Ingegerd Jansson
Lilli Japec

Kvalitetskriterier för statistik baserad på digitala data - vägledning

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 vägledningen tillsammans med rapporten Kvalitetskriterier för statistik baserad på digitala data—bakgrund 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 bakgrundsrapporten. Processen då statistik produceras baserat på digitala data illustreras och potentiella osäkerhetskällor som kan uppstå beskrivs övergripande i både 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 den här 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 den här 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

Digitala data, organiska data, big data, nya datakällor, found data eller ”it-is-what-it-is” data är några namn som används för att beskriva datakällor som genereras automatiskt snabbt och ofta som en biprodukt. Dessa data kan vara av väldigt olika karaktär, från data som är väl beskrivna och väldigt lika de traditionella administrativa källorna, till data som är högst ostrukturerade och med bristfälliga metadata. Vi har valt att kalla dessa data för digitala data (National Academies of Sciences, Engineering, and Medicine 2022).

I september 2021 fick SCB 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.”*

Den här vägledningen tillsammans med rapporten Kvalitetskriterier för statistik baserad på digitala data—bakgrund (2023) är resultatet av arbetet som gjorts på kvalitetsområdet inom ramen för regeringsuppdraget. I bakgrundsrapporten beskrivs bland annat kvalitetsbegreppet, de föreskrifter och den förordning som finns för kvalitet i officiell statistik samt hur det här arbetet förhåller sig till annat internt arbetet på SCB. Vidare ges även en sammanfattning av relevant nationell och internationell litteratur inom kvalitetsområdet som ligger till grund för den här vägledningen. Bakgrundsrapporten avslutas med en sammanfattning av de felkällor som uppstår då statistik produceras baserat på digitala data och en diskussion om de avvägningar som behöver göras vid beslut om att använda en digital datakälla.

Den här vägledningen inleds med en kort bakgrund som beskriver de föreskrifter och den förordning som ligger till grund för officiell statistik. Exempel på digitala data ges och en modell som beskriver de olika felkällor som kan finnas då statistik produceras baserat på digitala data presenteras. Detta följs av en beskrivning av de avvägningar som behöver göras innan ett beslut kan tas om att använda en digital datakälla. I de efterföljande avsnitten (1-8) beskrivs de olika felkällorna, exempel på felkällor ges och kvalitetsindikatorer beskrivs. De kvalitetsindikatorer som presenteras kan användas för att utvärdera kvalitet i data och några av dessa kan också användas för att löpande följa upp kvalitet i statistik i de fall man väljer att producera statistisk baserat på den digitala datakällan.

Bakgrund

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 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.

Digitala data

Digitala data kan vara av väldigt olika karaktär. UNECE (2014) grupperar den här typen av data som man kallar för big data i tre olika kategorier enligt Tabell 1.

Data som en statistikproducent får tillgång till kan se väldigt olika ut, det kan till exempel handla om stora mängder mikrodata eller någon form av aggregerade data. Gemensamt för många av dessa källor är att det kan finnas stora inslag av bearbetning och modellering när de hanteras. 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.

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

Tabell 1. Klassificering enligt UNECE (2014)

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. Om det finns flera önskvärda statistiska syften så kan det vara nödvändigt att göra mer än en utvärdering av vissa delar.

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.

Kvalitet i digitala data

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.

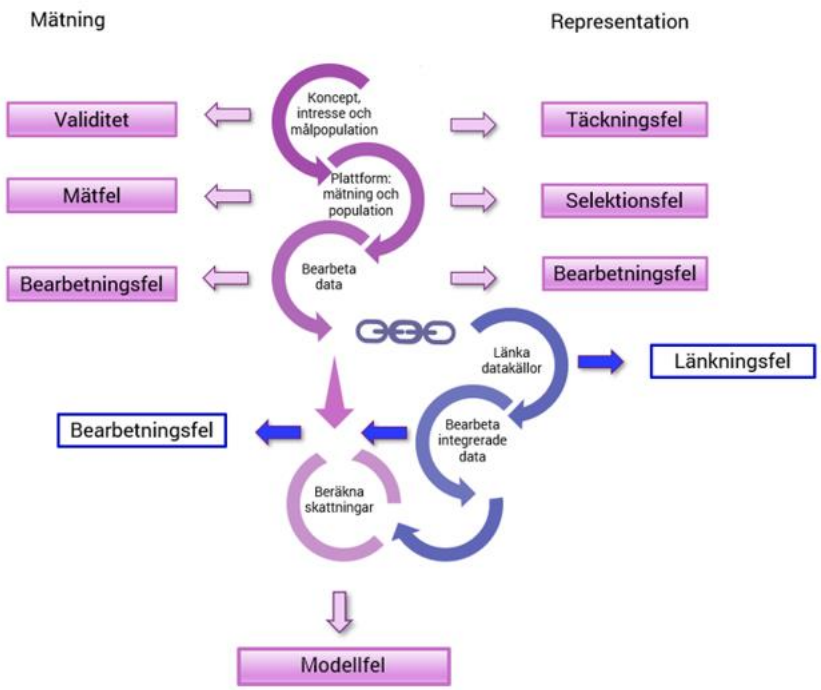
Beskrivningen av felkällor i digitala data (se Figur 1) är inspirerad av ramverk för TSE men också av Zhang (2012), Reid et al (2017) och Lothian et al (2019) där integrering av data är centralt, men beskrivningen här är medvetet förenklad. Ett begrepp är lånat från Sen et al (2022) men används i en 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 sammanfattar steg i de två dimensionerna och deras osäkerhetskällor. Figuren visar en process som börjar med en enskild digital källa 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.

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 mest troliga 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.

Felkällor, begrepp och indikatorer presenteras vidare i avsnitt 1-8. Vägledningen är upplagd enligt figuren nedan d.v.s. först presenteras osäkerhetskällor som påverkar representationen (täckningsfel, selektionsfel, bearbetningsfel och länkningsfel) och sedan felkällor som påverkar mätningen (validitet, mätfel, bearbetningsfel i den digitala källan och bearbetningsfel vid läkning).



Figur 1. Potentiella felkällor i digitala data

Avvägningar vid bedömning av en datakälla

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. Det är problematisk att använda ordet kriterier då det implicerar att det finns absoluta svar på vad som är tillräckligt bra data. Därför har arbete 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.

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 aspekter, 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.

Allt detta måste gå in i en samlad bedömning av om och i så fall hur digitala data kan användas.

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 den här vägledningen är alla de föreslagna indikatorerna tänkbara att ingå i en initial utvärdering av en datakälla. Listan med indikatorer ska ses som en bruttolista. Vidare föreslår vi även indikatorer som kan beräknas löpande i varje produktionsomgång.

Det viktigaste första steget är att göra en kvalitativ utvärdering av den digitala datakällan för att förstå exakt vad data innehåller d.v.s. vad som registreras och för vilka objekt (Indikator 1-A1 och 5-A2) samt vilka bearbetningar som görs. En sådan utvärdering kan ge viktig information om vilka andra felkällor som behöver studeras baserat på användarbehov. Vissa digitala data kan ha stora problem med täckningen t.ex. om vi endast har data från en mobiloperatör och vi vill dra slutsatser om Sveriges befolkning. I det fallet är det viktigt att studera indikatorer som har med täckningsfel att göra (Indikator1-B1). För andra datakällor kanske täckningen inte är något problem t.ex. data om sjöfartspositioner (AIS-data). I andra typer av digitala data så kan det vara problem med både täckning, validitet och mätfel t.ex. webbskrapning och då bör även viktiga indikatorer för dessa andra felkällor studeras (Indikator 5-A3, 6-A1, 6-A4). De är svårt att ge mer generella rekommendationer då både digitala data och användarbehov kan se så olika ut.

Det pågår, som litteraturgenomgången visar, en hel del arbete, och det är därför viktigt med en fortsatt omvärldsbevakning och att vid behov ompröva det här föreslagna indikatorerna.

Representation

Utgångspunkten är en intressepopulation, det vill säga en population med objekt som 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. I idealfallet stämmer målpopulationen överens med intressepopulationen, men troligt är att det finns skillnader, speciellt för datakällor där statistikproducenten inte alls eller bara delvis råder över designen för datainsamlingen.

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.

1. Täckningsfel—representation

I en undersökning definieras täckningsfel som skillnaden mellan mål- och rampopulation. Notera att de två populationer har samma objektstyp. Dessa skillnader leder till över- eller undertäckning dvs rampopulationen innehåller objekt som inte ingår i målpopulationen respektive rampopulationen saknar objekt som ingår i målpopulationen. Täckningen skattas genom länkning av datakällorna.

Objekten i en digital källa 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. I de fall plattformens (eller operatörens) population innehåller *samma objektstyp* som målpopulationen så kan plattformspopulationen ses som en motsvarighet till rampopulation ovan (ordet plattform är lånat från Sen et al 2022) och täckningsfelet skattas på motsvarande sätt.

I de fall då plattformspopulationen innehåller *en annan objektstyp* än målpopulationen så behöver data länkas för att få fram motsvarigheten till rampopulation ovan. När data från två källor, till exempel en digital källa och ett basregister, integreras är målpopulationen den integrerade mängden objekt, och ramenpopulationen består både av objekten i basregistret och objekten från plattformen. Registrets och plattformens objekt kan vara olika, till exempel kan plattformens objekt vara mätpunkter, sms eller webbannonser medan basregistrets objekt är personer, företag eller fastigheter. Via integreringen av data skapas ett statistiskt register med endast en typ av objekt. Skillnaden mellan målpopulationen och rampopulationen ger över- eller undertäckning. Både täckningen i basregistret och täckningen i den digitala källan bidrar.

Plattformspopulation

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 plattformspopulationens användar- eller kundbas inte är

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

Ett annat exempel på täckningsfel är om plattformspopulationen är en mobiloperatörs kunder och vi vill dra slutsatser om Sveriges befolkning (målpopulation). Avvikelsen mellan dessa två populationer kan leda till täckningsfel. En mobiloperatörs kunder kan vara koncentrerade till en viss del av landet eller till stor del utgöras av en yngre del av befolkningen. Att dra slutsatser om Sveriges befolkning baserat på den specifika mobiloperatörens data kan bli missvisande såvida man inte har kunskap om täckningsfelet och kan justera skattningarna. Ett exempel på övertäckning kan vara då man vill se hur befolkningen rör sig mellan olika platser över en viss tidperiod. Objekten är simkort och i det fall en person har flera mobilabonnemang d.v.s. simkort så kan det bli övertäckning.

1. Täckningsfel—representation

Förekommer täckningsfel?
Om ja:
A1. Har det gjorts någon utvärdering av hur data genereras och vilka objekt som registreras? Finns över- eller undertäckning?
A2. Hur ser kopplingen ut mellan målobjekt och plattformspopulationens objekt?
A3. Kan objekten i den digitala datakällan kopplas till ett basregister? (Se avsnitt 4 om länkningsfel)
A4. Görs någon justering för att hantera täckningsfel? Om ja, beskriv hur.

Indikatorer:
B1. Plattformens täckningsgrad av marknaden
B2. Något mått från en utvärdering av effekten av täckningsfel

Indikatorer som kan tas fram löpande: A1, B1

2. Selektionsfel—representation

Skillnad mellan rampopulationen och den observerad mängden ger selektionsfel. I de fall då målpopulationen och den observerade mängden har samma objektstyp så motsvarar rampopulationen plattformspopulationen. Målsättningen är att observera alla objekt i rampopulationen men det kan förkomma att kända objekt inte kan observeras. Slumpmässigt urval kan också förekomma och beskrivs i så fall separat.

Selektionsfel uppstår i digitala data om objekt som borde ingå i indata av någon anledning inte finns med. Det kan finnas fler orsaker till att *objekt* saknas. Det kan till exempel bero på att registreringar av objekt misslyckas eller att det finns eftersläpning i registreringen av objekt. Det kan också bero på att man endast har ett urval av objekt. Man kan ha gjort ett medvetet urval för att mängden data är för stor eller för att avgränsa en domän.

När mätvärden saknas i en digital källa så räknas vissa fall även det som ett representationsfel (se även avsnitt 6 om mätfel). Ett saknat mätvärde kan betyda att även objektet saknas, till exempel för en mobiltelefon som är avstängd registreras inga mätvärden (signaler) men inte heller objektet simkort. Ytterligare ett exempel på detta är en platsannons som inte kommer med i en webbskrapning vilket betyder att inte heller företaget som söker arbetskraft kommer med.

Det krävs en speciellt designad utvärderingsstudie för att avgöra om det finns selektionsfel och för att beräkna storleken på felet (De Waal et al 2019, Daas et al 2020).

2. Selektionsfel—representation

Förekommer selektionsfel?
Om ja:
A1. Vilken typ av selektionsfel finns det?
A2. När i genereringen eller registreringen av data uppstår selektionsfel?
A3. Görs någon justering för att hantera selektionsfel? Om ja, beskriver hur.

Indikatorer
B1. Något mått från en utvärdering av effekten av selektionsfelet

Indikatorer som kan tas fram löpande: A1

3. Bearbetningsfel i digitala data—representation

Bearbetningsfel som påverkar representationen uppstår i den digitala datakällan då man vill *ta bort* (t.ex. misstänkta dubbletter), *lägga till* (t.ex. data från ytterligare en operatör) eller på annat sätt *modifiera* objekt. Den här processen sker med hjälp av någon form av modell, till exempel för att identifiera och ta bort dubbletter. I det här steget används endast information som finns i den digitala datakällan (se även avsnitt 4 om länkningsfel då datakällor integreras).

Ett objekt definieras i digitala data som en *dubblett* när det finns minst en till registrering i indata som avser samma objekt. Man kan ta bort objekt om man misstänker att det handlar om dubbletter men det är inte alltid entydigt vad som är en dubblett. I andra fall så kan definitionen av en dubblett till exempel bero på tidsperioder eller vad som är objektet.

Ett annat exempel då objekt tas bort vid bearbetning av digitala data är vid webbskrapning om man misstänker att ett konto tillhör en bot (se avsnitt 6 om mätfel). Ibland kan man också imputera objekt, till exempel i mobiloperatörsdata om det varit ett avbrott och man imputera senast tillgängliga data för objekt.

Annonser om lediga jobb som förekommer precis samtidigt på flera webbportaler och som avser samma jobb är troligen dubbletter som inte är önskvärda att behålla. Om annonsen avser olika tidsperioder så är det eventuellt inte dubbletter eftersom annonsen kan ha lagts ut igen för att platsen inte blev tillsatt. Därför kan det vara relevant att inte ta bort misstänkta dubbletter utan i stället markera dom som dubbletter och lämna avgörandet till när data integreras eller i ett skattningsförfarande.

Notera att dubbletter även kan uppstå vid länkning av datakällor (se avsnitt 4).

3. Bearbetningsfel i digitala data—representation

A1. Beskriv vilken typ av bearbetning som görs och vilka modeller som används
A2. Finns det anledning att behålla dubbletter? Om ja, markeras de i så fall på något sätt?

Indikatorer
B1. Antalet/andelen objekt som är misstänkta dubbletter
B2. Antalet/andelen borttagna dubbletter

Indikatorer som kan tas fram löpande: B1, B2

4. Länkningsfel—representation

Det kan finnas flera skäl till att man vill integrera datakällor, till exempel för att utöka variabelmängden, för att få en koppling till målpopulationen eller för att utvärdera kvalitet i data. Det finns även olika strategier för att integrera datakällor som beror på vilken information som finns i de datakällor som ska integreras.

Det är troligt att de flesta digitala källor av intresse för SCB kommer behöva en koppling till något 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 med Telia som operatör så kan relevant

statistik skattas med data från endast Telia. Om inferensen avser mobilanvändning hos Sveriges befolkning så behövs både representativa data från en eller flera operatörer och en koppling till ett register över målpopulationen Sveriges befolkning.

Strategier för länkning

För att länka till ett befintligt register behövs relevanta länkningsvariabler. Helst ska dessa variabler vara unika identifierare i de digitala data som återfinns 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 geografiska områden (DeSo) till RTB.

När digitala data länkas till ett befintligt register så är det inte säkert att det går att länka alla objekt i basregistret till objekt i de digitala data. Det kan finnas olika skäl till att länkning inte fungerar. Länkningsinformation kan saknas i den digitala källan 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 statistik matchning. Då integreras datakällor som innehåller olika objekt. Det kan ske på mikro- eller makronivå (De Waal et al 2020).

Länkning för att skapa målobjekt

En inte helt ovanlig situation med digitala data är att dessa har en annan objektstyp än de objekt som man vill dra slutaster om. Genom länkning till ett basregister så definieras en ny objektstyp, målobjekt. Länkningen kan innebära en transformation av objekt i ett eller flera steg. En mätpunkt för eldata kan till exempel först behöva länkas till en person (via en adress) som i sin tur ingår i ett hushåll, eller till ett företag (via organisationsnummer) som genom en adress kan länkas till ett arbetsställe. Mätpunkter för el har i sig inget intresse för användarna, utan det är statistik över företagens eller hushållens elanvändning som efterfrågas. Dessa målpopulationer är inte definierade i den digitala källan. Plattformspopulationen är alla mätpunkter i Sverige. I praktiken

faller vissa bort av säkerhetsskäl eller eventuellt på grund av faktorer i leveransen och saknas därför i den observerade mängden.

Problem som uppstår vid länkning

Det är troligt att en viss andel av objekten i målpopulationen inte kan länkas entydigt. Det kan bero på felaktig länkningsinformation, men det kan även vara så att den information som finns att tillgå inte räcker för entydig länkning. Den andel av objekten som inte kan länkas entydigt behöver beskrivas ytterligare. Det kan till exempel vara att en mätpunkt i eldata kan länkas till fler än ett arbetsställe eller hushåll, eller att fler än ett företag ser ut att kunna ha gett upphov till samma annons om ett ledigt jobb.

Länkningsfel uppstår vid integrering av två eller fler datakällor om objekt saknas eller om objekt finns med fast de inte är av intresse.

Länkningsfel kan uppstå i olika situationer:

- Alla objekt i den digitala datakällan kan inte länkas till ett objekt i basregistret på grund av att
    - alla objekt inte har den information som krävs för en lyckad länkning eller informationen är av dålig kvalitet,
    - objektet inte finns i basregistret.
- Alla objekt i den digitala datakällan kan länkas till basregistret men basregistret innehåller objekt som inte finns i den digitala datakällan

Exemplet mobildata illustrerar några situationer. Simkort kan inte länkas till individer i RTB för att tillräcklig information om individerna inte finns i de digitala data SCB har tillgång till. Om SCB får mer data om individerna så kan det finnas simkort som hör till individer som inte är folkbokförda eller har samordningsnummer i Sverige. Om SCB bara fått data från en operatör så kommer det finnas många individer i RTB som inte har abonnemang hos den operatören (se avsnitt 1 om täckningsfel).

Länkningen kan skapa icke önskvärda dubbletter som behöver identifieras och rensas bort. Det kan vara fallet om flera objekt i den digitala datakällan kan länkas till samma objekt i basregistret. Två mobiler som är registrerade på samma person kan till

exempel vara en dubblett om individer är målobjekt, men de är kanske inte dubbletter om målobjektet är mobilmaster.

Om den digitala datakällan inte innehåller unikt identifierande variabler så krävs en modell som bestämmer hur datamängderna ska länkas. Det tillför osäkerhet till den integrerade objektsmängden.

4. Länkningsfel—representation

Görs länkning av datakällor?
Om ja:

A1. Vilka källor länkas? Finns uppgifter om täckningsfel i de datakällor som den digitala källan länkas till?
A2. Vilka variabler används för länkning?
A3. Finns det identifierande variabler?
A4. Beskriv modellen för länkning
A5. Hur hanteras objekt som inte kan länkas entydigt?

Indikatorer (se Bilaga 1)

B1. Andelen objekt som *inte* kunde länkas entydigt
B2. Antal/andel objekt av plattformspopulationen som *inte* kunde länkas entydigt
B3. Andel objekt av målpopulationen som *inte* kunde länkas entydigt, eventuellt efter redovisningsgrupper (se även avsnitt 1 om täckningsfel)
B4. Andelen av målpopulationen som kan länkas, eventuellt efter redovisningsgrupper
B5. Andel av plattformspopulationen som kan länkas
B6. Se litteraturen för mått för modellprestanda t.ex. UNECE (2021)

Se Bilaga 1 för illustration av de olika indikatorerna ovan.

Indikatorer som kan tas fram löpande: B1-B5

Mätning

I en undersökning fångas användarnas behov och krav genom intressevariabler och utifrån dessa formuleras målvariabler. Utifrån målvariablerna formuleras frågor. Här har statistikproducenten möjlighet att noggrant formulera frågor som ska mäta det koncept som användarna är intresserade av. I digitala data finns inte den möjligheten. 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.

5. Validitet—mätning

Problem med validitet uppstår då den operationalisering av konceptet som statistikproducenten valt inte helt överensstämmer med konceptet som användarna är intresserad av. Det kan hända att operationaliseringen inte fångar hela det önskvärda konceptet. 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.

Om användarna är intresserad av konceptet ”dagbefolkning” dvs att se var befolkningen befinner sig under dagen så kan det mätas på flera olika sätt. I det fall man har data från mobiloperatörer med signaler från mobiler och positionsdata, så behöver konceptet definieras tydligt.

I det fall användarna är intresserade av konceptet ”elförbrukning” så kan vi använda data från smarta elmätare för att mäta hur mycket el som förbrukas. Konceptet ligger väldigt nära det sätt vi valt att operationalisera det då vi använder elmätardata.

En mätteknisk utvärdering (Persson 2022) kan ge värdefull information om hur digitala data har genererats, vilka bearbetningar som gjorts och information som kan användas för att definiera konceptet. Det kan i vissa fall även vara motiverat med någon form av kvantitativ utvärderingsstudie.

5. Validitet–mätning

Finns avvikelser mellan konceptet och operationaliseringen av konceptet?

Om ja:

A1. Hur definieras konceptet som användarna är intresserade av?
A2. Vad mäts i den digitala datakällan?
A3. Vilken avvikelse finns mellan A1 och A2?
A4. Har det gjorts någon mätteknisk utvärdering?

Indikatorer:

B1. Någon form av mått på avstånd mellan koncept och målvariabel

6. Mätfel—mätning

Mätfel beror på mätinstrumentet eller användaren och uppstår då observerade värden är *felaktiga* eller då värden som borde registrerats *saknas*. Det kan även vara *outliers* som observerats. I digitala data så är mätinstrument tekniska enheter eller applikationer som registrerar olika typer av signaler eller händelser. Digitala data kan också vara text som användare skriver i form av inlägg i socialmedia som Twitter eller Facebook eller annan information som finns på nätet, till exempelpriser på varor.

Statistikproducenten kan oftast inte påverka hur mätvärden 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. När webbskrapning görs så kan mätfel uppstå på grund av att de nyckelord som man valt inte återspeglar det som man önskar mäta.

I vissa digitala data är risken för mätfel troligen ganska liten, till exempel 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.

Saknade värden

Ett *saknat* värde i digitala data skulle kunna uppstå på grund av tekniska problem, planerade avbrott eller att användaren stängt av den tekniska enheten och ingen signal eller position kommer då att registreras. I mobildata kan ett saknat värde bero på en överbelastad mobilmast, kallt väder eller ett fel på SIM-kortet. Det kan också bero på att användaren tillfälligt stängt av mobiltelefonen. Fallet ovan med webbskrapning kan leda till saknade värden. I de fall digitala data
sträcker sig över en längre tidsperiod så kan observationer för ett objekt finnas vid ett tillfälle men kan saknas vid nästa omgång. För saknade värden över tid så kan vi skatta indikatorer vid varje omgång. Saknade värden avser objekt som finns med vid båda tillfällena.

Felaktiga värden

Ett *felaktigt* värde kan uppstå på grund av planerade avbrott eller tekniska problem. Det kan till exempel handla om att ingen

position registrerats i data om sjöfartspositioner. Felaktiga värden kan också bero på att en individ eller en enhet registrerat fel värden, till exempel fel yrkeskod i platsannonsdata. Andra exempel på felaktiga värden i digitala data är vid webbskrapning där text på webben kan ha genererats av en bot eller i en jobbportal där företag lagt ut annonser om lediga jobb enbart i marknadsföringssyfte. Det är möjligt att ibland identifiera ett felaktigt värde om det är extremt högt eller lågt (se nedan om outliers). I de flesta andra fall så krävs en studie eller modell för att identifiera felaktiga värden.

Outliers

En *outlier* är en observation (eller en serie av observationer) vid en viss tidpunkt som avviker mycket från övriga observationer i data, som kan vara korrekt och som har stor påverkan på skattningar. Hur stor avvikelsen ska vara för att betraktas som mycket, eller hur stor påverkan definieras, varierar mellan datakällor och syfte. Det kan behövas en modell som avgör när en observation är en outlier. Eftersom en outlier kan vara korrekt så bör den inte ändras eller tas bort i digitala data, men den ska markeras för att senare kunna hanteras i skattningsförfarandet.

6 Mätfel—mätning

I. Förekommer felaktiga mätvärden?
Om ja:
A1. Vilken typ av felaktiga mätvärden finns det?
A2. Hur uppstår felaktiga mätvärden?
A3. Hur hanteras felaktiga mätvärden?

Indikatorer per redovisningsgrupp
B1. Andelen felaktiga mätvärden
B2. Andelen korrigerade mätvärden
B3. Andelen borttagna mätvärden

II. Saknas mätvärden?
Om ja:
A4. Varför saknas mätvärden?
A5. Görs någon justering för saknade mätvärden?

Indikatorer
B4. Andel saknade mätvärden per variabel
B5. Andel saknade mätvärden per redovisningsgrupp

III. Förekommer outliers?
Om ja:
A6. Beskriv en eventuell modell som används för att identifiera outliers. Vilka antaganden görs i modellen? Är det några antaganden som inte är uppfyllda? Är modellen robust?
A7. Hur markeras outliers?
A8. Hur hanteras outliers i skattningar?

Indikatorer
B6. Antal outliers
B7. Andel av det totala mätvärdet per redovisningsgrupp
B8. Se litteraturen för mått för modellprestanda t.ex. UNECE (2021)

Indikatorer som kan tas fram löpande: B1-B7

7. Bearbetningsfel i digitala data—mätning

Digitala data bearbetas genom att man *tar bort värden* (t.ex. outliers) eller *skapar nya värden* (t.ex. genom kodning). I det här steget används data som redan finns i den digitala datakällan för att göra bearbetningar. Bearbetningsfel som uppstår kan påverka skattningar. I det fall då bearbetning görs av dataägaren så behöver det finnas information om vilken bearbetning som gjorts och hur. Det mesta i bearbetningsfasen görs automatiskt med hjälp av en *modell*.

Nya värden skapas genom kodning. Den kodning som görs i det här steget handlar om att man använder information som finns i den digitala datakällan och klassificerar den enligt en befintlig standard eller en egenutvecklad sådan. Kodningsfel uppstår då en felaktig kod sätts. Ett exempel på kodning som görs i digitala data är när man har positionsdata från mobiloperatörer och vill göra en geografisk indelning till exempel enligt DeSo (demografiska statistikområden), en av flera olika standarder som används för geografisk indelning. Ett annat exempel på kodningsfel som kan uppstå är då man med hjälp av texten i platsannonser kodar de lediga jobben enligt SSYK (standard för svensk yrkesklassificering).

Notera att det är två typer av utvärdering som behöver göras dels utvärdering av modellen som används för att koda materialet och dels själva utfallet av kodningen vid varje tillfälle som man applicerar modellen. Modellen som används kan behöva justeras om utfallet av kodningen försämras över tid.

7 Bearbetningsfel i digital datakälla—mätning

Förekommer kodning?
Om ja:
A1. Vilken klassifikationsstandard används?
A2. Utvärdering av modellen som används: Beskriv modellen som används för kodning. Vilka antaganden görs i modellen? Är det några antaganden som inte är uppfyllda? Är modellen robust? Har någon kontrollkodning gjorts?
A3. Utvärdering av kodning: Finns det objekt som inte kunde kodas? Hur hanteras dessa objekt?

Indikatorer
B1. Andel värden som inte kunde kodas
B2. Se litteraturen för mått för modellprestanda t.ex. UNECE (2021)

Indikatorer som kan tas fram löpande: B1

8. Bearbetningsfel då datakällor integreras—mätning

Den bearbetning som sker på mätsidan då data integreras handlar om att skapa *nya variabler* eller att *lägga till värden (imputering)* baserat på variabler i det integrerade datasetet. Notera att den här bearbetningen skiljer sig från den tidigare då det finns mer information att tillgå i det integrerade datasetet än enbart i den digitala datakällan. Precis som tidigare så sker bearbetningen även här med hjälp av någon modell. De bearbetningsfel som kan uppstå är härledningsfel, kodningsfel, imputeringsfel och modellfel.

När flera datakällor integreras så kan nya variabler skapas baserade på variabler i det integrerade datasetet. Det kan handla om att man skapar en ny variabel som innehåller koder som baseras på olika kombinationer av värden på variabler i det integrerade datasetet. Det kan också vara så att man använder andra typer av statistiska modeller för att skapa en ny variabel. I båda dessa fall så görs antagande och om det är fel värden i någon av de ingående variablerna så kan *härledningsfel* uppstå i de härledda variablerna.

Imputering betyder att felaktiga eller saknade variabelvärden ersätts med andra värden, som antas ligga nära de ersatta värdena. De andra värdena baseras på en modell där det till exempel kan ingå observerade värden från tidigare omgångar, observerade värden i samma omgång, eller hjälpvariabler där det finns fullständig information. Imputeringsmodeller för digitala data diskuteras till exempel i UNECE (2021).

Värden som extraherats ur annonser, till exempel vilken typ av tjänst som söks eller hur många tjänster som utlyses i annonsen, bearbetas för att beräkna till exempel antalet lediga jobb i en viss bransch. Information från basregistret används för att koda bransch och eventuellt för att till imputera värden eller identifiera företag som är outliers, enligt någon modell.

8 Bearbetningsfel då datakällor integreras—mätning

I. Skapas nya värden genom härledning i det integrerade datasetet?
Om ja:
A1. Vilka variabler skapas och hur?

Indikatorer
B1. Andel objekt för vilka det inte gick att härleda den nya variabeln.

II. Förekommer imputering?
Om ja:
A2. Vad imputeras?
A3. Hur markeras imputerade variabelvärden?
A4. Utvärdering av modell: Beskriv modellen som används för imputering. Vilka antaganden görs i modellen? Är det några antaganden som inte är uppfyllda? Vilka modellvariabler används? Är modellen robust?
A5. Utvärdering av imputeringen: Finns det objekt där imputering inte gick att göra? Hur hanteras dessa objekt?

Indikatorer
B2. Antal/andel imputerade värden per variabel
B3. Se litteraturen för mått för modellprestanda t.ex. UNECE (2021)

Indikatorer som kan tas fram löpande: B1, B2

Referenser

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

Groves, R. M. (2011) Three eras of survey research. Public Opinion Quarterly, 75:5, pp 861-871.
https://doi.org/10.1093/poq/nfr057

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). Survey Methodology. Wiley Series in Survey Methodology.

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.

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

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

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.

Persson, A. (2022) Mätteknik 2.0.

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 (2016) Att utforma och förbättra en statistisk undersökning.

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

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

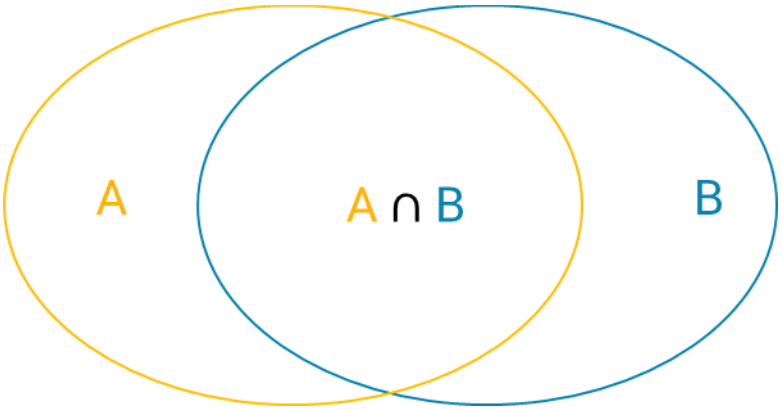
Sen, I., Flöck, F., Weller, K., Weiss, B. & Wagner, C. (2021) A total error framework for digital traces of human behavior on online plattforms. Public Opinion Quarterly 85:S1, pp 399-422. DOI: https://doi.org/10.1093/poq/nfab018

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

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

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

Bilaga 1 Länkning



Figuren illustrerar två datamängder som länkas, A är plattformspopulationen och B är målpopulationen. Snittet $A \cap B$ av A och B är den mängd objekt som kan länkas. Unionen

$$A \cup B = A + B - (A \cap B)$$

av A och B är totala mängden objekt.

Andel som kan länkas:

$$(A \cap B)/(A \cup B)$$

Andel som inte kan länkas:

$$(A + B - 2(A \cap B))/(A \cup B)$$

Andel av plattformspopulationen som inte kan länkas:

$$(A - (A \cap B))/A$$

Andel av målpopulationen som inte kan länkas:

$$(B - (A \cap B))/B$$

Andel av plattformspopulationen som kan länkas:

$$(A \cap B)/A$$

Andel av målpopulationen som kan länkas:

$$(A \cap B)/B$$