

## Tietosuoja ja tulosten tarkastusmenettely

Tässä ohjeessa kerrotaan tutkimuskäytössä olevan aineiston ja erityisesti siitä tuotettujen tulosteiden (taulukoiden, kuvioiden, tilastollisten mallien yms.) tietosuojasta sekä tulosteiden tarkastusmenettelystä.

Tietosuojaohjeet koskevat aineistojen käyttöä niin tutkimushankkeissa kuin mikrosimuloinnissakin. Tulosteiden tarkastusmenettely mikrosimuloinnissa kuitenkin poikkeaa muista tutkimushankkeista.

Mikäli tietosuoja-asiat herättävät kysymyksiä, tutkija voi ottaa yhteyttä Tutkijapalveluihin (tutkijapalvelut@stat.fi / mikrosimulointi@stat.fi).

### Tutkimusaineistojen ja tulosteiden tietosuoja

Tutkimusaineistojen tietosuojasta huolehtiminen on sekä Tilastokeskuksen että tutkijoiden tehtävä. Tilastokeskus huolehtii osaltaan aineistojen tietosuojasta ennen niiden luovutusta tutkimuskäyttöön sekä etäkäyttöympäristön tietoturvallisuudesta. Tutkijan on osaltaan huolehdittava tietosuojasta aineistonsa tutkimuskäytön ajan sekä julkaistessaan tutkimustulosteita.

Tutkija on vastuussa tietosuojan toteutumisesta julkaisemissaan tutkimustuloksissa. Nämä ohjeet on laadittu tukemaan vastuullista toimintaa tietosuoja-asioissa, ja ohjeilla pyritään estämään niin tahattomat kuin tahallisetkin tietosuarikkomukset. Tutkijapalveluilla on käytössä tutkimustulosten tarkastusmenettely, jolla myös tuetaan huolellisuutta tietosuoja-asioissa. Tarkastusmenettelystä kerrotaan tarkemmin osiossa Tulosteiden tarkastusmenettely.

### Miksi tietosuoja?

Tutkimusaineistojen tietosuojasta huolehtiminen on edellytys sille, että Tilastokeskus voi luovuttaa tilastotarkoituksiin keräämiään aineistoja tutkimuskäyttöön ja mikrosimulointiin. Tilastokeskuksen oikeus kerätä rekisteri- ja kyselytutkimusaineistoja tilastointia varten on taattu tilastolailla, kuten myös oikeus luovuttaa näitä tietoja tutkimuskäyttöön. Toisaalta laki velvoittaa myös huolehtimaan tietojen asiallisesta suojaamisesta, ja osa aineistoista sisältääkin hyvin arkaluonteisia tietoja.

Henkilötietojen käsittelyä ohjaa Euroopan unionin yleinen tietosuoja-asetus (EU) 2016/679 ja sitä täydentävä kansallinen tietosuojalaki (1050/2018). Tutkimushankkeen saamaan käyttölupaun liittyvän salassapitovelvoitteen mukaan tutkijan on huolehdittava siitä, että tutkimustuloksissa ei ole yksikötason tietoja, eli yksittäistä henkilöä tai yritystä koskevia tietoja tai mahdollisuutta niiden paljastumiseen. Käsitellessään tutkimuskäyttöön luovutettua aineistoa tai toimiessaan etäkäyttöympäristössä tutkijan on salassapitovelvoitteen mukaan huolehdittava myös siitä, ettei aineistoa paljasteta tai luovuteta taholle, jolla ei ole siihen käyttöoikeutta.

## Eri tulostetyyppien tietosuojaohjeet

Alla on esitetty ohjeita ja sääntöjä erilaisten tulostetyyppien tietosuojasta huolehtimiseen. Koska tutkimushankkeita ja erilaisia tulostetyyppejä on paljon erilaisia, ei jokaisen aineiston ja tulosteen tietosuojaaja voida arvioida ja ohjeistaa erikseen. Tämän takia yleisten ”nyrkkisääntöjen” antaminen on välttämätöntä.

Joidenkin Tilastokeskuksen aineistojen ja muiden viranomaisten tutkimuskäyttöön luovuttamien aineistojen osalta tietosuojaohjeet saattavat poiketa alla esitetyistä säännöistä. Nämä poikkeavat säännöt löytyvät Taika-katalogissa olevista aineistokuvauksista, ja tarvittaessa ne kirjataan myös käyttöluopapäätökseen.

Mikäli tutkija on jonkin tulosteen tietosuojasta epävarma tai haluaa pyytää tarkennusta annettuihin ohjeisiin, tulee hänen ottaa yhteyttä Tutkijapalveluihin ([tutkijapalvelut@stat.fi](mailto:tutkijapalvelut@stat.fi)).

## Otosaineistoista tuotetut tulosteet

Kaikille tulostetyypeille on yhteistä, että tulosten taustalla olevien havaintojen paljastumisriski on yleensä pienempi, jos käytetty aineisto on vain (satunnais) otos kokonaisaineistosta. Otannasta mahdollisesti johtuva pienempi paljastumisriski ei kuitenkaan tarkoita, että otosaineistoista tuotettuja tulosteita varten voitaisiin automaattisesti lieventää tässä ohjeessa esitettyjä suojaussääntöjä.

Jos havainnon paljastumisriskiä joudutaan kuitenkin arvioimaan jossain tulosteessa tapauskohtaisesti (esim. onko jakaumatunnusluku paljastava tai voiko kuvasta yksittäisen kuvapisteen taustalla olevaa havaintoa tunnistaa), niin tällöin otannan käyttö vaikuttaa arviointiin.

On huomioitava, että kokonaisaineisto ei aina tarkoita vain kaikkia Suomessa asuvia henkilöitä tai kaikkia Suomessa toimivia yrityksiä. Tietosuojan näkökulmasta ja havaintojen paljastumisriskiä arvioitaessa esim. tietyn toimialan kaikki yritykset muodostavat ennemminkin kokonaisaineston kuin otosaineiston, sillä yleensä on pääteltävissä, kuuluuko yksittäinen yritys ”otokseen”, joka on rajattu kattamaan tietyn toimialan kaikki yritykset.

## Frekvenssi- ja määrätaulukot

Taulukkomuotoon aggregoidut tiedot voivat sisältää henkilötietoja, yritystietoja tai molempia. Esimerkiksi työntekijä-työnantaja-aineistoissa on suojattava sekä henkilö- että yritystaso. Myös ammattiryhmiä kuvaavat tiedot saattavat epäsuorasti sisältää yritystietoja, jos (lähes) kaikki tiettyyn ammattiryhmään kuuluvat henkilöt ovat vain yhden (monopoli)yrityksen palveluksessa, jolloin kyseisen yrityksen tietosuojasta on huolehdittava.

Sekä yritys- että henkilötietoja sisältävien taulukoiden suojauksessa pääsääntönä on kynnysarvo 3. Toisin sanoen taulukon solun tai havaintoryhmän tiedot saa julkaista vain, jos tietojen taustalla on vähintään 3 (painottamatonta) havaintoa. Myöskään havaintojen lukumäärätietoa ei pienten solujen tai ryhmien osalta saa julkaista.

Kynnysarvosäännön käyttö pääsääntönä on helpoin tapa laskea julkaistaviin taulukoihin mahdollisesti kohdistuvaa yksittäisten havaintojen paljastumisriskiä. Havainnon tunnistaminen ryhmänsä ainoaksi tai harvinaiseksi tapaukseksi voi lisätä riskiä tunnistaa havainto ja/tai paljastaa lisää tietoja tästä havainnosta (esim. muiden samaa aineistoa tai aihetta käsittelevien taulukoiden ja tulosteiden avulla). Tämän takia kynnysarvoa tulee soveltaa, vaikka taulukossa julkaistaisiinkin pelkkiä lukumäärätietoja.

Edellä mainitun pääsäännön lisäksi on huomioitava seuraavat säännöt:

- Tuoreissa yritystiedoissa (alle 15 kuukautta viiteajankohdasta) on kynnysarvon lisäksi käytettävä dominanssisääntöä<sup>1</sup> (1,75). Dominanssisääntöä voidaan edellyttää käytettäväksi myös vanhemmissa yritystiedoissa tai muissa aineistoissa (esim. tulorekisteri). Dominanssisääntö edellyttää suojattavaksi esimerkiksi tiedot, joissa yhden yrityksen liikevaihto tai palkkasumma tai jokin muu aineiston muuttuja muodostaa yli 75 prosenttia tietyn toimialan (sektorin, alueen tms.) yritysten liikevaihdosta/palkkasummasta tms. tai jos esim. tietyn ammattiryhmän palkansaajista yli 75 prosenttia työskentelee jossain tietyssä yrityksessä.
- Toimipaikkatason tietoja suojatessa on mahdollisuuksien mukaan varmistettava myös yritystason suojaus, eli jokaisessa solussa on oltava toimipaikkoja vähintään kolmesta eri yrityksestä. Samoin on toimittava konserni-yritys-suhteiden kanssa.
- Hyödyketiedoissa (teollisuuden tuotantotilaston tuotteet sekä aineet ja tarvikkeet) yritysten lukumäärä on luottamuksellinen tieto kaikissa tuotantonimikkeissä.
- Henkilötietojen osalta tietosuojasta on huolehdittava erityisen tarkasti, jos taulukko sisältää arkaluonteisia henkilötietoja<sup>2</sup> (ml. EU:n yleisessä tietosuojasetuksessa<sup>3</sup> luetellut erityiset henkilötietoryhmät). Arkaluonteisten tietojen suojaamisessa saattaa olla tarpeellista käyttää suurempaa kynnysarvoa.
- Yritystiedoissa esiintyviin ammatinharjoittajiin sovelletaan samoja suojaussääntöjä, kuin muihinkin yritystietoihin, vaikka ammatinharjoittajat lähtökohtaisesti ovatkin henkilöitä.

Koordinaatti- tai ruututietojen julkaisuun liittyy erityisehtoja: Kokonaista ruutuaineistoa ei saa julkaista. Ruutuihin perustuvia tietoja voidaan ottaa ulos etäkäyttöjärjestelmästä ja julkaista silloin, kun ruudussa on vähintään 10 havaintoyksikköä. Julkaistavien tietojen tulee olla aggregoituja suuremmalle

<sup>1</sup> Dominanssisäännön ( $n,k$ ) mukaan taulukon soluarvoa ei voi julkaista, mikäli sen  $n$  suurinta havaintoa muodostavat vähintään  $k$  prosenttia solun kokonaisarvosta.

<sup>2</sup> Arkaluonteisia tietoja tässä ovat tiedot, jotka kuvaavat henkilön rotua tai etnistä alkuperää, poliittisia mielipiteitä, uskonnollista tai filosofista vakaumusta, ammattiliiton jäsenyyttä, terveyttä, seksuaalista käyttäytymistä tai suuntautumista, rikostuomioita ja rikkomuksia ja niihin liittyviä turvaamistoimia, kuolinsyytä, kieltä, kansalaisuutta, syntyperää tai synnyinmaata, tuloja, velkoja, varallisuutta, harvinaista ammattia tai muuta sosioekonomista asemaa, sosiaalihuollon tarvetta tai saatuja sosiaalihuollon palveluita, sosiaalihuollon tukitoimia tai muita etuuksia, tai tiedot henkilöön kohdistetuista hoitotoimenpiteistä tai niihin verrattavista toimista.

<sup>3</sup> Euroopan parlamentin ja neuvoston asetus (EU) 2016/679, 9 artikla

aluetasolle, suhtautettuja tai muulla tavalla käsiteltyjä, jotta henkilöitä ja asutokuntia ei voi tunnistaa.

## Erilaiset jakaumatunnusluvut

Minimi ja maksimi liittyvät yleensä yhteen havaintoon, joten niitä ei useimmiten voi julkaista. Esimerkiksi toimialan suurin yritys on yleensä mahdollista tunnistaa, joten siihen liittyvää liikevaihtotiedon maksimia ei voi julkaista. Mikäli minimiä tai maksimia ei voida yhdistää yksittäiseen havaintoon ja kynnysarvo toteutuu, ne voidaan julkaista.

Jakaumapisteet (pl. minimi ja maksimi), kuten esimerkiksi desiilit, muodostavat erikoistapauksen taulukosta, jossa solufrekvenssejä vastaavat jakaumapisteiden väliin jäävien havaintojen lukumäärät. Mikäli nämä lukumäärät ylittävät taulukoissa sovellettavan kynnysarvon 3, voidaan jakaumapisteet julkaista.

Moodi voidaan julkaista, mikäli (lähes) kaikki havainnot eivät saa samaa arvoa.

Keskiarvo, muut suhdeluvut ja jakaumatunnuslukujen korkeammat momentit (esim. varianssi) voidaan julkaista, mikäli niiden laskennassa on käytetty vähintään kolmea havaintoa.

Osuuksia julkaistaessa on kynnysarvon 3 toteuduttava kaikkien osuuksia muodostavien ryhmien osalta. Toisin sanoen, jos halutaan julkaista esim. naisten osuuden olevan 58 % koko populaatiosta, niin tuon 58 %, samoin kuin miesten 42 %, on sisällettävä vähintään kolme henkilöä. Ei siis riitä, että naisia ja miehiä on yhteensä koko populaatiossa vähintään 3.

## Muut numeeriset tulostetyypit

Indeksipisteluvut, korrelaatiokertoimet ja testisuureet (t, F,  $X^2$ , yms.) voidaan yleensä julkaista, mikäli niiden laskennassa on käytetty vähintään 10 havaintoa.

Regressiomallin voi kokonaisuudessaan julkaista, mikäli mallin taustalla on riittävästi havaintoja (vähintään 10) ja malli ei kuvaa aikasarjaa yhteen yritykseen tai henkilöön perustuvista havainnoista. Mallin yksittäisiä kertoimia voidaan yleensä aina julkaista.

Monimutkaisempien tilastollisten mallien numeeriset tulokset voidaan yleensä aina julkaista, mikäli mallin taustalla on riittävästi havaintoja ja malli ei kuvaa aikasarjaa yksittäiseen yritykseen tai henkilöön perustuvista havainnoista.

## Kuvat

Numeeristen tulosteiden tapaan myöskään kuvat eivät saa paljastaa yksittäisen havainnon tietoa. Aineistoista piirretyt kuvat ovat sallittuja, jos yksittäinen kuvapiste tai kuvan osa ei voi paljastaa sen taustalla olevaa yksittäistä havaintoa.

Pylväsdiagrammit ja muut luokitellun aineiston esittämiseen käytetyt kuvat ovat tyypillisesti sallittuja julkaistaviksi, kunhan kussakin luokassa on riittävästi

havaintoja. Tällaisen kuvan informaatio voidaan yleensä esittää myös taulukkomuodossa ja siihen voidaan soveltaa samoja tietosuojasääntöjä kuin muihinkin taulukkoaineistoihin (ks. yllä kohta Frekvenssi- ja määrätaulukot).

Jakaumakuvista tasoitettut tai riittävän karkealla asteikolla esitetyt jakaumat, histogrammit ja kertymäfunktioit ovat sallittuja. Osa jakaumakuvista voi sisältää tietoja poikkeavista havainnoista tai ääriarvoista, jotka tapauskohtaisesti voivat paljastaa yksittäisen havainnon tietoja. Ohjelmien piirtofunktioit merkitsevät usein automaattisesti mm. laatikkokuvaajiin (box plot) poikkeavat havainnot. Poikkeavat havainnot yksilöivät kuvat eivät sovi julkaistaviksi, ellei tutkija pysty hyvin perustelemaan, etteivät kuvaan merkityt poikkeavat havainnot ole tunnistettavissa.

Hajontakuvia käytetään tyypillisesti kahden jatkuvan muuttujan arvojen esittämiseen. Hajontakuvien pisteet kuvaavat lähtökohtaisesti kukin yksittäisen havainnon tietoja, minkä vuoksi se on tietosuojan kannalta edellisiä kuvatyyppejä hankalampi tapaus. Hajontakuvien tietosuojan arvioinnissa tutkijan tulee kiinnittää erityistä huomiota aineiston luonteeseen mm. otoksen koon, tiedon arkaluonteisuuden ja poikkeavien havaintojen esiintymisen kannalta. Hajontakuva ei täytä tietosuojavaatimuksia, jos siitä on suoraan nähtävissä tai helposti pääteltävissä esim. yksittäisen toimialan suurimman yrityksen tietoja.

## Suojausmenetelmät

Etäkäyttöjärjestelmästä ulosotettavista tiedostoista tai taulukoista ei saa olla mahdollista tunnistaa yksittäistä henkilöä tai yritystä koskevia tietoja. Paljastumisriskin sisältävät tiedot on suojattava suunnittelemalla tulosteiden sisältö tietosuojan kannalta hyväksyttäväksi esimerkiksi tarpeeksi karkeita luokituksia käyttämällä.

Taulukossa paljastumisriskin sisältävät solut voidaan suojata taulukon rakennetta muuttamalla, yksittäisiä soluarvoja tai kokonaisia rivejä peittämällä tai muuttamalla soluarvoja esimerkiksi pyöristämällä tai korvaamalla alkuperäinen soluarvo likimääräisellä satunnaisluvulla. Taulukon suojausmenetelmän valinnassa on syytä pyrkiä löytämään menetelmä, joka suojaa taulukkoa riittävästi, mutta säilyttää sen käyttötarkoituksen kannalta tärkeät ominaisuudet mahdollisimman hyvin.

Taulukon rakenteen muuttaminen tarkoittaa muuttujien määrän kontrollointia tai luokituksen muuttamista. Luokitusta muuttamalla taulukosta pyritään hävittämään paljastumisriskissä olevat solut yhdistämällä niitä sisältävät luokat muihin taulukon luokkiin. Luokituksen muuttaminen tarkoittaa usein käytännössä koko luokituksen karkeistamista.

Peittämiseen kuuluu ensisijainen, paljastumisriskissä olevien solujen peittäminen ja toissijainen peittäminen. Toissijaisella peittämisellä varmistetaan, ettei taulukon rivi- tai saraketotaalien avulla pystytä paljastamaan ensisijaisesti peitettyjen solujen arvoja. Peittäminen voidaan tehdä myös rivikohtaisesti. Jos taulukon johonkin rivitotaaliin kuuluu vain pieni määrä tilastoyksiköitä (vähemmän kuin käytetty kynnysarvo), peitetään kyseinen rivi kokonaisuudessaan huomioimatta sen eri soluissa olevien tilastoyksiköiden lukumäärää.

Lisätietoja tietosuojamenetelmistä löytyy esimerkiksi lopussa mainitun Tutkimusaineistot etäkäytössä -verkkokurssin materiaaleista.

## Tulosteiden tarkastusmenettely

Tutkijapalveluilla on etäkäytössä olevien aineistojen osalta käytössä tutkimustulosten tarkastusmenettely. Jos tutkija on epävarma tulosteen tietosuojasta, kannattaa hänen olla yhteydessä Tutkijapalveluihin jo ennen tulosteen tarkastukseen tai järjestelmästä ulos vientiä.

Tarkastusmenettelyn rooli on tukea tutkijan vastuullista toimintaa tutkimustuloksia koskevissa tietosuoja-asioissa. Tarkastusmenettelystä huolimatta tutkijalla on vastuu tutkimustulostensa tietosuojasta. Tarkastusmenettely antaa Tutkijapalveluille mahdollisuuden seurata tietosuojan toteutumista tutkimusaineistojen tuloksissa sekä huomata tarpeet tarjota lisäopastusta tietosuoja-asioissa.

Tarkastusmenettely toimii käytännössä eri tavalla tutkimushankkeiden etäkäytössä ja mikrosimulointimallin etäkäytössä. Molempien osalta tutkijan on kuitenkin huolehdittava, etteivät etäkäyttöjärjestelmästä ulos siirrettävät (tai siirrettäväksi toivottavat) tulosteet sisällä yksikkötason aineistoa tai mahdollisuutta yksittäistä havaintoa koskevien tietojen paljastamiseen.

Etäkäyttöjärjestelmästä tarkastukseen tai ulos siirrettäessä tulee noudattaa harkintaa. Järjestelmästä tulee siirtää tarkastukseen tai ulos vain tulosteita, jotka on tarkoitus julkaista. Toisin sanoen ns. välituloksien ja varsinkin (suurien) log-tyyppisten tiedostojen siirtoa on vältettävä. Tarkastukseen vietävien taulukoiden ja kuvioden tulisi olla sisällöltään siinä muodossa, missä ne on tarkoitus julkaista. Järjestelmästä ei voi saada ulos tulosteita, joita ei tietosuojan takia voitaisi julkaista.

Tulosteet on dokumentoitava huolellisesti, jotta tarkastajalle on selvää tulosteen tietosisältö. Tulosteissa on oltava näkyvissä taulukoiden, kuvien, tunnuslukujen jne. laskemisessa käytettyjen havaintojen lukumäärät. Jos tulostarkastuksesta halutaan ulos tietosuojaohjeista poikkeavia tietoja, tulee tietosuojan toteutuminen tulosteissa perustella hyvin.

Tulosteiden, erityisesti kuvien tiedostomuodon on myös oltava sellainen, ettei se aiheuta riskiä yksikkötason tietojen paljastumiseen. Tarkastettavaksi soveltuvia kuvaformaatteja ovat esimerkiksi:

- Bittikarttaformaatit
- PNG (Portable Networks Graphics)
- BMP (Bitmap)
- JPEG (Joint Photographic Experts Group)
- TIFF (Tagged Image File Format)
- Vektoriformaatit
- EPS (Encapsulated PostScript)
- PS (PostScript)
- PDF (Portable Document Format)
- SVG (Scalable Vector Graphics)

– WMF/EMF (Windows Metafile)

Stata-ohjelmassa yllä mainittuja kuvaformaatteja pystyy luomaan graph export -komennolla. SPSS-ohjelmassa kuvaformaatin saa valita Export output -toiminnossa. R-ohjelmassa tietoa piirtofunktioista saa komennolla help(grDevices). Tietyt kuvatyypit, kuten Statan gph-tiedostot, tallentavat lähtökohtaisesti kuvan piirtämiseen käytetyn aineiston, minkä vuoksi ne eivät välttämättä sovellu tarkastukseen ja ulos siirrettäviksi.

### Tutkimusaineistot etäkäytössä -verkkokurssi

Tilastokeskuksen internetsivuilla on julkaistu Tutkimusaineistot etäkäytössä -verkkokurssi osana Tilastokoulua

([https://tilastokoulu.stat.fi/verkkokoulu\\_v2.xql?page\\_type=ketusivu&course\\_id=tkoulu\\_tutki](https://tilastokoulu.stat.fi/verkkokoulu_v2.xql?page_type=ketusivu&course_id=tkoulu_tutki)).

Verkkokurssi antaa lisätietoja tutkimusaineistojen ja SISU-mikrosimulointimallin etäkäytöstä ja aineistojen tietosuojasta. Kurssi sisältää myös esimerkkejä taulukkoaineistojen suojauksesta.

Varsinkin etäkäyttöjärjestelmää ensimmäistä kertaa käyttävien tutkijoiden on suositeltavaa tutustua näiden ohjeiden ohella myös Tutkimusaineistot etäkäytössä -verkkokurssin materiaaleihin.