Uppdrag att främja delning och nyttiggörande av data för smart statistik

Slutrapport

2023-02-23
1.0
SCB Dnr A2021/3275
Marie Haldorson
Biträdande generaldirektör

Förord

SCB har fått regeringens uppdrag att främja delning och nyttiggörande av data för smart statistik med inriktning på datakvalitet, dataåtkomst och datatillgång. I uppdraget har ingått att lämna förslag till ny smart statistik baserad på delning av bl.a. mobilitetsdata samt hur smart statistik kan visualiseras på ett enkelt vis. Uppdraget har delredovisats i februari 2022 och slutrapporteras härmed.

Uppdraget har koordinerats av biträdande generaldirektör Marie Haldorson och genomförts av ett stort antal experter inom SCB under ledning av Lilli Japec och Ingegerd Jansson (kvalitetsindikatorer), Pieter Vlag och Ulf Durnell (mobilitetsdata), Sara Brinkberg och Linda Larsson (föreskrifter öppna data), Elma Jakupovic och Karin Hansson (visualisering) samt Mihaela Weideskog (riskanalys datasäkerhet).

SCB har tagit hjälp av Visualiseringscenter C, Linköpings universitet, för att utveckla visualiseringen av öppna data. Samråd har skett med Digg, främst i samband med digital arena. Arbetet med mobilitetsdata har skett i samarbete med Telia och Tre. Samtliga statistikansvariga myndigheter har involverats i arbetet med föreskrifter för öppna data. Trafikanalys och Tillväxtverket har även varit involverade i det utforskande arbetet att använda mobilitetsdata.

SCB 23 februari 2023

Joakim Stymne	Marie Haldorson
Generaldirektör	Biträdande generaldirektör

Innehåll

Sammanfattning

SCB har genomfört ett arbete bestående av olika insatser för att främja delning och nyttiggörande av data för smart statistik enligt regeringsuppdrag (I2021/02417). Arbetet bidrar till genomförandet av den nationella datastrategin (Regeringen 2021).

Uppdraget har bedrivits i fem parallella spår som redovisas i vart och ett av rapportens huvudkapitel:

- Kvalitetsindikatorer för digitala data
- Datasäkerhet
- Mobilnätsdata – ny statistik
- Öppna data – föreskrifter
- Öppna data – visualisering av statistik

Redovisningen avslutas med generella slutsatser och rekommendationer.

Kvalitetsindikatorer för digitala data

I takt med att olika typer av digitala data blir alltmer använda för statistikproduktion finns också ett ökat behov av att kunna bedöma dessa datakällors kvalitet. SCB har tagit fram en uppsättning kvalitetsindikatorer som kan användas för att bedöma kvaliteten i statistik som produceras baserat på digitala data. Arbetet med indikatorerna visar att datakvalitet bara kan utvärderas utifrån vad den ska användas till, det går inte att uttala sig generellt.

Kvalitetsindikatorer har valts framför kvalitetskriterier, eftersom kriterier kan tolkas som att det finns absoluta svar på vad som är tillräckligt bra data – vilket det inte gör. Kvalitetsindikatorerna kan ge en uppfattning om svagheter och felkällor som kan påverka statistiken, de kan också vara mätbara. De ger ett konkret faktaunderlag som visar om datakällan kan användas i statistikproduktionen. Då det aktuella uppdraget har avsett data som kan användas för att ta fram smart statistik så har arbetet avgränsats till att gälla digitala data.

Kvalitetsindikatorerna kommer att användas i SCB:s statistikproduktion och vara tillgängliga för alla statistikansvariga myndigheter. Listan med indikatorer (Bilaga 1) ska ses som en bruttolista. De är utformade med ett statistiskt synsätt och avser två huvudsakliga dimensioner, representation och mätning, med fyra typer av felkällor i data för respektive dimension. För representation kan det handla om täckningsfel, selektionsfel, bearbetningsfel och länkningsfel. För

mätning kan det handla om validitet, mätfel, bearbetningsfel per datakälla och för integrerade datakällor. För att upptäcka dessa fel finns förslag på olika indikatorer.

Ytterst är det hur statistiken ska användas som avgör vilka datakällor som är bäst lämpade för att ta fram tillräckligt bra statistik. Data kan vara bra för en användning men sämre för en annan. Ibland räcker det att bedöma datakvaliteten utifrån ett ganska snävt användningsområde, medan i andra fall ska data kunna användas brett för många olika syften.

Det viktigaste första steget är att göra en kvalitativ utvärdering av den digitala datakällan för att förstå exakt vilka data den innehåller. En sådan utvärdering kan ge viktig information om vilka 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.

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

Rapportavsnittet fördjupar resonemangen kring kvalitetskriterierna och ger en internationell översikt, liksom en övergripande beskrivning av de olika kriterierna. I Bilaga 1 beskrivs de åtta felkällorna mer i detalj med vägledande frågor som kan användas för att bedöma kvaliteten i en digital datakälla.

Datasäkerhet

SCB har genomfört en riskanalys med utgångspunkt i Diggs vägledning för att tillgängliggöra information, där risker identifierats kopplat till fyra av vägledningens sju principer:

- Princip 2, bedriv ett riskbaserat och systematiskt informationssäkerhetsarbete;
- Princip 3, tillgängliggör aktuell och uppdaterad information med användaren i centrum;
- Princip 5, använd villkor som främjar bred användning och
- Princip 6, dokumentera och beskriv information.

SCB har identifierat flera informations- och cybersäkerhetsrisker samt tagit fram åtgärdsförslag för riskhantering. Det behövs t.ex. tydligare rekommendationer och kunskap om gällande lagkrav vad gäller informationsklassning hos myndigheter. Det är också så att privata/kommunala aktörer inte alltid omfattas av samma lagkrav som statliga myndigheter, vilket kan resultera i att informationen inte hanteras korrekt.

Avsaknad av en gemensam begreppskatalog gör att samma term kan tolkas på flera olika sätt. Ett exempel är aggregerade uppgifter som inom säkerhetsskyddslagstiftningen betyder att flera olika typer av uppgifter samlas och tillsammans utgör ett nytt ökat skyddsvärde. I statistiksammanhang skulle detta benämnas integrerade eller kombinerade data. När data aggregeras till statistisk slås däremot exempelvis uppgifter från flera individer, grupper eller tidsperioder samman, vilket ger en lägre informationsklassning. MSB:s begreppskatalog (termbank för informationssäkerhet) bör utvecklas vidare och kunskap om denna katalog spridas inom offentlig och privat sektor.

Verksamheter bör inte hantera mer information än vad som krävs för uppdraget, det ska finnas etablerade och tydliga rutiner för hantering av data. Kombinerade data kan i vissa fall kan ge ett högre skyddsvärde än de enskilda delarna. Det behövs nationell vägledning om vilken organisation/aktör som bär ansvaret för att bedöma nivån på materialet som ska informationsklassas.

Rapportavsnittet innehåller en översikt med samtliga identifierade risker, i Bilaga 2 finns varje risk beskriven mer i detalj tillsammans med förslag på åtgärder.

Mobilnätsdata – ny statistik

Mobilnätsdata är en datakälla som utforskats av SCB och andra myndigheter, både nationellt och internationellt, under de senaste åren. Det finns potential att snabbt kunna följa förändrade resmönster och att se var mobiltelefoner befinner sig vid olika tidpunkter på dygnet. Under pandemin användes mobilnätsdata för att följa just detta. Samtidigt är det av högsta vikt att den här typen av data hanteras med högt dataskydd och de uppgifter som SCB fått testa är anonymiserade och aggregerade.

Uppdraget redovisar lärdomar från de senaste årens tester att använda mobilnätsdata för att ta fram statistik. Data från Telia och Tre har kunnat användas för att studera var dag- och nattbefolkning befinner sig (baserat på var mobilers SIM-kort är), liksom rörelser som kan översättas till resor. För att kunna tolka de stora datamängder som genereras krävs god kunskap om datagenereringsprocessen – även om den till stora delar sker hos mobilnätsoperatören. SCB behöver kunna garantera att statistiken är oberoende och har tagits fram med en reproducerbar metod.

SCB har testat att ta fram ny smart statistik för den dynamiska befolkningen, där det går att se säsongsmässiga skillnader i befolkningsaktiviteter – t.ex. under helger och under semesterperioder. Detta kan vara ett värdefullt, snabbt komplement till den registerbaserade befolkningsstatistiken som utgår ifrån folkbokföring.

SCB har även tillsammans med Trafikanalys och Tillväxtverket undersökt om det går att använda mobilnätsdata för att minska behovet av direktinsamling till befintlig statistik om pendling, turism och resvanor. Här har arbetet inte kommit så långt att det gått att konkretisera potentialen för mobilnätsdata. Andra myndighets-kontakter har visat att det finns intresse och potential inför ett eventuellt nästa steg.

Öppna data - föreskrifter

SCB har utformat justerade föreskrifter med tillhörande allmänna råd riktade till statistikansvariga myndigheter, med syfte att all officiell statistik ska gå att hitta via dataportal.se som öppna data. Alla statistikansvariga myndigheter har haft möjlighet att lämna synpunkter på den nya föreskriften, SCB har omhändertagit dessa så långt som möjligt.

SCB har parallellt med föreskriftsarbetet arbetat för att öka kunskapen hos övriga statistikansvariga myndigheter om vikten av att tillgängliggöra statistik som öppna data. Det har skett genom inspiration och erfarenhetsutbyte, t.ex. med externa intressenter som Wikimedia och genom att Digg har informerat om möjligheterna som dataportalen och den digitala arenan ger.

SCB har även delat med sig av exempel på tekniska lösningar såsom PxWeb och bildat ett användarforum för PxWeb och relaterade programvaror riktat till övriga statistikansvariga myndigheter. Forumet syftar till kunskapsdelning och erfarenhetsutbyte kring tekniska lösningar för att tillgängliggöra statistik. Den senaste versionen av PxWeb möjliggör automatiserad leverans av data från PxWeb till dataportal.se. Detta ska underlätta för statistikansvariga myndigheter att leva upp till den nya föreskriften och målsättningen att all officiell statistik ska gå att hitta via dataportal.se.

Rapportavsnittet kompletteras med konsekvensutredning vid regelgivning i Bilaga 3.

Öppna data – visualisering av statistik

Visualisering av öppna data har ingått i uppdraget och ett projektsamarbete mellan SCB och Visualiseringscenter C genom Linköpings universitet, LIU, har undersökt och testat hur en befintlig applikation, Sverige i siffror, kan dataförsörjas med öppna data från SCB och andra statistikansvariga myndigheter. Applikationen är i dagsläget installerad på s.k. touchbord som finns på Science center i Norrköping och på Universeum i Göteborg. Ett ytterligare resultat av arbetet är en designprototyp för en webbversion av Sverige i siffror som kan öka möjligheten för t.ex. skolor att ta del av visualiseringen av statistik genom att den kan användas på egna datorer eller surfplattor.

Slutsatser och rekommendationer

Rapporten avslutas med generella slutsatser och rekommendationer inför ytterligare insatser för att främja delning och nyttiggörande av data för smart statistik kopplat till två av den nationella datastrategins insatsområden. SCB lämnar följande rekommendationer:

1. SCB föreslår att regeringen tillför medel till SCB för allmänna företagsregistret enlig äskande i myndighetens budgetunderlag för 2024–2026.
2. SCB föreslår att regeringen ger SCB tillsammans med Visualiseringscenter C i uppdrag att implementera den framtagna designprototypen för Sverige i siffor i webbversion. Uppdraget bör omfatta aktiviteter för att sprida användningen av Sverige i siffror till grund- och gymnasieskolor, vilket därmed ökar nyttjandet av öppna data från SCB och andra myndigheter.
3. SCB föreslår att regeringen tillför medel till SCB för det tekniklyft som krävs för att öka takten i användningen av digitala och andra data enligt äskande i myndighetens budgetunderlag för 2024–2026.
4. SCB föreslår fortsatt dialog med Regeringskansliet i frågor som rör data för smart statistik, där SCB kan bidra med bred kompetens inom t.ex. dataförvaltning, datasäkerhet, juridik och tekniskt stöd. SCB ser även fram emot fortsatt dialog rörande SCB:s uppdrag relaterat till dataförvaltningsförordningen.

Termer och benämningar som används i rapporten

Term/benämning	Förklaring
Aggregering	Inom säkerhetsskyddslagstiftningen betyder aggregerade uppgifter att flera olika typer av uppgifter samlas och tillsammans utgör ett nytt ökat skyddsvärde. I statistiksammanhang skulle detta benämnas integrerade data eller kombinerade data. När data aggregeras till statistisk slås däremot exempelvis uppgifter från flera individer, grupper eller tidsperioder samman, vilket ger en lägre informationsklassning.
Creative Commons, CC-BY och CC0	Licenser som används för öppna data och ges ut av en amerikansk ideell organisation med samma namn. CC-BY innebär att de externa parter som vill utnyttja material som har publicerats, kan göra det på vilket sätt de vill, så länge en oförändrad vidarepublicering anger varifrån materialet kommer. Licensen CC0 ställer inte krav på att ange varifrån materialet kommer, vilket ytterligare underlättar vidareutnyttjande.
Dagbefolkning (mobilnätsdata)	Genomsnittligt antal personer som vistas i ett område mellan kl 09:00 och 15:00 baserat på aggregerad och anonymiserad SIM-kortsinformation. Denna definition av dagbefolkning skiljer sig från traditionell sysselsättningsstatistik där förvärvsarbetande dagbefolkning beskriver var individer arbetar enligt register.
Dataekosystem	En FN-rapport (FN 2021) beskriver ett dataekosystem som en modell för hur olika aktörer kan utbyta, producera och använda data. Data genereras på allt fler ställen (hos myndigheter, företag, organisationer och inom civilsamhället) och är en decentraliserad, digital resurs av stort värde för samhället. Att använda begreppet ekosystem visar att det är något som ständigt förändras, det tillkommer nya datakällor och nya användningsområden.
Digitala data	Digitala data avser data som genereras digitalt som en del av någon process men som har ett annat huvudsakligt syfte än att ta fram statistik. Kallas ibland även ”big data”, nya datakällor, organiska data, found data eller it-is-what-it-is data. Referens: National Academies of Sciences, Engineering, and Medicine 2022.
Imputering	Imputering innebär att saknade variabelvärden ersätts med värden som kan antas ligga nära de sanna värdena.
Inferens	Inferens betyder slutledning under osäkerhet, t.ex. när man vill dra slutsatser om en egenskap i en population där man bara har information om ett urval av populationen.
Intressepopulation	Intressepopulation är en population av objekt som motsvarar den användarna är intresserade av. Det kan till exempel vara företag, hushåll eller individer, eller delpopulationer av dessa.

Intressevariabler	Intressevariabler är variabler som motsvarar det som användarna är intresserade av. Det kan t.ex. vara inkomst för individer, omsättning för företagen etc.
Kalibrering	Kalibrering är en statistisk metod som används för att justera för bortfall.
Länkning	Länkning eller dataintegration innebär att man sammanfogar information från olika datakällor.
Målpopulation	Målpopulationen är den population som statistikproducenten valt att undersöka och dra slutsatser om.
Målvariabler	I en undersökning fångas användarnas behov och krav genom intressevariabler och utifrån dessa formuleras målvariabler. När uppgifter samlas in med hjälp av enkäter eller intervjuer formuleras frågor som kan mäta målvariabler. När digitala data används finns sällan möjligheten att påverka vilken information som samlas in.
Mätning	Mätning avser observationer och värden som samlas in, referens finns i SCB (2016).
Mätfel	Avvikelser från det sanna mätvärdet som beror på uppgiftslämnare, frågeformulär, intervjuare, teknisk enhet eller applikation som registrerar data.
Nattbefolkning (mobilnätsdata)	Genomsnittligt antal personer som vistas i ett område mellan kl 02:00 och 04:00 baserat på aggregerad och anonymiserad SIM-kortsinformation. Denna definition av nattbefolkning skiljer sig från traditionell sysselsättningsstatistik där förvärvsarbetande nattbefolkning beskriver var individer bor enligt register.
Representation	Representation avser hur väl målpopulationen avspeglas i tillgängliga data.
Skattning	Skattning eller estimat är det statistikvärde som räknas fram för den storhet man vill uttala sig om i en population.
Smart statistik	Smart statistik kan ses som en utökad roll för officiella statistiken i en värld genomsyrad med smarta tekniska lösningar. Smarta tekniker inkluderar automatiska, interaktiva lösningar som optimerar användningen av apparater och konsumentenheter i realtid. Själva statistiken kan sedan omvandlas via en smart teknisk lösning kopplad till smarta system till användbar information ur ett brett perspektiv i både tid och rum. Källa: Eurostat
Sverige i siffror	Applikationen Sverige i siffror används för att presentera visualiserad statistik på en fast installation, t.ex. ett touch-bord. Finns idag på Visualiseringscenter C i Norrköping och på Universeum i Göteborg.
Sverige i siffror webb	Ny iPad-anpassad och webbaserad version av Sverige i siffor som bygger på öppna data, finns ännu bara som designprototyp.
Säsongrensning	Säsongrensning är en statistisk metod som innebär att statistikvärden rensats på säsong- och kalendereffekter.

Validitet	Validitet handlar om i hur hög grad något mäter det man är intresserade av att mäta.
Viktning	Viktning är en statistisk metod för att justera för bortfall eller täckning.
Visualiseringscenter C	Visualiseringscenter C: ett konsortium bestående av Linköpings universitet, Norrköpings kommun m.fl. med mångårig erfarenhet av forskning och utveckling av visualiseringsmetoder och deras tillämpning inom en rad olika områden.

Kvalitetsindikatorer för digitala data

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 ”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.”

Bakgrund

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 digitala data. I september 2021 fick SCB dessutom detta regeringsuppdrag vilket resulterat i en uppsättning kvalitetsindikatorer (se Bilaga 1).

Utgångspunkten i arbetet med indikatorerna har varit att den kvalitet som avses är kvaliteten i den slutliga statistik som produceras med digitala data. Datakvalitet kan inte utvärderas fristående från datas användning i den slutliga statistiken. Användningen 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.

I föreliggande arbete läggs stor vikt vid länkning av digitala data 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 specifika 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.

Indikatorerna som föreslås ska vara mätbara så långt som möjligt, 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).

I fallet med digitala datakällor är SCB och andra statistikproducenter 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 kan användas både för en initial utvärdering och 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 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 metadata. Nya datakällor kan även vara data från administrativa register som inte funnits eller inte använts av statistikproducenten tidigare, till exempel

arbetsgivardeklarationer 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 elproduktion.

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

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.

Tabell 1. Klassificering enligt UNECE (2014)

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

Arbetet med regeringsuppdraget har resulterat i en vägledning (SCB 2023) som SCB kommer börja använda och som kommer vara tillgänglig för alla intresserade i det s.k. Statistikproduktionsstödet[1].

Under arbetets gång har diskussioner förts internt på SCB i olika grupper och med enskilda medarbetare. Utkast till dokument har cirkulerats vid två tillfällen för synpunkter. Arbetet har även presenterats för SCB:s Vetenskapliga råd samt på ett seminarium för statistikansvariga myndigheter.

I det följande kommer först en genomgång av relevanta arbeten internationellt, nationellt och internt inom SCB. Därefter beskrivs översiktligt den föreslagna strukturen för indikatorerna. I Bilaga 1 finns samtliga indikatorer redovisade.

[1] https://sps.scb.se/sites/vstod/statprod/Sidor/Processen.aspx

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 tagits fram 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 av 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.

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 kvalitetsindikatorer 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.

Kvalitetsindikatorerna – fördjupning kring hur de tagits fram

I detta avsnitt diskuteras begrepp och relevanta osäkerhetskällor när digitala data ska användas för statistik. Figur 1 visar vilka områden

kvalitetsindikatorerna täcker och i Bilaga 1 finns var och en av de åtta felkällorna beskrivna mer i detalj med en frågeguide och förslag på indikatorer.

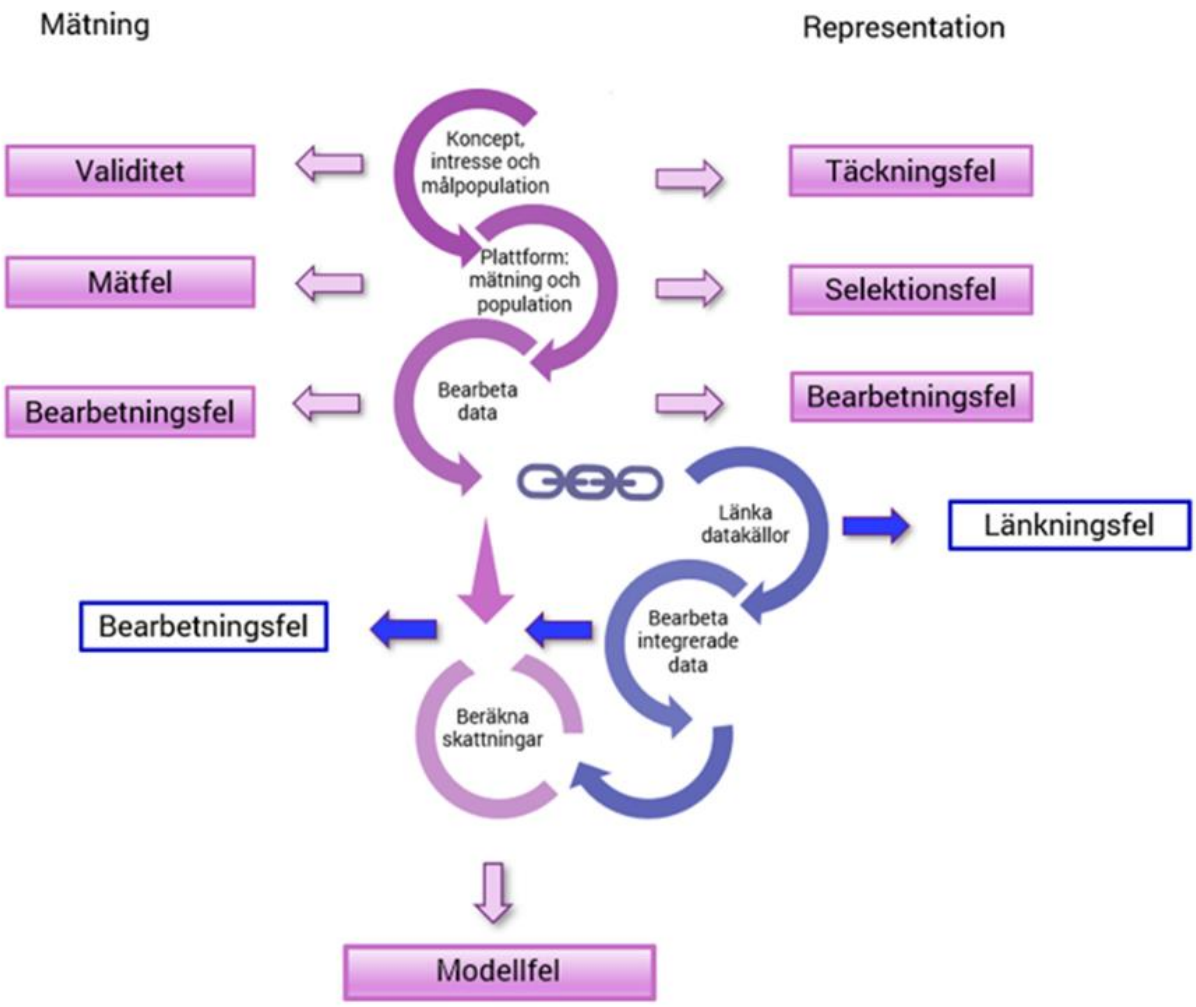
Inledning

Utgångspunkten är sammanfattningsvis följande:

- Tillförlitlighet enligt kvalitetsbegreppet som används för officiell statistik, samt relationen till Total Survey Error, 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, indikatorerna beskrivs mer ingående i Bilaga 1.

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 ramen (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. 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 kan identifieras 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 före detta kunder finns registrerade, så bidrar det till täckningsproblem.

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 troligen 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, pingar 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 länknings-variabler. 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 mobilnätsdata 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 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.

Diskussion

Det viktigaste första steget är att göra en kvalitativ utvärdering av den digitala datakällan för att förstå exakt vilka data den 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 Bilaga 1 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.

Alla 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. 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.

Litteraturgenomgången har visat att det pågår 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 bearbetade observations-

register (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. Internationellt finns det förslag på gemensamma ramverk för urvalsdata, administrativa data och digitala data (se till exempel Biemer 2016, Amaya et al 2020 eller Gootzen et al 2022). Detta är ett område som SCB på sikt skulle kunna arbeta vidare med.

Datasäkerhet

I uppdraget ingår att ”analysera hur man kan minska och hantera riskerna med aggregerade data samt att beakta de hot och risker som ett ökat databeroende innebär”. Informations- och cybersäkerhet är en central aspekt att beakta vid datadelning. Datadelning innebär komplexa och inte minst resurskrävande åtgärder både inom verksamheter som ska dela data och på nationell och internationell nivå.

Utgångspunkter för en riskanalys

SCB har genomfört en riskanalys baserat på Diggs vägledning för att tillgängliggöra information. Utifrån de sju övergripande principerna för tillgängliggörande av information har SCB identifierat flera informations- och cybersäkerhetsrisker kopplat till fyra av principerna.

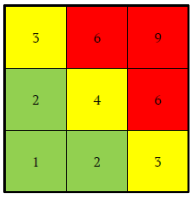
I riskanalysen tillämpas även SCB:s riktlinjer för riskhantering, vilka bygger på en modell för riskanalys med en tregradig skala för riskvärdering och matriser med tre färger för grafisk presentation av prioritet. En risk värderas utifrån sannolikhet och konsekvens.

Sannolikhet bedöms enligt följande:
1 = Låg sannolikhet
2 = Medelhög sannolikhet
3 = Hög sannolikhet

Konsekvens bedöms enligt följande:
1 = försumbar, lindrig
2 = kännbar
3 = allvarlig

Riskvärdet är lika med värdet för sannolikhet multiplicerat med värdet för konsekvens. Riskvärde värderas enligt följande:
1–2 = Lågt, bevaka men behöver inte åtgärdas
3–4 = Medelhögt, bevaka och åtgärda i ordinarie planerings eller beslutsprocess
6–9 = Högt, kritisk risk som ska åtgärdas omgående

Riskerna kan grafiskt åskådliggöras i en riskmatris med färger kopplade till de olika riskvärdena.



Röd innebär kritisk risk som behöver åtgärdas omgående, gul är medelstor risk där man bör planera för åtgärd och följa noga och grön är en låg risk som bevakas men inte behöver åtgärdas annat än om det är enkelt och inte medför några egentliga kostnader.

Riskanalysen avser data för smart statistik och begreppet smart statistik definieras här helt i enlighet med Eurostats definition, se begreppslista.

Riskanalysen har genomförts i flera workshoppar. I riskanalysgruppen ingick kompetenser inom informationssäkerhet, juridik, verksamhets-arkitektur, IT-arkitektur, processdokumentation, dataanalys och verksamhetsutveckling.

Sammanfattande riskanalys

SCB har genomfört en riskanalys och beskriver i Bilaga 2 vilka risker som har identifierats samt vilka åtgärder som föreslås för hantering av de allvarligaste riskerna. I tabell 2 finns en sammanfattande lista.

Tabell 2 – Riskanalys

Kort beskrivning	Sannolikhet	Konsekvens	Riskvärde
A. Risk att informationsklassning är en subjektiv bedömning från respektive organisation	3	2	6
B. Risk att information hanteras/skyddas på olika sätt beroende på att det finns olika lagkrav att beakta för olika organisationer.	2	3	6
C. Risk att olika aktörer nyttjar olika begrepp inom informationssäkerhet.	3	2	6
D. Risk på grund av olika lagkrav	3	2	6
E. Risk för kvalitetsbrister i data	1	3	3
F. Risk för leverans och överlåtelse av onödigt stora datamängder.	2	3	6
G. Risker när informationen kombineras eller samlas ihop (adderar ytterligare uppgifter)	2	3	6
H. Risk att tekniken hos sändande leverantör har begränsade möjligheter för att leverera information och det saknas möjlighet att påverka leverantören (felaktiga data skickas)	2	3	6
I. Risk att organisationer inte har tillräckliga resurser för att genomföra skyddsåtgärder	3	3	9
J. Risk att delning av viktig information är beroende av tillgänglighet hos leverantören av information	2	3	6
K. Risk att leverantören går i konkurs eller blir uppköpt, eller vill sluta leverera på grund av ny affärsstrategi	1	3	3
L. Risk för medveten manipulation av data hos leverantör för att påverka eller sabotera	1	3	3
M. Risk: Kostnadsaspekt – att man tar betalt eller ökar kostnaderna för att leverera data enligt säkerhetsperspektiv	2	2	4
N. Risk att säkerhetskrav försvårar användning och vidareutnyttjande	1	3	3
O. Metadata stämmer inte överens med innehållet i dataleveransen ur informationssäkerhetsperspektiv	1	2	2

Kort beskrivning av resultatet

Som nämnts tidigare är informations- och cybersäkerhet en central aspekt att beakta inom datadelning. Sett till resultaten från denna riskanalys kan en rad allvarliga konsekvenser förekomma i samband med att information delas mellan olika aktörer. De föreslagna åtgärderna avser aktiviteter som kan genomföras inom respektive organisation eller mellan olika organisationer, men också åtgärder som skulle behövas genomföras på nationell och internationell (EU) nivå.

Nedan följer en kortfattad beskrivning av de allvarligaste riskerna som har identifierats, de beskrivs tillsammans med lösningsförslag mer i detalj i Bilaga 2.

Tydligare rekommendationer och kunskap om gällande lagkrav

Eftersom dagens rekommendationer omfattar ett antal olika modeller som kan användas för informationsklassning kan det finnas risk att informationen inte får rätt skyddsåtgärder. På nationell nivå är det därför viktigt att det finns tydligare rekommendationer om en modell som ska användas för informationsklassning.

Risk finns också för felaktiga åtgärder (för svaga/för starka) samt onödigt långa startsträckor och resursslöseri för att reda ut vilka modeller som har använts och vilket lagkrav som gäller för vilken organisation.

Vidare noteras att privata/kommunala aktörer inte alltid omfattas av samma lagkrav som statliga myndigheter. Detta kan resultera i att informationen inte hanteras korrekt eftersom kunskapen och kompetensen gällande informationssäkerhet skiljer sig åt.

Gemensam begreppskatalog

Avsaknad av en gemensam begreppskatalog gör att samma begrepp kan tolkas på flera olika sätt. Vissa begrepp kan till och med användas på olika språk (svenska eller engelska) inom samma organisation.

Vi föreslår att MSB:s begreppskatalog (termbanken för informationssäkerhet) utvecklas vidare och kunskap om denna katalog sprids inom offentlig och privat sektor.

Dialog och tydliga rutiner för hantering av (stora) datamängder

Vid leverans av ogallrat data eller när verksamheten behöver experimentera med nya datakällor finns det risker att större datamängder än som motsvarar det faktiska behovet kommer in i verksamheten. Konsekvensen blir bland annat att information hanteras i onödan och att datadelningen riskerar att bryta mot gällande lagstiftning. Det är då viktigt att inom verksamheten finns etablerade och tydliga rutiner för hantering av data.

Hantering av kombinerade data
Kombinerade data kan i vissa fall ge ett högre skyddsvärde än de enskilda delarna. Det är således viktigt med transparens när data samlas ihop eller kombineras. Det behövs dessutom nationell vägledning om vilken organisation/aktör som bär ansvaret för att bedöma nivån på materialet som ska informationsklassas.

När möjlighet till kryptering saknas
Dataleverans via utländska molntjänster utan möjlighet till kryptering kan resultera i otillåten överföring till tredje land. Om leveranssättet hos respektive leverantör inte kan påverkas ska ingen information från respektive leverantör inhämtas.

Minimikrav vid resursbrist och avsaknad av incitament
Begräsningar i ekonomi och kompetens, samt olika lagkrav för offentliga/privata aktörer kan medföra att skyddsåtgärder inte kommer på plats i tillräcklig omfattning.

Som minimikrav är det viktigt att säkerställa att leverantören följer rekommendationer och krav från Nationell Cybersäkerhet eller ISO 27000.

Rutiner för hantering av data vid driftproblem eller sabotage
Vid driftproblem finns det risk att det inte är möjligt för leverantören att leverera i tid. Det kan också finnas risk för sabotage genom till exempel överbelastningsattacker (DDOS). Icke-funktionella krav gällande exempelvis krav på redundans och säkerhetskopiering ska då beaktas.

Mobilnätsdata – ny statistik

I uppdraget har ingått att ta fram förslag till ”ny, smart statistik baserad på delning av mobilitetsdata. SCB ska inhämta synpunkter från Trafikanalys, Tillväxtverket och privata aktörer. Förslaget ska även visa hur berörda myndigheter, så långt möjligt, kan ersätta direktinsamlade uppgifter med mobilitetsdata. Lösningsförslaget ska gå att applicera även på andra data som kan användas för att ta fram smart statistik.”

Arbetet med mobilitetsdata har resulterat i många insikter och lärdomar för SCB och övriga involverade (Trafikanalys, Tillväxtverket och teleoperatörer). Hur behöver datakällan vara beskaffad för att fungera för att få fram ny smart statistik? Är det faktiskt möjligt att ersätta direktinsamlad information från enkäter med mobilitetsdata och på så sätt göra befintlig statistik på nytt sätt? Vilken ny smart statistik kan vara lämplig att publicera baserat på mobilitetsdata? Vilka generella lärdomar går att dra som är applicerbara även på andra digitala datakällor?

Under uppdragets gång har SCB presenterat resultaten i några olika internationella sammanhang, vilka framgår av referenslistan.

Att lära känna en digital datakälla

Tidigare samarbete med mobilnätsoperatörer

Fyra mobilnätsoperatörer är verksamma i Sverige. Den största är Telia, som också är verksam i Danmark, Finland, Norge och de baltiska staterna. Telia har skapat en Telia Crowd Insights-plattform för att sälja aggregerade och anonymiserade mobilnätsdata). Data från Telia användes i Sverige under covid-19-pandemin för att snabbt visa befolkningsrörelserna efter pandemins utbrott.

SCB har samarbetat med Telia inom ramen för två projekt som föregått samarbetet i detta uppdrag, för att på sätt lära känna, bearbeta och analysera aggregerade signaldata för mobiltelefoner. Det första projektet som genomfördes 2020 visade att aggregerade och anonymiserade mobilnätsdata lämpar sig väl för att ta fram ny statistik om dynamisk befolkning och befolkningsrörelser. Sådan ny statistik kan komplettera befintlig pendlings-, turism- och resestatistik. Arbetet har också visat att den exakta nivån av förändringar i befolkningsrörelser var svår att kvantifiera, insynen i processerna skulle kunna förbättras liksom möjligheten att beskriva statistikens kvalitet (Vlag, P. 2021).

Det andra projektet genomfördes 2021 och var inriktad på att förbättra kvaliteten och insynen i processerna. Det resulterade i ökad kunskap om kvaliteten och förbättringar av metodiken när mobilnätsdata används (Vlag, P. et al. 2022a, Vlag P. et al. 2022b). Det tredje projektet, som drivits inom ramen för detta uppdrag, har pågått från september 2021 till december 2022 med fokus på

- villkor och kvalitetskriterier för användning av mobilnätsdata (och nya datakällor i allmänhet) för officiell statistik
- utveckling av ny ”smart” statistik
- förslag till minskning av statistiska undersökningar med hjälp av mobilnätsdata.

SCB hade hoppats att med regeringsuppdraget som grund kunna etablera ett samarbete även med övriga mobilnätsoperatörer, men efter flera försök tillkom endast en av dem, Tre. Upplägget för samarbetet var detsamma som i fallet med Telia, att SCB skulle få dataleveranser i utbyte mot kunskapsdelning. På så sätt kunde samarbetet drivas utan krav på marknadsmässig ersättning från Telia eller Tre.

Telia och Tre har tillsammans ungefär 50 % av marknaden, där Telia har en marknadsandel på 35–40 % och Tre har en marknadsandel på 10–15 %. Dessa marknadsandelar är ungefär konstanta över tiden på nationell nivå sedan 2019. Det finns regionala skillnader i marknadsandelar, delvis på grund av att inte alla mobilnätsoperatörer har ett nätverk som täcker hela Sverige.

Dataskydd och dataleveranser

Datasäkerhet och personuppgiftsskydd är viktiga frågor vid hantering av mobilnätsdata och alla mobilnätsoperatörer i Sverige har höga krav på dataintegritet. Även om uppgifter som SCB hanterar inom ramen för statistikverksamheten skyddas av statistiksekretess har uppdraget genomförts baserat på aggregerade uppgifter som lämnats till SCB. Den processinformation som lämnats skyddas även den av sekretess, eftersom den riskerar att avslöja konfidentiella affärsprocesser hos en enskild operatör.

Dataleveranserna har bestått av aggregerade data om **aktiviteter** och **rörelser/resor**. Definitionen av en aktivitet är en stationär signal mätt minst 40 minuter inom en timme i samma rutnätscell. En aktivitet kan relatera till att vara hemma, att vara på arbetet eller i skolan etc. Den kan pågå från 40 minuter upp till en dag. Genom att välja 40 minuters tröskel är aktiviteten unik för den timmen. Det kan därför tolkas som den huvudsakliga lokaliseringen av en mobiltelefon under en timme. Nattbefolkning har definierats som registrerade mobilaktiviteter mellan 02:00 och 04:00 och dagbefolkning har definierats som perioden 09:00 - 15:00.

Definitionen av en resa är en rörelse mellan två områden. Detta är fallet när den stationära signalen är mindre än 40 minuter i samma rutnätscell inom en timme. Generellt tillämpas detta så att rörelser i stadsmiljö kan identifieras om mobilanvändaren flyttar sig 300–500 meter, medan det behöver vara över 1 kilometer på landsbygden. Över vatten eller i områden med nästan inga invånare kan det vara svårt att identifiera rörelser. Resor kan relateras till dessa rörelser. Vid långväga resor behöver dock pauser och mellanstopp beaktas och det finns behov av implementerbara definitioner för pauser och mellanstopp för mobilnätsdata.

För att kunna använda mobilnätsdata tillsammans med andra data hos SCB innehöll leveranserna uppgifter om aktiviteter och rörelser på olika geografiska nivåer: rutnät, DeSo[2], kommun och län per timme från Telia och på kommunnivå från Tre. Dataleveranserna skedde månadsvis och har täckt in uppgifter för 2019–2022 från Telia och 2021–2022 från Tre. Data har analyserats på kommun- och länsnivå för utveckling av smart statistik, medan uppgifter på finare nivåer (rutnät och DeSo) har använts för kvalitetssäkring, detektering av avbrott och artificiella signaler i tidsserien och korrigering för dem.

Bearbetning av mobilnätsdata hos operatörerna

Varje dag genererar miljontals mobilanvändare miljarder datapunkter när de använder sina telefoner. Dessa data samlas in, bearbetas och placeras i en datasjö som ska användas i aktiviteter för t.ex. affärsändamål som att förbättra mobiloperatörernas nätverksprestanda. De data som genereras kommer både från aktiv användning och från passiv användning. Exempel på aktiv användning är att skicka ett textmeddelande eller ta emot data. Exempel på passiv användning kan vara om en telefon byter anslutning från en antenn till en annan, eller när en kontroll sker om telefonen fortfarande finns inom samma antenn. Data från passiv användning kallas ”signaldata” inom Telekombranschen. SCB har använt sådana signaldata för att skapa statistik, något som blivit vedertaget även internationellt.

Mobilnätsoperatörerna börjar med att kombinera dessa råa nätverksdata och infrastrukturdata i så kallade antennsignaler. Antennsignaler innehåller hundratals miljoner rader från alla enheter (SIM-kort) i nätverket, där en rad representerar en anslutning, händelse eller signal från enheten. Antennsignaler består av tre huvudkomponenter.

- en anonym nyckelidentifierare
- tidsstämpel (vid vilken punkt SIM-kortet lämnade ett spår i nätverket)

[2] https://www.scb.se/hitta-statistik/regional-statistik-och-kartor/regionala-indelningar/deso---demografiska-statistikomraden/

- cell-ID (vid vilket celltorn, eller Base Transceiver Station, lämnade SIM-kortet ett spår vid den givna tidsstämpeln).

En anonymisering av antennsignaler sker utifrån regelverket i GPDR och genomförs vid regelbundna tidsintervaller: Telia tillämpade tidigare 24 timmar, men har efter juridisk prövning förlängt från 24 timmar till 7 dagar från hösten 2022. Tre tillämpar 30 dagar. Detta begränsar hur länge man kan följa ett SIM-kort. GDPR styr också att mobiloperatörer inte använder kunddata i sin process.

Mobilnätsdata behandlades hos Telia och Tre i tre huvudsteg:

1. Databehandling av källdata: anonymisering, beräkning av ”missade” tidsluckor, filtrering för störningar i nätet och dataleveranser etc.
2. Geolokalisering som kopplar signaler från de mottagande antennerna till ett geografiskt rutnät.
3. Viktning av urvalet av mobiler till befolkningen i stort.

Som tidigare beskrivits så har utgångspunkten för samarbetet varit att SCB bara ska behöva ha den information som behövs för att kunna ta fram statistik. I tidigare projekt har insikter om processen hos Telia lett till att SCB kunnat bedöma hur mobilnätsdata kan användas och vilken kvalitet data har. Transparensen i hur statistiken tagits fram måste anpassas till skyddet av processen hos dataägarna.

Slutsatser kring mobilnätsdata som datakälla

SCB har arbetat med mobilnätsdata sedan 2020 och de viktigaste slutsatserna från arbetet med data från Telia och Tre, inklusive lärdomar från tidigare projekt med Telia är:

- Mobilnätsdata är en omfattande och komplex datakälla som kräver särskild beräkningskapacitet, trots att det som levereras till SCB handlar om aggregerade och anonymiserade data
- Det tar tid att förstå och beskriva mobilnätsoperatörernas processer, men det är nödvändigt för att förstå datakvaliteten. Detaljerade processbeskrivningar måste skyddas av sekretess.
- Det är en fördel att använda data från mobilnätsoperatörer med nationell täckning.
- Geolokalisering och att relatera mobila enheter till personer är viktiga steg vid bearbetning av data.
- Processtegen till och med geolokaliseringen behöver utföras av mobilnätsoperatören eftersom det kräver teknisk kunskap om operatörens antennsystem. Systemen är olika för olika aktörer.
- Av de fyra operatör som har eget nätverk i Sverige idag är Telia och Tre intresserade av fortsatt samarbete, övriga två har meddelat att de inte kan leverera data till SCB av ekonomiska skäl.

Det finns en komplett och detaljerad processbeskrivning av Telias processer som möjliggör användning av mobilnätsdata och det finns tidsserier baserade på Telias data från oktober 2018. Dataleveransen från Tre har kompletterat Telias data, men täcker en kortare tidsperiod och har inte samma detaljerade processbeskrivning.

Ny smart statistik

Nästa steg i uppdraget har varit att se om mobilnätsdata kan användas för att ta fram ny smart statistik. Statistik som bygger på mobilnätsdata faller in under definitionen av smart statistik, genom att den är baserad på digital information som har bearbetats i olika steg redan hos dataägaren innan den anonymiseras och aggregeras och skickas till SCB för ytterligare datakonsolidering och analys. Viktiga aspekter att beakta för smart statistik är: definierade användarbehov, datakvalitet, åtkomst till (bearbetade) digitala data, konsistens med traditionell statistik och att data ger kontinuerliga tidsserier.

Mobilnätsdata för statistikändamål

Enligt uppgifter från Post- och telestyrelsen, PTS, finns det 1,4 SIM-kort per capita i Sverige. Detta gör att det behövs viktning avseende antalet SIM-kort i förhållande till totalpopulationen, även om data från alla operatörer skulle finnas tillgängliga. Urvalet i viktningen är enheter med ett SIM-kort registrerat i Sverige, oavsett typ av användare eller typ av enhet.

Grunden för den viktmodell som används av Telia (och som SCB även använt för Tre-data) är relativt enkel:

- Den första aktiviteten under en 24-timmarsperiod betraktas som ”Hemma” (vissa undantag finns).
- Summan av ”Hemma” korrigeras för stationära enheter i en kommun
- Viktfaktorn w per kommun beräknas på följande sätt:
    - W = N/n med
    - N = befolkning enligt SCB:s befolkningsstatistik i en kommun
    - n = summan av ”Hemma” eller den så kallade baslinjen för en referensperiod. En referensperiod är en period utanför semestersäsongerna för vilken det kan antas att personer befinner sig på sin folkbokförda adress (=”Hemma”) under natten.

Viktfaktorer appliceras på den mobila enheten under en 24-timmarsperiod, oavsett om enheten flyttar till andra platser under dagen. Faktorn beror på placeringen av enhetens första signal under dagen. Samma viktfaktor gäller även vid förflyttningar från en plats till en annan. Viktproceduren upprepas var 24:e timme. Den kritiska

faktorn i processen är fastställandet av referensperioden, eller den s.k. baslinjen.

Huvudantagandet bakom detta tillvägagångssätt är att under vardagar utanför semesterperioder tillbringar de flesta personer och deras mobiltelefoner natten på sin hemadress och baslinjen (= korrigerad summa av enheter) kan därför relateras till befolkningsstatistiken.

SCB har identifierat ett antal kritiska faktorer vid beräkningen av viktfaktorn:

- stabilitet hos den så kallade baslinjen (Summan av ”Hemma”)
- den geografiska aggregeringsnivån för viktberäkningarna
- uppdateringsfrekvens för att uppdatera viktfaktorerna.

Viktfaktorn bör inte omräknas dagligen, så att verkliga variationer i övernattningar på en viss plats (t.ex. under ledigheter och helger) kan fångas upp. Den bör dock uppdateras regelbundet, för att hålla viktmodellen uppdaterad på grund av förändringar i antenntäckning och förändringar i andel mobiler.

Analyser av data för 2020 visade att en instabil baslinje för viktningen är den främsta orsaken till instabilitet och oförklarliga hopp i tidsserier. Instabiliteten berodde både på artificiella orsaker (förändringar i nätet som påverkade geolokaliseringen, temporära minskningar i tillhandahållandet av data på grund av fel och ändringar i nätverket, mindre förändringar i mobilnätsoperatörens kundandel) och verkliga förändringar i hur personer rört sig på grund av ledigheter.

En viktig slutsats blev att mobiluppgifterna visade befolkningsrörelser under arbetsperioder och semesterperioder som var mer varierade än vad som ursprungligen antogs i viktmodellen, på grund av ökad flexibilitet i semesterperioder, arbete på olika platser etc.

Iakttagelserna 2020 tyder på att kvaliteten på de viktade mobil-uppgifterna var tillräcklig för att upptäcka kortsiktiga trender i befolkningsrörelser på grund av ledigheter och förändrat beteende till följd av covid-19-pandemin. Observationerna från 2020 tyder också på att kvaliteten var otillräcklig för den officiella statistiken, eftersom den officiella statistiken ofta baseras på nivåuppskattningar och långsiktiga förändringar.

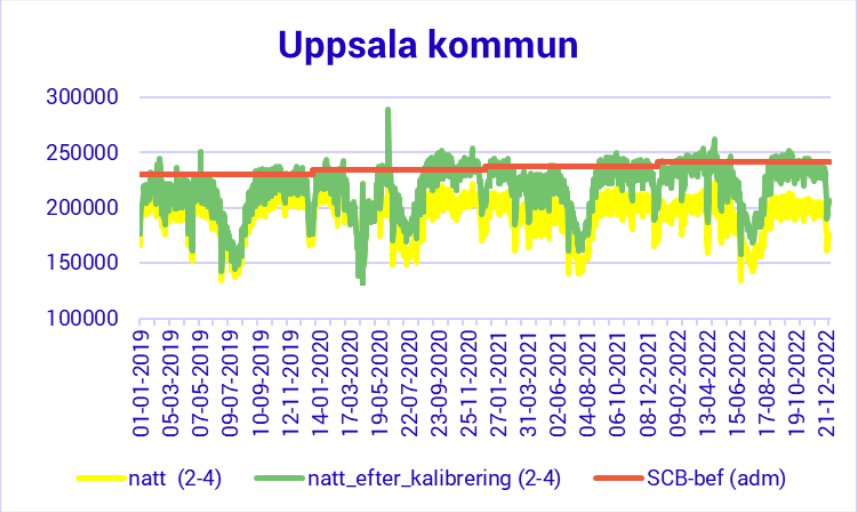
Kvalitetsförbättringar

SCB har föreslagit förbättringar av viktningsmetoden för mobilnätsdata som Telia delvis implementerat. En månadsvis referensrapport har tagits fram där de månadsvisa uppskattningar som härrör från mobilnätsoperatörernas data jämförs med relaterad officiell statistik från SCB såsom befolkningsstatistik, pendlingsstatistik, utbildnings- och inkomststatistik. Syftet är att skapa en kvalitetssäkrad

statistikprocess genom att kontrollera rimligheten hos mobilnätsoperatörernas uppskattningar och upptäcka nivåförändringar i tidsserier i ett tidigt skede. De ska även hjälpa till att upptäcka verkliga förändringar i befolkningsrörelser och utvärdera vilken ”smart statistik” som kan utvecklas ur mobilnätsdata.

Tidsserier med nivåförändringar kan inträffa även efter att viktnings-tekniken förbättrats. Dessa nivåförändringar är ofta svåra att förutsäga eftersom de kan orsakas av mänskligt beteende (t.ex. genom att personer använder mobiler för olika ändamål). Därför behöver en extra kvalitetskontroll läggas till i form av kalibrering.

Kalibreringsmodellen bygger på antagandet att aktiviteter från mobiler som genererar den genomsnittliga nattbefolkningen oktober-november ska motsvara nivån i SCB:s registerbaserade befolkningsstatistik. Om så inte är fallet behöver tidsserierna kalibreras så att de överensstämmer. Kalibreringsmetoden behöver utvecklas ytterligare för att upptäcka nivåförändringar i ett tidigt skede och korrigera för dessa förändringar. Ett exempel visas i figur 2. SCB har även utvecklat en liknande viktmodell för Tre-data.

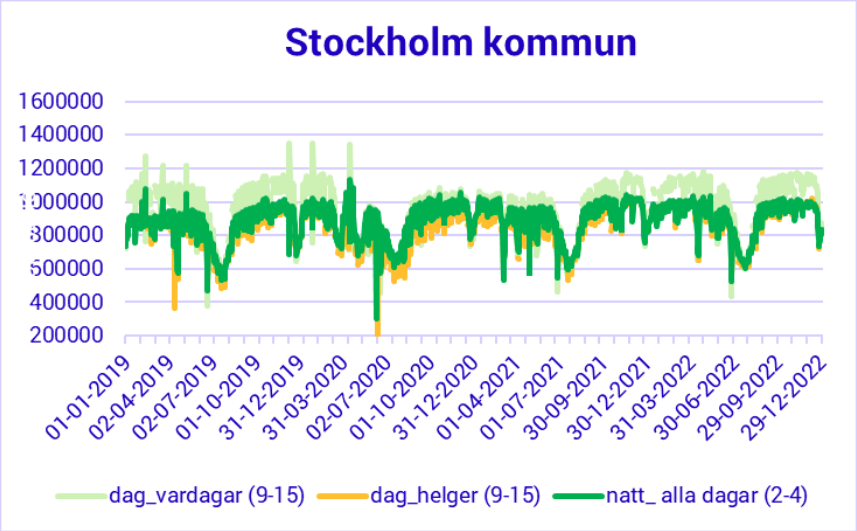


Figur 2 Nattbefolkningen i Uppsala kommun. Den gula linjen visar aktivitet för nattbefolkningen som härrör från operatörens standarddatabehandling. Den gröna linjen visar aktivitet för nattbefolkningen korrigerad för nivåförändringar (efter kalibrering). Den röda linjen visar SCB:s befolkningsstatistik.

Kurvan över nattbefolkning visar verkliga förändringar i Uppsala under sommar, jul och påsk. Dessa kan förklaras av människor som lämnar staden, mönstret liknar andra städer med studenter. Lägre nattpopulationer under helger generellt är sannolikt också relaterade till människor som lämnar staden (gäller t.ex. studenter eller veckopendlare). Detsamma gäller förmodligen de kraftigt minskade nattaktiviteterna i början av pandemin (mars 2020). Det går därför att

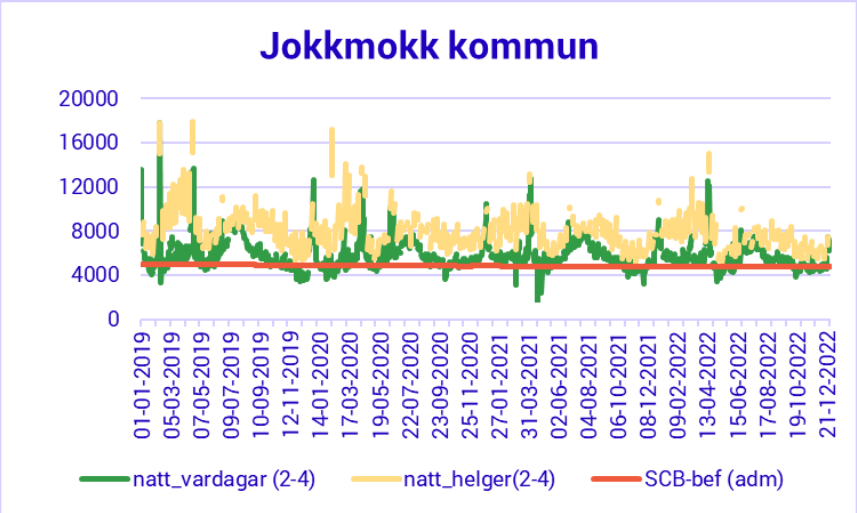
förklara att den genomsnittliga årliga nattbefolkningen i Uppsala kommun är lägre än SCB:s registerbaserade befolkningsstatistik (röd linje).

Figurerna 3 och 4 visar två tidsserier (2019–2022) med skattningar av dagliga nattpopulationer baserade på mobilnätsdata från Telia. Figur 3 visar Stockholms kommun och figur 4 visar Jokkmokks kommun.



Figur 3: Tidsserie för befolkningsaktiviteter för Stockholms kommun under tre olika tidsintervall: vardagar 09:00-15:00, helger 09:00-15:00 och nätter hela veckan 02:00-04:00.

Observera att före covid-19-pandemin är dagsbefolkningen under vardagar (ljusgrön linje) högre i Stockholms kommun. Detta kan man även se om man lägger till rörelser / resor, som tyder på pendling till Stockholm för arbete. Denna högre dagliga befolkning under vardagar försvinner under covid-19-pandemin och återhämtar sig bara långsamt under våren 2022. Under helgerna är dagbefolkningen i allmänhet något lägre än nattbefolkningen (gula linjen). Det finns också minder systematiska skillnader mellan den genomsnittliga nattbefolkningen under vardagar och helger i Stockholm – liksom i andra större städer. Med längre tidsserier kan dessa skillnader analyseras bättre.



Figur 4: Tidsserie för befolkningsaktiviteter för Jokkmokks kommun under två olika tidsintervall: vardagsnätter 02:00-04:00 och helgnätter 02:00-04:00.

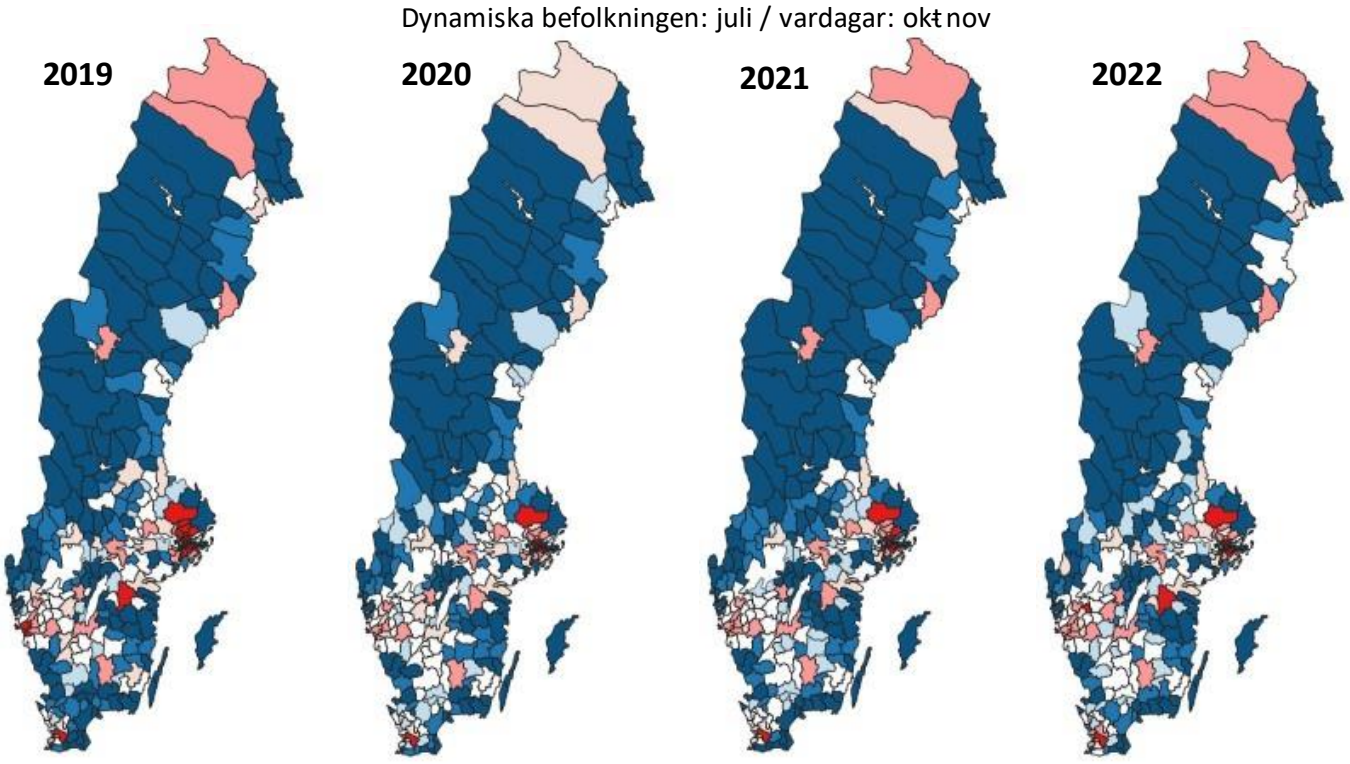
För Jokkmokk finns ett säsongsmönster i tidsserien med högre befolkningsaktivitet under senvinter/tidig vår än under sommaren. Men det mest slående inslaget i denna serie är den högre nivån på nattbefolkning under helgerna. Detta kan förklaras av människor som tillbringar helgerna i sitt hus i Jokkmokk men under veckan arbetar i städerna längs Bottenviken, eller i gruvorna i närliggande Gällivare och Kiruna. Det går att förklara att för Jokkmokk är det årliga genomsnittet för nattbefolkningen högre än SCB:s registerbaserade befolknings-statistik (röd linje). En liten nivåförändring upptäcks dock före och efter utbrottet av covid-19-pandemin, vilket försvårat en exakt kvantifiering.

Exemplen med Stockholm och Jokkmokk visar tydligt att det finns säsongsvariationer i var befolkningen befinner sig, något som inte går att fånga i den vanliga befolkningsstatistiken. För vissa kommuner är nattbefolkningen, baserad på mobilaktivitet, högre än befolknings-statistiken när man tittar månadsvis, kvartalsvis och ibland även som årsgenomsnitt. För andra kommuner ligger den lägre. Flera myndigheter har visat intresse för denna typ av analyser och fortsatt utveckling av statistik som jämför traditionell befolkningsstatistik med dynamisk befolkningsstatistik baserad på mobilnätsdata.

Användningsområden för dynamisk befolkningsstatistik

Genom att ta fram dynamisk befolkningsstatistik går det att visa säsongsmässiga skillnader i befolkningsaktiviteter per kommun, se figur 5. Kartorna illustrerar en befolkningsrörelse från städer och förorter till landsbygden under sommaren. Statistiken har både ett allmänintresse och kan konkret användas av t.ex. Trafikverket och för att dimensionera hälso- och sjukvård.

Förhållandet mellan dag-/nattpopulationer kan relateras till daglig pendling och visa effekter av pandemin i våra dagliga pendlingsmönster. Statistiken baseras i det fallet både på aktiviteter och rörelser.



Figur 5: Exempel på säsongsvariationer i nattbefolkningens aktiviteter per kommun. Jämförelse mellan juli (alla dagar) och oktober-november (vardagar) samma år.

Kartorna visar jämförelser för nattbefolkningen under juli och oktober-november (vardagar) samma år, d.v.s. semesterns effekter på var folkbokförda personer i Sverige verkligen befinner sig:

- mörkblå områden: 20+ % högre nattbefolkning i juli än i oktober-november
- blå områden: 10–20 % högre nattbefolkning i juli än i oktober – november
- ljusblå: 5–10 % högre nattbefolkningen i juli än i oktober – november
- ljusröd: 5–10 % lägre nattbefolkning i juli än i oktober – november.
- röd: 10–20 % lägre nattbefolkning i juli än i oktober – november
- mörkröd: -20 % lägre nattbefolkning i juli än i oktober – november.
- vit: oförändrad nattbefolkning.

Blå områden är kommuner med högre befolkningsaktivitet i juli. Dessa är i de flesta fall kommuner vid kusten, i turistområden i tätbefolkade kommuner med många fritidshus. Röda områden är kommuner med lägre befolkningsaktivitet i juli. Dessa är i de flesta fall städer och förorter eller i den nordligaste delen kommuner med gruvor. Effekterna av rekommendationer under pandemitiden är också detekterbara eftersom skillnaderna mellan nattbefolkning under juli och arbetsdagar i oktober - november var något mindre 2020, särskilt jämfört med 2019 och 2022.

Slutsatser kring mobilnätsdata för ny smart statistik

Mobilnätsdata är en värdefull datakälla för smart statistik om dynamiska populationer och befolkningsrörelser. En ansats skulle kunna vara att ta fram månads- eller veckostatistik om nattbefolkning och förhållandet mellan dag- och nattbefolkningens aktivitet baserad på mobilnätsdata samt relaterad resestatistik. Detta skulle i så fall inledningsvis publiceras som statistik under uppbyggnad, för att senare övergå till officiell statistik när kvaliteten bedöms stabil.

För att satsa på den här typen av statistik krävs att referensmetoderna för att upptäcka fel och korrigera nivån i statistiken löpande kan testas och utvecklas, både nationellt och internationellt. Det behöver också finnas en nytta av statistiken som motiverar kostnaderna för datainköp, eftersom mobilnätsoperatörerna använder mobilnätsdata för kommersiell verksamhet. Eventuella framtida partnerskap med mobilnätsoperatörer behöver fortsätta att bygga på en affärsmodell som inkluderar ömsesidig utveckling.

Smart statistik kan tas fram även av andra än SCB och den möjligheten skulle gynnas om det går att etablera en ”mobildatahubb” för statistik som alla statistikansvariga myndigheter kan nyttja. På så sätt skulle kostnaden för att kvalitetssäkra mobilnätsdata för statistik kunna delas mellan fler intressenter. Frågor om juridiska förutsättningar, finansiering och teknisk förmåga behöver lösas innan det går att realisera en sådan datahubb.

SCB välkomnar ett uppdrag från regeringen för att utreda sådana förutsättningar ur ett mer generellt dataperspektiv, där mobilnätsdata kan vara ett användningsfall. Detta utvecklas ytterligare i avsnittet med sammanfattande slutsatser och rekommendationer.

Kan mobilitetsdata ersätta direktinsamling till befintlig statistik?

Potentialen fortsatt svår att konkretisera

I uppdraget fanns planer på att ta fram förslag på hur resultaten av statistik om den dynamiska befolkningen i kombination med

relaterande datakällor (trafikmätningar, olika typer av öppna data) kan utvecklas till smart statistik. Sådan statistik skulle kunna komplettera nuvarande statistik om pendlare, turism och resvanor och minska behovet av direktinsamling.

SCB har tillsammans med Trafikanalys och Tillväxtverket, som ansvarar för officiell statistik där mobilnätsdata kan vara en intressant datakälla, identifierat användningsområden för officiell statistik baserad på mobilnätsdata.

Tabell 3: Statistik där mobilnätsdata kan vara aktuell som datakälla

Statistik	Ansvar	Officiell statistik idag
Dynamisk befolkning	SCB	Nej
Pendling	SCB	Ja
It-användning i företag	SCB	Ja
Inkvartering, turism	Tillväxtverket	Ja
Svenskars resande	Tillväxtverket	Nej
Resvanor	Trafikanalys	Ja
Fritidsfiske	Havs- och vattenmyndigheten.	Ja

Mobildata kan ha potential att ge uppskattningar av **antal resor** med bättre kvalitet, snabbare och mer detaljerat än dagens enkätundersökningar. Undersökningar om resvanor och svenskars resande skulle kunna utformas på ett nytt sätt, göras mer sällan och bara fokusera på kompletterande statistiska kärnfrågor, t.ex. om syftet med resan, som är mer stabilt över tid.

Det pågår diskussion om fortsatt arbete både hos Tillväxtverket och Trafikanalys, men det har inte blivit några konkreta förslag som kan inkluderas inom ramen för detta uppdrag.

Utmaningar vid integration av datakällor

SCB har parallellt med detta uppdrag bedrivit ett antal utvecklings-initiativ, som inte rör just mobilnätsdata som datakälla. Ett sådant som också levererat konkret resultat är att det nu finns en helt registerbaserad statistik om befolkningens arbetsmarknadsstatus (SCB – BAS). Det pågår även flera initiativ inom den ekonomiska statistiken med syfte att identifiera nya sätt att ta fram statistiken som kan minska uppgiftslämnarbördan för företag och organisationer. En ökad användning av tillgängliga datakällor, inklusive sådana som är aggregerade som är fallet med mobilnätsdata, ger SCB möjlighet att effektivisera statistikproduktionen och att öka statistikens kvalitet.

Täcker tillgängliga data användarbehoven?
Utgångspunkten för statistikproduktion baserad på integration av datakällor är överenstämmelsen mellan tillgängliga data och användarbehov. I idealfallet överensstämmer användarbehov, intressepopulation och intressevariabler med tillgängliga data och en urvalsundersökning kan då ersättas. Alternativt finns icke-urvalsbaserade data tillgängliga men de överensstämmer inte helt med användarbehov, det kan exempelvis vara så att målpopulationen inte överensstämmer med behovet, eller att variabelinnehållet är för litet. Även om den här informationen inte kan ersätta en urvalsundersökning kan den användas som hjälpinformation, i urvalsdragning eller för att komplettera variabelinnehåll.

Större variation i hur statistiken designas
En statistikproduktion baserad på integration av datakällor kan innebära en större variation i hur statistikproduktion utförs med avseende på undersökningsdesign och bearbetning. Vissa undersökningar kommer kunna ersättas med tillgängliga datakällor, men i många fall kommer urval behövas. Antingen för att komplettera variabelinnehåll eller för att hantera täckningsproblem i tillgängliga data. Därmed förändras urvalsundersökningens roll i vissa fall, från att vara den primära datakällan till att i stället användas för att berika och/eller validera tillgängliga datakällor.

God kunskap om tillgängliga datakällor
Kunskap om tillgängliga datakällor, hur de kan användas/kombineras i en undersökningsdesign är en förutsättning för att SCB bättre ska kunna nyttja tillgänglig information i statistikproduktionen. Det handlar om en kombination av ämneskunskap och kunskap om teori och tillämpning av undersökningsdesign. Det nya för SCB här är den ökade variationen i tillämpning av undersökningsdesign och den nära kopplingen till kunskap om hur tillgängliga datakällor kan användas.

Mer flexibel statistikproduktion
För SCB innebär en ökad användning av tillgängliga datakällor att statistikproduktionen behöver vara mer flexibel än idag. Statistikproduktionen behöver kunna ställa om efter nya förutsättningar i indata. En utmaning för SCB är att säkerställa att det finns förutsättningar för en effektiv och stabil produktion utan att varje undersökning blir ett stort resurskrävande utvecklingsprojekt.

Objektivitet och osäkerhetsmått
Officiell statistik ska vara objektiv, vilket ställer krav på hur den designas för att på ett tillförlitligt sätt mäta olika fenomen. Man kan tala om att statistiken måste ha en tillförlitlig inferens. En förutsättning för inferens är att statistiken baseras på ett statistiskt ramverk (teori), med ett ramverk kan statistikens kvalitet beskrivas. Med en ökande

användning av tillgängliga datakällor kommer icke-urvalsfel stå för en relativt större andel av statistikens osäkerhet. Och SCB behöver kunna beskriva osäkerheten och samtidigt använda metoder för att motverka den.

Utmaningar för SCB

En anpassning av dagens statistikproduktion kommer ta tid och innebär en utmaning för SCB, genom att den statistikproduktion SCB har idag är resultatet av metodutveckling som pågått under mycket lång tid. SCB behöver arbeta aktivt för att förändra statistikproduktionen successivt. En omställning till inferens baserad på integration av datakällor innebär att SCB behöver tänka på nya sätt i undersökningsdesign och bearbetning av data:

- Urvalsundersökningarnas roll blir mer varierad, från primär datakälla till komplement för att berika och/eller validera tillgängliga data
- Metoder för inferens finns, SCB behöver inte utveckla några nya metoder, arbetet handlar i stället om att identifiera möjligheter i nya data och att få förändrade undersökningsdesigner i produktion.
- Av den totala osäkerheten växer betydelsen av icke-urvalsfel och med det följer behov av att kunna kommunicera osäkerheten till användare och motverka effekten via undersökningsdesign och bearbetning.
- En redesign till produktion baserad på integration av datakällor kan förväntas pågå under lång tid, förflyttningen behöver därför organiseras på ett så effektivt sätt som möjligt och bygga på strukturer för att ta tillvara på lärdomar.
- Integration av tillgängliga datakällor för med sig en ökad variation i tillämpning av teori för undersökningsdesign
- Ökad användning av tillgängliga datakällor i kombination med urvalsdata för med sig nya behov av kompetens inom metodområdet. Samtidigt ökar behovet av en kombination av metod och ämneskompetens
- Dagens fabrikslika produktion behöver kunna överföras till en mer flexibel produktion, utan att varje undersökning blir ett eget stort resurskrävande utvecklingsprojekt.

Samarbete och samverkan

SCB har samarbetat och samverkat med ett antal olika parter i denna del av det aktuella uppdraget.

Ägare av mobilnätverken

Samarbete med Telia och Tre för att säkerställa dataleveranser och tillhörande dokumentation för uppdraget, återkoppla resultat av uppdraget och diskutera möjligheter för eventuellt fortsatt samarbete efter uppdraget. Via Telia kom vi även i kontakt med utländska

mobiloperatörer för att kunna jämföra hur kvaliteten i Telias processer står sig jämfört med andra nätoperatörer. Det visade sig att svenska operatörer ligger långt framme vad gäller transparens och utveckling jämfört med andra operatörer i Europa, vad gäller insamling och bearbetning av mobilnätsdata.

Myndigheter
Samverkan med Trafikanalys (statistikansvarig myndighet för transport- och resvanestatistiken) för att dela kunskap, undersöka möjligheter att ersätta frågor i resvanestatistiken, minska frekvens i direktinsamlade data.

Samverkan med Tillväxtverket (statistikansvarig myndighet för turismstatistik, bl.a. svenskars resande och inkvarteringsstatistik för informationsutbyte.

Samverkan med Trafikverket och Naturvårdsverket for att inventera användningsbehov av ny smart statistik.

Andra statistikbyråer
Samverkan med statistikbyråerna i Norge, Finland, Tyskland och Österrike för att inventera användningsbehov, utveckla standardbegrepp och kunskapsutbyte.

Öppna data – föreskrifter

Denna del av uppdraget har bestått av att i samarbete med representanter från övriga statistikansvariga myndigheter ta fram en föreskrift som reglerar hur officiell statistik ska tillgängliggöras som öppna data i maskinläsbara format. Målsättningen med föreskriften är att all officiell statistik ska gå att hitta via dataportal.se. En kartläggning av vilka resurser och stöd som behövs för att efterleva föreskriftens innehåll har även genomförts.

Bakgrund

SCB har tillsammans med övriga statistikansvariga myndigheter tagit fram en gemensam målbild för den officiella statistiken som Rådet för den officiella statistiken ställde sig bakom 2019. Målbilden har fått rubriken ”Vi beskriver Sverige” och fokuserar på tre områden, varav ett handlar om att det ska vara lätt att hitta rätt statistik. Målbilden tar sikte på 2025 och åtföljs av en rullande handlingsplan. Efter dialog mellan Digg och SCB, och som följd av målbildsarbetet, inkluderades ett förslag att ta fram styrande föreskrifter om öppna data i Diggs regeringsuppdrag om öppna data, öppen och datadriven innovation och AI (Digg 2019).

En aktivitet i den rullande handlingsplanen som genomfördes 2020/21 har handlat om att ge en nulägesbeskrivning av ett antal statistikansvariga myndigheters tillgängliggörande av statistik som öppna data (SCB 2021a).

En generell slutsats från båda dessa aktiviteter är att det finns behov av att stärka förmågan att tillgängliggöra statistik som öppna data så att det går lätt att hitta all officiell statistik på ett ställe, på dataportal.se.

Föreskriftsarbetet

Rådsarbetsgruppen för användbarhet och tillgänglighet[3] genomförde 2020 en översyn av de riktlinjer som fanns för elektronisk publicering. Resultatet av det arbetet visade att det var mest lämpligt att infoga de reglerna i gällande föreskrifter.

SCB utfärdade för drygt tjugo år sedan föreskrifter och allmänna råd för offentliggörande m.m. av officiell statistik (SCB-FS 2002:16). Föreskrifterna har ändrats 2016 (SCB-FS 2016-27) och 2020 (SCB-FS 2020:16). Inom ramen för regeringsuppdraget kunde SCB bedriva ett arbete med att införliva riktlinjerna i dessa föreskrifter och allmänna råd under hösten 2021 och våren 2022.

[3] https://scb.se/sam-forum/hem/radet/radets-arbetsgrupper/anvandbarhet-och-tillganglighet/

De nya föreskrifterna omfattar en ny paragraf som säger att den officiella statistiken ska tillgängliggöras som öppna data i maskinläsbar form och nya allmänna råd till denna paragraf.

De allmänna råden till denna paragraf säger att datamängder bör tillgängliggöras under licensen CC0. Rapporter, analyser och övrigt redaktionellt material som inte entydigt är öppna data bör tillgängliggöras under licensen CC-BY. Öppna standarder och format som möjliggör ett tillgängliggörande på dataportalen hos Myndigheten för digital förvaltning bör användas. Användarhandledningar och beskrivningar bör finnas tillgängliga.

Dessa allmänna råd har baserats på slutsatserna i en rättslig utredning om licensiering av öppna data som gjordes på SCB 2021 (SCB2021b). Den utredningen låg till grund för att SCB gick över från att licensiera sina öppna data enligt CC-BY till en differentierad licensiering på det sätt som återspeglas i förslaget till allmänna råd.

I skälen för SCB:s beslut om en mer differentierad licensiering refereras till att både Digg och Patent- och Registreringsverket (PRV) rekommenderar svenska myndigheter att ha licenser enligt Creative Commons. I rekommendationen har Digg och PRV angett att för statliga myndigheter kan CC0 kan vara lämpligt i flera fall, medan CC BY är lämpligt i vissa fall. Där framgår sammanfattningsvis följande:

- För information och data som skapas hos myndighet och som inte är föremål för upphovsrättsligt eller annat immaterialrättsligt skydd rekommenderas märkningen PDM eller CC0.
- För databaser som skapas hos myndigheter för myndighetsutövning rekommenderas märkningen CC0.
- För övrig information som skapas hos myndighet som är föremål för upphovsrättsligt skydd som verk eller prestation rekommenderas licensen CC BY 4.0.

Licenserna CC-BY, CC0 och PDM innebär öppen tillgång till ett verk eller en prestation. Användaren är inte begränsad i sin användning av verket eller prestationen eller för egna bearbetningar av dessa. Även CC BY-SA räknas som en öppen licens, men innebär att den som bearbetar materialet inte kan sätta någon annan licens på det bearbetade materialet. Andra licenser enligt Creative Commons innebär på olika sätt en villkorad användning, de kallas för delningslicenser.

Rekommendationen är inte precis anpassad efter de statistiska produkter som SCB framställer. Den pekar ändå mot att merparten av den statistik som SCB framställer snarare skulle ha licensen CC0 än CC BY. Även om CC0 innebär att källa inte behöver anges, utgör den inget hinder mot att en användare skulle ange SCB som källa.

Mot denna bakgrund har SCB gått över till CC0 för SCB:s öppna data, enklast avgränsat till maskinläsbara data (statistik och geodata) som tillgängliggörs i statistikdatabasen eller på SCB:s geodataplattform. Analyser, texter, fördjupande artiklar, visualiseringar och annat material har även fortsättningsvis CC-BY, liksom produkter som levereras på uppdrag ihop med SCB:s allmänna villkor för avtal och överenskommelser.

Genom att lägga in denna typ av differentiering i de allmänna råden för alla statistikansvariga myndigheter ser SCB att möjligheten ökar till enkel tillgång till officiell statistik som öppna data.

Process och fortsatt stöd

Alla statistikansvariga myndigheter har haft möjlighet att lämna remissynpunkter på den nya föreskriften med tillhörande konsekvensutredning (se Bilaga 3. SCB har omhändertagit synpunkterna så långt som möjligt. SCB:s interna arbetsgrupp har under arbetet haft stöd från ämnesexperter, Digg, jurister samt en undergrupp till arbetsgruppen för användbarhet och tillgänglighet för statistikansvariga myndigheter (AoT), som leds av SCB.

Från hösten 2021 när SCB fick uppdraget har arbetet i AoT fokuserat på processen med att ta fram den nya föreskriften genom samverkan och information, och genom att öka kunskapen om vikten av att tillgängliggöra statistik som öppna data. Det sistnämnda har exempelvis gjorts genom att statistikansvariga myndigheter som kommit längre i arbetet har fått dela med sig av framgångsfaktorer och externa intressenter som Wikimedia har inspirerat kring nyttan med vidareanvändning av officiell statistik, samt att Digg har informerat om möjligheterna som dataportalen och den digitala arenan ger.

SCB har även delat med sig av exempel på tekniska lösningar såsom PxWeb. Några statistikansvariga myndigheter har även uttryckt ett behov av stöd och resurser för att kunna uppfylla kraven på öppna data som ställs i föreskriften, bland annat genom ökad samverkan och kunskapsdelning mellan myndigheter som använder PxWeb för att tillgängliggöra sin statistik som öppna data.

PxWeb[4] används för att sprida statistik i dynamiska figurer från antingen s.k. PX-filer eller en SQL-databas. PxWeb tillhandahålls gratis till statliga myndigheter och kommuner samt till utomsvenska statistikbyråer och till internationella organisationer som presenterar statistik. I dialog med Digg har SCB lagt upp en artikel[5] på den nya

[4] https://www.scb.se/vara-tjanster/statistikprogram-for-px-filer/PxWeb/

[5] https://beta.dataportal.se/aktuellt/pxweb

betaversionen av dataportal.se, för att öka kännedomen och intresset för PxWeb.

SCB har som svar på det uttryckta behovet av ökad samverkan kring kunskapsdelning kring tekniska lösningar bildat ett användarforum för PxWeb och relaterade programvaror för att tillgängliggöra statistik. Deltagare i användarforumet består av intresserade från andra statistikansvariga myndigheter, och syftar till att utbyta erfarenheter och kunskaper och få stöd av varandra.

Under 2022 har SCB som ett led i arbetet med föreskriften även prioriterat utvecklingen av PxWeb genom att den senaste versionen PxWeb ger möjlighet att skapa en dcat-ap.xml-fil som möjliggör automatiserad leverans av data från PxWeb till dataportal.se. Detta ska underlätta för statistikansvariga myndigheter att leva upp till den nya föreskriften och målsättningen att all officiell statistik ska gå att hitta via dataportal.se.

Öppna data – visualisering av statistik

I uppdraget har ingått att SCB ska främja användningen av statistikansvariga myndigheters öppna data för visualisering i samarbete med Visualiseringscenter C och Dataspelsbranschen.

Bakgrund

I arbetet med målbilden för den officiella statistiken, som nämnts tidigare, har en viktig drivkraft för att all officiell statistik ska finnas som öppna data varit att det möjliggör återanvändning och visualisering. Statistikansvariga myndigheter har i allmänhet olika former av visualisering på den egna webbplatsen också, men när statistiken blir tillgänglig som öppna data kan vem som helst bygga visualiseringar och kombinera statistik med t.ex. kartor eller annan öppna data.

I uppdraget anges specifikt att samarbete ska ske med Visualiseringscenter C, ett konsortium bestående av Linköpings universitet, Norrköpings kommun m.fl. och har mångårig erfarenhet av forskning och utveckling av visualiseringsmetoder och deras tillämpning inom en rad olika områden, särskilt installationen som heter Sverige i siffror[6]. Centret har även en publik arena för vetenskapskommunikation med storskaliga visualiseringsmiljöer, som till exempel en domteater, samt utställningar med interaktiva installationer. Centret har rönt stor nationell och internationell uppmärksamhet för sin verksamhet och speciellt sina nydanande resultat inom vetenskapskommunikation till allmänheten.

Under perioden september 2022 till januari 2023 har en projektgrupp bestående av forskningsingenjörer Yin He, Ludvig Mangs och Måns Gezelius (Linköpings universitet, LIU), forskare Selcan Mutgan från Institutionen för analytisk Sociologi (IAS), pedagog Matilda Stafstedt från Visualiseringscenter C samt formgivarna/gränssnittsdesignerna Love Jacobsson och Anders Olsen från SCB:s kommunikationsenhet arbetat tillsammans i ett projekt för att utveckla Sverige i siffror. Projektet har avrapporterat till en styrgrupp ledd av sektionschefen på SCB:s kommunikationsenhet. Projektet har avrapporterat sitt resultat i en slutrapport (LIU 2023).

[6] SCB har tidigare haft ett eget projekt som gick under rubriken ”Sverige i siffror” för barn i skolåldern kring 6:e klass, men i detta sammanhang avses den visualiseringslösning som Visualiseringscenter C har tagit fram med samma namn.

Projektet har sammanfört SCB:s kompetenser inom statistik, kartor och digital tillgänglighet med LIU:s erfarenheter av att använda Sverige i siffror 1.0 i publik verksamhet samt utveckling av applikationen tillsammans med forskare på LIUs Institutionen för analytisk sociologi (IAS). Sverige i siffror används på två av landets främsta science center (Visualiseringscenter C i Norrköping och Vislab på Universeum i Göteborg) som nås av många hundra tusen besökare per år och Linköpings universitets didaktiska forskare har varit involverade.

Prioriterade målgrupper för Sverige i siffror idag är elever inom gymnasium och högstadium. En viktig aspekt av visualiseringen är det pedagogiska värdet i att unga ovana användare få möjlighet att själva kunna laborera med statistiken, vilket ökar förståelsen, ger nya insikter och väcker nyfikenhet.

Kontakter har även tagits med Dataspelsbranschen, mer specifikt företaget Mojang som ligger bakom spelet Minecraft. Dessa ledde inte vidare till ett konkret samarbete, utan SCB valde att fokusera på projektet med Visualiseringscenter C.

Utveckling av Sverige i siffror

Målsättningar för projektet var att utifrån den visualiseringskompetens som finns vid Visualiseringscenter C:

- utveckla ett attraktivt sätt att presentera SCB:s statistik i form av öppna data genom installationen Sverige i siffror
- utforska möjligheterna att förse Sverige i siffror med öppna data från andra statistikansvariga myndigheter samt
- ta fram en webbversion/Ipadversion av Sverige i siffror för ökad användning i undervisningssyfte.

Det är viktigt att förstå vilka som kommer vara de främsta användarna av en uppdaterad Sverige i siffror. Det finns en stor mängd data och många alternativ för att sammanställa, visualisera och förklara fakta genom statistik. Verktyget kommer med en webbversion att finnas tillgängligt på Internet och i publika platser och Visualiseringscenter C behöver ta höjd för att kunna utöka funktionalitet över tid.

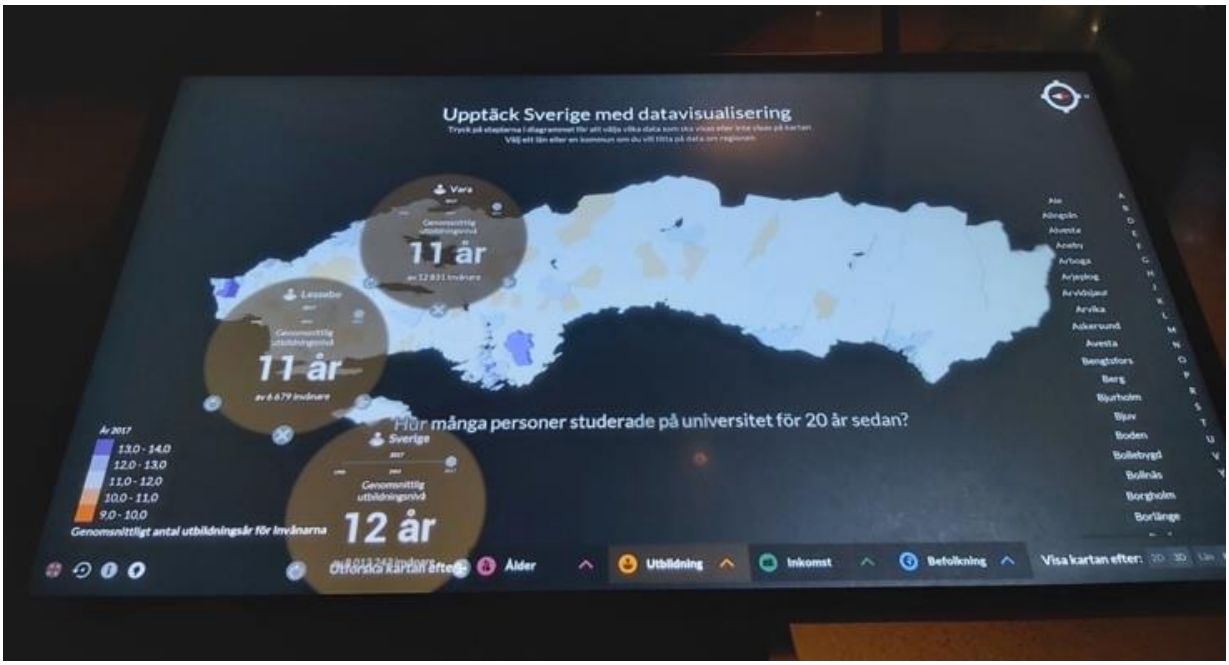
I stället för att rikta in sig på några få användargrupper skapades en portfölj av målgrupper. Det hjälper till att definiera innehållsval, kommunikation och användargränssnitt för de primära användarna som är elever i gymnasieklasser och besökare i publika kunskapsbaserade museer. Samtidigt kan man behålla en neutral design och en öppen struktur för att på sikt kunna expandera till fler användargrupper.

SCB har bidragit med en alternativ kartvisualisering som kan ersätta nuvarande 3D-lösning som SCB upplevde svår att läsa av grafiskt och förstå. SCB:s förslag var en tillgängligare visualisering med beskrivande

data i tabeller och diagram på fasta platser utanför kartan. SCB har även bidragit med expertkunskaper om hur den typ av statistisk som SCB produceras bäst visualiseras på ett användarvänligt sätt som uppfyller kraven på tillgänglighet enligt lagen om tillgänglighet till digital offentlig service (DOS).

Utställningsapplikation som bygger på öppna data

En del av projektet har handlat om att vidareutveckla den existerande applikationen Sverige i Siffror. Applikationen finns idag som en del av utställningarna på Visualiseringscenter C i Norrköping och på Universeum i Göteborg. Vidareutvecklingen innebar att modifiera applikationen för att kunna hantera öppna data från SCB:s statistikdatabas. Idag använder sig applikationen av data från en SCB-databas som inte är publikt tillgänglig.



Figur 6: Sverige i siffror, touch-bord.

En av de större utmaningarna, förutom att anpassa applikationen till andra sorters data än de som tidigare använts, är att generalisera både datas användning i applikationen (t.ex. filtrering) och dess format, så att flera typer av dataset ska gå att använda i applikationen. Tester visade att det är möjligt att anpassa Sverige i Siffror till ett nytt, mer flexibelt dataformat, baserat på ett exempel-dataset från SCB:s statistikdatabas. För att applikationen ska vara användbar för flera olika dataset (från SCB eller andra källor) behöver den kunna hantera och konvertera dataset till rätt format. Det behövs funktionalitet för att kunna byta mellan de olika dataseten som applikationen stödjer och det behövs förbättrad prestanda.

Den nya funktionaliteten för öppna data har implementerats och arbetet är dokumenterat på LIU:s Gitlab. Funktionaliteten har testats på

ett första dataset med statistik om utbildning från SCB:s statistikdatabas[7].

Ett antal förbättringsområden har identifierats för att SCB:s öppna data ska fungera optimalt för användning i Sverige i siffror. Det rör t.ex. begränsningar i hur mycket data som kan laddas hem samtidigt, att kunna filtrera i data med stöd av API:et och hantering av information i fotnoter. Man önskar också ett sätt att hitta vilka dataset som stämmer in på vissa urvalskriterier, t.ex. att de går att redovisa för kommuner och län.

På motsvarande sätt har förbättringsområden identifierats för Sverige i siffror, så att applikationen kan vara flexibel mot nya data så länge de baseras på regioner (kommun/län) samt årtal. Övriga filter, i detta fall ålder, utbildning och kön skulle i princip kunna vara vilken data som helst.

Vad gäller att undersöka möjligheten till att använda även andra myndigheters data som källa i Sverige i siffror så valde projektteamet att börja utredningen med webbversionen och sedan utvärdera de tekniska avvägningarna mellan Sverige i siffror med mer tillgång till öppna data eller att ersätta den med en storbildsoptimerad Sverige i siffror iPad/webbversion för touchboard i framtiden.

Designprototyp för en webbversion av Sverige i siffror

Som underlag för designval för en webbversion (Ipad-version) av Sverige i siffror gjordes en genomgång av SCB:s webbplats och andras webbplatser med geografiskt relaterade data (t.ex. IPCC WGI Interactive Altas, USA Census 2020 Map etc.). Genomgången visade att det finns enormt många olika designalternativ. Med många alternativ var det utmanande att välja primära användare som grund för designvalen. Projektet landade i studenter och besökare till publika kunskapsbaserade museum som primär målgrupp. Kompetensmässigt befinner sig webbversionen av Sverige i siffror på ingångs- eller mellannivå, vilket kan jämföras med SCB:s webbplats som är mer anpassad till avancerade användare.

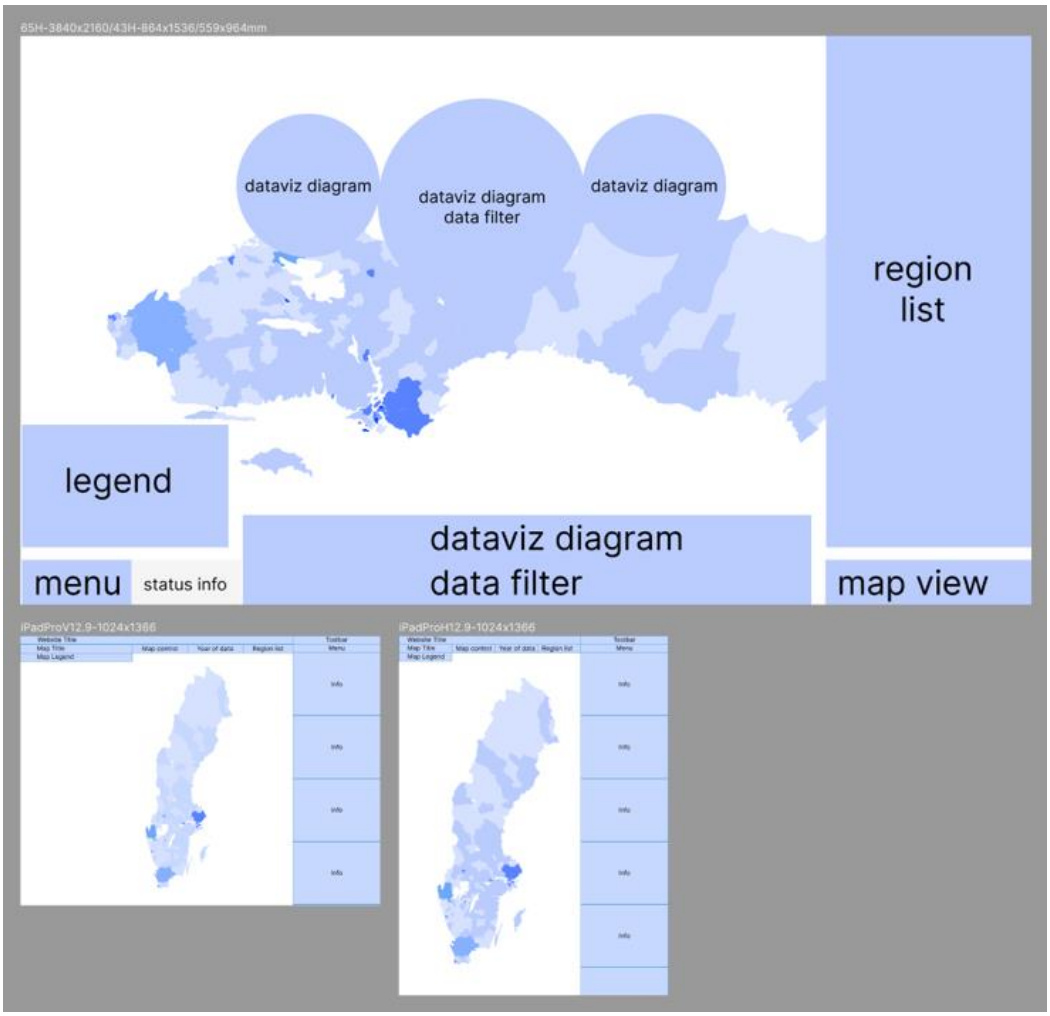
En utgångspunkt var att informationen och berättelserna i webbversionen för att skapa värde för användarna måste ha tre grundläggande dimensioner: plats, tid och demografi (t.ex. ålder och kön). Detta gör det enklare för användare att knyta an till ämnen baserat på sina preferenser.

De olika kompetenserna i projektet bidrog till att utveckla intressanta berättelser för att skapa engagemang och nyfikenhet till att lära sig mer, som skulle kunna implementeras i kommande versioner av Sverige i

[7] https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__UF__UF0506__UF0506B/Utbildning

siffror. På så sätt kan användare av en framtida webbversion baserad på designprototypen nå spännande sociala frågeställningar med hjälp av data och diagram. Det kan t.ex. handla om att visa mönster av könsrelaterad yrkessegregation, förändringar i infrastruktur över tid eller demokratisering av universitetsutbildningen.

Med tanke på användares olika datakunskaper, kognitiva kapacitet och intresseområden valdes interaktivitet för att bygga upp informationsnivåer från enkelt till mer djupgående. Sammanfattningsvis är utgångspunkten att försöka följa Design Principles for Tools to Support Creative Thinking (M. Resnick, et al.) när prototypen utformas, vilket är en lärdom från tidigare UI-tester med Sverige i siffror. Det ska vara enkelt att komma igång, högt i tak – skalbart och erbjudas flera ingångar.



Figur 7: Sverige i siffror, designprototyp för webbversion

En mer omfattande beskrivning av genomförande och förslag på ett koncept och klickbar prototyp för Sverige i siffror i en Ipad-/webbversion finns i projektrapporten (LIU 2023). Prototypen finns även

tillgänglig här: Figma-designprojektet - Sweden-in-numbers designprototyp[8]

Leveransen inom ramen för projektet är en designprototyp, men projektet har även tagit fram en kort beskrivning av implementationen av en webbaserad lösning av Sverige i Siffror med rekommendationer på ramverk och tekniska lösningar.

Valet av en webbaserad lösning handlar framför allt om att ge Visualiseringscenter C möjligheten att anpassa applikationen till ett flertal plattformar, vilket inkluderar alla enheter med en webbläsare så som en större touchskärm för utställning eller en mindre Ipad. Sidan kan med enkelhet offentliggöras och läggas ut på internet för en större utsträckning användare.

Fördelen med en webbsida är att den går att visa på många enheter, men det finns också möjligheter att exportera sidan till en Windows-applikation, exempelvis för utställning.

Figur 8: Sverige i siffror, exempel på hur en webbversion kan se ut på dator och Ipad.

Under utvecklingsperioden har test av öppna datakällor koncentrerats till ett av SCB:s öppna dataset. Utifrån dessa tester är bedömningen att det är fullt möjligt att använda öppna data både från SCB och andra myndigheter, men att det kräver anpassningar och visst manuellt arbete för att standardisera data.

[8] https://www.figma.com/file/LqoCFLqxpPQOxAy94JZvxk/Sweden-in-Number?node-id=1%3A3528&t=41azgh9m4NT247Mp-1 (kräver inloggning)

Slutsatser och rekommendationer

SCB har genom uppdraget att främja delning och nyttiggörande av data för smart statistik (I2021/02417) kunnat bidra inom två insatsområden i regeringens datastrategi från 2021: (1) ”Ökad tillgång till data” och (2) ”Öppen och kontrollerad datadelning”.

Insatsområde 1: Ökad tillgång till data

Mål: Tillgången till data som delas öppet eller kontrollerat ökar kontinuerligt. 2023 ska det finnas delade data för att möta prioriterade samhällsutmaningar samt inom en majoritet av de dataområden som definieras i EU:s datastrategi. Följande områden prioriteras: prioriterade samhällsutmaningar, europeiska gemensamma dataområden, områden med identifierade värdefulla datamängder samt områden där data efterfrågas av t.ex. företag.

SCB har genom föreskrifter om öppna data ökat tillgången till den officiella statistiken som öppna data och genom samarbetet med Visualiseringscenter C har steg tagits mot att också visualisera den statistik som ligger fritt tillgänglig som öppna data.

Ökad användning av statistik och företagsdata

Möjligheten att vidareutnyttja den officiella statistiken ökar genom att den läggs ut i maskinläsbart format och går att hitta på dataportal.se som öppna data. SCB har arbetat med att öka utbudet av öppna data sedan 2017, bland annat genom att tillgängliggöra mer statistik på låg regional nivå liksom att göra data enklare att återanvända (Haldorson 2022). Genom det aktuella uppdraget och möjligheten att utnyttja SCB:s föreskriftsrätt inom systemet för den officiella statistiken kommer det från 2024 gå att hitta och använda all officiell statistik som öppna data.

Nyttan med öppna data uppstår först när den används, det finns många goda exempel på hur statistik kan användas av olika målgrupper när den tillgängliggjorts som öppna data och sedan vidareförädlats (FN 2022). SCB avser att fortsätta stödja Digg i arbetet med dataportal.se och den digitala arenan, vilket också innebär att ge stöd till övriga statistik-ansvariga myndigheter vad gäller öppna data.

SCB:s allmänna företagsregister skulle skapa stor nytta om det kan tillgängliggöras avgiftsfritt. Registret regleras idag av en förordning som säger att SCB ska ta ut avgifter, vilket har försvårat bl.a. myndighetssamarbete kring företagsuppgifter. SCB har inte kunnat prioritera fram de anslagsmedel som krävs för att hålla god kvalitet i registret inom ram, utan har äskat om utökat anslag i budgetunderlaget.

SCB berörs av de genomförandebestämmelser som nu träder i kraft vad gäller särskilt värdefulla datamängder, kopplade till EU:s direktiv för öppna data. Statistik och företagsuppgifter är bland kategorierna för sådana datamängder som ska tillgängliggöras avgiftsfritt. Det är i dagsläget klart vilken statistik som omfattas, företagsuppgifter återstår att klargöra i detalj.

Rekommendation 1

SCB föreslår att regeringen tillför medel till SCB för allmänna företagsregistret enlig äskande i myndighetens budgetunderlag för 2024–2026.

Sverige i siffror på webben

Samarbetet mellan Visualiseringscenter C och SCB har varit lärorikt för båda parter, SCB har kunnat utforska sätt att visualisera öppna data och även fått möjlighet att synas i nya sammanhang (Science center i Norrköping och Universeum i Göteborg).

Utöver ett förfinat gränssnitt och möjligheten att läsa in nya källor av öppna data är projektet ett första steg i att möjliggöra att Sverige i siffror kan bäras ut från science centers till skolklasser. Den skulle kunna bli en brygga mellan lärare, skolbesök på science center och efterarbete i klassrummen efter ett inspirerande upplevelsebesök på ett science center.

Linköpings universitet har planerat att göra en vetenskaplig uppföljning av projektet tillsammans med berörda forskargrupper, t.ex. inom didaktisk forskning, för att utvärdera det som kommit fram genom samarbetet.

Rekommendation 2

SCB föreslår att regeringen ger SCB tillsammans med Visualiserings-center C i uppdrag att implementera den framtagna designprototypen för Sverige i siffor i webbversion. Uppdraget bör omfatta aktiviteter för att sprida användningen av Sverige i siffror till grund- och gymnasieskolor, vilket därmed ökar nyttjandet av öppna data från SCB och andra myndigheter.

Insatsområde 2: Öppen och kontrollerad datadelning

Mål 2023: Statliga myndigheter och statliga företag har en god förmåga att dela data både på ett öppet och kontrollerat sätt. Svenska företag har en god förmåga att dela data och är delaktiga i utvecklingen av och kan utnyttja de uppbyggda datamarknaderna. Offentliga data, inklusive forskningsdata, ska där så är lämpligt, vara så öppna som möjligt och så stängda som nödvändigt.

SCB har genom uppdraget kunnat visa på möjligheterna att ta fram ny, smart statistik baserat på mobilnätsdata. SCB har tagit fram

kvalitetsindikatorer för digitala data som ska användas för statistik, liksom genomfört en riskanalys baserat på Diggs vägledning för att tillgängliggöra information.

Relevant och oberoende statistik kräver säkrad datatillgång

SCB arbetar kontinuerligt med att utveckla den officiella statistiken, bland annat baserat på digitala data, så att statistiken kan fortsätta vara relevant och oberoende. Som arbetet med kvalitetsindikatorerna har visat finns det många aspekter att beakta innan en digital datakälla kan tas i bruk. Det behöver vara transparent hur data används och vilken kvalitet statistiken får när den baseras på integrerade data.

I takt med att administrativa och digitala data kan återanvändas alltmer minskar behovet att belasta framför allt företag och organisationer med att lämna uppgifter för statistik. En datakälla, t.ex. momsregister eller digitala årsredovisningar, kan användas till många olika statistikprodukter. Detta gör samtidigt statistikproduktionen känslig för förändringar i dessa datakällor. Dataförsörjningen till SCB från såväl myndigheter som privata dataägare behöver vara reglerad och stabil över tid.

Uppgifter som behövs för officiell statistik kan begäras in med stöd av lagen och förordningen om den officiella statistiken, men så länge statistik är under uppbyggnad – vilket t.ex. gäller för statistik baserad på mobilnätsdata – är SCB och andra statistikansvariga myndigheter hänvisade till frivilliga överenskommelser. Här kan finnas behov av förstärkt rättsligt stöd, t.ex. i linje med Norges nya statistiklagstiftning som anger att samråd ska ske med statistikmyndigheten vid förändringar i administrativa register. I dagsläget bedömer SCB dock att det är tillräckligt att kunna föreskriva om uppgiftslämnarplikt och att ingå frivilliga överenskommelser.

För att fullt ut kunna dra nytta av digitala datakällor tillsammans med administrativa data från myndigheter och företag pågår arbete med att ställa om SCB:s statistikproduktion. SCB har inrättat en särskild dataavdelning sedan hösten 2021, med uppgift att bl.a. effektivisera mottagningen av datakällor till SCB och möjliggöra ökad återanvändning av redan tillgängliga data för statistik (internt) och för statistik och forskning (externt). I SCB:s budgetunderlag görs äskanden för att kunna accelerera takten i uppbyggnaden av en teknisk plattform för datainsamling och datahantering. Digitala data ställer extra höga krav på teknisk bearbetningskapacitet.

Rekommendation 3

SCB föreslår att regeringen tillför medel till SCB för det tekniklyft som krävs för att öka takten i användningen av digitala och andra data enligt äskande i myndighetens budgetunderlag för 2024–2026.

SCB:s förmåga att bistå andra myndigheter med data

SCB:s verksamhet består idag främst av att ta fram officiell statistik inom utpekade statistikområden samt att på uppdrag ta fram statistik och tillhandahålla mikrodata åt myndigheter, forskare m.fl. SCB har i uppgift att samordna systemet för den officiella statistiken och det har nyligen etablerats en rådsarbetsgrupp för datafrågor för att ge statistikansvariga myndigheter ett forum för erfarenhetsutbyte kring dataförsörjning. SCB deltar också i arbetet inom Ena, både som grunddataproducent inom två av grunddatadomänerna och som kravställare på grunddata. Det finns stordriftsfördelar att data som ska användas för statistik kvalitetssäkras och förädlas på ett ställe för att sedan användas båda av SCB och av andra statistikansvariga myndigheter.

När det gäller digitala data som t.ex. mobilnätsdata finns också en kostnad för datainköp att beakta. SCB bedömer att ett bra sätt att komma vidare med smart statistik baserad på mobilnätsdata, förutom att delta i internationella projekt, kan vara att etablera en ”mobildatahubb” för statistik. Kostnaden om varje myndighet som behöver mobilnätsdata ska köpa in data och kvalitetssäkra den blir betydligt högre än om t.ex. SCB etablerar en sådan tjänst. Erfarenheterna från samarbete med Telia och Tre är positiva och det bör gå att sätta upp en mer långsiktig dataförsörjning baserad på ett partnerskap med mobilnätsoperatörer baserat på en affärsmodell som inkluderar ömsesidig utveckling.

Det pågår också diskussioner kring vilka myndigheter som ska få rollen som behörigt organ enligt dataförvaltningsförordningen. SCB kan se fördelar med att ge SCB en sådan roll för att kunna ge tekniskt stöd för att tillhandahålla en säker behandlingsmiljö för att ge tillgång till vidareutnyttjande av data.

Rekommendation 4

SCB föreslår fortsatt dialog med Regeringskansliet i frågor som rör data för smart statistik, där SCB kan bidra med bred kompetens inom t.ex. dataförvaltning, datasäkerhet, juridik och tekniskt stöd. SCB ser även fram emot fortsatt dialog rörande SCB:s uppdrag relaterat till dataförvaltningsförordningen.

Nationellt och internationellt samarbete i datafrågor

Uppdraget har inneburit att SCB kunnat fördjupa kunskapen om vilka indikatorer som behövs för att bedöma kvaliteten i digitala datakällor som används för statistik. Resultatet från uppdraget kommer kunna användas brett, både genom det statistikproduktionsstöd som SCB tillhandahåller för andra statistikansvariga myndigheter och genom dataportal.se / den digitala arenan. SCB kommer även fortsätta att bidra i arbetet med att bygga upp Ena, den förvaltningsgemensamma digitala infrastrukturen, inom SCB:s expertområden.

Det finns ännu inte några internationellt vedertagna kvalitets-indikatorer för digitala data, SCB kommer fortsätta följa och bidra till utvecklingen. Indikatorerna ska bl.a. presenteras på International Statistical Institutes stora statistikkonferens sommaren 2023. SCB arbetar även för att med hjälp av EU-finansierade utvecklingsprojekt öka förmågan att göra statistik baserat på digitala data, t.ex. kompetens kring mobilnätsdata.

Resultaten från riskanalysen baserat på Diggs vägledning för att tillgängliggöra information visar på flera områden där myndighets-samverkan behöver stärkas för ökad informationssäkerhet.

Bilaga 1: Kvalitets-indikatorer, utdrag ur vägledning

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 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.

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, pingar 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 före detta kunder finns registrerade, så bidrar det till täckningsproblem.

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 mobilnätsdata 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:

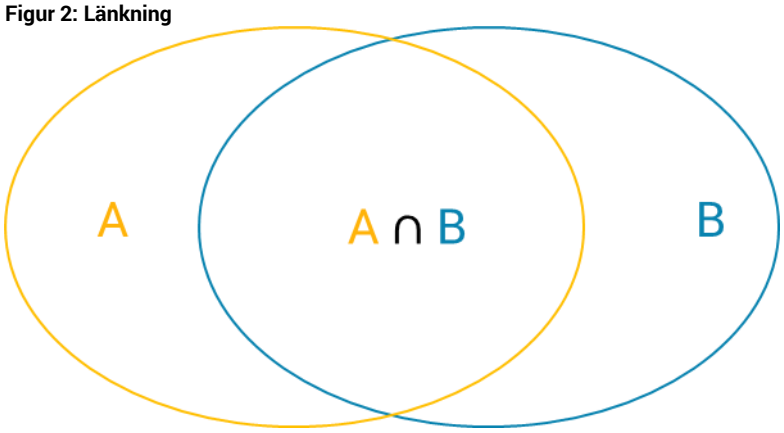
A1. Vilka källor länkas? Finns uppgifter om täckningsfel i de datakällor som den digitala källan länkas till?
A2. Vilka variabler används för länkning?
A3. Finns det identifierande variabler?
A4. Beskriv modellen för länkning
A5. Hur hanteras objekt som inte kan länkas entydigt?

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)

Indikatorer som kan tas fram löpande: B1-B5

Figur 2: 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:

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

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 mobilnätsdata 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

Bilaga 2: Resultat från riskanalysen

Riskbaserat och systematiskt informations-säkerhetsarbete

Enligt princip 2, ”Bedriv ett riskbaserat och systematiskt informationssäkerhetsarbete”, i Diggs vägledning för att tillgängliggöra information ska informationen som tillgängliggörs vara informations-klassad utifrån aspekterna konfidentialitet, riktighet och tillgänglighet.

Syftet med att klassa informationen är exempelvis att utreda: Vad kan konsekvenserna bli om informationen kommer obehöriga till del (*Konfidentialitet*)? Vad kan konsekvenserna bli om informationen är manipulerad eller förstörd (*Riktighet*)? Vad kan konsekvenserna bli om någon (som är behörig) inte får tillgång till informationen (*Tillgänglighet*)?

Informationsklassning används för att värdera information och skapa en organisationsgemensam bedömningsgrund för hur information ska hanteras. Genom att värdera information med hjälp av en klassningsmodell kan man identifiera vilka konsekvenser otillräckliga säkerhetsåtgärder skulle orsaka och utifrån det säkerställa att rätt åtgärder vidtas.

En organisation ska dessutom beakta att den tillgängliggjorda informationen kan ackumuleras och aggregeras med annan information vilket kan ge upphov till högre klassning. Informationen kan i brist på rätt klassning och otillräckliga skyddsåtgärder orsaka skador för exempelvis individers integritet, eller organisationer och samhällets säkerhet.

Inför informationsklassning ska följande aspekter beaktas:

- Innehåller informationen sekretessrelaterade uppgifter[9]?
- Innehåller informationen personuppgifter[10]?
- Innehåller informationen säkerhetsskyddsklassificerade information[11]?

---

[9] Konfidentialitet medför inte automatiskt sekretess, även om det kan finnas en koppling. Sålunda ska de två begreppen (konfidentialitet och sekretess) hållas åtskilda. Sekretess är enbart en benämning på den del av informationen som hamnar under Offentlighets- och sekretesslagen (2009:400) (OSL). Det ska även anmärkas att allmänna handlingar som troligtvis bedöms vara offentliga vid begäran om utlämning bör, trots detta, inte ges den lägsta konsekvensnivån (”ingen eller försumbar”) när det gäller konfidentialitet.

[10] Så hänger lagarna ihop | IMY

[11] Säkerhetsskydd - Säkerhetspolisen (sakerhetspolisen.se)

- Innehåller informationen upphovsrättsligt skyddande verk och andra prestationer som tredje man har rättigheterna till. Finns det stöd i lagen för publiceringen/datadelningen?
- Har informationsägaren gett sitt medgivande till publiceringen/datadelningen?
- Har hänsyn tagits till andra specifika krav i lagar, förordningar, föreskrifter, vägledningar eller annan författning som styr hur informationen får hanteras?

Risker som har identifierats enligt *Princip nr. 2*:

A. **Risk att informationsklassning är en subjektiv bedömning från respektive organisation**.
Risken idag är att det finns många olika typer av modeller för informationsklassning. I ***"Vägledning utforma klassningsmodell från MSB"*** [12] ges exempel på ett antal klassningsmodeller. Vissa modeller har exempelvis tre nivåer andra fem nivåer. Vidare börjar vissa modeller på till exempel K0 och andra på K1. Konsekvensen kan bli att informationen inte får rätt skyddsåtgärder (svagare eller starkare) genom att nivåerna tolkas olika och informationen får då inte rätt skyddsåtgärder.
En ytterligare konsekvens kan vara att lagkravet inte följs då rätt skyddsåtgärd inte appliceras för informationen. Det finns samtidigt en risk för lång startsträcka innan det nås samsyn om informationsklassning, med konsekvens att respektive projekt försenas.
Det ska även anmärkas att det är resurskrävande att genomföra informationsklassning inom en myndighet/organisation. Det är sålunda viktigt att riktlinjerna för informationsklassning är tydliga så att resultatet inte ska behöva omprövning ifall det inte stämmer med andra myndigheters/organisationers informationsklassning.

Sannolikhet: 3
Konsekvens: 2

Förslag på åtgärd: På nationell nivå borde det finnas tydligare rekommendationer om en klassningsmatris.
Det är viktigt att hålla en bra dialog mellan organisationer i samband med informationsklassning. Det behöver finnas en tydlig vägledning om informationsklassning och hur samsyn nås innan samarbetet börjar. Det är bra att arbeta med en

[12] vagledning-utforma-klassningsmodell_kommentarsperiod.pdf (informationssakerhet.se)

checklista med aktuell modell för informationsklassning som används inom varje organisation. Det blir således lättare att identifiera och jämföra informationsklassningsmodellerna som används inom berörda organisationer.

B. **Risk att information hanteras/skyddas på olika sätt beroende på att det finns olika lagkrav att beakta för olika organisationer**.
Vissa organisationer omfattas av **MSBFS 2020:7**[13] eller av den europeiska motsvarigheten **NIS-direktivet (EU) 2016/1148**[14] med mera. Det kan bland annat medföra ökade kostnader för en viss verksamhet ifall en organisation behöver anpassa sig för krav som den inte i nuläget behöver efterleva. Det finns även en risk att otillräckliga säkerhetsåtgärder appliceras.

Sannolikhet: 2
Konsekvens: 3

Förslag på åtgärd: Det är viktigt med utbildning avseende informations- och cybersäkerhet inom varje organisation. Det är betydelsefullt att det planeras tillräckligt med tid avseende samverkan i processen för informationsklassning. Vidare bör man identifiera tillämpliga lagstiftningar för respektive berörd verksamhet. Detta underlättar att beakta konsekvenserna för myndigheten/organisationen om den tidigare inte tillämpat respektive lagkrav, till exempel NIS-direktivet.

C. **Risk att olika aktörer nyttjar olika begrepp inom informationssäkerhet.**
Det kan medföra risk för feltolkning och felaktiga skyddsåtgärder (för starka/för svaga). Exempelvis används ibland svenska och ibland engelska begrep. Denna risk kan till och med finnas inom samma organisation där det används olika begrepp på olika språk. Dessa kan betyda och tolkas olika.

Sannolikhet: 3
Konsekvens: 2

Förslag på åtgärd: MSB har tagit fram en begreppskatalog, som förslagsvis kan utvecklas vidare, men framför allt krävs att kunskap om denna katalog sprids inom offentlig och privat sektor[15]. Det är med andra ord viktigt att det finns en

[13] Författning (msb.se)

[14] NIS-direktivet (msb.se)

[15] Termbanken för informationssäkerhet (informationssakerhet.se)

begreppskatalog som används inom statistikansvariga myndigheter.

Även inom EU kan begrepp och informationsklassningsmodeller behövas harmoniseras. Det finns idag vissa initiativ som genomförts där Eurostat haft initiativ för att jämföra modeller gällande informationsklassning[16].

Det finns också risker med begreppet ”aggregering” som används olika i olika vägledningar och föreskrifter. Inom säkerhetsskyddslagstiftningen betyder aggregerade uppgifter att flera olika typer av uppgifter samlas och tillsammans utgör ett nytt ökat skyddsvärde[17]. I aggregerad statistisk slås däremot exempelvis uppgifter från flera individer, grupper eller tidsperioder samman, ofta i syfte att öka överskådligheten och därigenom orsakar en lägre informationsklassning.

D. **Risk på grund av olika lagkrav**
För statliga myndigheter finns lagkrav på informationssäkerhet via till exempel MSBFS 2020:6 eller offentlighets- och sekretesslagen, OSL. För privata/kommunala aktörer gäller inte alltid samma lagstiftning. Det finns en risk att informationen inte hanteras korrekt eftersom kunskapen och kompetensen gällande informationssäkerhet och dess lagkrav skiljer sig åt.

Utan samma grund och krav för informationsklassning finns en risk att skyddsåtgärder prioriteras olika. Exempelvis kan ett företag mer prioritera risker ur ett affärsperspektiv samtidigt som det finns en risk att det inte tas höjd för ett nationellt perspektiv (Sveriges civila försvar).

Om uppgifter går att köpa från en leverantör kan detta ge en falsk trygghet eftersom det kanske uppfattas att informationen har ett lågt skyddsvärde. Risken är då att informationen hanteras på ett felaktigt sätt.

Sannolikhet: 3
Konsekvens: 2

[16] Home - Eurostat (europa.eu)

[17] https://www.riksdagen.se/sv/dokument-lagar/dokument/utredning-fran-riksdagsforvaltningen/kompletteringar-till-regelverket-om_H8A5URF3/html

Förslag på åtgärd: Det optimala vore att det finns ett gemensamt lagkrav inom informationssäkerhet på nationell nivå, men sådant kan vara svårt att införa.

Här behövs i första hand dialog om vilka perspektiv och krav som har beaktas i samband med informationsklassningen. Rätt skyddsåtgärder ska sedan införas, exempelvis kryptering. Dialogen behöver genomföras på rätt kompetensnivå, det vill säga att exempelvis säkerhetsansvarig, cybersäkerhetsansvarig ska vara inblandad/konsulterad. I annat fall ska informationen inte tas in.

E. **Risk för kvalitetsbrister i data**
Det kan exempelvis finnas risk att företag inte vill dela med sig information ur ett affärsperspektiv (medvetet eller omedvetet). Det kan också finnas risk att informationen är felaktig, genom att data har behandlats/förändrats/manipulerats av någon anledning. Konsekvensen kan bli att informationen inte är helt korrekt eller är missvisande. Detta kan resultera i att informationens riktighet har påverkats även om det inte behöver vara uppsåt att vilseleda, utan det kan ha inträffats med syftet att underlätta eller förenkla. Konsekvensen blir att både statistik och analyser blir felaktiga när användarna av data inte vet att data har uteslutits eller manipulerats av leverantören.

Organisationer behöver på ett tydligt sätt belysa hur materialet har påverkats för att möjliggöra anpassningen av statistiken/analysen. Detta ska förtydliga vilken population statistiken/analysen avser. Men det finns risk att det kan vara svårt att få insyn. Sannolikhet har satts till 1 här, men värdet kan också sättas högre baserat på erfarenheter från användning av mobilnätsdata t.ex.

Sannolikhet: 1
Konsekvens: 3

Förslag på åtgärd: Transparent dialog mellan mottagare och leverantör för att få korrekt metadata och/eller data. Dessutom kan det krävas jämförande studier mot annat liknande data.

F. **Risk för leverans och överlåtelse av onödigt stora datamängder.**
Det finns en risk att det saknas system som på ett enkelt sätt kan överföra data. Det finns en risk att ogallrat data (det vill säga data med onödigt mycket information) levereras och som mottagaren behöver gallra. Konsekvensen blir bland annat att information hanteras i onödan och att det dessutom riskerar datadelningen att bryta mot gällande lagstiftning.

Verksamheten kan ofta behöva experimentera med nya datakällor, då är risken att större mängd data, än det faktiska behovet, kommer in i verksamheten. Det är i så fall viktigt att ha rutiner för att kunna hantera detta på ett säkert och lagligt sätt, dokumentera riskbeslut och sedan, vid produktion, minska mängden information så att den anpassas till det verkliga behovet.

Sannolikhet: 2
Konsekvens: 3

Förslag på åtgärd: Det kräver gemensam dialog och syn på hur och i vilket format och mängd/aggregeringsnivå information ska levereras på. Tekniska åtgärder, exempelvis funktioner som rensar data direkt vid leverans kan implementeras. Datarensningen ska ske in i en skyddad miljö som är isolerad i olika skyddade segment. En ytterligare åtgärd är att gallra filerna direkt vid insamling.

G. **Risker när informationen kombineras eller samlas ihop (adderar ytterligare uppgifter)**.
Skyddsvärdet på information kan i dessa fall öka.
En annan risk som kan uppstå vid uppräkning till total population är att det saknas transparens om hur uppräkningen har gjorts.

Det ska särskilt beaktas att, när informationen kombineras eller samlas ihop, inte skadar individers integritet eller hotar samhällets säkerhet (Sveriges säkerhet/civilt försvar).

Sannolikhet: 2
Konsekvens: 3

Förslag på åtgärd: Inom offentliga myndigheter är det viktigt att ha dialog med beredskapsmyndigheter och sektorsansvariga myndigheter om huruvida informationens klassning, när data samlas ihop från olika källor, ökar i nivå vid informationsklassning. Det är viktigt att beakta om kombinerade data innebär fara för Sveriges säkerhet. Exempelvis insamlade data för energibolag som kompletteras med annan information.

Det är också viktigt med tydlig nationell vägledning och tydligt utpekande om ansvar om vem som kan bistå med att bedöma nivån på materialet som ska informations klassas.

H. **Risk att tekniken hos sändande leverantör har begränsade möjligheter för att leverera information och det saknas möjlighet att påverka leverantören (felaktiga data skickas)**. Exempelvis har en privat aktör endast möjligheter att leverera information genom en utländsk molntjänst utan möjlighet till kryptering. Uppgifterna från leverantören kan lyda under OSL och GDPR. Det blir en risk för otillåten överföring till tredje land samt att det kan medföra en risk att informationen inte får rätt nivå av kryptering.

Sannolikhet: 2
Konsekvens: 3

Förslag på åtgärd: Planera tillräckligt med tid för att kunna påverka leverantörer, ta även höjd för manuellt arbete. I sista hand avstå att inhämta information från leverantören. Internt i organisationen ska en aktiv dialog föras med juridiska instanser, arkitekter, arkivarier, dataskyddsombud och säkerhetsorganisationen. Inrätta interna beredningsgrupper som i samverkan belyser risker och förslag på rekommenderade åtgärder.

I. **Risk att organisationer inte har tillräckliga resurser för att genomföra skyddsåtgärder.**
Begräsningar i ekonomi och kompetens medför att skyddsåtgärder inte kommer på plats i tillräcklig omfattning. Konsekvensen blir att uppgifter med högt skyddsvärde inte skyddas tillräckligt. Inom offentlig verksamhet finns det olika lagar då vissa endast lyder under till exempel MSBFS 2020:7 som driver på införandet av gällande kompetens och teknik för Informationssäkerhet. Privata aktörer har inte alltid samma incitament.

Sannolikhet: 3
Konsekvens: 3

Förslag på åtgärd: Viktigt att den som levererar data har dialog med mottagaren om hur uppgifter kommer skyddas under leverans och när data kommer fram. Exempelvis att man följer rekommendationer och krav från Nationell Cybersäkerhetscenter[18] eller ISO 27000.

[18] Nationellt cybersäkerhetscenter (ncsc.se)

Aktuell och uppdaterad information med användaren i centrum

Enligt princip 3, ”Tillgängliggör aktuell och uppdaterad information med användaren i centrum”, ska tillgängliggörandet ske utan onödig fördröjning och bör inte undanhållas för att informationen inte är fullständig eller brister i kvalitet.

Följande risker har identifierats:

J. **Risk att delning av viktig information är beroende av tillgänglighet hos leverantören av information.**
Risken är om leverantören har driftproblem som medför att det inte är möjligt att leverera i tid eller risk för sabotage genom till exempel överbelastningsattacker (DDOS).

Sannolikhet: 2
Konsekvens: 3

Förslag på åtgärd: Vid avtalsskrivning ska även icke funktionella krav gällande till exempel krav på redundans, säkerhetskopiering med mera beaktas.

K. **Risk att leverantören går i konkurs eller blir uppköpt, eller vill sluta leverera på grund av ny affärsstrategi.**
Risken är att information inte längre tillgängliggörs på grund av att leverantören har ett nytt affärsfokus, eller av någon annan anledning inte längre vill eller kan medverka i samarbetet. Man kan till exempel anse att det blir för dyrt och att det inte ger något mervärde.

Sannolikhet: 1
Konsekvens: 3

Förslag på åtgärd: Risken kan vara svår att hantera. Ställer krav vid avtalsskrivande.

L. **Risk för medveten manipulation av data hos leverantör för att påverka eller sabotera.**
Det kan finnas syfte och uppsåt hos leverantörer att påverka marknaden eller politiska beslut. Detta kan leda till att man medvetet och systematiskt förmedlar en bild som inte överensstämmer med verkligheten.

Sannolikhet: 1
Konsekvens: 3

Förslag på åtgärd: Kräver att mottagaren har förmåga och

genomför kvalitetsanalyser av data, letar efter avvikelser mot till exempel tidigare perioder, jämför med andra interna och/eller externa publicerade data.

M. **Risk: Kostnadsaspekt – att man tar betalt eller ökar kostnaderna för att leverera data enligt säkerhetsperspektiv.**
Nya förändrade säkerhetskrav eller andra behov som påverkar informationen kan medföra att leverantören får ökade kostnader. Detta kan medföra att leverantören vill ha mer betalt, eller att leverantören inte längre är intresserad att bistå med data. Lagkrav och skyddsåtgärder är föränderliga och kräver att det regelbundet görs anpassningar.

Sannolikhet: 2
Konsekvens: 2

Förslag på åtgärd: Risken kan vara svår att hantera. Ställer krav vid avtalsskrivande men även en samsyn och förståelse från båda parter om vikten av att skydda informationen.

Villkor som främjar bred användning

Enligt princip 5, ”Använd villkor som främjar bred användning”, ska informationen publiceras/delas under villkor som inte i onödan begränsar möjligheterna till vidareutnyttjande. Utgångspunkten är att informationen tillgängliggörs utan krav på att användaren ska ansöka om tillstånd eller registrera sig. Information ska dessutom erbjudas avgiftsfritt om inte det framgår annat i lag, förordning eller särskilt beslut av regeringen. Om avgift tas ut ska beräkningsgrunden publiceras öppet och elektroniskt.

Här har nedanstående risker identifierats:

N. **Risk att säkerhetskrav försvårar användning och vidareutnyttjande.**
Risken innebär att säkerhetskraven är eller blir så hårda att det blir svårt att använda och bearbeta informationen. Exempelvis kan klassningen på informationen öka. Risken blir att resultatet inte uppfyller syftet då det inte kan användas som avsett.

Sannolikhet: 1
Konsekvens: 3

Förslag på åtgärd: Risken är svår att hantera. Krav kan förändras under resans gång allt eftersom informationen förändras och förädlas.

Dokumentation och beskrivning av information

Enligt princip 6, ”Dokumentera och beskriv information”, ska information som tillgängliggörs dokumenteras och beskrivas så att den är lätt att upptäcka, förstå och användas både av människor och maskiner.

Här har nedanstående risker identifierats:

O. **Metadata stämmer inte överens med innehållet i dataleveransen ur informationssäkerhetsperspektiv.**
Om vissa data har uteslutits eller manipulerats ur ett material och informationsklassningen görs utifrån metadata där denna behandling inte framgår, påverkas datamaterialets riktighet och det finns en risk att informationsklassningen blir felaktig. Den kan i detta fall bli för högt satt. Likväl kan det förekomma att det inte framgår av metadata att datamaterialet innehåller skyddsvärda data, vilket kan ge en för låg informationsklassning.

Sannolikhet: 1
Konsekvens: 2

Förslag på åtgärd: Inga åtgärder

Länksamling informationssäkerhet

Vägledning för att tillgängliggöra information | DIGG

KLASSA (skr.se)

Metodstöd för LIS (informationssakerhet.se)

Systematiskt informationssäkerhetsarbete (msb.se)

vagledning-utforma-klassningsmodell_kommentarsperiod.pdf (informationssakerhet.se)

Föreskrifter om säkerhetsåtgärder i informationssystem för statliga myndigheter (MSBFS 2020:7)

Nationellt cybersäkerhetscenter (ncsc.se)

Vägledning : säkerhetsåtgärder i informationssystem (msb.se)

Bilaga 3: Konsekvensutredning vid regelgivning

Denna bilaga återger den konsekvensutredning som gjorts i samband med förändringen av SCB:s föreskrifter och allmänna råd om tillgängliggörande av officiell statistik.

1 Inledning

Statistiska centralbyrån (SCB) har för avsikt att utfärda föreskrifter om tillgängliggörande av officiell statistik med stöd av 15 § förordningen (2001:100) om den officiella statistiken.

Rubriksättningen följer förordningen (2007:1244) om konsekvensutredning vid regelgivning.

2 Utredning enligt 6 §

2.1 Problemet och vad man vill uppnå

Tillgängliggörande av officiell statistik har hanterats genom riktlinjer. Det har fungerat, men innehållet i en riktlinje omfattar inte andra myndigheter. SCB vill förtydliga kraven på hur en statistikansvarig myndighet (SAM) tillgängliggör officiell statistik för att stärka hela systemet för den officiella statistiken och därmed göra det enklare för användarna. För att kraven ska slå igenom i hela statistiksystemet är en föreskrift mer effektiv. Dessutom ska SCB i regeringsuppdraget ”Uppdrag att främja delning och nyttiggörande av data för smart statistik” (Uppdrag att främja delning och nyttiggörande av data för smart statistik - Regeringen.se) säkerställa att den officiella statistiken kan användas som öppna data.

2.2 Vilka alternativa lösningar som finns för det man vill uppnå och vilka effekterna blir om någon reglering inte kommer till stånd

Alternativet är att uppdatera de riktlinjer för elektronisk publicering som finns från 2013. Effekten är otydlighet och ett svagare system för den officiella statistiken. Det i sin tur innebär att användarna får svårare att hitta och använda den officiella statistiken.

2.3 Vilka som berörs av regleringen

Alla statistikansvariga myndigheter. Det framgår av bilagan till förordningen (2001:100) om den officiella statistiken vilka myndigheter som är statistikansvariga myndigheter.

2.4 De bemyndiganden som myndighetens beslutanderätt grundar sig på

Enligt 16 § förordningen (2001:100) om den officiella statistiken:

> De statistikansvariga myndigheterna ska dokumentera och kvalitetsdeklarera officiell statistik. Myndigheterna ska också utan avgift offentliggöra och hålla sådan statistik allmänt tillgänglig i elektronisk form genom ett allmänt nätverk.
>
> Statistiska centralbyrån får, utöver av vad som följer av 15 §, meddela föreskrifter om verkställighet av bestämmelserna.
>
> Innan myndigheten meddelar sådana föreskrifter ska den höra övriga statistikansvariga myndigheter.

2.5 Vilka kostnadsmässiga och andra konsekvenser regleringen medför och en jämförelse av konsekvenserna för de övervägda regleringsalternativen

SCB gör bedömningen att flera av de förändrade delarna innebär ökade kostnader för att utveckla de nya delar som regleras i de nya föreskrifterna. I vissa fall innebär det att extra moment ska genomföras, vilket kräver extra resurser. I några fall tas moment bort i statistikproduktionsprocessen och då innebär det att färre resurser behöver nyttjas.

I den mån de statistikansvariga myndigheterna inte redan har en ingång till statistik på sin startsida behöver det läggas in (viss kostnadsökning).

Officiell statistik ska tillgängliggöras som öppna data, vilket kräver utveckling och nya former (kostnadsökning).

SAM ska meddela SCB förändringar för olika statistikprodukter för att hålla produktdatabasen aktuell (kostnadsökning).

Statistikprodukter som inte tillgängliggör nya siffror på scb.se behöver inte skicka in en kvalitetsdeklaration för publicering på produktsidan (kostnadsbesparing).

SAM behöver inte meddela SCB vid varje offentliggörande, i stället ska SCB meddelas om offentliggörandet inte gått enligt plan (kostnadsbesparing).

2.6 Bedömning av om regleringen överensstämmer med eller går utöver de skyldigheter som följer av Sveriges anslutning till Europeiska unionen

En genomgång har gjorts av de europeiska riktlinjerna för europeisk statistik (EU Code of Practice) och den föreslagna föreskriften går i linje med dessa.

2.7 Bedömning av om särskilda hänsyn behöver tas när det gäller tidpunkten för ikraftträdande och om det finns behov av speciella informationsinsatser

Att officiell statistik ska vara öppna data tar tid att genomföra har framkommit i diskussioner med SAM. Tidpunkten för ikraftträdande behöver ligga lite framåt i tiden och dialog med SAM har datum för genomförande planerats till 1 januari 2024. Informationsinsatser görs inom ramen för SCB:s samordningsroll.

Bilaga 4: Resursförbrukning

SCB har använt 3 559 874 kronor för genomförande av uppdraget under 2021 och 2022, som har redovisats mot anslaget 2:4 Informationsteknik och telekommunikation, anslagsposten 4 Informationsteknik under utgiftsområde 22 Kommunikationer. Medlen har använts främst till att finansiera SCB-medarbetares arbetstid, men även till inköp av mobilnätsdata från Telia och Tre (1 000 000 kronor) och för projektinsatser utförda av Visualiseringscenter C (500 000 kronor).

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

Digg (2019). Uppdrag att öka den offentliga förvaltningens förmåga att tillgängliggöra öppna data, bedriva öppen och datadriven innovation samt använda artificiell intelligens. Slutrapport i regeringsuppdraget I2019/0416/DF samt I2019/01020/DF (delvis), Dnr: 2019-139

FN (2021) Approaches to data stewardship, bakgrundsdokument till FN:s statistikkommission UNSC52. Framtaget av FN:s statistikdivision för HLG-PCCB.

FN (2022). Local-Level Statistics as Open Data: A User-centric Approach, bakgrundsdokument till FN:s statistikkommission UNSC53. Framtaget av UN Working group on open data. https://unstats.un.org/unsd/statcom/53rd-session/documents/BG-3w-OpenData-E.pdf

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

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.

Haldorson, M. (2022). Open Data - Increased Use of Official Statistics, Nordic Statistical Meeting in Reykjavik.

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)

LiU (2023). Projektrapport ”Sverige i siffror”, resultat från samarbete mellan Visualiseringscenter C och SCB vad gäller visualisering av öppna data.

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.

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). Data – en underutnyttjad resurs för Sverige: En strategi för ökad tillgång av data för bl.a. artificiell intelligens och digital innovation, Dnr I2021/02739.

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

Resnick, M. et al. (2005). Design Principles for Tools to Support Creative Thinking

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 (2021) SCB i det nationella dataekosystemet, nuläge våren 2021, Dnr A2021/2327.

SCB (2021a) med stöd av beredningsgrupp med representanter för andra statistikansvariga myndigheter: Tillhandahålla officiell statistik som öppna data. En nulägesbild 2021 med utgångspunkt i en utvärdering från Statskontoret 2018, Dnr A2022/0030

SCB (2021b) Standardlicenser för SCB:s upphovsrätt, gd-beslut 2021-05-25, Dnr A2021/1979

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

SCB-FS 2002:16. Statistiska centralbyråns föreskrifter och allmänna råd (SCB-FS 2002:16) för offentliggörande m.m. av officiell statistik.

SCB-FS 2016:27. Föreskrifter (SCB-FS 2016:27) om ändring i Statistiska centralbyråns föreskrifter och allmänna råd (SCB-FS 2002:16) för offentliggörande m.m. av officiell statistik,

SCB-FS 2020:16. Föreskrifter (SCB-FS 2020:16) om ändring i Statistiska centralbyråns föreskrifter och allmänna råd (SCB-FS 2002:16) för offentliggörande m.m. av officiell statistik,

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

Vlag, P. (2021). Quality improvement in mobile phone position data: a step forward towards use for official statistics, SCB-dokument tillgängligt på begäran

Vlag, P., Durnell, U., Malmros, J. (2022a). QUALITY IMPROVEMENT IN MOBILE PHONE POSITION DATA: A collaboration model between SCB and an operator, research paper Q-2022 conference

Vlag, P., Durnell, U., Malmros, J. (2022b). Mobil phone position data and official statistics, research paper Nordic Statistical Meeting

Vlag, P., Durnell, U., Malmros, J. (2022c). Experiences and challenges with access to privately owned data about mobile phone positions in Sweden, conference paper CES-meeting

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