Danmarks Statistik

Retningslinjerne for brug af Forskermaskiner

Institut for Folkesundhed

Aarhus Universitet

Institut for Folkesundhed
Datamanagement, nsb
Version 1.0
2014

1 Indholdsfortegnelse

2 Introduktion

Herunder er retningslinjerne for hjemtagning af informationer fra Danmarks Statistik (DST) refereret, tolket og forklaret med eksempler, anno 2014.

Det er vigtig, at instituttets medarbejdere og studerende til fulde forstår, hvilke betingelser der gælder for at der kan arbejdes på DSTs forskermaskiner og hjemtages informationer.

> **Bemærk, at der gælder helt særlige regler for dette arbejde og at arbejdsrutinerne er meget anderledes end det de fleste er vant til – derfor kan man ubevist kommer til at overtræde disse regler, som kan have fatale konsekvenser for dig som bruger og påføre de øvrige medarbejdere sanktioner i længere tid.**
>
> **Visse overtrædelser er endvidere omfattet af staffelovens bestemmelser!**

3 Hjemtagning af informationer fra DST

Reglerne for hjemtagelse af informationer:

> **Citeret fra DST "Hjemsendelse-af-filer-fra-forskermaskiner.pdf":**
>
> 1. ***Filerne må ikke indeholde identificerbare data, dvs data der indeholder enkeltrekords eller hvor virksomheder eller personer på anden måde er identificerbare.***
> 2. ***Filen må derfor fx ikke indeholde én variabel fra én observation fra et rådatasæt, lige meget hvad denne variabel indeholder.***
> 3. ***Optællinger, tabeller, output, kørselslogs, programmer og grafikfiler, må gerne hjemsendes. Alt sammen under forudsætning af, at de ikke indeholder identificerbare data.***

Den fulde fil er her: Link

Det fulde regelsæt og lovgivning kan læses her: http://www.dst.dk/da/OmDS/Lovgivning.aspx

Det er vigtigt at man sætte sig grundigt ind i de få ting der står her og desværre kan retningslinjerne læses og tolkes på flere måder. Derfor er der her givet en uddybning af betydningen af reglerne.

Fortolkninger af reglerne:

Pkt 1. Dette punkt er lige til – Ingen hele records, altså en observation, må hjemtages fra et rådatasæt. Eller en reduceret record med informationer, som kan identificerer en person eller et firma.

- En overtrædelse kunne f.eks. være en information, der siger: ”et firma med en omsæt på over 20 mia. kr. i Bjerringbro kommune”. Her er to variable nok, omsætning og kommune. Så ved alle, at det i dette eksempel drejer sig om Grundfos.

Pkt 2. Er mere listigt, der står faktisk; at ikke så meget som én celleværdi fra et rådatasæt må hjemtages uanset, hvad værdien af cellen er. Dvs. bare ét tal fra ét rådatasæt. Her skal man virkelig passe på!

- I eksemplet ovenover med Grundfos, havde vi den ene variable værdi sat til ”over 20 mia”. Hvis vi nu havde det præcise tal for omsætning fra rådata, f.eks. 22,6 mia. kr. som omsætningen var i 2012, så ville dette ene tal være en overtrædelse af hjemtagningsreglerne.

- pnr-nummeret, som indgår i mange datasæt, er flere steder beskrevet som ”et anonymiseret cpr-nummer, der er unikt for hvert projekt”. Dette er ikke længere tilfældet. pnr kan afkodes til cpr, hvis man kender algoritmen - og DST kender sammenhængen. pnr-numret må derfor ikke fremover forlade forskermaskinerne, da de ikke længere betragtes som anonyme - og de skal behandler som et cpr-nummer. Det samme gælder nu for recnum, famid etc. og nøgle værdier generelt (altså et observations id).

Pkt 3. Det lyder som om, at de nævnte filtyper gerne må hjemsendes, hvis ikke de ikke indeholder identificerbar data. Men, her skal man også passe på!

- I en do fil kan der f.eks. stå ”drop if pnr == 123.45”. UPS! En celle værdi! En overtrædelse!

- Logfiler fra Stata kan indeholder langt mere information end man lige tænker over – og mere information end man har set på skærmen!

- Man kunne nemt fristes til at plotte en variable, og hjemtage plottet, for at få et overblik. Her kan nogle signifikante værdier, uden for en plagemage af mange værdier, indeholde identificerbar data. Dette kunne f.eks. være et punkt-plot af omsætningen for alle virksomheder i Bjerringbro, hvor en stor virksomhed skeller sig markant ud fra alle de andre - og derved bliver identificerbar!

- Der må aldrig være mindre end f.eks. 5 råværdier i en hjemtagen beregnet værdi. Så vil man f.eks. lave et histogram over omsætningen i virksomhederne i Bjerringbro, så skal den søjle, som f.eks. indeholder Grundfos, indeholde mindst 4 andre virksomheder. Så gruppe må ikke hedde f.eks. over 20 mia. kr. i omsætning, men skal måske hedde over 100 mio. kr. i omsætning få at få nok virksomheder med i gruppen.

- Der står, at man må hente tabeller, som ikke indeholder rådata. Dette skal forstås som, at tabellen ikke må indeholde pnr-numre - og at f.eks. et beregnet gennemsnit i en variable skal dække over mindst 5 observationer af en variablen i rådatasættet.

3.1 Simple statistiske værdier

Simple statistiske værdier som minimum (min), maksimum (max) og median referer direkte til celle værdier i rådata – det er derfor ikke tilladt at henvis til disse værdier.

Det må generelt kun hentes aggregerede data. Dvs. beregnede data, som minimum er gennemsnitte af 5 observationer.

Der må heller ikke refereres til en antal, hvis antallet er mindre end 5.

3.2 Pas på med statistik programmerne

F.eks. et statistik program som STATA kan logge alle informationer - også informationer som ikke er set på skærmen!

Dette betyder at celle værdier fra radata står i disse log filer, hvis der f.eks. kommandoen ”codebook”, ”summarize”, ”list” etc. er blevet kaldt.

STATA do-filer kan lige ledes indeholde celleværdier fra rådata, f.eks. ”drop if pnr == 1234”.
Her er endda to overtrædelser både en pnr-variable og en celleværdi ”1234”.

Når du skriver f.eks. do-filer kan det være en ide, at skrive f.eks. ”drop if X < 100” i stedet for ”drop if X == 123.45”. Derved kan du undgå, at der direkte er celleværdier fra rådata i do-filen – som du så ikke må hjemtage. Selvfølgelig kun hvis valget ”100” ikke er lig en celleværdi.

Det kan derfor være en god ide kun at nedtage de endelige plots, som skal med i f.eks. en artikel - efter at man har gennemgået dem nøje. Samt et minimum af dokumentation i form af f.eks. do-filer og log-filer fra e.g. STATA, hvor man har gennemlæst filerne grundigt.

3.3 Vær også opmærksom på

Det kan være umådeligt svært, at undgår at overtræde reglerne fra DST, når man nedtager informationer, plots eller dokumentation for, hvad man har gjort for, at nå frem til sit resultat.

- En dokumentation kan f.eks. beskrives sprogligt i en tekst, frem for matematisk i en do-file. Derved kan man undgå at der er celleværdier fra rådata.

- Skal du tale med f.eks. din vejleder, om den proces, du er i gang med på DSTs forskermaskiner, så log på og vis tingene direkte på en DST forskermaskine i stedet for at hente informationer hjem.

  Bemærk, at din vejleder ikke nødvendigvis har tilladelse til at se data på fra DST!
  Hvilket så vil være et brud på fortroligheden vedr. dine data.

- Det kan være meget tidkrævende, at sikre sig, at reglerne ikke er overtrådt. Derfor er det gode råd, at lad være med at hjemtage information, med mindre det er stængt nødvendigt – og at du har gennemgået informationerne minutiøst inden du hjemtager dem.

- Du må aldrig f.eks. zippe eller bruge andre former for komprimering/kryptering af det du nedtager. Dette er i sig selv en overtrædelse af DSTs regler, da det bliver betragtet dom et forsøg på at undgå kontrol.

- Du må aldrig nedtage ”store mængder” filer af gangen.
  Dette er i sig selv en overtrædelse af DSTs regler, men der findes ingen definition på hvor mange filer ”store mængder” er.

- Arbejd kun inde i den lukkede verden på DSTs Forskermaskiner og lad være med at lave arbejdskopier eller medbringe identificerbare data til møder.

- Det er heller ikke tilladt at skrive af efter skærmen eller tage billeder af skærmen.

- Det er heller ikke tilladt at brug data til at finde frem til information om enkeltpersoner, som anvende til andre formål end det der er givet tilladelse til. F.eks. hvis du finder ud af at en "kendt person er med i dine data". Dette er ikke var en overtrædelse af DST reglerne, men en overtrædelse af den danske lovgivning.

- Når du har fået din godkendelse, så kontakt Datamanagement - og spøg, når du er i tvivl!

3.4 Overtrædelser og konsekvenser

Hvis du laver en fejl, endda meget ubevist, så rammer straffen alle medarbejder på instituttet i form af mindst 1 mdr. udelukkelses fra DST. Du lukker altså også alle andres projekter ned!

Der ud over får den enkelte, der var årsagen:

- 1. gang: 3 mdr. karantæne
- 2. gang: 12 mdr. karantæne
- 3. gang: Permanent udelukkelse via aftalen med DST.

Bevidste handlinger for at tilegne sig personfølsomme data med føre en politi anmeldelse med henvisning til Forvaltningslovens § 27, stk. 3 og Straffelovens §152.

Læse det fulde regelsæt via linket ovenfor, inden du går i gang…