18.3.2024

Dataskydd vid distansanvändning och kontrollförfarandet för resultat

Denna anvisning redogör för dataskyddet för forskningsmaterial som används på distans och i synnerhet för utskrifter som producerats utifrån det (tabeller, diagram, statistiska modeller o.d.) och kontrollförfarandet för utskrifter.

Dataskyddsanvisningarna gäller för användning av material såväl i forskningsprojekt som i mikrosimuleringar. Kontrollförfarandet för utskrifter vid mikrosimuleringar avviker dock från förfarandet vid övriga forskningsprojekt.

Om dataskyddet väcker frågor, kan du som forskare kontakta Forskartjänsterna (e-post: tutkijapalvelut@stat.fi/mikrosimulointi@stat.fi).

Dataskydd för forskningsmaterial och utskrifter

Såväl Statistikcentralen som forskarna har till uppgift att sköta dataskyddet för forskningsmaterial. Statistikcentralen sköter för egen del om materialets dataskydd innan det överlåts för forskningsanvändning och för datasäkerheten i distansanvändningsmiljön. Forskaren ska i sin tur ta hand om dataskyddet för sitt material under tiden för forskningsanvändningen och vid publicering av forskningsutskrifter.

Forskaren ansvarar för dataskyddet i fråga om de forskningsresultat som han eller hon publicerar. Dessa anvisningar har uppgjorts för att främja ansvarsfullt förfarande i dataskyddsärenden och syftet med anvisningarna är att hindra såväl oavsiktliga som avsiktliga dataskyddsförseelser. Forskartjänsterna använder ett kontrollförfarande för forskningsresultat, som också främjar aktsamhet i dataskyddsfrågor. En närmare redogörelse för kontrollförfarandet finns i avsnittet Överföring av utskrifter från distansanvändning.

Varför dataskydd?

Ett fungerande dataskydd är en förutsättning för att Statistikcentralen för forskningsanvändning och mikrosimulering ska kunna lämna ut forskningsmaterial som den samlat in för statistiska ändamål. Statistikcentralens rätt att samla in register- och enkätmaterial för statistikföring har tryggats i statistiklagen, såsom också rätten att lämna ut dessa uppgifter för forskningsanvändning. Å andra sidan innehåller lagen också ett åliggande att se till att uppgifterna skyddas på sakligt sätt och en del av materialet innehåller också väldigt känsliga uppgifter.

Behandlingen av personuppgifter styrs av Europeiska unionens allmänna dataskyddsförordning (EU) 2016/679 och den nationella dataskyddslagen (1050/2018), som kompletterar den. Enligt sekretessplikten i anknytning till användningstillståndet för ett forskningsprojekt ska forskaren se till att forskningsresultaten inte innehåller uppgifter på enhetsnivå, det vill säga att det inte är möjligt att röja uppgifter om en enskild person eller ett enskilt företag. Då forskaren behandlar material som överlämnats för forskningsanvändning eller då forskaren använder det i distansanvändningsmiljön, ska han eller hon enligt

sekretessplikten se till att materialet inte röjs eller överlåts till en aktör som saknar rätt att använda det.

Filer som lämnas in till ett distansanvändningsprojekt

Man kan be om att offentliga uppgifter överförs till ett projekt, om inga uppgifter är på enhetsnivå. Om ett material innehåller uppgifter på enhetsnivå kan det krävas användningstillstånd för det, även om materialet är offentligt. Ett material som kräver användningstillstånd tilläggs till ett projekt först efter tillståndsförfarandet.

Offentliga uppgifter som inte är på enhetsnivå och som kan överföras till ett projekt är t.ex. områdes- och kommunkodsklassificeringar eller kommunspecifika uppgifter, t.ex. folkmängd eller areal.

Offentliga uppgifter på enhetsnivå som kräver användningstillstånd är t.ex. offentlig upphandling.

Offentligt material som innehåller direkta identifierare kan inte överföras direkt till ett projekt. Sådant material kräver att användningstillståndet uppdateras.

Om du ber om att en fil överförs till ett projekt, uppge i meddelandet om det är fråga om offentliga uppgifter eller uppgifter som kräver användningstillstånd. Berätta också i meddelandet vilka uppgifter filen innehåller och om filen innehåller uppgifter på enhetsnivå.

Dataskyddsanvisningar för olika utskriftstyper

Nedan presenteras anvisningar och regler för att sköta dataskyddet för olika utskriftstyper. Eftersom det finns många olika forskningsprojekt och utskriftstyper, är det inte möjligt att bedöma och ge anvisningar om dataskyddet för varje material och utskrift separat. Därför är det nödvändigt att ge allmänna tumregler.

Dataskyddsanvisningarna kan avvika från de nedan presenterade reglerna i fråga om utlämnade av vissa material av Statistikcentralen och material som utlämnats av övriga myndigheter för forskningsanvändning. Dessa avvikande regler finns i materialbeskrivningarna i Taika-katalogen, och vid behov antecknas de också i beslutet om användningstillstånd.

Om en forskare är osäker på dataskyddet för en utskrift eller vill be om precisering av givna anvisningar, ska han eller hon kontakta Forskartjänsterna. (tutkijapalvelut@stat.fi).

Utskrifter som producerats från urvalsmaterial

Det är gemensamt för alla utskriftstyper att risken för röjande av observationer bakom resultaten i allmänhet är mindre, om materialet som används enbart är ett (slumpmässigt) urval ur totalmaterialet. En mindre risk för röjande som eventuellt beror på urvalet innebär dock inte att de dataskyddsregler som presenteras i denna anvisning automatiskt kan lindras för utskrifter som producerats utifrån urvalsmaterial.

Om det dock är nödvändigt att för en utskrift göra en fallspecifik bedömning av risken för att en observation röjs (t.ex. om ett fördelningsnyckeltal är avslöjande

eller om det är möjligt att identifiera en observation bakom en enskild bildpunkt på en bild), påverkar användningen av urvalet i så fall bedömningen.

Det ska observeras att ett totalmaterial inte alltid omfattar enbart alla personer som bor i Finland eller alla företag som verkar i Finland. Med tanke på dataskyddet och bedömningen av risken för röjande bildar alla företag inom till exempel en viss näringsgren snarare ett totalmaterial än ett urvalsmaterial, eftersom det i allmänhet är möjligt att dra en slutsats om huruvida ett enskilt företag hör till det ”urval” som avgränsats för att täcka alla företag inom en viss näringsgren.

Frekvens- och mängdtabeller

De uppgifter som aggregerats i tabellform kan innehålla personuppgifter, företagsuppgifter eller bägge två. Till exempel i arbetstagar-arbetsgivar-material ska såväl person- som företagsnivån skyddas. Också sådana uppgifter som beskriver olika yrkesgrupper kan indirekt innehålla företagsuppgifter, om (nästan) alla personer som hör till en viss yrkesgrupp tjänstgör enbart för ett (monopol)företag, vilket innebär att företagets dataskyddet ska tryggas.

I regel är tröskelvärdet 3 både i skyddet av tabeller som innehåller företagsuppgifter och i skyddet av tabeller som innehåller personuppgifter. Med andra ord är det tillåtet att publicera tabellceller eller uppgifter i en observationsgrupp enbart om åtminstone 3 (oviktade) observationer finns bakom uppgifterna. Inte heller information om antalet observationer får publiceras vad gäller små celler eller grupper. Närmare anvisningar om skyddsmetoder uppges i ett separat kapitel längre ner.

Användning av ett tröskelvärde som huvudregel är det enklaste sättet att räkna ut en potentiell risk för röjande av enskilda observationer som hänför sig till de tabeller som ska publiceras. Identifiering av en observation som det enda eller ett sällsynt fall i en grupp kan öka risken att en observation identifieras och/eller att fler uppgifter om observationen röjs (t.ex. med hjälp av andra tabeller eller utskrifter som behandlar samma material eller ämne). Därför ska tröskelvärdet tillämpas, trots att enbart mängduppgifter publiceras i en tabell.

Utöver ovan nämnda huvudregel ska följande regler beaktas:

- I färska företagsuppgifter (nyare än 15 månader från referenstidpunkten) ska man utöver tröskelvärdet tillämpa dominansregeln[1] (1,75). Det kan förutsättas att dominansregeln används också i äldre företagsuppgifter eller i annat material (t.ex. inkomstregistret). Dominansregeln förutsätter att man ska skydda till exempel uppgifter där ett företags omsättning eller lönebelopp eller en annan variabel i materialet står för över 75 procent av företagens omsättning/lönebelopp inom en viss näringsgren (sektor/område e.d.) eller om till exempel över 75 procent av löntagarna i en viss yrkesgrupp arbetar hos ett visst företag.

[1] Enligt dominansregeln (n,k) kan tabellens cellvärde inte publiceras, om dess n största observationer utgör minst k procent av cellens totala värde.

- Vid skydd av uppgifter på arbetsställesnivå ska man också i mån av möjlighet säkerställa skyddet på företagsnivå, det vill säga att det i varje cell ska finnas arbetsställen från minst tre olika företag. Förfarandet ska vara detsamma för koncern-företag-förhållanden.
- I tillgångsuppgifterna (information från statistiken över industriproduktion, material och förnödenheter) är antalet företag en konfidentiell uppgift i fråga om alla produktionsbenämningar.
- Dataskyddet av personuppgifter ska skötas särskilt noggrant, om tabellen innehåller känsliga personuppgifter[2] (inkl. de särskilda kategorier av personuppgifter som räknats upp i EU:s allmänna dataskyddsförordning). I skyddet av känsliga uppgifter kan det vara nödvändigt att använda ett högre tröskelvärde.
- Vad gäller yrkesutövare som förekommer i företagsuppgifterna tillämpas samma skyddsregler som för övriga företagsuppgifter, trots att yrkesutövare i princip är personer.

Publicering av koordinat- eller rutuppgifter är förknippade med särskilda villkor: Rutmaterial i sin helhet får inte publiceras. Uppgifter som baserar sig på rutor får tas ut ur distansanvändningssystemet och publiceras då det finns minst 10 observationsenheter i en ruta. Den information som ska publiceras måste endera sammanställas till en större regional nivå, ges som relationstal eller behandlas på annat sätt så att personer och bostadshushåll inte kan identifieras.

Olika fördelningsnyckeltal

Minimum och maximum är i allmänhet förknippade med en observation, varför de oftast inte kan publiceras. Till exempel kan det största företaget inom en näringsgren i allmänhet identifieras, och därför kan relaterade uppgifter om maximiomsättningen inte publiceras. Om minimum eller maximum inte kan förenas med en enskild observation eller om tröskelvärdet överskrids, kan de publiceras.

Fördelningspunkterna (exkl. minimum och maximum), såsom deciler, utgör ett specialfall av tabellen där antalet observationer som blir mellan fördelningspunkterna motsvarar cellfrekvenserna. Om dessa tal överskrider det tröskelvärde på 3 som ska tillämpas på tabellerna, kan fördelningspunkterna publiceras.

Typvärdet kan publiceras, om (nästan) alla observationer inte får samma värde.

[2] Uppgifter som är känsliga avser här uppgifter som beskriver en persons ras eller etniska ursprung, politiska åsikter, religiösa eller filosofiska övertygelse, medlemskap i fackförbund, hälsa, sexuella beteende eller inriktning, dom i brottmål och förseelser eller säkringsåtgärder i anknytning till dem, dödsorsak, språk, nationalitet, härkomst eller födelseland, inkomster, skulder, förmögenhet, ovanliga yrke eller annan socioekonomisk ställning, behov av socialvård eller erhållna socialvårdstjänster, stödåtgärder eller andra förmåner inom socialvården, eller uppgifter om vårdåtgärder eller därmed jämförbara åtgärder som gäller personen.

Medeltalet, andra relationstal och de högre momenten i fördelningsnyckeltalen (t.ex. varians) kan publiceras om man vid beräkningen av dem har använt minst tre observationer.

Vid publiceringen av andelarna ska tröskelvärdet 3 överskridas vad gäller alla grupper som bildar andelar. Med andra ord, om man vill publicera till exempel att kvinnornas andel är 58 procent av hela populationen, ska dessa 58 procent, liksom också männens 42 procent innehålla åtminstone tre personer. Det är med andra ord inte tillräckligt att det i hela populationen finns åtminstone totalt 3 kvinnor och män.

Andra numeriska utskriftstyper

Indextal, korrelationskoefficienter och teststorheter (t, F, X^2 o.d.) kan i allmänhet publiceras, om åtminstone 10 observationer använts i beräkningen av dessa.

Regressionsmodeller kan publiceras i sin helhet om det ligger tillräckligt med observationer bakom modellen (åtminstone 10) och modellen inte beskriver tidsserien av observationer av ett företag eller en person. Enskilda koefficienter i modellen kan i allmänhet alltid publiceras.

Numeriska resultat i mer komplicerade statistiska modeller kan i allmänhet alltid publiceras, om tillräckligt med observationer finns bakom modellen och modellen inte beskriver tidsserien av observationer om ett enskilt företag eller en enskild person.

Bilder

I likhet med numeriska utskrifter, får inte heller bilder avslöja uppgifter om enskilda observationer. Bilder som ritats utgående från materialet är tillåtna, om en enskild bildpunkt eller en del av bilden inte kan röja en enskild observation bakom den.

Stapeldiagram och andra figurer för presentation av klassificerat material är vanligen tillåtna att publiceras, så länge det finns en tillräcklig mängd observationer i varje klass. Informationen i en sådan bild kan i allmänhet också presenteras i tabellform och det är möjligt att på den tillämpa samma dataskyddsregler som på annat tabellmaterial (se ovan i punkten Frekvens- och mängdtabeller).

Fördelningar, histogram och kumulativa fördelningsfunktioner som har utjämnats eller presenterats på en tillräckligt grov nivå är tillåtna. En del av fördelningsbilderna kan innehålla uppgifter om avvikande observationer eller extremvärden, vilka från fall till fall kan röja uppgifter om en enskild observation. Programmens ritfunktioner märker ofta ut avvikande observationer automatiskt bland annat i låddiagram (box plot). Bilder som specificerar avvikande observationer lämpar sig inte för publicering, såvida forskaren inte kan motivera att de avvikande observationer som antecknats på bilden inte kan identifieras.

Spridningsbilder används vanligen för presentation av värden hos två kontinuerliga variabler. Punkterna i spridningsbilder beskriver i princip uppgifter om varje enskild observation, varför de i jämförelse med föregående bildtyper är

ett besvärligare fall med tanke på dataskyddet. I bedömningen av dataskyddet för spridningsbilder ska forskaren rikta särskild uppmärksamhet mot materialets natur, med tanke på bland annat urvalets storlek, informationens känsliga natur och förekomsten av avvikande observationer. Spridningsbilden uppfyller inte dataskyddskraven, om det utifrån dessa inte direkt kan ses eller enkelt dras en slutsats till exempel om det största företaget inom en enskild näringsgren. Dataskyddet för en spridningsbild kan förbättras genom att klassificera det bakomliggande materialet på så sätt att det i spridningsdiagrammet finns flera observationer bakom en punkt. Ett alternativt sätt, där man i vissa fall kan minska röjanderisken för enheter, är att lägga till slumpmässigt brus (jittering) till punkter i spridningsdiagrammet.

Skyddsmetoder

Filer eller tabeller som tas ut ur distansanvändningssystemet får inte innehålla en möjlighet att identifiera en enskild person eller ett enskilt företag. Uppgifter som omfattas av risk för röjande ska skyddas genom att planera utskrifternas innehåll på ett sätt som är godtagbart med tanke på dataskyddet, genom att använda till exempel grova klassificeringar.

De tabellceller som omfattar en risk för röjande kan skyddas genom att ändra tabellstrukturen, genom att täcka enskilda cellvärden eller hela rader eller genom att ändra cellvärden till exempel genom att avrunda eller ersätta det ursprungliga cellvärdet med ett ungefärligt slumpmässigt tal. I valet av skyddsmetod för tabellen finns det skäl att sträva efter att hitta en metod som täcker tabellen tillräckligt, men att bevara dess egenskaper som är viktiga för användningsändamålet på ett maximalt bra sätt.

Ändring av strukturen på tabellen innebär kontroll av antalet variabler eller ändring av klassificeringen. Genom att ändra klassificeringen strävar man efter att ta bort celler förknippade med en risk för röjande genom att kombinera klasser som innehåller dem med andra klasser i tabellen. Att ändra klassificeringen betyder ofta i praktiken att hela klassificeringen görs grövre.

Till täckning hör primär täckning av celler som ligger i risk för att röjas och sekundär täckning. Med sekundär täckning säkerställs att man inte med hjälp av tabellens rad- eller kolumntotaler kan avslöja de cellvärden som täckts primärt. Täckningen kan också göras radvis. Om en radtotal i tabellen innehåller bara en liten mängd statistiska enheter (färre än tröskelvärdet), täcks denna rad i sin helhet utan att beakta antalet statistiska enheter i de olika cellerna.

Närmare information om dataskyddsmetoderna finns till exempel i materialet i webbkursen Tutkimusaineistot etäkäytössä (ung. Distansanvändning av forskningsmaterial), som nämns i slutet av anvisningen.

Överföring av utskrifter från distansanvändning

Vad gäller material som är i distansanvändning använder forskartjänsterna ett kontrollförfarande med stickprover för forskningsresultat och manuellt kontrollförfarande.

Manuellt kontrollförfarande

Forskningsresultat som producerats i distansanvändning granskas innan uppgifterna överlåts och det är inte möjligt att själv överföra filer från distansmiljön till den egna arbetsstationen, utan dataöverföringen sker med en separat begäran per e-post via (tutkijapalvelut@stat.fi). Granskningen sker inom 1–2 arbetsdagar.

Filer som begärs för extrahering ska sparas på O-disken, och de ska kunna tolkas tydligt och vara förenliga med de separata dataskyddsreglerna. Användaren ansvarar för att filerna är förenliga med reglerna.

Sänd ett e-postmeddelande till ovan nämnda adress och beskriv datainnehållet i de filer som du begärt ut i meddelandet.

Antalet observationer per cell ska vara synligt i tabellerna, liksom också antalet observationer som använts i beräkningen av estimat och nyckeltal. Också antalet observationer bakom bilderna ska framgå av utskriften.

Om antalet observationer inte framgår av de egentliga utskrifterna, uppge det i ditt meddelande. Nämn distansprojektets kod i e-postmeddelandet (till exempel x01).

Utskrifterna kommer efter granskningen automatiskt som ett e-postmeddelande inom cirka 15 minuter från granskningen (i vissa fall har meddelandena styrts till mottagarens mapp för skräppost, som det lönar sig att granska om du inte hör något om utskrifterna)

Kontrollförfarande med stickprover

När forskaren vill ta ut resultat från distansanvändningsmiljön fyller han eller hon i ett formulär på arbetsbordet i Fiona om de resultat som ska tas ut. Därefter kan resultaten antingen kontrolleras manuellt i förväg eller, om forskaren är föremål för stickprovskontroller, fås omedelbart per e-post.

Alla resultat från nya användare kontrolleras i förväg. Sannolikheten för att omfattas av förfarandet med stickprovskontroller ökar i takt med att användaren genomgått flera godkända förhandskontroller efter varandra. Förseelser och fel nollställer situationen och förhandskontrollerna börjar från början. Som sanktion för allvarliga och upprepade förseelser omfattas personen inte längre av förfarandet med stickprovskontroller. Också andra åtgärder vidtas till följd av dataskyddskränkningar.

Om en forskare är osäker på en fråga som gäller dataskyddet av en utskrift, lönar det sig för hen att vara i kontakt med Forskartjänsterna redan innan resultaten tas ut ur systemet. Att ta kontakt och utreda dataskyddet på förhand inverkar inte på forskarens möjligheter att bli omfattad av förfarandet med stickprovskontroller.

Om utskrifter som överförs

Forskaren har ansvar för dataskyddet för sina forskningsresultat. Forskaren ska se till att de utskrifter som överförs (eller som önskas överföras) från distansanvändningssystemet inte innehåller material på enhetsnivå eller en möjlighet att en uppgift innehållande en enskild observation röjs.

Utskrifter från distansanvändningssystemet ska tas ut efter prövning. Endast sådana utskrifter som är avsedda att publiceras ska tas ut ur systemet. Med andra ord ska överföring av så kallade mellanresultat och i synnerhet (stora) loggfiler undvikas. Tabellerna och diagrammen bör till sitt innehåll vara i den form som de avses publiceras. Det är inte möjligt att ur systemet ta ut sådana utskrifter som på grund av dataskyddet inte kan publiceras.

Antalet observationer som använts i beräkningen av tabeller, bilder, nyckeltal osv. ska vara synliga i utskrifterna. Om man vill ta ut sådana uppgifter ur distansanvändningssystemet som avviker från dataskyddsanvisningarna, men som inte kan röja uppgifter om en individ, ska forskartjänsterna kontaktas innan en begäran om utskrift görs.

Filformatet för utskrifter, i synnerhet för bilder, ska vara sådant att det inte orsakar risk för att uppgifter på enhetsnivå röjs. Filen ska kunna öppnas för granskning med verktygen i Fiona. De bildformat som lämpar sig för granskning utgörs av till exempel:

- Bitkarteformat
- PNG (Portable Networks Graphics)
- BMP (Bitmap)
- JPEG (Joint Photographic Experts Group)
- TIFF (Tagged Image File Format)
- Vektorformat
- EPS (Encapsulated PostScript)
- PS (PostScript)
- PDF (Portable Document Format)
- SVG (Scalable Vector Graphics)
- WMF/EMF (Windows Metafile)

I programmet Stata kan ovan nämnda bildformat skapas med kommandot graph export. I programmet SPSS kan figurformatet väljas i funktionen Export output. I R-programmet får man information om ritfunktionen med kommandot help(grDevices). Vissa figurtyper, såsom gph-filer i Stata, sparar i regel det material som använts när figuren ritats och därför lämpar de sig nödvändigtvis inte för överföring till granskning och extrahering.

Ytterligare anvisningar om dataskydd

Webbhandboken Tutkimusaineistot etäkäytössä (ung. Distansanvändning av forskningsmaterial) ger mera information om distansanvändningen av forskningsmaterial och om användning av SISU-mikrosimuleringsmodellen och dataskyddet av material. Kursen innehåller också exempel på skydd av tabellmaterial.

I synnerhet för forskare som för första gången använder distansanvändningssystemet rekommenderas att de, vid sidan om dessa anvisningar, också tar del av webbkursens material.