivalenssiluokassa esiintyy vaihtelua luottamuksellisten muuttujien suhteen. Vaihtelun tarkoituksena on estää k-anonymiteetissä mahdollinen luottamuksellisen muuttujan ominaisuusrikomus. l-diversiteetistä on olemassa useampia versioita, mutta tässä tutkielmassa käytetään *entropia l-diversiteettiä* (määritelmä 16), sillä se on yksi yleisimmistä l-diversiteetin käytetyistä versioista. [7]

Määritelmä 16. *Entropia l-diversiteetti*

Aineisto M toteuttaa entropia l-diversiteetin, jos luottamuksellisella muuttujalla S, sen arvojoukolla $\{s_1, \ldots, s_h\}$, positiivisella reaaliluvulla l ja jokaisella ekvivalenssiluokalla q_i pätee

$$-\sum_{s \in S} p(q_i, s) \log(p(q_i, s)) \geq \log(l),$$

jossa $p(q_i, s) = n_{(q_i,s)} / \sum_{s' \in S} n_{(q_i,s')}$ kertoo ekvivalenssiluokan havaintoyksiköiden osuuden, joiden luottamuksellinen muuttuja on s. [7]

Entropia l-diversiteetin toteuttavan aineiston jokaisessa ekvivalenssiluokassa on vähintään l eri luottamuksellisen muuttujan arvoa. Ominaisuutta on havainnollistettu esimerkissä 7 ja kuvassa 9.

Esimerkki 7. *2-diversiteetti*

Alkuperäinen, esimerkissä 6 esitelty aineisto (a) ja sen 2-diversifioitu versio (b), kun kvasitunnisteeseen valitaan ikä sekä sukupuoli, ja päänsärky on luottamuksellinen muuttuja. Aineistosta (a) tunnistetaan eksivalenssiluokat, joita karkeistetaan keskiarvon ja moodin avulla, minkä seurauksena muuttujien arvojoukot pienenevät. Viides havaintoyksikkö – joka 2-anonymiteetin tapauksessa yhdistettiin naisten ekvivalenssiluokkaan – on nyt yhdistetty miesten ekvivalenssiluokkaan, jotta 2-diversiteetti toteutuu molemmissa luokissa luottamuksellisen muuttujan, eli päänsäryn osalta.

(a) Alkuperäinen aineisto.

Ikä	Sukupuoli	Päänsärky
18	Mies	0
19	Mies	0
30	Nainen	1
25	Nainen	0
20	Muu	1

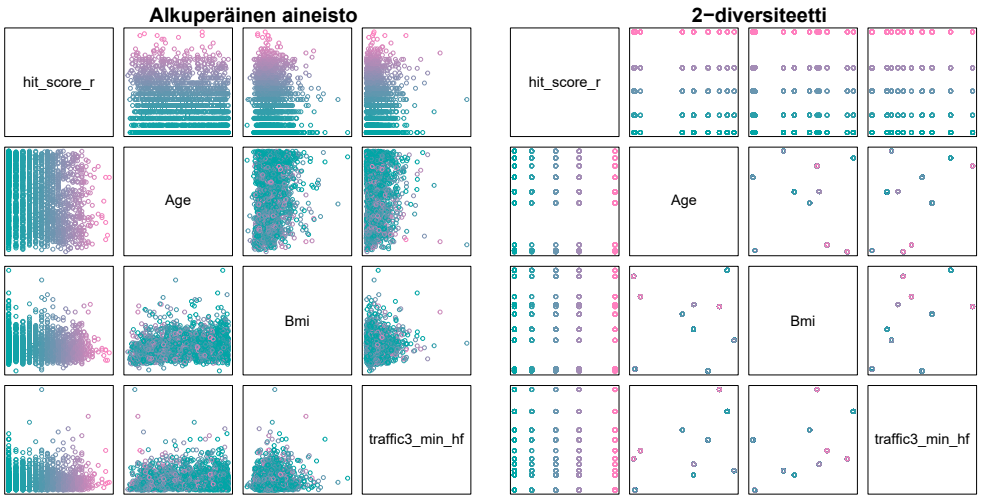
(b) 2-diversifioitu aineisto.

Ikä	Sukupuoli	Päänsärky
19	Mies	0
19	Mies	0
27,5	Nainen	1
27,5	Nainen	0
19	Mies	1

Jos aineistossa on useampi luottamuksellinen muuttuja, ei riitä, että aineisto toteuttaa määritelmän 16 jokaiselle luottamukselliselle muuttujalle erikseen, vaan määritelmän 16 täytyy toteutua jokaiselle muuttujalle muut luottamukselliset muuttujat huomioon ottaen. Jokaista luottamuksellista muuttujaa käsitellään kerrallaan aineiston ainoana luottamuksellisena muuttujana, ja loput luottamukselliset muuttujat lisätään kvasitunnisteeseen. Jos jokaiselle luottamukselliselle muuttujalle aineisto toteuttaa määritelmän 16 tässä tapauksessa, sanotaan aineiston toteuttavan l-diversiteetin. [7]

Määritelmä 17. *Usean luottamuksellisen muuttujan l-diversiteetti*

Olkoot aineistolla M luottamukselliset muuttujat $S_1, \ldots, S_m$ ja kvasitunniste $\{Q_1, \ldots, Q_k\}$. Aineiston sanotaan toteuttavan l-diversiteetin, jos kaikilla $j = 1, \ldots, m$ aineisto toteuttaa l-diversiteetin kun S_j on ainoa luottamuksellinen muuttuja ja joukko $\{Q_1, \ldots, Q_k, S_1, \ldots, S_{j-1}, S_{j+1}, \ldots, S_m\}$ on kvasitunniste. [7]



Kuva 9: Alkuperäisen aineiston ja 2-diversifioidun aineiston päänsäryn vakavuuden (hit_score_r), iän, painoindeksin ja puheluminuuttien (traffic3_min_hf) hajontakuvio, kun päänsäryn vakavuus -pistemäärä ja itseraportoitu masennus (ei kuvassa) valitaan luottamuksellisiksi muuttujiksi ja loput sarakkeet muodostavat kvasitunnisteen. 2-diversifioitu aineisto tiivistyy muutamiin pisteisiin ja jatkuvaa luottamuksellista muuttujaa joudutaan karkeistamaan. Vaaleanpunainen väri tarkoittaa suurta arvoa muuttujalla hit_score_r kyseisessä aineistossa ja sininen pientä arvoa.

Kvasitunnisteen koon ja luottamuksellisten muuttujien määrän kasvaessa ekvivalenssiluokkien koko välttämättä kasvaa, jotta l-diversiteetti toteutuisi [7]. Usean luottamuksellisen muuttujan tapauksessa jatkuvien luottamuksellisten muuttujien arvojoukkoa saatetaan joutua karkeistamaan, ettei ekvivalenssiluokista tule liian pieniä (kuva 9). Tämä puolestaan johtaa suurempaan muuttujan arvojoukon karkeistamiseen ja sitä kautta käytettävyyden heikkenemiseen. l-diversiteetti kärsii k-anonymiteetin ongelmien lisäksi luottamuksellisten muuttujien valinnasta, johon ei myöskään ole yleispätevää menetelmää. l-diversiteetti soveltuu siis kaikenlaisille aineistoille, mutta luottamuksellisten muuttujien määrän kasvaessa käsiteltyjen aineistojen käytettävyys heikkenee.

3.3 Spektraalinen anonymisointi

Määritelmässä 19 esitetty *Spektraalinen anonymisointi* (eng. spectral anonymization) on Laskon ja Vinterbon [9] vuonna 2010 kehittämä menetelmä, joka perustuu havaintoon, että aineiston anonymisointi voidaan toteuttaa aineiston alkuperäisen kannan sijaan jossain *spektraalisessa kannassa* (eng. spectral basis). Sopivasti valittu spektraalinen kanta voi parantaa menetelmän tuottamaa yksityisyyttä, keventää tarvittavia laskutoimituksia tai helpottaa muuttujien suuren määrän tuomia ongelmia (eng. curse of dimensionality). Tässä tutkielmassa spektraalisen kannan tuottamiseen käytetään *pääakselihajotelmaa* (eng. singular value decomposition) (määritelmä 18). [9]

Määritelmä 18. *Pääakselihajotelma*

Olkoon $M_{n\times p}$ alkuperäinen aineisto. Nyt

$$M = UDV^T$$

on sen pääakselihajotelma, jossa $U_{n\times n}$ ja $V_{p\times p}$ ovat ortogonaalisia matriiseja, joiden sarakkeet ovat matriisien MM^T ja M^TM ominaisvektorit ja $D_{n\times p}$ on diagonaalimatriisi, jonka nollasta poikkeavat diagonaalialkiot ovat edellisten matriisien nollasta poikkeavien ominaisarvojen neliöjuuret. [9]

Määritelmä 19. *Spektraalinen anonymisointi*

Olkoon M aineisto, jonka pääakselihajotelma on $M = UDV^T$. Olkoon $f : \mathbb{R}^{a\times b} \to \mathbb{R}^{a\times b}$ jokin funktio, joka anonymisoi annetun aineiston. Merkitään anonymisoitua aineistoa $\hat{M} := f(M)$. Nyt

$$\begin{aligned}
\hat{M} &= f(U)DV^T,\\
\hat{M} &= UDf(V^T),\\
\hat{M} &= f(UD)V^T,\\
\hat{M} &= Uf(DV^T)
\end{aligned}$$

ovat spektraalisesti anonymisoituja aineistoja. [9]

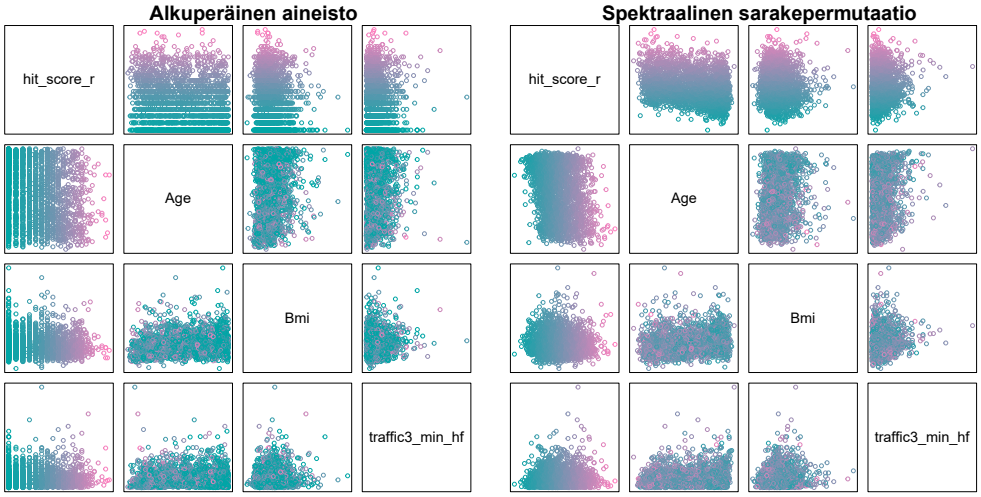
Anonymisointifunktio kannattaa kohdentaa eri pääakselihajotelman osaan riippuen käytettävästä anonymisointifunktiosta. Ortogonaalisen matriisin U sarakkeet ovat korreloimattomia, jolloin sen sarakkeita voidaan käsitellä toisistaan riippumatta. Tämä on hyödyllinen ominaisuus, kun anonymisointifunktiona käytetään esimerkiksi kohinan lisäystä tai sarakekohtaista permutaatiota. Spektraalinen anonymisointi pääakselihajotella toimii parhaiten, kun aineiston rakenteet ovat lineaarisia. [9]

Spektraaliset menetelmät soveltuvat kaikille aineistoille, mutta niitä kannattaa käyttää erityisesti, jos kvasitunnistetta tai luottamuksellisia muuttujia ei voida syystä tai toisesta määrittää tai niiden osuus aineistosta on suuri. Jos aineistossa on taas paljon julkisia muuttujia, spektraaliset menetelmät tekevät turhaa työtä käsitellessään kaikkia muuttujia. Lisäksi on hyvä huomata, että riippuen anonymisointiin käytettävästä funktiosta, ei spektraalinen anonymisointi välttämättä anna samanlaisia yksityisyyteen liittyviä ominaisuustakeita kuten k-anonymiteetti tai l-diversiteetti, vaan yksityisyydestä tulee varmistua erikseen käyttämällä esimerkiksi luvussa 2.6 esitettyjä menetelmiä.

3.3.1 Spektraalinen sarakepermutaatio

Sarakepermutoinnissa (eng. cell swapping) matriisin jokainen sarake korvataan satunnaisella permutaatiolla kyseisestä sarakkeesta. Alkuperäiseen aineistoon käytettynä permutointi säilyttää muuttujakohtaiset momentit, mutta rikkoo muuttujien väliset riippuvuudet. Jos permutointia käytetään pääakselihajotelman ortogonaaliseen matriisiin U, voidaan muuttujien väliset riippuvuudet säilyttää likimain, sillä matriisin U korrelaatiomatriisi on vakiota vaille identiteettimatriisi. Tällöin sarakekohtainen permutointi säilyttää alkuperäisten muuttujien korrelaatiorakenteen myöskin likimain (kuva 10). Tätä menetelmää kutsutaan *spektraaliseksi sarakepermutoinniksi* (eng. spectral swapping). Jos matriisista M vähennetään sarakekeskiarvot ennen pääakselihajotelmaa ja anonymisointia, säilyvät lisäksi muuttujakohtaiset keskiarvot, varianssit ja kovarianssit, kunhan keskiarvot lisätään takaisin anonymisoituun matriisiin $\hat{M}$. Lisäksi menetelmä säilyttää pääkomponenttien likimääräiset suunnat. [9]

Spektraalinen permutointi on yksinkertainen ja nopea menetelmä, eikä hyperparametreja – kuten k tai l – tarvitse valita. Lisäksi identiteettirikkomuksen riski on pieni, sillä todennäköisyys, että satunnainen permutaatio palauttaa jonkin identtisen tai lähes identtisen rivin on pieni. Menetelmä ei kuitenkaan säilytä kaikkea aineiston rakennetta, sillä matriisin U permutointi ei täysin säilytä sen korrelaatiorakennetta.



Kuva 10: Alkuperäisen aineiston ja spektraalisen sarakepermutaatio -aineiston hajontakuvio. Spektraalisesti permutoidun aineiston muoto pysyy likimain samanlaisena, mutta esimerkiksi muuttujassa hit_score_r esiintyvä diskreettiys katoaa, sillä muuttujaa käsitellään jatkuvana. Vaaleanpunainen väri tarkoittaa suurta arvoa muuttujalla hit_score_r kyseisessä aineistossa ja sininen pientä arvoa.

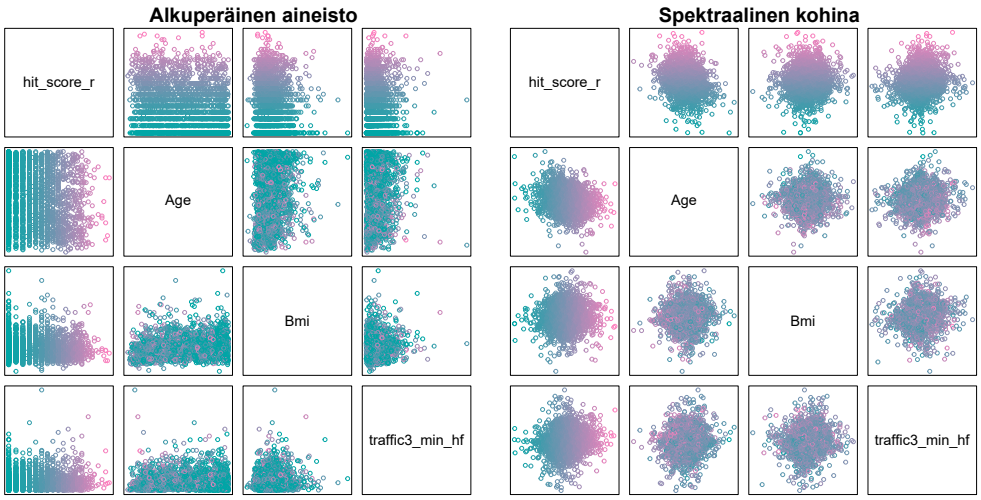
3.3.2 Spektraalinen kohina

Kohinan lisääminen (eng. additive noise) on anonymisointimenetelmä, jossa aineiston M jokaiseen havaintoon lisätään satunnaista kohinaa jostain jakaumasta. Riittävän anonymiteetin saavuttaminen vaatii yleensä paljon kohinaa, varsinkin jos muuttujien lukumäärä on suuri ja niiden välillä esiintyy riippuvuuksia. Suuri kohinan määrä taas peittää aineistossa olevat rakenteet. [9]

Spektraalisessa kohinassa kohinaa lisätään pääakselihajotelman ortogonaalisen matriisin U sarakkeisiin. Tällöin kohinan lisäyksessä ei tarvitse kiinnittää huomioita muuttujien välisiin riippuvuussuhteisiin, sillä ortogonaalisen matriisin sarakkeet ovat korreloimattomia. Kohinaa voidaan lisätä erilaisista jakaumista, eikä kaikkiin sarakkeisiin tarvitse lisätä kohinaa samaa määrää tai edes samasta jakaumasta.

Tässä tutkielmassa matriisin U sarakkeen U_j arvoihin lisätään kohinaa Laplace$(0, b_j)$-jakaumasta, jonka hajontaparametri $b_j = \frac{|\max_i U_j^i - \min_i U_j^i|}{4}$, jossa $i \in \{1, \ldots, n\}$ ja U_j^i on matriisin U arvo rivillä i sarakkeella j. Hajontaparametri b_j on suurempi enemmän hajontaa sisältävillä sarakkeilla ja pienempi vähemmän hajontaa sisältävillä sarakkeilla. Tämä auttaa peittämään mahdolliset oudokit, mutta välttää liiallista kohinan lisäämistä sarakkeilla, joissa oudokkeja ei esiinny. Tässä tutkielmassa spektraalisessa sarakepermutoinnissa tehty sarakekeskiarvojen vähennys ja lisäys tehdään myös spektraalisessa kohinassa, sillä se havaittiin samankaltaisuutta parantavaksi toimenpiteeksi. Kohinajakauma ja hajontaparametrin laskutapa valittiin kokeilemalla eri yhdistelmiä, kuten normaalijakaumaa, Laplace-jakaumaa, keskihajontaa ja tutkielmaan valittua maksimi-minimi-erotusta. Normaali- ja Laplace-jakauman välillä ei ollut merkittävää eroa, mutta maksimi-minimi -erotus antoi paremman yksityisyydensuoja verrattuna keskihajontaan. Kuvassa 11 tarkastellaan

spektraalisen kohinan lisäämisen vaikutusta aineiston hajontakuvioon.



Kuva 11: Alkuperäisen aineiston ja spektraalisella kohinalla suojatun aineiston hajontakuviot. Kohinan lisääminen johtaa muuttujien riippuvuusrakenteiden muuttumiseen. Vaaleanpunainen väri tarkoittaa suurta arvoa muuttujalla hit_score_r kyseisessä aineistossa ja sininen pientä arvoa.

Spektraalinen kohina on suhteellisen suoraviivainen ja laskennallisesti kevyt menetelmä. Sen haasteena on kuitenkin sopivan kohinajakauman sekä kohinan määrän valinta; liian vähäinen kohina ei suojaa aineistoa riittävästi ja sopimaton jakauma tai liiallinen määrä sen sijaan heikentää aineiston käytettävyyttä tarpeettoman paljon.

3.4 RSA

Aineistoja voidaan anonymisoida sekoittavien ja karkeistavien menetelmien lisäksi *kryptografisia* menetelmiä käyttäen. Kryptografisten menetelmien käyttöä anonymisointiin esittelivät Rabbi ym. [11] vuonna 2019. Kryptografisissa menetelmissä lähtöjoukon arvot kuvataan maalijoukkoon *yksisuuntaisella funktiolla* (eng. one-way function). Yksisuuntaisuus tarkoittaa, että lähtöjoukosta maalijoukkoon kulkeminen on paljon helpompaa kuin maalijoukosta lähtöjoukkoon kulkeminen. Tämä siis tarkoittaa, että aineiston salaaminen on laskennallisesti paljon kevyempää kuin salauksen purkaminen ilman oikeaa *avainta*. Tässä tutkielmassa kryptografisena menetelmänä käytetään Rivest–Shamir–Adleman- eli RSA-algoritmia. RSA on *salaisen* ja *julkisen avaimen* (eng. private and public key) salausmenetelmä, jossa julkinen avain voi olla kaikkien tiedossa ja salainen avain on nimensä mukaisesti pidettävä salassa. RSA-avainten generointi esitellään algoritmissa 1, joka perustuu lähteisiin [17, 18]. Lähetettävä viesti salataan vastaanottajan julkisella avaimella ja tämä salaus on purettavissa vain julkista avainta vastaavalla salaisella avaimella. RSA:n perustana toimii havainto, että oikein valittuna kolmella suurella kokonaisluvulla n, e ja d pätee

$$(m^e)^d \equiv m \pmod{n}$$

kaikilla kokonaisluvuilla m, $0 \leq m < n$. Lisäksi, kokonaislukujen n ja e ollessa julkisia, on vaikea päätellä kokonaislukua d. [11, 19]

Algoritmi 1 RSA-avaimien generointi

1: Valitse satunnaisesti kaksi suurta alkulukua $p \neq q$ toisistaan riippumatta. Kokonaisluvut p ja q ovat pidettävä salassa.
2: Laske $n = pq$. Kokonaisluku n toimii moduluksena sekä salaisessa että julkisessa avaimessa.
3: Laske $\lambda(n) = \text{pyj}(p-1, q-1)$. Myös $\lambda(n)$ on pidettävä salassa.
4: Valitse kokonaisluku e siten, että $2 < e < \lambda(n)$ ja $\text{syt}(e, \lambda(n)) = 1$. Kokonaisluku e on osa julkista avainta.
5: Laske $d \equiv e^{-1} \pmod{\lambda(n)}$. Kokonaisluku d on osa salaista avainta.

Viestin m salaaminen tapahtuu julkista avainta (e, n) käyttäen seuraavasti:

$$c \equiv m^e \pmod{n}.$$

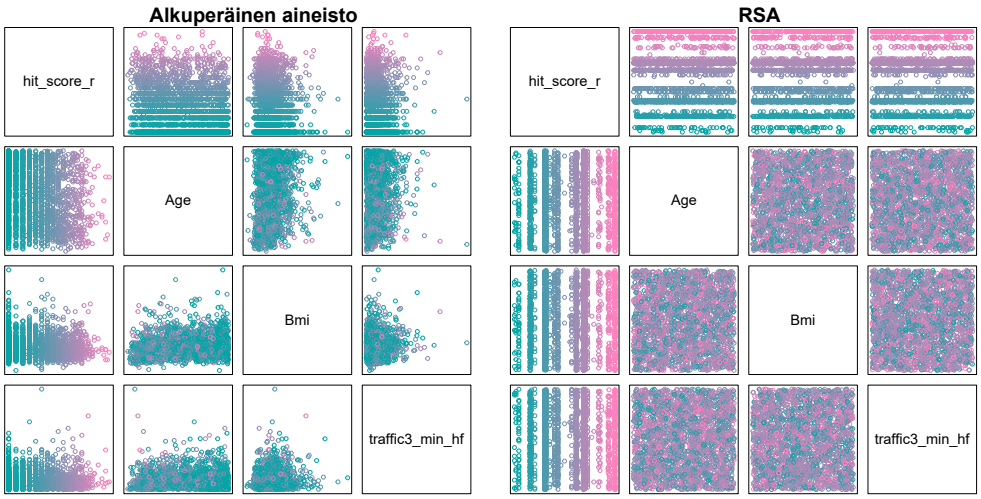
Salatun viestin c salauksen purkaminen tapahtuu salaista avainta (d, n) käyttäen seuraavasti:

$$c^d \equiv (m^e)^d \equiv m \pmod{n}.$$

Tässä tutkielmassa aineiston jokainen havainto muutetaan kokonaisluvuiksi, jonka jälkeen ne salataan erikseen samaa RSA-avainta käyttäen. Lisäksi, aineisto minmax -skaalataan salaamisen jälkeen, eli jokaisesta sarakkeesta vähennetään sarakkeen pienin arvo ja erotus jaetaan suurimman ja pienimmän arvon erotuksella. Minmax -skaalaus toimii lisäsuojana aineistolle, sillä skaalaus ei ole kääntyvä ilman tietoa sarakekohtaisista minimeistä ja maksimeista. [11]

RSA on altis ominaisuusrikkomukselle varsinkin julkisen avaimen ollessa kaikkien tiedossa. Jos salattavan sarakkeen lähtöjoukko on pieni, on sarakkeen kaikki mahdolliset arvot helppo käydä kokeilemalla läpi. Esimerkiksi sukupuolen tai kokonaislukuna ilmoitetun iän salaus olisi helppo murtaa kokeilemalla, jos salaamiseen käytetty julkinen avain on hyökkääjän tiedossa. Salaamiseen käytetty julkinen avain kannattaa siis pitää salassa, vaikka sekään ei takaa suojaa ominaisuusrikkomusta vastaan. RSA:lla salatun aineiston kategoristen muuttujien salaus voidaan tietyin ehdoin murtaa myös ilman tietoa käytetystä julkisesta avaimesta: jos hyökkääjä esimerkiksi tietää jonkin kategorisen muuttujan jakauman populaatiotasolla ja aineisto on edustava otos tästä populaatiosta, voi hyökkääjä päätellä sarakkeen tarkat arvot vertaamalla populaatiotason ja salattujen arvojen vallitsevuutta.

Aineiston salaaminen RSA:lla tuhoaa aineiston muuttujien arvojoukot, sillä menetelmä kuvaa jokaisen havainnon M_j^i näennäissatunnaiseksi kokonaisluvuksi (kuva 12). Vierekkäisetkin arvot voivat kuvautua hyvin kauas toisistaan ja kaukaiset arvot hyvin lähelle toisiaan, eli arvojoukon järjestyksellä ei ole enää merkitystä salaamisen jälkeen. Tämän takia RSA soveltuu käytännössä vain nominaalisten muuttujien salaamiseen, sillä se kadottaa muuttujien suunnan ja suuruuden, jolloin numeerisille muuttujille suunnattujen analyysien ja mallinnusmenetelmien soveltaminen ei



Kuva 12: Alkuperäisen aineiston ja RSA-aineiston hajontakuvio. Aineiston arvojoukko muuttuu täysin, ja muuttujien väliset riippuvuudet katoavat, sillä RSA-algoritmi kuvaa jatkuvat muuttujat näennäissatunnaisluvuiksi. Vastaavasti osittain diskretisoitunut muuttuja hit_score_r muuttuu selvemmin kategoriseksi (raidoittuminen kuvassa). Vaaleanpunainen väri tarkoittaa suurta arvoa muuttujalla hit_score_r kyseisessä aineistossa ja sininen pientä arvoa.

ole enää mielekästä. Esimerkiksi lineaarisen regressiomallin tapauksessa jatkuvien muuttujien parametrien estimointi perustuisi pseudosatunnaislukuihin ja estimaatit olisivat siis käytännössä aina nollia. Kategoriset muuttujat pysyvät ennallaan sillä merkittävällä muutoksella, että tasojen arvoja ei voi yhdistää alkuperäisiin muuttujien arvoihin. Jos RSA esimerkiksi kuvaa arvon mies $\rightarrow$ 0 ja arvon nainen $\rightarrow$ 1, pysyvät estimaattien arvot samoina, mutta salatun tason nimestä yksin ei voi päätellä mitä alkuperäistä tasoa mikäkin luku tarkoittaa. Tulosten tulkinnasta tulee siis anonymisoinnin takia vaikeaa. Menetelmä soveltuu täten aineistolle silloin, kun sillä on tarkoitus tehdä jotain muuta kuin parametrista estimointia, kuten esimerkiksi luokittelua.

4 Empiirinen tarkastelu

Tässä luvussa tarkastellaan luvussa 3 esiteltyjen anonymisointimenetelmien vaikutusta COSMOS-tutkimuksen [15] aineistoon. Tutkimuksessa seurattiin ihmisten matkapuhelimen käyttöä neljän vuoden ajan ja tämän lisäksi heillä teetätettiin kysely mahdollisista oireista ja sekoittavista tekijöistä sekä tutkimuksen alussa että lopussa. Tutkimuksessa selvitettiin matkapuhelimen käytön yhteyttä päänsärkyyn, tinnitukseen ja kuulonalenemaan.

Tarkastelua varten COSMOS-aineistosta muodostettiin kaksi eri aineistoa vastemuuttujan mukaan riippuen siitä, miten käytettävyyttä mitattiin. Aineiston muuttujat sekä niiden tyypit ja roolit on esitelty taulukossa 8. Puuttuvista havainnoista johtuen lopullisissa aineistoissa on hieman eri määrä havaintoyksiköitä. Aineistossa 1 vastemuuttujana on jatkuva päänsäryn vakavuuden pistemäärä ja aineisto koostuu 3065 havaintoyksiköstä ja seitsemästä muuttujasta. Aineistossa 2 vastemuuttujana on binäärinen viikottaisen päänsäryn indikaattorimuuttuja ja aineisto koostuu 3057 havaintoyksiköstä ja seitsemästä muuttujasta.

Taulukko 8: COSMOS-aineiston kuvaus.

Muuttujan selite	Muuttujan nimi	Tyyppi	Rooli
Päänsäryn vakavuuden pistemäärä	Hit_score_r	Jatkuva	Vaste 1
Viikottainen päänsärky	Headache_FU	Binäärinen	Vaste 2
Puheluminuutit	Traffic3_min_hf	Jatkuva	Altiste
Ikä	Age	Jatkuva	Kovariaatti
Sukupuoli	Gender	Binäärinen	Kovariaatti
Painoindeksi	Bmi	Jatkuva	Kovariaatti
Itseraportoitu masennus	Depression	Binäärinen	Kovariaatti
Itseraportoitu hyvinvointi	Sf12gr	Binäärinen	Kovariaatti

Tarkastelun kohteena oli anonymisointimenetelmien vaikutus aineistojen samankaltaisuuteen (luku 2.4), käytettävyyteen (luku 2.5) ja yksityisyydensuojaan (luku 2.6). Ideaalinen anonymisointimenetelmä pitää aineiston mahdollisimman samankaltaisena alkuperäisen aineiston kanssa, samalla säilytten aineiston käytettävyyden ja suojaten erilaisilta tietosuojarikkomuksilta. Aineistojen anonymisointia varten ohjelmoitiin R-paketti *anon*, jonka lähdekoodi löytyy GitHub-palvelusta [20]. Taulukossa 9 esitellään menetelmien soveltamisessa käytetyt kvasitunnisteet ja luottamukselliset muuttujat.

Taulukko 9: Käsiteltäväksi valittavat luottamukselliset muuttujat ja kvasitunniste menetelmäkohtaisesti. Päänsäryllä viitataan sekä viikottaiseen päänsärkyyn että päänsäryn vakavuuden pistemäärään. Takaliite "pieni" viittaa siihen, että kvasitunnisteeseen ei ole valittu kaikkia muuttujia.

Aineisto	Luottamukselliset muuttujat	Kvasitunniste
5-anonymiteetti	-	Kaikki muuttujat
5-anonymiteetti pieni	-	Ikä, sukupuoli, bmi
2-diversiteetti	Masennus, päänsärky	Loput muuttujat
2-diversiteetti pieni	Masennus, päänsärky	Ikä, sukupuoli, bmi
Spekt. kohina	-	-
Spekt. sarakepermutaatio	-	-
RSA	-	-

Aineistot 1 ja 2 jaettiin satunnaisesti opetus- ja testiaineistoihin 80 % / 20 % jaolla. Paljon laskentatehoa vaativille k-anonymiteetille ja l-diversiteetille jako tehtiin 10 kertaa ja nopeammin ajettaville spektraalisille menetelmille sekä RSA:lle 100 kertaa. Luvuissa 2.4-2.6 esitellyt mittarit samankaltaisuudelle, käytettävyydelle ja yksityisyydelle laskettiin jokaisella jaolla erikseen ja mittareiden tuloksista laskettiin keskiarvot. Samankaltaisuus ja yksityisyydensuoja estimoitiin pelkästään opetusaineistojen perusteella, mutta käytettävyyttä estimoitaessa tarvittiin lisäksi myös testiaineistoja. Regressiomallit sovitettiin anonymisoiduille opetusaineistoille, mutta mallien selitysasteet ja AUC-pistemäärät laskettiin käsittelemättömällä testiaineistolla menetelmien vertailun mahdollistamiseksi. Poikkeuksena tähän on RSA, jonka selitysaste ja AUC-pistemäärä estimoitiin anonymisoidulla testiaineistolle. Kaikki tulokset esitetään kolmen merkitsevän numeron tarkkuudella ja luvut, joiden kymmenpotenssimuodon eksponentti on pienempi kuin -15 on pyöristetty nolliksi.

4.1 Samankaltaisuus

Anonymisoitujen opetusaineistojen samankaltaisuutta käsittelemättömien opetusaineistojen kanssa tarkasteltiin luvussa 2.4 esitettyjen samankaltaisuustaulukoiden avulla.

Päänsäryn vakavuuden pistemäärälle (aineisto 1) spektraalinen sarakepermutaatio tuotti selvästi parhaimmat tulokset (taulukko 10). Keskiarvot, varianssit ja korrelaatiot eivät juuri poikenneet alkuperäisistä. Vähiten samankaltaiset aineistot tuottivat spektraalinen kohina, joka epäonnistui varianssien säilyttämisessä ja RSA, joka epäonnistui muun muassa keskiarvojen säilyttämisessä. Molemmat 5-anonymiteetit ja 2-diversiteetit aliarvioivat variansseja, mutta säilyttivät keskiarvot. Keskiarvot säilyivät täysin, sillä tässä tutkielmassa numeeriset kvasitunnisteen muuttujat karkeistettiin keskiarvoa käyttäen.

Taulukko 10: Aineiston 1 opetusaineistojen samankaltaisuustaulukko. Luvut ovat keskiarvoja optusaineistojen välillä lasketuista suureista. Nollaa lähellä olevat arvot implikoivat suurempaa samankaltaisuutta. Paras menetelmä (mitattuna itseisarvoisten estimaattien summana) on lihavoitu.

Aineisto	E[S_μ]	kh[S_μ]	E[S_{σ^2}]	kh[S_{σ^2}]	E[S_ρ]	kh[S_ρ]
5-anonymiteetti	0	0	-0,406	0,333	0,0575	0,108
5-anonymiteetti pieni	0	0	-0,116	0,222	0,00333	0,0399
2-diversiteetti	0	0	-0,429	0,400	0,116	0,149
2-diversiteetti pieni	0	0	-0,174	0,255	0,0173	0,0739
Spekt. kohina	0,00104	0,0383	4,78	4,04	0,0352	0,0748
Spekt. sarakeperm.	0	0	-0,0000481	0,00913	-0,000180	0,0187
RSA	-2,73	2,92	-0,797	0,451	-0,0439	0,138

Viikottaiselle päänsärylle (aineisto 2) päti pitkälti samanlaiset tulkinnat kuin aineistolle 1. Spektraalinen sarakepermutaatio pärjäsi samankaltaisuudessa parhaiten, molemmat 5-anonymiteetit sekä 2-diversiteetit keskinkertaisesti ja RSA sekä spektraalinen kohina huonoiten.

Taulukko 11: Aineiston 2 opetusaineistojen samankaltaisuustaulukko. Luvut ovat keskiarvoja optusaineistojen välillä lasketuista suureista. Nollaa lähellä olevat arvot implikoivat suurempaa samankaltaisuutta. Paras menetelmä (mitattuna itseisarvoisten estimaattien summana) on lihavoitu.

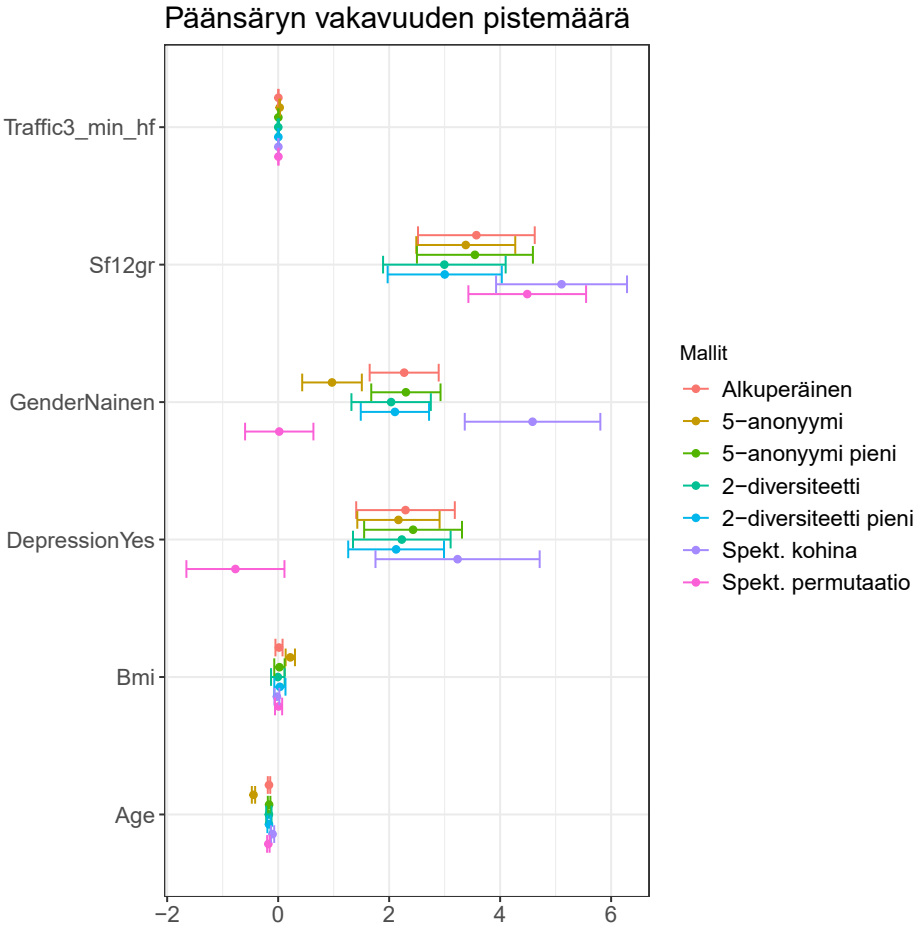
Aineisto	E[S_μ]	kh[S_μ]	E[S_{σ^2}]	kh[S_{σ^2}]	E[S_ρ]	kh[S_ρ]
5-anonymiteetti	0	0	-0,395	0,347	0,0961	0,0736
5-anonymiteetti pieni	0	0	-0,151	0,254	0,0113	0,0526
2-diversiteetti	0	0	-0,416	0,352	0,121	0,104
2-diversiteetti pieni	0	0	-0,169	0,234	0,0139	0,0532
Spekt. kohina	0	0,0400	5,30	4,53	0,00866	0,0764
Spekt. sarakeperm.	0	0	-0,0000232	0,00647	-0,0000815	0,0181
RSA	-2,08	2,91	-0,746	0,505	-0,0933	0,0657

4.2 Käytettävyys

Anonymisoinnin vaikutusta aineistojen käytettävyyteen tarkasteltiin regressiomallien avulla. Opetusaineistot anonymisoitiin, jonka jälkeen niille sovitettiin regressiomallit, joita testattiin käsittelemättömällä testiaineistoilla. Vertailun suorittamiseksi, regressiomallit sovitettiin myös anonymisoimattomille opetusaineistoille. Käytettävyyden mittareina toimivat mallien parametrien keskimääräiset piste-estimaatit sekä niiden 95 %:n luottamusvälit opetusaneistoilla, selitysaste testiaineistolla (lineaarinen regressio) tai AUC-pistemäärä testiaineistolla (logistinen regressio). Mitä lähempänä nämä ovat alkuperäisen aineiston vastaavia tunnuslukuja, sitä käyttökelpoisempi anonymisoitu aineisto on. RSA-opetusaineistojen testiaineistot käsiteltiin RSA:lla selitysasteiden ja AUC-pistemäärien laskemisen mahdollistamiseksi. RSA-aineistolle sovitetun regressiomallien piste-estimaatteja ja luottamusvälejä ei ole mie-

lekästä tulkita kategorisille muuttujille, sillä tasojen nimet muuttuvat salauksen takia näennäissatunnaisluvuiksi ja tästä syystä vastaavia kuvia ei raportoida.

Päänsäryn vakavuuden osalta (aineisto 1) 5-anonymiteetti pieni sekä molemmat 2-diversiteetit säilyttivät parhaiten opetusaineistojen käytettävyyttä, spektraaliset menetelmät toimivat keskinkertaisesti ja RSA sekä 5-anonymiteetti suoriutuivat huonoimmin, kun tarkastellaan taulukkossa 12 ja kuvaajassa 13 esiteltyjä tuloksia. Muuttujien piste-estimaatit ja niiden luottamusvälit vaihtelivat (kuva 13) kaikilla menetelmillä hieman, mutta eniten spektraalisilla menetelmillä kategorisilla muuttujilla. Selitysaste heikkeni 5-anonymiteetillä ja RSA:lla eniten, spektraalisilla menetelmillä keskinkertaisesti ja molemmilla 2-diversiteeteillä sekä 5-anonymiteetti pienellä vähiten.



Kuva 13: Lineaarisen regression parametrien 95 %:n luottamusvälit opetusaineistoilla. Esitetyt piste-estimaatit ja luottamusvälit lasketaan keskiarvona opetusaineistoille sovitetuista regressiomalleista. Alkuperäisestä poikkeavat piste-estimaatit tai luottamusvälit kertovat käytettävyyden laskusta. RSA-aineistolle sovitetun mallin parametreja ei kuvata.

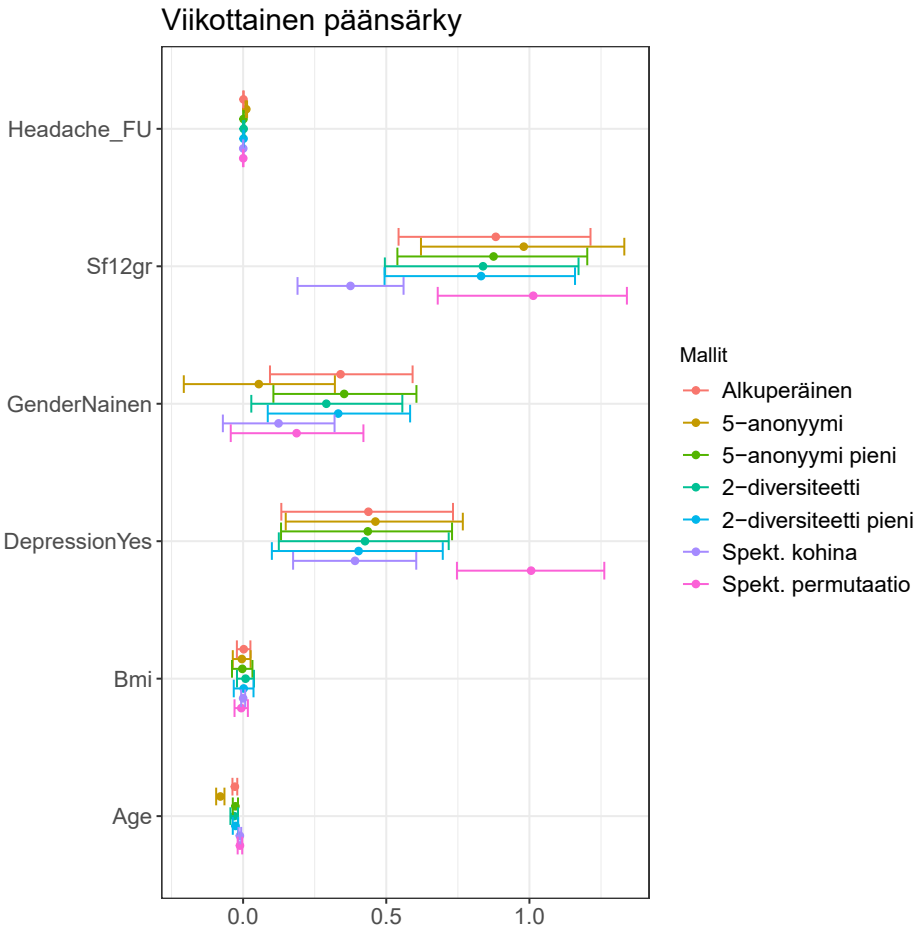
Taulukko 12: Alkuperäisille ja anonymisoiduille opetusaineistoille sovitettujen lineaaristen regressiomallien selitysasteiden keskiarvo testiaineistoilla. Alkuperäisestä laskenut selitysaste kertoo laskeneesta käytettävyydestä. Menetelmä, jonka arvo on lähinnä alkuperäisen selitysastetta, on lihavoitu.

Aineisto	Selitysaste
Alkuperäinen	0,147
5-anonymiteetti	-0,221
5-anonymiteetti pieni	0,140
2-diversiteetti	0,141
2-diversiteetti pieni	0,136
Spektraalinen kohina	0,0998
Spektraalinen sarakepermutaatio	0,102
RSA	0,00418

Viikottaiselle päänsärylle (aineisto 2) päti pitkälti samankaltaiset tulkinnat kuin aineistolle 1. 5-anonymiteetti pieni sekä molemmat 2-diversiteetit säilyttivät parhaiten käytettävyyttä, spektraaliset menetelmät toimivat keskinkertaisesti ja RSA sekä 5-anonymiteetti suoriutuivat huonoimmin. Piste-estimaatit ja luottamusvälit vaihtelivat (kuva 14) eniten taas spektraalisilla menetelmillä, ja 5-anonymiteetin sekä RSA:n AUC-pistemäärät (taulukko 13) laskivat menetelmistä eniten.

Taulukko 13: Alkuperäisille ja anonymisoiduille opetusaineistoille sovitettujen logististen regressiomallien AUC-pistemäärien keskiarvo testiaineistoilla. Alkuperäisestä laskenut AUC-pistemäärä kertoo laskeneesta käytettävyydestä. Pistemäärät perustuvat liitteen A kuvaajiin 17-23. Menetelmä, jonka AUC-pistemäärä on lähinnä alkuperäistä, on lihavoitu.

Aineisto	AUC
Alkuperäinen	0,670
5-anonymiteetti	0,633
5-anonymiteetti pieni	0,657
2-diversiteetti	0,654
2-diversiteetti pieni	0,670
Spektraalinen kohina	0,650
Spektraalinen sarakepermutaatio	0,640
RSA	0,606

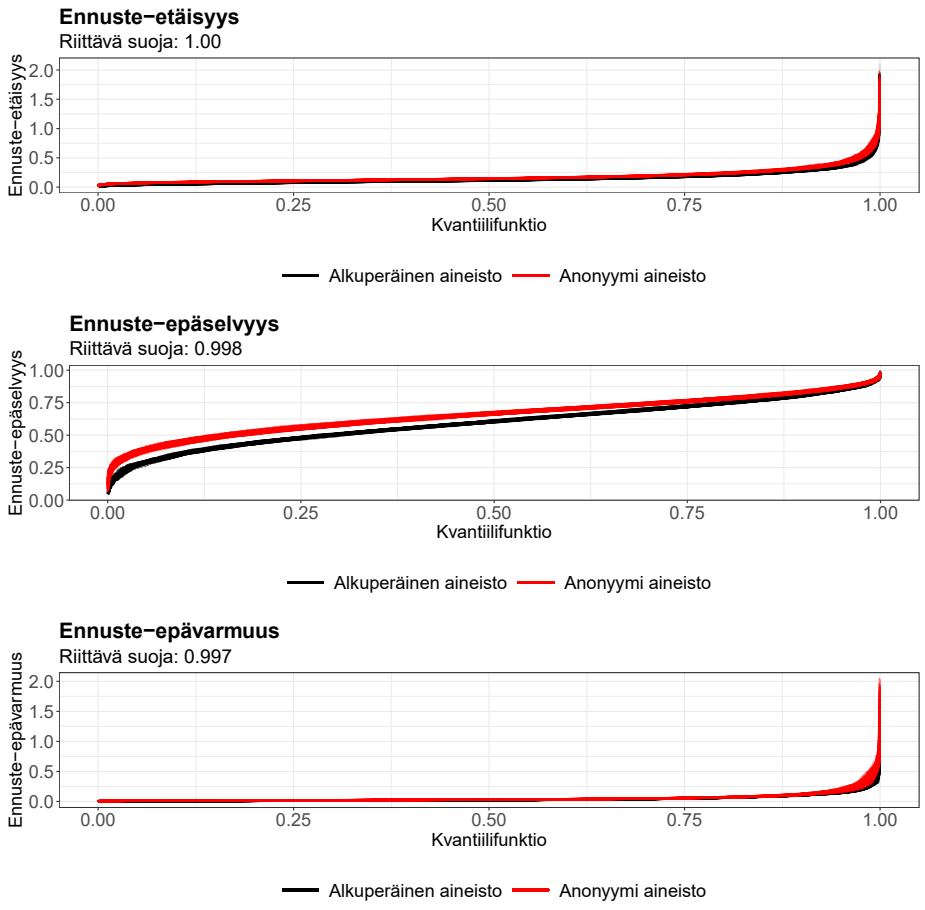


Kuva 14: Logistisen regression parametrien 95 %:n luottamusvälit opetusaineistoilla. Esitetyt piste-estimaatit ja luottamusvälit lasketaan keskiarvona opetusaineistoille sovitetuista regressiomalleista. Alkuperäisestä poikkeavat piste-estimaatit tai luottamusvälit kertovat käytettävyyden laskusta. RSA-aineistolle sovitetun mallin parametreja ei kuvata.

4.3 Yksityisyys

Kuvissa 15 ja 16 on esitetty yksityisyyden osalta parhaiten suorituneiden menetelmien ennustekuvaajat. Muiden menetelmien vastaavat kuvaajat löytyvät liittestä B. Taulukoissa 14 ja 15 on esitetty menetelmien yksityisyydensuojan numeeriset tunnusluvut.

Päänsäryn vakavuuden pistemäärän osalta (aineisto 1) yksityisimmät aineistot tuotti spektraalinen sarakepermutointi, jonka ennuste-etäisyyden, -epäselvyyden ja -epävarmuuden kuvaajat esitellään kuvassa 15. Molemmat 2-diversiteetit tuottivat hyvän suojan identiteettirikkomusta vastaan, mutta epäonnistuvat ennuste-epävarmuuden mittarilla (liite B, kuvat 26 ja 27). Molemmat 5-anonymiteetit tuottivat heikomman suojan sekä identiteettirikkomusta että päättelyrikkomusta vastaan (liite B, kuvat 24 ja 25). RSA:n yksityisyydensuojaa ei voi arvioida näillä mittareilla.



Kuva 15: Spektraalinen sarakepermutaatio -opetusaineistojen ennuste-etäisyydet, ennuste-epäselvyydet ja ennuste-epävarmuudet. Riittävän suojan kriteeri lasketaan keskiarvona opetusaineistojen kriteerien arvoista. Kriteeri toteutuu ennuste-etäisyydelle ja melkein -epäselvyydelle sekä -epävarmuudelle.

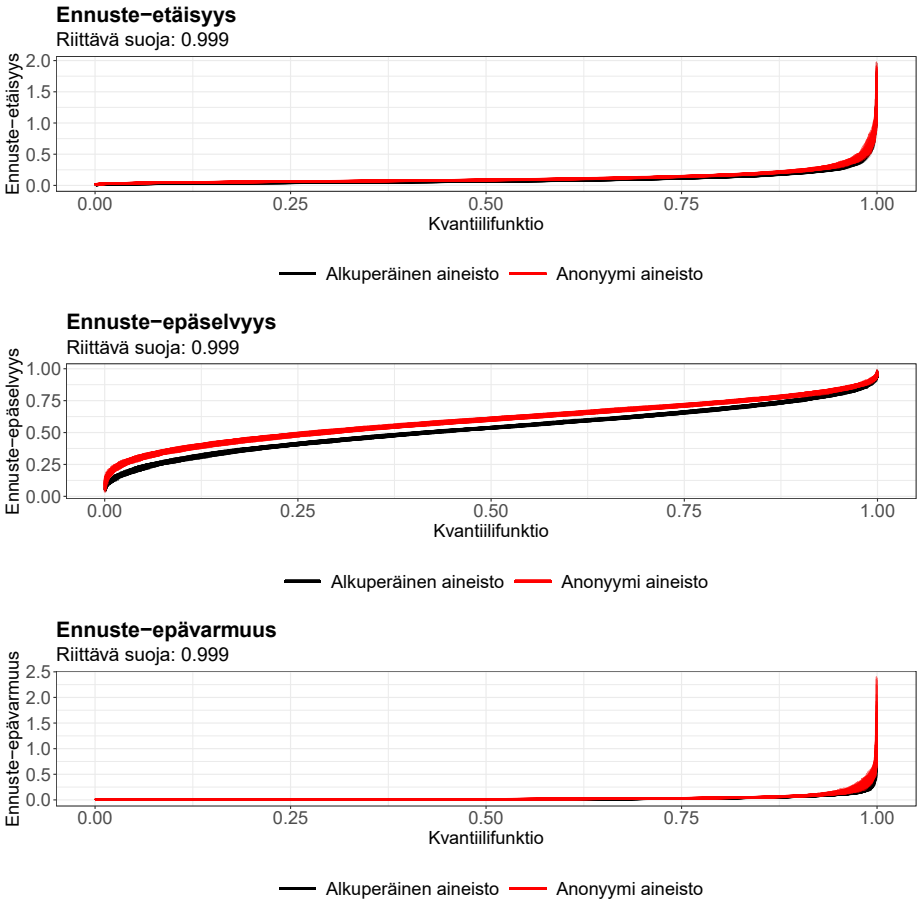
Taulukko 14: Alkuperäisten ja anonymisoitujen opetusaineistojen tunnistusasteiden ja päättelyrikkomustunnuslukujen keskiarvot. Tuunistusaste estimoitiin kaikille menetelmille oletuksella, että kvasitunnisteeseen kuuluu ikä, sukupuoli ja painoindeksi. Paras menetelmä (mitattuna parhaiden arvojen lukumäärällä, tasapelissä tunnistusaste ratkaisee) on lihavoitu.

Aineisto	Tunnistusaste	Ennuste-etäisyys	-epäselvyys	-epävarmuus
Alkuperäinen	1,00	-	-	-
5-anonymiteetti	0,00672	1,00	1,00	0
5-anonymiteetti pieni	0,0102	0,126	0,596	0,0176
2-diversiteetti	0,00198	1,00	1,00	0,000300
2-diversiteetti pieni	0,00318	0,998	0,983	0,00890
Spekt. kohina	0,00254	0,995	0,997	0,694
Spekt. sarakeperm.	0,000391	1,00	0,998	0,997

Viikottaiselle päänsärylle (aineisto 2) päti pitkälti samat päätelmät kuin aineistolle 1. Yksityisimmät aineistot tuotti jälleen spektraalinen sarakepermutointi, jonka ennuste-etäisyyden, -epäselvyyden ja -epävarmuuden kuvaajat esitellään kuvassa 16. Molemmat 2-diversiteetit ja 5-anonymiteetit tuottivat hyvän suojan identiteettirikkomusta vastaan, mutta epäonnistuvat päättelyrikkomuksen mittareilla (liite B, kuvat 29-32), eikä RSA:n yksityisyydensuojaa voi arvioida tutkielman mittareilla.

Taulukko 15: Alkuperäisten ja anonymisoitujen opetusaineistojen tunnistusasteiden ja päättelyrikkomustunnuslukujen keskiarvot. Tuunistusaste estimoitiin kaikille menetelmille oletuksella, että kvasitunnisteeseen kuuluu ikä, sukupuoli ja painoindeksi. Paras menetelmä (mitattuna parhaiden arvojen lukumäärällä, tasapelissä tunnistusaste ratkaisee) on lihavoitu.

Aineisto	Tunnistusaste	Ennuste-etäisyys	-epäselvyys	-epävarmuus
Alkuperäinen	1,00	-	-	-
5-anonymiteetti	0,00802	1,00	1,00	0
5-anonymiteetti pieni	0,00943	0,960	0,894	0,0238
2-diversiteetti	0,00452	1,00	1,00	0,0127
2-diversiteetti pieni	0,00602	0,994	0,964	0,0243
Spekt. kohina	0,00277	0,991	0,998	0,756
Spekt. sarakeperm.	0,000364	0,999	0,999	0,999



Kuva 16: Spektraalinen sarakepermutaatio -opetusaineistojen ennuste-etäisyydet, ennuste-epäselvyydet ja ennuste-epävarmuudet. Riittävän suojan kriteeri lasketaan keskiarvona opetusaineistojen kriteerien arvoista. Kriteeri toteutuu kaikilla mittareilla melkein.

5 Pohdintaa

Tässä tutkielmassa yleisesti käytössä olevia (k-anonymiteetti ja l-diversiteetti) sekä muita lupaavalta vaikuttaineita anonymisointimenetelmiä (spektraalinen anonymisointi ja RSA) tarkasteltiin kootusti samoilla mittareilla Suomen lainsäädäntö mielessä pitäen, eikä tällaista vertailua ole tiedettävästi aiemmin kirjallisuudessa esitetty. Lisäksi kirjallisuudessa anonymisointimenetelmien empiirinen tarkastelu on keskittynyt pääosin aineistojen käytettävyyden mittaamiseen. Myöskään valittujen menetelmien empiiristä yksityisyydensuojaa ei ole kootusti tarkasteltu, tai ainakaan julkaisuja aiheesta ei löydetty. Taulukossa 16 on esitetty parhaiten kullakin mitatulla osa-alueella suoriutunut menetelmä kummallekin tutkielmassa sovelletulle aineistolle.

Taulukko 16: Parhaiden anonymisointimenetelmien yhteenvetotaulukko.

Mittari	Aineisto 1	Aineisto 2
Samankaltaisuus	Spekt. sarakeperm.	Spekt. sarakeperm.
Selitysaste / AUC-pistemäärä	l-diversiteetti	l-diversiteetti pieni
Piste-estimaatit ja luottamusvälit	k-anonymiteetti pieni	k-anonymiteetti pieni
Identiteettirikkomuksen suoja	Spekt. sarakeperm.	Spekt. sarakeperm.
Päättelyrikkomuksen suoja	Spekt. sarakeperm.	Spekt. sarakeperm.

Anonymisointimenetelmät suoriutuivat samankaltaisuuden, käytettävyyden ja yksityisyyden mittareilla vaihtelevasti. Kaikki menetelmät säilyttivät tyypillisesti muuttujien keskiarvot RSA:ta lukuun ottamatta, joka laski huomattavasti muuttujien keskiarvoja. Molemmat k-anonymiteetit ja l-diversiteetit pienensivät odotettavasti muuttujien variansseja, sillä muuttujien karkeistamiseen käytettiin tutkielmassa keskiarvoa. Spektraalinen kohina nosti muuttujien varianssia ja spektraalinen sarakepermutointi säilytti varianssit kutakuinkin entisellään, kuten menetelmien odotettiinkin tekevän. Korrelaatioissa menetelmillä oli vaihtelua molempiin suuntiin, ja kaikista menetelmistä spektraalinen sarakepermutointi säilytti tulokset parhaiten.

k-anonymiteetti pieni (kolme muuttujaa kvasitunnisteessa), l-diversiteetti (kaksi luottamuksellista muuttujaa ja loput kvasitunnisteessa) sekä l-diversiteetti pieni (kaksi luottamuksellista muuttujaa ja kolme muuttujaa kvasitunnisteessa) tuottivat käytettävyydeltään parhaat aineistot. Regressiomallien parametrit ja niiden luottamusvälit muistuttivat eniten alkuperäisen aineiston vastaavia tunnuslukuja ja selitysasteet sekä AUC-pistemäärät olivat laskeneet vähiten. k-anonymiteetti (kaikki muuttujat kvasitunnisteessa) taas laski varsinkin selitysastetta huomattavasti. Spektraaliset menetelmät säilyttivät selitysasteen ja AUC-pistemäärän hyvin, mutta epäonnistuivat kategoristen muuttujien piste-estimaattien ja niiden luottamusvälien säilyttämisessä. Tämä johtuu todennäköisesti pääakselihajotelmasta ja sen epäsopivuudesta käsitellä ei-jatkuvia muuttujia. RSA laski selitysastetta ja AUC-pistemäärää huomattavasti ja aiheutti kategoristen muuttujien piste-estimaattien tulkittavuuden menetyksen.

Spektraaliset menetelmät – varsinkin spektraalinen sarakepermutaatio – tuottivat parhaan suojan sekä identiteetti- että päättelyrikkomusta vastaan. Huomattava on, että päättelyrikkomuksen mittareiden kehittäjät kehittivät myös

spektraaliset menetelmät. Molemmat l-diversiteetit tuottivat hyvän suojan identiteettirikkomusta vastaan, mutta epäonnistuivat päättelyrikkomusta arvioivalla ennuste-epävarmuuden mittarilla. Tämä siis tarkoittaa, että l-diversiteettiaineistoilla viiden parhaan arvauksen arvojoukot muistuttivat toisiaan. Molemmat k-anonymiteetit tuottivat heikoimman suojan identiteettirikkomusta vastaan ja varsinkin 5-anonymiteetti pieni epäonnistui suojaamaan päättelyrikkomusta vastaan. Heikompi yksityisyydensuoja l-diversiteettiin verrattuna johtuu todennäköisesti pienemmistä ekvivalenssiluokista ja vaihtelun puutteesta luottamuksellisissa muuttujissa.

Kaikilla osa-alueilla yhteensä parhaiten toimiva malli oli spektraalinen sarakepermutaatio, joka säilytti jatkuvien muuttujien samankaltaisuuden erinomaisesti, säilytti kiitettävän osan aineiston käytettävyydestä ja tuotti yksityisyydensuojaltaan parhaat aineistot. Ero spektraalisten aineistojen samankaltaisuudessa ja käytettävyydessä viittaa siihen, että pääakselihajotelmaa hyödyntävät spektraaliset menetelmät eivät sovellu yhtä hyvin kategoristen muuttujien käsittelyyn, kuin jatkuvien muuttujien käsittelyyn. Tulokset yksityisyydensuojalle olivat samansuuntaisia kuin alkuperäisessä julkaisussa [9]: spektraalinen sarakepermutointi tuotti paremman suojan päättelyrikkomusta vastaan kuin kohinan lisääminen ja k-anonymiteetti. Spektraalisia menetelmiä ei ole kirjallisuudessa käsitelty alkuperäisen julkaisun [9] lisäksi, tai julkaisuja aiheesta ei löydetty.

RSA tuotti selvästi huonoimmat aineistot. RSA-aineistojen käytettävyys ja samankaltaisuus olivat kehnoimmat, eikä aineistojen yksityisyyttä voitu arvioida tämän tutkielman empiirisillä mittareilla. Alkuperäisessä julkaisussa [11] menetelmän yksityisyydensuojan tarkastelu jäi toteamukseen, että kryptografinen salaus yksinään on riittävä toimenpide riittävän yksityisyydensuojan takaamiseen. Tutkielmassa kuitenkin huomattiin, että pelkkä kryptografinen salaus ei luvussa 3.4 esitettyjen syiden takia ole riittävä toimenpide anonyymin tiedon saavuttamiseen. Tämän tutkielman ja alkuperäisen julkaisun [11] tulosten perusteella – jossa käytettävyyttä mitattiin satunnaismetsällä – RSA sopii paremmin aineistolle, jolla ei ole tarkoitus tehdä parametrista estimointia.

k-anonymiteetti ja l-diversiteetti voivat olla paikallaan aineistoille, joissa kvasitunniste ja luottamukselliset muuttujat eivät kata valtaosaa aineiston muuttujista: kaikkien muuttujien sisällyttäminen joko kvasitunnisteeseen tai luottamuksellisiin muuttujiin johti odotetusti aineiston muuttujien varianssien pienenemiseen ja käytettävyyden laskuun. Jos kaikkia muuttujia ei sisällytetty kvasitunnisteeseen tai luottamuksellisiin muuttujiin, ei varianssien pieneneminen ollut niin suurta, mutta aineistojen yksityisyydensuoja kärsi tästä. Tulokset ovat yhteneviä aiemman julkaisun [7] kanssa, jossa k-anonymiteetin ja l-diversiteetin käytettävyys mitattuna muun muassa Kullbackin-Leiblerin poikkeamalla laski lukujen k ja l kasvaessa.

Tässä tutkielmassa käsiteltiin vain muutamia anonymisointimenetelmiä; myös differentiaalisen yksityisyyden [21] ja täysin homomorfisten salausalgoritmien [22] (verrattuna osittain homomorfinen RSA) soveltuvuutta anonyymien aineistojen tuottamiseen voisi tutkia. Lisäksi, aineistoja oli vain yksi ja käytettävyyttä mitattiin vain regressiomalleilla; menetelmät saattaisivat tuottaa erilaisia tuloksia eri tilastollisilla menetelmillä. Tutkielmassa myös keskityttiin vain identiteetti- ja päättelyrikkomusten riskien estimoimiseen, eikä esimerkiksi ominaisuusrikkomuk-

sen riskiä arvioitu, vaikka sen mahdollisuutta k-anonyymin aineiston tapauksessa sivuttiinkin. Tutkielmassa kuitenkin esiteltiin, miten l-diversiteetti estää tämän k-anonymiteetissä mahdollisen kvasitunnisteen kautta tapahtuvan ominaisuusrikkomuksen. Osallisuusrikkomus ei vaikuttanut olevan COSMOS-aineiston tapauksessa uhka, mutta sen riskiä voisi arvioida esimerkiksi δ-läsnäololla (eng. δ-presence) [23].

Anonymisointi muuttaa aineistoa usealla tavalla ja se ei todennäköisesti ole paras tapa tuottaa toisiolain mukaisia anonyymejä tuloksia. Anonymisointi saattaa kuitenkin olla sopiva menetelmä tietoturvallisten aineistojen tuottamiseen yleiseen julkaisuun, sillä avoimen tieteen periaatteiden mukaisesti tutkimuksessa käytetettävät aineistot tulisi olla mahdollisimman saatavilla. Anonyymin aineiston voisi näiden periaatteiden mukaisesti julkaista artikkelin mukana suuntaa antavana aineistona, jota muut tutkijat voisivat käyttää omien hypoteesien rakentamiseen ja tutkimiseen, sekä julkaistun tutkimuksen tulosten toistamiseen. Jälkimmäisessä tapauksessa tietysti tulee valita sellainen anonymisointimenetelmä, joka säilyttää julkaisun tutkimustulokset. Anonyymin aineiston tutkiminen voisi toimia myös esiaskeleena tietoluvalle; jos aineisto vaikuttaa mielenkiintoiselta tai siitä saadaan alustavia tuloksia, voisi tutkija pyytää alkuperäisen tutkimusaineiston käyttöönsä. Anonymisointia kuitenkin hankaloittaa se, että anonyymille tiedolle ei ole olemassa Findatan määrittelemiä mittareita tai niiden raja-arvoja, joiden avulla anonymisoija voisi itse todeta, onko tieto anonyymiä vai ei [4]. Tämän osalta siis kaivattaisiin yhä enemmän lisätutkimusta sekä viranomaislinjauksia tutkimustuloksiin nojaten.

Viitteet

[1] Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer ja Peter Paul de Wolf. *Statistical disclosure control*. Wiley Series in Survey Methodology. Wiley, 2012, s. 2–6. DOI: 10.1002/9781118348239.

[2] Euroopan parlamentin ja neuvoston asetus (EU) 2016/679. *Luonnollisten henkilöiden suojelusta henkilötietojen käsittelyssä sekä näiden tietojen vapaasta liikkuvuudesta ja direktiivin 95/46/EY kumoamisesta (yleinen tietosuoja-asetus)*. https://eur-lex.europa.eu/legal-content/FI/TXT/PDF/?uri=CELEX:32016R0679&from=EN. Luettu: 1.6.2023. 2016.

[3] Eduskunta. *Laki sosiaali- ja terveystietojen toissijaisesta käytöstä 552/2019*. https://www.finlex.fi/fi/laki/alkup/2019/20190552. Luettu: 1.6.2023. 2019.

[4] Findata. *Anonyymien tulosten tuottaminen*. https://findata.fi/palvelut-ja-ohjeet/anonyymien-tulosten-tuottaminen/. Luettu: 2.6.2023.

[5] Latanya Sweeney. "k-anonymity: a model for protecting privacy". *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10 (05 lokakuu 2002), s. 557–570. ISSN: 0218-4885. DOI: 10.1142/S0218488502001648.

[6] Tietosuojavaltuutetun toimisto. *Pseudonymisointi ja anonymisointi*. https://tietosuoja.fi/pseudonymisointi-anonymisointi. Luettu: 27.9.2023.

[7] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke ja Muthuramakrishnan Venkitasubramaniam. "l-diversity: privacy beyond k-anonymity". *ACM Transactions on Knowledge Discovery from Data* 1 (1 maaliskuu 2007). ISSN: 15564681. DOI: 10.1145/1217299.1217302.

[8] Olga Vovk, Gunnar Piho ja Peeter Ross. "Methods and tools for healthcare data anonymization: a literature review". *International Journal of General Systems* (2023). ISSN: 15635104. DOI: 10.1080/03081079.2023.2173749.

[9] Thomas A. Lasko ja Staal A. Vinterbo. "Spectral anonymization of data". *IEEE Transactions on Knowledge and Data Engineering* 22 (3 maaliskuu 2010), s. 437–446. ISSN: 10414347. DOI: 10.1109/TKDE.2009.88.

[10] Raza Imam, Qazi Mohammad Areeb, Abdulrahman Alturki ja Faisal Anwer. "Systematic and critical review of RSA based public key cryptographic schemes: past and present status". *IEEE Access* 9 (2021), s. 155949–155976. DOI: 10.1109/ACCESS.2021.3129224.

[11] Fazle Rabbi, Amin Aminifar, Yngve Lamo ja Ka I Pun. *A practical methodology for anonymization of structured health data*. URL: https://www.researchgate.net/publication/337325212.

[12] Abdul Majeed ja Sungchang Lee. "Anonymization techniques for privacy preserving data publishing: a comprehensive survey". *IEEE Access* 9 (2021), s. 8515. ISSN: 21693536. DOI: 10.1109/ACCESS.2020.3045700.

[13] Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan ja Adam Smith. "Composition attacks and auxiliary information in data privacy". Teoksessa: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '08. Las Vegas, Nevada, USA: Association for Computing Machinery, 2008, s. 265–273. ISBN: 9781605581934. DOI: 10.1145/1401890.1401926. URL: https://doi.org/10.1145/1401890.1401926.

[14] K.E. Emam, L. Mosquera ja R. Hoptroff. *Practical synthetic data generation: balancing privacy and the broad availability of data*. O'Reilly Media, 2020. ISBN: 9781492072713.

[15] Anssi Auvinen, Maria Feychting, Anders Ahlbom, Lena Hillert, Paul Elliott, Joachim Schüz, Hans Kromhout, Mireille B Toledano, Christoffer Johansen, Aslak Harbo Poulsen, Roel Vermeulen, Sirpa Heinävaara, Katja Kojo, Giorgio Tettamanti ja COSMOS Study Group. "Headache, tinnitus and hearing loss in the international cohort study of mobile phone use and health (COSMOS) in Sweden and Finland". *International Journal of Epidemiology* 48.5 (heinäkuu 2019), s. 1567–1579. ISSN: 0300-5771. DOI: 10.1093/ije/dyz127. eprint: https://academic.oup.com/ije/article-pdf/48/5/1567/30800553/dyz127.pdf. URL: https://doi.org/10.1093/ije/dyz127.

[16] Alan Agresti. *Foundations of linear and generalized linear models Wiley series in probability and statistics*. Wiley, 2015.

[17] T Chandra Segar ja R Vijayaragavan. *Pell's RSA key generation and its security analysis*. 2013. DOI: 10.1109/ICCCNT.2013.6726659.

[18] Kannan Balasubramanian. “Variants of RSA and their cryptanalysis”. Teoksessa: *2014 International Conference on Communication and Network Technologies*. 2014, s. 145–149. DOI: 10.1109/CNT.2014.7062742.

[19] R. L. Rivest, A. Shamir ja L. Adleman. “A method for obtaining digital signatures and public-key cryptosystems”. *Commun. ACM* 21.2 (helmikuu 1978), s. 120–126. ISSN: 0001-0782. DOI: 10.1145/359340.359342. URL: https://doi.org/10.1145/359340.359342.

[20] Johannes Rajala. *anon: anonymization tools for micro data*. R package version 0.1.0. URL: https://github.com/rajalah71/anon.

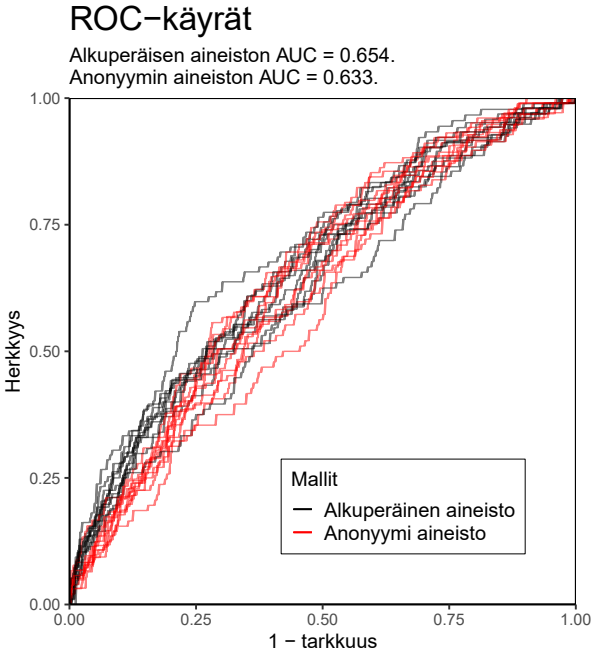
[21] Cynthia Dwork. “Differential privacy”. Teoksessa: *Automata, languages and programming*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, s. 1–12. ISBN: 978-3-540-35908-1. DOI: https://doi.org/10.1007/11787006_1.

[22] X. Yi, R. Paulet ja E. Bertino. *Homomorphic encryption and applications*. SpringerBriefs in Computer Science. Springer International Publishing, 2014, s. 47–66. ISBN: 9783319122298. URL: https://books.google.fi/books?id=OgA6BQAAQBAJ.

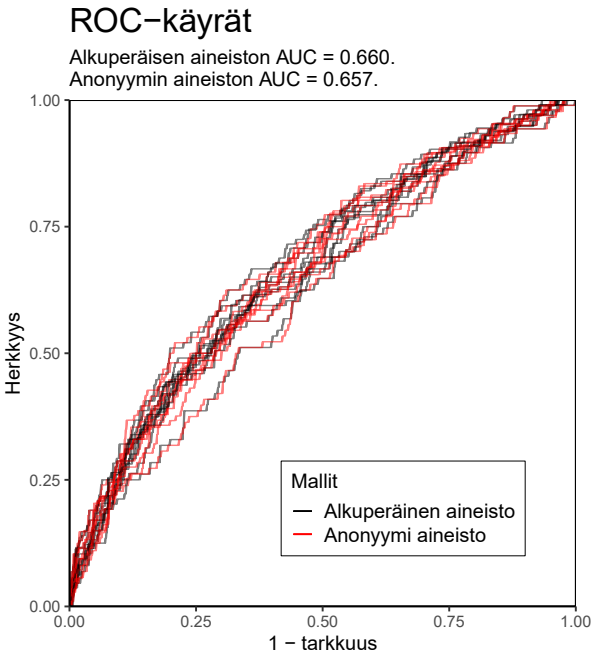
[23] Mehmet Ercan Nergiz ja Chris Clifton. “δ-presence without complete world knowledge”. *IEEE Trans. on Knowl. and Data Eng.* 22.6 (kesäkuu 2010), s. 868–883. ISSN: 1041-4347. DOI: 10.1109/TKDE.2009.125. URL: https://doi.org/10.1109/TKDE.2009.125.

[24] R Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria, 2023. URL: https://www.R-project.org/.

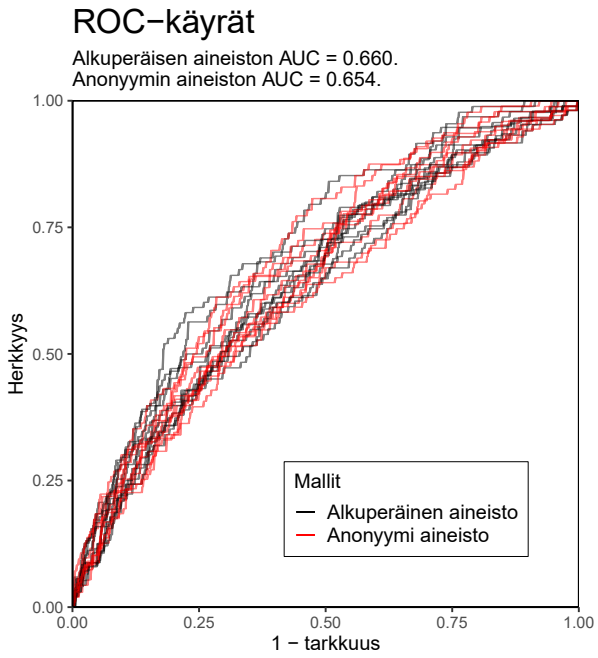
A



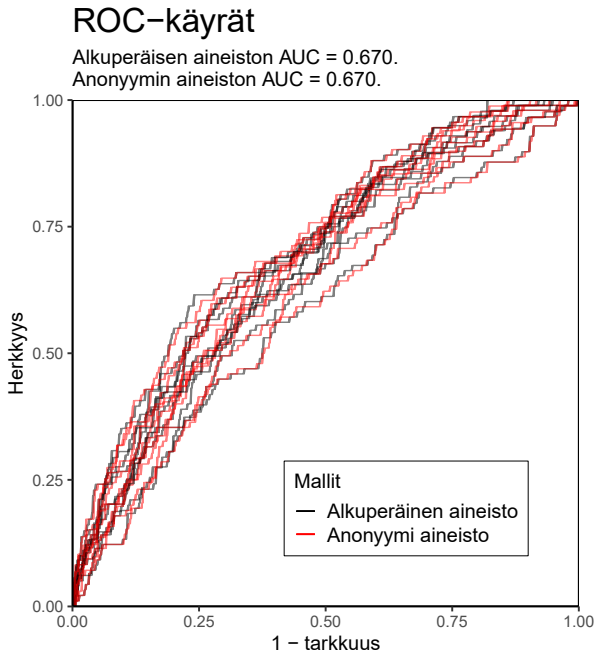
Kuva 17: ROC-käyrä ja AUC-pistemäärä 5-anonymiteetti-opetusaineistoille (10 kpl).



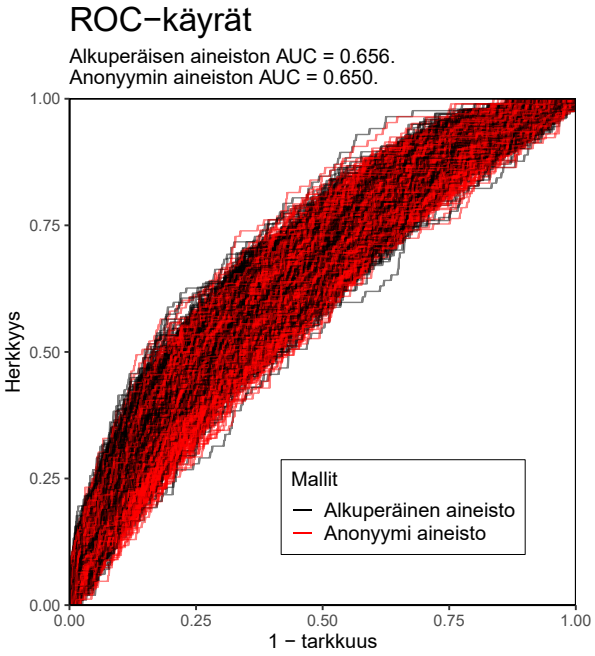
Kuva 18: ROC-käyrä ja AUC-pistemäärä 5-anonymiteetti pieni -opetusaineistoille (10 kpl).



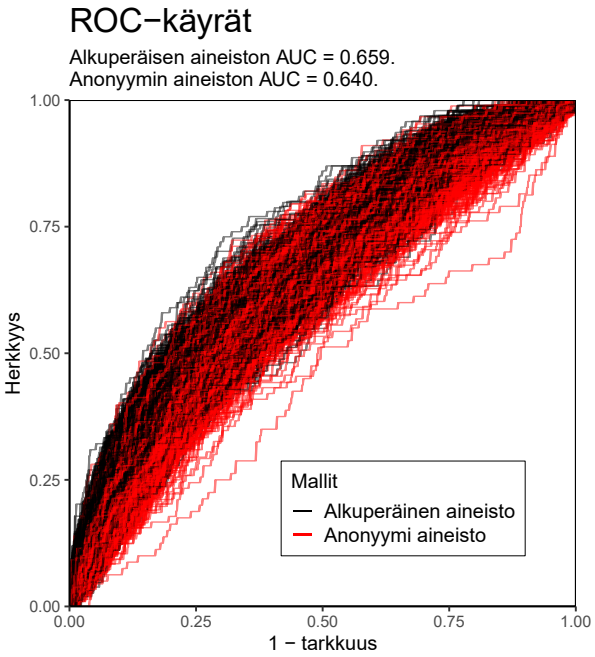
Kuva 19: ROC-käyrä ja AUC-pistemäärä 2-diversiteetti-opetusaineistoille (10 kpl).



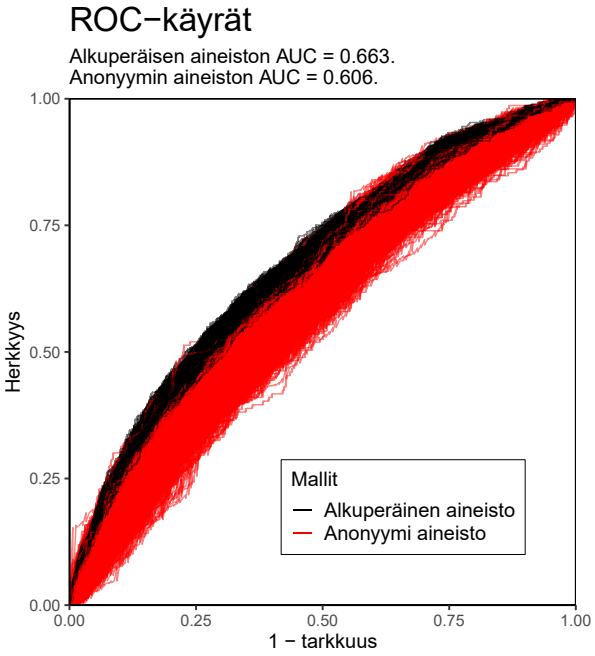
Kuva 20: ROC-käyrä ja AUC-pistemäärä 2-diversiteetti pieni -opetusaineistoille (10 kpl).



Kuva 21: ROC-käyrä ja AUC-pistemäärä spektraalinen kohina -opetusaineistoille (100 kpl).

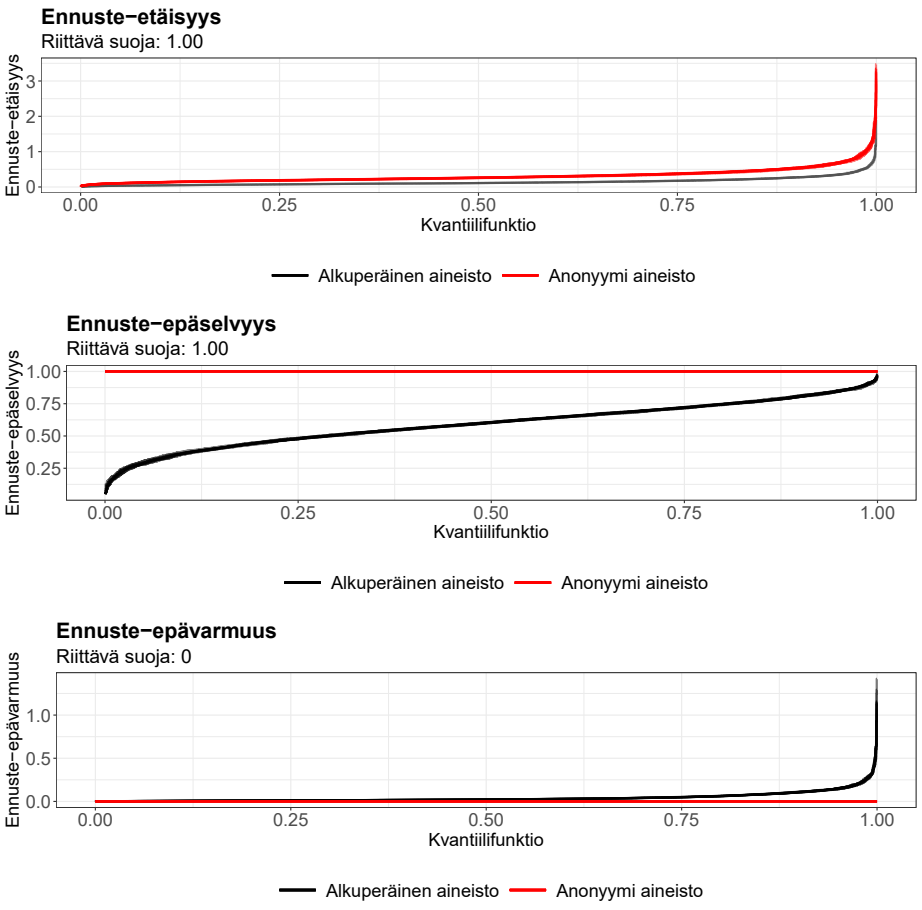


Kuva 22: ROC-käyrä ja AUC-pistemäärä spektraalinen sarakepermutaatio -opetusaineistoille (100 kpl).

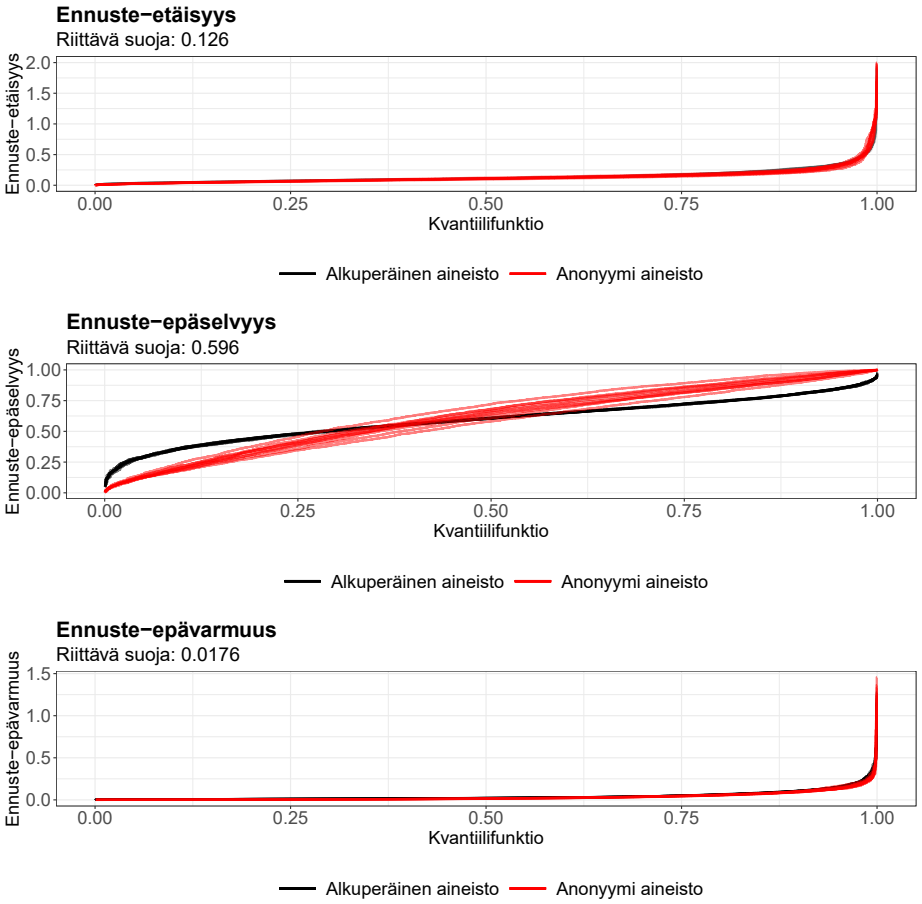


Kuva 23: ROC-käyrä ja AUC-pistemäärä RSA-opetusaineistoille (100 kpl).

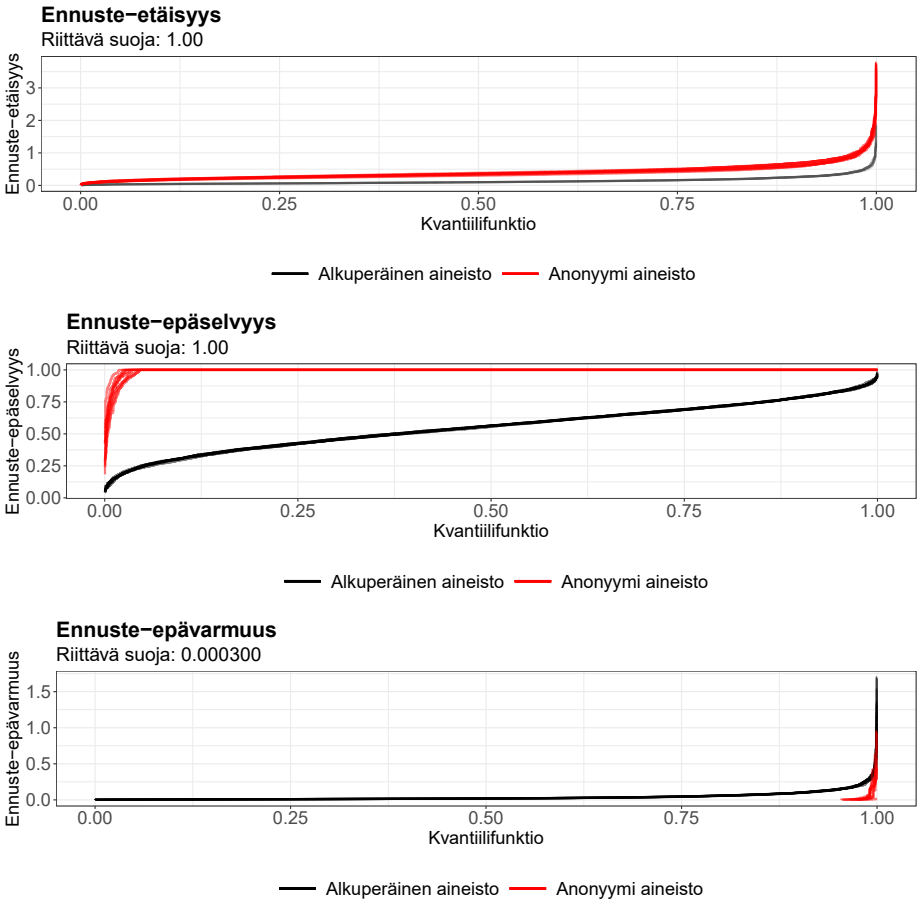
B



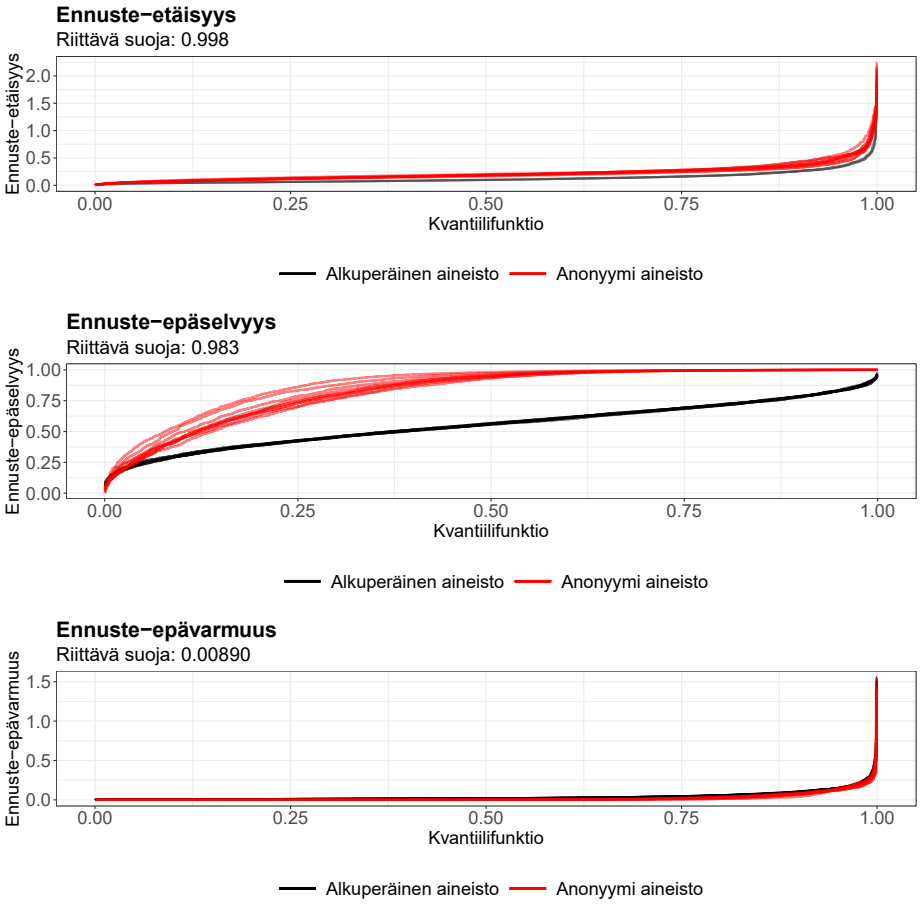
Kuva 24: 5-anonymiteetti-opetusaineistojen ennuste-etäisyydet, ennuste-epäselvyydet ja ennuste-epävarmuudet. Riittävän suojan kriteeri lasketaan keskiarvona 10 opetusaineiston kriteerin arvosta. Kriteeri toteutuu täysin ennuste-etäisyydelle ja -epäselvyydelle, muttei lainkaan -epävarmuudelle.



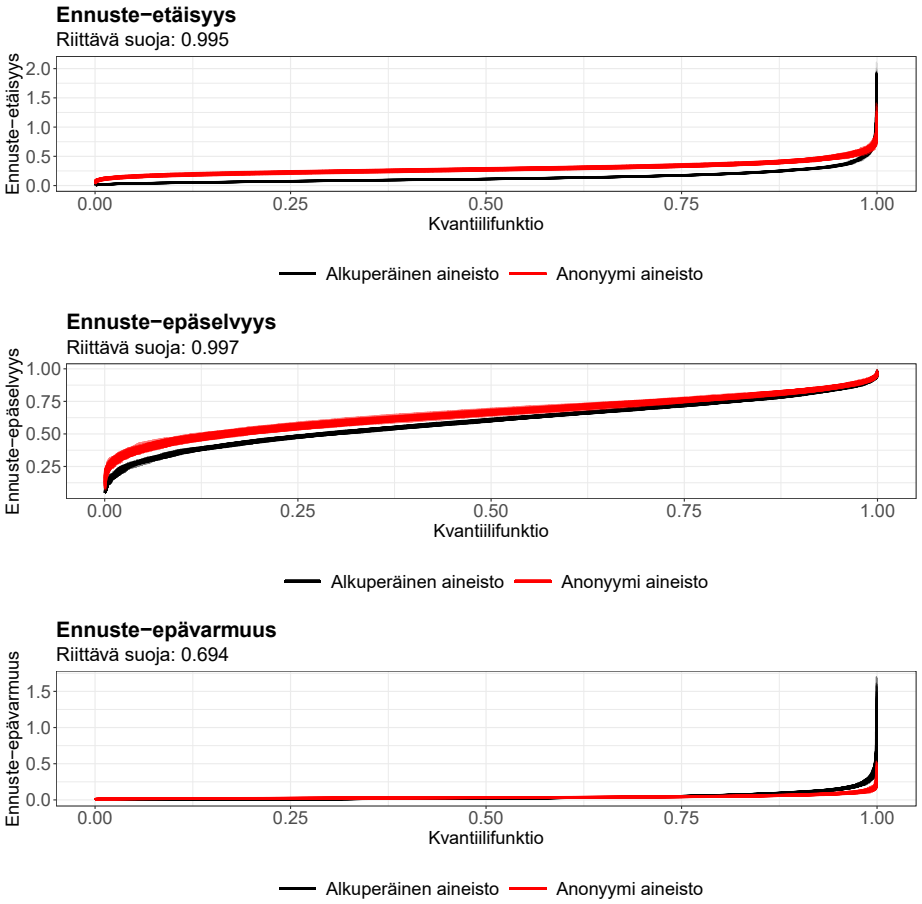
Kuva 25: 5-anonymiteetti pieni -opetusaineistojen ennuste-etäisyydet, ennuste-epäselvyydet ja ennuste-epävarmuudet. Riittävän suojan kriteeri lasketaan keskiarvona 10 opetusaineiston kriteerin arvosta. Menetelmä ei toteuta riittävän suojan kriteeriä millään mittarilla.



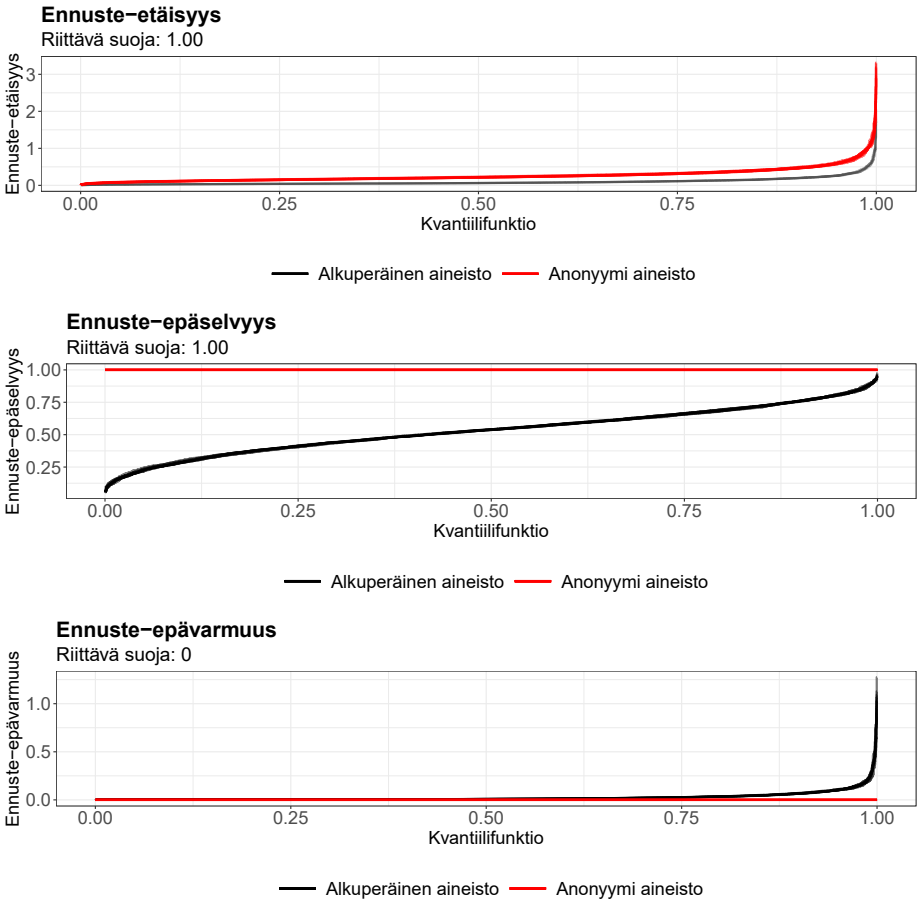
Kuva 26: 2-diversiteetti-opetusaineistojen ennuste-etäisyydet, -epäselvyydet ja -epävarmuudet. Riittävän suojan kriteeri lasketaan keskiarvona 10 opetusaineiston kriteerin arvosta. Kriteeri toteutuu ennuste-etäisyydelle ja -epäselvyydelle, muttei -epävarmuudelle.



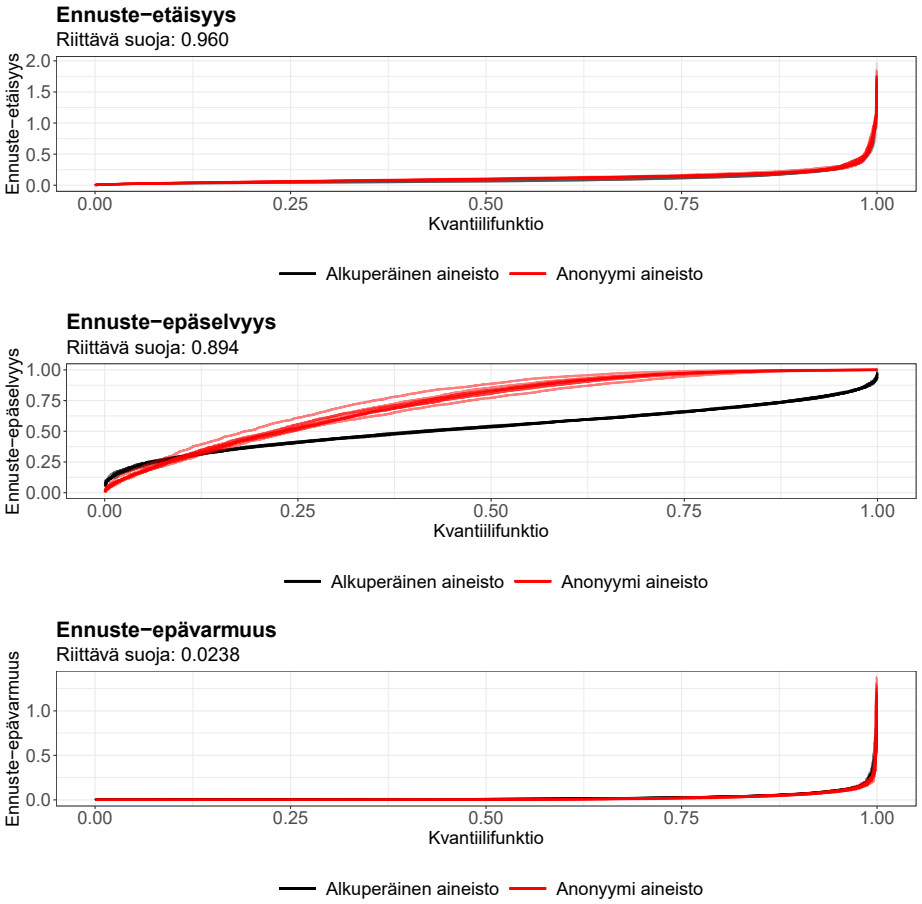
Kuva 27: 2-diversiteetti pieni -opetusaineistojen ennuste-etäisyydet, ennuste-epäselvyydet ja ennuste-epävarmuudet. Riittävän suojan kriteeri lasketaan keskiarvona 10 opetusaineiston kriteerin arvosta. Kriteeri toteutuu melkein ennuste-etäisyydelle ja -epäselvyydelle, muttei lainkaan -epävarmuudelle.



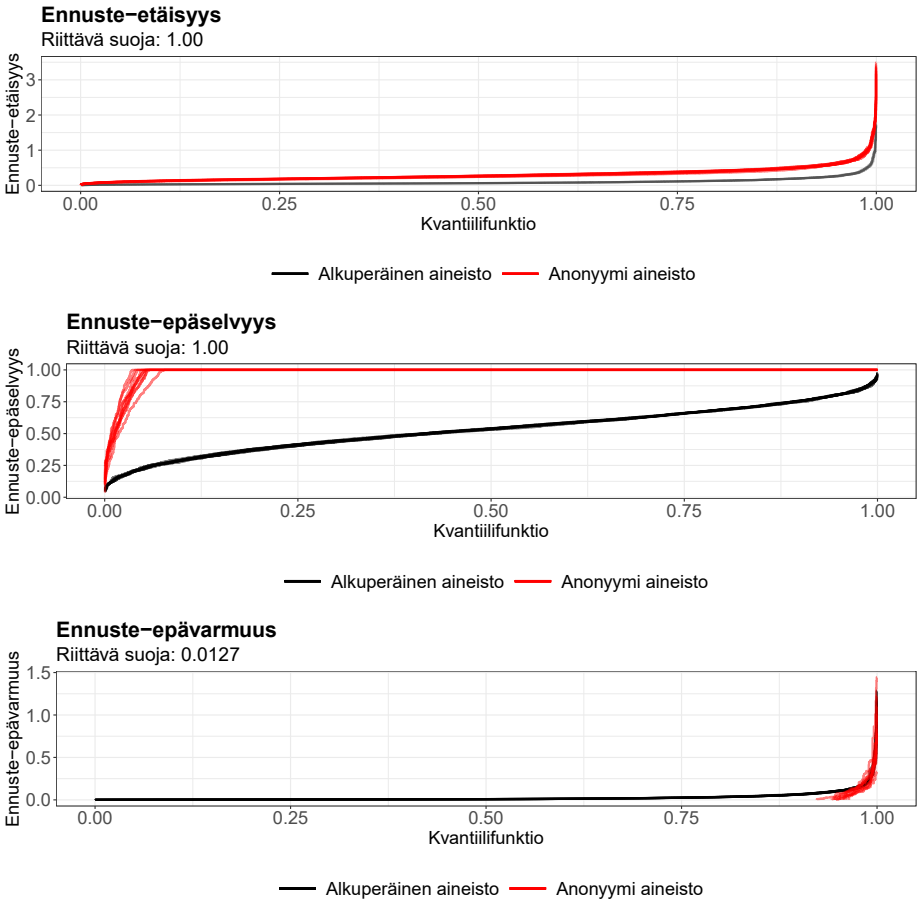
Kuva 28: Spektraalinen kohina -opetusaineistojen ennuste-etäisyydet, ennuste-epäselvyydet ja ennuste-epävarmuudet. Riittävän suojan kriteeri lasketaan keskiarvona 100 opetusaineiston kriteerin arvosta. Kriteeri toteutuu melkein ennuste-etäisyydelle ja -epäselvyydelle, muttei -epävarmuudelle.



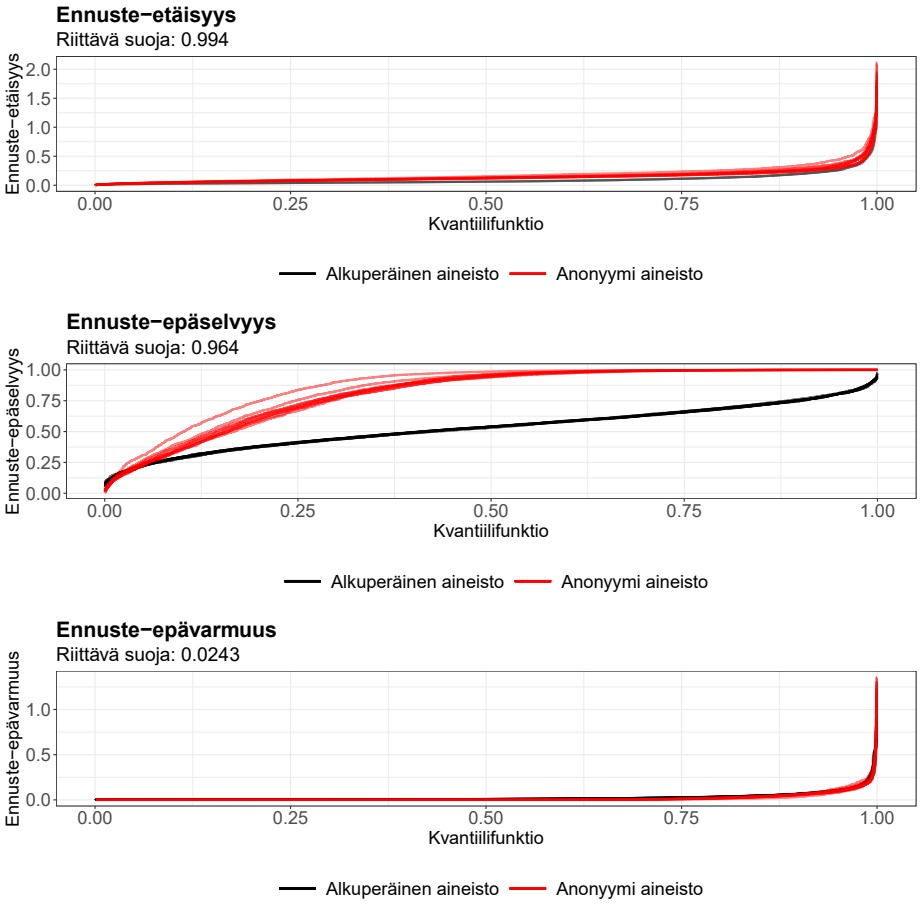
Kuva 29: 5-anonymiteetti-opetusaineistojen ennuste-etäisyydet, ennuste-epäselvyydet ja ennuste-epävarmuudet. Riittävän suojan kriteeri lasketaan keskiarvona 10 opetusaineiston kriteerin arvosta. Kriteeri toteutuu täysin ennuste-etäisyydelle ja -epäselvyydelle, muttei lainkaan -epävarmuudelle.



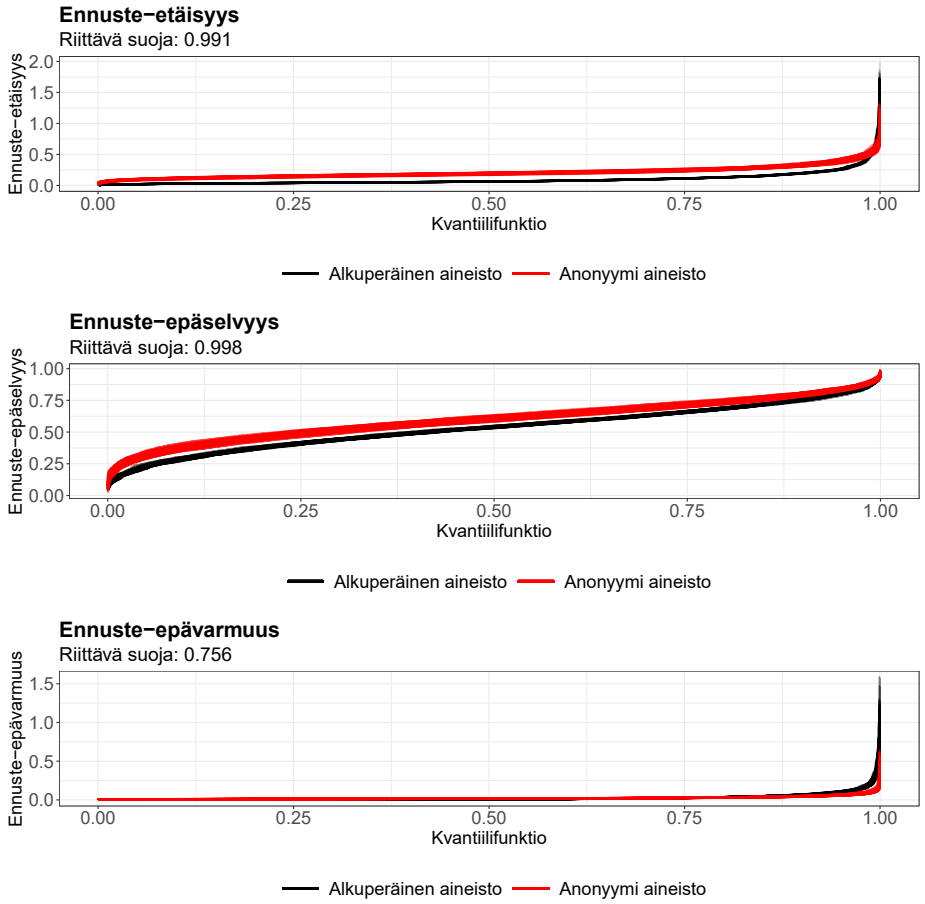
Kuva 30: 5-anonymiteetti pieni -opetusaineistojen ennuste-etäisyydet, -epäselvyydet ja -epävarmuudet. Riittävän suojan kriteeri lasketaan keskiarvona 10 opetusaineiston kriteerin arvosta. Menetelmä ei toteuta riittävän suojan kriteeriä millään mittarilla.



Kuva 31: 2-diversiteetti-opetusaineistojen ennuste-etäisyydet, -epäselvyydet ja -epävarmuudet. Riittävän suojan kriteeri lasketaan keskiarvona 10 opetusaineiston kriteerin arvosta. Kriteeri toteutuu ennuste-etäisyydelle ja -epäselvyydelle, muttei -epävarmuudelle.



Kuva 32: 2-diversiteetti pieni -opetusaineistojen ennuste-etäisyydet, -epäselvyydet ja -epävarmuudet. Riittävän suojan kriteeri lasketaan keskiarvona 100 opetusaineiston kriteerin arvosta. Kriteeri toteutuu melkein ennuste-etäisyydelle ja -epäselvyydelle, muttei -epävarmuudelle.



Kuva 33: Spektraalinen kohina -opetusaineistojen ennuste-etäisyydet, -epäselvyydet ja -epävarmuudet. Riittävän suojan kriteeri lasketaan keskiarvona 100 opetusaineiston kriteerin arvosta. Kriteeri toteutuu melkein ennuste-etäisyydelle ja -epäselvyydelle, muttei -epävarmuudelle.

C

Tutkielman teossa on käytetty R-ohjelman versiota 4.3.2 [24] sekä RStudion versiota 2023.09.0 + 463.

```
write_data_lin = function(n = 10){

  COSMOS_HDS2 <- read.csv("C:/Users/Johannes/Desktop/Gradu/DATA/COSMOS_HDS2.csv")

  cols = c("hit_score_r",
           "Age", "Gender",   "sf12gr", "Depression",  "Bmi",

           "traffic3_min_hf")

  subset = COSMOS_HDS2[, cols]

  data_omit = na.omit(subset)
  empty_cells <- data_omit == "" | data_omit == " "
  data_clear = data_omit[empty_cells] <- NA
  data_omitted = na.omit(data_clear)
  droprows = c("73", "74", "187", "317", "412", "417", "851", "1375", "1445", "1613
      ", "1626", "1661", "1677", "1887", "2455", "2594", "3226")
  data = na.omit(data_omit[!(rownames(data_omit) %in% droprows), ])

  datalist = list()
```

```
  for(i in 1:n){
    index = sample(c(TRUE, FALSE), nrow(data), TRUE, prob = c(0.8, 0.2))
    train = data[index, ]
    test = data[!index, ]
    datalist[[i]] = list(train = train, test = test)
  }

  return(datalist)
}

datalist = write_data_lin()

setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

kanons = list()
index = 1
for(pair in datalist){
  anon = kAnon(pair$train, 5)
  item = list(train = pair$train, anon = anon, test = pair$test)
  kanons[[index]] = item
  index = index + 1
}

# save list to disk
setwd("C:/Users/Johannes/Desktop/Gradu/DATA/lin_kanon")
saveRDS(kanons, "kanons.rds")
```

```
datalist = write_data_lin()

setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

kanons = list()
index = 1
for(pair in datalist){
  anon = kAnon(pair$train, 5, c("Bmi", "Age", "Gender"))
  item = list(train = pair$train, anon = anon, test = pair$test)
  kanons[[index]] = item
  index = index + 1
}

# save list to disk
setwd("C:/Users/Johannes/Desktop/Gradu/DATA/lin_kanon_real")

saveRDS(kanons, "kanons_real.rds")
```

```
datalist = write_data_lin()

setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

kanons = list()
index = 1
for(pair in datalist){
  score_general = sensitive_generalizer(pair$train$hit_score_r, subsets = 6)
  pair$train$hit_score_r = score_general
  anon = lDiversity(pair$train, c("hit_score_r", "Depression"), l=2)
  item = list(train = pair$train, anon = anon, test = pair$test)
  kanons[[index]] = item
  index = index + 1
}

# save list to disk
setwd("C:/Users/Johannes/Desktop/Gradu/DATA/lin_ldiv")

saveRDS(kanons, "ldiv.rds")
```

```
datalist = write_data_lin()

setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

kanons = list()
index = 1
for(pair in datalist){
```

```
  score_general = sensitive_generalizer(pair$train$hit_score_r, subsets = 6)
  pair$train$hit_score_r = score_general
  anon = lDiversity(pair$train, c("hit_score_r", "Depression"), l=2, c("Bmi", "Age"
      , "Gender"))
  item = list(train = pair$train, anon = anon, test = pair$test)
  kanons[[index]] = item
  index = index + 1
}

# save list to disk
setwd("C:/Users/Johannes/Desktop/Gradu/DATA/lin_ldiv_real")

saveRDS(kanons, "ldiv_real.rds")
```

```
datalist = write_data_lin(n = 100)

setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

kanons = list()
index = 1
for(pair in datalist){
  anon = spectral(pair$train, function(x) sensitive_noise(x, 5))
  item = list(train = pair$train, anon = anon, test = pair$test)
  kanons[[index]] = item
  index = index + 1
}

# save list to disk
setwd("C:/Users/Johannes/Desktop/Gradu/DATA/lin_noise")

saveRDS(kanons, "noise.rds")
```

```
datalist = write_data_lin(n = 100)

setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

kanons = list()
index = 1
for(pair in datalist){
  anon = spectral(pair$train, cell_swap)
  item = list(train = pair$train, anon = anon, test = pair$test)
  kanons[[index]] = item
  index = index + 1
}

# save list to disk
setwd("C:/Users/Johannes/Desktop/Gradu/DATA/lin_swap")

saveRDS(kanons, "swap.rds")
```

```
datalist = write_data_lin(n = 100)

setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

kanons = list()
index = 1
for(pair in datalist){
  train_indicies = rownames(pair$train)
  combined = rbind(pair$train, pair$test)
  anon = encrypt(combined)
  train = anon[rownames(anon) %in% train_indicies, ]
  test = anon[!(rownames(anon) %in% train_indicies), ]
  item = list(train = pair$train, anon = train, test = test)
  kanons[[index]] = item
  index = index + 1
}

# save list to disk
setwd("C:/Users/Johannes/Desktop/Gradu/DATA/lin_rsa")

saveRDS(kanons, "rsa.rds")
```

```
write_data_bin = function(n = 10){

```

```
  COSMOS_HDS2 <- read.csv("C:/Users/Johannes/Desktop/Gradu/DATA/COSMOS_HDS2.csv")

  cols = c("headache_FU",
           "Age", "Gender",   "sf12gr", "Depression",  "Bmi",

           "traffic3_min_hf")

  subset = COSMOS_HDS2[, cols]

  data_omit = na.omit(subset)
  empty_cells <- data_omit == "" | data_omit == " "
  data_clear = data_omit[empty_cells] <- NA
  data_omitted = na.omit(data_clear)
  droprows = c("73", "74", "187", "317", "412", "417", "851", "1375", "1445", "1613
      ", "1626", "1661", "1677", "1887", "2455", "2594", "3226")
  data = na.omit(data_omit[!(rownames(data_omit) %in% droprows), ])

  # encode response variable from 0,1 to no,yes
  data$headache_FU = ifelse(data$headache_FU == 1, "Yes", "No")

  datalist = list()

  for(i in 1:n){
    index = sample(c(TRUE, FALSE), nrow(data), TRUE, prob = c(0.8, 0.2))
    train = data[index, ]
    test = data[!index, ]
    datalist[[i]] = list(train = train, test = test)
  }

  return(datalist)
}

datalist = write_data_bin()

setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

kanons = list()
index = 1
for(pair in datalist){
  anon = kAnon(pair$train, 5)
  item = list(train = pair$train, anon = anon, test = pair$test)
  kanons[[index]] = item
  index = index + 1
}

# save list to disk
setwd("C:/Users/Johannes/Desktop/Gradu/DATA/bin_kanon")
saveRDS(kanons, "kanons.rds")
```

```
datalist = write_data_bin()

setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

kanons = list()
index = 1
for(pair in datalist){
  anon = kAnon(pair$train, 5, c("Bmi", "Age", "Gender"))
  item = list(train = pair$train, anon = anon, test = pair$test)
  kanons[[index]] = item
  index = index + 1
}

# save list to disk
setwd("C:/Users/Johannes/Desktop/Gradu/DATA/bin_kanon_real")

saveRDS(kanons, "kanons_real.rds")
```

```
datalist = write_data_bin()

setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

kanons = list()
index = 1
for(pair in datalist){
  anon = lDiversity(pair$train, c("headache_FU", "Depression"), l=2)
```

```
  item = list(train = pair$train, anon = anon, test = pair$test)
  kanons[[index]] = item
  index = index + 1
}

# save list to disk
setwd("C:/Users/Johannes/Desktop/Gradu/DATA/bin_ldiv")

saveRDS(kanons, "ldiv.rds")
```

```
datalist = write_data_bin()

setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

kanons = list()
index = 1
for(pair in datalist){
  anon = lDiversity(pair$train, sensitiveAttributes =  c("headache_FU", "Depression
      "), l = 2, quasiIdentifiers = c("Bmi", "Age", "Gender"))
  item = list(train = pair$train, anon = anon, test = pair$test)
  kanons[[index]] = item
  index = index + 1
}

# save list to disk
setwd("C:/Users/Johannes/Desktop/Gradu/DATA/bin_ldiv_real")

saveRDS(kanons, "ldiv_real.rds")
```

```
datalist = write_data_bin(n = 100)

setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

kanons = list()
index = 1
for(pair in datalist){
  anon = spectral(pair$train, function(x) sensitive_noise(x, 5))
  item = list(train = pair$train, anon = anon, test = pair$test)
  kanons[[index]] = item
  index = index + 1
}

# save list to disk
setwd("C:/Users/Johannes/Desktop/Gradu/DATA/bin_noise")

saveRDS(kanons, "noise.rds")
```

```
datalist = write_data_bin(n = 100)

setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

kanons = list()
index = 1
for(pair in datalist){
  anon = spectral(pair$train, cell_swap)
  item = list(train = pair$train, anon = anon, test = pair$test)
  kanons[[index]] = item
  index = index + 1
}

# save list to disk
setwd("C:/Users/Johannes/Desktop/Gradu/DATA/bin_swap")

saveRDS(kanons, "swap.rds")
```

```
datalist = write_data_bin(n = 100)

setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

kanons = list()
index = 1
for(pair in datalist){
  train_indicies = rownames(pair$train)
```

```
  combined = rbind(pair$train, pair$test)
  anon = encrypt(combined)
  train = anon[rownames(anon) %in% train_indicies, ]
  test = anon[!(rownames(anon) %in% train_indicies), ]
  item = list(train = pair$train, anon = train, test = test)
  kanons[[index]] = item
  index = index + 1
}

# save list to disk
setwd("C:/Users/Johannes/Desktop/Gradu/DATA/bin_rsa")

saveRDS(kanons, "rsa.rds")
```

```
#data-------------
train = read.csv("C:/Users/Johannes/Desktop/Gradu/DATA/train_lin", row.names=1)
swap_lin = read.csv("C:/Users/Johannes/Desktop/Gradu/DATA/swap_lin", row.names=1)
rsa_lin = read.csv("C:/Users/Johannes/Desktop/Gradu/DATA/rsa_lin", row.names=1)
plotcols = c("hit_score_r", "Age", "Bmi", "traffic3_min_hf")
setwd("C:/Users/Johannes/Desktop/Gradu/Gradun kuvat/esims/scatter")
color_palette <- colorRampPalette(c("#00a6a6", "#ff80bf"))
point_colors <- color_palette(100)[cut(train$hit_score_r, 100)]


#original---------------
pdf("original_lin_esim.pdf")
plot(train[, plotcols], xaxt='n', yaxt='n', col = point_colors )
title("Alkuperäinen aineisto", cex.main = 1.4)
dev.off()

#rsa--------------
point_colors <- color_palette(100)[cut(rsa_lin$hit_score_r, 100)]
pdf("rsa_lin_esim.pdf")
plot(rsa_lin[, plotcols], xaxt='n', yaxt='n', col = point_colors)
title("Menetelmä 1", cex.main = 1.4)
dev.off()


#swap-------------
point_colors <- color_palette(100)[cut(swap_lin$hit_score_r, 100)]
pdf("swap_lin_esim.pdf")
plot(swap_lin[, plotcols], xaxt='n', yaxt='n', col = point_colors)
title("Menetelmä 2", cex.main = 1.4)
dev.off()
```

```
train = read.csv("C:/Users/Johannes/Desktop/Gradu/DATA/train_lin", row.names=1)
kanon_lin <- read.csv("C:/Users/Johannes/Desktop/Gradu/DATA/kanon_lin", row.names
    =1)
ldiv_lin <- read.csv("C:/Users/Johannes/Desktop/Gradu/DATA/ldiv_lin", row.names=1)
noise_lin = read.csv("C:/Users/Johannes/Desktop/Gradu/DATA/noise_lin", row.names=1)
swap_lin = read.csv("C:/Users/Johannes/Desktop/Gradu/DATA/swap_lin", row.names=1)
rsa_lin = read.csv("C:/Users/Johannes/Desktop/Gradu/DATA/rsa_lin", row.names=1)

plotcols = c("hit_score_r", "Age", "Bmi", "traffic3_min_hf")

setwd("C:/Users/Johannes/Desktop/Gradu/Gradun kuvat/Lin/scatter")

# Create a color palette based on the continuous variable 'z'
color_palette <- colorRampPalette(c("#00a6a6", "#ff80bf"))

# Generate colors based on the gradient
point_colors <- color_palette(100)[cut(train$hit_score_r, 100)]  # Adjust the
    number of colors as needed
pdf("original_lin.pdf")
plot(train[, plotcols], xaxt='n', yaxt='n', col = point_colors)
title("Alkuperäinen aineisto", cex.main = 1.5, line = 1.5)
dev.off()

# -------------- kanon

# Generate colors based on the gradient
point_colors <- color_palette(100)[cut(kanon_lin$hit_score_r, 100)]  # Adjust the
    number of colors as needed
pdf("kanon_lin.pdf")
plot(kanon_lin[, plotcols], xaxt='n', yaxt='n', col = point_colors)
title("5-anonymiteetti", cex.main = 1.5, line = 1.5)
dev.off()

```

```
#------------- ldiv

# Generate colors based on the gradient
point_colors <- color_palette(100)[cut(ldiv_lin$hit_score_r, 100)]  # Adjust the number of colors as needed
pdf("ldiv_lin.pdf")
plot(ldiv_lin[, plotcols], xaxt='n', yaxt='n', col = point_colors)
title("2-diversiteetti", cex.main = 1.5, line = 1.5)
dev.off()

#------------- noise

# Generate colors based on the gradient
point_colors <- color_palette(100)[cut(noise_lin$hit_score_r, 100)]  # Adjust the number of colors as needed
pdf("noise_lin.pdf")
plot(noise_lin[, plotcols], xaxt='n', yaxt='n', col = point_colors)
title("Spektraalinen kohina", cex.main = 1.5, line = 1.5)
dev.off()

#------------- swap

point_colors <- color_palette(100)[cut(swap_lin$hit_score_r, 100)]  # Adjust the number of colors as needed
pdf("swap_lin.pdf")
plot(swap_lin[, plotcols], xaxt='n', yaxt='n', col = point_colors)
title("Spektraalinen sarakepermutaatio", cex.main = 1.5, line = 1.5)
dev.off()

#------------- rsa

# Generate colors based on the gradient
point_colors <- color_palette(100)[cut(rsa_lin$hit_score_r, 100)]  # Adjust the number of colors as needed
pdf("rsa_lin.pdf")
plot(rsa_lin[, plotcols], xaxt='n', yaxt='n', col = point_colors)
title("RSA", cex.main = 1.5, line = 1.5)
dev.off()

setwd("C:/Users/Johannes/Desktop/Gradu/anon")
```

```
setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()
ISLR::Default
index = sample(c(TRUE, FALSE), replace = TRUE, size = length(Default), prob = c(0.7, 0.3))
train = Default[index, ]
test = Default[!index, ]
noise = spectral(train, cell_swap)

lm = glm(as.factor(default) ~., family = binomial(), data = train)
noiselm = glm(as.factor(default) ~., family = binomial(), data = noise)

setwd("C:/Users/Johannes/Desktop/Gradu/Gradun kuvat")
roc_plot(lm,  noiselm, test)

dev.copy2pdf(file = "testiroc.pdf")
```

```
# Function to concstruct desired plots for examples

prediction_plot_esim = function(prediction_all_output, k, n = 10000, dist = euc_dist){

  # Helper function to make vertical lines if needed
  vert_maker = function(y){
    # If only one unique value is present, return a sequence from 0 to 1 instead
    table = table(y)
    if(length(table) == 1){
      vert = seq(0,1,length.out = length(y))
      return(vert)
    }
    # Else return the original data unchanged
    return(y)
  }

  # estimate cdf from data and calculate n equidistant points from it
  ecdf_points <- function(data) {
    # Create an ECDF function
```

```
  ecdf_func <- ecdf(data)

  # Generate a sequence of values for x-axis
  x <- seq(min(data), max(data), length.out = n)

  # Calculate the estimated CDF values for the x-axis values
  y <- ecdf_func(x)

  # return x and y
  return(list(x,y))

}

quantile_distance_0_1_calc = function(og, anon, n = 1000){
  # calculate the proportion of points in anon quantile function that have no smaller value thank in og quantile function
  amount = 0

  # Estimate quantile functions
  og_q = quantile(og, probs = seq(0, 1, 1/n))
  anon_q = quantile(anon, probs = seq(0, 1, 1/n))

  # Calculate the proportion
  for(i in 1:n){
    if(og_q[i] <= anon_q[i]){
      amount = amount + 1
    }
  }
  return(amount/n)
}

# prediction_distance
og_list = lapply(seq_along(prediction_all_output), function(x) ecdf_points(prediction_all_output[[x]]$original$prediction_distance))
ref_list = lapply(seq_along(prediction_all_output), function(x) ecdf_points(prediction_all_output[[x]]$reference$prediction_distance))

erotus_list = lapply(seq_along(prediction_all_output), function(x) quantile_distance_0_1_calc(prediction_all_output[[x]]$original$prediction_distance, prediction_all_output[[x]]$reference$prediction_distance))
mean_erotus = mean(unlist(erotus_list))
format = formatC(signif(mean_erotus, digits=3), digits=3, format="fg", flag="#")

# plot with ggplot2
p_distance = ggplot(data = NULL) +
  lapply(seq_along(prediction_all_output), function(x) geom_line(aes(x = vert_maker(og_list[[x]][[2]]), y = og_list[[x]][[1]], colour = "Alkuperäinen aineisto"), size = 1)) +
  lapply(seq_along(prediction_all_output), function(x) geom_line(aes(x = vert_maker(ref_list[[x]][[2]]), y = ref_list[[x]][[1]] + rnorm(length(ref_list[[x]][[1]]), 0.5, 0.001), colour = "Anonyymi aineisto"), size = 1)) +
  scale_colour_manual(name = "Aineisto", values = c("Alkuperäinen aineisto" = "black", "Anonyymi aineisto" = "red")) +
  labs(x = "Kvantiilifunktio", y = "Ennuste-etäisyys", title = "Ennuste-etäisyys", subtitle = paste0("Riittävä suoja: ", signum(1,3))) +
  theme_bw() +
  theme(legend.position = "bottom", legend.title = element_blank(), legend.text = element_text(size = 25), legend.key.size = unit(2, "cm"),
        legend.key = element_rect(fill = "transparent", colour = "transparent"), plot.title = element_text(size = 25, face = "bold"), axis.title = element_text(size = 24),
        axis.text = element_text(size = 24), plot.subtitle = element_text(size = 24))

print(p_distance)

# prediction_ambiguity
og_listt2 = lapply(seq_along(prediction_all_output), function(x) ecdf_points(prediction_all_output[[x]]$original$prediction_ambiguity))
ref_list2 = lapply(seq_along(prediction_all_output), function(x) ecdf_points(prediction_all_output[[x]]$reference$prediction_ambiguity))

erotus_list = lapply(seq_along(prediction_all_output), function(x) quantile_distance_0_1_calc(prediction_all_output[[x]]$original$prediction_ambiguity, prediction_all_output[[x]]$reference$prediction_ambiguity))
mean_erotus = mean(unlist(erotus_list))
format = formatC(signif(mean_erotus, digits=3), digits=3, format="fg", flag="#")

```

```
  # plot with ggplot2
  p_ambiguity = ggplot(data = NULL) +
    lapply(seq_along(prediction_all_output), function(x) geom_line(aes(x = vert_
        maker(og_listt2[[x]][[2]]), y = og_listt2[[x]][[1]], colour = "Alkuperäinen
         aineisto"), size = 1)) +
    lapply(seq_along(prediction_all_output), function(x) geom_line(aes(x = vert_
        maker(ref_list2[[x]][[2]]), y = ref_list2[[x]][[1]]+ rnorm(length(ref_list
        [[x]][[1]]), 0.3, 0.001), colour = "Anonyymi aineisto"), size = 1)) +
    scale_colour_manual(name = "Aineisto", values = c("Alkuperäinen aineisto" = "
        black", "Anonyymi aineisto" = "red")) +
    labs(x = "Kvantiilifunktio", y = "Ennuste-epäselvyys", title = "Ennuste-epä
        selvyys", subtitle = paste0("Riittävä suoja: ", signum(1,3))) +
    theme_bw() +
    theme(legend.position = "bottom", legend.title = element_blank(), legend.text =
         element_text(size = 25), legend.key.size = unit(2, "cm"),
          legend.key = element_rect(fill = "transparent", colour = "transparent"),
              plot.title = element_text(size = 25, face = "bold"), axis.title =
              element_text(size = 24),
          axis.text = element_text(size = 24), plot.subtitle = element_text(size =
              24))

  print(p_ambiguity)


  # prediction_uncertainty
  og_listt3 = lapply(seq_along(prediction_all_output), function(x) ecdf_points(
      prediction_all_output[[x]]$original$prediction_uncertainty))
  ref_list23 = lapply(seq_along(prediction_all_output), function(x) ecdf_points(
      prediction_all_output[[x]]$reference$prediction_uncertainty))

  erotus_list = lapply(seq_along(prediction_all_output), function(x) quantile_
      distance_0_1_calc(prediction_all_output[[x]]$original$prediction_uncertainty,
       prediction_all_output[[x]]$reference$prediction_uncertainty))
  mean_erotus = mean(unlist(erotus_list))
  format = formatC(signif(mean_erotus, digits=3), digits=3, format="fg", flag="#")

  # plot with ggplot2
  p_uncertainty = ggplot(data = NULL) +
    lapply(seq_along(prediction_all_output), function(x) geom_line(aes(x = vert_
        maker(og_listt3[[x]][[2]]), y = og_listt3[[x]][[1]], colour = "Alkuperäinen
         aineisto"), size = 1)) +
    lapply(seq_along(prediction_all_output), function(x) geom_line(aes(x = vert_
        maker(ref_list23[[x]][[2]]), y = ref_list23[[x]][[1]]+ rnorm(length(ref_
        list[[x]][[1]]), 0.5, 0.001), colour = "Anonyymi aineisto"), size = 1)) +
    scale_colour_manual(name = "Aineisto", values = c("Alkuperäinen aineisto" = "
        black", "Anonyymi aineisto" = "red")) +
    labs(x = "Kvantiilifunktio", y = "Ennuste-epävarmuus", title = "Ennuste-epä
        varmuus", subtitle = paste0("Riittävä suoja: ", signum(1,3))) +
    theme_bw() +
    theme(legend.position = "bottom", legend.title = element_blank(), legend.text =
         element_text(size = 25), legend.key.size = unit(2, "cm"),
          legend.key = element_rect(fill = "transparent", colour = "transparent"),
              plot.title = element_text(size = 25, face = "bold"), axis.title =
              element_text(size = 24),
          axis.text = element_text(size = 24), plot.subtitle = element_text(size =
              24))



  print(p_uncertainty)

}

swap = readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/swap.rds")

setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()
setwd("C:/Users/Johannes/Desktop/Gradu/Gradun kuvat/EsimDistances")


pdf("preddist.pdf", width = 12, height = 12)
prediction_plot_esim(swap[1], 5)
dev.off()

setwd("C:/Users/Johannes/Desktop/Gradu/anon")


setwd("C:/Users/Johannes/Desktop/Gradu/DATA/")
# read RDS
```

```
kanon <- readRDS("./lin_kanon/kanons.rds")
kanon_real <- readRDS("./lin_kanon_real/kanons_real.rds")
ldiv <- readRDS("./lin_ldiv/ldiv.rds")
ldiv_real = readRDS("./lin_ldiv_real/ldiv_real.rds")
noise <- readRDS("./lin_noise/noise.rds")
swap <- readRDS("./lin_swap/swap.rds")
rsa <- readRDS("./lin_rsa/rsa.rds")


model_list = list("kanon" = kanon,
                  "kanon_real" = kanon_real,
                  "ldiv" = ldiv,
                  "ldiv_real" = ldiv_real,
                  "noise" = noise,
                  "swap" = swap,
                  "rsa" = rsa)


# Calculate similarity
lapply(model_list, meansAll_list)

# Write RDS to disk
setwd("C:/Users/Johannes/Desktop/Gradu/DATA/")
saveRDS(model_list, "lin_data.rds")

setwd("C:/Users/Johannes/Desktop/Gradu/anon/")
```

```
setwd("C:/Users/Johannes/Desktop/Gradu/anon/")
devtools::load_all()

setwd("C:/Users/Johannes/Desktop/Gradu/DATA/")
# read RDS
kanon <- readRDS("./bin_kanon/kanons.rds")
kanon_real <- readRDS("./bin_kanon_real/kanons_real.rds")
ldiv <- readRDS("./bin_ldiv/ldiv.rds")
ldiv_real = readRDS("./bin_ldiv_real/ldiv_real.rds")
noise <- readRDS("./bin_noise/noise.rds")
swap <- readRDS("./bin_swap/swap.rds")
rsa <- readRDS("./bin_rsa/rsa.rds")


model_list = list("kanon" = kanon,
                  "kanon_real" = kanon_real,
                  "ldiv" = ldiv,
                  "ldiv_real" = ldiv_real,
                  "noise" = noise,
                  "swap" = swap,
                  "rsa" = rsa)


# Calculate similarity
lapply(model_list, meansAll_list)

# Write RDS to disk
setwd("C:/Users/Johannes/Desktop/Gradu/DATA/")
saveRDS(model_list, "bin_data.rds")

setwd("C:/Users/Johannes/Desktop/Gradu/anon/")
```

```
setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

yticks = c("Traffic3_min_hf",  "Sf12gr","GenderNainen","DepressionYes","Bmi","Age")
modelnames = c("Alkuperäinen", "Anonyymi")

lin_model_list <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/lin_model_list.rds"
    )

setwd("C:/Users/Johannes/Desktop/Gradu/Gradun kuvat/esims/confint")
pdf("confint_esim.pdf")
coeff_plot_list(lin_model_list[5], modelnames)+
  theme(legend.text = element_text(size = 12), axis.text.y = element_text(size =
      12), axis.text.x = element_text(size = 12), plot.title = element_text(size =
      16), plot.subtitle = element_text(size = 14)) +
  scale_x_discrete(labels = rev(yticks)) +
  labs(title = NULL)
dev.off()
```

```r
# Käytettävyys yksityisyys
ri = c(0,0.2, 1)
util = c(0,0.85, 1)

pdf("KäytettävyysYksityisyys.pdf")
plot(util, ri, xlab = "Käytettävyys", ylab = "Tietosuojarikkomuksen todennäköisyys"
     , main = "Yksityisyyden ja käytettävyyden vaihtokauppa", pch = 16, xlim = c
     (0,1), xaxt = "n", yaxt = "n", cex.main = 1.3, cex.lab = 1.3)
axis(1, at = c(0, 1), lab = c("Pieni", "Suuri"), cex.axis = 1.3)
axis(2, at = c(0, 1), lab = c("Pieni", "Suuri"), cex.axis = 1.3)
abline(h = 0.3, lty = 3)
text(0.3, 0.3, "Suurin sallittu riski", pos = 3, cex = 1.3)
text(util, ri, c("Aineistoa ei julkaista", "Anonymisoitu aineisto", "Alkuperäinen
     aineisto"), pos = c(4,2,2), cex = 1.3)
dev.off()
```

```r
setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

lin_model_list <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/lin_model_list.rds"
    )

yticks = c("Traffic3_min_hf",  "Sf12gr","GenderNainen","DepressionYes","Bmi","Age")
modelnames = c("Alkuperäinen", "5-anonyymi", "5-anonyymi pieni", "2-diversiteetti",
     "2-diversiteetti pieni", "Spekt. kohina", "Spekt. permutaatio")

setwd("C:/Users/Johannes/Desktop/Gradu/Gradun kuvat/lin/confint")
pdf("confint_lin.pdf")
coeff_plot_list(lin_model_list[-7], modelnames)+
  theme(legend.text = element_text(size = 12), axis.text.y = element_text(size =
      12), axis.text.x = element_text(size = 12), plot.title = element_text(size =
      16), plot.subtitle = element_text(size = 14)) +
  scale_x_discrete(labels = rev(yticks)) +
  labs(subtitle = "Päänsäryn vakavuuden pistemäärä")
dev.off()
```

```r
setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

bin_model_list <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/bin_model_list.rds"
    )

yticks = c("Headache_FU",  "Sf12gr","GenderNainen","DepressionYes","Bmi","Age")
modelnames = c("Alkuperäinen", "5-anonyymi", "5-anonyymi pieni", "2-diversiteetti",
     "2-diversiteetti pieni", "Spekt. kohina", "Spekt. permutaatio")

setwd("C:/Users/Johannes/Desktop/Gradu/Gradun kuvat/bin/confint")
pdf("confint_bin.pdf")
coeff_plot_list(bin_model_list[-7], modelnames)+
  theme(legend.text = element_text(size = 12), axis.text.y = element_text(size =
      12), axis.text.x = element_text(size = 12), plot.title = element_text(size =
      16), plot.subtitle = element_text(size = 14)) +
  scale_x_discrete(labels = rev(yticks)) +
  labs(subtitle = "Viikottainen päänsärky")
dev.off()
```

```r
lin_model_list <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/lin_model_list.rds"
    )

setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

og = lapply(seq_along(lin_model_list[[1]][[1]]), function(i) lin_model_list
    [[1]][[1]][[i]])
og_test = lapply(seq_along(lin_model_list[[1]][[3]]), function(i) lin_model_list
    [[1]][[3]][[i]])
og_R2 = testRsquared_list(og, og_test, 1)

kanon = lapply(seq_along(lin_model_list[[1]][[2]]), function(i) lin_model_list
    [[1]][[2]][[i]])
kanon_test = lapply(seq_along(lin_model_list[[1]][[3]]), function(i) lin_model_list
    [[1]][[3]][[i]])
kanon_R2 = testRsquared_list(kanon, kanon_test, 1)

kanon_real = lapply(seq_along(lin_model_list[[2]][[2]]), function(i) lin_model_list
    [[2]][[2]][[i]])
kanon_real_test = lapply(seq_along(lin_model_list[[2]][[3]]), function(i) lin_model
    _list[[2]][[3]][[i]])
```

```
kanon_real_R2 = testRsquared_list(kanon_real, kanon_real_test, 1)

ldiv = lapply(seq_along(lin_model_list[[3]][[2]]), function(i) lin_model_list
    [[3]][[2]][[i]])
ldiv_test = lapply(seq_along(lin_model_list[[3]][[3]]), function(i) lin_model_list
    [[3]][[3]][[i]])
ldiv_R2 = testRsquared_list(ldiv, ldiv_test, 1)

ldiv_real = lapply(seq_along(lin_model_list[[4]][[2]]), function(i) lin_model_list
    [[4]][[2]][[i]])
ldiv_real_test = lapply(seq_along(lin_model_list[[4]][[3]]), function(i) lin_model_
    list[[4]][[3]][[i]])
ldiv_real_R2 = testRsquared_list(ldiv_real, ldiv_real_test, 1)

noise = lapply(seq_along(lin_model_list[[5]][[2]]), function(i) lin_model_list
    [[5]][[2]][[i]])
noise_test = lapply(seq_along(lin_model_list[[5]][[3]]), function(i) lin_model_list
    [[5]][[3]][[i]])
noise_R2 = testRsquared_list(noise, noise_test, 1)

swap = lapply(seq_along(lin_model_list[[6]][[2]]), function(i) lin_model_list
    [[6]][[2]][[i]])
swap_test = lapply(seq_along(lin_model_list[[6]][[3]]), function(i) lin_model_list
    [[6]][[3]][[i]])
swap_R2 = testRsquared_list(swap, swap_test, 1)

rsa = lapply(seq_along(lin_model_list[[7]][[2]]), function(i) lin_model_list
    [[7]][[2]][[i]])
rsa_test = lapply(seq_along(lin_model_list[[7]][[3]]), function(i) lin_model_list
    [[7]][[3]][[i]])
rsa_R2 = testRsquared_list(rsa, rsa_test, 1)

print(c(og_R2, kanon_R2, kanon_real_R2, ldiv_R2, ldiv_real_R2, noise_R2, swap_R2,
    rsa_R2))
```

```
bin_model_list = readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/bin_model_list.rds")

setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

setwd("C:/Users/Johannes/Desktop/Gradu/Gradun kuvat/Bin/roc")

pdf("roc_kanon.pdf")
roc_plot_list(1, bin_model_list[[1]][[1]], bin_model_list[[1]][[2]], bin_model_list
    [[1]][[3]])
dev.off()

pdf("roc_kanon_real.pdf")
roc_plot_list(1, bin_model_list[[2]][[1]], bin_model_list[[2]][[2]], bin_model_list
    [[2]][[3]])
dev.off()

pdf("roc_ldiv.pdf")
roc_plot_list(1, bin_model_list[[3]][[1]], bin_model_list[[3]][[2]], bin_model_list
    [[3]][[3]])
dev.off()

pdf("roc_ldiv_real.pdf")
roc_plot_list(1, bin_model_list[[4]][[1]], bin_model_list[[4]][[2]], bin_model_list
    [[4]][[3]])
dev.off()

pdf("roc_noise.pdf")
roc_plot_list(1, bin_model_list[[5]][[1]], bin_model_list[[5]][[2]], bin_model_list
    [[5]][[3]])
dev.off()

pdf("roc_swap.pdf")
roc_plot_list(1, bin_model_list[[6]][[1]], bin_model_list[[6]][[2]], bin_model_list
    [[6]][[3]])
dev.off()

test_list = lapply(seq_along(bin_model_list[[7]][[1]]), function(i) bin_model_list
    [[7]][[1]][[i]][["data"]])

pdf("roc_rsa.pdf")
roc_plot_list(1, bin_model_list[[7]][[1]], bin_model_list[[7]][[2]], test_list, bin
    _model_list[[7]][[3]])
```

```
dev.off()
```

```
setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

lin_data <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/lin_data.rds")

# Write a list object for each distances object on each anoymization method to a
    file

kanon = prediction_all_list(lin_data[[1]], 5)
saveRDS(kanon, "C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/kanon.rds")

kanon_real = prediction_all_list(lin_data[[2]], 5)
saveRDS(kanon_real, "C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/kanon_real.
    rds")

ldiv = prediction_all_list(lin_data[[3]], 5)
saveRDS(ldiv, "C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/ldiv.rds")

ldiv_real = prediction_all_list(lin_data[[4]], 5)
saveRDS(ldiv_real, "C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/ldiv_real.
    rds")

noise = prediction_all_list(lin_data[[5]][1:100], 5)
saveRDS(noise, "C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/noise.rds")

swap = prediction_all_list(lin_data[[6]][1:100], 5)
saveRDS(swap, "C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/swap.rds")

rsa = prediction_all_list(lin_data[[7]][1:100], 5)
saveRDS(rsa, "C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/rsa.rds")

# load all the list objects and write them into a single file
kanon = readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/kanon.rds")
kanon_real = readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/kanon_real
    .rds")
ldiv = readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/ldiv.rds")
ldiv_real = readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/ldiv_real.
    rds")
noise = readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/noise.rds")
swap = readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/swap.rds")
rsa = readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/rsa.rds")

lin_distances = list(kanon, kanon_real, ldiv, ldiv_real, noise, swap, rsa)
saveRDS(lin_distances, "C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/lin_
    distances.rds")
```

```
setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

bin_data <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/bin_data.rds")

# Write a list object for each distances object on each anoymization method to a
    file

kanon = prediction_all_list(bin_data[[1]], 5)
saveRDS(kanon, "C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/kanon.rds")

kanon_real = prediction_all_list(bin_data[[2]], 5)
saveRDS(kanon_real, "C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/kanon_real.
    rds")

ldiv = prediction_all_list(bin_data[[3]], 5)
saveRDS(ldiv, "C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/ldiv.rds")

ldiv_real = prediction_all_list(bin_data[[4]], 5)
saveRDS(ldiv_real, "C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/ldiv_real.
    rds")

noise = prediction_all_list(bin_data[[5]][1:100], 5)
saveRDS(noise, "C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/noise.rds")

swap = prediction_all_list(bin_data[[6]][1:100], 5)
saveRDS(swap, "C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/swap.rds")

rsa = prediction_all_list(bin_data[[7]][1:100], 5)
saveRDS(rsa, "C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/rsa.rds")
```

```
28
29 # load all the list objects and write them into a single file
30 kanon = readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/kanon.rds")
31 kanon_real = readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/kanon_real
       .rds")
32 ldiv = readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/ldiv.rds")
33 ldiv_real = readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/ldiv_real.
       rds")
34 noise = readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/noise.rds")
35 swap = readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/swap.rds")
36 rsa = readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/rsa.rds")
37
38 bin_distances = list(kanon, kanon_real, ldiv, ldiv_real, noise, swap, rsa)
39 saveRDS(bin_distances, "C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/bin_
       distances.rds")
```

```
1 swap <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/swap.rds")
2 rsa <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/rsa.rds")
3 noise <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/noise.rds")
4 ldiv_real <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/ldiv_real.
      rds")
5 ldiv <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/ldiv.rds")
6 kanon_real <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/kanon_
      real.rds")
7 kanon <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/lin_distances/kanon.rds")
8
9 setwd("C:/Users/Johannes/Desktop/Gradu/anon")
10 devtools::load_all()
11
12 setwd("C:/Users/Johannes/Desktop/Gradu/Gradun kuvat/lin/distances")
13
14 # Plotting the distances
15 pdf("lin_kanon.pdf", width = 12, height = 12)
16 prediction_plot_list(kanon, 5)
17 dev.off()
18
19 pdf("lin_kanon_real.pdf", width = 12, height = 12)
20 prediction_plot_list(kanon_real, 5)
21 dev.off()
22
23 pdf("lin_ldiv.pdf", width = 12, height = 12)
24 prediction_plot_list(ldiv, 5)
25 dev.off()
26
27 pdf("lin_ldiv_real.pdf", width = 12, height = 12)
28 prediction_plot_list(ldiv_real, 5)
29 dev.off()
30
31 pdf("lin_noise.pdf", width = 12, height = 12)
32 prediction_plot_list(noise, 5)
33 dev.off()
34
35 pdf("lin_swap.pdf", width = 12, height = 12)
36 prediction_plot_list(swap, 5)
37 dev.off()
```

```
1 swap <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/swap.rds")
2 rsa <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/rsa.rds")
3 noise <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/noise.rds")
4 ldiv_real <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/ldiv_real.
      rds")
5 ldiv <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/ldiv.rds")
6 kanon_real <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/kanon_
      real.rds")
7 kanon <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/bin_distances/kanon.rds")
8
9 setwd("C:/Users/Johannes/Desktop/Gradu/anon")
10 devtools::load_all()
11
12 setwd("C:/Users/Johannes/Desktop/Gradu/Gradun kuvat/Bin/distances")
13
14 # Plotting the distances
15 pdf("bin_kanon.pdf", width = 12, height = 12)
16 prediction_plot_list(kanon, 5)
17 dev.off()
18
19 pdf("bin_kanon_real.pdf", width = 12, height = 12)
20 prediction_plot_list(kanon_real, 5)
```

```
dev.off()

pdf("bin_ldiv.pdf", width = 12, height = 12)
prediction_plot_list(ldiv, 5)
dev.off()

pdf("bin_ldiv_real.pdf", width = 12, height = 12)
prediction_plot_list(ldiv_real, 5)
dev.off()

pdf("bin_noise.pdf", width = 12, height = 12)
prediction_plot_list(noise, 5)
dev.off()

pdf("bin_swap.pdf", width = 12, height = 12)
prediction_plot_list(swap, 5)
dev.off()
```

```
setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

lin_data <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/lin_data.rds")

# Calculate ri_rate for all the lin_data objects
kanon = reidentification_rate_list(lin_data[[1]], c("Age", "Gender", "Bmi"))
saveRDS(kanon, "C:/Users/Johannes/Desktop/Gradu/DATA/lin_ri/kanon.rds")

kanon_real = reidentification_rate_list(lin_data[[2]], c("Age", "Gender", "Bmi"))
saveRDS(kanon_real, "C:/Users/Johannes/Desktop/Gradu/DATA/lin_ri/kanon_real.rds")

ldiv = reidentification_rate_list(lin_data[[3]], c("Age", "Gender", "Bmi"))
saveRDS(ldiv, "C:/Users/Johannes/Desktop/Gradu/DATA/lin_ri/ldiv.rds")

ldiv_real = reidentification_rate_list(lin_data[[4]], c("Age", "Gender", "Bmi"))
saveRDS(ldiv_real, "C:/Users/Johannes/Desktop/Gradu/DATA/lin_ri/ldiv_real.rds")

noise = reidentification_rate_list(lin_data[[5]][1:100], c("Age", "Gender", "Bmi"))
saveRDS(noise, "C:/Users/Johannes/Desktop/Gradu/DATA/lin_ri/noise.rds")

swap = reidentification_rate_list(lin_data[[6]][1:100], c("Age", "Gender", "Bmi"))
saveRDS(swap, "C:/Users/Johannes/Desktop/Gradu/DATA/lin_ri/swap.rds")

sapply(c(kanon, kanon_real, ldiv, ldiv_real, noise, swap), signum, 3)
```

```
setwd("C:/Users/Johannes/Desktop/Gradu/anon")
devtools::load_all()

bin_data <- readRDS("C:/Users/Johannes/Desktop/Gradu/DATA/bin_data.rds")

# Calculate ri_rate for all the bin_data objects
kanon = reidentification_rate_list(bin_data[[1]], c("Age", "Gender", "Bmi"))
saveRDS(kanon, "C:/Users/Johannes/Desktop/Gradu/DATA/bin_ri/kanon.rds")

kanon_real = reidentification_rate_list(bin_data[[2]], c("Age", "Gender", "Bmi"))
saveRDS(kanon_real, "C:/Users/Johannes/Desktop/Gradu/DATA/bin_ri/kanon_real.rds")

ldiv = reidentification_rate_list(bin_data[[3]], c("Age", "Gender", "Bmi"))
saveRDS(ldiv, "C:/Users/Johannes/Desktop/Gradu/DATA/bin_ri/ldiv.rds")

ldiv_real = reidentification_rate_list(bin_data[[4]], c("Age", "Gender", "Bmi"))
saveRDS(ldiv_real, "C:/Users/Johannes/Desktop/Gradu/DATA/bin_ri/ldiv_real.rds")

noise = reidentification_rate_list(bin_data[[5]][1:100], c("Age", "Gender", "Bmi"))
saveRDS(noise, "C:/Users/Johannes/Desktop/Gradu/DATA/bin_ri/noise.rds")

swap = reidentification_rate_list(bin_data[[6]][1:100], c("Age", "Gender", "Bmi"))
saveRDS(swap, "C:/Users/Johannes/Desktop/Gradu/DATA/bin_ri/swap.rds")

sapply(c(kanon, kanon_real, ldiv, ldiv_real, noise, swap), signum, 3)
```