Statistisk sentralbyrå
Statistics Norway

Prinsipper og retningslinjer for dataeditering

Ane Seierstad og Aslaug Hurlen Foss

TALL SOM FORTELLER
NOTATER / DOCUMENTS
2023/49

I serien Notater publiseres dokumentasjon, metodebeskrivelser, modellbeskrivelser og standarder.

Publisert: 8.november 2023

ISBN 978-82-587-1833-5 (elektronisk)
ISSN 2535-7271 (elektronisk)

Standardtegn i tabeller	Symbol
Ikke mulig å oppgi tall Tall finnes ikke på dette tidspunktet fordi kategorien ikke var i bruk da tallene ble samlet inn.	.
Tallgrunnlag mangler Tall er ikke kommet inn i våre databaser eller er for usikre til å publiseres.	..
Vises ikke av konfidensialitetshensyn Tall publiseres ikke for å unngå å identifisere personer eller virksomheter.	:
Desimaltegn	,

Forord

SSB fikk som en del av Eurostats «Peer review» som ble gjennomført høsten 2021, en anbefaling om å jobbe for mer standardisering når det gjelder bruk av metoder i statistikkproduksjonen (anbefaling 4, improvement action 4.1). For å følge opp denne anbefalingen har Seksjon for metoder satt i gang et arbeid med å finne metodeområder og metoder som er egnet for standardisering.

Dataeditering er et metodeområde som angår svært mange statistikker, og hvor mange seksjoner bruker betydelig ressurser. Derfor er dette et område hvor en standardisering vil kunne gi en gevinst med hensyn på kvalitet og effektiv ressursbruk. Prinsippene og retningslinjene vil være et grunnlag for utvikling av metoder, kode og tekniske løsninger for dataeditering framover.

Prinsippene og retningslinjene for dataeditering som beskrives i dette notatet, er et utgangspunkt for det videre arbeidet med å standardisere dataeditering i SSB. Forslaget er sendt på høring i organisasjonen, og Standardutvalget har gitt sin tilslutning til det. Prinsippene og retningslinjene ble vedtatt i DM 17. oktober 2023.

Statistisk sentralbyrå, 24. oktober 2023

Arvid Olav Lysø

Sammendrag

Dette notatet beskriver prinsipper og retningslinjer for dataeditering i SSB. Dataeditering er kontroll, gransking og retting av data. Det blir også kalt datavasking.

Prinsippene for dataeditering er generelle og skal være gyldige for all statistikkproduksjon.

Det foreslås ni prinsipper for editering i SSB. Disse gir overordnede føringer for når og hvordan editering skal gjøres. Editering i SSB skal automatiseres så mye som mulig, men det må være mulig med menneskelige kontroller og intervensjon i prosessene.

Retningslinjene er en konkretisering av prinsippene. Det er tenkt at retningslinjene skal passe for flest mulig statistikker, og vi har derfor valgt å skille på person- og næringsstatistikk i enkelte prosesser. Retningslinjene er utformet slik at de kan brukes som en håndbok for god praksis for editeringsarbeid. Det angis hva som skal kontrolleres og hvordan. Retningslinjene inneholder også de anbefalte kvalitetsindikatorene for de ulike delene av editeringsprosessen.

Innhold

Forord **3**

Sammendrag **4**

1. Innledning **6**

1.1. Om prinsippene og retningslinjene 6

1.2. Dataeditering - definisjon og formål 6

2. Prinsipper **7**

3. Retningslinjer **9**

3.1. Kontroll av kildedata 10

3.2. Populasjonsediting og editering av systematiske feil 11

3.3. Selektiv editering 13

3.4. Makroeditering 15

3.5. Siste kontroller før publisering 15

3.6. Dokumentasjon av editering 16

3.7. Evaluering av editering 16

Referanser **17**

Vedlegg A: Feilkoder for fødselsnummer **18**

Vedlegg B: DUF-nummer **19**

1. Innledning

1.1. Om prinsippene og retningslinjene

SSB har hatt prinsipper for dataeditering i forskjellige versjoner tidligere. De første prinsippene for SSB ble publisert som «de ti bud» i Håndbok for datarevisjon (SSB, 2005). Deretter har det vært laget prinsipper i forbindelse med arbeidet med metodikk for modernisering av statistikkproduksjonen (Bråthen, Seierstad og Foss, 2020), som så har blitt endret etter tilbakemeldinger på kurs i dataeditering. De reviderte prinsippene som presenteres i dette notatet, bygger på tidligere versjoner og Eurostats prinsipper for validering (Eurostat, 2020). Prinsippene er generelle og skal være gyldige for all statistikkproduksjon.

I tillegg til prinsipper, er det nå laget retningslinjer for dataeditering. Retningslinjene er en konkretisering av prinsippene. Det er tenkt at de generelle retningslinjene skal passe for flest mulig statistikker, men vi har valgt å skille mellom person- og næringsstatistikk i enkelte prosesser. Statistikkproduksjon kan variere mye og være kompleks, og de samme retningslinjene vil derfor ikke alltid passe til alle statistikker, med ulike typer kildedata. I utarbeidelsen av retningslinjer for dataeditering har vi tatt utgangspunkt i den internasjonale prosessmodellen for dataeditering (UNECE, 2019). Denne modellen er generell og beskriver de forskjellige prosesstegene i data-editering. Vi anbefaler å først kjøre hele produksjonsprosessen automatisk, og deretter gå tilbake og sjekke alle prosesstegene, se prosessmodell for modernisering av statistikkproduksjon (Bråthen, Seierstad og Foss, 2020)

I tillegg til retningslinjer, angis forslag til kvalitetsindikatorer. Det er da tatt utgangspunkt i anbefalte kvalitetsindikatorer for offisiell statistikk (Foss og Haugen, 2023). Dette er ment som en hjelp til å velge relevante kvalitetsindikatorer. Indikatorene som er foreslått skal være praktiske og effektive, og gi god informasjon om produksjonsprosessen.

Prinsippene og retningslinjene som beskrives i dette notatet skal gi et grunnlag for utvikling av kode i SSB framover, både for fellesfunksjoner og for kode som utvikles særskilt for en enkelt statistikk. Parallelt med kodeutvikling, må rutiner for editering jevnlig oppdateres for å ivareta ny kunnskap og nye krav til statistikken. At data fra etablerte kilder endres, og at datavolum og tilfanget av nye kilder endres over tid, forsterker behovet for å kontinuerlig utvikle og forbedre editering.

1.2. Dataeditering - definisjon og formål

Dataeditering er kontroll, gransking og retting av data. Prosessen består av en rekke aktiviteter som tar sikte på å vurdere rimeligheten til dataene, identifisere potensielle problemer og deretter utføre handlinger for å avhjelpe de identifiserte problemene.

Dataeditering omfatter hele produksjonsprosessen for statistikk, slik den er beskrevet i prosess-modellen GSBPM (SSB, 2019). Modellen beskriver alle stegene i produksjonen for å transformere data fra inndata til statistikk. Dataediteringen er en del av denne produksjonsprosessen. Noen ganger kan stegene som innebærer dataeditering grupperes for å danne et fast ledd i kjeden, slik som for delprosessene 5.3 "kontrollere og validere" og 5.4 "editere og imputere". Andre ganger kan editering gjøres som en del av et annet steg eller delprosess.

Formålet med dataeditering er å sikre tilstrekkelig god kvalitet på publisert statistikk. Det er umulig å sjekke at alle data er korrekte; selv ikke en nøyaktig gjennomgang av alle data fra skjema eller

registre, kan garantere gode data. Kvaliteten måles på sluttproduktene, dvs. på publiseringsnivå for statistikkene eller eventuelt gjennom bruk av mikrodata. Leveranser av mikrodata til brukere som for eksempel forskningsmiljøer, seksjon for nasjonalregnskap og Eurostat, stiller strenge krav til korrekte og konsistente data på enhetsnivå, noe som igjen øker behovet for en mer omfattende editeringsprosess. Effekten av editering på statistikkens nøyaktighet må vurderes opp mot andre kvalitetskriterier som aktualitet og relevans. I tillegg må man alltid vurdere effekten opp mot kostnader og ressursbruk.

Vi har forsøkt å bruke terminologien slik som den er brukt i prosessmodellen for statistikk-produksjon (GSBPM) og prosessmodell for dataeditering (GSDEM). Begrepet editering vil i dette dokumentet bli brukt for hele prosessen med kontroll, gransking og retting av data. For prosessen med å korrigere data slik som beskrevet i prosess 5.4 "editere og imputere" i prosessmodellen vil vi bruke terminologien korrigere og imputere.

2. Prinsipper

1. God kunnskap om fagfeltet for statistikken og bakgrunnen til kildedata, er grunnlaget for en god editeringsprosess

Med god kunnskap om fagfeltet er det mulig å vurdere om statistikken samsvarer med forvent-ninger til utvikling og kunne forklare eventuelle avvik fra denne. God kunnskap om kildedata gjør at en vet hvilke feil som kan forekomme og kan sette opp kontroller som fanger opp dette.

2. Formålet med dataediteringen skal være klart formulert

En klar målsetning med dataediteringen er viktig for å prioritere riktig. Dette kan gjøres ved å definere hvilke tabeller som skal publiseres og hvilken usikkerhetsmargin disse kan ha, hvis det er mulig å beregne. Det er viktig å ha et makroperspektiv i statistikkproduksjonen og fokusere på det som påvirker sluttresultatet. For å sikre personuavhengig behandling av data, bør det finnes en instruks som beskriver hva statistikkprodusenten skal gjøre i prosessen dataeditering.

3. Gode data inn er best

Gode data inn til SSB sikrer best kvalitet i statistikken. Det er den som er nærmest kilden som kan rapportere data best. For administrative data går dataene gjennom flere ledd før de kommer til SSB. Det er mest effektivt å få gode data inn, fordi det sparer tid på korrigering av data. Det bør derfor prioriteres å sette inn tiltak for å få inn gode data fremfor å korrigere data i etterkant. God dialog med oppgavegivere og dataeiere er et viktig tiltak for å få inn gode data. Dersom det oppdages feil som påvirker statistikken mye, bør vi spørre om å få levert inn nye tall fra oppgavegiver. For administrative data er tilbakemelding på kvalitetsindikatorer et viktig tiltak for langsiktig forbedring av data.

4. Kontroller alltid data

Selv om vi har tillit til at data er sjekket før de mottas, skal dataene likevel alltid kontrolleres. Vellykket datautveksling er et felles ansvar. Dette kan ikke gjøres uten en rimelig grad av tillit og forståelse for hverandres utfordringer. Det er den som leverer data som er ansvarlig for at nøyaktigheten er tilfredsstillende ut ifra deres behov. Den som mottar data, er ansvarlig for å

kontrollere data ut ifra behovene til statistikken som skal produseres og gi dataeier nyttige tilbakemeldinger som kan bidra til å heve datakvaliteten.

5. Jo tidligere, jo bedre

Prosessen for dataediteringen må utformes slik at feil kan bli oppdaget så raskt som mulig. Da kan korrigering utføres på det stadiet hvor kunnskapen er tilgjengelig. Jo raskere feil oppdages i en produksjonskjede, jo enklere og mer effektivt er det å rette dem. Når feil blir oppdaget og rettet tidlig, vil de resterende prosessene blir mindre påvirket av feil.

6. Kontrollene, kontrollutslagene og endringene skal være veldokumenterte

Kontrollene for en statistikk må være klart og entydig definert. De må være godt dokumentert slik at de kan kommuniseres til dataeier og til andre som bruker datasettet eller statistikken. Dette gjør at det oppnås en felles forståelse mellom de ulike involverte aktørene om hva som er implementert i produksjonsprosessen og hvordan statistikken er laget.

Resultatet av kontrollene, kontrollutslagene, må være klart og entydig definert og dokumentert for det datasettet som er under granskning.

Endringene som blir gjort på datasettet, både manuelle og maskinelle, skal dokumenteres. Det som er gjort må beskrives for å sikre en felles forståelse av resultatet.

Veldokumenterte kontroller, kontrollutslag og endringer er basisen for å lage kvalitetsindikatorer for dataeditering.

7. Automatiser editeringsprosessen så mye som mulig

Automatisering av editeringsprosessen innebærer å sette opp kontroller som kjøres automatisk, samt gjøre automatiske korrigeringer av data ved hjelp av logiske regler eller ved statistiske imputeringsmetoder. I tillegg bør det settes opp automatiske makrokontroller, det vil si kontroller av tabellene som publiseres. Automatiserte prosesser gjør produksjonen mer effektiv, men slike prosesser må overvåkes.

8. Effektiviser editeringsarbeidet

God bruk av selektiv editering og å jobbe ut fra et makroperspektiv, hjelper til med å effektivisere statistikkprodusentenes editeringsarbeid. I tillegg effektiviseres editeringsarbeidet ved å legge godt til rette for oppgaver statistikkprodusenten har. Dette er for eksempel administrasjon av kontroller og automatiske opprettinger, vurdering av kvaliteten i data og sluttprodukt og å finne og eventuelt korrigere innflytelsesrike feil. Grafikk kan gi en rask oversikt over resultater, trender og strukturer i data. Drilling, det vil si tilgang til dypere nivåer suksessivt i hierarkiske data og figurer, hjelper til med å forklare data og avgjøre om det er behov for å korrigere.

9. Dataediteringen bør evalueres

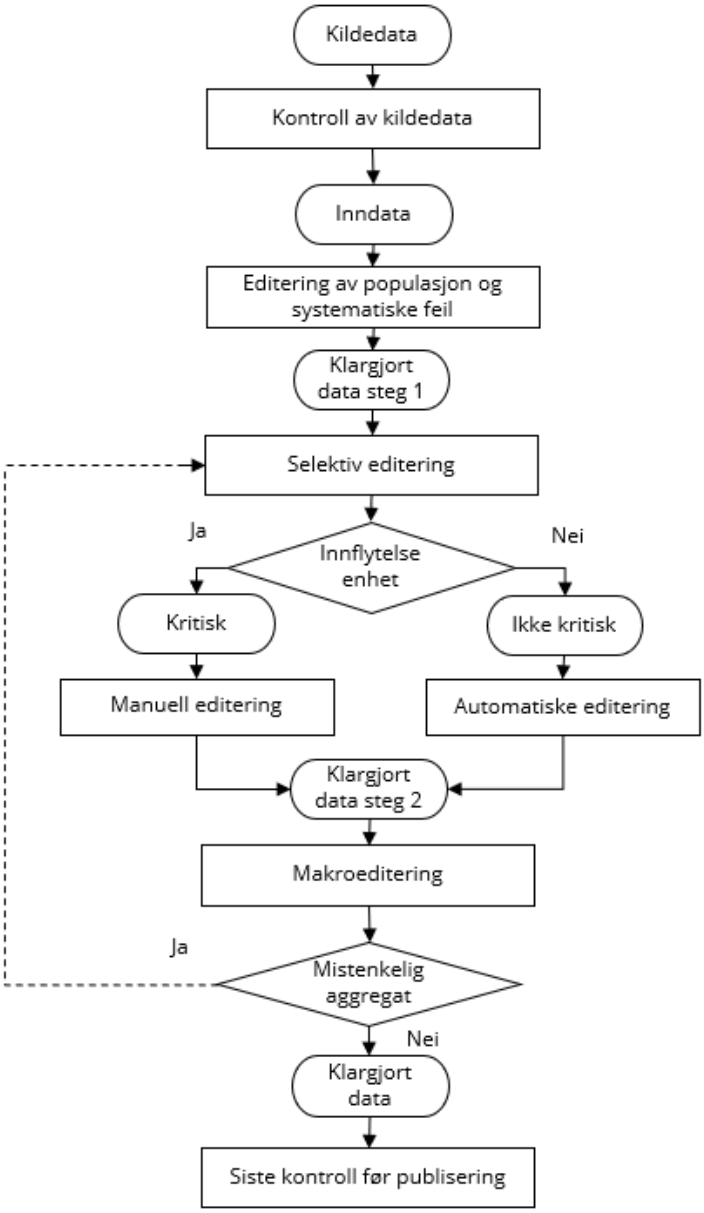
Evaluering av dataediteringen ved hjelp av kvalitetsindikatorer er viktig for å samle kunnskap slik at forbedringstiltak kan settes inn. Evaluering bidrar til kontinuerlig forbedring av produksjons-prosessen og av data. Kvalitetsindikatorer, slik som effekten av editering, kontrollutslagsrate og editeringsrate, kan være input til en slik evaluering. En vurdering av kostnader og ressursbruk bør

også inngå i en evaluering. Hvis tiltak blir satt inn, slik at feil ikke oppstår igjen eller at kontrollene blir mer treffsikre, vil prosessen bli mer effektiv og kvaliteten på statistikken bli høyere.

3. Retningslinjer

Retningslinjene tar utgangspunkt i prosessmodellen for dataeditering (UNECE, 2019). Det er laget forskjellige slike modeller for ulike typer statistikk. Modellen som vises i figur 3.1 er basert på modellen for næringsstatistikk. I prosessmodellen fra UNECE er det også laget en modell for husholdningsstatistikk, og persondelen av denne modellen er nesten helt lik den for nærings-statistikk. Modellen er en forenkling av virkeligheten, men viser likevel godt hovedtrekkene i hvilke prosesser som er inkludert i dataeditering. I retningslinjene er det i tillegg tatt med kontroll av kildedata og statistikkbanktabeller, som henholdsvis kommer før og etter prosessene i UNECE-modellen, noe som gjenspeiles i figur 3.1.

Figur 3.1 Prosessmodell for dataeditering basert på modell for næringsstatistikk



3.1. Kontroll av kildedata

Kontroll av kildedata vil avhenge av om SSB foretar datainnsamling selv direkte fra oppgavegiver eller om SSB mottar data fra noen andre som har samlet dem inn. I dette notatet er det kalt sekundære kildedata når kildedata er samlet inn av andre, dette vil for eksempel være data fra Skatteetaten eller Tolletaten. Når SSB har samlet inn data selv er det her kalt primære kildedata, dette vil for eksempel være data som er samlet inn gjennom systemene CATI, Altinn, Designer og lignende.

Kontroll av primære kildedata

I primære kildedata er det SSB som bestemmer, det vil si hvem som skal rapportere, hva som skal samles inn av informasjon og hvilke kontroller disse dataene skal tilfredsstille for at de skal bli sendt inn.

Kontroller bør legges inn i skjemaer eller ved innsending av data. Kontrollene bør testes slik at en vet om de fungerer og ikke hindrer innsendingen.

Retningslinjer:

- Kontroller at variabler har riktig format.
- Kontroller at datacellene har verdier.
- Kontroller at verdiene er i gyldig verdiområde.
- Hvis det kontrolleres for mulige feil, bør treffsikkerheten være høy og det åpnes for dialog med oppgavegiver om riktig verdi.

Anbefalte kvalitetsindikatorer:

- Tidsbruk skjemautfylling

Kontroll av sekundære kildedata

Data som SSB mottar fra dataeiere eller henter inn fra sekundærkilder på annet vis (for eksempel ved webscraping), skal kontrolleres raskt for å oppdage feil tidlig. Dersom vi har et samarbeid med dataeier, må feilene meldes tilbake til dataeier.

Retningslinjer:

- Ha et godt samarbeid med dataeier som dekker temaer som:
  - Dataleveranse
  - Metadata
  - Datakvalitet
  - Informasjon om statistikkproduktet som skal bli laget av disse dataene
- For administrative data skal det være en samarbeidsavtale med eieren av data. Denne avtalen skal regulere dataleveranse og kvalitet.
- For privateide sekundærkilder som SSB henter inn skal det finnes en samarbeidsavtale med eieren av data. Denne avtalen skal regulere dataleveranse og kvalitet.
- Kontroller at datasettet har riktig struktur og at variabler har riktig format.
- Kontroller at datasettet inneholder alle enheter og observasjoner som skal være med.
- Kontroller at datacellene inneholder verdier.

Anbefalte kvalitetsindikatorer:

- Avvik data mottatt: Avvik i antall mottatte enheter sammenlignet med tidligere perioder. For korttidsstatistikker er det forrige periode eller tilsvarende periode forrige år.
- Andel ugyldige data: Andel data pr variabel som ikke samsvarer med databeskrivelsen.

ID-editering
Identifiseringsnummer er svært viktig i statistikkproduksjon. Det blir brukt til å administrere populasjoner og sette sammen data fra forskjellige kilder. Målet for denne prosessen er å sjekke og dokumentere kvaliteten av identifiseringsnummer på mottatte data, samt identifisere endringer i identifikatorer over tid.

Generelle retningslinjer:

- ID-nummer skal kontrolleres hvis det kan forekomme feil.
- Manglende og feil ID-nummer skal erstattes med riktig ID-nummer hvis det finnes.
- Manglende og feil ID-nummer skal dokumenteres.

Retningslinjer for kontroll av personidentifikator:

- Kontroll av gyldig personidentifikator gjøres ved kobling mot nyeste SNR-katalog. SNR-katalogen inneholder alle fødselsnummer (FNR) / D-nummer (DNR) som noen gang har vært gyldige, slik at historikken tas vare på.
- Utgåtte identifikatorer erstattes med nyeste FNR/DNR for å bevare stabil ID.
- Ugyldige identifikatorer skal dokumenteres med standardiserte feilkoder, se vedlegg 1.

Det finnes også DUF-nummer som er for personer som har søkt om asyl eller oppholdstillatelse i Norge, se vedlegg 2 for nærmere beskrivelse.

Retningslinjer for kontroll av organisasjonsnummer:
10.
- Kontroll av gyldig organisasjonsnummer gjøres ved kontroll mot Virksomhets- og foretaksregisteret (VoF).

Anbefalte kvalitetsindikatorer:

- Andel ugyldige identifikasjonsnummer, en variant av kvalitetsindikatoren: andel ugyldige data.
- Andel korrigerte identifikasjonsnummer, en variant av editeringsrate.

3.2. Populasjonsediting og editering av systematiske feil

Populasjonsediting
En populasjon er samlingen av de enheter statistikken skal gi informasjon om. Det kan skilles mellom to typer populasjoner: målpopulasjon og undersøkelsespopulasjon. Målpopulasjonen er den populasjonen man ønsker å lage statistikk for, dette er en teoretisk størrelse. Undersøkelsespopulasjonen er den populasjonen informasjonen kan skaffes for i undersøkelsen, også kalt totalpopulasjon. Populasjonen er også tid- og stedfestet. I prosessen med å lage undersøkelsespopulasjonen inngår det også editering av den (Hustoft og Sæbø, 2006). Heretter kaller vi undersøkelsespopulasjonen for populasjonen.

Næringsstatistikk

I næringsstatistikk oppdateres populasjonen på grunnlag av endring i klassifikasjonsvariablene som definerer populasjonen eller ved oppdatering av gyldighetstidspunktet for disse. I noen tilfeller defineres også en populasjon ut ifra størrelse ved antall ansatte, omsetning eller lignende. En oppdatering av disse kan således også gi en endring i populasjonen. Det er viktig med fokus på referansetidspunktet/-perioden som en statistikk skal omfatte og tilstrebe best mulig opplysning om enhetene slik de var på dette tidspunktet. Det er viktig at sammenhenger mellom enheter slik som virksomhet, foretak og konsern, er riktig. I tillegg skal det nå lages statistikk basert på statistisk enhet/foretak. Statistisk enhet må i noen tilfeller lages ut fra flere juridiske enheter eller ved å splitte opp juridiske enheter. I noen tilfeller er populasjonen ukjent, det vil si at den ikke kan dannes direkte ut fra VoF. I slike tilfeller må populasjonen dannes ut fra annen tilgjengelig informasjon og fagkunnskap.

Retningslinjer næringsstatistikk:

- Det bør kontrolleres hva som er hovednæring dersom enheten også har en binæring. En rekke kjennetegn kan avgjøre hva som er hovednæring, for eksempel antall ansatte, omsetning eller konsesjonstype.
- For viktige enheter bør det følges med på nyheter for å ha god oversikt over fusjoner og fisjoner.

Personstatistikk

SSB gjør ingen endringer i Folkeregisteret, men det blir laget kvalitetsindikatorer som blir oversendt Skatteetaten og disse kan føre til endringer. Husholdninger blir laget i SSB ut ifra boligopplysninger om hver person. Denne dannelsen av husholdninger blir editert, og det er laget en internasjonal modell for editering av husholdningsstatistikk (UNECE, 2019). I tillegg lager SSB statistikk tilknyttet aktiviteter som blir registrert på personer, slik som opplysninger om utdanning og sosialhjelp. Populasjonen er ofte ukjent og må derfor gå gjennom en dataediteringsprosess. Det kan være manglende innrapportering, og det kan forekomme dobbelrapportering om personer. I tillegg kan det være ugyldige fødselsnummer, se avsnittet om editering av ID-nummer.

Retningslinjer personstatistikk:

- Størrelsen på populasjonen må kontrolleres for å se om den er rimelig i forhold til tidligere perioder eller andre datakilder.
- Det skal sjekkes for dubletter, og hvis de finnes så skal de ryddes opp i.
    - Identisk: Alle verdiene som er rapportert inn er identiske.
    - Revidert: Noen av verdiene som er rapportert inn er endret.
    - Partiell: Verdiene blir delt opp og rapportert i forskjellige omganger.

Anbefalte kvalitetsindikatorer:

- Overdekningsrate: Andel enheter som ikke tilhører målpopulasjonen.
- Andelen dubletter. En dublett er når enheten i undersøkelsen har flere forekomster i innrapporteringen.

Editering av systematiske feil

Åpenbare feil

Åpenbare feil er observasjoner som har klart uriktige verdier. Det kan være at verdien er utenfor gyldighetsområdet, slik som at antall timer arbeidet overstiger antall timer i døgnet. Det kan også være ugyldige kombinasjoner av variabler, slik som at en femåring er registrert som student.

Retningslinjer:

- Kontroller at alle numeriske variabler er innen gyldig verdiområde.
- Kontroller at alle kategoriske variabler har gyldige kategorier.
- Kontroller at variablene er komplette.
- Kontroller logiske sammenhenger som skal være mellom variabler.
- Reglene for kontroller skal dokumenteres og kommuniseres til oppgavegiver og brukere av datasettet.
- Hvis det er sikkert at missing betyr 0 (null), så korriger til 0.
- Hvis de logiske sammenhengene mellom variabler er feil og det er kjent hva som er korrekt, skal disse reglene brukes til å korrigere verdier automatisk.
- Reglene for korrigering skal dokumenteres og kommuniseres til oppgavegiver og brukere av datasettet.
- Endringene som blir gjort i data skal logges.

Anbefalte kvalitetsindikatorer:

- Kontrollutslagsandel: Andel utslag av kontroller i forhold til antall enheter.
- Editeringsandel: Andel endringer av variabler i forhold til antall enheter.
- Treffsikkerhet: Andel endringer av variabelen i forhold til antall kontrollutslag.

Systematiske feil

Systematiske feil er gjennomgående feil som trekker i én retning og som forekommer for mange enheter i datagrunnlaget. Det kan være flere systematiske feil i en og samme undersøkelse. Typiske systematiske feil er for eksempel at en del av populasjonen oppgir beløp med moms og resten av populasjonen oppgir beløp uten moms, eller at noen beløp blir oppgitt i euro i stedet for kroner.

Retningslinjer:

- Viktige variabler bør kontrolleres for systematiske feil, enten ved hjelp av sammenligning med en annen kilde eller periode, eller ved bruk av fordeling.
- Bruk grafikk, slik som punktdiagram; det kan vise trender i data som tyder på systematiske feil.
- Klare systematiske feil som er lette å oppdage bør rettes automatisk.

3.3. Selektiv editering

Selektiv editering handler om å identifisere de feil som sterkest påvirker hovedresultatene. Selv om en verdi er innenfor det aksepterte området, kan den fortsatt være unøyaktig. Metoden prioriterer verdier som ser mest mistenkelige ut og som har høy innflytelse på resultatet. Ved å fokusere på disse potensielle feilene, kan man begrense kostnadene ved manuell retting, samtidig som nøyaktigheten i statistikken opprettholdes.

Retningslinjer:

- De enhetene og verdiene som har høy innflytelse og er mistenkelige (for eksempel ved stor endring fra fjoråret) bør plukkes ut til manuell inspeksjon. Dette er mest relevant for næringsstatistikk.
- De enhetene og verdiene som er mistenkelige og har lav innflytelse bør korrigeres automatisk.
- For kontroll mot forrige periode bør det brukes metoder som tar hensyn til at det ofte er trend i data og at små verdier kan ha stor prosentvis endring. Derfor anbefaler vi metoder som robust regresjon og Hidiroglou-Berthelot-metoden for å kontrollere mot forrige periode.

Anbefalte kvalitetsindikatorer:

- Kontrollutslagsrate: Andel utslag av kontroller i forhold til antall enheter.

Manuell editering

Manuell editering, internasjonalt også kalt interaktiv editering, innebærer at mikrodata kontrolleres, og om nødvendig korrigeres manuelt, ved hjelp av ekspertkunnskap eller andre kilder. Det er de observasjonene som er plukket ut i selektiv editering som bør bli manuelt kontrollert og korrigert. Det kan også være nødvendig med manuell korrigering for å trene modeller for automatisk korrigering.

Retningslinjer:

- Bare enheter med høy innflytelse bør bli manuelt kontrollert og korrigert.
- Innflytelsesrike observasjoner som man ikke finner en god automatisk metode for, bør bli manuelt editert.
- Ressursene som kreves for å bygge automatisk korrigering må vurderes opp mot kostnadene ved manuell korrigering.
- Det bør foreligge en instruks for manuell editering, slik at det sikrer en personuavhengig behandling av data.
- Korrigeringene som blir gjort skal bli logget.

Anbefalte kvalitetsindikatorer:

- Editeringsandel: Andel endringer av variabler i forhold til antall enheter.

Automatisk editering

Automatisk editering er prosessen med å oppdage og behandle feil og manglende verdier helt automatisk, uten menneskelig innblanding. Ved automatisk korrigering blir det satt inn verdier for manglende opplysninger eller man erstatter verdier man tror er feil. Det blir også kalt imputering.

Retningslinjer:

- Bruk fagkunnskap og vurder om det kan settes opp regler for automatisk kontroll og korrigering av verdier.
- Det anbefales å bruke regelbasert imputering for systematiske feil og statistiske modeller for tilfeldige feil.
- Imputerte verdier skal ha samme format som variabelen.
- Imputeringen skal kontrolleres.
- Imputeringen skal flagges og endringen skal dokumenteres.

Anbefalte kvalitetsindikatorer:

- Editeringsandel: Andel endringer av variabler i forhold til antall enheter.
- Usikkerhet: Variasjonskoeffisient for modellbasert imputering.

3.4. Makroeditering

Makroeditering er å analysere aggregater eller beregninger på data for hele populasjonen med det formål å identifisere deler av datasett som kan inneholde potensielt innflytelsesrike feil.

Retningslinjer:

- Det skal vurderes om aggregatene er plausible gitt historisk forløp, og eventuelt ut ifra historiske forløp i annen statistikk, eller annen informasjon om samme eller tilgrensende emneområder.
- Plausibiliteten av aggregatet skal vurderes i lys av utviklingen i avledede størrelser, for eksempel forholdstall.
- Ved endringer utenom det som anses som normalt skal datagrunnlaget bli undersøkt for å enten finne den potensielle feilen eller kunne forklare endringen.
- Aggregatene bør vurderes opp mot tilsvarende aggregater for sammenlignbare land.
- Aggregater kan være gjenstand for editering når de inngår i et system der visse relasjoner må holde.

3.5. Siste kontroller før publisering

Kontroll av konfidensialitet i tabeller

Tabellene som skal bli publisert må kontrolleres for å se om de tilfredsstiller kravene for konfidensialitet. Hvis tabellene ikke tilfredsstiller kravene, skal celler undertrykkes med godkjente funksjoner angitt av metodeseksjonen. Disse funksjonene primær- og sekundærundertrykker celler i tabellene.

Kontroll av statistikkbanktabeller

Når dataene er ferdig kontrollert og godkjent, lages tabeller som skal til statistikkbanken. For disse tabellene er det regler som skal følges som de i statistikkbanken kontrollerer. Det mest effektive er at statistikkprodusenten kjører tilsvarende kontroller før tabellene blir sendt til banken, da blir det raskere å gjøre endringer hvis noe er feil.

Retningslinjer for kontroller av statistikkbanktabeller:

- Statistikkprodusenten skal kjøre kontroller på tabellene før de sendes til statistikkbanken.
- De som mottar tabellen, skal kontrollere at tabellene er i henhold til reglene for statistikkbanken.
- Det som skal kontrolleres er:
    - Antall deltabeller og antall kolonner i hver deltabell.
    - Korrekt formatering på tidskolonner.
    - Korrekte koder i prikkekolonner.
    - Kun bruke kategoriske koder som er registrerte i statistikkbanken fra før.
    - Kun unike kombinasjoner av tid og kategoriske koder, ikke duplikater over disse.
    - Kun tall (eller ingenting) for statistikkvariabler.
    - Alle deltabeller skal inneholde de samme periodene.

Kontroll av tabeller til Eurostat

De fleste statistikker rapporterer tabeller til Eurostat eller andre internasjonale organisasjoner. Disse tabellene gjennomgår en validering før de blir godkjent. Reglene for validering er ofte utarbeidet i fellesskap for hvert fagområde. Utveksling av data og validering blir ofte gjort med en felles standard kalt SDMX (Statistical Data and Metadata eXchange) som er en standard for lagring av data og beskrivelse av dens struktur, betydning og innhold. Reglene i dette systemet er maskinlesbare, og de kan da lastes inn i en editeringsløsning. I R-pakken Validate er det utarbeidet funksjoner som laster disse reglene rett inn i editeringsløsningen (van der Loo, 2023).

Retningslinjer for kontroller av tabeller til Eurostat:

- Statistikkprodusenten bør kjøre disse kontrollene før tabellene sendes til Eurostat.

3.6. Dokumentasjon av editering

Editeringsprosessen skal dokumenteres, slik at vurderinger, korrigeringer og forklaringer er tilgjengelige også i ettertid. Dette sikrer en felles forståelse for statistikken, både for de som leverer data til statistikkproduksjonen, de som produserer statistikken og de som bruker statistikken.

Dokumentasjonen skal gjøres i tråd med gjeldende generelle prinsipper for statistikkproduksjon, kode og behandling av data i SSB. Særlig relevant for editering er lagring av obligatoriske datatilstander (SSB, 2021).

Retningslinjer:

- Manuell og automatisk editering skal dokumenteres, slik at effekten av all korreksjon kan analyseres. Det skal logges om endringen er gjort manuelt eller med hvilken funksjon, tidspunkt for endringen og verdi før og etter. Ved manuell editering skal det i tillegg som et minimum logges hvem som editerte og eventuelt årsak.
- Produksjonskode skal lagres og versjoneres, slik at all automatisk kontroll og korreksjon kan kjøres på nytt.
- Relevante datasett skal lagres, slik at hele editeringsprosessen kan gjenskapes.

3.7. Evaluering av editering

Dataediteringen skal evalueres jevnlig for kontinuerlig forbedring av prosess og data. Det er særlig viktig å kvalitetssikre automatisk korrigering og imputering. Evaluering av produksjonsprosessen og datakvaliteten er viktig for å kunne samle kunnskap slik at forbedringstiltak kan settes inn. Hvis tiltak blir satt inn slik at feil ikke oppstår igjen, vil prosessen bli mer effektiv og kvaliteten på statistikken høyere.

Retningslinjer:

- Dataediteringen skal evalueres jevnlig.
- Kvalitetsindikatorene bør automatiseres og samles i en rapport.
- Kvalitetsindikatorene bør vurderes og analyseres.
- Det bør utformes en tiltaksplan for prosesser som kan forbedres basert på analysen av kvalitetsindikatorer. Det kan for eksempel være tettere samarbeid om kildedata, justering av kontroller eller implementering av nye metoder for kontroll og imputering.

Referanser

Bråthen, Seierstad og Foss (2020). Metodikk for modernisering av statistikkproduksjonen. Notater 2020/21, https://www.ssb.no/teknologi-og-innovasjon/artikler-og-publikasjoner/metodikk-for-modernisering-av-statistikkproduksjonen

Eurostat, data og metadata exchange: https://cros-legacy.ec.europa.eu/content/data-and-metadata-exchange_en

Eurostat (2018). Methodology for data validation 2.0, https://ec.europa.eu/eurostat/cros/system/files/ess_handbook_-_methodology_for_data_validation_v2.0_-_rev2018_0.pdf

Eurostat (2020). ESS vision 2020 validation, principles for data validation, https://ec.europa.eu/eurostat/cros/content/principles_en

Foss og Haugen (2023). Anbefalte kvalitetsindikatorer i offisiell statistikk. Notater 2023/xx, «notat kommer»

Hustoft og Sæbø (2006). Sentrale begreper knyttet til metadata – til bruk i SSBs felles metadatasystemer. Interne dokumenter 2006/2. https://www.ssb.no/a/metadata/definisjoner/begreper.pdf

UNECE (2019). Generic Statistical Data Editing Model, Version 2.0, https://statswiki.unece.org/display/sde/GSDEM

SSB (2005). Datarevisjon Kontroll, gransking og retting av data. Anbefalt praksis. Statistisk sentralbyrås håndbøker 84, https://www.ssb.no/a/histstat/ssh/ssh_84.pdf

SSB (2019). Generic Statistical Business Process Model GSBPM, version 5.1, January 2019. Norsk oversettelse. Notater 2019/43, https://www.ssb.no/befolkning/artikler-og-publikasjoner/generic-statistical-business-process-model-gsbpm

SSB (2021). Datatilstander i SSB. Interne dokumenter 2021/17

Vedlegg A: Feilkoder for fødselsnummer

Dette er feilkodene som er foreslått for fødselsnummer (FNR) og DNR.

Tabell A.1 Feilkoder for fødselsnummer

Kode	Tekst
A1	fnr(1-6) = ddmmåå og fnr(7-11) = nnnnn kunstig/ugyldig
A2	fnr(1-6) = ddmmåå og fnr(7-11) = blanke
A3	fnr(1-6) = ddmmåå og fnr(7-11) = 00000/0____
A4	fnr(1-6) = ddmmåå og fnr(7-11) = nnn + blanke
A5	fnr(1-6) = ddmmåå og fnr(7-11) = nnnn + blank
A6	fnr(1-6) = ddmmåå og fnr(7-11) = xxxxx
B1	dnr(1-6) = ddmmåå og dnr(7-11) = nnnnn kunstig/ugyldig
B2	dnr(1-6) = ddmmåå og dnr(7-11) = blanke
B3	dnr(1-6) = ddmmåå og dnr(7-11) = 00000/0____
B4	dnr(1-6) = ddmmåå og dnr(7-11) = nnn + blanke
B5	dnr(1-6) = ddmmåå og dnr(7-11) = nnnn + blank
B6	dnr(1-6) = ddmmåå og fnr(7-11) = xxxxx
C1	fnr(1-11) = ddmmååännnnn (+50 i mm) kunstig/ugyldig
C2	fnr(1-11) = ddmmååännnnn (+20 i mm) kunstig/ugyldig
D1	fnr(1-5) = dmmåå og fnr(6-11) = 99999_
D2	fnr(1-5) = dmmåå og fnr(6-11) = 000000/00000_
D3	fnr(1-5) = 00000 og fnr(6-11) = ddmmåå
E	fnr(1-11) = 00000000000/0_/blanke
F1	fnr(1-4) = åååå og fnr(5-11) = blanke
F2	fnr(1-6) = 00åååå og fnr(7-11) = blanke
F3	fnr(1-11) = annen feil

Vedlegg B: DUF-nummer

Personer som har søkt om asyl eller oppholdstillatelse i Norge får DUF-nummer i UDIs datasystem, og dette nummeret brukes til de får oppholdstillatelse og DNR. DUF(1-4) = åååå og DUF(5-12) = nnnnnnnn. åååå er årstallet vedkommende søkte om asyl/opphold + det nummeret vedkommende ble registrert med i UDIs datasystem. Ved gjentatte søknader beholdes samme DUF-nummer som ved første gangs søknad.