Statistisk sentralbyrå
Statistics Norway

Metodikk for modernisering av statistikkproduksjonen

TALL SOM FORTELLER
NOTATER / DOCUMENTS
2020/ 21

Remy Bråthen, Ane Seierstad og Aslaug Hurlen Foss

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

Publisert 6. mai 2020

ISBN 978-82-587-1129-9 (elektronisk)
ISSN 2535-7271 (elektronisk)

Standardtegn i tabeller	**Symbol**
Tall kan ikke forekomme	.
Oppgave mangler	..
Oppgave mangler foreløpig	…
Tall kan ikke offentliggjøres	:
Null	-
Mindre enn 0,5 av den brukte enheten	0
Mindre enn 0,05 av den brukte enheten	0,0
Foreløpig tall	*
Brudd i den loddrette serien	—
Brudd i den vannrette serien	|
Desimaltegn	,

Forord

Dette notatet presenterer en bearbeidet versjon av metodikken som er utarbeidet i forbindelse med modernisering av statistikkproduksjonen. All metodikk i notatet tar utgangspunkt i internasjonale standarder, artikler i publiserte tidsskrifter, på konferanser eller i andre nasjonale statistikkbyråer. Når vi har brukt dette materialet har vi i dette notatet tatt med referansene og lenkene til hvor og når vi har hentet informasjonen.

Statistisk sentralbyrå, 16. april 2020

Arvid Olav Lysø

Sammendrag

Det første som ble utarbeidet i arbeidet med å modernisere statistikkproduksjonen er hvilke prinsipper en moderne produksjon bør vektlegge. Disse prinsippene gjelder temaer som faglig forankring, metadata, dokumentasjon, arkivering, automatisering, dataeditering, kvalitet, evaluering og effektive metoder.

Det grunnleggende for å bygge en statistikkproduksjon er å ha tilgang til funksjoner. Nødvendige funksjoner har blitt identifisert og kategorisert her/i dette notatet. Funksjoner må være tilgjengelig for statistikkproduksjonen ved at de enten finnes i et programmeringsspråk man har tilgang til, kan bygges av statistikk-produsenten selv eller kan tas i bruk via bruk av «internasjonale biblioteker», slik som for eksempel Cran for de som bruker programmeringsspråket R.

Deretter har vi sett på hvordan funksjoner kan settes sammen til en prosess. De fleste prosesser har minst tre prosessteg - forprosess, hovedprosess og etterprosess. Forprosessen er ofte en tilrettelegging av data for hovedprosessen, mens etterprosessen er en kontroll av at hovedprosessen har gjort det den skal med tilfredsstillende kvalitet.

Data går bearbeides gjennom hele produksjonsprosessen. En datatilstand er et datasett med metadata på et gitt punkt i produksjonsprosessen. Hovedtilstandene er de viktigste datatilstander i produksjonsprosessen: rådata, inndata, klargjorte data, statistikkdata og utdata. Vi har laget en oversikt over hvilke delprosesser som inngår i utarbeidelsen av hver hovedtilstand og hvilke kvalitetskriterier hver av hovedtilstandene har.

Deretter har vi laget forslag til hvilke kvalitetsmål som kan beregnes for å kunne måle om kvalitetskravene til hovedtilstandene er oppfylt. Kvalitetsmålene er kategorisert etter tema: Metadata, populasjon-enhet og variabler. Vi har i tillegg sett på hva som skal til for å tilrettelegge et statistikksystem for å kunne lage kvalitetsmålene. Det viktigste er å ta vare på informasjon som genereres underveis, fra de ulike prosessene, slik at de i etterkant kan brukes til å beregne kvalitetsmål.

Til slutt presenterer vi en prosessmodell som legger vekt på automatisering, makroperspektiv, kvalitet og datatilstander. Modellen et sett av prosesser satt i rekkefølge, der den menneskelige interaksjonen med dataene er skilt klart ut fra det som kan bli gjort maskinelt. I denne modellen har vi integrert kvalitet i alle prosesser og produkter som inngår. Dette er en metode for å sikre god kvalitet på sluttproduktene.

Innhold

1. Prinsipper for fasene klargjøre og analysere

Det er laget et målbilde for fasene *klargjøre* og *analysere*; en visjon av hvordan framtiden vil se ut. For å tydeliggjøre målbildet er det utarbeidet prinsipper som skal fungere som et sett med felles retningslinjer for hvordan statistikk bør produseres. De kan også brukes som en huskeliste for design av statistikkproduksjonen.

Tabell 1.1 Prinsipper for fasene klargjøre og analysere

Navn	Fag- og emnekunnskap
Prinsipp	Planleggingen og gjennomføringen av statistikkproduksjonen må baseres på fagkunnskap.
Begrunnelse	Fagkunnskap om emnet statistikken handler om, er en forutsetning for å planlegge og utføre en god klargjørings- og analyseprosess og for å lage et statistikkprodukt av god kvalitet.
Konsekvens	Statistikksystemet skal tilby funksjoner og løsninger som dekker behovene de enkelte statistikkprodusentene har for å lage en effektiv produksjonsprosess av god kvalitet. For å få dette til, må personer med fagkunnskap involveres i utviklingen av statistikksystemet.
Navn	Fleksibel datastruktur
Prinsipp	Datastrukturen skal ikke være til hinder for enkelte prosessteg eller for hele arbeidsflyten.
Begrunnelse	Datastruktur er et viktig element inn og ut av funksjoner.
Konsekvens	Statistikksystemet må tilrettelegge datastrukturen slik at funksjonene utføres korrekt, men samtidig slik at helheten i prosessen blir ivaretatt. Strukturering av data kan foregå inni funksjonen der det er hensiktsmessig.
Navn	Metadata og prosessdata
Prinsipp	Metadata og prosessdata er like viktig som data.
Begrunnelse	God oversikt over data og metadata inkl. prosessdata gjennom hele produksjonsprosessen er avgjørende for å kunne gjøre veloverveide valg og for å kunne vurdere kvaliteten i statistikkproduksjonen og sluttproduktet.
Konsekvens	Metadata og prosessdata må være lett tilgjengelige for de som produserer statistikk. Metadata og prosessdata må kunne oppdateres og tas vare på for alle manuelle og automatiske handlinger utført på data fra inndata til publisering. Kvalitetsindikatorer bør være en del av prosessdata, og skal beregnes og oppdateres for hvert steg der det er relevant.
Navn	Dokumentasjon av metoder, funksjoner og prosesser
Prinsipp	Metoder, funksjoner og prosesser skal dokumenteres og all slik dokumentasjon må gjøres lett tilgjengelig.
Begrunnelse	Informasjon om hva som går inn i, ut av og blir utført i de ulike metodene, funksjonene og prosessene er viktig for å kunne lage en selvbetjent løsning med mulighet for mer automatiske prosesser.
Konsekvens	Statistikksystemet må tilby god dokumentasjon av metoder, funksjoner og prosesser som tilbys brukeren. Dokumentasjonen må være enkelt tilgjengelig, søkbar og tilknyttet de aktuelle metoder, funksjoner og prosesser. I tillegg må statistikksystemet tilrettelegge for en enkel tilgang til nasjonale og internasjonale retningslinjer/beste praksis.
Navn	Arkivering av data skal automatiseres
Prinsipp	Data og tilhørende metadata må langtidslagres etter gjeldende prinsipper for informasjonsforvaltning, og denne prosessen må gjøres mest mulig automatisert.
Begrunnelse	Sikring og automatisering av langtidslagring fører til at data lagres og at arkivloven (Lovdata, 2018) blir overholdt. Dette sikrer at dataene kan gjenbrukes i framtiden for å lage historiske analyser.
Konsekvens	For statistikksystemet betyr dette at den må tilby en løsning som sikrer at prosessen med langtidslagring av data og metadata blir gjennomført. Prosessen må automatiseres slik at det blir minst mulig manuelle operasjoner.
Navn	Automatisering av prosesser
Prinsipp	Stegene i klargjørings- og analyseprosessen skal kunne automatiseres.
Begrunnelse	Økt bruk av automatiserte steg i produksjonsprosessen vil gjøre produksjonen mer effektiv.
Konsekvens	Funksjonaliteten i statistikksystemet må være bygget opp slik at systemet er i stand til å gjenta brukernes handlinger i den rekkefølgen det er ønskelig.
Navn	Utvikle effektive funksjoner og metoder
Prinsipp	Funksjoner og metoder som kan effektivisere prosessen eller erstatte manuelt arbeid skal benyttes.
Begrunnelse	En effektiv metode eller funksjon kan erstatte mange steg i en prosess eller gjøre de raskere.
Konsekvens	Det må legges til rette for at effektive funksjoner og metoder kan utvikles i SSB. Disse funksjonene og metodene skal være tilgjengelig for brukerne av statistikksystemet f.eks. via et metodebibliotek.

Navn	Mer bruk av grafikk
Prinsipp	Grafikk skal benyttes som en integrert del av all statistikkproduksjonen.
Begrunnelse	Grafikk kan gi oversikt over trender og strukturer i data. Grafikk kan også gi oversikt over produksjonsløp og kvalitet.
Konsekvens	Statistikksystemet skal tilby grafikk som en integrert del av produksjonsprosessen. Statistikksystemet må både tilby mulighet for selvbetjening av grafikk og som en standardisert del av en prosess.
Navn	Tidlig korrigering
Prinsipp	Feil som med sikkerhet er feil bør kontrolleres og korrigeres så tidlig som mulig i prosessen.
Begrunnelse	Kontrollering og korrigering tidlig fører til at de resterende prosessene ikke blir påvirket av feilen.
Konsekvens	Statistikksystemet må tilrettelegge slik at kontrollering og korrigering kan skje så tidlig som mulig i prosessen.
Navn	Makroperspektiv
Prinsipp	Produksjonsprosessen bør i størst mulig grad fokusere på de aggregerte sluttprodukt og prioritere å kontrollere de verdiene som påvirker det mest.
Begrunnelse	Fokus på det som påvirker sluttproduktet mest gir overordnet perspektiv på statistikken og hjelp til å prioritere hva som er viktig og med det sikre en effektiv ressursbruk.
Konsekvens	Statistikksystemet må gi brukerne mulighet til å se sluttproduktet og resultater på ulike aggregeringsnivåer når det er behov for det underveis i produksjonen. Det må være mulig å lete etter mistenkte feil ved å «drille» seg nedover i nivåer. Statistikksystemet må tilby funksjoner for å selektere verdier som har stor innflytelse på sluttproduktet.
Navn	Sirkulær prosess
Prinsipp	Klargjøringsprosessen bør være en sirkulær prosess, der en enkelt kan bevege seg gjennom alle prosesser fram til sluttproduktet flere ganger i løpet av produksjonsprosessen.
Begrunnelse	En sirkulær prosess vil gjøre det mulig å styre prosessen etter hvordan de ulike stegene påvirker hverandre, og hvordan de til syvende og sist påvirker sluttproduktet.
Konsekvens	Det må være enkelt å gå fra et prosessteg til et annet, slik at det er mulig å kjøre hele eller deler av produksjonsprosessen flere ganger for å se hvilken betydning eventuelle tiltak man gjør kan få. For eksempel kan man tenke seg at man starter med å lage et tidlig estimat (eventuelt en framskrivning av fjorårets tall) på sluttproduktet, lager et eller flere estimater underveis og slutter når estimatet av sluttproduktet har nådd den kvaliteten man ønsker.
Navn	Kvalitetsmål
Prinsipp	Kvaliteten i statistikkproduksjonen skal måles ved hjelp av kvalitetsmål.
Begrunnelse	Det må utarbeides kvalitetsmål for hver hovedprosess i statistikkproduksjonen. Kvalitetsmålene gir statistikkprodusenten et verktøy til å styre, overvåke og evaluere produksjonsprosesser. Kvalitetsmål inngår også som bekrivelse av produktkvalitet, og bør være en del av opplysningene i "Om statistikken" og i rapporter til Finansdepartementet og Eurostat.
Konsekvens	Statistikksystemet må tilby kvalitetsmål gjennom hele produksjonsprosessen og samle kvalitetsmålene i en rapport ved prosesslutt. Statistikksystemet må lagre kvalitetsmålene og vise dem til statistikkprodusenten. Statistikksystemet må legge til rette for muligheten for å lage egne kvalitetsmål tilpasset sin statistikk.
Navn	Evaluering
Prinsipp	Produksjonsprosessen og datakvalitet skal evalueres jevnlig.
Begrunnelse	Evaluering av produksjonsprosessen og datakvalitet er viktig for å kunne samle kunnskap slik at forbedringstiltak kan settes inn. Hvis tiltak blir satt inn, ved for eksempel at feil ikke oppstår igjen, vil prosessen bli mer effektiv og kvaliteten på statistikken bli høyere.
Konsekvens	Statistikksystemet må tilrettelegge for evaluering av produksjonsprosess og datakvalitet. Det må være mulig å sammenligne data før og etter ulike steg i prosessen, slik at effekten av hvert steg kan evalueres. Det bør også være mulig å foreta sammenligninger med tidligere perioder for å se hvordan eventuelle forbedringer/endringer slår ut på statistikken. Kvalitetsmål blir ofte brukt til å evaluere.
Navn	Standardisering og gjenbruk av prosesser
Prinsipp	Prosesser som mange benytter skal bli standardisert og gjenbrukes av aktuelle statistikker.
Begrunnelse	Standardisering og gjenbruk av prosesser gjør at bygging av statistikkprosessen blir mer effektiv og det blir færre komponenter å vedlikeholde.
Konsekvens	Statistikksystemet må tilby standardiserte prosesser. Gjenbrukbare prosesser må identifiseres og bygges.
Navn	Samordning av felles data og editeringsprosess
Prinsipp	Data som blir brukt til å lage flere statistikker skal samordnes og editeringsprosessen skal være felles lengst mulig.
Begrunnelse	Samordning av felles data og editeringsprosess fører til at statistikken som blir laget ut fra felles datakilde blir mest mulig sammenlignbar.
Konsekvens	Statistikksystemet må tilby løsninger for å håndtere felles data og editeringsprosess for data som skal deles av mange.

2. Identifisering og kategorisering av funksjoner

Vi har identifisert og kategorisert funksjoner det er behov for ved bearbeiding av data i produksjon av statistikk. Dette arbeidet er konsistent med den generiske statistiske informasjonsmodellen: «Generic Statistical Information Model» (GSIM) versjon 1.1 og SSBs Informasjonsmodell (SSB-IM). Disse modellene samsvarer med virksomhetsmodellen: Generic Statistical Business Process Model (GSBPM) versjon 5.1. I GSIM og SSB-IM er navnet «forretningsfunksjoner» brukt istedenfor «funksjoner». Kategoriseringsarbeidet vårt følger i stor grad samme tilnærming som den generiske statistiske dataediteringsmodellen: Generic Statistical Data Editing Model (GSDEM) versjon 1.0. Hovedskillet mellom denne funksjonskategoriseringen og GSDEM er at vår kategorisering beskriver funksjoner tilknyttet all bearbeiding av data, mens GSDEM er fokusert på editeringsfunksjoner. I vår funksjonskategorisering dekker hovedkategoriene ***korrigere*** og ***kontrollere*** i stor grad det vi finner som editeringsfunksjoner i GSDEM, mens hovedkategoriene, ***beregne, strukturere og presentere*** dekker de resterende funksjonene som brukes i bearbeiding av data.

Funksjoner kan kombineres i et nær sagt uendelig antall kombinasjoner. Det er ofte dette som er årsaken til at mange oppfatter sitt statistikkprodukt som unikt. Selv om kombinasjonene av funksjoner i et gitt statistikkprodukt er unikt for det produktet, så er ikke selve funksjonene som produktet består av unike. De samme funksjonene vil vi finne igjen i en rekke statistikkprodukter. Det er nettopp dette poenget som lar oss standardisere produksjon av statistikk; ved at vi ikke standardiserer produksjonsløpene, men byggeklossene som produksjonsløpene består av.

Hver funksjon har i sin kjerne prosessmetoder (GSIM/SSB-IM) som bidrar med metodene som gir funksjonen. En funksjon kan bestå av en rekke prosessmetoder. Eksempelvis vil funksjonen ***fordelingskontroll*** kunne ha både HB-metoden og Regresjonsanalyse som prosessmetode.

De kategoriene av funksjoner som beskrives er ikke låst fast til noen bestemt fase av produksjonsprosessen, men må ses på som byggeklosser som kan brukes i ulike deler av produksjonen og som lar oss bygge et mangfold av produksjonsløp. Målet er, som sagt, ikke å standardisere produksjonsløpene, men byggeklossene, her representert med kategorier av funksjoner. Slik sikrer vi fleksibilitet i produksjonen samtidig som vi standardiserer.

GSBPM beskriver prosessene i produksjon av statistikk, funksjonene bidrar til å realisere prosessene. Funksjoner har en løs tilknytning til prosessene, der en funksjon kan brukes i en rekke sammenhenger og ikke er bundet til en enkelt prosess.

For eksempel vil ***logisk kontroll*** kunne brukes i flere faser av produksjonsprosessen:

Sjekk av gyldig verdi (for eksempel x>0):

- Sjekk av x>0 ved innrapportering av data, kanskje allerede i skjema. Fase: ***Samle inn***.
- Sjekk av x>0 når data har kommet inn, som en del av editeringen. Tas da ut til videre behandling. Fase: ***Klargjøre***.
- Sjekk av at x>0 som en siste kvalitetssjekk før publisering. Fase: ***Analysere***.

Arbeidet med å kategorisere funksjoner har bestått av å identifisere hoved-kategorier av funksjoner og så bryte disse videre ned i underkategorier av funksjoner.

Hovedkategoriene er følgende kategorier:

- Kontrollere
- Korrigere
- Strukturere
- Beregne
- Presentere

2.1. Kontrollere

Hovedkategorien ***kontrollere*** består av funksjoner for å finne mulige feil og mangler i data. Funksjoner for å se på tabeller og grafikk, som en metode for å kontrollere data, er ikke tatt med da dette dekkes av funksjoner i hovedkategori ***presentere***. Funksjonene i kategorien ***kontrollere*** kan resultere i seleksjon av enheter/variable uttrykt ved en boolsk variabel eller et kvalitetsmål (poeng-funksjon) som kan brukes til vurdering om verdien er korrekt. I GSDEMs er ***kontrollere*** delt opp i to hovedkategorier ut fra om funksjonen skal brukes til vurdering eller seleksjon. Vi har ikke gjort denne oppdelingen fordi det til dels er de samme funksjonene som kan gå igjen i begge kategoriene. I hovedsak er det bare hva resultatet skal brukes til som er forskjellig i de to kategoriene. I tillegg til funksjoner for å finne feil, er informasjon om enhet og variabel et viktig virkemiddel for å vurdere om verdien er korrekt.

Tabell 2.1 Kategorier av funksjoner i hovedkategori kontrollere

Kategori	Underkategori	Beskrivelse	Resultat av metoden
Logisk kontroll	Logisk kontroll av kategoriske verdier	Et sett av gyldige kategoriske verdier er definert for variabelen.	Boolsk verdi
	Logisk kontroll av numeriske verdier	Et gyldig intervall for verdiene er definert for variabelen.	
Enhetskontroll	Dublettkontroll		Boolsk verdi
Fordelingskontroll	Univariat kontroll	Kontroll basert på fordelingen til en variabel, for eksempel kvartilmetode.	Poengsum og/eller boolsk verdi
	Bivariat kontroll	Kontroll basert på korrelasjonen mellom to variabler, for eksempel HB-metoden.	
	Multivariat kontroll	Kontroll basert på fordelingen av mange variabler samtidig. For eksempel regresjon med mange variabler og/eller Mahalanobisdistansen.	
Innflytelseskontroll	Andel av total	Enhetens andel av totalen for en variabel.	Poengsum og/eller boolsk verdi
	Andel av endringstall	Enhetens andel av endringen for en variabel.	
	Innvirkning på beregning	Enhetens påvirkning på beregninger som snitt og indeks for en variabel.	
	Multivariat innflytelse	Samlet poengsum for hvordan variablene påvirker sluttproduktet, for eksempel poengfunksjon.	

2.2. Korrigere

Hovedkategorien ***korrigere*** består av funksjoner for å endre verdier som er rapportert inn eller lage verdier som mangler. Det omfatter ikke å lage nye variabler, funksjonaliteten for dette er lagt til kategorien ***beregne.***

Tabell 2.2 Kategorier av funksjoner i hovedkategori korrigere

Kategori	Underkategori	Beskrivelse
Manuell korrigering	Rekontakt	Spørre om ny verdi fra oppgavegiver.
	Verdierstatning	Bruke en verdi fra en annen kilde.
	Verdiopprettelse	Sette verdi ut fra kunnskap om emnet.
Regelbasert imputering	Funksjonsimputering	Beregne en verdi ut fra en deterministisk funksjon av andre verdier.
	Logisk imputering	Utlede en verdi fra logiske regler.
	Historisk imputering	Bruke en verdi fra tidligere periode.
Modellbasert imputering	Median-imputering	Bruke medianen til variabelen.
	Gjennomsnitts-imputering	Bruke gjennomsnittet til variabelen
	Regresjons-imputering	Predikere en verdi med en regresjonsmodell.
Donorimputering	Tilfeldig donor-imputering	Velge en donor tilfeldig.
	Sekvensiell donorimputering	Velge donor sekvensielt.
	Nærmeste nabo-imputering	Velge en donor ut fra en avstandsfunksjon.
	Proxy-imputering	Adoptere en enhet fra en relatert enhet.
Konsistensjusteringer	Balansering	Justere basert på konsistensbetingelser.
	Prorating	Justere blokker av ek sisterende verdier for å oppnå konsistens.
	Forholdskorrigere donorimputering	Imputere en donorverdi med forholdsjustering for å oppnå konsistens.

2.3. Str0075kturere

Hovedkategorien ***Strukturere*** består av funksjoner for å utlede nye datasett basert på ett eller flere eksisterende datasett, der hvert enkelt datasett inneholder nøkkel-informasjon som gjør at datasettene kan kobles sammen, og kan knyttes opp mot enhetene. Dette stiller krav til grunndatasettene som skal sammenstilles eller struktureres.

Tabell 2.3 Kategorier av funksjoner i hovedkategori strukturere

Kategori	Underkategorier	Beskrivelse
Aggregere	Summere	Legge sammen tall
	Summere hierarkisk	Legge sammen tall i en hierarkisk struktur
Koble	Koble eksakt	Koble sammen to datasett ved hjelp av koblingsnøkkel eller identifiseringsvariabler slik at datasettet består av variablene fra begge datasett. Prosessmetoder: indre, ytre, venstre og høyre.
	Koble probabilistisk	Koble sammen to datasett ut fra hvilke enheter som sannsynligvis hører sammen uten koblingsnøkkel eller identifiseringsvariable.
Sammensette		Slå sammen to eller flere datasett.
Pivotere		Gjøre om rader til kolonner og omvendt.
Filtrere	Filtrere variabler	Filtrering av variable ved å angi hvilke variabler som skal være med i det nye datasettet.
	Filtrere enheter	Filtrere ut enheter ved hjelp av gitte betingelser.
Sortere		Sortere variabel fra minst til størst eller omvendt
Utvalgstrekke		Trekke et utvalg av enheter basert på utvalgsplan. Prosessmetoder: • Enkelt tilfeldig • Stratifisert • Systematisk • Klynge • PPS • Totrinns • Selvveiende
Formatere	Datasett-format	Omgjøre data til et annet dataformat, for eksempel til semikolonseparert fil.
	Variabel-format	Omgjøre format på variabler: numerisk, heltall, desimaltall, tekst, dato og tid, avrunding.

2.4. Beregne

Hovedkategorien ***beregne*** består av funksjoner knyttet til å kalkulere en ny verdi basert på eksisterende verdier. Dette vil si at vi lager nye variabler basert på en utregning av variabler vi allerede har. Beregningene kan være både enkle, som for eksempel en sum av variabler, eller mer komplekse beregninger som indikatorer, indekser og statistiske metoder som sesongjustering.

Tabell 2.4 Kategorier av funksjoner i hovedkategori beregne

Kategori	Underkategori	Beskrivelse
Avledet variabel Avledet variabel er en variabel som er beregnet ut fra eksisterende variabler.	Avledet numerisk variabel	Avledet numerisk variabler blir laget ut fra aritmetiske operatorer mellom flere numeriske variabler.
	Avledet kategorisk variabel	Avledet kategorisk variabel blir til ved for eksempel å slå sammen to kategoriske variabler (konkatenering), dele opp en kategorisk variabel eller ved å lage kategorier av numeriske variabler ved hjelp av betingelser.
	Avledet variabel fra variabelnavn	Hvis et variabelnavn består av to sammensatte navn, kan disse splittes og det kan bli laget to variabler.
Avledet enhet Avledet enhet er en enhet som er beregnet ut fra andre enheter og variabler	Lage overordnet enhet ut fra underordnete enheter	Sammen med de avledete enhetene vil det ofte bli laget tilhørende avledete variabler. Eksempel: lage foretak ut fra virksomheter i foretaket.
	Lage underordnede enheter ut fra overordnet enhet	Eksempel: imputere virksomheter ut fra foretak.
	Masseimputering	Imputere alle verdier for en enhet samlet.
Fordelingsberegninger	Rang	Rangeringsnummeret av en verdi i en gruppe.
	Snitt	Beregne gjennomsnitt i en gruppe.
	Antall	Telle opp antall i en gruppe.
	Minimum	Finne minimumsverdien i en gruppe.
	Maksimum	Finne maksimumsverdien i en gruppe.
	Kvartiler	Finne kvartilverdiene i en gruppe, inklusiv median.
Matriseberegninger	Vektor	Beregninger som er gjort på vektorer.
	Matrise	Beregninger som trenger å operere på både rader og kolonner. Matriseoperasjoner: addisjon og subtraksjon, skalarmultiplikasjon (matrisemultiplikasjon, transponering, invers).
	Flerdimensjonal matrise	Beregninger på hyperdimensjonale kuber
Vektberegninger	Modellbasert	Beregner vekter for hver enhet basert på en modell av data. Eksempler på modeller er ratemodell, vanlig regresjon og homogen modell.
	Designbasert	Beregner vekter ut fra trekksannsynlighet og frafall.
	Vektkalibrering	Justerer vekter slik at totalene stemmer med kjente verdier.
Indeksberegninger	Personindekser	Eksempler: levekårsindeks, likestillingsindeks.
	Økonomiske indekser	Eksempler: volumindeks, verdiindeks, prisindeks og tjenesteprisindeks.
	Elementærindeks	Prosessmetoder: • Carli • Dutrot • Jevans
	Aggregeringsmetode over elementærindeks	Prosessmetoder: • L-type • P-type • Fisher
	Deflatere	Beregne fastpris.
	Kjeding	Beregninger for å knytte eldre og nyere indekser sammen

Kategori	Underkategori	Beskrivelse
Sesongjustering	X-12-ARIMA	Programvare som utfører sesongjustering med utgangpunkt i en metode som bruker en kombinasjon av filtre og arimamodell.
	TRAMO-SEATS	Programvare som utfører sesongjustering ved modellbasert metode.
	Multivariat	Metoder som simultant sesongjusterer flere serier på en gang.
Konfidensialitet	Undertrykking	Undertrykking er en metode som sikrer at informasjon om individuelle enheter ikke kan avsløres på grunnlag av tabeller eller mikrodatasett, ved å primær- og sekundærprikke data.
	Tilfeldig avrunding	Tilfeldig avrunding sikrer at informasjon om individuelle enheter ikke kan avsløres på grunnlag av tabeller eller mikrodatasett ved hjelp av tilfeldig avrunding.
	Støylegging	Legge til tilfeldig støy på antall og numeriske summer.
Kvalitetsberegninger	Usikkerhetsberegninger	Funksjoner for å beregne usikkerheten, slik som Monte Carlo-simuleringer, Jack-knife-metoder og lignende.
Analyseberegninger	Regresjonsanalyse	Regresjonsanalyse er en samlebetegnelse på statistiske analysemetoder som har til mål å beskrive sammenhengen mellom én eller flere uavhengige variabler *(x_1, x_2, og så videre)* og en avhengig variabel (y).
	Tidsserieanalyse	Tidsserieanalyse er en samlebetegnelse på statistisk analysemetode som har til formål å beskrive variasjonsmønsteret i en tidsserie med målsetning om å bedre forstå mønsteret, hva som er hendt, og derved mulighet for å fremskrive det videre forløp for en eller flere perioder
	Forløpsanalyse	Forløpsanalyse er en samlebetegnelse på statistisk analysemetode av forløp av hendelser, varighet og levetider.
	Romlig analyse	Romlig analyse er en samlebetegnelse på analysemetoder som er avhengige av teknikker og teknologier som måler den relative form, avstand og plassering av objekter.
Algoritmer og maskinlæring		Algoritmer og maskinlæring er grupper av funksjoner som utfører en endelig serie av operasjoner for å løse et problem eller flere problemer.
Utvalgsplan	Utvalgsmetoder	En utvalgsplan er en plan for hvordan å trekke et utvalg av enheter. Eksempler på prosessmetoder: • Enkelt tilfeldig utvalg • Stratifisert utvalg • Systematisk utvalg • Klyngeutvalg • PPS-utvalg • Totrinnsutvalg • Selvveiende utvalg
	Allokeringsmetoder	Allokeringsmetoder er metoder for å fordele, eller allokere, utvalget på strataene. Eksempler: proporsjonal allokering og optimal allokering.
	Stratifiseringsmetoder	Metoder for å finne homogene grupper. Cluster-analyse.

2.5. Presentere

Hovedkategorien ***presentere*** består av funksjoner for visning av data. Det mest vanlige er å vise tall i form av en tabell. Det har ikke blitt laget underkategorier for kategorien tabell, da funksjonaliteten i visningene av en tabell er det viktigste. I en tabellvisning bør det være mulighet for å kunne selektere og se deler av datasettet. Dette er spesielt viktig når det fins mange rader og kolonner. Det er nyttig med et fast oppsett på tabeller som ofte blir brukt. Ved inspeksjon av data er det mange som synes det er effektivt å kunne drille i tabeller mellom makro- og mikro-nivå i

statistikken. Markering av data med forskjellige egenskaper kan ofte hjelpe til når tallene skal bli vurdert. Det kan være markering av for eksempel manglende verdi, imputert verdi eller ekstremverdi.

En annen måte å framstille tall på er gjennom grafikk. Grafikk viser data og kan presentere mange tall på liten plass. Grafikk kan avslør dataene på forskjellige detaljnivåer. Grafikk kan vise variasjon og skjevheter i datasettet, som ikke ville blitt oppdaget uten. Det fins mye funksjonalitet tilknyttet grafikk som kan være svært nyttig i statistikkproduksjon, for eksempel bruk av farger for å markere forskjeller, mulighet til å lage mange figurer samtidig, informasjon som popper opp når musa holdes over et punkt, mulighet for å selektere deler av grafen, drille mellom grafer på forskjellige aggregeringsnivå og gå fra et markert punkt i grafikk til datatabellen for å korrigere feil. Grafikk krever ofte en strukturering av data tilpasset grafen som skal bli laget. Metadata er viktig for å kunne tolke grafikken, det vil si titler, akse-merking, forklaringer av symboler.

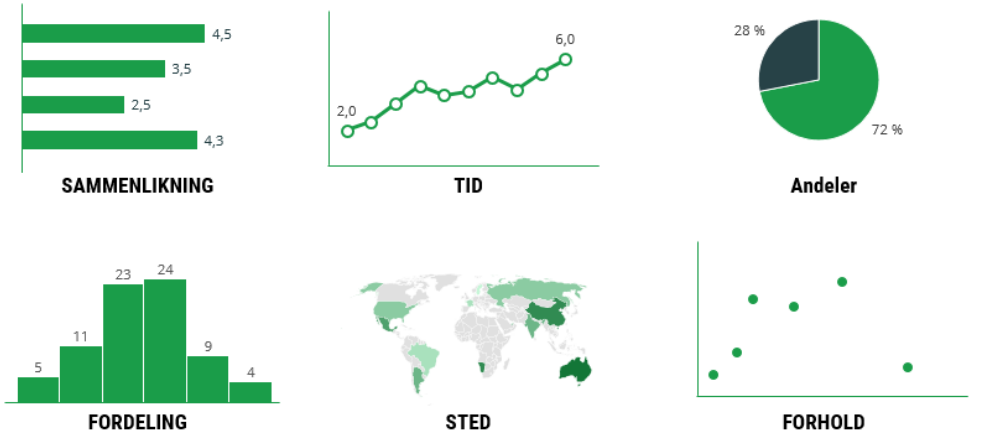
Det finnes mange forskjellige typer grafer, avhengig av hva en ønsker å studere. De forskjellige grafene kan deles inn i forskjellige underkategorier. En inndeling av grafer som ofte er brukt er: sammenligning, fordeling, andeler, forhold, sted og tid. Komplekse grafer kan bli laget ved å kombinere de ulike underkategoriene.

I statistikksystemer brukes tekst ofte i kommunikasjon med brukeren om hvordan en prosess har kjørt ofte i form av en log eller en feilmelding. Tekst brukes også for å dokumentere og forklare kode i et program.

Tabell 2.5 Kategorier av funksjoner i hovedkategori presentere

Kategori	Underkategori	Beskrivelse
Tabell		Visning av tallene i en tabell
Tekst		Kommunikasjon til brukerne med en tekst
Grafikk	Sammenligning	Figurer som sammenligner størrelser, ved for eksempel forskjellige typer stolpediagram.
	Fordeling	Figurer som viser hvilke verdier en variabel har og hvilke verdier som forekommer ofte.
	Andeler	Figurer som viser andelen av forskjellige kategorier, for eksempel forskjellige former for kakediagram.
	Forhold	Figurer som viser sammenheng mellom variabler for eksempel x-y-plot viser korrelasjonen mellom x og y.
	Sted	Figurer som viser tallene geografisk fordelt.
	Tid	Figurer som blir brukt til å vise utvikling over tid, for eksempel forskjellige typer linjediagram.

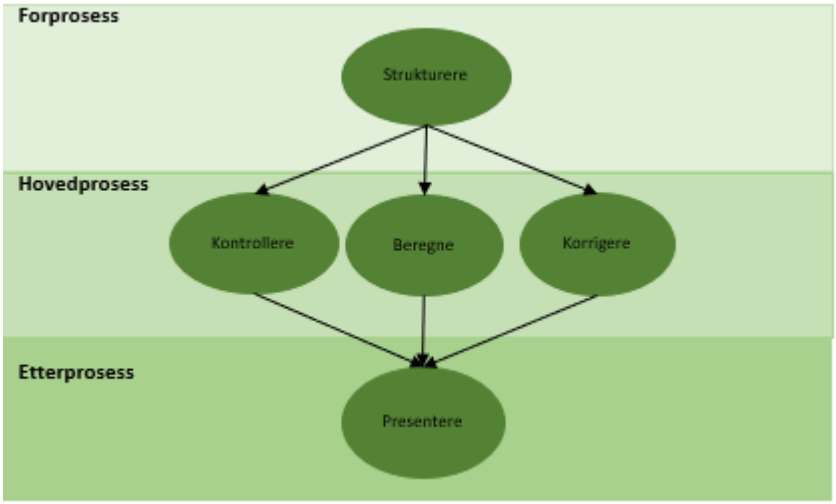
Figurer 2.5 Underkategorier av grafikk



3. Sammensetting av funksjoner til prosesser

De fleste prosesser består av en samling av flere funksjoner. Det finnes en del fellestrekk i hvordan prosesser blir bygd opp. Ofte starter prosessen med å strukturere data slik at de tilfredsstiller kravene til hovedprosessen. En slik forprosess kan også ha kontroll av sider ved data som må være oppfylt for å kunne kjøre funksjonene i hovedprosessen. Det kan f.eks. være minimum antall observasjoner for å kunne kjøre regresjon eller kontroll av om det er dubletter før data skal kobles. Hovedprosessen består ofte av funksjoner fra hovedkategoriene ***kontrollere, korrigere og beregne***. Resultatet av hovedprosessen blir presentert i tekst, tabeller og grafikk. Resultatet innbefatter også kvalitetsmål og logger i form av tekst som forteller om hvordan prosessen er gjennomført.

Figur 3.1 Generell sammensetning av funksjoner til prosess



4. Metodebibliotek

4.1. Definisjon av metodebibliotek

Et metodebibliotek er et arkiv som har til oppgave å oppbevare, katalogisere og gjøre tilgjengelig metoder for statistikkproduksjonen. Med metoder menes her alle slags teknikker, operasjoner eller funksjoner som kan bli utført på et datasett i statistikkproduksjonen, det omfatter altså ikke bare statistiske metoder. Metoder kan være både det som er definert som prosessmetoder og prosessteg i GSIM, så lenge det er gjenbrukbart av flere statistikker. Metodebiblioteket bør også gi statistikkprodusenten informasjon om innholdet i metodene og hvordan bruke disse metodene.

4.2. Målsetting med metodebibliotek

Målsetting med biblioteket er å stille til disposisjon alle typer metoder som trengs for å dekke behovet statistikkprodusenten har for å kunne finne, forstå og bruke metodene som er nødvendig for å lage et produksjonsløp for sin statistikk.

4.3. Prinsipper for metodebibliotek

Dette er forslag til prinsipper som utgjør rammeverket til et metodebibliotek. Dette rammeverket skal gi føringer, støtte og verktøy i arbeidet med å holde "orden i eget hus". Målsetningen med prinsippene er å kunne øke evnen til samhandling på tvers av statistikkproduksjonene. Prinsippene gir generelle føringer for biblioteket, men sier ikke noe om hvilke metoder som skal ligge i metodebiblioteket. Samtidig bør

det tas hensyn til at de prinsippene som ikke blir implementert i første omgang kan bli utviklet i framtiden. Prinsippene sier ikke noe om hvordan selve metoden skal lages, for eksempel hvilke statistiske metoder som bør velges framfor andre.

Tabell 4.3 Prinsipper for metodebibliotek

Navn	Tilgjengelighet
Prinsipp	Alle metoder skal være tilgjengelig der de trengs.
Begrunnelse	For å holde orden på hva vi har av metoder bør vi putte dem i en eller flere samlinger, slik at vi kan forvalte dem.
Konsekvens	Man må kunne aksessere et system hvor metodene kan forvaltes og gjøres tilgjengelig for statistikkprodusenten når de er utviklet.
Navn	Identifikasjon
Prinsipp	Alle metoder skal ha en identifikasjon.
Begrunnelse	For å kunne forvalte metoder må de ha en unik id.
Konsekvens	I det administrative systemet må alle metoder få en identifikasjon, slik at det er mulig å skille metodene/stegene fra hverandre.
Navn	Navn på metode
Prinsipp	Alle metoder skal ha et meningsbærende navn.
Begrunnelse	For å kunne forstå hva metoden gjør må den ha et meningsbærende navn
Konsekvens	I det administrative systemet må alle metoder få et meningsbærende navn, slik at det er mulig å forstå hva metoden gjør.
Navn	Kort beskrivelse av metoden
Prinsipp	Alle metoder skal ha en kort beskrivelse av hva de gjør.
Begrunnelse	For å kunne forstå det meningsbærende navnet, må den ha en kort beskrivelse.
Konsekvens	I det administrative systemet må alle metoder få en kort forklaring, slik at det er mulig å forstå i litt mer detalj hva metoden gjør.
Navn	Uavhengige metoder
Prinsipp	Metodene skal være uavhengige komponenter.
Begrunnelse	For å kunne sette en metode inn i forskjellige prosessteg, slik som oppsett med metoder parallelt, sekvensielt eller innen grupper, må metoden være uavhengig av prosessmønstrene de inngår i.
Konsekvens	Metoden må tilgjengeliggjøres som en egen komponent uavhengig av de prosessmønstre metoden som oftest inngår i.
Navn	Kategorisering
Prinsipp	Alle metoder hører til minst en funksjon.
Begrunnelse	For å holde orden på metoder som kan løse samme behov på forskjellig måte må metodene tilhøre minst en forretningsfunksjon.
Konsekvens	I det administrative systemet må det tilrettelegges for at metoder blir kategorisert etter minst en forretningsfunksjon, slik at det er mulig for statistikkprodusenten å velge forskjellige metoder som løser en forretningsfunksjon.
Navn	Søkbar
Prinsipp	Alle metoder skal være søkbare.
Begrunnelse	For å kunne finne den riktige metoden på en effektiv måte må det være mulig å søke etter metoder ved hjelp av forskjellige kriterier.
Konsekvens	I det administrative systemet må det finnes mulighet for å søke etter metoder og legge inn "nøkkelord" og synonymer som ofte er brukt, slik at statistikkprodusenten enklere kan finne riktig metode.
Navn	Brukerdokumentasjon – innhold
Prinsipp	Alle metoder skal ha en beskrivelse av hva de gjør.
Begrunnelse	For å kunne forstå en metode er det viktig at det er en beskrivelse av hva metoden gjør.
Konsekvens	En forklarende brukerdokumentasjon som er lenket sammen med metoden må tilgjengeliggjøres.
Navn	Brukerdokumentasjon – kjøre
Prinsipp	Alle metoder skal ha en beskrivelse av hvordan de skal kjøres.
Begrunnelse	For å kunne kjøre en metode er det viktig å ha en instruks for hvordan den skal settes opp i systemet.
Konsekvens	En brukerdokumentasjon som er lenket sammen med metoden for hvordan de skal bli kjørt må tilgjengeliggjøres.
Navn	Parametre - metadata og verdiområdet

Prinsipp	Parametre til en metode skal ha: meningsbærende navn, kort forklaring og et verdiområde.
Begrunnelse	For å kunne forstå og sette opp parametre til en metode.
Konsekvens	Metadata om parametre og hvilke verdiområdet de har må være tilgjengelig.
Navn	Parametre - default verdi
Prinsipp	Parametre til en metode skal ha en default verdi.
Begrunnelse	For lettere å kunne finne riktig parameterverdi skal metoder ha defaultverdier.
Konsekvens	Default parametre må være tilgjengelig.
Navn	Metadata – prosessoutput
Prinsipp	Variabler som oppstår i en metode, skal ha metadata. Variabler som går gjennom en metode, skal bevare navn og metadata.
Begrunnelse	For å kunne forstå og bruke output fra en metode er metadata om variablene nødvendig.
Konsekvens	Metadata må kunne bevares, og metadata må kunne lages i de metodene de oppstår.
Navn	Prosesslogg
Prinsipp	Alle metoder skal ha en prosesslogg
Begrunnelse	For å kunne kontrollere kjøringen at metoden er blitt utført korrekt er prosessloggen nødvendig.
Konsekvens	Det må være mulig å se prosessloggen.
Navn	Feilmelding i prosesslogg
Prinsipp	Når metoden ikke fungerer, skal det komme en feilmelding i prosessloggen.
Begrunnelse	For å kunne finne feil er feilmelding i prosessloggen et godt hjelpemiddel.
Konsekvens	Det må være mulig å se en feilmelding i prosessloggen.
Navn	Prosessmåling
Prinsipp	Når metoden har kvalitetsmål skal dette komme ut som en prosessmåling i prosessoutputen.
Begrunnelse	Kvalitetsmål er viktig for å kunne vurdere data og metode.
Konsekvens	Det må legges til rette for prosessmåling som en output fra en metode.
Navn	Attributtkomponent
Prinsipp	Attributtkomponent skal bli laget for metoder fra funksjoner innen korrigere og kontrollere data, samt konfidensialitet.
Begrunnelse	Attributtkomponentene er en dokumentasjon av prosessen og det vil bli beregnet kvalitetsmål basert på dem.
Konsekvens	Det må legges til rette for å ha attributtkomponenter som output fra en metode, og disse må følge med data.

4.4. Forvaltningsprinsipper av metodebibliotek

For at metodebiblioteket skal fungere etter hensikten, må det fastsettes noen forvaltningsprinsipper.

Tabell 4.4 Prinsipper for forvaltning av metodebibliotek

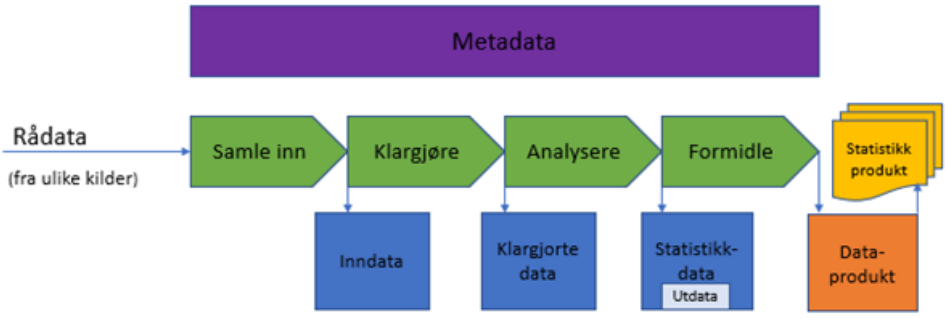
Navn	Ansvar for metodebiblioteket og innholdet
Prinsipp	Det må være et avklart forvaltningsansvar for det administrative systemet og innholdet i biblioteket.
Begrunnelse	For å kunne vedlikeholde og utvikle metodebiblioteket må det være avklart hvem som er ansvarlig for metodebiblioteket og for innholdet i det.
Konsekvens	Det administrative systemet må legge til rette for å differensiere hvem som har forvaltningsansvar for den enkelte metode som er implementert i biblioteket.
Navn	Versjonering
Prinsipp	Endringer av metoder skal dokumenteres og ved store endringer skal det lages nye versjoner.
Begrunnelse	Feil og forslag til forbedringer vil forekomme i metodene, og derfor må endringer bli håndtert av systemet.
Konsekvens	Det administrative system må ha dokumentasjon av programkode og kunne versjonere metoder som er lagt inn.
Navn	Kommunikasjon av endringer
Prinsipp	Når endringer av metoder får konsekvenser for statistikkprodusenten skal de bli varslet.
Begrunnelse	Når endringer av metoder fører til at produksjonsprosessen stopper, må statistikkprodusenten bli varslet slik at de kan få gjort nødvendige endringer slik at de kan produsere statistikken.
Konsekvens	Det må lages en mulighet til å varsle statikkprodusentene ved endringer av metoder
Navn	Oversikt innhold
Prinsipp	Innholdet i metodebiblioteket skal være kategorisert og søkbart.
Begrunnelse	For de som forvalter skal metodene være lette å finne ved at de er kategoriserte og søkbare.
Konsekvens	Det administrative systemet må legge til rette for at metoder er søkbare og kategoriserte.
Navn	Oversikt bruk
Prinsipp	Det skal lages statistikk over bruk av innholdet i metodebiblioteket
Begrunnelse	For å kunne planlegge kurs og utvikling av nye metoder er statistikk over metoder i bruk et godt grunnlag.
Konsekvens	Det administrative systemet bør kunne lage en enkel oversikt over antall metoder som er i bruk og hvilke statistikkprodukter som bruker dem.

5. Datatilstander

5.1. Bakgrunn for identifikasjon av hovedtilstander for data

I SSBs informasjonsforvaltning (Hustoft, 2018) er det foreslått å identifisere fem hovedtilstander for data og metadata i produksjonsprosessen som skal lagres: rådata, inndata, klargjorte data, statistikkdata og utdata, se figur 5.1. Hovedtilstandene er knyttet opp mot produksjonsfasene i virksomhetsmodellen. Inndata, klargjorte data, statistikkdata og utdata er output fra disse fasene. Utdata er delmengden av statistikkdata som skal bli publisert eksternt. Data bearbeides og endres gjennom hele produksjonsprosessen. En datatilstand er et datasett med metadata på et gitt punkt i produksjonsprosessen. Hovedtilstandene for data er de viktigste datatilstandene i produksjonsprosessen, og det er knyttet kvalitetskriterier til hver av disse hovedtilstandene. I produksjon av statistikk vil dataene endre seg fra innsamling til publisering og ha ulik grad av kvalitet underveis. Når data har oppnådd høy nok kvalitet skal de lagres og gjøres internt tilgjengelige for gjenbruk, det er da data har oppnådd en hovedtilstand. Identifiseringen og lagringen av disse tilstandene vil bidra til at produksjonsprosessene kan gjenskapes og kontrolleres når det er nødvendig, og gi grunnlag for lovbestemt ivaretakelse av arkivverdige data.

Figur 5.1 Hovedtilstander og virksomhetsmodellen (GSBPM)



5.2. Formål

Det er flere formål med å definere forskjellige datatilstander i statistikkproduksjonen som en helhetlig verdikjede for stegvis håndtering av informasjon. Datatilstander kan sees på som et konsept som skal bidra til:

- Å fastsette hvilke data som skal langtidslagres, dokumenteres og versjoneres.
- Å kunne gjenskape og kontrollere statistikkprodukter
- Å kunne ivareta lovbestemmelser om arkivering av data
- Å gjenbruke data
- Økt kvalitet i både data og prosesser i produksjonen
- Kunnskap om kvaliteten på data
- Åpenhet og etterrettelighet i bearbeiding av data
- Tydeliggjøring av eierskap, ansvar og formålet med data
- Felles forståelse av data som baserer seg på felles struktur og beskrivelser
- Informasjonssikkerhet. Tydeligere krav til dokumentasjon og lagring bidrar til bedre informasjonssikkerhet.

5.3. Fastsetting av hovedtilstandene

Hovedtilstander defineres relativt til den enkelte statistikkproduksjonen, og er data som går ut av en prosess (se figur 5.1).

Hovedtilstanden *rådata* er eksterne data med metadata som er tatt imot av SSB uten at det er gjort noe med dem. *Inndata* er rådata som er transformert over til en standardisert lesbar form som statistikkprodusentene kan ta i bruk.

I fasen ***samle inn*** skal det ikke bare samles inn data fra eksterne; en del av prosessen er også å sammenstille all nødvendig informasjon for en statistikkproduksjon. Det vil si sammenstille interne datakilder for den aktuelle statistikkproduksjonen. De fleste statistikker bruker mange datakilder i produksjon av statistikk.

Fellesbegrepet for all data som går inn i en klargjøringsprosess har vi valgt å kalle *Startdata*. Det viktige her er at *startdata* ikke er en hovedtilstand i seg selv men består av data med ulike hovedtilstander. Det kan være *inndata, klargjorte data* eller *statistikkdata*.

Klargjorte data vil være resultatet av klargjøringsprosesser som skal til for at statistikk skal kunne produseres med det kvalitetsnivået vi har behov for. *Klargjorte data* er et eller flere mellomprodukt som både utgår fra felles

klargjøring og den spesifikke klargjøringsprosessen. Når klargjorte data utgår fra en prosess for felles klargjøring, vil den utgjøre grunnlaget for en spesifikk klargjøringsprosess i etterkant. Klargjorte data fra den spesifikke klargjøringen utgjør grunnlaget for å lage Statistikkdata. Det er ingen skiller mellom hvilke funksjoner og underliggende prosesser som skaper klargjorte data i felles og spesifikk klargjøringsprosess. Skillet mellom felles klargjøring og spesifikk klargjøring har kun betydning ut fra behovet for å ta vare på ulike klargjorte datatilstander og organisering av klargjøringsarbeidet. I hvilken grad hovedtilstandene er egnet for deling er ikke knyttet til prosessene, men kvaliteten på datatilstanden.

Statistikkdata vil være resultatet av analyseprosessen av de klargjorte dataene.

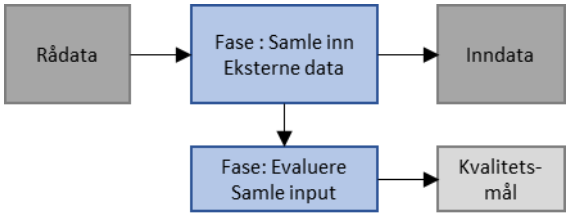
Utdata er de dataene vi sender ut av SSB. *Utdata* er en delmengde av *statistikkdata* som er sikret for ekstern bruk.

Alle prosesser har en kvalitet, og den kan beskrives med fase 8 i GSBPM; ***Evaluere,*** og da spesielt delprosess 8.1*: Samle input til evalueringen*. Vi har valgt å bruke begrepet kvalitetsmål, fordi det er det anbefalte begrepet på norsk fra Direktoratet for forvaltning og IKT (DIFI, 2017)

Inndata

Inndata er data som er mottatt av SSB av en ekstern kilde eller leverandør og transformert slik at de oppfyller SSBs krav til lagringsstruktur og har de nødvendige metadata.

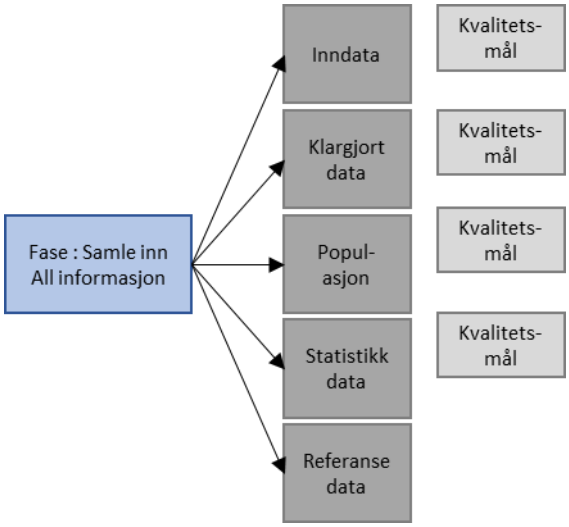
Figur 5.3.1 Inndata: output fra fasen Samle inn, eksterne kilder



Startdata

Startdata er en samling av alle data, hentet både internt og eksternt, som blir brukt i produksjon av en statistikk. Det er ikke en hovedtilstand i seg selv, men hvert datasett i startdatasamlingen har en hovedtilstand. Unntaket er referansedata, som er annen informasjon som blir brukt i statistikkproduksjonen, slik som årsrapporter og informasjon funnet på internett. Ifølge GSBPM er hovedmålet i fasen ***samle inn***, å samle eller hente inn all nødvendig informasjon.

Figur 5.3.2 Startdata: output fra fasen Samle inn, hente inn all nødvendig informasjon



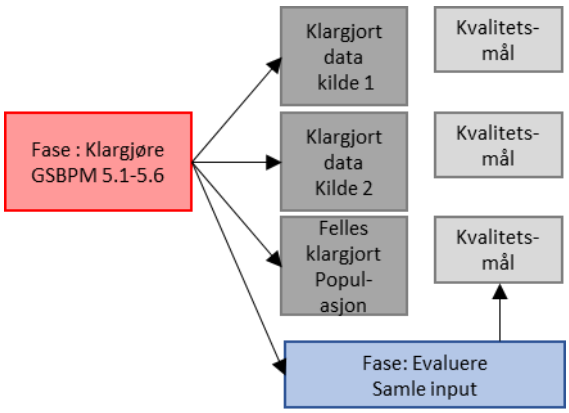
Klargjorte data

Klargjorte data er data som har gått gjennom en foredlingsprosess med kontroller, korrigeringer og andre relevante prosesser i klargjøring. Klassifiseringer og avledede variabler og enheter er en del av *klargjorte data* når det er grunnlaget for statistikkproduktet. For utvalgsdata skal *klargjorte data* inneholde vektene for oppblåsing mot populasjonen. For data som har blitt imputert, enten på grunn av partielt frafall eller enhetsfrafall, skal verdiene av imputeringen være en del av *klargjorte data*. *Klargjorte data* er et eller flere mellomprodukter av data som danner grunnlaget for å kunne produsere og beregne statistikkproduktene som skal publiseres. Delprosesser som er inkludert for å komme fram til klargjorte data er: 5.1-5.6 i GSBPM (5.1 *Integrere data*, 5.2 *Klassifisere og kode*, 5.3 *Kontrollere og validere*, 5.4 *Editere og imputere*, 5.5 *Avlede nye variabler og enheter*, 5.6 *Beregne vekter*). Kalkulering av aggregater for publisering er noe som inngår i produksjon av *statistikkdata*, men aggregering som funksjon er noe som både skjer i klargjøring og analyse.

Hva som defineres som *klargjorte data* må defineres ut fra produksjonsprosessen til produktene som skal skapes. Et enkelt statistikkprodukt vil kunne ha flere forskjellige datasett som hver for seg utgjør klargjorte datasett og til sammen utgjør de klargjorte dataene til produktet.
Statistikken er basert på mange datakilder:

- I de fleste statistikker er det en hovedkilde (skjema eller register). Øvrige kilder blir brukt enten til klargjøring av data fra hovedkilden, eller for påkobling av informasjon til statistikkformål, eller begge deler. Klargjøring av data fra hovedkilden skjer typisk ved at man kobler sammen data fra de ulike kildene, og basert på de sammenkoblede dataene kjører kontroll, beregning og korrigering på dataene fra hovedkilden.
- I noen tilfeller er statistikken basert på flere hovedkilder, da vil hver datakilde ha sine egne *inndata* som går gjennom en klargjøringsprosess før de blir satt sammen til ett datasett. Hver datakilde har sin egen datatilstand av *klargjorte data*. I tillegg vil det sammensatte datasettet være en annen tilstand av *klargjorte data*. Et datasett basert på flere datakilder må ofte gjennomgå en harmoniseringsprosess for å kunne bli brukt i statistikkproduksjon.
- I noen tilfeller vil datakildene som inngår i statistikken være *klargjorte data* fra andre statistikkprodukt. Det sammensatte datasettet blir definert som *klargjorte data* etter at det har gått gjennom en editeringsprosess for å harmonisere de forskjellige kildene.

Figur 5.3.3 Klargjorte data output fra fasen klargjøre, multikilder

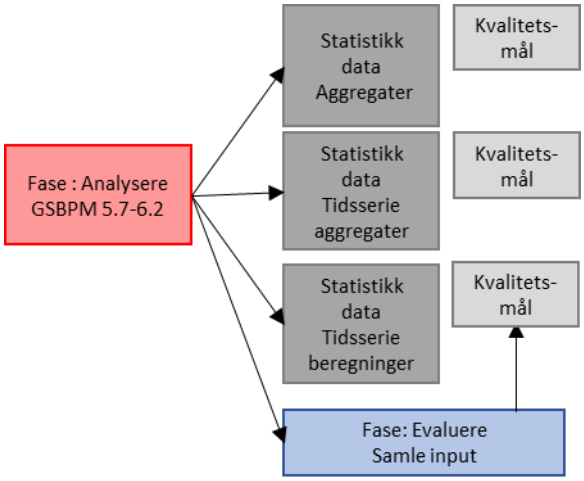


Fellesklargjorte data er data som flere skal bruke til å lage statistikk av, der det er en felles prosess med klargjøring av datagrunnlaget inntil et visst punkt. Hvilke enkeltprosesser som inngår i klargjøringsprosessen må avtales ut fra behovet til de fleste statistikkene. De mest dominerende prosessene i fellesklargjøring vil ofte være dataeditering og dataintegrering. For å sikre effektivitet og sammenlignbarhet mellom statistikker, er det viktig at så mye som mulig blir gjort i det *fellesklargjorte datasettet* og at de fleste bruker dette datasettet til å bygge den videre spesifikke statistikkprosessen på.

Statistikkdata

Statistikkdata er klargjorte data som er på publiseringsnivå eller har blitt aggregert slik at de er på nivået som skal publiseres. Beregninger slik som utledning av relative tall, indekser, indikatorer og sesongjustering er en del av statistikkdata. Integrasjon av andre datakilder og historiske data som skal til for å lage grunnlaget for beregningene, inngår i prosessen. *Statistikkdata* blir kontrollert, men korrigeringen av verdier skjer hovedsakelig i *klargjorte data*. *Statikkdata* er data som er klar til publisering, bortsett fra at datasettet ikke er sikret for ekstern bruk. Delprosesser som er inkludert for å komme fram til statistikkdata er: 5.7 *Beregne aggregater*, 5.8 *Ferdigstille datafiler*, 6.1. *Utarbeide produktutkast*, 6.2. *Kvalitetssikre produkter*

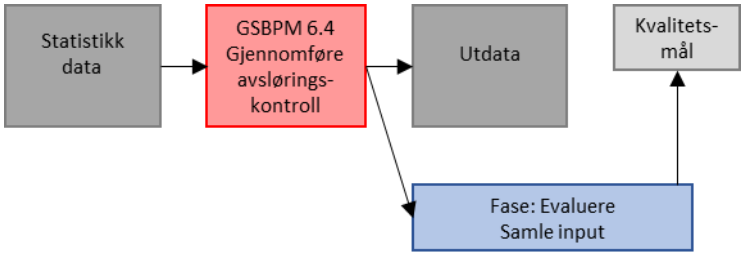
Figur 5.3.4 Statistikkdata output fra fasen analysere



Utdata

Utdata er en delmengde av *statistikkdata* som er sikret for ekstern bruk og klar til publisering. Det vil si at s*tatistikkdata* har gått gjennom delprosessen 6.4 *Gjennomføre avsløringskontroll*.

Figur 5.3.5 Utdata konfidensialitetssikret statistikkdata



5.4. Kvalitetskrav til hovedtilstandene av data

Data bør både være veldefinerte og stabile for å kunne bli gjenbrukt. Med stabilitet menes at kvaliteten på hovedtilstanden av data endrer seg lite (Struijs, Camstra, Renssen & Braaksma, 2013). I dagens statistikkproduksjon forbedres kvaliteten kontinuerlig, ofte uten en klar standard. Ved å ha kvalitetskrav basert på kvalitetsmål, kan alle lettere informeres om hvilken kvalitet data har. Kvalitetsmålene skal måle kvaliteten for hele datasett, men vil ofte være basert på status for hvert enkelt datapunkt. Hovedtilstandene kan ha forskjellige kvalitetsversjoner relatert for eksempel til foreløpige og endelige tall. Når data har god nok kvalitet kan hovedtilstanden låses, det vil si lagres og bli gjort tilgjengelig for andre. Alle hovedtilstandene kan være i endring samtidig. Det normale vil være å fastsette en versjon av *inndata* en liten periode før de andre hovedtilstandene. *Klargjorte data* og *statikkdata* vil som regel fastsettes i en versjon samtidig noen dager før publisering. De stabile datatilstandene har kvalitetsmål knyttet til seg som er avgjørende for om datatilstanden er av god nok kvalitet og kan brukes i publisering eller deles med andre. Kvalitetsmålene sier noe om både prosessen fram til datatilstanden og selve kvaliteten knyttet til dataene.

Kvalitetskrav til hovedtilstandene er en bearbeidet versjon av egenskapene ved datatilstander slik som beskrevet i «Statistics Netherlands Architecture; Business and Information Model» (Huigen, Bredero, Dekker & Renssen, 2009). I tabell 5.4 er kvalitetskravene listet opp etter hovedtilstandene. Kvalitetskravene er inndelt i temaer for lettere å holde oversikt. Temaene som er valgt er: deling, metadata, enhet, periode, sted og andre variabler. Temaet «deling» er kvalitetskrav som er viktige for å kunne dele data mellom statistikkproduksjoner. I dette temaet er det også inkludert generelle kvalitetskrav. I temaet «metadata» inngår kvalitetskrav som gjelder beskrivelser og informasjon om variabler, kodelister og enheter. Temaet «enhet» omhandler kvalitetskrav som gjelder om det er de riktige enhetene som er med i undersøkelsen. Temaet «periode» omhandler kvalitetskrav som gjelder referansetidspunkt for statistikken, er den klart og tydelige definert. Eller har den blitt samordnet, når statistikken består av forskjellige kilder. Kvalitetskrav som gjelder geografisk stedsfestning ligger under temaet «sted». Teamet «andre variabler» er kvalitetskrav som gjelder generelle variabler.

Tabell 5.4 Kvalitetskrav til hovedtilstandene

	Inndata	Klargjorte data	Statistikkdata	Utdata
Deling	-Tilstrekkelig metadata til å vurdere om data er hensiktsmessig å bruke til et gitt formål	-Kvalitetsmål som angir bearbeidingsgrad -Tilstrekkelig metadata til å vurdere om data er hensiktsmessig å bruke til et gitt formål	-Kvalitetsmål som angir bearbeidingsgrad -Tilstrekkelig metadata til å vurdere om data er hensiktsmessig å bruke til et gitt formål	
	-Kontroller opp mot avtale/kontrakt -Sjekker konsistens mot avtalte strukturer og medfølgende metadata fra kilden	-Samordnet datakilder og sikret mot feilkoblinger -Utvalgsdata har inkludert vekter som er mest mulig optimale -Minimum behandlingstid	-Minimum totalfeil -Minimum behandlingstid	-Minimum totalfeil -Minimum behandlingstid -Sjekket kriteriene for konfidensialitet
Metadata	-Data og metadata er lest, formatert og strukturert til SSBs standard, med færrest mulig feil -Alle koder er dokumentert med et tilhørende kodeverk	-Klare definisjoner av variable, koder og enheter	-Klare definisjoner av variabler, koder og aggregater	-Tilfredsstillende forklaring og begrunnelse -Klare definisjoner av variabler, koder og aggregater
Enhe	-Undersøkelses-populasjonen er så fullstendig som mulig	-Statistikk-populasjonen er så fullstendig som mulig	-Full dekning på aggregerte populasjonsnivå	-Full dekning på aggregerte populasjonsnivå
eriode	-Referansetidspunkt eller -periode er klart definert	-Referansetidspunkt eller -periode er klart definert eller har blitt samordnet	-Referansetidspunkt eller -periode er klart definert	-Referansetidspunkt eller -periode er klart definert
Sted	-Stedfesting eller den romlige dimensjonen er klart definert	-Stedfesting eller den romlige dimensjonen er klart definert iht. SSBs definisjonskrav	-Den romlige dimensjonen er klart definert iht. SSBs definisjonskrav	-Den romlige dimensjonen er klart definert iht. SSBs definisjonskrav
Andre variabler	-Kildeavhengige definisjoner av variabler	-Datasettet er mest mulig fullstendig i forhold til verdier -Ingen store inkonsistenser i variable fra hver kilde -Minimale målefeil i variabelverdier	-Minimalt antall inkonsistenser i variabler mellom tabeller og i tid	-Ingen inkonsistenser i variabler mellom tabeller og i tid

6. Kvalitet i prosesser og hovedtilstander

6.1. Innledning

Dette kapittelet gir en konkretisering av hvilke kvalitetsmål som kan beregnes for å kunne måle om kvalitetskravene til hovedtilstandene er oppfylt, se kapittel 4. For prosesser har vi valgt at hovedkilde for kvalitetsmålene skal være basert på: Quality Indicators for the Generic Statistical Business Process Model (GSBPM) - For Statistics derived from Surveys and Administrative Data Sources, version 2. (UNECE, 2017). I tillegg skal kvalitetsmål være i overensstemmelse med den norske standarden for beskrivelse av kvalitet på datasett som er beskrevet av «Direktoratet for forvaltning og IKT» (DIFI, 2017). Når det er mulig skal beregningsmetodene følge beskrivelsen som er gitt i ESS guidelines for the implementation of the ESS quality and performance indicators (EUROSTAT, 2014).

For å kunne måle kvalitet må kvalitetsmål fra alle prosesser samles inn. Kvalitetsmål kan være på svært detaljert nivå og må i slike tilfeller være input til beregning av en indikator på et høyere nivå som er mer anvendbar for en statistikkprodusent. Kvalitetsmål kan inneholde alt fra prosessdata tilknyttet den

enkelte funksjon til mål fra større prosesser som er basert på mange funksjoner til selvrapportering fra statistikkprodusent.

6.2. Avgrensing av kvalitetsmålene

Kvalitetsmålene som er foreslått er avgrenset til kvalitetsmål som kan beregnes fra eksisterende datasett i produksjonsløpet eller som output fra prosesser. Det vil si at de ikke inkluderer kvalitetsmål som vurderer om en variabel dekker det konseptet vi ønsker å måle eller kvalitetsmål som måler om enhetene tilhører målpopulasjonen som vi ønsker å måle. For å gjøre slike analyser kreves det en høy grad av faglig vurdering fra statistikkprodusenten og gode hjelpedata for å kunne beregne kvalitetsmål. Slike kvalitetsmål er viktige, men er vanskelig å automatisere og er derfor utelatt her.

Kvalitetsmålene som blir nevnt her skal gjelde for alle typer data; vi har derfor utelatt kvalitetsmål som bare gjelder utvalgsundersøkelser eller administrative data. Kvalitetsmål som gjelder bare en gruppe eller en enkelt statistikk kan bli laget som del av klargjøringsgrunnlaget. Slike spesialsydde kvalitetsmål bør kunne bli tatt med i tillegg til kvalitetsmålene som her er foreslått.

Det finnes svært mange kvalitetsmål og målene har mange dimensjoner (per variabel, per gruppe, totalt, antall og andel). Mange av disse detaljerte målene er nødvendige for statistikkprodusenten for å sikre kvaliteten, men er ikke like hensiktsmessige som kvalitetsmål knyttet til datatilstand. De detaljerte kvalitetsmålene kan bli laget i rapporter til statistikkprodusenten. Mens et utvalg av de viktigste målene blir valgt til å være kvalitetsmål tilknyttet datatilstanden.

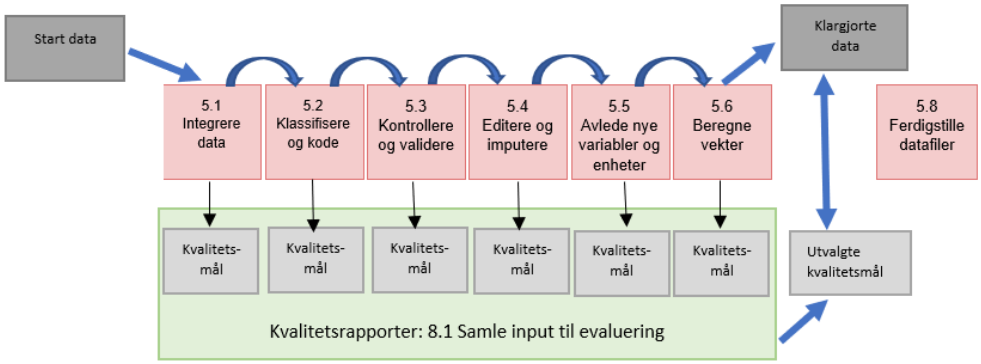
6.3. Prosesskvalitet – kvalitet tilknyttet prosesser

Mål på kvalitet bør framkomme som output av prosesser sammen med selve resultatet av prosessen. Kvalitetsmålene som er definert for virksomhetsmodellen GSBPM (UNECE, 2017) bør bli benyttet. I mange tilfeller er disse kvalitetsmålene abstrakte og må gjøres operasjonell for den aktuelle statistikkproduksjonen. I tillegg må kvalitetsmålene være nyttige for statistikkprodusenten, slik at de blir aktivt brukt. Kvalitetsmål som ikke blir brukt har liten verdi. Det bør derfor brukes tid på å finne gode kvalitetsmål for en statistikk.

Funksjoner i samme hovedkategori bør ha så likt kvalitetsmål som mulig slik at de kan sammenlignes for å kunne vurdere hvilken funksjon som er best. For eksempel bør kvalitetsmål fra nærmeste nabo-imputering kunne sammenlignes med kvalitetsmål fra modellbasert imputering.

Sammen med prosessen for å lage statistikk må en prosess som måler kvaliteten løpe parallelt, som en skyggeprosess. Det vil si at kvalitet har et eget prosessløp på lik linje med prosessløpet for statistikken.

Figur 6.3.1 Prosesskvalitet



6.4. Produktkvalitet – kvalitet tilknyttet datatilstander

Kvalitetsmålene knyttet til datatilstandene er samlet etter tema for hva de beskriver kvalitet på. Følgende temaer er valgt: deling, metadata, enheter, periode, sted og andre variabler. Noen typer variabler er viktige og det er derfor skilt ut som eget tema; dette er periode og sted. I mange tilfeller vil kvaliteten i klassifiseringsvariabelen, slik som næring og sektor, ha svært stor betydning for statistikken og kan bli skilt ut som eget tema.

Når det gjelder kvalitetsdimensjonene har vi valgt å følge "Retningslinjer for europeisk statistikk" (SSB, 2017) og ikke «Spesifikasjon for beskrivelse av kvalitet på datasett» fra DIFI (DIFI,2017). Hovedforskjellen er at dekning/kompletthet er tatt ut som en egen kvalitetsdimensjon fra dimensjonen «nøyaktighet og pålitelighet» i DIFI sin versjon. Det er i tillegg litt forskjellige ord som er valgt på kvalitetsdimensjonene. Ved utarbeidelse av kvalitetsmålene må det være å mulig å konvertere de til kvalitetsmål etter standard fra DIFI.

Foreslåtte kvalitetsmål er beskrevet i tabell 5.4.1 og 5.4.2. I første kolonne står navnet på kvalitetsmålet, deretter kommer en kolonne med en forklarende tekst til kvalitetsmålet. Kvalitetsdimensjonene, slik de er beskrevet i retningslinjer for europeisk statistikk, er i kolonne 3. Den neste kolonnen angir hvilke data eller datatilstander som går inn i beregningene. I noen tilfeller vil kvalitetsmålet være et direkte prosessmål fra en funksjon; da er det angitt hvilken delprosess kvalitetsmålet er output fra. Den siste kolonnen gir bemerkninger til kvalitetsmålet eller skisserer hvordan målet kan beregnes. Hvis beregningsmetoden fins i retningslinjer fra ESS (EUROSTAT, 2014), er dette angitt med ESS og kode for kvalitetsmålet som er brukt i retningslinjene.

Tabell 6.4.1 Kvalitetsmål for deling, metadata og enheter

	Kvalitetsmål	Forklaring	Kvalitets-dimensjon	Datakilder/ prosess	Bemerkninger og beregnings-metode
Deling	Gyldighetsrom datasett	Tidsrom data er gyldig for	Tilgjengelig og klarhet	Innfylling tekst	Tidsangivelse. Foreløpig, endelig eller reviderte tall
	Aktualitet	Graden av "ferskhet" av datasettet, for en spesifikk brukskontekst	Aktualitet og punktlighet	Inndata, klargjorte data, statistikkdata, utdata	Tid mellom opprettelse av datasett og referansetidspunkt for innholdet i datasettet ESS-QPI TP1:TP2
	Bruksformål datasett	Fritekstbeskrivelse av hva datasettet er opprettet/innsamlet for.	Tilgjengelighet og klarhet	Innfylling tekst	Ikke-kvantitativ/ fritekst-beskrivelse
	Egnethet datasett	Fritekstbeskrivelse av hva datasettet er, og ikke er, egnet til	Tilgjengelighet og klarhet	Innfylling tekst	Ikke-kvantitativ/ fritekst-beskrivelse
Metadata	Andel med variabel-beskrivelse	Andel variabler som har variabelbeskrivelse av totalt antall variabler	Tilgjengelighet og klarhet	Inndata, klargjort data, statistikkdata, utdata	ESS-QPI AC3
	Andel med kodeliste	Andel kategoriske variabler med kodeliste av totalt antall kategoriske variabler.	Tilgjengelighet og klarhet	Inndata, klargjort data, statistikkdata, utdata	ESS-QPI AC3
Enhet	Kompletthet av enhets identifikator	Andel enheter som har en enhets identifikator av totalt antall enheter	Nøyaktighet og pålitelighet	Inndata, klargjort data, statistikkdata, utdata	ESS-QPI A5
	Andel manglende enheter	Forholdet mellom antall enheter som mangler og antall enheter som skulle være med i datasettet.	Nøyaktighet og pålitelighet	Inndata mot undersøkelses-populasjon	ESS-QPI A4
	Andel dubletter	Andel enheter som har flere enn én forekomst av samme opplysning av totalt antall unike enheter	Nøyaktighet og pålitelighet	Inndata, klargjort data, statistikkdata, utdata	Antall dubletter/ totalt antall unike enheter
	Andel imputerte enheter	Andel imputerte enheter av totalt antall enheter i datasettet (klargjorte data)	Nøyaktighet og pålitelighet	Fase Klargjøre. Delprosess Avlede nye variabler og enheter.	ESS-QPI A7
	Koblingsrate	Andel av enheter som er koblet fra hvert datasett i en kobling	Nøyaktighet og pålitelighet	Fase Klargjøre Delprosess Integrere data Prosessmåling	ESS-QPI A3 Antall enheter koblet/antall enheter
	Andel enheter fjernet	Andel enheter fjernet av totalt antall enheter i klargjorte data	Nøyaktighet og pålitelighet	Klargjorte data mot Inndata	ESS-QPI A2 Antall enheter i inndata og ikke i klargjorte data/antall enheter i klargjorte data
	Andel enheter lagt til	Andel enheter lagt til av totalt antall enheter i klargjorte data	Nøyaktighet og pålitelighet	Klargjorte data mot Inndata	Antall enheter i klargjort data og ikke i inndata/antall enheter i klargjorte data

Tabell 6.4.2 Kvalitetsmål for periode, sted og andre variabler

		Kvalitetsmål	Forklaring	Kvalitets-dimensjon	Datakilder/ prosess for beregning	Bemerkninger eller beregnings-metode
Periode		Kompletthet av periodevariabel	Forholdet mellom antall enheter som ikke mangler verdi i periodevariabel og totalt antall enheter som er med i datasettet	Nøyaktighet og pålitelighet	Inndata, klargjortedata, statistikkdata, utdata	ESS-QPI A5 Antall ikke manglende verdier/ totalt antall enheter
Sted		Kompletthet av stedfestingsvariabler	Andel enheter som har stedfesting av totalt antall enheter	Nøyaktighet og pålitelighet	Inndata, klargjortedata, statistikkdata, utdata	ESS-QPI A5 Antall ikke manglende verdier/ totalt antall enheter
Andre variabler		Andel manglende verdier	Forholdet mellom antall verdier som mangler og antall enheter som skulle ha verdi, per variabel	Nøyaktighet og pålitelighet	Inndata, klargjortedata, statistikkdata, utdata	ESS-QPI A5 Antall manglende verdier/ totalt antall enheter
		Andel imputerte/korrigerte verdier	Forholdet mellom antall imputerte/korrigerte verdier og antall enheter som skulle ha verdier, per variabel	Nøyaktighet og pålitelighet	Fase Klargjøre, delprosess Editere og imputere Attributtkomponent	ESS-QPI A7 Antall imputerte verdier / totalt antall enheter
		Andel kontrollerte verdier	Forholdet mellom antall kontrollerte verdier og antall verdier	Nøyaktighet og pålitelighet	Fase Klargjøre, delprosess Kontrollere og validere Attributtkomponent	ESS-QPI A7 Antall kontrollerte verdier / totalt antall enheter
		Variasjonskoeffisient	Usikkerhet i aggregater på grunn av imputering/vekting	Nøyaktighet og pålitelighet	Fase Klargjøre, delprosess Editere og imputere. Funksjoner: modell- og donor-imputering Prosessmåling.	ESS-QPI A1 Andel usikkerheten utgjør av estimatet
		Revisjonsendring	Verdiendringen i prosent mellom versjoner av samme datasett	Nøyaktighet og pålitelighet	Fase Analysere, statistikkdata/utdata og tidligere versjon av de samme datasettene	
		Tidsserierevisjon	Gjennomsnittlig absolutt revisjoner (MAR) Relativ gjennomsnittlig absolutt revisjoner (RMAR) Gjennomsnittlig revisjon (MR)	Nøyaktighet og pålitelighet	Fase Analysere, Tidsseriedatasett og tidligere versjon av samme datasett Funksjon: Beregne/ Sesongjustere Prosessmåling	ESS-QPI A6

6.5. Tilrettelegging av statistikksystem for kvalitetsmål

Noen kvalitetsmål kan lages ut fra variabler som er målekomponenter i datasettet. Variablene som er målekomponenter er de variablene som har de observerte verdiene for en spesifikk enhet i et datasett. Når vi lager kvalitetsmål ut fra målekomponenter er det altså ikke er nødvendig med noe ekstra informasjon for å kunne lage kvalitetsmålet. I andre tilfeller må en sørge for at data som skal brukes for å finne kvalitetsmålet blir generert av prosessen. Dette kan skje på mange forskjellige måter avhengig av system og programmeringsspråk som blir benyttet. En måte er å sørge for at prosessmetoder har en prosessmåling som en output fra metoden. Det vil si at det kommer et kvalitetsmål direkte ut av en metode, i tillegg til resultat som skal brukes i statistikkproduksjonen. Hvis det blir brukt et programmeringsspråk der dette ikke er mulig, er det ofte mulig å lage en egen funksjon som beregner det samme kvalitetsmålet. En annen måte å fange data som er nødvendig for å lage kvalitetsmål, er å lage variabler som er attributtkomponenter som beskriver hva som er gjort med data. Det kan for eksempel være en dummyvariabel som forteller om en målevariabel er imputert eller ikke. Kvalitetsmål kan da lages som er funksjon av denne attributtkomponenten.

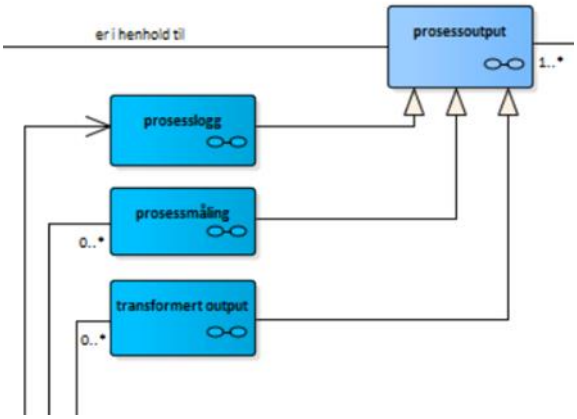
Det kan være smart å standardisere attributtkomponenter slik at en generisk funksjon kan benyttes for å lage kvalitetsmålet basert på attributtkomponenter. Her er forslag til hvilke attributtkomponenter som bør bli laget:

- Kontrollere
    - Statusvariabel som angir hvilke verdier som er markert som mulige feil som følge av kontroll av data. Den må knyttes til både enhet og variabel (målekomponent).
    - Statusvariabelen kan også angi hvilken instans av kontrollmetode som har markert en mulig feil.
- Korrigere
    - Statusvariabel som angir hvilke verdier som er korrigert. Den må knyttes til både enhet og variabel (målekomponent).
    - Statusvariabelen kan også angi hvilken instans av en kontrollmetode som har korrigert verdien. Viktig å skille manuell og automatisert prosessmetode.
- Validere - manuelt
    - Statusvariabel som angir hvilke verdier eller enheter som er: Ikke sjekket, kontrollert og godkjent, kontrollert og korrigert osv.
    - Statusvariabel kan også angi kilde for godkjenning/korrigering: statistikkprodusent, informasjonsleverandør, sekundærkilde (årsrapport, media, hjemmeside), osv.
    - Statusvariabel med tilleggsinformasjon om bakgrunnen for valget som er tatt, i form av en tekst.
- Konfidensialitet
    - Statusvariabel som angir hvilke verdier som skal undertrykkes, skille mellom primær og sekundær-undertrykking.

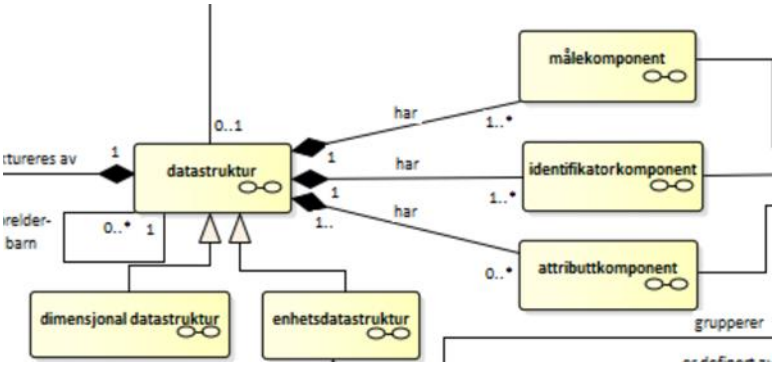
Kvalitetsmål er i de fleste tilfeller beskrivelser av kvalitet for totaler og aggregater i statistikken. Det vil si at det ikke beskriver kvaliteten til en enhet i datasettet. Det er derfor ikke hensiktsmessig å knytte kvalitetsmål som variabler til et datasett. I evaluering av en statistikk kan det bli laget kvalitetsmål som gjelder enheter; dette blir ofte brukt for å kunne sette inn tiltak for å heve kvaliteten hos de enheter og grupper av enheter som rapporterer mye feil. Kvalitetsmål vil også relatere seg til hvilken enhet beregningene er basert på. Det vil si at kvalitetsmålet vil relatere seg både til populasjon og variabler i en statistikk.

Ved hjelp av informasjonsmodellen er det mulig å beskrive generisk hvordan kvalitetsdata kan genereres i statistikkproduksjonen. Det viktigste er muligheten for å få beregnet prosessmålinger som en output fra en metode og lagre disse, se figur 6.5.1. Den andre måten er ved hjelp av å lage statusvariabler som er attributtkomponenter tilknyttet datasettet, se figur 6.5.2.

Figur 6.5.1 Kvalitetsmål som prosessmåling fra en prosessmetode, GSIM-forretningsgruppen



Figur 6.5.2 Variabler som er attributtkomponenter i et datasett, GSIM-strukturgruppen

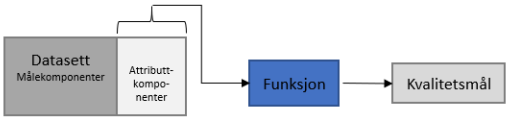


For kvalitetsmål som kompletthet og andelen dubletter kan kvalitetsmålet bli beregnet direkte fra datasettet, se figur 6.5.3.

Figur 6.5.3 Kvalitetsmål beregnet fra målekomponentene i ett datasett

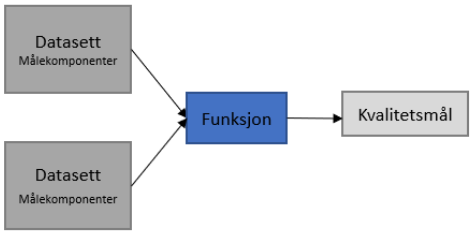
For kvalitetsmål som andelen imputerte eller kontrollerte verdier kan kvalitetsmålet lages ut fra statusvariabler som er attributtkomponenter i datasettet, se figur 6.5.4.a

Figur 6.5.4 Kvalitetsmål beregnet fra attributtkomponentene i ett datasett



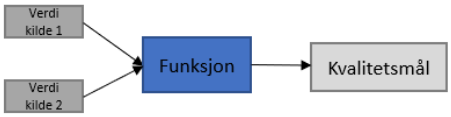
For kvalitetsmål som andelen enheter som er fjernet eller lagt til datasettet kan kvalitetsmål lages ut fra sammenligning av to datasett. For eksempel inndata og klargjorte data. Se figur 6.5.5 for illustrasjon.

Figur 6.5.5 Kvalitetsmål beregnet fra to datasett



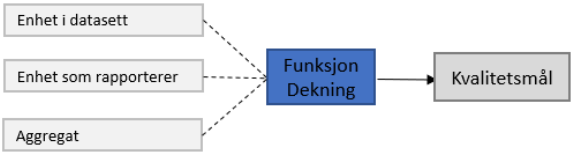
For kvalitetsmål slik som aktualitet kan kvalitetsmålet bli beregnet ut fra to verdier, for eksempel referanseperiode for statistikken og tidspunkt for publisering. Se figur 6.5.6 for illustrasjon

Figur 6.5.6 Kvalitetsmål beregnet fra to verdier



For kvalitetsmål som gjelder enhet er det viktig å angi enhetstypen som kvalitetsmålet gjelder for. I noen tilfeller vil vi ha en oversikt over hvilke enheter som finnes og som det skal skaffes informasjon om, det vil si at vi har en kjent undersøkelsespopulasjon. I mange tilfeller vil populasjonen av enhetene være ukjent på forhånd. Det vil ofte forekomme når en informasjonsleverandør rapporterer hendelser, slik som arbeidsforhold eller utbetaling av sosialhjelp.

Figur 6.5.7 Kvalitetsmål beregnet fra to verdier



7. Prosessmodell for modernisert produksjon

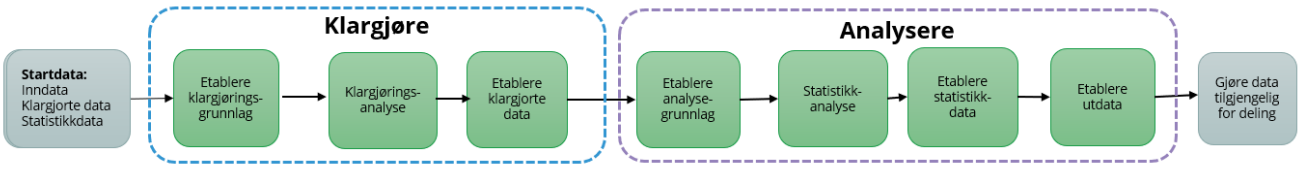
7.1. Innledning

Vi har laget en generell prosessmodell for klargjørings og analysefasene basert på den internasjonale virksomhetsmodellen: Generic Statistical Business Process Model (GSBPM) v5.1. Grunnen til dette, er at det i modernisering trengs en mer detaljert modell enn den generiske virksomhetsmodellen. GSBPM er en generell modell som beskriver statistikkproduksjonen på et høyt abstraksjonsnivå. Modellen som er laget for modernisering av statistikkproduksjonen tar hensyn til prinsippene for fasene klargjøre og analyse (kapittel 1). Et av de viktigste prinsippene er økt grad av automatisering av prosesser. Dette er ikke noe som framkommer som et viktig prinsipp av virksomhetsmodellen. Et annet prinsipp er at statistikk-produsenten skal kunne se sluttresultatet raskt i makroperspektiv, slik at de kan de prioritere å gjøre tiltak på det som har størst innflytelse. Innføring av mer automatisering og makroperspektiv i statistikkproduksjonen, vil føre til en effektivisering. I tillegg til dette vil det være et større fokus på kvalitet i statistikkproduksjonen, og at det skal styres etter kvalitetsmål. Deling av data og prosesser er økende og må tilrettelegges for i en modernisert statistikkproduksjon. Prosessmodellen for fasen klargjøre og analyse er beskrevet i dette kapitlet tar hensyn til disse prinsippene om automatisering, makroperspektiv, kvalitet og deling av data, noe ikke den generelle virksomhetsmodellen gjør.

Produksjonsprosessen skal fokusere mest mulig på kvaliteten på sluttproduktene, og prioritere kontroller av verdiene og prosessene som påvirker det mest. For å oppnå dette, skilles automatisering og menneskelig interaksjon tydelig i denne prosessmodellen. Først kjøres produksjonsprosessene automatisk. Alle prosesser skal ha kvalitetsmål, som samtidig skal være nyttige for produsenten av statistikk. Etter at den automatiske prosessen er kjørt, blir sluttproduktene og kvalitets-rapportene analysert av mennesker. Når det oppdages vesentlige feil og mangler i data eller prosesser, gjøres det tiltak for å rette dette, enten ved manuell redigering eller ved å bygge en ny delprosess. Deretter blir produksjonsprosessene kjørt på nytt. I modellen identifiseres hovedtilstander som er viktige og grunnlag for statistikkproduksjonen: det vil si klargjorte data, statistikkdata og utdata slik som beskrevet i kapittelet om datatilstander. Når prosessen har nådd høy nok kvalitet, lagres og dokumenteres alle data sammen med et utvalg av kvalitetsmål. Da kan data deles med interne og eksterne brukere av dataene. Med denne modellen har vi integrert kvalitet i alle prosesser og produkter som en måte å sikre god kvalitet på sluttproduktene.

Dette er ikke en revidert virksomhetsmodell, men en modell for modernisering som bygger på virksomhetsmodellen. Vi har foreslått prosesser under klargjøring og analyse som alle statistikker går gjennom. Disse prosessene er igjen delt opp i prosessteg. De prosessteg som alle statistikker går gjennom har fått en egen boks, mens de prosessteg som ikke alle bruker er samlet i en felles boks. Prosessene i fasene klargjøre og analysere består først i å etablere et datagrunnlag, deretter å analysere det og til slutt etablere hovedtilstander. Etablering av de ulike datagrunnlagene er en automatisert prosess. Resultatene fra denne automatiserte prosessen analyseres deretter. Analyseprosessen i klargjøring er å vurdere kvaliteten på datagrunnlaget, og hvis datagrunnlaget ikke har tilstrekkelig kvalitet, avdekke mulige feil og eventuelt korrigere disse. Dette er en prosess som krever høy grad av menneskelig interaksjon med statistikksystemet. Tabeller, lister og figurer av data blir vist til statistikkprodusenten som skal vurdere innholdet. Drilling i figurer og tabeller kan være effektive metoder for å avdekke feil. Feil kan manuelt korrigeres, eller det kan designes en prosessmetode som kan korrigere feilen som en del av den automatiserte prosessen. I etablering av hovedtilstandene for data velges de data som skal brukes videre i prosessen og som skal deles med andre.

Figur 7.1 Prosessmodell for modernisering på hovednivå

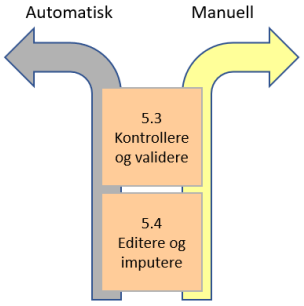


7.2. Etablere klargjøringsgrunnlag

Input til denne prosessen er *startdata* der all informasjon er samlet for å kunne kjøre klargjøring. Det inkluderer data, referansedata og metadata. *Klargjorte data* består av data som har gått gjennom de aktuelle delprosessene: 5.1 *Integrere data*, 5.2 *Klassifisere og kode*, 5.3 *Kontrollere og validere*, 5.4 *Editere og imputere*, 5.5 *Avlede nye variabler og enheter*, 5.6 *Beregne vekter*. Evaluering av disse delprosessene i form av kvalitetsrapporter inngår også i klargjøringsgrunnlaget. I denne prosessmodellen har vi delt disse delprosessene i to; det som kan kjøres automatisk er i prosessen ***etablere klargjøringsgrunnlag*** og det som krever menneskelig interaksjon er i prosessen ***klargjøringsanalyse***. Resultatet fra ***etablere klargjøringsgrunnlag*** vurderes i ***klargjøringsanalysen***. Dette kan være selve

resultatet av prosessen, logg for hvordan prosessen er kjørt, kvalitetsmål og kvalitetsrapporter fra evaluering. Det er spesielt delprosessene 5.3 *Kontrollere og validere* samt 5.4 *Editere og imputere* som generelt har krevd høy grad av menneskelig interaksjon. Det er derfor viktig å automatisere disse prosessene i så stor grad som mulig for å spare resurser. Automatiserte prosesser bør alltid kunne ettergås, for å evaluere prosessen.

Figur 7.2.1 Skille automatiske og manuelle prosesser



Delprosessen 5.3 *Kontrollere og validere* har vi splittet ytterligere i to prosesser: *Kjøre kontroller* og *Lage tabeller for kontrollformål*. Grunnen til dette var at i analyser vi har gjort av statistikkproduksjonen var dette to klart adskilte prosesser, og del av produksjonen av alle statistikker. Prosessteget *Kjøre kontroller* bruker funksjoner fra kategorien ***kontrollere*** med underkategoriene: *Logiske kontroller*, *Fordelingskontroller*, *Innflytelseskontroller og Enhetskontroller*. Mens *Lage kontrolltabeller* bruker funksjoner fra kategoriene ***strukturere*** og ***beregne***. Kontrolltabeller er ofte aggregater med beregning av endring fra forrige periode.

Våre analyser av statistikkproduksjonen viste at strukturering av data var funksjoner som ofte ble brukt. For enklere å kunne designe statistikkproduksjon ble derfor strukturering foreslått som eget prosessteg.

Etablere klargjøringsgrunnlaget inkluderer alltid følgende prosessteg:

- Kontrollere og validere (GSBPM 5.3) - Kjøre kontroller og lage kontrolltabeller
- Samle input til evaluering (GSBPM 8.1) - Lage kvalitetsrapporter

Etablere klargjøringsgrunnlaget kan også inkludere følgende prosessteg:

- Integrere data (GSBPM: 5.1) - Integrere
- Omforme datastruktur og variabler - Strukturere
- Klassifisere og kode (GSBPM: 5.2) - Klassifisere
- Editere and imputere (GSBPM: 5.4) - Autokorrigere
- Avlede nye variabler og enheter (GSBPM: 5.5) - Avlede
- Beregne vekter (GSBPM: 5.6) - Lage vekter
- Andre beregninger – Beregne
- Lage grafikk

7.3. Klargjøringsanalyse

I ***klargjøringsanalyse*** blir klargjøringsgrunnlaget vurdert manuelt av statistikkprodusenten. F.eks resultater fra kontroller som er kjørt og annen kvalitetsinformasjon dette må samtidig vurderes opp mot innflytelsen de har på statistikken. Det vil si at statistikkprodusenten hele tiden har et makroperspektiv i vurderingene som blir gjort. I tillegg må kvaliteten vurderes opp mot de kvalitetsmål som er satt som akseptable for statistikken. Grafikk er et nyttig hjelpemiddel for gjennomføring av denne prosessen.

Når det blir funnet feil og mangler i data har statistikkprodusenten behov for å finne årsaken til feilene. Det innebærer ofte å se på data i forskjellige prosessteg og på forskjellige aggregeringsnivåer. Det er også ofte nyttig å sjekke dataene opp mot referansedata for å finne forklaring eller mulig ny verdi; referansedata kan være andre datakilder, registre og informasjon fra internett. Denne prosessen har vi kalt *Inspisere*.

Når årsaken til feilen er funnet, må statistikkprodusenten ha mulighet til å sette inn tiltak for å rette opp feilen. Dette kan gjøres ved å manuelt korrigere verdier eller det kan designes og bygges en ny prosess som automatisk korrigerer dette når den blir kjørt. Denne delprosessen har vi kalt: *Tiltak*.

Klargjøringsanalyse inkluderer følgende prosessteg:

- Vurdere - analysere klargjøringsgrunnlaget for å vurdere om datakvaliteten møter de krav som er satt. (GSBPM: 5.3 Kontrollere og validere)
    - Se og vurdere kontrolltabeller
    - Se og vurdere kontroller som har slått ut
    - Se og vurdere output fra prosessteg
- Inspisere - avdekke årsak til utilstrekkelig kvalitet.
    - Sjekke data i ulike prosesser
    - Drille i tabeller og grafer
    - Sjekke data mot referansedata
- Tiltak - korrigere klargjøringsgrunnlaget for å øke kvaliteten.
    - Korrigere manuelt - korrigering direkte i datasettet (GSBPM: 5.4 Editere and imputere)
    - Gå tilbake til planleggingsfasen og bygge en automatisk korrigering. (GSBPM 2. Planlegge og 3. Bygge)

7.4. Etablere klargjorte data

Denne prosessen består i å bestemme hva som er *klargjorte data* og etablere *kvalitetsmål* tilknyttet hovedtilstanden. Dette er delprosess 5.8 *Ferdigstille datafiler* i GSBPM: 5.8.

Inkluderer følgende prosessteg:

- Velge klargjorte data - velge klargjorte data fra klargjøringsgrunnlaget
- Velge kvalitetsmål-velge kvalitetsmål som er generert av data og prosessen data har gått gjennom i klargjøring, og knytte valgte kvalitetsmål til klargjorte data.

7.5. Etablere analysegrunnlag

I prosessen ***etablere analysegrunnlag*** blir data tilrettelagt for å gjøre en analyse med formidlingsfokus. Dette er en automatisert prosess, deretter skal resultatet tolkes og forklares i prosessen ***statistikkanalyse***. Tilretteleggingen av data omfatter beregning av aggregater og laging av tabeller for analyseformål. I denne modellen tilsvarer det delprosess 5.7 *Beregne aggregater* og delprosess 6.1 *Utarbeide produktutkast* i GSBPM. Ofte i denne prosessen blir det laget tidsserier av beregnede aggregater og beregninger blir utført på disse tidsseriene. Prosessen kan også inkludere mange av prosessene som er under fasen ***klargjøre***, men nå på de aggregerte tallene. Det vil si 5.1 *Integrere data*, 5.*3 Kontrollere og validere*, 5.4 *Editere og imputere* og 5.5 *Avlede nye variabler og enheter* i GSBPM. Det er i denne delprosessen at indikatorer, indeksberegninger og sesongjusteringer blir utført.

Etablere analysegrunnlaget inkluderer følgende prosessteg:

- Beregne aggregater (GSBPM 5.7) og Utarbeide produktutkast (GSBPM 6.1) - Lage tabeller
- Samle input til evaluering (GSBPM 8.1) - Lage kvalitetsrapporter

Etablere analysegrunnlag kan inkludere følgende prosessteg:

- Integrere data (GSBPM: 5.1) - Integrere
- Omforme datastruktur og variabler - Strukturere
- Editere and imputere (GSBPM: 5.4) - Autokorrigere
- Kontrollere og validere (GSBPM: 5.3) – Kontrollere
- Avlede nye variabler og enheter (GSBPM: 5.5) - Avlede
- Lage indikatorer (GSBPM 6.1)
- Lage indekser (GSBPM 6.1)
- Sesongjustere (GSBPM 6.1)
- Beregne (GSBPM 6.1)
- Lage grafikk

7.6. Statistikkanalyse

Prosessen innebærer analyse av data med fokus på hva som skal formidles om statistikken. Prosessen *tolke og forklare* viser analysegrunnlaget i form av tabeller og grafer, for at statistikkprodusenten skal kunne etablere en god forståelse av hva dataene kan si og ikke si. Dette tilsvarer delprosessene *Kvalitetssikre produkter* og *Tolke og forklare produkter* (GSBPM: 6.2. og 6.3). Prosessen *teste hypoteser* er en mulighet til å mer fritt kunne analysere datagrunnlaget for å teste ut mulige teorier.

Inkluderer følgende prosessteg:

- Tolke og forklare- vurdere analysegrunnlaget
- Teste hypoteser - bruk av analyseverktøy for å teste og forstå data. Arbeid med data, grafikk og metoder for å forstå statistikk, underliggende sammenhenger og samfunnsutvikling.

7.7. Etablere statistikkdata

Etablere statistikkdata er en automatisert prosess der man velger hva som er statistikkdata og hvilke kvalitetsdata som skal knyttes til denne hovedtilstanden.

Inkluderer følgende prosessteg:

- Velge *statistikkdata* -velge hvilke data som er statistikkdata ut fra analysegrunnlaget
- Velge *kvalitetsdata* -velge kvalitetsdata som er generert av data og prosessen data har gått gjennom i tidligere prosesser og prosessteg. Knytte valgte kvalitetsdata til statistikkdata.

7.8. Etablere utdata

Etablere utdata er å velge hvilke data som skal publiseres etter at *statistikkdata* er sikret i henhold til konfidensialitet. Dette tilsvarer delprosess *Ferdigstille produkter* GSBPM: 6.5.

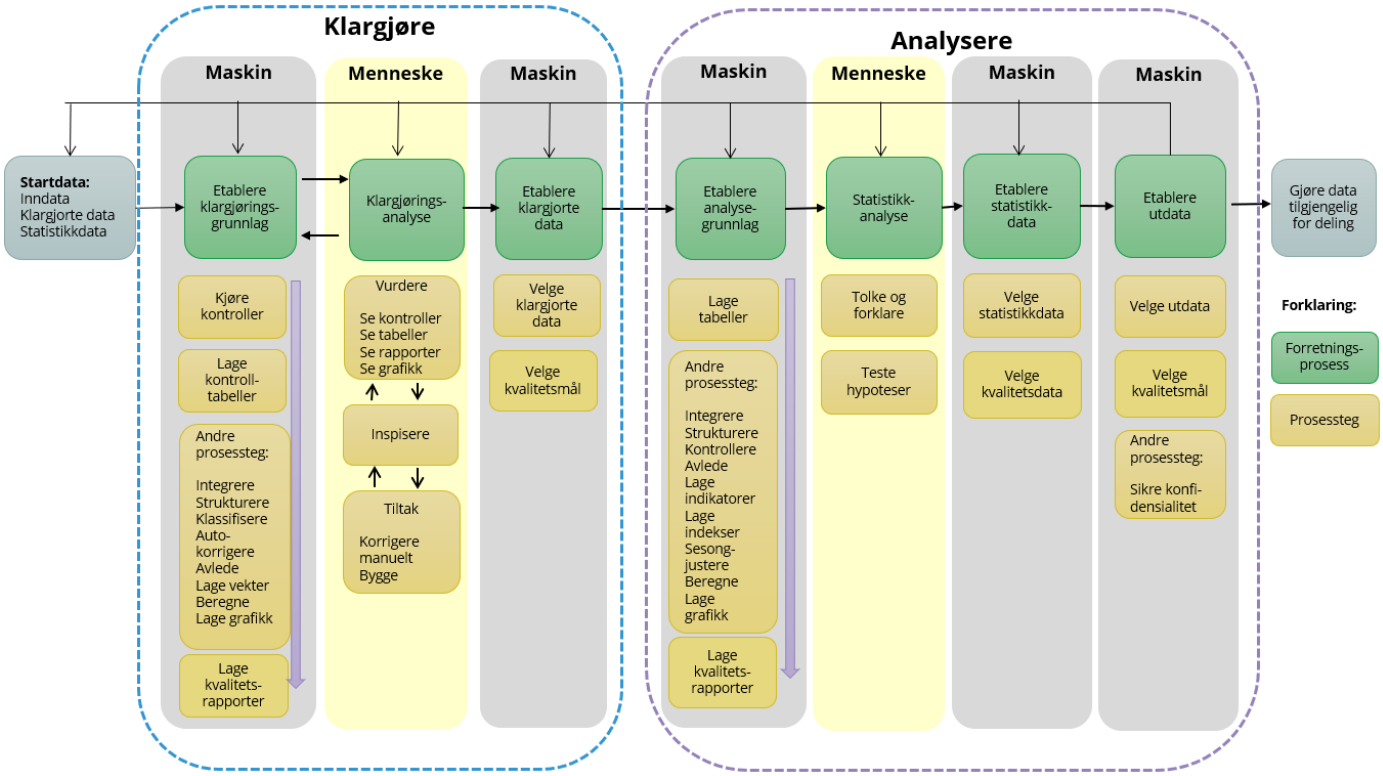
Inkluderer følgende prosessteg:

- Velge utdata - velge hvilke data som er utdata ut fra statistikkdata
- Velge kvalitetsmål-velge kvalitetsmål som er generert av data og prosessen data har gått gjennom i tidligere prosesser og prosessteg. Knytte valgte kvalitetsmål til utdata.

Kan inkludere følgende prosessteg:

- Sikre konfidensialitet

Figur 7.8.1 Prosessmodell for en modernisert statistikkproduksjon



Referanser

DIFI. (2017, september). Spesifikasjon for beskrivelse av kvalitet på datasett. Direktoratet for forvaltning og IKT. Versjon 1.0. Hentet 6. desember fra https://doc.difi.no/data/kvalitet-pa-datasett/

EUROSTAT. (2014). ESS GUIDELINES FOR THE IMPLEMENTATION OF THE ESS QUALITY AND PERFORMANCE INDICATORS (QPI). Hentet 6. Desember fra https://ec.europa.eu/eurostat/documents/64157/4373903/02-ESS-Quality-and-performance-Indicators-2014.pdf/5c996003-b770-4a7c-9c2f-bf733e6b1f31

EUROSTAT/SSB (2017, november). Retningslinjer for europeisk statistikk. Hentet 9.desember fra: https://www.ssb.no/omssb/lover-og-prinsipper/retningslinjer-for-europeisk-statistikk/_attachment/367184?_ts=166c47ec3f0

Huigen, R., Bredero, R., Dekker W. & Renssen, R. (2009). Statistics Netherlands Architecture; Business and Information Model. Discussion paper (09018). Statistics Netherlands. Hentet 4. desember fra https://www.cbs.nl/-/media/imported/documents/2009/12/2009-18-x10-pub.pdf

Hustoft. (2018, oktober). Informasjonsforvaltning i SSB. Internt dokument. Hentet 3. desember 2019 fra https://wiki.ssb.no/display/MAS2/Informasjonsforvaltningsnotat?preview=/96632857/115279486/Informasjonsforvaltning%20_h%C3%B8ringsokt.docx

Lovdata (2018). Lov om arkiv. Hentet 11. mars 2019 fra https://lovdata.no/dokument/NL/lov/1992-12-04-126#KAPITTEL_2

SSBs informasjonsmodell (SSB-IM), Spesifikasjon. Internt dokument i SSB. Hentet 3. desember 2019 fra https://wiki.ssb.no/display/VIR/SSB+IM+-+Spesifikasjon

Struijs, P., Camstra, A., Renssen, R. & Braaksma, B. (2013). Redesign of Statistics Production within an Architectural, Framework: The Dutch Experience. Journal of Official Statistics, Vol. 29, No. 1, 2013, pp. 49–71, DOI: 10.2478/jos-2013-0004 hentet 4. desember fra https://www.degruyter.com/downloadpdf/j/jos.2013.29.issue-1/jos-2013-0004/jos-2013-0004.pdf

UNECE. (2013, desember). Generic Statistical Information Model (GSIM): Specification Version 1.1. Hentet 3. desember 2019 fra https://statswiki.unece.org/download/attachments/97356610/GSIM%20Specification%201__1.docx?api=v2

UNECE. (2015, oktober). Generic Statistical Data Editing Models GSDEMs Version 1.0. Hentet 3. desember 2019 fra https://statswiki.unece.org/download/attachments/255492643/Generic%20Statistical%20Data%20Editing%20Models%20v1_0.pdf?version=1&modificationDate=1560240758846&api=v2

UNECE. (2017, oktober). Quality Indicators for the Generic Statistical Business Process Model (GSBPM) - For Statistics derived from Surveys and Administrative Data Sources, versjon 2. Hentet 6. Desember fra https://statswiki.unece.org/download/attachments/185794796/Quality%20Indicators%20for%20the%20GSBPM%20-%20For%20Statistics%20derived%20from%20Surveys%20and%20Administrative%20Data%20Sources_Final.pdf?api=v2

UNECE. (2019, januar). Generic Statistical Business Process Model GSBPM Version 5.1 Hentet 3. desember 2019 fra https://statswiki.unece.org/download/attachments/185794796/GSBPM%20v5_1.docx?version=1&modificationDate=1554283783707&api=v2