Kvalitetskriterier för statistik baserad på digitala data – bakgrund

Innehåll

Sammanfattning

SCB har under ett antal år studerat nya typer av datakällor, till exempel data från elmätare, platsannonser och mobiloperatörer. Vi har valt att kalla dessa för digitala data (National Academies of Sciences, Engineering, and Medicine 2022). SCB har sett ett behov av att ta fram kriterier för att bedöma kvaliteten i den här typen av data. Den här rapporten tillsammans med Kvalitetskriterier för statistik baserad på digitala data—vägledning är en del av regeringsuppdraget ”Uppdrag att främja delning och nyttiggörande av data för smart statistik” (Regeringen 2021).

Bestämmelser om kvalitet i den officiella statistiken föreskrivs i SCB:s föreskrift om den officiella statistiken (SCB-FS 2016:17), med stöd av16 § 2 förordningen (2001:100) om den officiella statistiken. Kvaliteten i en slutprodukt ska beskrivas i en kvalitetsdeklaration med utgångspunkt i de fem huvudkomponenterna Relevans, Tillförlitlighet, Aktualitet och punktlighet, Tillgänglighet och tydlighet samt Jämförbarhet och samanvändbarhet. Vägledningen ska ses som ett komplement till ovanstående. Det är kvalitetskomponenterna Relevans och Tillförlitlighet som är i fokus i det här arbetet, då beskrivningar av de övriga komponenterna inte behöver kompletteras med någon ny indikator. I vägledningen är det kvaliteten i den slutliga statistik som produceras med digitala data som avses.

En genomgång av relevant litteratur har gjorts inom ramen för det här uppdraget och sammanfattas i den här rapporten. Processen då statistik produceras baserat på digitala data illustreras och potentiella osäkerhetskällor som kan uppstå beskrivs övergripande i den här rapporten och vägledningen. I vägledningen beskrivs osäkerhetkällorna mer utförligt. Beskrivningen av processen och osäkerhetskällor bygger på ramverk för Total Survey Error (Groves och Lyberg 2011) men också på Zhang (2012), Reid et al (2017) och Lothian et al (2019) där integrering av data är centralt. Några begrepp är lånade från Sen et al (2022) men används i en mer generell mening. Så långt som möjligt är begrepp och definitioner anpassade till det språkbruk som SCB använder, och avvikelser från det är tänkt som kompletteringar.

I vägledningen föreslås mätbara kvalitetsindikatorer. Dessa bygger på bland annat ett EU-arbete där man tagit fram

indikatorer på kvalitet då administrativa data integreras (KOMUSO 2019). Det viktigaste första steget när man har digitala data är att göra en kvalitativ utvärdering för att förstå exakt vad data innehåller. En sådan utvärdering ger information om vilka osäkerhetskällor som behöver studeras närmare baserat på användarbehov. För vissa digitala data kan det stora problemet vara täckningen medan andra typer av digitala data kan ha problem med både täckning, validitet och mätfel. Det är svårt att ge mer generella rekommendationer för hur indikatorer ska bedömas då både digitala data och användarbehov kan se väldigt olika ut.

Den uppsättning kvantitativa mått som föreslås i vägledningen är en av flera aspekter som statistikproducenten behöver ta ställning till för att bedöma om en digital datakälla kan användas. I de övervägningar som görs behöver även de övriga kvalitetskomponenterna som till exempel aktualitet och punktlighet vägas in. Ytterligare dimensioner att ta med i avvägningen är huruvida den digitala datakällan bedöms vara hållbar över tid samt hur datakällan ska användas. Slutligen behöver kostnader och uppgiftslämnarbördan också vägas in.

Det här arbetet har diskuterats internt på SCB vid flera tillfällen och i flera olika grupper. Många medarbetare har lämnat värdefulla synpunkter under arbetets gång och på olika utkast av vägledningen och bakgrundsrapporten. Arbetet har även presenterats för SCB:s Vetenskapliga råd och även där har värdefulla synpunkter på arbetet framförts. Slutligen har arbetet presenterats på ett seminarium för Statistikansvariga myndigheter.

Inledning

SCB har under ett antal år studerat nya typer av datakällor, till exempel data från elmätare, platsannonser och mobiloperatörer. Data från dessa källor kan skilja sig på flera sätt från data som kommer från urvalsundersökningar eller administrativa källor. SCB har därför sett ett behov av att ta fram kriterier för att bedöma kvaliteten i nya typer av data. I september 2021 fick SCB dessutom ett regeringsuppdrag (Regeringen 2021) att bland annat studera data från mobiloperatörer. I uppdraget ingår att ”*SCB ska lämna förslag till lämpliga kvalitetskriterier för data, primärt ur*

ett statistikperspektiv och utifrån den utveckling som sker nationellt och internationellt, särskilt inom EU.”

Vidare sägs i regeringsuppdraget *”För att fullt ut kunna dra nytta av data inom den förvaltningsgemensamma digitala infrastrukturen, eller hos privata dataägare, är det avgörande att kunna bedöma datakvaliteten med stöd av olika kvalitetskriterier. Kriterierna blir vägledande för om olika datakällor är lämpliga som underlag för statistikframställning och i vilken utsträckning de kan ersätta uppgifter som idag samlas in via enkäter eller intervjuer.”*

I detta arbete är utgångspunkten att den kvalitet som avses är kvaliteten i den slutliga statistik som produceras med nya typer av data. Även om syftet är att avgöra datakvalitet så kan en fullständig utvärdering inte göras fristående från datas användning i de slutliga skattningarna. Användningen av statistiken avgör vilka krav som behöver ställas på datakvaliteten. Producenter av officiell statistik har dessutom inte rätt att samla in data om det inte finns en tänkt användning, och i praktiken är det troligen alltid en idé om användningen som ligger bakom att en datakälla alls undersöks.

Samtidigt behöver det vara tydligt vilka delar i utvärderingen som är generella för olika användningar, och vilka delar som är specifika för en viss skattning. I föreliggande arbete läggs stor vikt vid länkning med befintliga register. Länkningen är viktig eftersom det är troligt att nyttan av data blir begränsad annars, men även för att den framtida statistikproduktionen kommer bygga mycket på att integrera flera datakällor (se till exempel De Waal et al 2020). Utgångspunkten är inte nödvändigtvis att en ensam datakälla förväntas täcka hela behovet, utan snarare att olika datakällor kommer behöva kombineras med andra register, eller kompletteras med urvalsundersökningar.

Resultatet av en sådan länkning kan troligen användas i flera produkter och därför bör länkningen göras och utvärderas enbart på ett ställe i organisationen, och med gemensamma kvalitetskrav. Motsvarande gäller för kodning och annan bearbetning som uppfyller gemensamma behov för olika användning. Hela eller delar av ett register med länkade källor kan sedan bearbetas vidare för specifik användning och med olika krav på kvaliteten i de slutliga skattningarna, på motsvarande sätt som befintliga register redan används.

Ordet kriterier kan ge intryck av det finns absoluta svar på vad som är tillräckligt bra data. Det gör det inte, utan de indikatorer som föreslås här syftar till att beskriva vissa aspekter av kvaliteten. Det ger information om var det finns problem. Hur dessa ska hanteras och när data är tillräckligt bra för att tas i produktion varierar och beror, förutom vad data ska användas till, även på andra aspekter som till exempel kostnader och uppgiftslämnarbörda.

Målsättningen är att presentera mätbara indikatorer, vilket är i linje med hur kvalitetsindikatorer definieras i till exempel Statistics New Zealand (2016), Reid et al (2017), De Waal et al (2019) och UNECE (2021). En annan typ av indikatorer, som inte diskuteras vidare i denna rapport, kan vara om det finns dokumentation eller om en viss process är implementerad. Tanken är att indikatorerna där det behövs ska kunna komplettera den fastställda kvalitetsdokumentationen för officiell statistik (SCB 2020, SCB 2019).

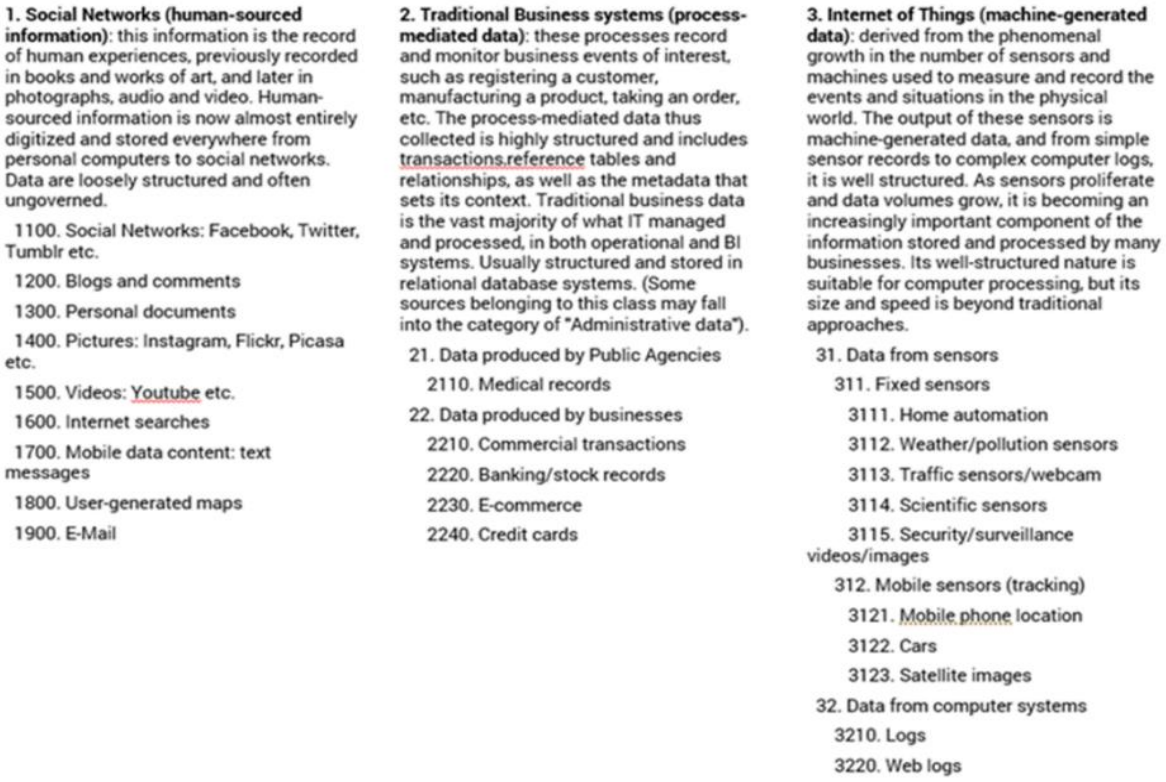
Syftet med indikatorer är generellt att ge en förenklad men samtidigt relevant och heltäckande bild av ett nuläge. Risken finns att det blir en förenklad bild eller att indikatorn missar målet. I fallet med nya typer av datakällor är SCB dessutom fortfarande i en utvecklingsfas. Indikatorerna bör därför regelbundet utvärderas och vid behov kompletteras eller på annat sätt ändras.

Indikatorerna kan vara relativt enkla att ta fram och skulle därför kunna användas för en initial utvärdering men även i produktion i varje produktionsomgång. Det gäller till exempel andelen imputerade värden eller andelen kodade värden. Andra indikatorer kan handla om att beskriva kvaliteten i en modell, och då görs det bara när modellen tas fram eller utvärderas, till exempel en modell för imputering eller kodning. Vissa indikatorer kan kräva specialstudier eller experiment, till exempel för att utreda mätfel. Då är det lämpligt att designa en speciell urvalsundersökning för just detta ändamål som utförs med lämplig periodicitet.

Processen att ta fram indikatorer är i sig central i kvalitetsarbetet. Den ger en möjlighet att tänka igenom vad data står för och hur de ska användas vidare. För en diskussion om metoder för indikatorer, deras betydelse och användning, se Radermacher (2020).

De nya typerna av data utgör inte en homogen grupp av datakällor. De kan vara av väldigt olika karaktär, från data som är väl beskrivna och väldigt lika de administrativa källor som redan

används, till data som är högst ostrukturerade och med bristfälliga



1. Social Networks (human-sourced information): this information is the record of human experiences, previously recorded in books and works of art, and later in photographs, audio and video. Human-sourced information is now almost entirely digitized and stored everywhere from personal computers to social networks. Data are loosely structured and often ungoverned.

1100. Social Networks: Facebook, Twitter, Tumblr etc.
1200. Blogs and comments
1300. Personal documents
1400. Pictures: Instagram, Flickr, Picasa etc.
1500. Videos: Youtube etc.
1600. Internet searches
1700. Mobile data content: text messages
1800. User-generated maps
1900. E-Mail

2. Traditional Business systems (process-mediated data): these processes record and monitor business events of interest, such as registering a customer, manufacturing a product, taking an order, etc. The process-mediated data thus collected is highly structured and includes transactions,reference tables and relationships, as well as the metadata that sets its context. Traditional business data is the vast majority of what IT managed and processed, in both operational and BI systems. Usually structured and stored in relational database systems. (Some sources belonging to this class may fall into the category of "Administrative data").

21. Data produced by Public Agencies
2110. Medical records
22. Data produced by businesses
2210. Commercial transactions
2220. Banking/stock records
2230. E-commerce
2240. Credit cards

3. Internet of Things (machine-generated data): derived from the phenomenal growth in the number of sensors and machines used to measure and record the events and situations in the physical world. The output of these sensors is machine-generated data, and from simple sensor records to complex computer logs, it is well structured. As sensors proliferate and data volumes grow, it is becoming an increasingly important component of the information stored and processed by many businesses. Its well-structured nature is suitable for computer processing, but its size and speed is beyond traditional approaches.

31. Data from sensors
311. Fixed sensors
3111. Home automation
3112. Weather/pollution sensors
3113. Traffic sensors/webcam
3114. Scientific sensors
3115. Security/surveillance videos/images
312. Mobile sensors (tracking)
3121. Mobile phone location
3122. Cars
3123. Satellite images
32. Data from computer systems
3210. Logs
3220. Web logs

metadata. Nya datakällor kan även vara data från administrativa register som inte funnits eller inte använts av statistikproducenten tidigare, till exempel arbetsgivardelarationer på individnivå (AGI). Avgränsningen mot administrativa data är inte tydlig, det är snarare en glidande skala där administrativa data ligger i den ena änden och till exempel data från mobiloperatörer i den andra. Andra exempel på nya datakällor som SCB undersökt på senare tid är webbaserade portaler för annonser om lediga jobb och digital mätning av elförbrukning och -produktion.

UNECE (2014) grupperar det man kallar för big data i tre olika kategorier enligt Tabell 1.

Tabell 1. Klassificering enligt UNECE (2014)

De olika typerna av källor gör det svårt att sätta ett gemensamt namn på de data och datakällor som avses i denna rapport. Big data, nya datakällor, organiska data , found data, digitala data eller it-is-what-it-is data är några förslag som förekommit de senaste åren. Här kallas de för digitala data (National Academies of Sciences, Engineering, and Medicine 2022). Det avser data som

genereras digitalt som en del av någon process men som har ett annat huvudsakligt syfte än data för statistik. Namnet i sig är inte avgörande och avsikten är endast att inte behöva tynga rapporten med omständliga formuleringar.

Denna rapport är en del av resultatet av arbetet med både regeringsuppdraget och SCB:s egna behov. Rapporten syftar till att sätta de föreslagna indikatorerna i ett sammanhang och ge en bakgrund till den mer detaljerade beskrivningen av indikatorerna i Kvalitetskriterier för statistik baserad på digitala data - vägledning (SCB 2023). I första delen av rapporten relateras arbetet till relevanta arbeten internationellt, nationellt och internt inom SCB. I den andra delen beskrivs översiktligt den föreslagna strukturen för indikatorerna.

Beskrivningar av kvalitet

Kvalitetsbegreppet och digitala data

Kvalitetsbegreppet är centralt för den officiella statistiken. Bestämmelser om kvalitet i den officiella statistiken föreskrivs i SCB:s föreskrift om den officiella statistiken (SCB-FS 2016:17), med stöd av16 § 2 förordningen (2001:100) om den officiella statistiken. Kvaliteten i en slutprodukt ska beskrivas i en kvalitetsdeklaration med utgångspunkt i de fem huvudkomponenterna Relevans, Tillförlitlighet, Aktualitet och punktlighet, Tillgänglighet och tydlighet samt Jämförbarhet och samanvändbarhet. Som stöd för att dokumentera statistiken finns en handbok (SCB 2020).

En motsvarande mall och handbok finns även för dokumentation av framställningen och kvaliteten i statistiska register, DOKSTAR (SCB 2019). I framställningen av ett statistiskt register är slutprodukten inte statistik, utan ett slutligt observationsregister, det vill säga ett register som innehåller alla data för att framställa den planerade statistiken.

De indikatorer som föreslås för digitala data är avsedda att, vid behov, komplettera den redan befintliga kvalitetsdokumentationen, oavsett om slutprodukten är ett statistiskt register eller färdig statistik. Alla kvalitetskomponenter är relevanta, men alla påverkas inte av att datakällan är en annan typ än de traditionella, i meningen att det behövs kompletterande indikatorer. Framför allt gäller det, som nämndes i inledningen, i de avslutande stegen av produktionen.

Det är kvalitetskomponenterna Relevans och Tillförlitlighet som är i fokus, då beskrivningar av de övriga komponenterna inte behöver kompletteras med någon ny indikator. I Tillförlitlighet ingår osäkerhetskällorna urval, ramtäckning, mätning, bortfall, bearbetning och modellantaganden. Till stor del är samma osäkerhetskällor relevanta för digitala datakällor, men det finns skillnader framför allt jämfört med urvalsundersökningar men även jämfört med administrativa data. Täckningsproblematiken är högst relevant, liksom mätning, bearbetning och modeller.

Alla indikatorer är inte nödvändigtvis relevanta för alla datakällor eller användningar, så de presenterade indikatorerna ska ses som

en bruttolista. Indikatorerna ska kunna fungera för att utvärdera och beskriva kvaliteten innan en datakälla tas i produktion, och även när den är i produktion.

När den digitala datakällan skapas finns (eventuellt) ett syfte med att generera data, eller så är data en biprodukt av någon annan process, men det finns troligen inget statistiskt syfte som stämmer med statistikproducentens syfte. Det bestäms av statistikproducenten utifrån användarnas behov och behöver preciseras för att kunna utvärdera datakällan innan den tas i produktion. Det finns samtidigt en uttalad målsättning att data som SCB tar in ska ha en bred användning. Om det finns flera önskvärda statistiska syften så kan det vara nödvändigt att göra mer än en utvärdering av vissa delar.

Det kan finnas stora inslag av bearbetning och modellering när digitala data hanteras, till exempel för att imputera värden, för att länka mellan digitala data och befintliga register, eller för att koda text. Ordet modell förekommer i flera olika betydelser. Alla modeller orsakar osäkerhet i de slutliga skattningarna. En statistisk modell kan ligga till grund för till exempel bearbetning av data, det kan handla om modeller för kodning, textanalys, imputering, outlierhantering eller länkning. För stora datamängder kan det handla om modeller för maskininlärning. Andra typer av modeller är fördelningsmodeller baserade på ämneskunskap om till exempel företag, arbetsmarknaden eller elmarknaden. Modell kan även syfta på en analysmodell där samband analyseras och förklaras, eller en regressionsmodell för skattningar.

Andra arbeten på SCB

Det pågår andra arbeten på SCB som berör dokumentation av datakällor eller processer för att effektivare utnyttja data. Rapporten Mätteknik 2.0 beskriver hur det mättekniska arbetet ska fokuseras och bedrivas framöver när datakällor där inte statistikproducenterna själva designat den datagenererande processen blir viktigare i statistikproduktionen. För datakällor där statistikproducenten inte har kontroll över hur data genereras så är det viktigt att försöka få en uppfattning om i vilket syfte data sparas och hur det går till, innan data hamnar hos statistikproducenten. Kunskap om detta är underlag för till exempel möjligheten att möta användarnas krav på statistiken och förståelsen av mätfel.

I SCB:s löpande produktion lanseras begreppen Berett observationsregister (BOR) och slutligt observationsregister (SOR). De avser olika hållpunkter (HP), definierade som avstämningspunkter i statistikproduktionsprocessen där datamängder finns beskrivna:

- (HP0) Indata
- (HP1) Tillrättalagda indata
- (HP2) Berett observationsregister (BOR)
- (HP3) Slutligt observationsregister (SOR)
- (HP4) Statistikvärden

Ett arbete är även gjort med att ta fram en mall för att dokumentera ett BOR. Ett BOR kan baseras på en eller flera externa datakällor som integreras och vidarebearbetas. Ett arbete är även gjort med att ta fram en mall för att dokumentera ett BOR. Dokumentationen utgör ett levande dokument till skillnad från Kvalitetsdeklarationen och DOKSTAR som avser en årgång.

Hur de föreslagna indikatorerna relaterar till hållpunkterna behöver utredas vidare, men är inte en del av föreliggande uppdrag.

Kriterier för kvalitet i litteraturen

Kvalitet i undersökningar beskrivs ofta utifrån Total Survey Error (TSE), ett ramverk för totalfel som beskriver osäkerhetskällor i direktinsamlade data (se till exempel Groves och Lyberg 2010). Ramverket beskriver två dimensioner relevanta för kvaliteten i de slutliga skattningarna, mätning och representation. Mätning avser observationer och värden, medan representation avser objekt och populationer. I SCB (2016) finns en svensk översättning och anpassning till SCB:s begreppsapparat.

Kvalitetsramverk och -indikatorer för tillförlitligheten i integrerade, administrativa och digitala datakällor har föreslagits i flera tidigare arbeten. Alla bygger på och utvecklar TSE, vissa enbart för en enskild datakälla och vissa för integrerade data. Några exempel tas upp nedan.

Amaya et al (2020) presenterar ett ramverk som syftar till att identifiera, beskriva och förstå osäkerhetskällor i både strukturerade och ostrukturerade data. Det är en generalisering av tidigare arbete (Biemer 2016) där den traditionella TSE-modellen bearbetats för att kunna appliceras på alla typer av strukturerade

data. Urvalsundersökningar och administrativa data betraktas som strukturerade, medan de digitala datakällorna kan vara ostrukturerade.

Sen et al (2021) går den motsatta vägen och har tagit fram ett ramverk som är specifikt för en viss typ av digitala data; data från sociala medier. Hurtado Bodell et al (2022) är också ett exempel på ett ramverk för en specifik typ av data. Författarna föreslår ett konceptuellt ramverk för att värdera kvaliteten i textdata i tre dimensioner som de benämner total corpus error, corpus comparability, och corpus reproducibility. Den första dimensionen avser tillförlitligheten i skattningar baserade på textdata.

De ovan nämnda exemplen tar inte specifikt upp integrering av datakällor. Administrativa data behöver i regel alltid integreras med andra data, och detsamma gäller för digitala datakällor.

Laitila et al (2011) diskuterar problemet att undersöka kvalitet i ett register som är avsett för bred användning. Kvalitet ska bedömas i relation till användningen och kvaliteten i den slutliga statistiken, men författarna menar att det inte är möjligt för ett register med bred användning. Författarna skriver ”Quality assessment of a register should instead focus on available information on the administrative register and on information that is based on a systematic analysis of the administrative source” (sidan 9). De föreslår indikatorer när den administrativa källan integreras antingen med ett basregister eller med andra undersökningar.

ESSnet-projektet Quality of multisource statistics (KOMUSO) tog fram ett ramverk för kvalitet i statistik med flera källor, och speciellt administrativa källor (Brancato et al 2019). Det som författarna benämner som big data eller administrativa data med big data-egenskaper omfattas inte av ramverket. I De Waal et al (2019) beskrivs i detalj hur kvalitetsmåtten som projektet tog fram ska beräknas. Ett stort antal mått föreslås, och det grundläggande är att beräkna MSE, dvs varians och bias. Många av måtten kräver speciella studier.

De Waal et al (2020) utvecklar vidare riktlinjer för statistik med flera källor och inkluderar även nya datakällor. Artikeln presenterar inte ett allomfattande ramverk utan ger riktlinjer för olika typer av kombinationer av datakällor som kan uppstå och föreslår metoder för hur de kan hanteras. Författarna diskuterar speciellt åtta typer av kombinationer som de anser är de mest vanligt förekommande.

Projektet ESSnet Big Data II tog fram kvalitetsriktlinjer specifikt för nya datakällor (Quaresma et al 2020). Skillnader mot ramverket som togs fram i KOMUSO motiveras bland annat med att användningen av administrativa data är betydligt mer mogen och därför kan fokusera på kvalitetsmått för färdig statistik, medan nya datakällor fortfarande måste fokusera mest på indata och processen fram till ett strukturerat register. En utmaning var att formulera kriterier som är generella nog att vara relevanta för alla typer av nya data eftersom dessa kan vara väldigt olika. Daas et al (2020) ger en översikt över metoder för big data och knyter an till kvalitetsriktlinjerna från projektet.

Gootzen et al (2022) konstaterar också att kvalitetsramverk avsedda för urvalsdata och administrativa data inte räcker för nya datakällor. Målsättningen i deras arbete är att konstruera ett ramverk för kvalitet för alla tre typer av datakällor. Ramverket ger inte kvantitativa kriterier utan består av klassificeringar i olika dimensioner och kategorier.

Zhang (2012) gör en anpassning av TSE-ramverket för administrativa data och delar upp beskrivningen av kvaliteten i två faser. I första fasen beskrivs osäkerhetskällor för en enskild datamängd. I den andra fasen beskriver osäkerhetskällor som kan uppstå när två eller flera datamängder integreras. Integrering kan vara att samma eller överlappande objektsmängder läggs ihop i syfte att utöka antalet variabler eller för att skapa nya statistiska objekt. En annan situation är när likadana variabelmängder för ej överlappande objektsmängder kommer från olika leverantörer.

I båda faserna i Zhang (2012) finns en del som avser objekt (Representation) och en del som avser variabler (Measurement). Under insamling och bearbetning av variabler och objekt kan fel uppstå i båda faserna. Dessa behöver beskrivas och mätas för att ge en uppfattning om den totala kvaliteten i den integrerade datamängden.

Reid et al (2017) tillämpar ramverket enligt Zhang (2012) och lägger till en tredje fas där den slutliga kvaliteten i skattningar färdiga för publicering beskrivs. Det är i den tredje fasen som till exempel viktberäkningar, punktskattningar, förändringsskattningar, variansskattningar och säsongrensning görs. Den tredje fasen är i mycket hög grad beroende av specifika användarbehov. Ramverket är implementerat på statistikbyrån i Nya Zealand (Statistics New Zealand 2016).

Lothian et al (2019) utvidgar ramverket ytterligare för att även beskriva användningen av vad de kallar ”it-is-what-it-is”-data, det vill säga data där statistikproducenten inte har kontroll över insamlingsprocessen. För att utvärdera representativitet behöver data relateras till en målpopulation. En stor del av Lothian et al (2019) handlar om att i detta syfte skapa basregister.

Slutligen kan nämnas en rapport som tagits fram inom ett initiativ drivet av UNECE HLG-MOS (UNECE 2021). Här presenteras bland annat ett kvalitetsramverk för att utvärdera hur algoritmer presterar när maskininlärning används för produktion av officiell statistik. Rapporten är även en bra introduktion till maskininlärning generellt, med många relevanta exempel.

Kvalitet i digitala data

Inledning

I detta avsnitt diskuteras begrepp och relevanta osäkerhetskällor.

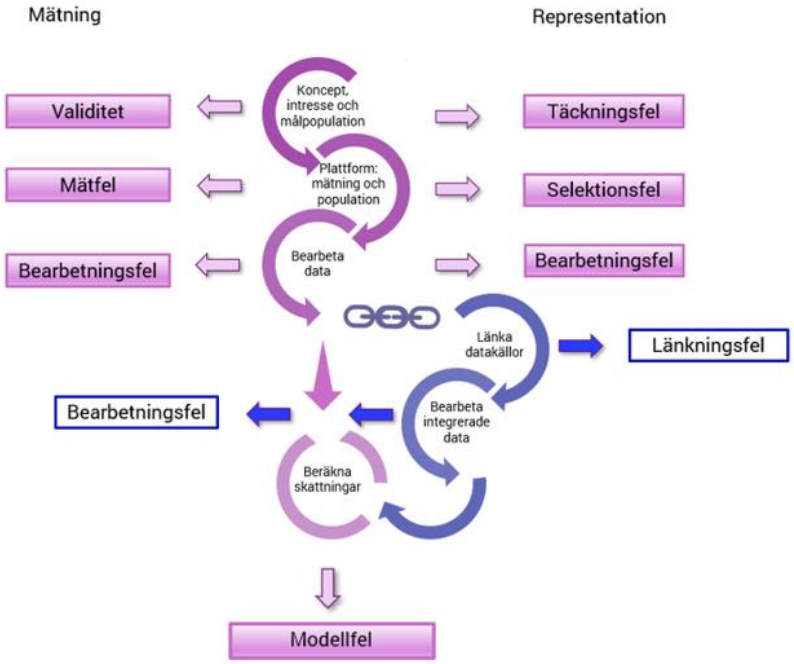
Utgångspunkten är sammanfattningsvis följande:

- Tillförlitlighet enligt Kvalitetsbegreppet samt relationen till TSE är grundläggande även för statistik baserad på digitala data.
- Integrering av digitala datakällor med befintliga register är centralt.
- De förslag som presenteras i uppdraget kompletterar den fastställda kvalitetsdokumentationen för statistik baserad på urvals- och administrativa data.

Beskrivningen av felkällor i digitala data (se Figur 1) är inspirerad av ramverk för TSE men också av Zhang (2012), Reid et al (2017) och Lothian et al (2019) där integrering av data är centralt, men beskrivningen här är medvetet förenklad. Ett begrepp är lånat från Sen et al (2022) men används i en mer generell mening än enbart data från sociala medier. Så långt som möjligt är begrepp och definitioner anpassade till det språkbruk som SCB använder, och avvikelser från det är tänkt som kompletteringar.

Utgångspunkten är att det finns ett koncept som avser något användarna är intresserade av att veta och som statistikproducenten vill mäta, och en population som motsvarar detta intresse. Operationaliseringen, det vill säga processen att göra om konceptet till något mätbart, hanteras i dimensionen Mätning. Motsvarande process att specificera en population som går att observera hanteras av dimensionen Representation.

Figur 1. Potentiella felkällor i digitala data



Figur 1 sammanfattar steg i de två dimensionerna och deras osäkerhetskällor. Figuren visar en process som börjar med en enskild digital datakälla och som kan sluta i någon skattning baserad enbart på den digitala källan. Mer troligt är dock att data integreras (länkas) med data från någon annan källa, till exempel ett basregister. Länkningen ger integrerade data som används för att skatta målstorheter. Detaljer och begrepp diskuteras vidare nedan.

Representation

Intressepopulation är en population av objekt som motsvarar det användarna är intresserade av. Det kan till exempel vara företag, hushåll eller personer, eller delpopulationer av dessa. Målpopulationen är den population som statistikproducenten valt att undersöka och dra slutsatser om. Målpopulationen kan stämma överens med intressepopulationen, men det kan finnas skillnader, speciellt för datakällor där statistikproducenten inte alls eller bara delvis råder över designen för datainsamlingen.

För en urvalsundersökning krävs en rampopulation som urvalet dras från. Skillnaden mellan mål- och rampopulation ger över- eller undertäckning. I framställningen av ett statistiskt register kan det finnas ett ramförfarande om det statistiska registret i sin tur är baserat på ett befintligt statistiskt register (SCB 2019), till exempel om en delpopulation från Registret över

totalbefolkningen (RTB) vidarebearbetas till ett eget statistiskt register. Objekten i en digital datakälla kan inte alltid tydligt avgränsas som en ram vid ett givet tillfälle, men det kan gå att avgränsa objektens möjlighet att generera data via en operatör, till exempel en mobiloperatör eller en elnätsoperatör, eller en plattform, till exempel en portal för annonser om lediga jobb. Plattformens (eller operatörens) population kan ses som en motsvarighet till rampopulation (ordet plattform är lånat från Sen et al 2022).

När data från två källor, till exempel en digital källa och ett basregister, integreras är målpopulationen den integrerade mängden objekt, och rampopulationen består både av objekten i basregistret och objekten från plattformen. Registrets och plattformens objekt kan vara av olika typ, men via integreringen av data skapas ett statistiskt register med endast en typ av objekt.

Skillnaden mellan intressepopulationen och den integrerade mängden observerade objekt påverkas av flera delar.

- Skillnaden mellan intressepopulationen och målpopulationen är i första hand teoretisk och inte mätbar utan en speciellt designad undersökning. Med källor där statistikproducenten i liten utsträckning kan påverka datainsamlingen så finns det en risk att denna skillnad är betydande.
- Skillnaden mellan målpopulationen och rampopulationen ger över- eller undertäckning. Både täckningen i basregistret och täckningen i den digitala datakällan bidrar.
- Täckningen skattas genom länkning av datakällorna. Skillnad mellan rampopulationen och den observerad mängden ger selektionsfel. Målsättningen är att observera alla objekt i ramen men det kan förkomma att kända element i ramen inte kan observeras. Slumpmässigt urval kan också förekomma och beskrivs i så fall separat.

Plattformspopulationen/ramen kan bestå av en eller flera plattformars eller operatörers användare/kunder, till exempel flera elnätbolag, flera mobiloperatörer eller annonser som levereras från fler än en jobbportal. Om det är möjligt görs ett arbete för att standardisera leveranser så mycket som möjligt, och automatiska kontroller av format och liknande görs vid leveranser. Om det fortfarande finns olikheter mellan data levererat av olika leverantörer så är det viktigt att dokumentera det. Om plattformspopulationens användar- eller kundbas inte är

representativ för den målpopulation som undersöks, till exempel om data inte är tillgängliga från alla mobiloperatörer eller om före detta kunder finns registrerade, så bidrar det till under- respektive övertäckning.

Länkning

Det är troligt att de flesta digitala datakällor av intresse för SCB kommer behöva en koppling till något befintligt register, och ofta ett basregister. Behovet av länkning beror på vilken population det är av intresse att göra inferens till. Om inferensen till exempel endast avser mobilanvändare från en viss operatör så kan relevant statistik skattas med data från endast denna operatör. Om inferensen avser mobilanvändning hos Sveriges befolkning (eller en del av befolkningen) så behövs både representativa data från en eller flera operatörer och en koppling till ett register över populationen Sveriges befolkning.

Den digitala källan och det befintliga registret kan ha olika typer av objekt, till exempel kan plattformens objekt vara mätpunkter, sms eller webbannonser medan basregistrets objekt är personer, företag eller fastigheter. Länkningen görs då troligen i flera steg för att skapa en integrerad datamängd med samma typ av objekt, och med mätvärden som aggregeras i ett eller flera steg till de objekt som finns i den integrerade mängden.

För att länka till ett befintligt register behövs relevanta länkningsvariabler. Helst ska dessa variabler vara unika identifierare i de digitala data och återfinnas i befintliga register. Det är den situation som råder i survey- och (oftast) administrativa data. De unika identifierarna behöver inte alltid identifiera objekt unikt för att möjliggöra en länkning, till exempel har SCB testat att länka mobildata aggregerat på geografiska områden till befolkningstotaler på motsvarande områden via RTB.

När digitala data länkas till ett befintligt register så är det inte säkert att det går att länka alla objekt i registret till objekt i de digitala data. Det kan finnas olika skäl till att länkning inte fungerar. Länkningsinformation kan saknas i den digitala datakällan eller i det befintliga registret. Då krävs i stället en modell för att länka datakällor. En modell är även nödvändig om kopplingen mellan datakällor inte görs på objektsnivå.

Länkning med unikt identifierande variabler i två (eller fler) datakällor som ska integreras kallas deterministisk länkning. Om det finns icke unika identifierande variabler i källorna så kan metoder för probabilistisk länkning eller

maskininlärningsmodeller för länkning användas. En annan form av integrering är så kallad statistisk matchning. Då integreras datakällor som innehåller olika objekt. Det kan ske på mikro- eller makronivå (De Waal et al 2020).

Mätning

I en undersökning fångas användarnas behov och krav genom intressevariabler och utifrån dessa formuleras målvariabler. Utifrån målvariablerna formuleras frågor. De observerade variabelvärdena (svaren) kan avvika systematiskt från målvariablerna.

För digitala data kan det vara relevant att utgå från ett koncept som en benämning på det som användarna är intresserade av, om det inte direkt går att formulera intresset i konkreta variabler. Det beror på datakällans innehåll och struktur. Ett koncept avser något som är mer abstrakt än en eller flera intressevariabler. Konceptet behöver först definieras för att möta användarnas intresse och sedan operationaliseras genom en eller flera mätbara målvariabler (se till exempel Persson 2016 eller SCB 2016). Eftersom innehållet i digitala data inte påverkas av statistikproducenten så är det inte alltid en tydlig process från koncept eller intresse till observerade variabler. Utgångspunkten kan snarare vara vilka data som finns tillgängliga och hur dessa kan passa med ett tänkt syfte. I litteraturen (se till exempel Persson 2016, SCB 2016 eller Hox 1997) skiljer man på ett teoridrivet respektive ett empiriskt eller datadrivet angreppssätt.

Det kan hända att operationaliseringen inte fångar hela det önskvärda konceptet, vilket ger problem med (begrepps)validiteten. Säg till exempel att konceptet vakans operationaliseras genom att samla in annonser om lediga jobb via portaler. Det går dock inte alltid i annonser att särskilja vakanser från andra lediga jobb, som det gör genom att ställa frågor till arbetsgivare. I data från jobbportaler kommer det därför att ingå tjänster som inte är vakanser, till exempel vikariat som inte ska tillträdas omedelbart, och det koncept som fångas är lediga jobb, inte vakanser.

Det behöver inte vara ett problem att inte hela konceptet eller intresset fångas med endast en datakälla. Den framtida statikproduktionen förväntas bygga mer och mer på integrering av olika datakällor (De Waal et al 2020). Om flera källor tillsammans förväntas täcka hela konceptet så är det viktigt att för varje datakälla veta hur eventuella gap mellan koncept och mätbara variabler ser ut, samt vilka andra källor som kan täcka det.

Mätfel kan uppstå redan innan digitala data hämtas in, det vill säga datakällan har misslyckats med att fånga den information den syftar till eller innehåller felaktiga värden. Statistikproducenten kan oftast inte påverka hur observationer har genererats, men kan i vissa fall påverka vilka data som hämtas in. I fallet med platsannonser så behövs till exempel en algoritm som väljer annonser baserat på ord eller textdelar och därmed påverkas vilka data som hämtas in. Om alla befintliga data tankas direkt från tekniska system så finns däremot ingen eller väldigt liten sådan påverkan. Mätvärden kan behöva bearbetas innan (bearbetade) värden integreras med andra data eller målstorheter skattas. Text som extraheras ur annonser måste till exempel kodas till relevanta kategorier som går att använda för statistik. Konceptet elförbrukning kvantifieras som elförbrukning, och den mäts genom att förbrukningen läses av direkt från elmätare. Det är inte ett mätförfarande som SCB kan påverka eller designa, och risker för mätfel finns men är inte så stora.

När digitala data integreras med en annan datakälla (till exempel ett basregister) så sker ingen ytterligare mätning. Däremot kan det vara nödvändigt med bearbetning för att harmonisera med registret, härleda nya variabler, imputera, koda eller göra något annat enligt någon modell.

Skattning

I det slutliga skattningssteget går de två dimensionerna representation respektive mätning ihop.

Skattningar kan beräknas direkt på digitala data eller på integrerade data, och det troligt är att de flesta digitala datakällor som SCB vill använda kommer behöva integreras med befintliga register. Skattningar kan kräva modeller för att till exempel justera för bias eller säsongrensa tidserier. I detta steg finns inga kvalitetskrav som är specifika för digitala data och kvalitetsbeskrivningen behöver inte kompletteras med någon ny indikator.

Två exempel

Nedan beskrivs kortfattat två exempel på datakällor som SCB undersökt på senare år och som kan sägas representerar varsin ände på den glidande skalan över digitala data.

Med data från mobiloperatörer så vill SCB mäta var människor befinner sig under dagen och/eller under natten. Det stämmer inte med koncepten dag- respektive nattbefolkning, då dagbefolkningen avser var människor arbetar medan

nattbefolkningen avser var människor bor. Intressepopulationen är Sveriges befolkning och den rampopulation som SCB normalt definierar då är den folkbokförda befolkningen men mobildata hänförs snarare till en rörlig population. SCB får data från två leverantörer, Telia och Tre. De täcker tillsammans inte alla mobilanvändare i Sverige. Levererade data är aggregerade till geografiska områden och SCB rår inte över den bearbetning som görs innan data levereras och har inte heller kunskap om hur den görs exakt. Målpopulationen skiljer sig även från intressepopulationen genom att barn under 7 år inte antas ha egna mobiler. Det går inte att länka objekt på individnivå, men det går att integrera aggregerade data på geografiska områden, det vill säga koppla aggregerade mobiltelefondata till aggregerade data från den registrerade befolkningen vid en viss tidpunkt.

Två modeller har tagits fram för att skatta antalet arbetade timmar i svensk ekonomi (Lennartsson och Gullberg Brännström 2022). Individer i AGI kopplas ihop med företag i FDB. Målpopulationen består av individer och företag som går att matcha genom information om vilka individer som är anställda på vilket företag, samt referensperiod. Data från AGI, FDB och konjunkturlönestatistiken (KL) kombineras sedan varvid designbaserade skattningar tas fram (som input till modellerna) för KL:s målpopulation. Under olika modellantaganden (för att bland annat hantera KLP:s cut-off) appliceras dessa designbaserade skattningar på AGI:s målpopulation (som är större än KL:s) och modellbaserade skattningar tas fram. När data integreras så påverkas osäkerheten i de slutliga skattningarna av flera olika saker, till exempel olika målpopulationer och värden skattade i tidigare led.

Diskussion

Uppdraget har varit att ta fram kvalitetskriterier för digitala data som kan användas för att bedöma datakvaliteten och utvärdera om data är lämpliga för statistikframställning. Ordet kriterier kan ge intrycket att det finns absoluta svar på vad som är tillräckligt bra data, vilket inte är fallet. I stället har arbetet fokuserat på indikatorer, och i möjligaste mån mätbara indikatorer. Dessa indikatorer ger en uppfattning om svagheter och felkällor som kan påverka de slutliga skattningarna och är tänkta att fungera som en del av ett underlag för en utvärdering av om datakällan kan användas i produktionen.

Det viktigaste första steget är att göra en kvalitativ utvärdering av den digitala datakällan för att förstå exakt vad data innehåller. En sådan utvärdering kan ge viktig information om vilka andra felkällor som behöver studeras baserat på användarbehov. För vissa digitala data kan det stora problemet vara täckningen medan andra typer av digitala data kan ha problem med både täckning, validitet och mätfel. Det är svårt att ge mer generella rekommendationer för hur indikatorer ska bedömas då både digitala data och användarbehov kan se väldigt olika ut.

De föreslagna indikatorerna kan kräva speciella utvärderingsstudier av till exempel mätfel, men det finns även indikatorer som är relativt enkla att ta fram. Om datakällan så småningom kommer att ingå i statistikproduktionen så kan de enklare indikatorerna beräknas löpande i varje produktionsomgång. Genom att följa hur indikatorerna uppträder över tid kan det vara möjligt att få en indikation på när nya utvärderingsstudier behöver göras eller om modeller behöver ses över. I SCB (2023) markeras indikatorer som kan beräknas löpande.

Olika aspekter av kvalitet behöver alltid vägas samman för att avgöra om en datakälla kan användas, oavsett typ av data. Kvalitetsindikatorer och -beskrivningar ger några aspekt, medan andra aspekter är uppgiftslämnarbörda och kostnader.

Användarnas krav på vad statistiken ska visa har stor betydelse, till exempel hur finfördelad redovisningen ska vara eller om det är nivå- eller förändringsskattningar som är viktigast. Data kan vara bra för en användning men sämre för en annan, och syftet kan

vara en specifik användning eller användning för så många användningsområden som möjligt. Det spelar också roll hur statistikproducenten tänkt använda data, till exempel om det är direkt i skattningar, som hjälpinformation i design, för att ta fram helt ny statistik eller för att förbättra befintlig statistik.

Dessa aspekter är inte specifika för användningen av digitala data, men olika aspekter kan vara mer eller mindre viktiga beroende på datakälla. Ofta anges kortare produktionstid och minskad uppgiftslämnarbörda som viktiga orsaker till att använda digitala data. Samtidigt så finns andra aspekter som behöver hanteras med digitala data. Hållbarheten över tid kan vara en riskfaktor där både tillgången och reliabiliteten över tid kan påverkas. Statistikproducenten har inte kontroll över om dataägaren till exempel gör ändringar som betyder att datagenereringen eller kvaliteten påverkas, väljer att lägga ner delar av sin verksamhet eller bestämmer sig för att sätta ett pris på data. Om användarnas krav ändras så har statistikproducenten inga eller begränsade möjligheter att justera datakällan efter det.

Dataägaren och statistikproducenten behöver ha ett tydligt och detaljerat avtal. Det pågår även arbete med lagförslag på EU-nivå som har till syfte att underlätta användning och delning av data och som inkluderar privata dataägare.

Allt aspekter som nämns ovan måste gå in i en samlad bedömning av om och i så fall hur digitala data kan användas. Utvärderingen av kvaliteten i en digital datakälla är en del av SCB:s godkännandeprocess för nya datakällor. Denna process beskrivs i rapporten Godkännandeprocessen för nya datakällor 1.0. I steg tre (av fem), en grundlig undersökning av data, ingår bedömning av datakvalitet och analys av testdata. I steg 1 av processen görs en initial översiktlig bedömning och i steg 2 ses eventuella juridiska begränsningar över, kontakter knyts med dataägare och det görs en viss specificering av innehållet i data. I steg 4 tas beslut ifall datakällan ska tas in, baserat på en helhetsbedömning av nytta och risker och i steg 5 skrivs avtal med dataägare.

Det pågår, som litteraturgenomgången visar, en hel del internationellt arbete, och det är därför viktigt med en fortsatt omvärldsbevakning och att vid behov ompröva de här föreslagna indikatorerna. Det är viktigt att samordna arbetet med annan utveckling av kvalitetsarbetet på SCB, till exempel Mätteknik 2.0 och dokumentation av BOR, samt med redan fastställda dokumentationsmallar som Kvalitetsdeklaration och DOKSTAR.

Den framtida statistikproduktionen kommer bygga mycket på att integrera flera datakällor (se till exempel De Waal et al 2020). Detta ligger i linje med det SCB benämner Statistikproduktion 4.0, där den framtida statistikproduktionen förväntas utgå från befintliga data, i register eller andra källor, som kompletteras med direktinsamling vid behov. Ett gemensamt ramverk för urvalsdata, administrativa data och digitala data skulle vara önskvärt.

Referenser

Amaya, A., Biemer, P. P. & Kinyon, D. (2020). Total error in a big data world: adapting the TSE framework to big data. Journal of Survey Statistics and Methodology 8:1, pp. 89-119. https://doi.org/10.1093/jssam/smz056

Biemer, P., P. (2016). Errors and Inference, in Big Data and Social Science: A Practical Guide to Methods and Tools, eds. I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, and J. Lane, pp. 265–297, Boca Raton: CRC Press.

Brancato, G., Ascari, G., Krapavickaite, D., Alexander, P.J. & Waldner, C. (2019). Eurostat ESSnet KOMUSO Quality in multisource statistics, Work package 1, Quality guidelines for multisource statistics (QGMSS) qgmss-v1.1_1.pdf (europa.eu)

Daas, P, Maslankowski, J., Salgado, D., Quaresma, S., Tuotu, T., Di Consiglio, L., Brancato, G., Righi, P., Six, M. & Kowarik, A. (2020). Eurostat ESSnet Big Data II Work package K, Methodology and quality, Deliverable K9: Revised version of the methodological report. WPK_Deliverable_K9_Revised_version_of_the_methodological_report_2020_11_17_Final.pdf (europa.eu)

De Waal, T., Van Delden, A. & Scholtus, S. (2019). Eurostat ESSnet KOMUSO Quality in multisource statistics, Quality measures and indicators, Complete Overview of Quality Measures and Calculation Methods (QMCMs) qmcms_examples_overview_1.pdf (europa.eu)

De Waal, T., Van Delden, A. & Scholtus, S. (2020). Multi-source statistics: Basic situations and methods. International Statistical Review, 88:1, pp 203-228. https://doi.org/10.1111/insr.12352

Gootzen, Y., Daas, P. & Van Delden, A. (2022). Quality Framework for combining survey, administrative and big data for official statistics. Paper presented at Q2022, Vilnius. (PDF) Quality Framework for combining survey, administrative and big data for official statistics (researchgate.net)

Groves, R. M. & Lyberg L. (2010). Total Survey Error: Past, present, and future. Public Opinion Quarterly 74:5, pp 849-879. DOI: https://doi.org/10.1093/poq/nfq065

Hox, J. J. (1997). From Theoretical Concept to Survey Question. In Survey Measurement and Process Quality, Lyberg et al (red). Wiley.

Hurtado Bodell, M., Magnusson, M., & Mützel, S. (2022). From Documents to Data: A Framework for Total Corpus Quality. Accepterad för publicering i Socius. https://osf.io/preprints/socarxiv/ft84u/

Laitila, T., Wallgren, A. & Wallgren, B. (2011). Quality assessment of administrative data. R&D Methodology Reports, Statistics Sweden. i9426en.pdf (fao.org)

Lothian, J., Holmberg, A. & Seyb, A. (2019). An evolutionary schema for using "it-is-what-it-is" data in official statistics. Journal of Official Statistics 35:1, pp 137-165. DOI: https://doi.org/10.2478/jos-2019-0007

Lennartsson, D., & Gullberg Brännström, S. (2022). Model estimation of number of hours worked. Paper presented at Nordic Statistical Meeting 2022, Reykjavik.

National Academies of Sciences, Engineering, and Medicine (2022). Toward a 21st Century National Data Infrastructure: Mobilizing Information for the Common Good. Washington, DC: The National Academies Press. https://doi.org/10.17226/26688

Persson, A. (red) (2016). Frågor och svar, om frågekonstruktion i enkät- och intervjuundersökningar. Statistiska centralbyrån.

Quaresma, S., Maslankowski, J., Salgado, D., Ascari, G., Brancato, G., Di Consiglio, L., Righi, P., Tuotu, T., Daas, P., Six, M. & Kowarik, A. (2020). Eurostat ESSnet Big Data II Work package K, Methodology and quality, Deliverable K3: Revised version of the quality guidelines for the acquisition and usage of big data. WP3_Deliverable_K3_Revised_Version_of_the_Quality_Guidelines_for_the_Acquisition_and_Usage_of_Big_Data_Final_version.pdf (europa.eu)

Radermacher, W. J. (2020). Official Statistics 4.0. Verified facts for people in the 21st century. Springer.

Regeringen (2021). Uppdrag att främja delning och nyttiggörande av data för smart statistik. Uppdrag att främja delning och nyttiggörande av data för smart statistik - Regeringen.se

Reid, G., Zabala, F. & Holmberg, A. (2017). Extending TSE to administrative data: A quality framework and case studies from Stats NZ. Journal of Official Statistics 33:2, pp 477-511. DOI: https://doi.org/10.1515/jos-2017-0023

SCB (2023) Kvalitetskriterier för statistik baserad på digitala data - vägledning.

SCB (2020) Kvalitet för den officiella statistiken - en handbok. Version 2:2. Statistiska centralbyrån. Kvalitet för den officiella statistiken – en handbok, version 2:2 (scb.se)

SCB (2019) Det statistiska registrets framställning och kvalitet – en handbok.

SCB (2016) Att utforma och förbättra en statistisk undersökning.

Sen, I., Flöck, F., Weller, K., Weiss, B. & Wagner, C. (2021). A total error framework for digital traces of human behavior on online plattforms. Public Opinion Quarterly 85:S1, pp 399-422. DOI: https://doi.org/10.1093/poq/nfab018

Statistics New Zealand (2016). Guide to reporting on administrative data quality. Guide to reporting on administrative data quality (stats.govt.nz)

UNECE. (2021). Machine learning for official statistics. ECECESSTAT20216.pdf (unece.org)

UNECE (2014). A Suggested Framework for the Quality of Big Data. Deliverables of the UNECE Big Data Quality Task Team. Big Data in Official Statistics - Big Data in Official Statistics - UNECE Statswiki

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica 66:1, pp 41-63. DOI: https://doi.org/10.1111/j.1467-9574.2011.00508.x