Statistics Norway

Statistisk sentralbyrå

Datarevisjon

Kontroll, gransking og retting av data

Anbefalt praksis

Forord

Kontroll og retting av data er viktig for kvalitetssikring av all statistikkproduksjon, økonomisk statistikk og personstatistikk, fra tellinger, utvalgsundersøkelser eller registre. Dette arbeidet kan være svært ressurskrevende, og mange bruker mye av sin tid på kontroll og retting av data - eller revisjon som det også benevnes. Revisjonsarbeidet må gjøres effektivt ved at automatiske rutiner erstatter manuelt arbeid der det er hensiktsmessig, mens manuelle kontroller prioriteres til viktige områder.

Dokumentasjon og analyse av revisjonsprosessen vil øke forståelsen av hvordan ulike kontroll- og retteprosedyrer påvirker statistikkdata og publiserte tall, og vil dermed være et godt grunnlag for å utvikle og teste ut revisjonssystemer. Håndboken er laget for å inspirere til kontinuerlig evaluering og forbedring av revisjonsopplegg gjennom å bedre kvalitet på inngående data og effektivisere revisjonsrutinene.

Formålet med håndboken er å gi en anbefalt praksis for revisjon. Den dekker revisjons-/ kontrollarbeidet gjennom prinsipper og metoder, og er ment for alle undersøkelser, tellinger, utvalg og registerbaserte person- og bedriftsundersøkelser. I forhold til håndboken fra 1998 omhandler den, bl.a. nye statistiske rutiner, grafiske kontrollmetoder og IT-revisjonssystemer. Håpet er at den kan bli et hjelpemiddel for alle som er involvert i kontroll/revisjon og statistikkproduksjon, både nye medarbeidere, revisjonsstaben, revisjonsansvarlige, statistikkansvarlige og ledere, og at alle gjør sitt til at gode rutiner blir synlige og tilgjengelig for andre.

Håndboken er utarbeidet av en prosjektgruppe i Seksjon for statistiske metoder og standarder med Anne Sofie Abrahamsen som prosjektleder og Jan Bjørnstad som styringsgruppe og ansvarlig forskningssjef. I tillegg til prosjektleder har prosjektgruppen bestått av Aslaug Hurlen Foss, Anna-Karin Mevik, Ane Seierstad, Arnfinn Schjalm og Anne Vedø. Prosjektgruppen har samarbeidet tett med flere fagseksjoner.

Statistisk sentralbyrå, 28. oktober 2005

Øystein Olsen

Innhold

1. Innledning

1.1. Formål med håndboken

Håndbok i datarevisjon inneholder anbefalte metoder til kontroll- og revisjonsarbeidet for kvalitetssikring av data i SSB. Den dekker arbeidet med gransking, kontroll og eventuell retting av data. Annen form for bearbeiding av data frem til publisert statistikk, som f.eks. imputering og estimering, behandles ikke her.

Formålet med håndboken er å vise metoder for effektiv kontroll, hvordan effekten av ulike kontroll/revisjonsmetoder kan analyseres og hvordan revisjonsprosessen kan dokumenteres. Dette er forutsetninger for å kunne utvikle og opprettholde effektive kontroll- og revisjonsprosesser som gir tilstrekkelig god kvalitet på data.

Håndboken kan brukes som en oppslagsbok om revisjonsmetoder, flagging og annen dokumentasjon. Den bør være sentral ved utvikling av nye revisjonsrutiner og med tanke på kontinuerlig vurdering og forbedring av datafangst og databearbeiding. Den kan fungere som en lærebok for medarbeidere uten erfaring fra kontroll og revisjon, mens erfarne medarbeidere kan finne metoder for å måle virkningen av eget arbeid.

1.2. Hva er revisjon?

Revisjon brukes her om kontroll av data, identifisering og behandling av feil og mistenkelige verdier (f.eks. ekstremverdier) i data som brukes som grunnlag for statistikkproduksjon. Hele produksjonsprosessen fra datainnsamling frem til publisering berøres eller påvirkes av revisjon. Data kan være fra utvalgsundersøkelser, totaltellinger eller fra ulike administrative registre.
Revisjon innebærer å

- avsløre og korrigere feil i data fra oppgavegiver/registereier
- identifisere problemområder i datainnsamlingen
- identifisere problemområder og måle kvaliteten på databearbeidingsprosessen

Problemområder kan identifiseres gjennom å måle effekten av revisjonsprosessen, f.eks. effekten av ulike kontrollmetoder.

1.3. Formålet med revisjon

Revisjon, kontroll og retting av data fra oppgavegivere er en viktig del av arbeidet for å sikre kvaliteten. Feil i grunnlagsdata kan påvirke resultatene slik at brukernes tillit til statistikkene skades.
Formålet med revisjon er å sikre tilstrekkelig god kvalitet på publisert statistikk gjennom å

- øke kvaliteten på datagrunnlaget for statistikkproduksjon
- forbedre rutiner for datainnsamling slik at det blir bedre kvalitet på inngående data
- bedre kvaliteten på administrative registre som brukes som datagrunnlag, gjennom samarbeid med registereier

1.4. Hvor mye skal korrigeres?

Det er umulig å sikre at alle data er korrekte. Selv ikke en nøyaktig gjennomgang av alle data fra skjema eller registre kan garantere gode data. Det er heller ikke nødvendig å kontrollere alt like nøye. Små feil kan ofte oppveie hverandre slik at de ikke påvirker totalen, og usikkerheten som skyldes utvalgsfeil og/eller populasjonsfeil har ofte større betydning enn små avvik i data. For mye revisjon kan til og med medføre nye feil i datasettet.

Kvaliteten måles på sluttproduktene, f.eks. på publiseringsnivå for statistikkene, eventuelt gjennom fokus på detaljert statistikk eller bruk av mikrodata. Marginale endringer fra videre revisjon sett i forhold til andre usikkerhetsfaktorer må være grunnlaget for vurdering av om revisjonen er tilstrekkelig for det enkelte statistikkproduktet.

1.5. Effektiv revisjon

1.5.1. Bruk av ressurser

Det brukes fortsatt mye ressurser på revisjon selv om utviklingen har gått mot maskinelle kontroller og begrenset, målrettet manuell behandling. Ressursene til revisjon styres mot problemområder eller viktige enheter, som kan variere avhengig av type statistikk. For eksempel vil noen få enheter innen økonomisk statistikk stå for det aller meste av endringer som følge av revisjon. Innsatsen styres da mot kontroll av store tall og/eller viktige enheter. Et viktig satsingsområde er avsløring av systematiske feil, det vil si at mange oppgavegivere gjør samme type feil. Det kan f.eks. forekomme at mange rapporterer etter andre standarder enn vi ber om.

1.5.2. Dokumentasjon

Dokumentasjon av revisjonsarbeidet har betydning for å sikre og dokumentere kvaliteten på det statistikkdatasettet revisjonen resulterer i. Slik dokumentasjon gjør det også mulig å bruke erfaringene fra revisjonen til å forbedre datafangsten og revisjonsarbeidet i senere undersøkelser.

1.5.3. Gode data inn gir best resultat

Kvaliteten på inngående data overvåkes gjennom revisjonsopplegget, og det søkes stadig forbedring av datainnsamlingen. Færre feil i inngående data kan redusere kostnader og produksjonstid og samtidig øke datakvaliteten. En viktig del av revisjonen blir å finne årsaken til feilrapportering og dermed gi bidrag til bedre datainnsamling.

Statistikk over rettinger under revisjonen kan avsløre problemområder for oppgavegiverne. Spørsmålsformulering og veiledning er midler vi selv har kontroll over i egne skjemabaserte undersøkelser. Prosessen for forbedring blir atskillig tyngre for administrative data hvor alle endringer må fremmes gjennom registereier. Desto viktigere er det å opprettholde nær dialog med registereier.

Utvikling innen informasjons- og kommunikasjonsteknologi gir nye muligheter innen effektivisering av datainnsamling og datarevisjon. Dette kan utnyttes til bedre samordning og samhandling med dataleverandører for å dekke krav til statistikken. Elektronisk datainnsamling gir muligheter til å flytte datarevisjonen tidligere i prosessen, før dataene overføres til statistikkbyrået. Oppgavegivere (inkludert eiere av administrative registre og intervjuere) kan vurdere sine data ved hjelp av kontrollrutiner som aktiveres før overføringen skjer.

1.5.4. Revisjonssystem

For at revisjonen skal bli effektiv må den integreres i en kontinuerlig oppfølging/forbedringsprosess for hele undersøkelsen. Det er sammenheng mellom statistisk revisjon og andre deler av produksjonsprosessen. Revisjonsarbeidet, automatisk revisjon og manuell oppfølging må planlegges

og styres systematisk. Statistiske metoder må rettes også mot revisjon. Mål for effekten av revisjon brukes som verktøy for planlegging og oppfølging av revisjonsprosessen. Sammenlignende resultater fra ulike revisjonssystem kan vise brukerne hvordan systemene virker.

1.6. De ti bud om datarevisjon

Forklaring: Ti punkter for effektive rutiner ved granskning, kontroll og endring av data

Bud 1: Formålet med revisjonsarbeidet må være klart og tydelig definert
Forklaring: Med klart formål blir det her ment: på hvilket nivå (totalt, undergruppe, enhet) skal data ha en god kvalitet og hvor stor feilmargin kan bli akseptert på dette nivå. Se kapittel 1.

Bud 2: Dokumenter revisjonsprosessen og merk endringer som blir gjort av data.
Forklaring: Rådata skal lagres. I tillegg bør revisjonsprosessen og endring av data bli dokumentert slik at det er mulig å evaluere revisjonsarbeidet. Se kapittel 2 og Håndbok i datalagring.

Bud 3: Sett deg godt inn i bakgrunnen til datasettet slik at du vet hvilke feil som kan forekomme og hva disse feilene kan skyldes.
Forklaring: Grunnen til feil fra oppgavegiver er ofte misforståelser eller manglende tilgang på data. Registerfeil og feil under databearbeiding fører også til feil i data. Feilene kan være åpenbare (absolutte eller logiske feil) eller sannsynlige (ekstremverdier eller innliggere) og tilfeldige eller systematiske. Se kapittel 3.

Bud 4: Bruk effektive kontrollmetoder.
Forklaring: Kontrollmetodene skal avdekke flest mulig verdier som er feil, men samtidig få med færrest mulig verdier som er korrekte. Se kapittel 4.

Bud 5: Visualisering av datasettet kan hjelpe til med å avdekke feil.
Forklaring: Ved bruk av grafikk kan det oppdages feil som er vanskelig å oppdage ved vanlige kontroller. Grafikk kan bli brukt til å sette gode grenser for automatiske kontroller. Se kapittel 5.

Bud 6: Rett bare til riktig verdi eller til en klart riktigere verdi.
Forklaring: Logiske feil må bli rettet, mens sannsynlige feil må bli vurdert. Kontakt med oppgavegiver kan ofte være nyttig for å finne den riktige verdien. Se kapittel 6.

Bud 7: Gjør deg kjent med grunnlaget for administrative data.
Forklaring: Formålet med innsamling av administrative data påvirker kvaliteten. Sørg for god kontakt med registereier. Se kapittel 7.

Bud 8: Utnytt IT-systemer effektivt ved gjenbruk.
Forklaring: Generelle IT-system for revisjon vil kunne gi gode løsninger for mange og samtidig spare utviklingsressurser. Se kapittel 8.

Bud 9: Revisjonsprosessen skal evalueres for kontinuerlig forbedring av data.
Forklaring: Evaluering av revisjonsprosessen gir grunnlag for å bedre kvaliteten på inngående data og effektivisering av revisjonsrutinene. For evalueringsmetoder, se kapittel 9

Bud 10: Revisjonssystemet skal være personuavhengig.
Forklaring: Erfaringer og kunnskap skal så langt som mulig integreres i revisjonssystemet slik at datakvaliteten ikke er avhengig av hvilken person som har revidert datasettet.

2. Dokumentasjon av data og revisjonsprosessen

For evaluering og kvalitetssikring av hele revisjonsopplegget samt det endelige datasettet er det nødvendig å dokumentere og lagre informasjon om data og revisjonsprosessen. Sentrale elementer er dokumenterte datafiler, revisjonsinstruks, spesifikasjon av dataprogram for maskinelle revisjonsrutiner og flagging (merking) av revisjonsprosessen på fil. Revisjonsprosessen skal også dokumenteres i "Om statistikken".

2.1. Dokumenterte datafiler

Både originale og ferdigreviderte data skal lagres. Det vil være betydelige forskjeller mellom de ulike statistikkene angående hvor stort avvik det er mellom originale data og reviderte data, både avvik i totaltallene og antall rettede dataelementer. Det må derfor vurderes i hvert tilfelle om både originale og reviderte data skal lagres fullt ut (stor grad av dobbeltlagring) eller om originale og reviderte data bare skal lagres når de er forskjellig; dvs. bare korrigerte felt. I kapittel 9 vises beregning av effekten av revisjon ved bruk av originale og ferdig reviderte datafiler.

For lagring av datafiler henvises det til Håndbok i datalagring på Unix i Statistisk sentralbyrå. I tillegg til å lagre datafilene, må det lagres metadata om filene.

2.2. Revisjonsinstruks

En revisjonsinstruks er først og fremst beregnet på internt bruk for saksbehandlerne. Det er en detaljert beskrivelse av revisjons- og kodingsprosessen. Instruksen må være på plass når revisjonen starter, men kan deretter være et "dynamisk dokument" som forbedres og suppleres etter hvert som kontrollørene/revisorene vinner erfaring. Revisjonsinstruksen skal være med på å gjøre revisjonen mindre personavhengig.

I instruksen skal det være detaljert spesifikasjon av alle typer kontroller som utføres. Det bør også gis veiledning om hvordan en feil verdi skal rettes, og hvordan det skal avgjøres om en mistenkelig verdi skal rettes eller ikke. I tillegg kan det være hensiktsmessig å ha med kilder for nyttig bakgrunnsinformasjon for kontrollør/revisor (f.eks. definisjoner og lover som gjelder den aktuelle statistikken).

> ***Eksempel fra småviltjakt***
> "Instruks for feilretting av småviltskjema 2002/2003" er et eksempel på en revisjonsinstruks. Den inneholder generell informasjon om skjemaet, populasjonen og svarfristen, om innskanning, skjermbilder og hvordan de fysiske skjemaene er oppbevart og sortert, samt kilder for nyttig bakgrunnsinformasjon for revisorene, nemlig om jakt og jakttider. Det gis presis veiledning om innfylling av manglende fylkesnummer, riktignok slik at revisorene i noen tilfeller er henvist til en egen vurdering, men da med et klart vurderingstema. Det samme gjelder godkjenning av angitt jaktutbytte, der både vurderingstema og holdepunkter for vurderingen er angitt. Hvor ryddig skjemaet er utfylt er nevnt som et holdepunkt for påliteligheten når andre vurderinger etterlater tvil.
>
> Visse kontroller for innskanningsfeil er omtalt i instruksen samt hvilke koder som er gyldige i visse felt.

2.3. Spesifikasjon av dataprogram

Dataprogrammene for de automatiserte delene av revisjonsprosessen definerer presist hva som skjer i disse trinnene av revisjonsprosessen, men slike programmer er ofte tunge og arbeidskrevende å sette seg inn i. Kravspesifikasjonene til programmene inneholder imidlertid den essensielle informasjonen på en mer brukervennlig måte, og må tas vare på som dokumentasjon.

2.4. Flagging

For å kunne dokumentere revisjonsarbeidet er det nyttig med flagging, dvs. merking av enheter eller variabelverdier som forteller hvordan dataene er vurdert og behandlet i revisjonsprosessen. Flaggingsdataene gir grunnlag for statistikk knyttet til variable, oppgavegivere eller revisjonsaktiviteter, og kan dermed identifisere variable, revisjonsaktiviteter eller (grupper av) oppgavegivere som krever nærmere oppmerksomhet. Flaggingsdata kan også brukes til å simulere resultatet av alternative (forenklede) revisjonsopplegg, og til å forbedre produksjonsprosessen over tid, f.eks. knyttet til systematisk kvalitetsarbeid.

Flaggingen kan skje i forbindelse med eller signalisere at:

- en dataprosedyre utføres
- en kontroll gir utslag
- en verdi identifiseres som feil eller mistenkelig
- ny informasjon innhentes
- en verdi endres
- en verdi bekreftes eller godkjennes (f.eks. utligger)

Variabelflagg gir supplerende informasjon til en variabelverdi ved å beskrive variabelverdiens status mer nøyaktig enn det fremgår av denne alene. For eksempel, ni situasjoner som kan skilles med et variabelflagg er følgende: (i) verdi er ikke innlest (for denne enheten), (ii) verdien er fra originaldata og ikke mistenkt (den ønskede situasjon), (iii) verdien er innlest som blank, (iv) verdien er innlest fra originaldata, men mistenkt å være feil, (v) verdien er innlest fra originaldata, har vært mistenkt, men er blitt bekreftet/akseptert/godkjent, (vi) feil verdi er innlest fra originaldata, (vii) verdien er rettet, ikke identisk med verdien i originaldata, (viii) verdien er satt til missing (verdien fra originaldata er fjernet), (ix) verdien er lest inn som blank, men skulle vært utfylt.

Flagging av videre behandling er beskrevet i kapittel 6.4.

Flaggingen må være integrert i revisjonsprosessen. Flagging må være mest mulig automatisert, slik at den ikke krever ekstra arbeid og slik at det ikke oppstår feil i flaggingen. Både feilkontroller og videre behandling, automatisk eller manuell, må flagges. Flagging av feilkontroller og behandling videre under revisjonen er vesentlig for å kunne evaluere revisjonsopplegg. Et flaggingsopplegg som gir informasjon om revisjonsprosessen bør som et minimum registrere og ta vare på to typer informasjon for hver enkelt observasjon i datasettet:

- hvilke kontroller som har gitt utslag.
- hvilke variable som er endret i løpet av revisjonsprosessen.

Det første punktet kan behandles ved å la flagging ha følgende verdier:

- kontroll ikke utført
- kontroll utført uten utslag
- kontroll utført med utslag.

Eksempler på type kontroller kan være:

- Har skjema kommet inn - enhetsfrafall
- Er skjema fullstendig utfylt - partielt frafall (markeres pr. variabel)
- Feil i rapporterte data?

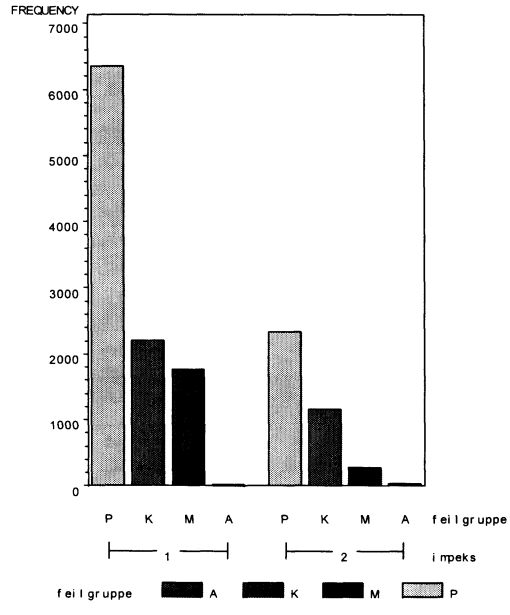
Eksempel på flagging av kontroller i utenrikshandelsstatistikken

Det kjøres maskinelle kontroller av alle varelinjer før dataene legges inn i revisjonsbasen. Hver kontroll har sin egen kode som legges til datafilen. Enkelte varelinjer kan dermed få flere feilkoder. Kodene for logiske og mulige feil har mange varianter (over 500) som kan deles i 4 grupper: A, K, M og P:

A - ugyldig variabelverdier - logiske feil, må rettes
K - kontroll av gitte land, varenummer eller høye verdier
M - maskekontroll - for gitte kombinasjoner (mange) av vare, land, organisasjonsnummer og/eller transport
P - ulike priskontroller

Alle varelinjer med feilkode går gjennom en manuell kontroll for eventuell oppretting. Figur 4.5.1 viser antall feilkoder fordelt på feiltype og import (impeks = 1) og eksport (impeks = 2) for oktober måned.

Figur 2.4.1. Antall feilkoder fordelt på import/eksport og feiltype



2.5. Om statistikken

"Om statistikken" er enkel dokumentasjon beregnet på brukerne. Her hører også omtale av revisjonen med. Den bør forklare i hovedtrekk på hvilke måter feil blir identifisert og rettet og gi brukerne et realistisk bilde av revisjonsinnsatsen og i hvilken grad feil kan tenkes å gjenstå.

2.6. Sammendrag av dokumentasjon

Data som skal bevares/lagres:

- Rådata
- Ferdigreviderte data
- Flaggingsdata

I tillegg skal revisjonsprosessen dokumenteres ved hjelp av

- statistikk over flaggingsdata
- revisjonsinstruks - detaljert beskrivelse av revisjonsprosessen
- spesifikasjon av dataprogram - dokumentasjon av de automatiserte revisjonsrutinene
- "Om statistikken" - enkel dokumentasjon beregnet på brukerne av statistikken

3. Feil i data

Det er flere grunner til at data inneholder feil som gjør revisjon nødvendig. Feil i data kan oppstå hos oppgavegiver eller registereier, eller under databearbeidingen.

3.1. Feil i data fra oppgavegiver

Feil i innkomne data fra oppgavegiver kan skyldes:

- Oppgavegiver misforstår spørsmålene
- Oppgavegiver har problem med å fremskaffe korrekte data
    - data er ikke tilgjengelig
    - avvikende definisjoner, f.eks. i oppgavegivers regnskapssystem
- Oppgavegiver rapporterer for feil statistisk enhet
    - feil i populasjonsavgrensning
    - data ikke tilgjengelig
    - selvseleksjon
- Manglende motivasjon fra oppgavegiver

3.1.1. Oppgavegiver misforstår spørsmålene

Utforming av skjema og veiledning har stor betydning for kvaliteten på inngående data. Data fra revisjonsprosessen er viktig informasjon for justering/videreutvikling av skjema, skjemavariable og veiledning, slik at oppgavegiver bedre skal forstå hva det blir spurt om og hva vi er ute etter å kartlegge.

Et eksempel på misforståelser er enhetsfeil, det vil si at man f.eks. gir data i hele kroner istedenfor i 1000 kroner når dette er angitt i skjemaet.

Misforståelser kan også skyldes ulik forståelse av begrep man bruker når man skal kommunisere med oppgavegiver om mulige feil. Dette kan gjelde også for tilsynelatende enkle begrep.

> ***Eksempel på ulik forståelse av begrepet "pris".***
> Fra seksjon for utenrikshandel sendte man forespørsel til tollvesenet angående mistenkelige verdier på pris, og mente enhetspris (pris pr. enhet). Etter uforståelige spørsmål/svar viste det seg at tollerne hadde oppfattet "pris" som ensbetydende med fakturaverdi, og denne verdien var det ikke noe mistenkelig ved.

3.1.2. Manglende motivasjon fra oppgavegiver

Det kan være krevende å motivere oppgavegiverne til å gi oppgaver av god kvalitet, men at skjema er tilpasset oppgavegiver og at formålet med undersøkelsen kommer klart fram, er forhold som bidrar til bedre motivasjon. Dette er ikke minst viktig når data i hovedsak samles inn av andre (register), og som inkluderer enkelte spørsmål til statistikkformål.

> ***Eksempel - motivasjon***
> Innsamlinger som skjer med "to-trinns-hjemler" av faglov og statistikklov (slik som data for kommunehelsetjenesten) kan virke motiverende for oppgavegiver, fordi deres daglige virksomhet er styrt av den samme fagloven (kommunehelseloven). Rapporteringen kan da anses som en integrert del av virksomheten og innebygd i de kontrollmekanismer man allerede har akseptert.

3.1.3. Avvikende definisjoner

Det forekommer at vi spør om kjennemerker som har definisjoner som kan avvike noe fra variable oppgavegiver bruker. Oppgavegiver kan da ha problemer med å svare nøyaktig på spørsmålet og velger heller (bevisst eller ubevisst) å bruke egne definisjoner direkte. Dette kan føre til skjevheter. Der er viktig å kjenne oppgavegivers egne data før man spør, og om nødvendig få kjennskap til forholdene via en særskilt kartlegging. Ofte kan det lønne seg å endre definisjonen etter oppgavegiverens data og heller foreta korreksjoner i det samlede materialet etterpå.

> ***Eksempel regnskap***
> Innen økonomisk statistikk må man være oppmerksom på definisjoner som brukes i foretakenes regnskap.

> ***Eksempel helsevern***
> Kommunens organisering av tjenester/ virksomheter som miljørettet helsevern, sosialt forebyggende arbeid og hjemmetjenester ytet fra eldreinstitusjon er noen av de organisatoriske "nøtter" som seksjon 330 har måttet finne statistiske løsninger på i årenes løp, og der ufullstendig kjennskap til organiseringen (og den enkelte etats forutsetninger for å gi data) representerte særlige utfordringer.

3.2. Feil i administrative data

Spesielt ved bruk av *administrative data* må vi tilpasse oss de definisjoner som ligger i de administrative registre. Vi kan likevel bruke vår innflytelse som er hjemlet i Statistikkloven om å påvirke innholdet i offentlige registre (§ 3-2) og undersøkelser som skal utføres av forvaltningsorgan (§ 3-3). SSB har også inngått en avtale med registereierne om å gjøre innholdet mest mulig egnet til statistisk bruk. Det er likevel under oppbygging av *nye* registre vi har størst mulighet i å lykkes med dette og derfor bør være mest aktive (se mer i kapittel 7.2).

> ***Eksempel***
> Et eksempel på avvikende definisjoner er sysselsetting. I Arbeidstakerregister, som er en av hovedkildene til registersysselsetning, er kriteriene for å bli registrert i registeret: Arbeidsforhold med 'minst fire timers gjennomsnittlig arbeidstid per uke', mens definisjonen i AKU (Arbeidskraftundersøkelsen) er 'minst en times arbeid i referanseuka'.

3.2.1. Feil i registeropplysninger

I registre som Det sentrale folkeregisteret (DSF) og Bedrifts- og foretaksregisteret (BOF) kan det forekomme feil og mangler. Et problem er at enheter som skulle vært i registrene ikke er der, eller at registrene har enheter som ikke skulle vært der. Ved utvalgsundersøkelser basert på disse registrene vil

disse problemene bli mer tydelige. I tillegg kan det forekomme feil i variabler tilknyttet enhetene, f.eks. adresse.

I DSF er det mange som ikke har riktige adresseopplysninger. Det kan skyldes at man har flyttet uten å sende flyttemelding, at det er mangler ved den registrerte bostedsadressen (f.eks. at bolignummeret mangler) eller at man faktisk bor på en annen adresse enn den formelle adressen i folkeregisteret. I den siste gruppen inngår studenter som nå har lov til å velge om de skal være folkeregistrert hos foreldrene eller på studiestedet.

Feil næringskode i BOF kan medføre at foretak får tilsendt skjematyper som ikke passer for virksomheten foretaket driver. I tillegg vil våre ulike undersøkelser ha behov for ulike statistiske enheter for samme konsern/foretak. Vår inndeling av foretaket i bedrifter kan være dårlig tilpasset foretakets interne organisering og det er heller ikke alltid at den fullstendige inndelingen går fram for foretaket. Dette medfører at vi får inn oppgaver fra andre enheter enn det som var forutsatt. Feilklassifisering kan føre til skjevheter som har like stor betydning som f.eks. utvalgsusikkerhet.

3.3. Statistisk enhet

Populasjoner er dynamiske. Korrekte oppdateringer av registre må til for å utelukke enheter som har besvart skjema eller finnes i administrative registre, men ikke tilhører populasjonen.

I personstatistikk er enheten den enkelte person eller husholdning. I næringsstatistikk er de vanligste enhetene foretak eller bedrift, men også bransjeenhet og lokal enhet brukes. Oppgavegiver kan komme til å svare for andre enheter enn de er bedt om, og det er ikke alltid klart hvilke enheter de svarer for.

Noen opplysninger, f.eks. finanskostnader, kan være lett tilgjengelig på foretaksnivå, men ikke på bedriftsnivå. Avvik mellom ønsket enhetsnivå og tilgjengelig enhetsnivå er en hyppig årsak til feil og bør unngås.

En annen enhet som stadig er blitt viktigere er adresse. Det gjelder spesielt for SSBs basisregistre BOF, GAB og BESYS. Kvaliteten på adresse er avgjørende for å kunne kople sammen nevnte basisregistre og for å utnytte registrene optimalt.

3.4. Feil fra databearbeiding

Feil som oppstår under databearbeidingsprosessen kan være:

- Feil oppstår under optisk lesing
- Feil i registeropplysninger - statistisk enhet
- Andre kodefeil og bearbeidingsfeil

3.4.1. Feil oppstår under optisk lesing

Ved optisk lesing av skjema kan det bli problemer med at enkelte sifre er vanskelig å skille fra hverandre, og at det er vanskelig å se komma ved bruk av desimaler hvor det forventes hele tall. Det finnes et opplegg for verifisering som retter opp det meste. Se mer om optisk lesing i kapittel 8.1.2.

3.4.2. Feil i registeropplysninger - statistisk enhet

Feil i registeropplysninger kan føre til både enhetsfeil i utvalg og populasjoner og gale verdier på registervariable som brukes til f.eks. oppblåsning.

3.4.3. Andre kodefeil og bearbeidingsfeil

Manuell inntasting av data gir mulighet for feil. Feil i data kan også følge av feil i programmer brukt ved elektronisk databearbeiding. Det forekommer også at retting under revisjon kan føre til nye eller større feil.

3.5. Typer feil

Feil i data kan deles inn i ulike kategorier:

- åpenbare feil og sannsynlige feil
- tilfeldige feil eller systematisk feil (målefeil)

Kontroll- og rettemetodene vil avhenge av kategori. Åpenbare feil er tydelige og må rettes opp, mens det for sannsynlige feil vil være en vurdering hvorvidt avvikene er av så stor betydning at de skal kontrolleres og eventuelt rettes. Systematiske feil er vanskeligere å oppdage enn tilfeldige feil, men svært viktige å avsløre/rette opp.

3.5.1. Åpenbare feil - logiske feil

Data som helt sikkert er gale, karakteriseres som åpenbare feil eller absolutte feil. Det kan være ugyldige verdier eller svikt i logiske sammenhenger. Feil i identifikasjonsvariable hører inn her. Feil skala, f.eks. "1 000 - feil" kan også være så tydelige at de anses som absolutte feil. Absolutte feil kan påvirke statistikken på en måte som gjør det åpenbart at det er feil i data.

Det er som regel lett å avsløre åpenbare feil, de skal kunne identifiseres ut fra skjema og registervariable for den aktuelle enhet. Det er imidlertid ikke alltid like klart hvilke variable som er gale, når det oppdages feil ved kombinasjoner av avhengige variable. Feilsøking og korreksjon av absolutte feil krever likevel sjelden de store ressursene.

Eksempler - åpenbare feil
(i) Eksport av varer i utenrikshandelsstatistikken aksepterer ikke Norge som bestemmelsesland.
(ii) Summen av alle underposter for inntekt er forskjellig fra total inntekt.
(iii) Negative verdier på kostnader som etter definisjonen skal være positive

3.5.2. Sannsynlige feil

Sannsynlige feil er verdier som ut fra gitte kriterier virker mistenkelige. Det er ikke alltid vi kan avgjøre om mistenkelige verdier er korrekte eller gale. Det er hovedsakelig to forhold som fører til mistanke om feil:

- Store avvik fra forventet verdi
    - ekstremverdi
    - manglende verdi (blank) eller null
- Systematisk feilrapportering

3.5.3. Avvik fra forventet verdi

Ekstremverdi
Store avvik fra forventet verdi avsløres ved sammenligning av verdier for de enkelte variable mot registervariable, tidligere rapporteringer eller samme variable for sammenlignbare enheter i samme undersøkelse. Store avvik vurderes enten som feil og kan rettes, eller de anses som korrekte verdier og klassifiseres som utliggere. Utliggere behandles særskilt ved imputering og estimering.

Innligger
Innliggere er observasjoner som er feil, men har verdier som kan se normale ut. De lar seg ikke avsløre like lett som utliggere. En vanlig måte for å avsløre innliggere er å vurdere flere variable fra samme

undersøkelse sammen, eller å vurdere mot registervariable som forventes å ha en sammenheng med den aktuelle variabelen, eller verdien til samme variable i en tidligere undersøkelse. Ved slike grep kan en innligger fremstå som ekstremverdi. Det kan oppstå problem hvis det er flere variable som defineres feil for samme enhet, sammenhengen mellom dem kan da synes rimelig uansett. Revisjon kan også utføres på mer detaljert nivå hvor en innligger blant alle enhetene kan skille seg ut som utligger blant en gruppe mer homogene enheter.

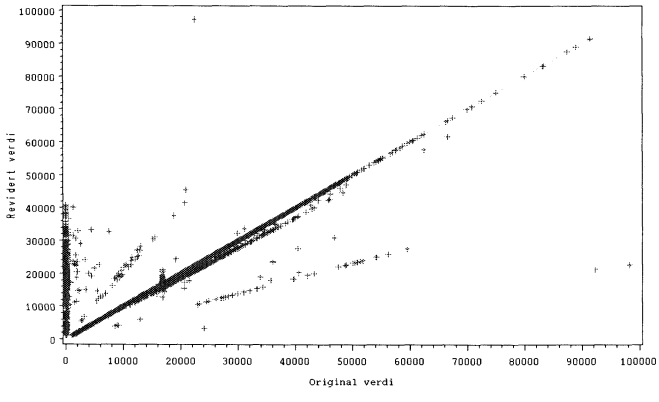
Manglende verdi eller null
Det er normalt at poster ikke fylles ut og at de dermed blir stående som blanke eller med f.eks. 0 som defaultverdi. I slike tilfeller er det ikke alltid lett å vite om verdien virkelig er 0 eller om det er manglende utfylling. (Ved flagging kan det markeres om verdien, her 0 eller blank, kommer fra rådata eller er satt inn senere i revisjonsprosessen. Dette løser imidlertid ikke et eventuelt problem med å avgjøre hvordan rådataverdien skal tolkes.)

3.5.4. Systematiske feil

Systematisk feilrapportering fra flere enheter kan skyldes definisjonsfeil, at oppgavegivere feiltolker spørsmålene eller tilpasser egne definisjoner. Dette er en type målefeil og vil ofte ikke være lett å avsløre da dataene ikke nødvendigvis skiller seg ut. Denne type feil vil ofte være innliggere. Denne type feil kan ha betydelig påvirkning på totalresultatet. Det er derfor viktig å avsløre systematiske feil.

Eksempel - endring i avtalt lønn
Plott av revidert mot original verdi av avtalt lønn viser klare mønstre. Figur 3.5.1 viser data fra 1998 hvor observasjoner med originalverdi over 100 000 eller missing er fjernet.

Figur 3.5.1. Sammenheng mellom original og revidert verdi av avtalt lønn



Observasjonene samler seg rundt den diagonale linja der urevidert og revidert verdi er like, og på den vertikale linja gjennom null. I tillegg ser vi flere linjer, blant annet en linje under 45-graderslinja, der verdiene blir omtrent halvert ved revisjon, og en linje over, der verdiene blir omtrent doblet. Mønsteret, hver linje, i dette plottet skyldes at flere observasjoner har samme type feil, som f.eks. at det er brukt feil periode ved rapportering av lønn.

Eksempel på misforståelse av begrep
Et eksempel fra utenrikshandelen er systematisk misforståelse av vektbegrepet i varekode for motorkjøretøy for last. Vektangivelse er gitt ved f.eks. "over 20 tonn" og gjelder egentlig lastekapasitet (last + sjåfør), men oppfattes ofte å gjelde vekten av kjøretøyet. Dette fører til bruk av galt varenummer.

3.6. Sammendrag - feilkilder og feiltyper

Feil i innkomne data kan oppstå

- hos oppgavegiver
  - oppgavegiver misforstår spørsmålene
  - data er vanskelig tilgjengelig
  - oppgavegiver er lite motivert
- hos registereier
  - manglende oppdatering av variable
- under databearbeiding
  - optisk lesing
  - koding og dataoverføring
  - beregninger

Det skilles gjerne mellom

- åpenbare feil - dvs. data som helt sikkert er gale
- sannsynlige feil - data som ut fra gitte kriterier virker mistenkelige

Systematiske feil er vanskelige å identifisere ved datarevisjon, men kan ha stor innflytelse på resultatene.

4. Kontrollmetoder

Tradisjonell revisjon består av maskinelle kontroller hvor feil og mistenkelige verdier markeres for manuell retting etter feillister. Ofte rettes det direkte i revisjonsbasen via en revisjonsapplikasjon. Det har utviklet seg muligheter for mange og omfattende kontroller via maskinelle rutiner. Kontrollene gjelder lovlige verdier og logiske sammenhenger. Det kontrolleres mot andre poster innen skjema, mot andre enheter i undersøkelsen, mot registervariable, sammenlignbare datasett og tidligere rapporterte data fra samme enhet. Kontrollene kan legges på enhetsnivå og på aggregerte data.

Et maskinelt feilidentifiseringssystem bør være en del av et større IT-system for klargjøring av data, identifisering av feil og eventuelt imputering av manglende verdier. Se kapittel 8 om IT-revisjonssystem. Systemet identifiserer (sannsynlige) feil i data ut fra et sett med rutiner og regler. Data som ikke blir identifisert som feil eller sannsynlige feil, blir ikke behandlet i revisjonen. Feil i mikrodata rettes automatisk eller etter manuell kontroll, eventuelt etter kontakt med oppgavegiver.

For sannsynlige feil er det viktig at kontroller og grenser for godkjennelse/manuell behandling tilpasses slik at viktige feil avsløres samtidig som videre kontroll av korrekte verdier og ubetydelige feil begrenses. Derved utvikles kontrollmetoder som sikrer tilstrekkelige gode data til lavest mulig kostnad. Revisjonssystemet bør være mest mulig personuavhengig.

4.1. Hvor og når

Revisjonen bør være en integrert del av databearbeidingsopplegget fra dataene kommer inn til statistikken er produsert. Feil bør i utgangspunktet bli rettet tidligst mulig i prosessen. Se mer om revisjon i forbindelse med elektroniske skjema og optisk lesing i kapittel 8.1.

4.1.1. Kontroller før data kommer inn

I elektroniske skjema er det nå mulig å legge inn kontroller som oppgavegiver får se når skjemaet blir utfylt. Oppgavegiver kan dermed selv korrigere sine svar før skjemaet blir sendt til SSB. I telefon- og besøksintervju er det intervjueren som får se kontrollene og som kan spørre oppgavegiver om svaret er korrekt. I administrative data er samarbeid med dataeier viktig Det kan være mulig å legge inn kontroller som kan rette opp feil før dataene blir sendt til SSB. Ved optisk lesing eller manuell punching av papirskjema finnes det ikke samme mulighet for at oppgavegiver kan korrigere svaret, men det kan bli lagt inn andre kontroller.

Kontroller før dataene sendes til SSB, bygger ofte på erfaringer fra revisjon av data. Det samme gjelder veiledning for utfylling av skjema Kvaliteten på skjema, veiledning og eventuelle kontroller må rutinemessig vurderes og eventuelt forbedres. Oppgavegivere bør ha mulighet for å ta kontakt med SSB for å sjekke uklarheter.

4.1.2. Revisjon av innkomne data

Revisjonsprosessen starter med kontroll av at skjema er utfylt i tilstrekkelig grad. Deretter blir dataene kontrollert for åpenbare og sannsynlige feil. Når åpenbare feil er rettet, blir det enklere å skille ut sannsynlige feil. Ofte kan man ha felles kontrollrutiner for avsløring av både åpenbare og sannsynlige feil. Etter retting av feil kan det foretas makrokontroller av aggregerte og estimerte verdier før godkjenning og publisering.

4.2. Ulike typer data

Data kan grupperes i forskjellige kategorier. Hvilken kategori de tilhører har betydning for hvordan de skal kontrolleres og rettes.

Identifikasjonsdata

Dette er variable som identifiserer enheten. Fødselsnummer og organisasjonsnummer er de mest vanlige identifikasjonsvariable i henholdsvis personstatistikk og foretaksstatistikk, men det vil også være andre variable som identifiserer enheten. Identifikasjonsvariable bør alltid være riktige.

Kategoriske data

Dette er variable som beskriver egenskaper ved en enhet. Egenskapene er kategorisert i klasser. Klassene kan ha numeriske verdier, men det har ingen mening å regne på disse verdiene, f.eks. fylkeskode. Kontroll av gyldige verdier vil hovedsakelig være absolutte kontroller. Kategoriske data kan også være omtalt som kvalitative data.

Numeriske data

Dette er variabler som uttrykker mengde eller kvantitet. Det er målbare variabler som f.eks. antall rom i et hus eller omsetning. Det gir mening å regne på numeriske variable, for å finne f.eks. gjennomsnittlig antall rom eller total omsetning i en bransje. Numeriske data kan også være omtalt som kvantitative data.

Kartdata

Kartdata omfatter i prinsippet alle stedfestede data, og beskriver i enkelhet egenskaper eller fenomener som kan lokaliseres eller avgrenses til punkt, linjer eller flater i forhold til et kjent geografisk rammeverk (et koordinatsystem).

Datoer
Datoer brukes på mange måter i statistikkproduksjonen, som å definere populasjoner eller generere nye variable. Datoer betinger egne kontroller, men kan også brukes i andre kontroller og automatisk revisjon.

4.3. Kontrollnivå

Det skilles tradisjonelt mellom kontroll av data på mikronivå hvor hver enkelt observasjon kontrolleres for seg selv, og makronivå hvor den enkelte enheten vurderes i forhold til hele datasettet (utvalget, tellingen, registeret). Utviklingen går mot at større deler av revisjonen utføres på makronivå. Kontroller på mikro- og makronivå kan supplere hverandre. Begge metoder kan brukes i samme undersøkelse, men vanligvis separat. Det er en klar fordel om datasettet er korrigert for absolutte feil (mikronivå) før det kontrolleres på makronivå.

4.3.1. Mikronivå

Dette innebærer at hver enhet blir kontrollert for seg, ofte også felt for felt. Enhetene blir sjekket for absolutte feil og mulige feil (vurderingskontroller). Det kan f.eks. være aktuelt å revidere på mikronivå ved detaljert statistikk på lavt nivå, når utvalget er lite eller ved førstegangsundersøkelser.

Kontroller på mikronivå er effektive for å rette opp absolutte feil.

Ulempen med mikrokontroller er at det brukes like mye tid på å kontrollere og rette små og store feil i datasettet siden det utføres samme type kontroller for alle enheter. Det er imidlertid mulig å differensiere mellom store og små enheter ved å bruke ulike avviksgrenser, både for prosentvise endringer og for absolutte endringer, avhengig av enhetens størrelse. For kvalitative verdier er det vanskeligere med slik differensiering.

4.3.2. Makronivå

Makrokontroll eller selektiv revisjon innebærer en utvelgelse av enheter til videre kontroll på mikronivå etter en samlet vurdering av enhetene, enten som kontroll på aggregert nivå eller ved andre kontroller som omfatter hele eller store deler av datasettet. Kriteriene for videre kontroll bygger på modeller (grafiske eller formelbaserte) for hvor viktige de enkelte avvik er for aggregerte estimater.

Kontroller på makronivå gir en god oversikt over datamaterialet og gir et godt grunnlag for å få frem mistenkelige verdier, men også absolutte feil i materialet kan avdekkes. For avklaring av tvilsomme forhold må enkeltenheter som forårsaker uventete resultater, identifiseres og rettes opp.

Et problem med makrokontroller er at enkelte makrokontroller krever at hele datamaterialet er tilgjengelig. Dette gjelder særlig metoder med aggregerte totaler, mens andre makrokontroller fungerer også om en stor andel av besvarelsene er kommet inn. Det er ikke uvanlig at oppgaver (skjema), også for viktige enheter, kommer inn til dels lenge etter fristen eller etter at revisjonsarbeidet har startet, f.eks. skjemabasert årsstatistikk. Ved makrokontroller kan feil på to eller flere enheter delvis oppveie hverandre og dermed ikke synes. Slike feil behøver ikke nødvendigvis ha betydning for statistikken, unntatt på mer detaljert nivå enn kontrollnivået.

4.4. Makrometoder - selektiv revisjon

Makrometoder for feilsøking er teknikker for å identifisere mistenkelige verdier i datamaterialet til videre kontroll av saksbehandler. De krever at hele datamaterialet er tilgjengelig maskinelt. Metodene søker å peke ut de avvikende verdiene som påvirker totaltallene mest. Dette sikrer oss at det ikke

brukes for mye tid på kontroll av mindre viktige avvik. Ulike makrometoder vil lett identifisere ekstremverdier (utliggere, outliers).

Det finnes flere metoder for feilsøking på makronivå:

- Aggregert metode
- Top-Down-metode (topp-ned, størst-først)
- Hidiroglou-Berthelot-metoden
- Grafisk revisjon

Selektiv revisjon innebærer at en konsentrerer revisjonsinnsatsen til feil som kan ha stor innvirkning på totalresultatene, dvs. på aggregert nivå. Et kontrollopplegg kan blinke ut slike mistenkelige verdier og prioritere enheter for nærmere undersøking. Kontroll og eventuell oppretting av de høyest prioriterte tilfellene kan gjøres ved en iterativ prosess med nye runder av kontroller og nye feilmeldinger etter hver runde med oppretting. Selektiv revisjon med konsekvent prioritering av viktige feilmeldinger er effektiv ressursbruk. Generelt for alle undersøkelser vil revisjonsinnsatsen være begrenset, og denne metoden sikrer oss at det er de minst viktige feilene som eventuelt ikke blir rettet når en må sette sluttstrek for kontroll av materialet. Selektiv revisjon egner seg også bra for publisering av foreløpige tall.

Det er viktig å være klar over at betydningen av mindre enheter vanligvis vil være forskjellig i en totaltelling og i en utvalgsundersøkelse. I en totaltelling vil en liten enhet bare representere seg selv og bety lite i totalbildet. I en undersøkelse med f.eks. én prosent utvalg, vil den tilsvarende enheten telle 100 ganger sin egen verdi. Feil i enheten vil derfor ha langt større konsekvenser enn i en totaltelling. Selektive revisjonsmetoder må derfor ta hensyn til vekt ved estimering. Hensikten med totaltellinger er imidlertid ofte å gi god statistikk på lavt geografisk nivå, og da vil små feil kunne medføre betydelige utslag.

Selektiv revisjon egner seg spesielt godt for undersøkelser der en har god bakgrunn for vurdering av mistenkelige feil, som f.eks. periodiske undersøkelser, mens det er mindre aktuelt ved f.eks. demografiske undersøkelser.

Selektive revisjonssystemer baseres på forutsetninger. Metodene må testes ut før innføring. Modeller og parametre må kontrolleres regelmessig. Det bør legges inn en kvalitetskontroll for å sikre kvalifisering av kontrollører og oppfølging av metoden.

Aktiv bruk av selektiv revisjon vil kunne begrense ressursbruken til revisjon. Frigitte ressurser må styres inn i andre deler av undersøkelsesprosessen og/eller analyser.

> ***Eksempel - Yrkeskoding***
> Arbeidsgiver skal rapportere yrke til Arbeidstakerregisteret. Seksjon for arbeidsmarked (S260) har nå utviklet kontrollmetoder hvor bedrifter med feil i yrkeskodingen forsøkes identifisert maskinelt, for dermed å få rettet opp større grupper av ansattes yrkeskoder. Hovedsakelig baserer kontrollmetodene seg på å sammenligne fordelingen av yrkeskodene for bedrifter med samme næring og omtrent samme størrelse. Ulike statistiske metoder brukes for å sammenligne frekvensfordelinger.

4.4.1. Aggregert metode

Aggregert revisjon er mer et overordnet prinsipp for feilsøking enn et konkret verktøy. Metoden går i korthet ut på å foreta feilsøking først på aggregert nivå. Aggregerte verdier som er mistenkelige ut fra valgte kontrollfunksjoner, flagges for videre kontroll av alle underliggende data. Dersom aggregerte verdier ikke gir grunn til mistanke om feil, aksepteres underliggende data direkte.

Aggregert metode er revisjon i to eller flere trinn:

- Feilsøking på aggregert nivå for å kartlegge mistenkelige (tabell)celler eller grupper av enheter
- Feilsøking på mindre aggregert nivå for videre kartlegging av mistenkelige (tabell)celler eller grupper av enheter. Gjentas inntil mikronivå.
- Feilsøking på mikronivå i de mistenkelige cellene.

Kontrollfunksjonene, både på aggregert og mikro nivå, må for utvalgsundersøkelser bygge på vektede verdier av variablene som er under revisjon. Kontrollene kan være manuell gjennomgang av variable, vektet og sortert, som faller utenfor valgte akseptgrenser. Eksempler på kontroller kan være differenser og forholdstall.

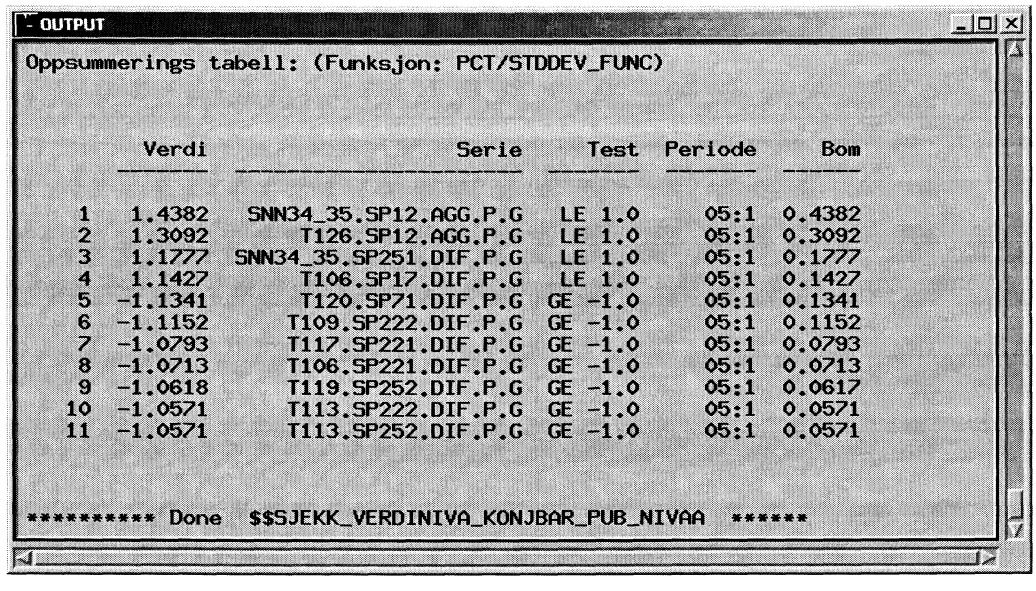
Eksempel
En vanlig form for aggregert makrokontroll er sammenligning før publisering, av beregnede tabeller for den aktuelle perioden med tabeller fra forrige periode.

4.4.2. Top-Down-metode

Dette er også et mer generelt prinsipp enn et konkret verktøy. Metoden går ut på å prioritere de viktigste enhetene fra toppen, f.eks. de største enhetene målt for en bestemt variabel (omsetning) eller de største positive og negative endringer (vektet) fra forrige periode. Alle data for enhetene vurderes, og ved korrigering beregnes nye aggregerte verdier på ønsket nivå. Prosessen fortsetter til feilrettingene er så små at de ikke påvirker resultatene.

Eksempel - Makrokontroll i FAME
For seksjon for økonomiske indikatorer (S240) er det laget en FAME-makrokontrollapplikasjon hvor brukeren selv kan definere en rekke makrokontroller. I dette eksemplet vises en makrokontroll for Konjunkturbarometeret for industrien, sesongjusterte tall. Makrokontrollen dekker 17 utvalgte spørsmål for 10 ulike publiseringsnivå (i alt 170 ulike tidsserier etter næring, spørsmål, serietype, utvalg/populasjon, og justeringstype). For hver serie blir beregnet verdi for siste periode sjekket. Bare seriene med de mest ekstreme endringer i forhold til endringer i tidligere perioder blir listet opp. Listen fra kontrollen kommer ut i et outputvindu, sortert etter avtagende absoluttverdi på testobservatoren som her er lik

$$\tau_t = endringsrate_t/std(endringsrate); \quad \text{hvor } endringsrate_i = \frac{x_i - x_{i-1}}{x_{i-1}} * 100 \text{ og } i=(t,...,t\text{-}40)$$


OUTPUT

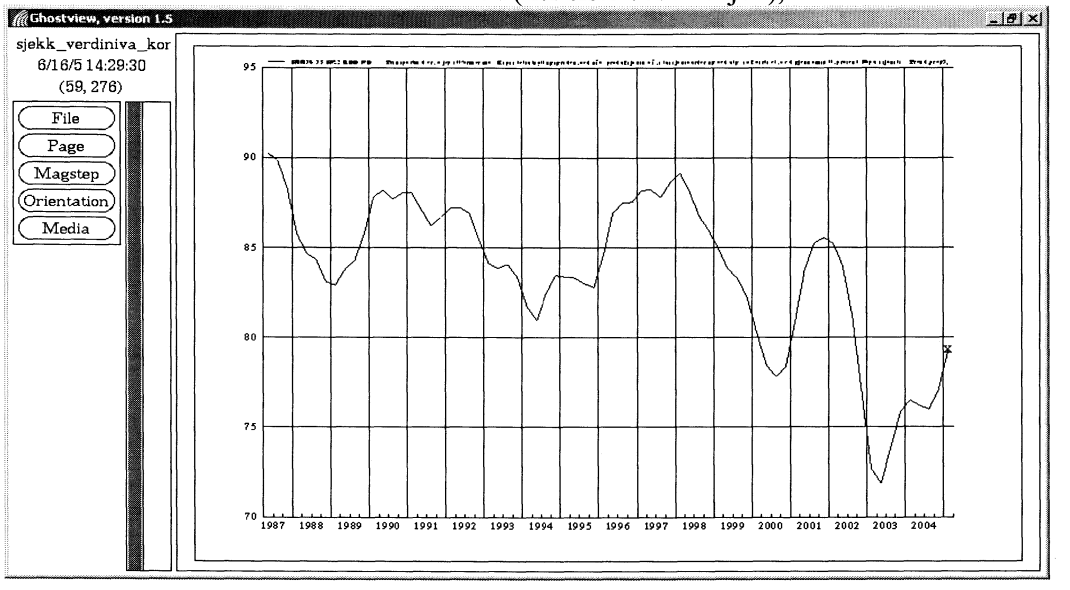
Oppsummerings tabell: (Funksjon: PCT/STDDEV_FUNC)

	Verdi	Serie	Test	Periode	Bom
1	1.4382	SNN34_35.SP12.AGG.P.G	LE 1.0	05:1	0.4382
2	1.3092	T126.SP12.AGG.P.G	LE 1.0	05:1	0.3092
3	1.1777	SNN34_35.SP251.DIF.P.G	LE 1.0	05:1	0.1777
4	1.1427	T106.SP17.DIF.P.G	LE 1.0	05:1	0.1427
5	-1.1341	T120.SP71.DIF.P.G	GE -1.0	05:1	0.1341
6	-1.1152	T109.SP222.DIF.P.G	GE -1.0	05:1	0.1152
7	-1.0793	T117.SP221.DIF.P.G	GE -1.0	05:1	0.0793
8	-1.0713	T106.SP221.DIF.P.G	GE -1.0	05:1	0.0713
9	-1.0618	T119.SP252.DIF.P.G	GE -1.0	05:1	0.0617
10	-1.0571	T113.SP222.DIF.P.G	GE -1.0	05:1	0.0571
11	-1.0571	T113.SP252.DIF.P.G	GE -1.0	05:1	0.0571

********** Done $$SJEKK_VERDINIVA_KONJBAR_PUB_NIVAA ******

Kolonnene i tabellen viser verdien på τ, identifikasjon av de seriene der det er funnet ekstreme endringsrater, grenseverdi som ble brutt, siste periode og avvik mellom "Verdi" og "Test". Det er plukket ut 11 serier der siste punkt i tidsserien faller utenfor valgte grenseverdier $-1 < \tau < 1$.

Den observasjonen som ble definert som mest ekstrem, er verdien i 1. kvartal 2005 i serien SNN34_35.SP12.AGG.P.G som gjelder Offshore-relatert virksomhet inkl. transportmiddelindustri, Kapasitetsutnyttingsgrad, Veid gjennomsnitt, Populasjonsestimat, Trendserie. I tillegg til tabellen får man frem grafene over de seriene som ble plukket ut i kontrollen. Verdien som slo ut i kontrollen (her siste observasjon), er markert med en X:



Med denne kombinasjonen av aggregert metode og top-down tilnærming vil man raskt kunne oppdage eventuelle feil i datamaterialet som gir ekstreme endringsrater i forhold til hva som er vanlig for indikatoren. Når slike ekstremer er identifisert, kan man enkelt finne frem til de mikrodata som forårsaker den ekstreme verdien. Eventuelle feil i oppblåsingsrutiner kan også identifiseres ved denne formen for makrokontroll.

Dette eksemplet viser kun en definert kontroll i FAME-makrokontrollapplikasjonen. Det er definert en rekke andre ekstremkontroller for andre statistikker ved seksjon 240, der det f.eks. er hensiktsmessig å benytte annen beregning av testobservatoren τ og dermed andre grenseverdier. Denne formen for makrokontroll benyttes både til revisjon av serier, for å få en objektiv oversikt over hva som er sentrale indikatorer, og for omtale i pressemeldingen.

4.4.3. Poengfunksjoner - selektiv revisjon av datasett med mange variable

Det er flere metoder som kan brukes ved utvelging av enheter til automatisk retting eller manuell kontroll når hver enhet har mange variable. Man kan velge ut enheter fra ett generelt størrelsesmål. Alternativt kan man beregne et mål for hvor viktig mulige feil fra en enhet er, og ved hjelp av vekting av avvikene for hver variabel beregne en poengfunksjon. Det finnes flere metoder for selektiv revisjon ved hjelp av poengfunksjoner.

I Canada har de utviklet en poengfunksjon DIFF for poengberegning av enheter (se Latouche og Berthelot, 1992). Enhetene velges ut til manuell kontroll på grunnlag av beregnede poeng, en sum som avhenger av endring, vekt, størrelse, viktighet og feilmarkering for hver variabel. (Se vedlegg 10.3) Poengfunksjonen er først og fremst utviklet for kontroller mot mulige feil i kontinuerlige variable. Men den kan også brukes for logiske feil, gitt at man har et fungerende system for automatisk imputering.

4.4.4. Kvartilmetode

Det er flere metoder som går under benevnelsen kvartilmetode. (Hidiroglou-Berthelot-metoden, som omtales i neste avsnitt, er ett eksempel). Dette er metoder for å velge ut enheter til videre kontroll av numeriske variable, og brukes ofte i korttidsstatistikk. Felles for metodene er at de tar utgangspunkt i robuste observatorer (median og kvartiler) for forholdet mellom to målinger. (Den ene målingen er av den variabelen som skal sjekkes, mens den andre kan være av samme variabel på et tidligere tidspunkt). Hvis dette forholdet er symmetrisk fordelt for enhetene som sjekkes, brukes kvartilene til forholdstallene til å lage et intervall av akseptable verdier. En enhet tas så ut til videre kontroll hvis dens forholdstall ligger utenfor intervallet. Et eks. på et slikt intervall er

$$\left[q_{0.5} - k \cdot (q_{0.75} - q_{0.25}) \ , \ q_{0.5} + k \cdot (q_{0.75} - q_{0.25})\right],$$

der $q_{0.5}$, $q_{0.25}$ og $q_{0.75}$ står for henholdsvis median, 1. og 3. kvartil (til forholdstallene). Parameteren k styrer bredden på intervallet, og må fastsettes ut fra erfaring.

I en situasjon hvor forholdet ikke er symmetrisk fordelt, må det transformeres for å oppnå bedre symmetri. Dermed er det kvartilene til de transformerte forholdstallene som benyttes til å lage et intervall av akseptable verdier. Et eks. på intervall for denne situasjonen er

$$\left[q_{0.5} - c \cdot d_L \ , \ q_{0.5} + c \cdot d_U\right],$$

der $d_L = \max(q_{0.5} - q_{0.25}, A)$ og $d_U = \max(q_{0.75} - q_{0.5}, A)$, og hvor $q_{0.5}$, $q_{0.25}$ og $q_{0.75}$ nå står for median, 1. og 3. kvartil til de transformerte forholdstallene. Parametrene c og A må fastsettes ut fra erfaring. (A brukes for å unngå vanskeligheter som ellers kunne oppstå når kvartilavstandene er veldig små).

4.4.5. Hidiroglou-Berthelot-metoden (HB-metoden)

HB-metoden er en statistisk feilsøkningsprosedyre basert på egenskapene til dataene. Metoden er utviklet ved Statistics Canada og er fullstendig maskinell når den først er programmert og testet. Den tar utgangspunkt i forholdet mellom to målinger, f.eks. måling av en variabel i to påfølgende perioder. Deretter transformeres dette forholdet to ganger, og grenser for godkjennelse eller videre kontroll beregnes ut fra median og kvartilavstand for den transformerte variabelen. Dermed påvirkes grensene for sannsynlige feil lite av utliggere samtidig som de følger naturlige svingninger i det opprinnelige forholdet. Formlene for transformasjonene er gitt i vedlegg 10.2.

HB - metoden innebærer bruk av 3 parametre **A** , **C** og **U** som settes på forhånd etter en analyse av feil og rettinger under revisjon.

Parameter	Forklaring
U	brukes for å ta hensyn til nivå på variablene. Hvis verdien for U økes, blir et mindre antall ekstremer relatert til lave verdier og flere relatert til høye verdier
C	brukes for å kontrollere bredden i konfidensintervallet. Hvis verdien for C økes, blir konfidensintervallet større og antall ekstremer som fanges mindre. Denne parameteren er meget viktig, og må settes på grunnlag av empiri
A	har virkning bare hvis medianen i de transformerte variablene er forskjellig fra null samtidig som avvikene mellom median og kvartiler for de transformerte variablene er svært små. Det kan ellers føre til altfor mange markeringer for utliggere. A-faktoren er i det kanadiske opplegget anbefalt satt til 0.05. Større verdi for A medfører et mindre antall ekstremer.

I tillegg til parametrene er stratifiseringen også et meget viktig moment for å optimalisere egenskapene ved HB-metoden. Homogene strata må velges samtidig som antall observasjoner som inngår i hvert strata er stort nok for å beregne robuste referanser (medianen, kvartiler).

Metoden har flere fordeler. Kontrollen av dataene er maskinell. Det settes ikke konstante grenser for hva som aksepteres eller ikke, men egenskapene ved fordelingen til dataene utnyttes. Ved hjelp av parametere kan man begrense antall observasjoner som må undersøkes og styre hvor stor vekt som skal legges på de store enhetene i forhold til de små. Metoden vil imidlertid feile dersom median og en av kvartilene er identiske.

I Statistisk sentralbyrå er metoden blant annet brukt innen Produksjonsindeksen og Ordrestatistikken for industrien, samt husleieundersøkelsen i Konsumprisindeksen. Den er også benyttet i Abrahamsen og Seierstad (2004). Metoden er spesielt egnet til periodiske undersøkelser, men kan også brukes på tverrsnittsdata og på tvers av undersøkelser. I periodiske undersøkelser kan man f.eks. se på forholdet til samme variabel foregående eller tilsvarende periode foregående år. I tverrsnittsdata, f.eks. priser i ulike områder av landet, kan man beregne forholdet relativt til priser i ett gitt område.

Etter beregninger og aggregeringer vil det foretas makrokontroller på seriene, noe som kan utløse ytterligere kontroller og feilretting på mikronivå.

Eksempel på bruk av Hidiroglou-Berthelot-metoden

Seksjon for økonomiske indikatorer (S240) bruker Hidiroglou-Berthelot-metoden som hovedverktøy ved feilidentifisering. For noen statistikker kombineres HB-metoden med andre metoder (Topp-Down, Tjebysjefs ulikhetsmetode), og i enkelte tilfeller brukes HB-metoden oppdatert i forhold til egenskapene ved datamaterialet. HB-metoden brukes ved feilidentifisering av alle indikatorer bortsett fra den kvartalsvise investeringsstatistikken. Den brukes som hovedmetode for de månedlige PI (Produksjonsindeksen), PPI (Produsentprisindeks) og KPI. Vi kan si at metoden fungerer optimalt etter at den blir tilpasset til hver enkelt statistikk. Når det gjelder kvartalsvis statistikk er det bare KOLS (Kvartalsvis ordre- og lagerstatistikk) som bruker den. Egenskapene ved data for KIS (Kvartalsvis investeringsstatistikk) og KBAR (Konjunkturbarometer) gjør at det brukes andre metoder for disse to statistikkene.

Felles for S240 er at data hentes regelmessig (månedlig eller kvartalsvis) fra et fast utvalg (rulleres en gang i året). Dette gjør det mulig å benytte kvotene som referanser for å identifisere ekstremer, og det er dette som er kjernen i HB-metoden. La oss beskrive hvordan HB-metoden brukes i KPI:

Den månedlige konsumprisindeksen beregnes ved å innhente ca. 40 000 priser hver måned. Disse er prisene for ca. 1 000 typer varer som rapporteres fra ca. 2 000 butikker. For å identifisere observasjoner som skal kontrolleres, beregnes forholdet mellom prisen som ønskes kontrollert i periode m og den som allerede er kontrollert i periode m-1, $R_i = p_{i,m} / p_{i,m-1}$.

Kvotene R_i transformeres slik at både oppgangen og nedgangen i priser behandles på samme måte. Prisendringene kobles med prisnivå gjennom en ny transformasjon og en parameter $\boldsymbol{U}$. Ved hjelp av parameter $\boldsymbol{U}$ kan vi styre i hvor stor grad vi skal fokusere på prisnivå når vi skal behandle prisendringene.

Transformerte prisvariable stratifiseres etter konsumgrupper, og for hvert stratum beregnes 1. kvartil, medianen og 3. kvartil av den transformerte variable. Verdien for medianen og kvartilene brukes for å bestemme hvilke observasjoner som skal behandles som ekstremer. Alle

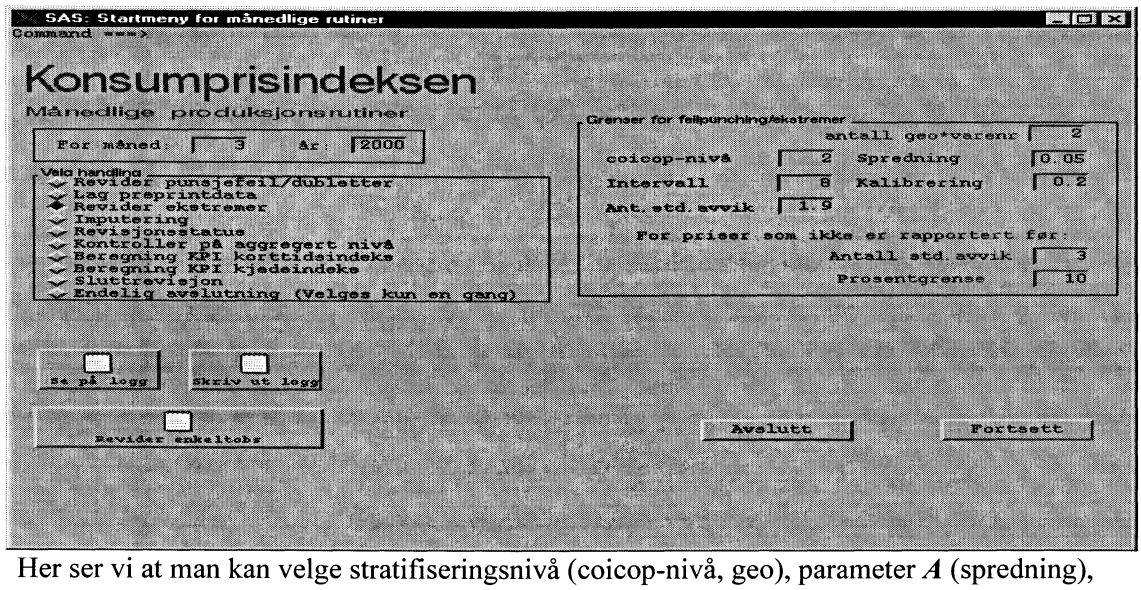
verdier av transformert variabel som avviker mer fra medianen enn en konstant C multiplisert med maks av
- avstanden mellom median og kvartil
- en konstant A multiplisert med medianen,

skal behandles som ekstremer.

Parameteren C påvirker avstanden fra median til grensene for feilmerking, mens hensikten med A er å unngå problemene som oppstår når avstanden mellom kvartiler og median er meget liten. C og A er parametre som bør estimeres.

Ofte er det slik at store endringer fra forrige periode allikevel er korrekte. Dette er typisk ved produkter som er preget av store sesongmønster. For å unngå at slike observasjoner dukker opp i våre kontroller beregner vi også algoritmen med utgangspunkt i kvotene p_m/p_{m-12}. Kun observasjoner som slår ut som ekstremer i begge tilfeller, går videre til behandling.

Følgende bilde viser sammensetning av parametere i revisjonen av KPI :



Her ser vi at man kan velge stratifiseringsnivå (coicop-nivå, geo), parameter A (spredning), parameter U (kalibrering), parameter C (intervall). Dessuten velges antall standardavvik for både ny og gammel prisrapportering (ved bruk av Tjebysjefs ulikhetsmetode i tillegg til HB-metoden).

Valg av parametere er basert på erfaringene ved de månedlige revisjonene, og stort sett ligger parametrene fast etter at flere alternativer er testet. For å illustrere betydningen av parametervalget, har vi beregnet ekstremer for KPI og for produksjonsindeksen (PI) med utgangspunkt i datamaterialet for februar og mars 2004. Følgende tabeller viser disse resultatene:

Effekten av parameter C (A=0.05 , U=0.5)

		Antall ekstremer		
Statistikk	Antall observasjoner	C=8	C=12	C=16
KPI	30 191	322	249	204
PI	1 200	61	46	40

*Effekten av parameter **U** (**A**=0.05 , **C**=12)*

		Antall ekstremer		
Statistikk	Antall observasjoner	***U***=0.0	***U***=0.5	***U***=1.0
KPI	30 191	244	249	282
PI	1 200	36	46	57

*Effekten av parameter A (**U**=0.5 , **C**=12)*

		Antall ekstremer		
Statistikk	Antall observasjoner	***A***=0.05	***A***=0.20	***A***=0.90
KPI	30 191	249	243	239
PI	1 200	61	45	43

Vi ser at parameter ***C*** betyr mest, mens parameter ***A*** betyr minst for begge statistikkene.

Kontroll av HB-metoden
Når median=1. kvartil og/eller median=3. kvartil, kollapser denne HB-metoden. En slik fordeling er vanlig når det er en sterk konsentrasjon rundt en enkelt verdi. I pristilfellet vil dette trolig kunne være situasjonen ved endringstall fra forrige måned, hvor de fleste kvoter blir 1 (dvs. de som ikke har endring). Bruk av parameter ***A*** kan representere en hjelp i denne situasjonen.

Et annet alternativ er å stratifisere kvotene på en slik måte at antall observasjoner per strata er stort nok til å tillate oss å overse kvotene=1, dvs. vi kontrollerer kun prisene som har endret seg fra forrige periode. En slik løsning brukes nå ved feilidentifisering av skannerdata hvor antall observasjoner er enormt stort og bare en liten andel viser prisendringene.

Egenskapene med datamaterialet har mye å si for å vurdere i hvor stor grad HB-metoden bør prioriteres.

4.4.6. Seleksjon basert på anslått revidert verdi

Dette er en metode for å identifisere store feil i en numerisk variabel, foreslått av Mevik (2005). Med stor feil mener vi at avviket, eller avstanden, mellom original og revidert verdi er stor. Mer presist definerer vi en stor feil ved at

$$|revidert\ verdi - original\ verdi| > \alpha + \beta \cdot revidert\ verdi\,,$$

der $\alpha \geq 0$ og $0 \leq \beta < 1$. (Tegnet $|\ |$ betyr absoluttverdi, dvs. at $|3| = 3$ mens $|-3| = 3$). Velger vi $\beta > 0$, betyr det at vi aksepterer større avvik jo større den reviderte verdien er. Hvor store α og β skal være, blir en avveiing mellom hvor nøyaktig tall vi ønsker og hvor mye resurser som skal brukes til retting.

Ideen er nå å lage en automatisk rutine som anslår, eller gjetter på, den reviderte verdien (før selve opprettingen av eventuelle feil tar til), og så plukke ut enheter hvor avviket mellom original og anslått verdi er stor. Dvs. vi plukker ut enheter hvor

$$|anslått\ verdi - original\ verdi| > \alpha + \beta \cdot anslått\ verdi\,.$$

Hvordan vi skal anslå den reviderte verdien, avhenger av hvordan variabelen blir revidert. Den automatiske rutinen må derfor utvikles i samarbeid med de som utfører revisjonen.

Hvor treffsikker denne metoden er, avhenger av den anslåtte verdien. Jo bedre vi klarer å gjette på den reviderte verdien, jo flere av de store feilene vil bli plukket ut (samtidig som færre småfeil plukkes ut).

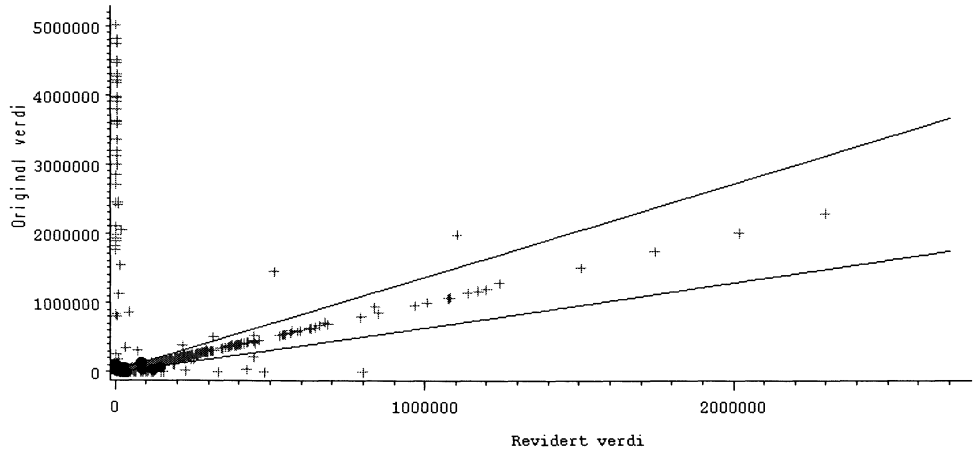
Dette er med andre ord en metode som egner seg for variable hvor det finnes tilgjengelig informasjon som kan brukes til å anslå den reviderte verdien.

Eksempel med data fra strukturstatistikk for industri

Strukturstatistikk for industri er en årlig utvalgsundersøkelse som blant annet skal gi en oversikt over produksjonsverdi og produksjonskostnader for bedriftene innen bergverksdrift og industri. Her skal vi se på en variabel som gjelder bedriftens produksjonsinntekter vedrørende salg av egenproduserte varer. En sammenligning av original og revidert verdi for de 2298 bedriftene i 2001-utvalget som tilhører et enbedriftsforetak, viser at denne variabelen er blitt rettet for samtlige av bedriftene. Men de aller fleste rettingene gjelder småfeil. Hvis vi f.eks. setter $\alpha = 15\ 000\ 000$ og $\beta = 0.35$, er det bare 153 av feilene som regnes som store, og disse utgjør hele 99,3% av den totale feilen. (Total feil = $\sum_{alle\ bedrifter} |revidert\ verdi - original\ verdi|$, dvs. summen av bedriftenes absolutte feil).

Ved revisjon av strukturstatistikken benyttes flere datakilder, f.eks. fjorårets reviderte verdi, Næringsoppgaven og årets Produksjonsstatistikkskjema. I denne omgang har vi veldig enkelt bare benyttet oss av fjorårets reviderte verdi, og brukt denne som et anslag på årets reviderte verdi (dvs. *anslått verdi = fjorårets reviderte verdi*). Selv med dette enkle anslaget klarer vi å plukke ut 134 av de 153 store feilene. Dvs. at det bare er 19 store feil som ikke plukkes ut. Totalt plukkes 225 bedrifter ut til videre sjekk, og feilen til disse utgjør 99,2% av den totale feilen. Dette er illustrert i figur 4.4.1 som viser plott av original verdi mot revidert verdi for salg av egenproduserte varer.

Figur 4.4.1. Plott av original verdi mot revidert verdi for salg av egenproduserte varer



De to heltrukne linjene i plottet markerer grensene for når en feil er stor; store feil er markert enten over den øverste linjen eller under den nederste linjen. De 19 store feilene som ikke blir tatt ut til retting, er markert med •. (For å gjøre figuren mer lesbar er ikke alle bedriftene markert i figuren).

4.5. Sammendrag - kontroller

Logiske kontroller og rettinger som bare avhenger av data for den enkelte enhet bør gjøres først. Det bedrer grunnlaget for kontrollgrenser og selektiv gransking. Sluttkontroll gjøres på aggregerte tall, og mistenkelige verdier der undersøkes ved å se på aggregerte tall på lavere nivå og til slutt på enkeltenheter.

Metoder for selektiv revisjon går ut på å prioritere enheter for videre kontroll. Prioriteringen kan gjøres etter

- størrelse (vektet eller ikke vektet),
- avvik (fra forventet verdi eller forrige periode)
- annet (f.eks. nye enheter)

Rettingene kan avsluttes når de ikke gir vesentlig utslag.

5. Grafisk revisjon

5.1. Generelt

Grafisk analyse gir muligheter for en bedre forståelse av datasettet. Metodene er enkle å bruke og gir raskt et visuelt bilde av datamaterialet. Vi ser fordelingen for de enkelte variable og sammenhengen mellom dem. Dermed ser vi også mønstre og avvik fra det normale.

Grafiske metoder kan avsløre feil som konvensjonelle metoder ikke kan avsløre. Områder hvor grafiske metoder har klare fordeler fremfor tradisjonelle metoder:

- makrorevisjon før publisering
- revisjon av små engangsundersøkelser
- utvikling av metoder (grenser) for selektiv revisjon
- feil verdi innenfor normalt variasjonsområde (innliggere) som er vanskelige å avsløre ved tradisjonelle metoder

Grafiske metoder har også sine ulemper. De viktigste problemområdene er:

- dårlig egnet for datasett med mange feil hvor de fleste enhetene må rettes
- dårlig egnet hvis datasettet består av mange nøkkelvariable - det blir uoversiktlig med mange grafer som alle anses viktige
- avslører ikke inkonsistens for én observasjon
- personavhengig - subjektiv

Det vil derfor ofte være nyttig å kombinere grafiske metoder med mer tradisjonelle revisjonsmetoder.

5.2. Grafiske metoder

De mest brukte grafene er histogram/stolpediagram, boksplott, punktdiagram og linjediagram. Histogram/stolpediagram egner seg best til å avsløre absolutte feil som f.eks. ulovlige verdier. Boksplott kan brukes på flere nivå innen revisjonsprosessen. De egner seg godt for sluttkontroller av aggregerte verdier, men kan også brukes på mikronivå hvor de gir mulighet for å avsløre ekstremverdier. I tillegg er boksplott et godt verktøy ved fastsettelse av grenseverdier for tradisjonelle

revisjonsmetoder. Punktdiagram egner seg for å avsløre ekstremverdier og innliggere. Linjediagram brukes sjeldnere, de passer bare for tidsserier.

5.2.1. Grafiske sluttkontroller

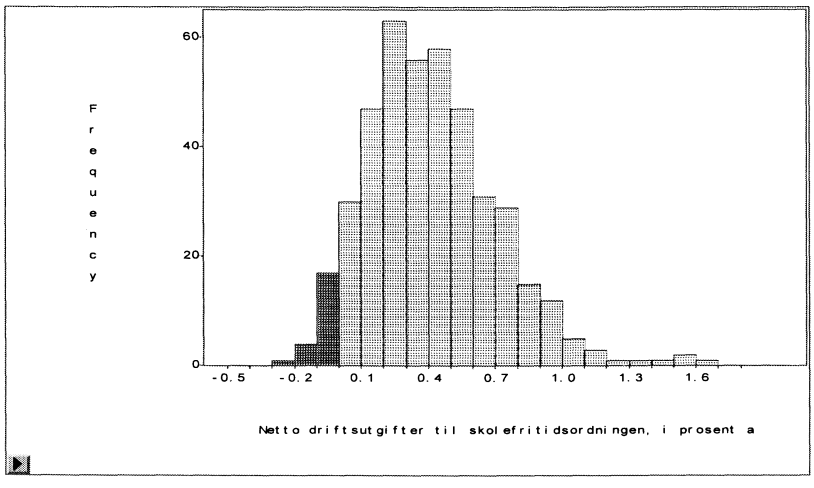
Grafisk revisjon egner seg godt til makrokontroller før publisering. Her vises noen eksempler på grafisk revisjon ved bruk av SAS - Insight på nøkkeltall som publiseres i KOSTRA (Kommune-Stat-rapportering). Disse eksemplene er hentet fra Foss (2003).

Eksempel på histogram

I figur 5.2.1 er det laget et histogram som viser hvordan andelen (%) av driftsutgifter som brukes til skolefritidsordningen, fordeler seg. Alle kommunene bruker mellom -0,3 og 1,7 % (x-aksen) og høyden på stolpene angir antall kommuner som bruker angitt andel.

Av figur 5.2.1 ser vi at noen kommuner har fått negative driftsutgifter til skolefritidsordningen, det vil si at de har hatt inntekter. Dette virker mistenkelig og bør undersøkes.

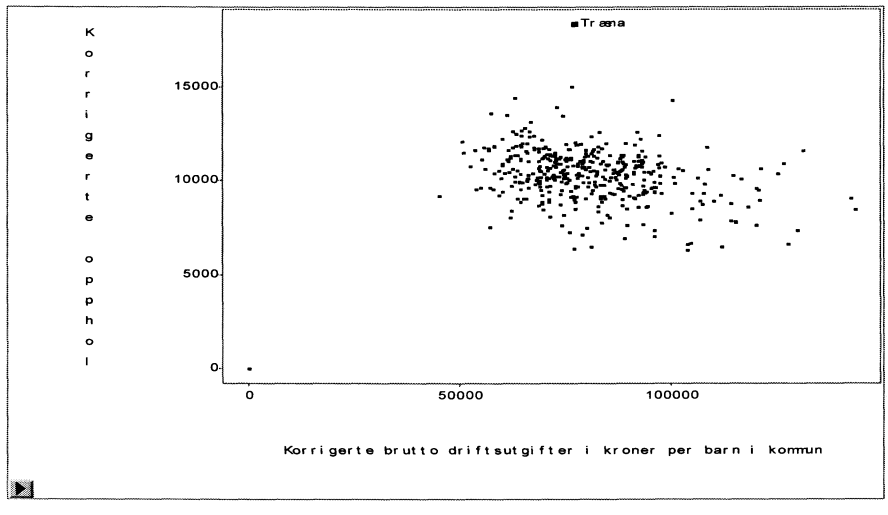
Figur 5.2.1. Stolpediagram med markering av ugyldige verdier



Eksempel på punktdiagram (X-Y plott)

I figur 5.2.2 er de to nøkkeltallene 'Korrigerte brutto driftsutgifter i kroner per barn i kommunale barnhager' og 'Korrigerte oppholdstimer per årsverk i kommunale barnehager' plottet mot hverandre. Hvert punkt i plottet tilsvarer verdiene på disse to nøkkeltallene for en kommune.

Figur 5.2.2. Punktdiagram av to nøkkeltall

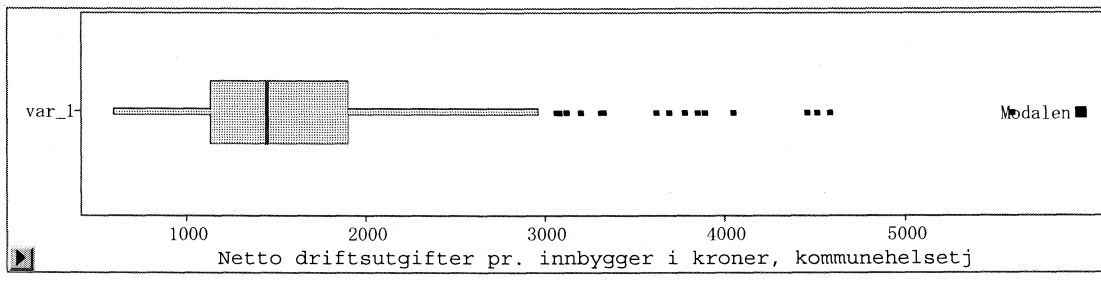


Av dette plottet ser vi at nøkkeltallene sprer seg fint ut i en ellipse og viser at det ikke er noe klar sammenheng mellom disse nøkkeltallene. Det er én kommune som har oppgitt 0 på begge nøkkeltallene, mens Træna har en svært høy verdi på nøkkeltallet 'Korrigerte oppholdstimer per årsverk i kommunale barnehager'.

Eksempel på boksplott

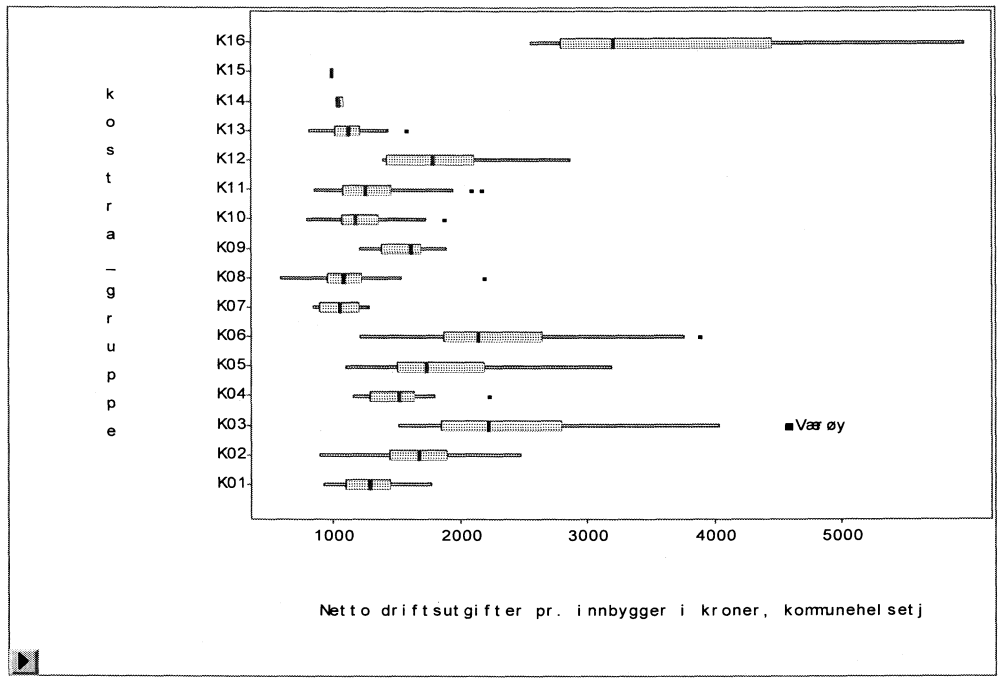
Boksplott viser fordelingen av enkeltvariable, med markering for median, øvre- og nedre kvartil og ekstremverdier. Ekstremverdier angis som punkter som ligger utenfor de heltrukne smale områdene. I figur 5.2.3 er det laget et boksplott av nøkkeltallet 'Netto driftsutgifter pr. innbygger i kroner, kommunehelsetjenesten'. Av dette plottet ser vi at det er en del kommuner som skiller seg ut for dette nøkkeltallet.

Figur 5.2.3. Boksplot av nøkkeltall



Ekstremverdier som ses her, behøver ikke være ekstreme i andre sammenhenger. De kan f.eks. være normale verdier innefor en gruppe av enheter. For å kontrollere dette kan vi se på figur 5.2.4 som viser et boksplot for hver KOSTRA-gruppe. KOSTRA-grupper er en inndeling av kommuner etter befolkningsstørrelse, frie disponible inntekter og bundne kostnader.

Figur 5.2.4. Boksplot av nøkkeltall etter KOSTRA-grupper



De fleste av de kommunene som så ut til å være ekstremverdi i figur 5.2.3, tilhører KOSTRA-gruppe 16, som er de 10 kommunene med høyeste frie disponible inntekter per innbygger. Siden

denne gruppen har høye disponible inntekter, blir det stor variasjon i hvor mye de bruker på drift av skoler. Derimot må Værøy karakteriseres som ekstremverdi siden verdien ligger langt fra de andre i den samme gruppen.

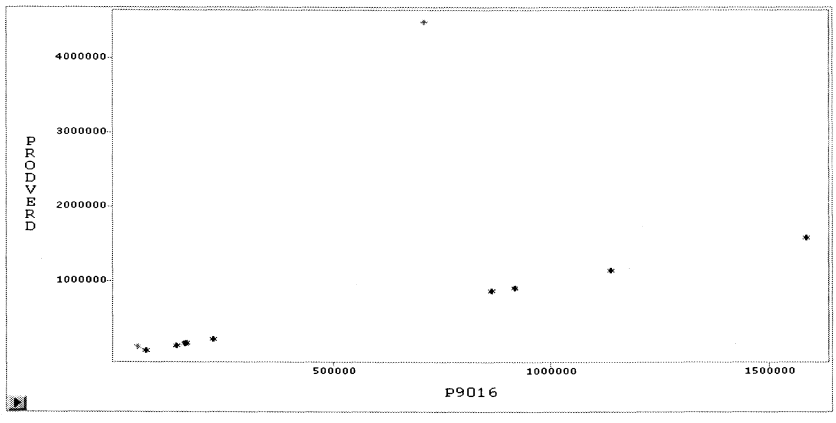
5.2.2. Ekstremverdier i mikrodata

Ekstremverdier avsløres vanligvis best ved plott, X-Y plott eller boksplott, hvor uvanlige verdier lett sees.

Eksempel på ekstremverdier

Figur 5.2.5 viser sammenhengen mellom produksjonsverdi (prodverd) fra spørreskjema i forhold til produksjonsverdi (P9016) fra regnskapsregisteret for aksjeselskap. Hvert punkt gjelder et enbedriftsforetak, alle innen samme næring. Observasjonen med høyest produksjonsverdi (over dobbelt så stor som den nest høyeste produksjonsverdien) er merket med +, og vil fremkomme som ekstrem selv om vi bare ser på produksjonsverdi fra skjema.

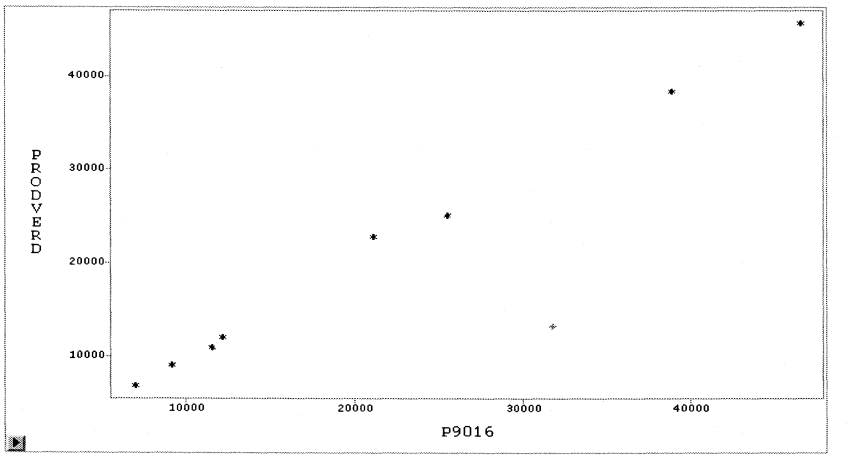
Figur 5.2.5. Ekstremverdi (utligger) i X-Y plott



Eksempel på innligger

Figur 5.2.6 viser sammenhengen mellom produksjonsverdi (prodverd) fra spørreskjema i forhold til produksjonsverdi (P9016) fra regnskapsregisteret for aksjeselskap. Hvert punkt gjelder et enbedriftsforetak, alle innen samme næring. Observasjonen merket med + kommer ikke frem som ekstrem hvis vi vurderer enten produksjonsverdi fra skjema alene eller produksjonsverdi fra regnskapsregisteret alene. Dette er et eksempel på innligger som tydelig fremstår som utligger når produksjonsverdiene fra spørreskjema og fra regnskapsregisteret ses i sammenheng.

Figur 5.2.6. Innligger som fremstår som utligger ved X-Y plott



5.2.3. Analyse av revisjonsrutiner

Ekstremverdier etter ulike definisjoner, rettede verdier og godkjente verdier kan illustreres grafisk. Plott kan brukes til vurdering av hvorvidt reglene for videre kontroller i revisjonssystemet er effektive, det vil si at de fanger opp flest mulig feil og færrest mulig akseptable verdier.

Eksempler på analyse av revisjonsrutiner - Hidiroglou-Berthelot metoden

Hvor mange og hvilke verdier som blir plukket ut som mistenkelige ved HB-metoden, avhenger av verdiene på parametrene ***U*** og ***C***. Høyere verdi på ***C*** vider ut grensene for godkjennelse, og dermed blir færre verdier markert som mistenkelige. Figurene 5.2.7 og 5.2.8 illustrerer hvordan valget av ***U*** (som styrer hvor mye vekt vi legger på størrelsen av tallet) påvirker hvilke verdier som blir ansett som mistenkelige. Begge grafene viser leger-i-alt i 2001 langs x-aksen og forholdstallet

$$forhold = \frac{leger\_i\_alt\ 2002\ urevidert}{leger\_i\_alt\ 2001\ revidert}$$

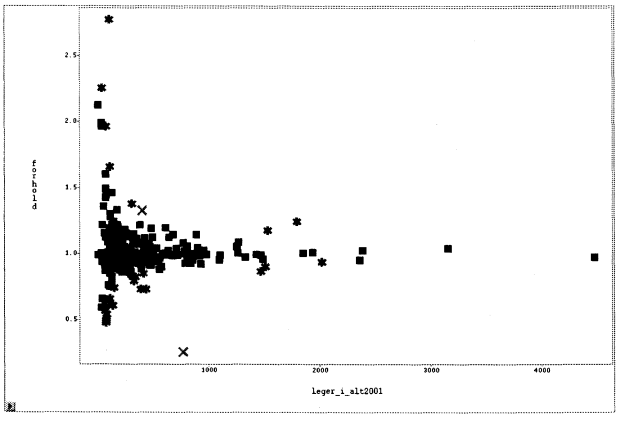
langs y-aksen. Dette forholdstallet vil være lik 1 hvis det ikke er noen endring, og vil ved normale endringer variere noe mer for små verdier enn for høye verdier.

Figur 5.2.7: Forholdet mellom originale tall i 2002 og reviderte tall fra 2001.

* - verdier som ble plukket ut av HB-metoden, når *U=0,7* og *C=15*

× - verdier plukket ut av HB-metoden, som også var blitt endelig endret i løpet av 2002

■ - verdier som ikke markeres

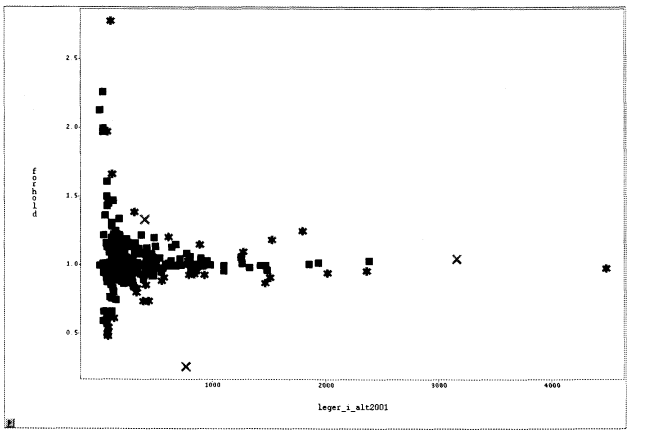


Figur 5.2.8: Forholdet mellom originale tall i 2002 og reviderte tall fra 2001.

* - verdier som ble plukket ut av HB-metoden, når *U=1* og *C=15*

× - verdier plukket ut av HB-metoden, som også var blitt endelig endret i løpet av 2002

■ - verdier som ikke markeres



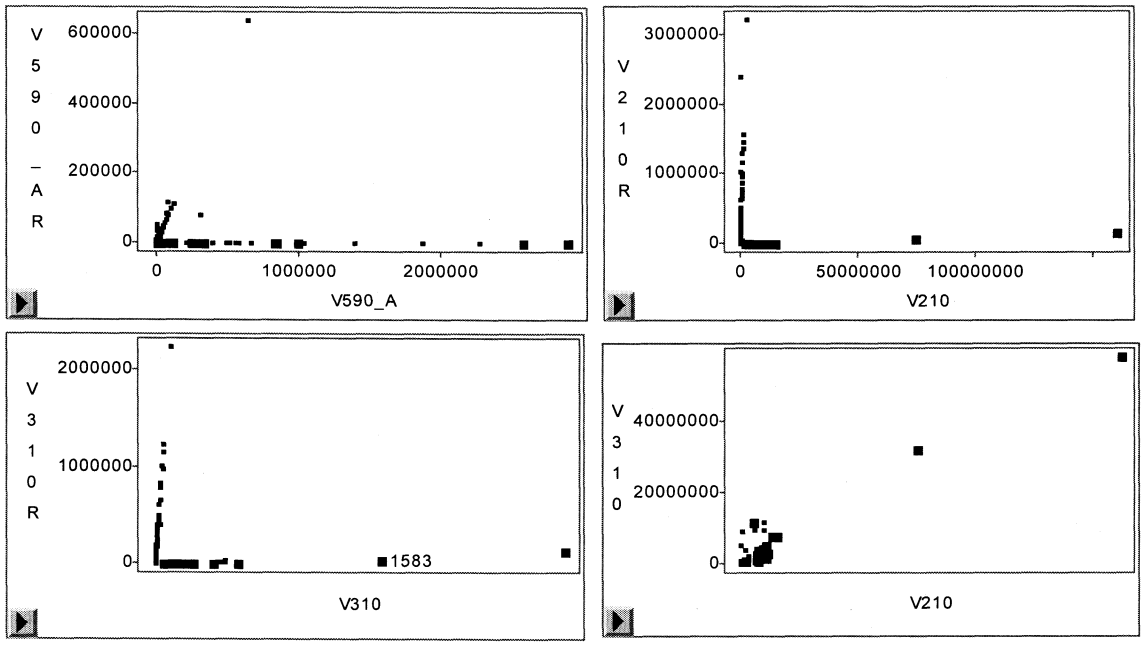
5.3. Interaktive metoder/SAS-Insight

Man kan arbeide med mange grafer samtidig. Markeringer som gjøres i en graf, kommer frem i alle grafer fra samme datasett. Det betyr at vi kan følge enkelte valgte enheter eller grupper av enheter. Det kan f.eks. være av interesse å se om en enhet som skiller seg ut med en utliggerverdi for en variabel også vil fremtre som utligger for flere variable.

Eksempel fra analyse av revisjon av strukturstatistikken industri

Tusenfeil fra en oppgavegiver forekommer ofte på flere variable samtidig. Figur 5.3.1 viser plott av originale data (x-aksen) og reviderte data (y-aksen) for investeringer (V590_A), salg av egenproduserte varer (V210) og råvarekostnader (V310), samt et plott av sammenheng mellom råvarekostnader og salg av egenproduserte varer. Tusenfeil kan identifiseres på de tre første grafene som de punktene som ligger langs x-aksen. Tusenfeil for "salg av egenproduserte varer" er markert (V210/V210_R > 900). Vi ser at også de aller fleste tusenfeil for "råvarer" markeres samtidig. Mange, men på langt nær alle tusenfeilene for investeringer markeres også.

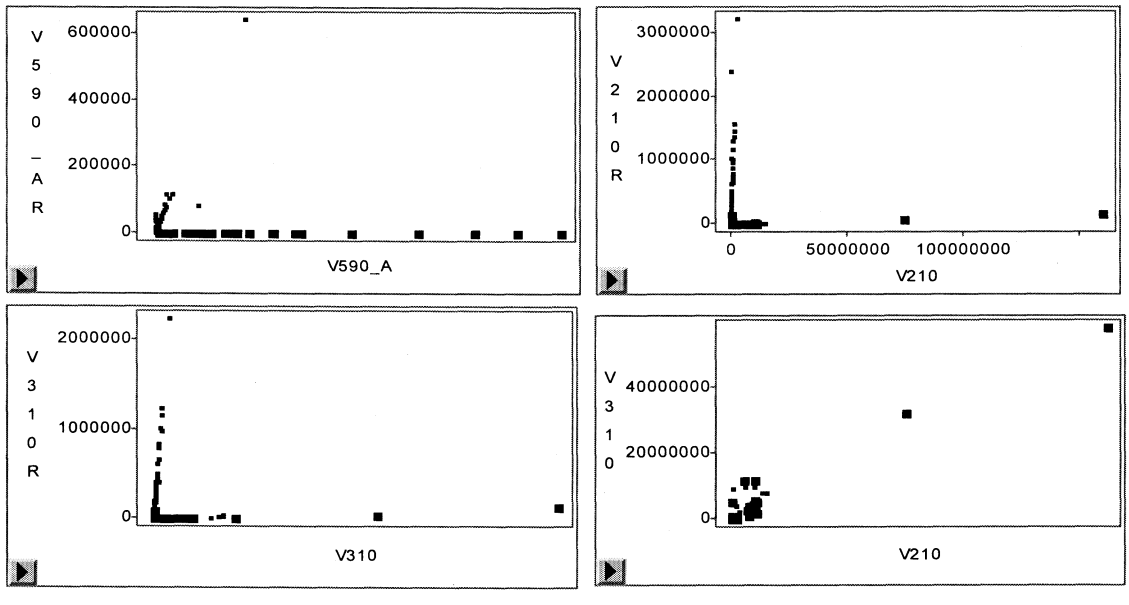
Figur 5.3.1. Markering av tusenfeil i "salg av egenproduserte varer"



Det er klar sammenheng mellom "salg av egenproduserte varer" - V210, og "råvarer" - V310 uavhengig av hvorvidt tusenfeil er markert.

Markering av tusenfeil for investeringer viser at oppgaver med tusenfeil i investeringer enten har tusenfeil i "salg av egenproduserte varer" og "råvarer" eller har svært lave verdier for disse variablene, se figur 5.3.2.

Figur 5.3.2. Markering av tusenfeil i "investeringer"



5.4. Sammendrag

Grafisk revisjon går ut på å presentere data i forskjellige typer diagrammer der mistenkelige verdier eller mønstre er lette å oppfatte ved visuell inspeksjon. Det er lett å se

- ugyldige verdier (f.eks. stolper for negative verdier)
- opphopning av verdier langs linjer (f.eks. periodefeil for lønninger eller tusenfeil)
- ekstremverdier (f. eks i gruppevise boksplott)
- isolerte enheter (punkter utenfor punktsvermen i X-Y-plott)

Grafikk kan med fordel brukes ved utvikling og kontroll av revisjonskriterier (kontrollgrenser)

6. Feilretting – endelig reviderte data

6.1. Retting av feil - korrigering

Når man under revisjon finner feil eller mistenkelige verdier i dataene, kan feil verdier erstattes med riktig verdi, eller i det minste en verdi som er riktigere enn den opprinnelige. Dette forutsetter at det er mulig å finne den riktige (eller riktigere) verdien. Det kan ofte være vanskelig. Tilbakekontakt til oppgavegiver er ressurskrevende både for SSB og oppgavegiver og bør forbeholdes viktige enheter eller systematiske feil. Det er mulig å automatisere tilbakemelding til oppgavegiver om mistenkelige verdier, se eksempel i kapittel 8.2 om databaser og revisjonssystemer.

Hvor mye som rettes vil avhenge av hvordan data brukes. Retting er normalt nødvendig bare for feil som har innflytelse på publiseringsnivå, tilfeldige feil kan oppveies av andre tilfeldige feil. Det er heller ikke nødvendig å rette tilfeldige feil i data hvis de er ubetydelige i forhold til utvalgsfeil. For mye retting kan medføre problemer, som forsinket publisering og feil bruk av ressurser. I tillegg kan rettinger føre til skjevhet. Det er ofte lettere å oppdage for store verdier enn for små verdier, eller det kan forekomme at systematiske rettemetoder innfører feilaktige mønstre i dataene.

Når man finner det nødvendig å rette data selv om man ikke har de riktige verdiene, kan man enten beregne en verdi fra andre variable på skjema eller fra andre kilder. Stokastiske imputeringsmetoder medfører ny usikkerhet i datamaterialet. Usikkerhetsberegninger krever derfor at slike imputeringsmetoder skilles fra vanlig revisjon.

6.2. Manglende verdier - frafall

Manglende verdier oppstår ved frafall i utvalgsundersøkelser eller tellinger og manglende informasjon i registre. Manglende data oppstår fordi

- enheten mangler - enhetsfrafall (enheter i populasjonen mangler fullstendig, f.eks. pga. opphør, nekting, sen datainngang, underdekning i registre e.l.)
- begrenset informasjon om enheten - partielt frafall (enheten i populasjonen er med i datamaterialet, men en del opplysninger mangler; ufullstendig oppgave)
- manglende verdi istedenfor en null (det er vanlig å ikke fylle ut poster der man ikke har noe, men det er ikke alltid like lett å vite om en blank post skal være 0 eller manglende)

Håndtering av frafall regnes vanligvis som en del av estimeringsproblemet. Enhetsfrafall korrigeres vanligvis ved vekting, men beregning av verdier brukes også. Partielt frafall korrigeres enten ved logiske beregninger, beregninger fra eksterne kilder eller stokastiske imputeringsmetoder. Stokastisk imputering (dvs. sette inn verdier for blanke poster basert på en estimert statistisk fordeling) behandles i en egen håndbok om frafall som er under utarbeidelse.

Revisjonsmetodene som brukes for beregning av manglende verdier eller retting av gale data, avhenger av hvilken informasjon som er tilgjengelig. Man forsøker å utnytte annen informasjon til å finne rimelige verdier for de manglende variablene. I enkelte tilfeller kan verdier beregnes fra andre kilder (register etc) og gi tilstrekkelig gode data. Mange av disse metodene kan legges inn som automatiske beregninger ved manglende data eller feil i data. De typiske imputeringsmetodene som brukes til feilretting er:

- Manuell imputering

Hvis omfanget av manglende data er begrenset, er det mulig å sette inn manglende verdier manuelt på bakgrunn av beregnede verdier, kontakt med oppgavegiver eller god fagkunnskap. Her bør man være oppmerksom på faren for personavhengighet.

- Deduktiv eller deterministisk imputering - logisk retting

Manglende verdi(er) for en enhet fastsettes fra andre data for samme enhet på grunnlag av logiske regler. F.eks. kan en sumpost beregnes som summen av underpostene (hvis disse er oppgitt). En oppgitt sumpost kan fordeles på underposter etter gitte "nøkkeltall". Logiske grunner tilsier at en blank post skal være lik 0. Hvis det går fram av oppgaven at en person er mor til barn, kan manglende avkryssing for kjønn settes til kvinne.

- Gjenbruk eller Cold-deck imputering

Her beregnes manglende verdier på grunnlag av data for samme enhet fra foregående undersøkelse. Den enkleste form for cold-deck imputering er ren kopiering av f.eks. priser fra forrige undersøkelse, men det kan legges inn mer avanserte beregninger.

- Donor eller Hot-deck imputering

Denne metoden beregner manglende verdier på grunnlag av data for en annen enhet i samme undersøkelse; en enhet som i en eller forstand ligner mest på «mottakende» enhet, f.eks. enhet i samme stratum, foretak i samme bransje eller personer i samme aldersgruppe.

6.3. Godkjennelse

Ikke alle enheter som kontrollmetodene flagger som mulige feil, skal rettes, dels fordi verdiene er bekreftet riktige eller fordi de aksepteres under usikkerhet, eller fordi eventuelle feil er ubetydelige. Men fordi de en gang har blitt merket som ekstremverdier, skal det merkes hvorvidt de er blitt vurdert og på hvilket grunnlag de er godkjent.

Eksempel på koder for godkjennelse eller annen behandling av mistenkelige verdier
Ved revisjon av utenrikshandelsdata kan revisorene rette selv når de finner feil, eller sende spørsmål tilbake til tollere. De har også maskinelle rutiner for omberegninger og masseendringer som merkes etter kodelisten:

G	-	Godkjent av revisor	O	-	Omberegning
T	-	Godkjent av toller	M	-	Masseendring
R	-	Rettet av revisor	U	-	Rettet revisor - godkjent toller
E	-	Rettet av toller	S	-	Rettet toller - rettet revisor

G - godkjent av revisor er den absolutt mest brukte koden. Den brukes for opptil 70 - 80 % av varelinjer markert for mulige feil.

6.4. Flagging

Alle enheter/verdier som er flagget for kontrollmetode (frafall, absolutte eller mulige feil), må også flagges for videre behandling. Flaggingsvariable skal gi opplysning om hvilken behandling en observasjon eller variabelverdi har, slik som "godkjent", "rettet", "bekreftet ved ny kontakt" eller andre koder. Eksempel på behandlinger - som til en viss grad vil variere for ulike statistikker er

- rettet maskinelt
- vurdert manuelt
    - godkjent
    - rettet
    - godkjent etter tilbakekontakt med oppgavegiver
    - rettet etter tilbakekontakt med oppgavegiver
- ikke vurdert eller videre behandlet
    - ubetydelig enhet
    - dårlig tid

Verdier som er imputert, skal også merkes på fil. Dette er viktig både for eventuell senere analyse av dataene og for kvalitetssikring av produksjonsprosessen. Det er vesentlig å flagge at kontroller er utført også når det ikke er feil, særlig ved arbeidskrevende manuelle kontroller.

Det er viktig at ekstremverdier (outliers) merkes hvis de er godkjent som korrekte verdier, - ikke bare hvis de blir rettet. Dette har betydning for imputerings- og estimeringsprosessene.

Eksempel på flagging i Boligtellingen 2001
I Folke- og boligtellingen ble det laget en enkel flaggerutine for revisjon av boligskjemaet. På grunn av det store datamaterialet ble kontrollene og endringene hovedsakelig gjort maskinelt. Kategoriene for flagging var: kontrollert men godkjent, rettet, opplysninger hentet fra GAB-registeret, imputert partielt, imputert for hele enheten og en enkel korrigering uten å hente informasjon utenfor besvarelsen.

Spørsmål fra boligtellingen 2001 etter revisjonskode. Prosent

Spørsmål om bygningen	I alt	Ikke endret	God-kjent[1]	Rettet	GAB	Imputert (partielt)	Imputert (enhet)	Enkel[2]
7. Når (intervall) ble bygningen eller huset du bor i bygget?	100	83,5	0,0	7,2	0,0	2,3	7,0	0,0
8. Kan du oppgi et mer nøyaktig byggeår?	100	45,6	9,3	0,8	26,6	5,1	7,0	5,6
9.1. Har bygningen kjeller?	100	78,5	0,0	13,2	0,0	1,2	7,0	0,0
9.2. Har bygningen underetasje?	100	54,8	0,0	36,3	0,1	1,7	7,0	0,0
10. Hvor mange etasjer har bygningen eller huset du bor i?	100	91,1	0,0	0,0	0,2	1,7	7,0	0,0
11. Er det heis i bygningen?	100	89,2	0,0	0,1	2,1	1,6	7,0	0,0

[1] En kontroll har slått ut, men verdien har blitt godkjent.

[2] En enkel korrigering av verdien som er gjort uten å søke informasjon utenfor selve besvarelsen.

Eksempel på flagging i Grunnskolens informasjonssystem (GSI)

I forbindelse med at data blir registrert på GSI-skjemaet settes det elektroniske spor. Disse sporene består av informasjon om endringer som er gjort i dataregistreringen, og de blir samlet i en såkalt loggfil. Denne filen inneholder informasjon om hvem som har gjort endringer, når det har skjedd endringer, hva man har endret til og hva man har endret fra. Denne informasjonen ble analysert i 2002 og brukt til å endre skjemaet (Roll-Hansen et al, 2002).

Endring av tall til enten lavere eller høyere verdi etter institusjon

Institusjon som har utført endringen	Totalt antall endringer	Lavere nytt tall. Prosent	Høyere nytt tall. Prosent
Land	119	93,28	6,72
Fylke	1 024	37,89	62,11
Kommune	11 217	36,00	64,00
Bydel	757	43,99	56,01

6.5. Når er det revidert nok? - overediting

Revisjonen legges opp etter statistikkens (dataenes) bruksnivå, om det er totaltall, detaljert statistikk eller mikrodata. Men uansett bør man begrense kontroller, manuelt arbeide og rettinger mest mulig.

Elektronisk databehandling gir mulighet for mange kontroller og dermed mange mistenkte feil i data. Disse blir deretter vurdert av kontrollør/revisor, ofte manuelt og dels med tilbakekontakt til oppgavegiver. Man tar for gitt at det er mulig å finne ut hva som er feil og finne en riktig(ere) verdi. Dette kan føre til overforbruk av ressurser uten at kvaliteten på data øker. Problem som kan oppstå ved for mye revisjon:

- Sen publisering, aktualiteten blir dårligere
- Ressurser brukes på ubetydelige endringer - for høye kostnader
- Viktige feil vil ikke alltid oppdages blant et svært stort antall flaggede feilmeldinger
- Skjevhet - da det ofte er lettere å oppdage tall som er for store enn tall som er for små
- Oppgavegiver har ofte ikke mulighet til å gi bedre tall ved tilbakekontakt
- Korreksjoner kan være feil, det kan til og med forekomme at retting gjør feilen større
- Mange rettinger/imputeringer kan føre mønster fra modeller som brukes under imputering/korreksjon inn i data. Revisorene vil tilpasse egne rettinger til hva de tror.

Særlig for bedrifts- og foretaksundersøkelser med sine svært skjeve fordelinger, vil en liten andel av feilene dominere total endring. Estimatene kan forbli tilnærmet uendret selv om revisjonsarbeidet reduseres betydelig.

Det er viktig å begrense flagging av ubetydelige feil mest mulig. Bare en liten del av tradisjonelt flaggede feil bør behandles manuelt.

Kontrollmetodene bør rette fokus mot mulige alvorlige feilkilder.

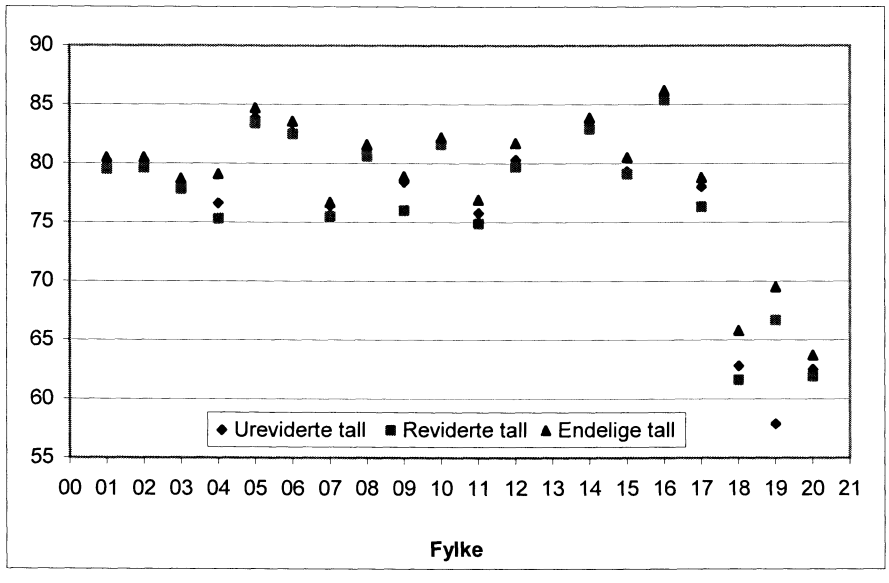
Eksempel på behov for detaljert revisjon

Fagdepartementets oppfølging av unge beboere på eldreinstitusjoner forutsetter spesielt stor presisjon i underlagsdata.

Eksempel på unødvendig retting
Priskontroll av utenrikshandelsstatistikk førte til markering for mulige feil i vel 100 varelinjer for en gitt vare. Vekten rettes for 75 av disse varelinjene. Rettingen førte til en økning på 0,4% av total vekt for alle feilmarkerte varelinjer med dette varenummeret.

Eksempel på revisjon som kan føre til nye feil
Informasjon om elever som har bestått eksamen kommer først inn som summariske tall gjennom KOSTRA og senere inn på individnivå til den nasjonale utdanningsdatabasen (NUDB). I sammenligning mellom urevidert KOSTRA tall, revidert KOSTRA tall og tall fra NUDB er det tydelig at for de fleste fylker fører revisjonen til at tallene får lavere verdi, mens de endelige tallene fra NUDB ligger høyere enn de ureviderte tallene. (Se figur 6.5.1). Dette vil si at i mange tilfeller ligger de ureviderte tallene nærmere de endelige tallene enn de reviderte tallene. Dataene har blitt revidert for mye i negativ retning. (Imidlertid korrigerer revisjonen et stort avvik (i fylke 19). Det er godt mulig at revisjonsrutinen kan beholde evnen til slike korreksjoner selv om den skulle bli justert for å unngå feilaktige nedjusteringer.)

Figur 6.5.1. Andelen elever med bestått eksamen. Ureviderte tall, reviderte tall og endelige tall. Prosent



6.6. Sammendrag

- Frafall
  - Enhetsfrafall håndteres oftest ved vekting
  - Partielt frafall håndteres ved imputering
- Mistenkt feilaktige verdier kan bli
  - bekreftet eller godkjent
  - rettet med imputeringsmetode
  - rettet på annen måte
- All behandling flagges
  - for dokumentasjon av datakvalitet
  - for vurdering av revisjonsopplegget

7. Administrative data og registerdata

Gjenbruk av administrative data utgjør en hoveddel av SSBs datafangst. Grunnen til at vi velger å bruke administrative data i så stor grad, er hovedsakelig god tilgang på relevante data og å begrense oppgavebyrden for våre oppgavegivere. Nesten all statistikk som blir produsert i Statistisk sentralbyrå, har enten et utgangspunkt i, eller henter informasjon fra en eller flere administrative datakilder. Noen statistikker blir produsert direkte fra ett eller flere registre, mens enkelte andre statistikker bare kopler på informasjon fra registre. De fleste utvalgsundersøkelser trekker utvalget fra ett av populasjonsregistrene. Filuttrekk fra oppgavegivers fagsystem (f.eks. regnskap-, lønn- og personalsystem) kan brukes i stedet for utfylling av skjema. Det er programvareleverandørene som legger til rette for datauttrekkene etter spesifikasjoner fra statistikkseksjonen.

Den omfattende bruken av administrative data krever gode kontroll- og revisjonsmetoder. Gode kontroller krever god kjennskap til formålene for innsamling av administrative dataene. Da er det mulig å rette oppmerksomheten om de data som er viktig for statistikken, men som har falt på siden av de administrative formålene. Ved mottatte administrative data er vi avhengig av den kvalitetskontroll som er foretatt av dataeier. Denne kvalitetssikringen kan være god eller dårlig, men den vil uansett primært være rettet mot kontroll for administrative formål. Slike kontroller *kan* være til begrenset nytte for bruk av materialet til statistiske formål. Kontakt og samarbeid med dataeier er derfor svært viktig.

> ***Eksempel - Data fra Skattedirektoratet***
> Skattedirektoratet vil være opptatt av om utliknet skatt og nettogrunnlaget for dette er riktig, men mindre opptatt av om bruttooppgavene for inntekter og kostnader og underspesifikasjoner er korrekte.
>
> ***Eksempel - Innsatsstyrt finansiering av sykehustjenester***
> Innsatsstyrt finansiering av sykehustjenester (ISF) har ført til en "politisk" dreining av datagrunnlaget for pasientstatistikken i retning av forhold som er av betydning for ISF.

Revisjonen av administrative data, og da særlig registre, avviker fra revisjon av skjemaundersøkelser på mange måter. Datamengden kan ofte være svært stor. Det kan da være vanskelig å kontakte enhetene i registeret ved feil eller mistanke om feil. Vi kan ikke rette i selve registeret, bare i vårt register til statistikkformål. Data fra ulike registre kan gi motstridende informasjon om samme variabel.

7.1. Sentrale registre og administrative data

Sentrale administrative data og registre som blir brukt i SSB
Skattedirektoratet eier en rekke register, hvorav de viktigste er:

- Det sentrale folkeregister (DSF)
- Liknings- og selvangivelsesregister
- Lønns- og trekkoppgaveregisteret (LTO)
- Momsregister

I tillegg er flere viktige administrative registre plassert blant Brønnøysundregistrene. Det gjelder blant annet:

- Enhetsregisteret (ER)
- Regnskapsregisteret
- Oppgaveregisteret

Toll- og avgiftsdirektoratet eier:

- Register over utførsel og innførsel av varer (TVINN- registeret)

Rikstrygdeverket eier:

- Arbeidsgiver- og arbeidstakerregister (Aa-registeret)

Statens kartverk eier:

- Register over grunneiendom, adresser og bygninger (GAB)

Felles for disse registrene er at vi mottar dataene for alle enhetene samlet og på en maskinell form. Det er ofte lite hensiktsmessig med omfattende mikrokontroll for å blinke ut spesielle enheter ved mottak. I de fleste tilfellene er det ofte heller ikke aktuelt å ta direkte kontakt med den enkelte enhet.

Statistiske populasjonsregistre som blir driftet i SSB
Enhetsregisteret er ett av tre administrative basisregistre som utgjør det som omtales som infrastrukturen i informasjonssamfunnet. De to andre er Det sentrale Folkeregister og Registeret over grunneiendommer, adresser og bygninger/boliger. Disse utgjør selve fundamentet for all gjenbruk av samfunnets data siden enhetene som inngår er entydig identifisert og registrene er Masterregistre. Basisregistrene inngår i et samspill med tre tilsvarende registersystemer i SSB. Disse registrene utgjør en felles populasjon og er kjernen i SSBs statistikksystemer.

Populasjonsregistrene i SSB er:

- Bedrifts- og foretaksregister (BoF)
- Register over grunneiendom, adresser og bygninger SSB sin versjon (SSB-GAB)
- Befolkningsstatistikksystemet (BESYS)

Andre viktige registre som blir driftet i SSB er

- Dødsårsaksregisteret, eid av Nasjonalt folkehelseinstitutt
- Nasjonal utdanningsdatabase- NUDB

7.2. Kontakt og samarbeid med dataeier

Et godt samarbeid med registereier vil kunne gi bedre kontroller og øke kvaliteten på data for registereier i tillegg til at vi får bedre data inn til SSB. Statistikkloven gir oss muligheter for å påvirke innholdet i offentlige registre og de standardene som ligger til grunn. Det kan likevel ha begrenset verdi å utvide et register med nye kjennemerker hvis de bare skal brukes til statistiske formål. Faren er stor for at slik informasjon ikke blir tilstrekkelig kvalitetssikret av registeransvarlig. Vi må derfor gjøre en streng prioritering av våre ønsker om nye kjennemerker og gjennom dialog med registeransvarlig sikre oss at disse kjennemerkene får en tilstrekkelig god kvalitet. Dette betyr først og fremst at statistikkdata blir kvalitetssikret innholdsmessig, men også at datamaterialet blir kontrollert og rettet for trivielle summefeil mv.

> ***Eksempel på samarbeide med Toll- og avgiftsdirektoratet***
> Toll- og avgiftsdirektoratet bruker priskontroller utviklet i SSB. SSB følger FNs retningslinjer når det gjelder statistikken over utenrikshandel med varer. FN har utarbeidet en manual for beste praksis for hvordan retningslinjene skal følges. I manualen gir FN konkrete anbefalinger om kontakten og samarbeidet mellom statistikkmyndighet og tollmyndighet.

7.3. Kontroll av administrative data

7.3.1. Kontroll uten bruk av andre kilder

Registrene blir ofte kontrollert for ekstremverdier og absolutte feil, slik som ulovlige verdier og logiske feil. Dette kan bli gjort uten bruk av eksterne kilder. Stor datamengde tilsier at det i hovedsak blir brukt maskinelle kontroller og korrigeringer. Det kan være mulig å spørre registereier om verdier vi finner urimelige. Noen ganger kan det føre til at registereier retter i sitt register eller legger inn nye kontroller for å unngå at slike feil oppstår igjen.

Eksempel - Kontroll av fødtefilen fra befolkningsstatistikken (Brørs et al 2000)
Det er omtrent 60 000 observasjoner på fødtefilen, og det er forholdsvis få endringer som blir gjort etter at kontrollene er gjennomført. Dette er de kontrollene som blir utført på fødtefilen:

- **Kommunenummer** - blir kontrollert maskinelt for gyldig nummer.
- **Alder** - blir kontrollert for ekstreme tilfeller.
- **Dødfødte** - blir kontrollert for at antallet ser rimelig ut.
- **Flerfødslene** - blir kontrollert for at koden for fødselstypen stemmer.
- **Ekteskapets varighet** - blir kontrollert for at det ser rimelig ut.
- **Statsborgerskap** - blir kontrollert for gyldige koder.

7.3.2. Kontroll mot andre registre

Manglende enheter og andre feil i registre kan bli avslørt ved å koble registre sammen. Noen registre bør inneholde de samme enhetene, men det kan også være slik at et register inneholder bare deler av et annet register eller at to registre helst ikke skal ha felles enheter. Når enkelte enheter forekommer bare i et (eller et fåtall) registre, kan det skyldes enhetsfrafall i enkelte registre og fører til partielt frafall ved kobling av registre. Verdier blir da ofte imputert.

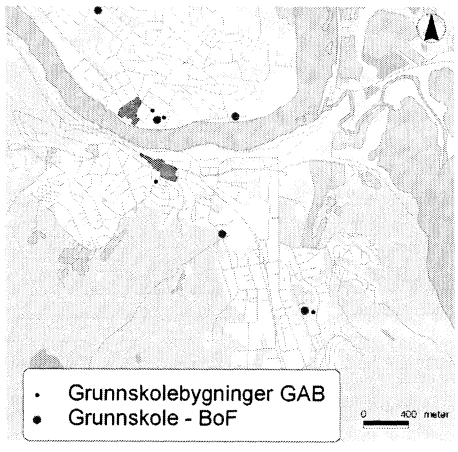
Samme variabel kan forekomme i flere registre. Det er viktig at hver variabel bare blir revidert én gang. Da må det samarbeides om hvem som kontrollerer hver variabel og hvor den eventuelt korrigeres. Når statistikk produseres ved kobling mellom flere registre, må det etableres rutiner for behandling av variable med ulike og muligens motstridende verdier fra ulike registre

Eksempel - Kontroll av Dødsårssaksregisteret ved bruk av dødsemeldinger fra DSF
I Dødsårsaksregisteret blir det kontrollert at alle døde er kommet med, ved at de medisinske dødsmeldingene fra legene sjekkes mot dødsmeldinger fra DSF. Ved manglende medisinsk dødsmelding blir det sendt purring. Ved manglende svar på purring blir dødsårsaken registrert som ukjent. (Det kan forekomme forskjell mellom antall døde i en årgang fra dødsårsaksregisteret og fra befolkningsseksjonen på grunn av forskjeller i avgrensning og databearbeiding.)

Eksempel - Kontroll av Bedrifts- og foretaksregisteret
Bedriftspopulasjonen i BoF kvalitetssikres gjennom direkte kontakt med enheter i forbindelse med datafangsten til strukturstatistikkene og gjennom kobling mot administrative registre i Overvåkingssystemet for bedrifter i BoF.

Eksempel - Kontroll av GAB ved hjelp av BoF
Sampresentasjon av stedfestede registre kan avsløre mangler. Figur 7.3.1 viser eksempler på stor avstand mellom skolebygning og bedrift, noe som indikerer feil eller mangler.

Figur 7.3.1. Manglende stedfesting og feil plassering i GAB og BoF



Eksempel - Registersysselsetting
I utarbeidelsen av den registerbaserte sysselsettingsstatistikken innhentes informasjon om arbeidsforhold fra en rekke registre. Hovedkilden for data om lønnstakere er Rikstrygdeverkets Aa register. I tillegg til jobbrelatert informasjon som arbeidstid, yrke og hvilken bedrift personene er ansatt i, inneholder registeret også en datering av arbeidsforholdet. Denne dateringen kontrolleres mot informasjon fra Aetats register over registrert ledige ved arbeidskontorene (ARENA - registeret). Hvis man finner at en person er registrert med både et aktivt arbeidsforhold og et aktivt ledighetsforhold på samme tidspunkt, vil dateringen på arbeidsforholdet i gitte tilfeller bli justert.

7.3.3. Kontroll mot utvalgsundersøkelser

I enkelte tilfeller fins det utvalgsundersøkelser som spør om den samme informasjonen som finnes i registeret. Hvis denne informasjonen er sammenlignbar, er det mulig å kontrollere registeret mot utvalgsundersøkelsen. Feil i registeret kan da bli identifisert og metoder for å utbedre denne feilen kan utvikles. Når registret er av god nok kvalitet, kan det etter hvert erstatte informasjonen i utvalgsundersøkelsene.

7.4. Sammendrag

Viktige områder ved revisjon av administrative data er:

- God kontakt og samarbeide med dataeier
- God forståelse av formålene med data og definisjonene av variablene
- Avsløre feilkilder og viktige feil i variable
- Oppdage manglende enheter (frafall)
- Kontroll mot andre registre og datakilder

8. IT-revisjonssystemer

IT-revisjonssystem dekker datakontroll/retting i alle ledd fra utfylling av elektroniske skjema, via mottak av skjema, til kontroll/retting av logiske og mulige feil, ved hjelp av IT-systemer som er bygd opp for hvert enkelt statistikkområde. De store revisjonssystemene er ofte laget i Oracle, mens kontroller ofte er laget i SAS. På enkelte statistikkområder med få enheter er også Excel benyttet til

databearbeiding, men dette er mer og mer sjeldent. Bruk av generelle revisjonssystemer kan føre til store besparelser av IT-resurser og effektivisering av revisjonsprosessen.

Store datamengder og/eller kort databearbeidingsperiode krever automatiserte revisjonsrutiner, både innen feilsøking og retting. Personstatistikk, som ofte har store datasett og hvor det ikke er mulig med tilbakekontakt med oppgavegiver, forutsetter stor grad av automatisk revisjon. Det var også der det startet, med Fellegi og Holt 1976 - automatisk avsløring av feil og et system for retting, slik at færrest antall variable endres inntil kontrollreglene tilfredsstilles.

8.1. Elektroniske kontroller av skjema

Mye av datafangsten skjer nå gjennom elektronisk datainnsamling. Det er en målsetting at alle undersøkelser i SSB skal ha et elektronisk innrapporteringsalternativ, men i mange undersøkelser er det mulig å levere skjema eller data på flere måter. I de fleste elektroniske skjemaene, inkludert Blaise som brukes ved telefon- og besøksintervju, er det lagt inn kontroller og forklaringer.

8.1.1. Utfylling og mottak av elektroniske skjemaer

Kontrollene ved utfylling er spesifisert av fagansvarlig. Det kan derfor variere fra skjema til skjema hvor mange kontroller det er. Kontrollene kan være harde, det vil si det er umulig å sende skjemaet før det er rettet opp. Kontrollene kan også være myke, det vil si at det kommer en advarsel om at tallet er unormalt, men den hindrer ikke at skjema kan bli sendt inn. Fordelen med kontroll av skjemaet hos brukerne er at oppgavegiver får mulighet til å rette opp feil direkte. Oppgavegiver vil da ha enkel tilgang til definisjoner og forklaringer som er lagt inn på de enkelte spørsmål. Det blir derfor mindre feil i besvarelsene og mindre belastning med revisjon og kontakt tilbake til oppgavegiver. Slike kontroller kan imidlertid være til irritasjon og forsinkelse for brukeren hvis grunnlaget for at kontrollen slår ut er dårlig. Eventuelt bør det være mulig å "overstyre" kontroller dersom oppgavegiver har vurdert grunnlaget for kontrollen og samme grunnlag fører til nye varsellamper senere i skjemaet.

I mottak av elektroniske skjemaer i SSB er det ofte noen enkle kontroller. Det blir kontrollert om overføringen av skjemaet har gått bra og om hovedvariabler er fylt ut. Ved feil/mangler blir det enten sendt beskjed tilbake til oppgavegiver eller til fagseksjonen om mangelfulle skjema, avhengig av hva slags prosedyrer som er valgt.

8.1.2. Optisk lesing av papirskjema

Optisk lesing (skanning) brukes ved de fleste papirskjemaundersøkelser. Optisk lesing består egentlig av tre prosesser: Selve skanningen (fotografering av skjemasidene), tolkning av det som blir skannet og til slutt verifisering. Verifisering er en manuell bearbeiding/kontroll av det som programmet ikke har greid å tolke eller er usikker på tolkningen av. Programvaren tolker avkryssing, tall og tekst.

I verifiseringen kan en legge inn kontroller. Kontrollene går på: korrekt eller gyldig avkryssing (f.eks. flere kryss i svaralternativer hvor det er tillatt med ett kryss), summering, konsistenskontroller og kontroll på at verdier ligger innenfor et nærmere definert intervall.

Oppgavegiverne har ved hovedutsendingen blitt orientert om at ufullstendig utfylt skjema vil bli sendt tilbake. Vanligvis blir dette gjort i tilfeller der rapporteringen har store og betydelige mangler – gjerne gjennomgående manglende eller feil koding, manglende fordeling av variable på bedrifter osv. Formålet er å få inn mest mulig korrekte opplysninger, samtidig som tilbakesending har en oppdragende effekt på oppgavegiverne.

8.1.3. Manuell registrering og koding av skjema

Det finnes fremdeles noen skjema som ikke er tilpasset optisk lesing. Disse må registreres manuelt i en egen registreringsrutine, eller direkte i et fagsystem. På samme måte som ved optisk lesing, kan en

legge inn kontroller. Det foretas konsistenskontroller innad i et skjema og kontroller mot data fra forrige rapporteringsperiode. En del opplagte feil og mangler blir rettet umiddelbart, mens en del andre feil blir kontrollert mot oppgavegiver eller registereier.

Det meste av kodingen i SSB foregår maskinelt. Av ulike årsaker er det ikke mulig å kode alle enheter (personer) maskinelt. Det er egne rutiner for manuell koding for disse. Ved kodingen benytter en oppslag i andre registre og datakilder, eller en kan hente nøkkelinformasjon fra andre kilder opp i kodeapplikasjonen. Her må en ofte akseptere en større grad av "ubesvart", siden det er problemenhetene som blir lagt ut til manuell koding.

I en del tilfeller koder oppgavegiver selv (f.eks. en del yrkeskoding). I databearbeidingen blir dette kontrollert og rettet med bakgrunn i tidligere rapportering, erfaring og ved å kontakte oppgavegiver på nytt.

8.1.4. Telefon- og besøksintervju med bruk av elektroniske skjemaer (Blaise)

I telefon- og besøksintervju blir elektroniske skjemaer som er programmert i Blaise benyttet. Dette programmet administrerer i tillegg intervjuere og intervjuobjektene. Slik som i andre elektroniske skjemaer kan det bli lagt inn både kontroller og forklaringer på hvert enkelt spørsmål. Kontrollene kan både være harde slik at svaret ikke blir godtatt, eller myke slik at intervjueren bare får en advarsel om merkelig verdi. Det som blir kontrollert, er gyldige verdier, konsistens mot tidligere verdier på skjema, kontroll mot registeropplysninger og kontroll mot tidligere svar i panelundersøkelser. Mengden av kontroller kan variere mye fra undersøkelse til undersøkelse. Det avhenger ofte av resurser og hvor god tid det er til å utføre undersøkelsen. Oppdragsgiver sammen med prosjektledelse (og prosjektgruppe) bestemmer hvilke kontroller som skal være i skjemaet.

8.1.5. Data fra ulike rapporteringskanaler

I stadig flere undersøkelser mottar SSB data gjennom ulike rapporteringskanaler. Det kan være ulike kombinasjoner av telefonintervju, elektroniske skjema, papirskjema og/eller filuttrekk fra administrative systemer. Dataene vil da ha ulik kvalitet. Data fra elektroniske skjemaer har blitt underlagt en del kontroll, mens papirskjemaene ikke har gått gjennom noen form for kontroll. Det er et mål at alle dataene som tilfaller statistikkseksjonene (f.eks. i et fagsystem), har gått gjennom samme kontroll. I elektroniske skjemaer er det logiske kontroller som summeringer og at det ikke er ført alfanumeriske tegn i et numerisk felt. Dette er enkle kontroller som også kan tas i verifisering av skannede papirskjema. Der det er større feil og mangler som opplagt ikke kan rettes etter en instruks, må dataene overføres med feil til statistikkseksjon for videre inspeksjon. I tilfeller der en legger inn vurderingskontroller i elektroniske skjemaer, er dette vanskeligere å følge opp i verifiseringen av papirskjema. Operatørene som arbeider med å verifisere skjema, kan ikke ha samme kunnskap om datagrunnlaget som den som fyller ut skjema.

8.2. Databaser og revisjonssystemer

De fleste statistikker av et visst omfang har et eget datasystem for revisjon. Dette gjelder både skjema som har kommet inn elektronisk eller på papir og administrative data. Funksjonaliteten til disse datasystemene er ofte forskjellige. Slike revisjonssystemer innholder ofte automatiske kontroller og rettinger. I tillegg inneholder det mulighet for manuelle kontroller og rettinger. Det kan også innholde mulighet for revisjon på aggregert nivå, prioritert revisjon, system for tilbakemelding til oppgavegiver og mulighet for revisjon mot tidligere årganger eller mot andre datakilder. Alle slike datasystemer for revisjon bør inneholde et system for dokumentasjon av revisjonsprosessen. Det vil si hvilke kontroller som er slått ut og hva som ble gjort med disse, samt at både originale og reviderte verdier lagres.

Eksempel - Automatisk revisjon av Folke- og boligtellingen 2001
Det er laget et omfattende opplegg for å kontrollere de svarene som er avgitt på boligskjemaene. Opplegget bygger på logiske kontroller og beregning av sannsynlige sammenhenger mellom svarene på forskjellige spørsmål. En skulle f.eks. på spørsmål 18 først oppgi hvilket intervall arealet til boligen lå innenfor og deretter skrive opp det eksakte arealet. Da er det ganske opplagt at det eksakte arealet må ligge i intervallet som en allerede har svart for. Andre kontroller kunne være mer statistiske. Det bør f.eks. være en sammenheng mellom arealet og antall rom i en bolig. Det var flere spørsmål om rommene i boligen: Kjøkken, antall soverom, andre oppholdsrom og bad og WC. Dette er utnyttet til både å kontrollere det totale antallet rom mot arealet og sammenhengen mellom svarene på de fire typene rom i boligen. En har også i noen tilfeller utnyttet andre kilder, f.eks. antall rom ifølge GAB registeret.

Ved partielt frafall ble det for noen spørsmål beregnet svar på grunnlag av svaret på et tilhørende spørsmål (det måtte da være en logisk sammenheng mellom spørsmålene). En har også her utnyttet opplysninger fra GAB registeret i noen tilfeller. Manglende svar på andre spørsmål ble automatisk imputert via "nærmeste nabo" på grunnlag av likhet i svaret på andre relevante spørsmål. En oppsummering av hva som er gjort på datarevisjon i FoB2001 forteller at det meste bygger på såkalt hot deck (samme datakilde) og bruk av nærmeste nabo. I tillegg er det litt cold deck (en annen datakilde).

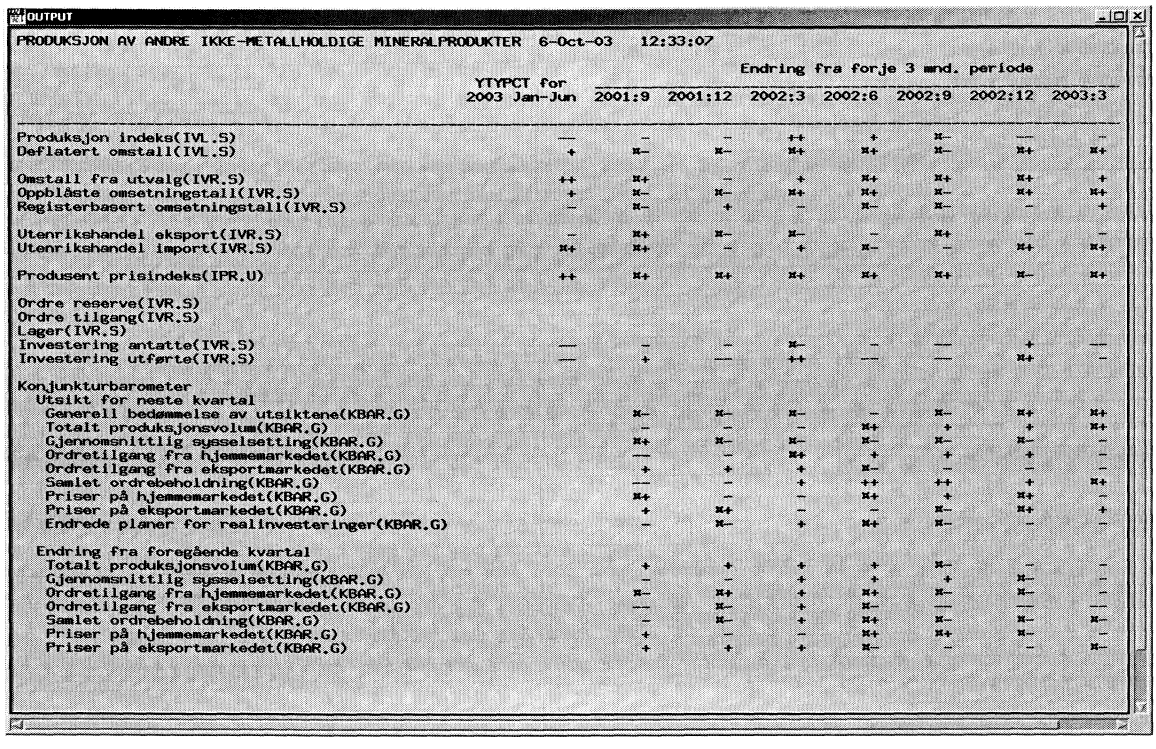
Eksempel - Database for å sammenstille ulike konjunkturindikatorer
Seksjon for økonomiske indikatorer (S240) har hatt behov for å utvikle et verktøy som viser sammenhengen mellom de ulike konjunkturindikatorene. Hovedmål:

- Visualisere sammenhengen mellom ulike statistikker
- Systematisering av tolkning av endringstall som publiseres
- Generere resultater til konjunkturrapport som brukes i arbeidsmøte med kvartalsvis nasjonal regnskap

For å få til dette har det blitt etablert en FAME-database med lik nomenklatur for alle indikatorer. I denne basen er samtlige indikatorer indeksert, og for de kvartalsvise/terminvise indikatorer er seriene fordelt på månedlige tall. Basen oppdateres fortløpende etter hvert som indikatorer blir publisert for en ny periode. I skjermbildet kan bruker styre hvilke tabeller som skal genereres på bakgrunn av valgt næringsaggregat eller konjunkturrapport, periode, frekvens (månedlig eller kvartalsvis endring) og innhold (tall eller symbol) i tabellen.

Dette systemet gir en mulighet for å vurdere resultater fra egen statistikk satt i en større sammenheng. Tabellene kan danne basis for å finne inkonsistens mellom økonomisk teori og statistikk. På den måten kan det gi indikasjoner på serier der det kan ha oppstått feil eller der man bør se nærmere på enkelte næringsaggregater. Tabellene gir også mulighet for en mer objektiv fremstilling av gradering mellom ulike endringstall, f.eks. hva er en stor oppgang, uendret osv. Næringstabellene gir sammenhengen mellom indikatorer fra Konjunkturbarometeret og tilhørende variable fra kvantitativ statistikk. Nedenfor er det gjengitt et eksempel på en næringstabell med symboler. Symbolene er definert på følgende måte:

++ : stor oppgang
\+ : svak oppgang
¤ : ikke signifikant endring
\- : svak nedgang
-- : stor nedgang
Symbolet for uendret, ¤, viser i tillegg fortegnet til endringstallet; ¤- og ¤+



OUTPUT

PRODUKSJON AV ANDRE IKKE-METALLHOLDIGE MINERALPRODUKTER 6-Oct-03 12:33:07

	YTYPCT for 2003 Jan-Jun	Endring fra forje 3 mnd. periode 2001:9	2001:12	2002:3	2002:6	2002:9	2002:12	2003:3
Produksjon indeks(IVL,S)	—	–	–	++	+	¤–	—	–
Deflatert omstall(IVL,S)	+	¤–	¤–	¤+	¤+	¤–	¤+	¤+
Omstall fra utvalg(IVR,S)	++	¤+	–	+	¤+	¤+	¤+	+
Oppblåste omsetningstall(IVR,S)	+	¤–	¤–	¤+	¤+	¤–	¤+	¤+
Registerbasert omsetningstall(IVR,S)	–	¤–	+	–	¤–	¤–	–	+
Utenrikshandel eksport(IVR,S)	–	¤+	¤–	¤–	–	¤+	–	–
Utenrikshandel import(IVR,S)	¤+	¤+	–	+	¤–	–	¤+	¤+
Produsent prisindeks(IPR,U)	++	¤+	¤+	¤+	¤+	¤+	¤–	¤+
Ordre reserve(IVR,S)								
Ordre tilgang(IVR,S)								
Lager(IVR,S)								
Investering antatte(IVR,S)	—	–	–	¤–	–	—	+	—
Investering utførte(IVR,S)	—	+	—	++	—	—	¤+	–
Konjunkturbarometer								
Utsikt for neste kvartal								
Generell bedømmelse av utsiktene(KBAR,G)		¤–	¤–	¤–	–	¤–	¤+	¤+
Totalt produksjonsvolum(KBAR,G)		–	–	–	¤+	+	+	¤+
Gjennomsnittlig sysselsetting(KBAR,G)		¤+	¤–	¤–	¤–	¤–	¤–	–
Ordretilgang fra hjemmemarkedet(KBAR,G)		—	–	¤+	+	+	+	–
Ordretilgang fra eksportmarkedet(KBAR,G)		+	+	+	¤–	–	–	–
Samlet ordrebeholdning(KBAR,G)		—	–	+	++	++	+	¤+
Priser på hjemmemarkedet(KBAR,G)		¤+	–	–	¤+	+	¤+	–
Priser på eksportmarkedet(KBAR,G)		+	¤+	–	–	¤–	¤+	+
Endrede planer for realinvesteringer(KBAR,G)		–	¤–	+	¤+	¤–	–	–
Endring fra foregående kvartal								
Totalt produksjonsvolum(KBAR,G)		+	+	+	+	¤–	–	–
Gjennomsnittlig sysselsetting(KBAR,G)		–	–	+	+	+	¤–	–
Ordretilgang fra hjemmemarkedet(KBAR,G)		¤–	¤+	+	¤+	¤–	¤–	–
Ordretilgang fra eksportmarkedet(KBAR,G)		—	¤–	+	¤–	—	—	—
Samlet ordrebeholdning(KBAR,G)		–	¤–	+	¤+	¤–	¤–	¤–
Priser på hjemmemarkedet(KBAR,G)		+	–	–	¤+	¤+	¤–	–
Priser på eksportmarkedet(KBAR,G)		+	+	+	¤–	–	–	¤–

YTYPCT= Year to year per cent: Endring i forhold til samme periode i fjor for ujusterte tall

$$YTYPCT = \frac{X_t - X_{t-12}}{X_{t-1}} \cdot 100$$

En slik tabell gir et raskt overblikk over situasjonen for en næring, basert på et sett med indikatorer der symbolene er standardisert i henhold til seriens variasjon. (Det synes mer komplisert hvis man ser på endringstallene.)

Tabellen over gir et fullstendig oversikt over industrinæringen NACE 26: Produksjon av andre ikke-metallholdige mineralprodukter. I første del av tabellen oppsummeres de indikatorene som estimeres i S240 pluss to serier fra S270. I den andre delen av tabellen oppsummeres resultatene fra de spørsmålene som hentes kvartalsvis via konjunkturbarometeret.

Når vi ser på resultatene i siste kvartal 2001 (kolonne 2001:12), er det klare sammenhenger. De fleste kvantitative indikatorer viser svak eller ikke signifikant nedgang, og konjunktur-barometeret gir oss det samme bilde. Seriene for investeringer bør tolkes separat. Vi ser klar sammenheng mellom antatte og utførte investeringer (bortsett fra første kvartal 2002 - kolonne 2002:3). Samtidig ser vi en negativ trend for investeringene i den perioden som vises i tabellen.

Ser vi på siste kvartal i 2002 (kolonne 2002:12), ser vi at produksjonsindeksen skiller seg klart fra de øvrige indikatorene. Den viser stor nedgang, mens resten av indikatorene viser svak eller ikke signifikant oppgang.

8.2.1. Utvikling av generelle revisjonssystem

De fleste revisjonssystem bygger på samme typer av kontroller og bearbeiding etter feilmeldinger. Samkjøring av revisjonssystemene vil spare IT-resurser og utnytte utviklede applikasjoner bedre. Dette kan f.eks. være applikasjoner for grafisk revisjon, aggregering av data, tilbakemelding til

oppgavegiver, selektiv revisjon, imputering, flagging m.m. Revisjonssystemer som er knyttet sammen, gjør det enklere bare å ha et sted for revisjon av variable som benyttes i flere statistikkområder.

Utvikling tar tid. Man prøver seg frem innenfor nye områder. Testing og kontroll før metodene tas inn i statistikkproduksjonen er viktig. Dessuten bør det med visse mellomrom vurderes om kontrollene fortsatt er aktuelle eller bør endres. Eksemplet om elevtall i kapittel 6.5.1 er en situasjon der det er naturlig å vurdere om noe kan vinnes på en endring i revisjonsopplegget.

Bruk av generelle revisjonssystemer kan føre til store besparelser av IT-resurser og effektivisering av revisjonsprosessen. Det er laget to generelle revisjonssystemer som de fleste statistikkområder i SSB kan benytte, ett for KOSTRA (KOmmune-STat-RApportering) og ett for IDUN (Informasjons- og DataUtveksling med Næringslivet). Disse systemene er generelle og har derfor blitt tatt i bruk på flere statistikkområder. Felles for disse systemene er at både skjema og revisjonsapplikasjonene blir generert ut fra metadata.

Forutsetninger for at oppbygging av generelle revisjonssystemer skal fungere er:

- Kompetanse ved utarbeiding av revisjonssystem
- De samme personene arbeider innenfor ulike produkter og kan se likheter (statistisk metode og IT)
- Metodene er tilgjengelige for utviklere
- Standard programvare
- Kompetanseoverføring
- Dokumentasjon

Revisjonssystemene som blir bygd opp, bør ha en meldingsdatabase som tar vare på alle rettinger og imputeringer. Et slikt system finnes for populasjonsdatabasene BoF, Besys og SSB-GAB. Et alternativ til dette er logføring av alle endringer som blir gjort under revisjonen. Databasene må også kunne generere historikken i data, det vil si ta vare på gamle verdier når de blir erstattet med nye.

Eksempel -Revisjonsdatabase i KOSTRA

I de elektroniske skjemaene i KOSTRA blir det lagt inn kontroller. Kontrollene blir laget ut fra den enkelte fagseksjons erfaringer fra revisjonsprosessen fra året før. Kontrollene er myke, det vil si at det er mulig å levere skjemaet selv om en kontroll har slått ut og verdien ikke har blitt endret. Når skjemaene blir sendt, får brukere en beskjed om overføringen har gått bra. Ved mottak av skjemaene kontrollers hovedvariable ved rutiner definert av fagansvarlig. Hvis skjemaet ikke blir godkjent, blir denne informasjonen lagt ut i en log på Internett og brukeren må sende inn skjemaet på nytt etter korrigering av feil. Deretter kan skjemaene bli lastet inn i revisjonsdatabasen GenRev, generelt revisjonssystem. Denne basen er en metadatastyrt revisjonsløsning. Det vil si at revisjonsapplikasjonen i hovedsak blir generert automatisk ut fra metadata. Med noen få steg, definisjon av skjema og fagfeltnavn, kan man få et revisjonssystem for et nytt område i drift. Denne basen består av en hovedmodul og noen tilleggsmoduler. Funksjonen til hovedmodulen er innhenting av data, visning av data og revisjon av dataene. I visning av data er det mulig å søke etter enheter, sortere datasettet, vise enheter som er godkjent eller ikke. Kontroller i hovedmodulen er noe som hvert enkelt fagsystem må få definert for seg, og som må programmeres og legges inn i systemet. I tillegg er det laget et eget merknadsfelt der det er mulig å skrive inn kommentarer. Det er også mulig å analysere datasettet i det grafiske verktøyet SAS-insight. Det er nå laget tre tilleggsmoduler til GenRev. En av modulene inneholder aggregering av data slik at det er mulig å studere dataene på flere administrative nivåer. En annen modul er for tilbakemelding til oppgavegiver på e-post. I tillegg er det laget en modul for justering av kontrollene som er lagt inn i Genrev. Dette gjør det enklere for fagpersoner å endre en kontroll hvis de oppdager at grensene de har valgt, er satt feil. Denne

databasen, GenRev, er svært generelt laget og har derfor blitt tatt i bruk på andre statistikkområder enn KOSTRA. I vedlegg 10.4 er noen av skjermbildene i GenRev vist.

Eksempel - Tilbakemeldingssystem i GenRev, KOSTRA
Elektronisk datainnsamling i KOSTRA muliggjør elektronisk tilbakemeldingssystem. Tilbakemeldingssystemet er en forlengelse av GenRev (Generelt Revisjonssystem) og gir saksbehandler på SSB mulighet til på en enkel måte å følge opp datainnsamling med oppfølgingsspørsmål til oppgavegiver.

Oppfølgingsspørsmål blir sendt ut med e-post og inneholder en tekst som er tilpasset det aktuelle fagområdet og undersøkelsen. Videre inneholder det en lenke som tar oppgavegiver til en side på SSBs web-server der oppgavegiver presenteres for de spørsmål SSB har til opprinnelig innsendte opplysninger. For hvert spørsmål tar oppgavegiver stilling til om opprinnelig verdi faktisk var korrekt eller om det var feil. Dersom opprinnelig verdi var korrekt, blir det bedt om en kort forklaring. Ellers må oppgavegiver angi en ny korrekt verdi. Når oppgavegiver er ferdig blir dataene lagret på SSBs webserver og oppgavegiver får en kvittering. Svarene blir gjort tilgjengelig for saksbehandler som benytter dem til å avslutte og godkjenne innrapporteringen.

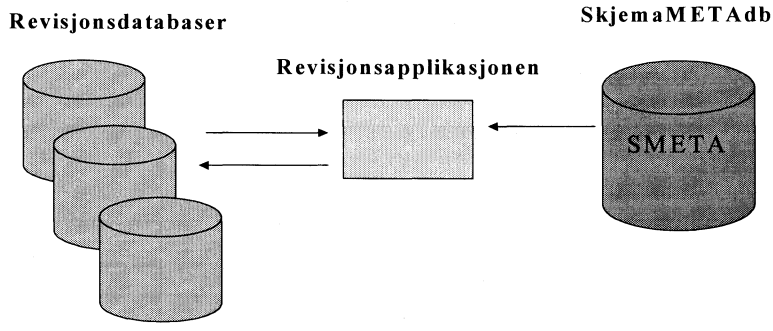
For saksbehandler på SSB gir tilbakemeldingssystemet en god oversikt over alle utestående spørsmål og hvor langt revisjonsprosessen er kommet (f.eks. antall kontroller som har slått ut, antall spørsmål som skal til oppgavegiver, om e-post og eventuelt purring er sendt og om svar er kommet inn, både totalt og pr. oppgavegiver). Purring på tidligere utsendte e-post skjer etter en gitt frist dersom det gjenstår ubesvarte spørsmål til oppgavegiveren.

Eksempel - Revisjonsdatabase tilknyttet IDUN
IDUN er et prosjekt for elektronisk innrapportering av data fra Næringslivet til SSB. Det er et mål at IDUN skal integreres i AltInn. AltInn er et samarbeid mellom Skatteetaten, Brønnøysundregistrene og Statistisk sentralbyrå, som sammen tilbyr elektronisk innlevering av skjemaer for næringslivet. I forbindelse med IDUN er det utviklet et felles revisjonssystem.

Fordelene med dette systemet er:

- Et generelt og fleksibelt system. Selv om skjema eller variabeldefinisjonene endrer seg fra en periode til en annen, trenger ikke dette systemet endringer i databasetabellene eller i revisjonsapplikasjonen.
- Det kreves relativt lite resurser og ikke veldig høy kompetanse innen Oracle for å utvikle en ny revisjonsdatabase klar til produksjon. Grunnen til dette er at samme datamodell og revisjonsskjermbilde blir brukt for alle lokale systemer.
- Felles kontrollsystem for alle revisjonssystemer. Gjenbruk av moduler gir enklere utvikling og lettere vedlikehold.
- Fagseksjonen bestemmer og styrer hvilke kontroller som skal kobles til skjema/variabler samt inputparametrene til kontrollene.
- Trenger ingen endringer i systemet ved ny periode.
- Beholder historisk informasjon/data i databasen.

Grunnlaget for revisjonsdatabasene er SkjemaMETAdatabasen (SMETA) til IDUN. Når skjemaene defineres gjennom SMETA, vil skjemadefinisjoner og kontroller bli lagret sentralt i SMETA. Denne informasjonen blir brukt i revisjonssystem. Lokale revisjonssystem inneholder bare delregisterinformasjon, datatabeller og tabeller som inneholder kontrollresultater. Alle lokale revisjonsbaser har samme struktur, noe som gir mulighet for standardiserte skjermbildemoduler.



Revisjonsbasen består av en kopling mellom enhet, metadatakoder og fakta. Den generelle revisjonsapplikasjonen består av forskjellige typer kontroller. Kontrollene kan bli koplet til skjema på forskjellige nivåer: variabel, spørsmål eller skjema. Kontrollene og resultatene av kontrollene blir vist i egne tabeller. Når kontrolltypen først er laget, kan fagpersonene selv legge til flere slike kontroller eller endre grensene for kontrollene. Det fins mulighet for logføring av endringer som blir gjort i revisjonsdatabasene.

I revisjonsbasen er det utviklet mulighet for følgende type kontroller:

- Logiske kontroller
- Sumkontroller
- Kontroll mot tidligere perioder (år, kvartal, måned)
- Kontroll mot andre databaser

Det er 145 forskjellige skjemaer som kommer inn elektronisk gjennom IDUN, men det er foreløpig bare tre skjemaer som har tatt i bruk revisjonsbasen i IDUN. Dette er Olje og gass, Forskning og utvikling og Utenrikstransaksjoner, utenlandske datterselskaper.

8.3. Sammendrag

Viktige trekk ved IT-revisjonssystemer er:

- Viktige kontrollene bør ligge så nær brukeren som mulig
- Elektroniske kontroller av skjema
- Databaser og revisjonssystemer lagrer dokumentasjon av revisjonsprosessen
- Gjenbruk gjennom utvikling av Generelle revisjonssystemer

9. Måling av effekter av revisjon

Basert på flagging under revisjonsprosessen og originale og reviderte verdier (eventuelt fra ulike stadier av revisjonsprosessen), kan det lages automatiske prosedyrer for analyser av hva som skjer under revisjonen. Temaene kan være:

- Hvilke feilkontroller avslører flest feil?
- Hvilke feilkontroller fører til flest manuelle kontroller?
- Hvor stor andel av de manuelle kontrollene fører til endring - avhengig av feilmarkering?

- Hvilke feilkontroller avslører de største feilene?
- Virkninger av rettinger - avhengig av feilmarkering.
- Hvilke variable rettes oftest?

9.1. Indikatorer i revisjonsprosessen

Revisjonsprosessen bør være under jevnlig evaluering, både for å vurdere om ressurser til revisjon står i rimelig forhold til effekten av revisjon og for å vurdere om nye metoder eller teknikker vil kunne gi bedre avkastning. Endringer i selve undersøkelsen, inkludert tilgang til administrative data, utvikling av statistiske metoder og nye programverktøy er viktige grunner for å vurdere nye rutiner.

Flaggingsdata og statistikk basert på flagging er nødvendig for en god evaluering av revisjonsprosessen. Flaggingsdata/statistikk kan brukes både ved overvåking over tid og analyse ved endringer. Indikatorer kan brukes som kvantitative mål for revisjonsprosessen. Beregning av indikatorer kan enten bli gjort ved hjelp av flagging og/eller ved annen god dokumentasjon av revisjonsprosessen.

Sentrale mål er kostnadene til revisjon, effekten av revisjon på sluttresultatet og treffsikkerheten i kontrollopplegg. Men mulighetene er mange, og det kan lages indikatorer tilpasset det enkelte statistikkprodukt.

9.1.1. Kostnadsindikatorer

Innsamling av data og bearbeiding av data er separate aktivitetskoder i Produktregisteret. Dette kan utnyttes til beregning av kostnader knyttet til revisjon innen de enkelte statistikkprodukter.

- Timeverk brukt på revisjon
  - utførte timeverk til revisjon i alt
  - utførte timeverk som andel av totalt utførte timeverk på statistikkproduktet
  - utførte timeverk pr. oppgaveenhet

Grensedragningen mellom de ulike aktivitetene kan være uklar, men ikke vanskeligere enn at saksbehandlernes egne anslag bør være tilstrekkelig for dette formålet.

9.1.2. Frekvensindikatorer

Et mål for effektiv revisjon er at kontrollmetodene er målrettet, det vil si at flest mulig av feilene oppdages og færrest mulig av akseptable enheter/verdier belaster revisjonsressursene videre. Et mål for dette har vi ved indikatorene:

- Markeringsfrekvens
- Treffsikkerhet
- Endringsfrekvens

Markeringsfrekvens angir hvor ofte feil eller mistenkelige verdier markeres. Hvis n er totalt antall enheter og a er antall markeringer, blir markeringsfrekvensen I_M lik:

$$I_M = \frac{a}{n}$$

Treffsikkerhet angir hvor stor andel av mistenkelige feil blir rettet. La antall rettinger være b. Da er treffsikkerheten lik:

$$I_T = \frac{b}{a}$$

Endringsfrekvens angir hvor stor andel som rettes, og vil være lik produktet av de to andre frekvensene.

$$I_E = \frac{b}{n}$$

Hver av disse indikatorene kan beregnes for hver

- kontrollmetode - hvor effektive er de enkelte kontrollmetodene?
- variabel - er det ekstra mye feil på enkelte variable?
- enhet - dårlige rapportører?
- ledd i revisjonsprosessen - automatisk retting, rettet manuelt, rettet av oppgavegiver etc

Kontrollmetode og variable henger ofte sammen. Hver kontrollmetode kan gjelde en variabel eller en gruppe variable. Det kan være en eller flere kontrollmetoder for hver variabel. Mange feil fra en oppgavegiver /stor andel av mange variable kan tyde på missforståelser og bør følges opp. Treffsikkerhet for manuelle kontroller kan være av interesse for ressursfordeling. Tilbakemelding til oppgavegiver er også ressurskrevende. Hvis oppgavegiver blir kontaktet og dette blir dokumentert, er det mulig å lage indikatorer for tilbakemeldingsfrekvens og treffsikkerhet på tilbakemeldingen.

Det kan fort bli mange indikatorer hvis det er mange kontroller og mange variable. Grafikk kan brukes til å analysere hvilke indikatorer som skiller seg ut.

Kontrollmetoder

Markeringsfrekvensen forteller hvor mange enheter i undersøkelsen en bestemt kontroll slår ut for, mens treffsikkerheten forteller hvor mange opprettinger som blir foretatt på grunnlag av feilmeldinger fra denne kontrollen.

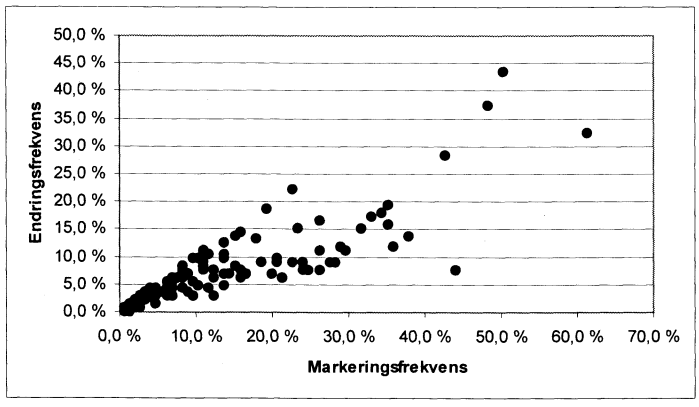
"En effektiv kontroll har en høy treffsikkerhet, kombinert med relativt lav markeringsfrekvens."

Høy treffsikkerhet kombinert med høy markeringsfrekvens, tyder på at det er noe galt med variabelen. Det bør da bli undersøkt om noe kan bli endret slik at denne feilen ikke oppstår ved den neste undersøkelsen.

Lav treffsikkerhet indikerer at kontrollen ikke er særlig effektiv, spesielt hvis markeringsfrekvensen er relativt høy. Det kan tyde på at mulighetsområdet (min/max-verdier) er definert for snevert slik at for mange enheter faller ut på kontrollen.

> ***Eksempel på indikatorer - markerings- og endringsfrekvens fra KOSTRA***
> I skjema 44 (Voksenpsykiatriske institusjoner) i KOSTRA er indikatorene markeringsfrekvens, endringsfrekvens og treffsikkerhet laget. Figur 9.1.1 viser plott av samhørende verdier for markeringsfrekvens og endringsfrekvens for hver enkelt kontroll av variable fra dette skjemaet. Punktene lengst unna diagonalen har minst treffsikkerhet.

Figur 9.1.1. Markeringsfrekvens mot endringsfrekvens. Kontroller



Tabell 9.1.1 viser verdiene på indikatorene markeringsfrekvens, endringsfrekvens og treffsikkerhet for noen utvalgte kontroller.

Tabell 9.1.1. Indikatorer i revisjonsprosessen for KOSTRA. Utvalgte kontroller

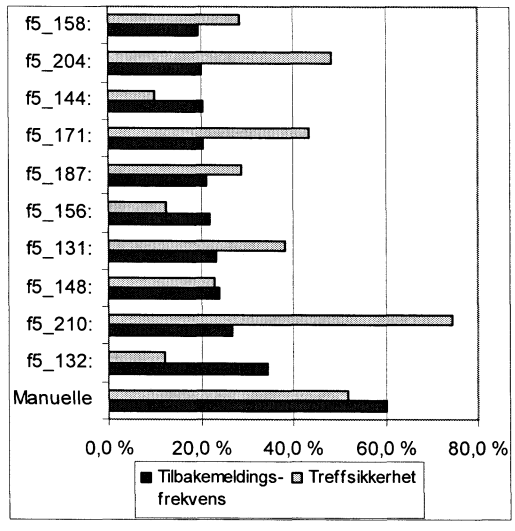
Kontroll		Markerings-frekvens	Endrings-frekvens	Treffsikkerhet
id	type			
f5_204:	F - logisk kontroll innen skjema	48,3 %	37,2 %	77,1 %
f5_210:	F - logisk kontroll innen skjema	50,3 %	43,4 %	86,3 %
f5_132:	G - kontroll mot fjorårets tall	44,1 %	7,6 %	17,2 %
f5_213:	E - svar må oppgis (obligatorisk)	11,0 %	11,0 %	100,0 %
f5_025:	D - sumkontroll	1,4 %	1,4 %	100,0 %

Indikatorene viser både kontroller på enkeltposter i skjemaet og kontroller av sammenhenger. Tolkning av indikatorene:

- Kontroll f5_204 - kontroll av logisk sammenheng i skjema, slår ut for omtrent halvparten av skjemaene, og 37 prosent av variablene blir rettet opp. Hele 77 prosent av kontrollene fører til en endring. Denne kontrollen er god, men grunnen til den høye andelen med feil burde kanskje bli kontrollert.
- Kontroll f5_210 - kontroll av logisk sammenheng i skjema, slår ut også for omtrent halvparten av skjemaene, og 43 prosent av variablene blir endret. Denne kontrollen har en treffsikkerhet på 86 prosent. Denne kontrollen er god og det er ikke nødvendig med en endring av den. Det bør derimot bli vurdert om skjema skal bli endret slik at det kanskje blir færre opprettinger.
- Kontroll f5_132 - endring fra fjorårets tall, slår ut for 44 prosent av skjemaene, men det fører bare til endring av 8 prosent av tilfellene. Denne kontrollen burde bli endret slik at treffsikkerheten ble større enn 17 prosent.
- Kontroll f5_213 - svar mangler og må oppgis er en absolutt kontroll. Det vil si at for hver gang kontrollen slår ut, fører det til en endring. Det er i tillegg forholdsvis få ganger kontrollen har slått ut.
- Kontroll f5_025 er en sumkontroll hvor alle feil må rettes. Den slår ut for 1,4 prosent av skjemaene.

Eksempel på tilbakemeldingsfrekvens og treffsikkerhet
KOSTRA har et system for tilbakemelding til oppgavegiver (se eksempel kapittel 8.2.1). Figur 9.1.2 viser tilbakemeldingsfrekvenser og treffsikkerhet fra utvalgte kontroller av KOSTRA-skjema 44 (Voksenpsykiatriske institusjoner).

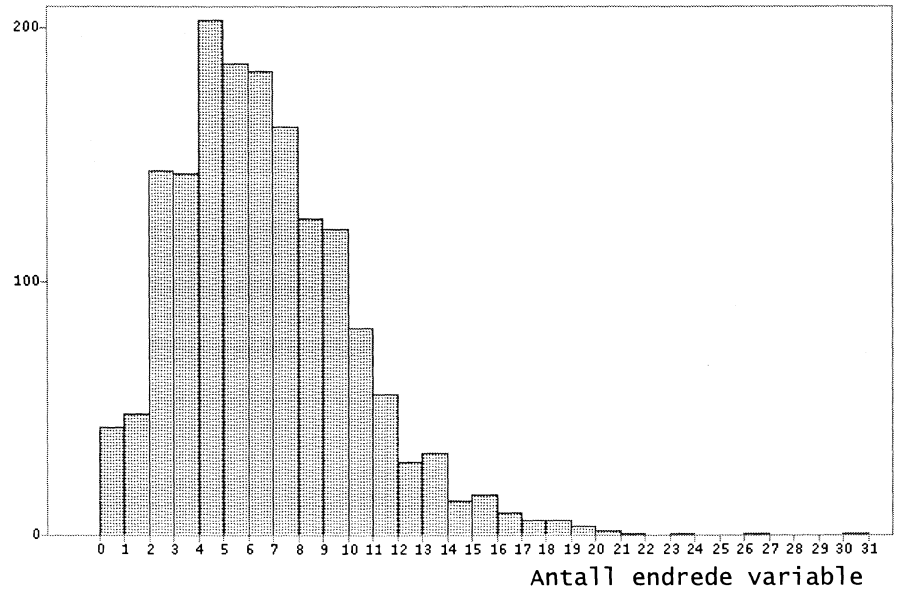
Figur 9.1.2. Tilbakemeldingsfrekvens og treffsikkerhet av tilbakemeldinger



Kontrollene vist her, er de 11 kontrollene som hadde størst andel med tilbakemeldinger til oppgavegiver. De fleste av disse gjelder endring fra fjoråret, bare 2 gjelder logiske kontroller innen skjema. Sammen med kontrollen er det også vist treffsikkerheten av kontrollen. Kontroll f5_132 - stor endring fra fjorårets tall, førte til tilbakemelding på i 35 prosent av tilfellene. Men det var bare 11 prosent av disse som fikk endret verdi. Kontroll f5_210 - logisk kontroll innen skjema, var det 27 prosent som fikk en tilbakemelding på, og det førte til en endring i 75 prosent av tilfellene.

Eksempel på antall endringer pr. enhet
Spørreskjemaet i industristatistikken - struktur ble rettet for en stor andel av oppgavegiverne. Skjemaet inneholder 90 variable, men det var vanligvis få variable som ble rettet på hvert skjema. Figur 9.1.3 viser fordelingen av antall variable rettet pr. bedrift.

Figur 9.1.3. Poster endret under revisjon



9.1.3. Verdiindikatorer

Størrelsen på endringer under revisjon kan være et bedre mål for effekten av revisjon enn antall endringer, f.eks. for økonomiske variable. Her kan vi for hver variabel beregne indikatorene:

- Endringsandel

Endringsandel angir hvor mye som endres i forhold til totaltall (eventuelt feilmarkeringer). Hvis X er total verdi for en variabel og x er endret verdi, blir endringsandelen I_V lik:

$$I_V = \frac{x}{X}$$

Indikatorene kan beregnes for hver

- kontrollmetode - hvor effektive er de enkelte kontrollmetodene
- ledd i revisjonsprosessen - automatisk retting, rettet manuelt, rettet av oppgavegiver etc.

9.2. Effekt av revisjon

Vi bruker betegnelsen statistikkdata for de endelig reviderte data, siden dette er data som de publiserte statistikkene er basert på. Når man skal se på effekten av revisjon, er det viktig å være klar over at statistikkdata ikke nødvendigvis er de korrekte verdiene (fasit).

9.2.1. Sammenligne statistikkdata og rådata for numeriske variable

For numeriske variable er det bl.a. interessant å se på endring pr enhet:

Statistikkdata - Rådata

Hvis denne er positiv, betyr det at den opprinnelige verdien til enheten er endret til en høyere verdi. Hvis den er negativ, betyr det at den opprinnelige verdien er endret til en lavere verdi. En klar overvekt av enten positive eller negative endringer, når vi ser på alle enhetene, kan tyde på skjevheter enten i det innkomne datamaterialet (f.eks. avvik i forhold til definisjon) eller i kontroll- og opprettingsrutinene.

Størrelsen på akseptable endringer av en variabel kan variere med verdien. Det gjør det interessant å se på prosentvis endring pr enhet:

$$\frac{|\text{Statistikkdata - Rådata}|}{\text{Statistikkdata}} * 100$$

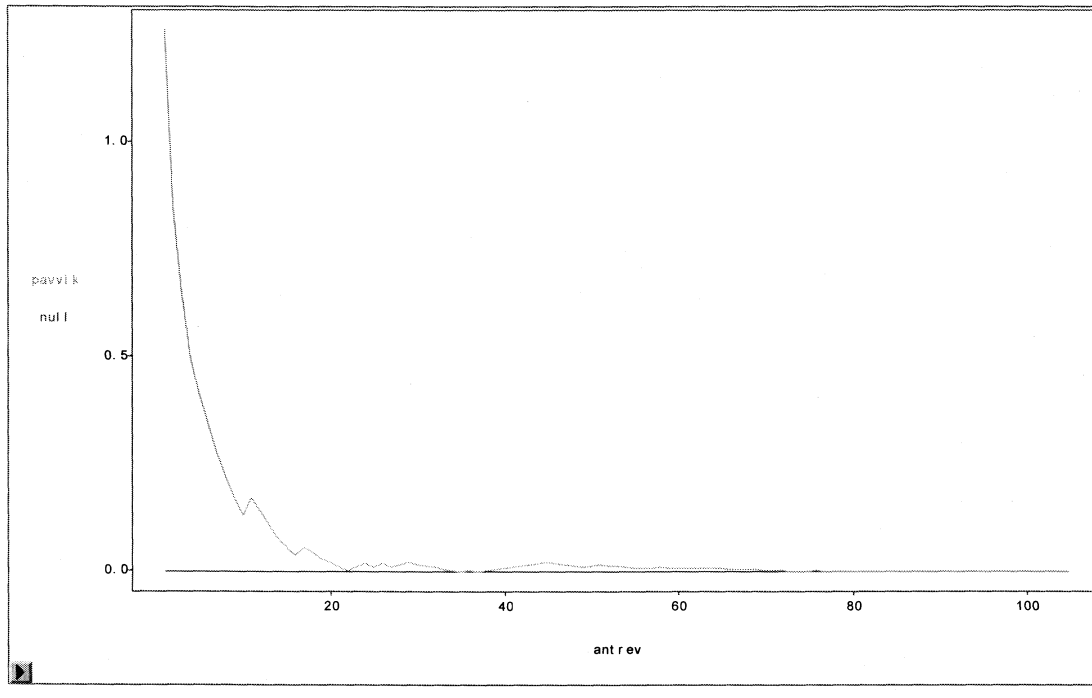
En effektiv metode for å studere enkeltobservasjoners betydning for totalresultatet er å lage en graf av utviklingen etter hvert som én og én enhet rettes, med de største absoluttendringene først. Hvis det er mange numeriske variable, kan det bli for omfattende å gjøre dette for alle variablene. Da kan man velge ut de variablene som er viktigst.

Prosentvis avvik fra revidert total vil sees som en kurve som starter med y-verdi lik total endring (%), og ender ved y-verdi = 0. En bratt kurve med avflating betyr at noen få enheter betyr mye for totalresultatet. Det er viktig at disse enhetene fanges opp tidlig i kontrollrutinene og rettes. Kurve med 45° helling betyr at alle opprettede enheter betyr like mye. Dette er sjelden tilfelle i praksis.

Eksempel fra revisjon av utenrikshandelsdata

Det er i alt 2 575 varelinjer som er merket med en feilkode en gitt måned. For 105 av disse blir verdien endret. Figur 9.2.1, hvor varelinjene er sortert etter avtagende absoluttendring, viser at når de 20 største endringene er rettet opp, er avviket til total verdi svært lite.

Figur 9.2.1. Avvik (i %) fra revidert total verdi etter hvert som varelinjer revideres, når varelinjer med størst endring i verdi revideres først



9.2.2. Sammenligne statistikkdata og rådata for kategoriske variable

Med en kategorisk variabel menes en variabel som bare kan ha et begrenset antall verdier, og hvor hver verdi svarer til en bestemt egenskap. Det kan være klare mønstre i endringer av kategoriske variable.

Eksempel - varenummerendring fra utenrikshandel

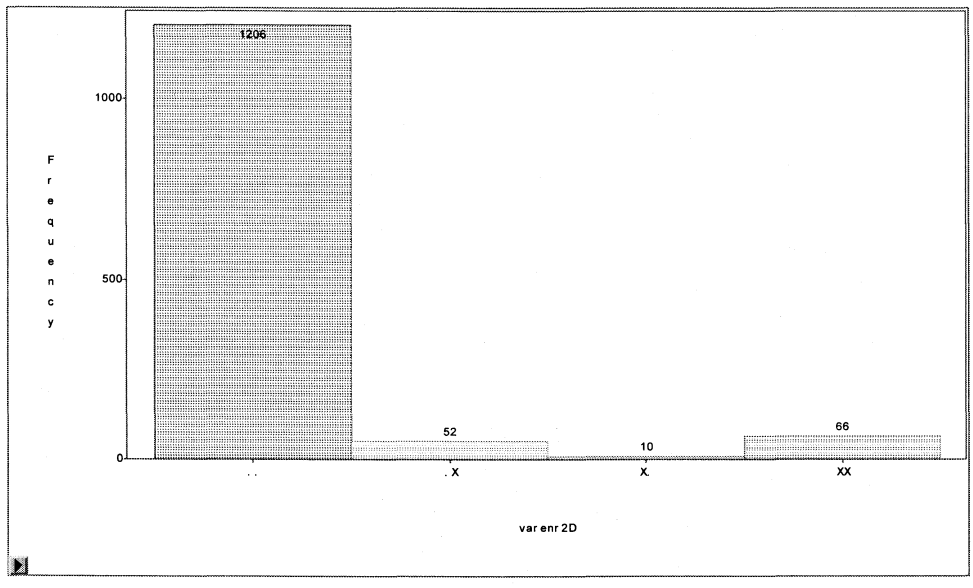
Varenummer er den variabelen som endres flest ganger. En måned ble varenummer endret for i alt 1 334 varelinjer av totalt 810 000 varelinjer. De mest omfattende endringer som innebærer endring i de 2 første sifrene i varenummeret, forekom derimot bare for 128 varelinjer. Antall endringer etter hvor mange av de første sifrene som ble endret, angis i tabellen nedenfor.

Tabell 9.2.1. Antall varelinjer med endring av varenummer

Antall siffer	Antall varelinjer i alt
2	128
4 (inkludert de 2 første)	342
6 (inkludert de 4 første)	768
Alle	1 334

Figur 9.2.2 viser fordeling av varenummerendringer for de 1334 varelinjene som har fått endret varenummer under revisjon. Endring i de 2 første sifrene er markert med 'XX' (begge sifrene), '.X ' (2. siffer) og 'X.' (1. siffer). De fleste endringene gjelder bare de 6 siste sifrene, markert med '..'.

Figur 9.2.2. Varenummerendring - 2 siffer.



Eksempel - bestemmelsesland fra utenrikshandel
Det er klare mønstre i retting av bestemmelsesland i utenrikshandelsstatistikken. De vanligste endringene en gitt måned vises i tabell 9.2.2.

Tabell 9.2.2. Bestemmelsesland endres oftest slik

Bestemmelsesland originalt		Bestemmelsesland rettet		Antall varelinjer
Landkode	Land	Landkode	Land	
YU	Jugoslavia	CS	Serbia og Montenegro	16
KP	Nord-Korea	KR	Sør-Korea	4
DM	Dominicia	DO	Dominikanske Rep.	2
SL	Sierra Leone	SK	Slovakia	2

9.2.3. Parallell revisjon og koding

En teknikk som kan brukes for kvalitetssikring, er å la to grupper revidere og kode det samme materialet enten ved å bruke de samme metoder eller to forskjellige metoder. Slike gjentakelser av prosessen gjøres meget sjeldent. Det skyldes at kostnadene er høye ved slikt dobbeltarbeid, både i form av tid og ressurser. Dette kan derfor ikke anbefales som en permanent ordning, men kan være aktuelt som enkeltstående tiltak for å teste ut revisjonsarbeidet.

I forbindelse med store tellinger har denne teknikken blitt brukt under navnet "acceptance sampling". En kontrollerer her arbeidet løpende ved å trekke et tilfeldig utvalg av et kodet materiale og foretar kodingen på nytt. Dersom de feilkodinger en finner utgjør en andel som er større enn en på forhånd bestemt prosentandel, legges alt tilbake til omkoding. Størrelsen på utvalget, samt hvor mye feil en skal akseptere, bestemmer den endelige kvaliteten på kodingen. Ofte starter en opp med høye ambisjoner for deretter å redusere disse for å få arbeidet gjort innen den fastsatte tidsramme.

I tillegg til å kontrollere kvaliteten til en bestemt undersøkelse gir slike opplegg betydelig innsikt i kodeprosessen og kvaliteten til kodeinstruksen og opplæring av koderne. Slik innsikt er også nyttig for opplegget av andre statistiske undersøkelser.

9.2.4. SAS-program for sammenligning av originale og reviderte data

Det er lagd fem SAS-programmer som kan brukes til å sammenligne reviderte data med originale. Programmene kan lett tilpasses ulike datasett (SAS- data originalt og revidert fra samme undersøkelse) og er tilgjengelig for alle. Programmene heter sammenligning.sas, perEnhet.sas, perVariabel.sas, kategoriskVariabel.sas og numeriskVariabel.sas. Programmene er lagt ut på fellesområdet på Unix: $FELLES/sasprog/sammenligning og dokumenteres på Byrånettet under Faglig, IT, Programvare, SAS, Nyttige SAS-programmer. Programmene ligger også sammen med programbeskrivelsen på området Q:\DOK\Revprosj\program\SASprogram.

Programmene gir

- en kort rapport om likheter og ulikheter mellom det originale og det reviderte datasettet
- antall endringer som er gjort
    - per enhet
    - per variabel
- grafikk generert automatisk
- datasett koblet for videre analyse
- egne analyser for kategoriske og numeriske variable

For en mer detaljert beskrivelse av programmene med eksempler, se vedlegg10.1.

10. Vedlegg

10.1. SAS-programmer for sammenligning av originale og reviderte data

På Unix: $FELLES/sasprog/sammenligning ligger det fem SAS-program som kan brukes til å sammenligne reviderte data med originale. En grundig beskrivelse av programmene finnes på Byrånettet under Faglig, IT, Programvare, SAS, Nyttige SAS-programmer. Samme program og beskrivelse ligger på området Q:\DOK\Revprosj\program\SASprogram. Programmene heter:

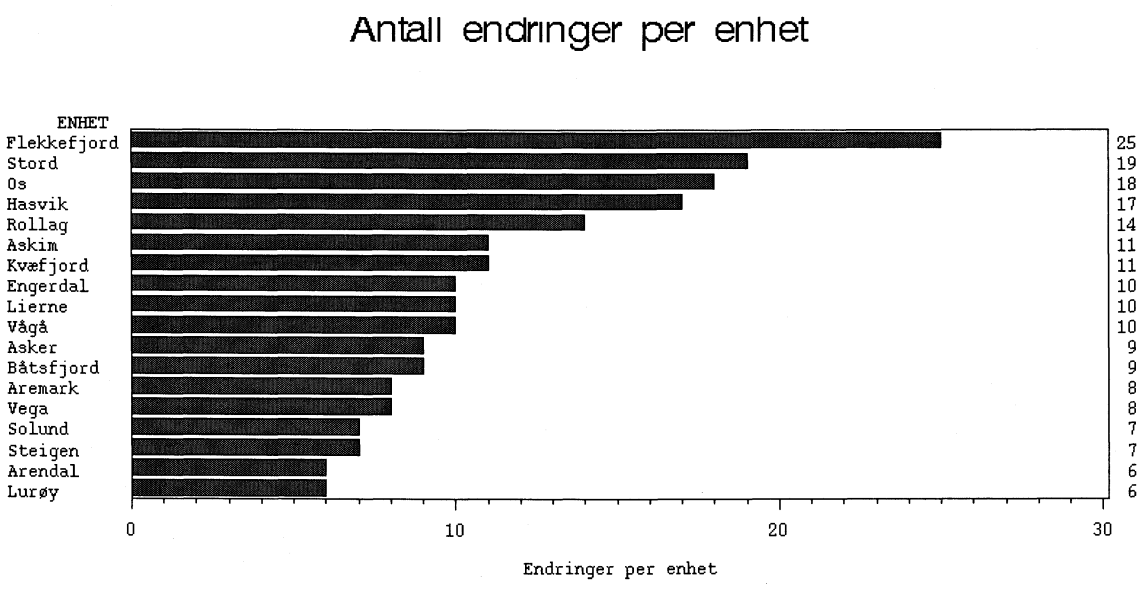
- sammenligning.sas
- perEnhet.sas
- perVariabel.sas
- kategoriskVariabel.sas
- numeriskVariabel.sas

Programmet sammenligning.sas

Dette programmet lager en kort rapport over likheter og ulikheter mellom det originale og det reviderte datasettet. I tillegg lages de fire datasettene "org_og_revdata", "statistikk", "kun_orgdata" og "kun_revdata". Datasettet "org_og_revdata" er en kobling av originale og reviderte data, slik at originale og reviderte verdier er gitt i samme datasett. Datasettet er egnet som utgangspunkt for videre sammenligning i SAS/INSIGHT. Datasettet "statistikk" gir deskriptiv statistikk for de numeriske variablene i "org_og_revdata". Datasettene "kun_orgdata" og "kun_revdata" består av eventuelle enheter som kun fins i henholdsvis originalt eller revidert datasett.

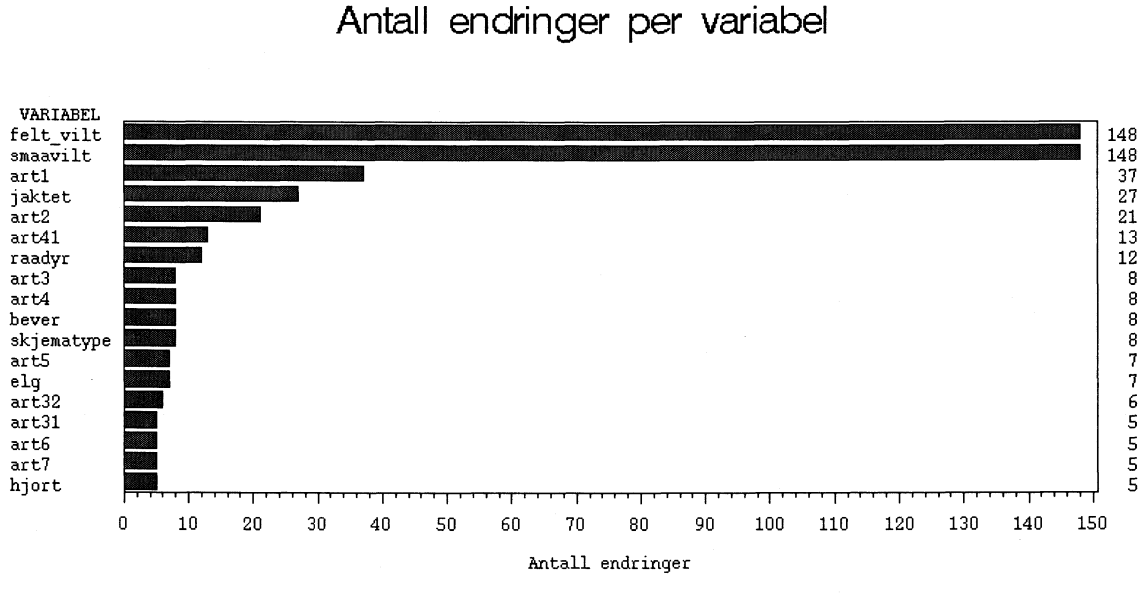
Programmet perEnhet.sas
Programmet teller opp hvor mange endringer som er gjort per enhet, dvs. hvor mange av variablene til en enhet som er endret. Det opprettes et datasett "perEnhet" med disse tallene. Programmet lager også et stolpediagram som visert antall endringer til de 18 enhetene med flest endringer. Figur 10.1.1 viser et slikt stolpediagram for en KOSTRA-undersøkelse (enhet er kommune).

Figur 10.1.1. Eks. på stolpediagrammet som lages av programmet perEnhet.sas



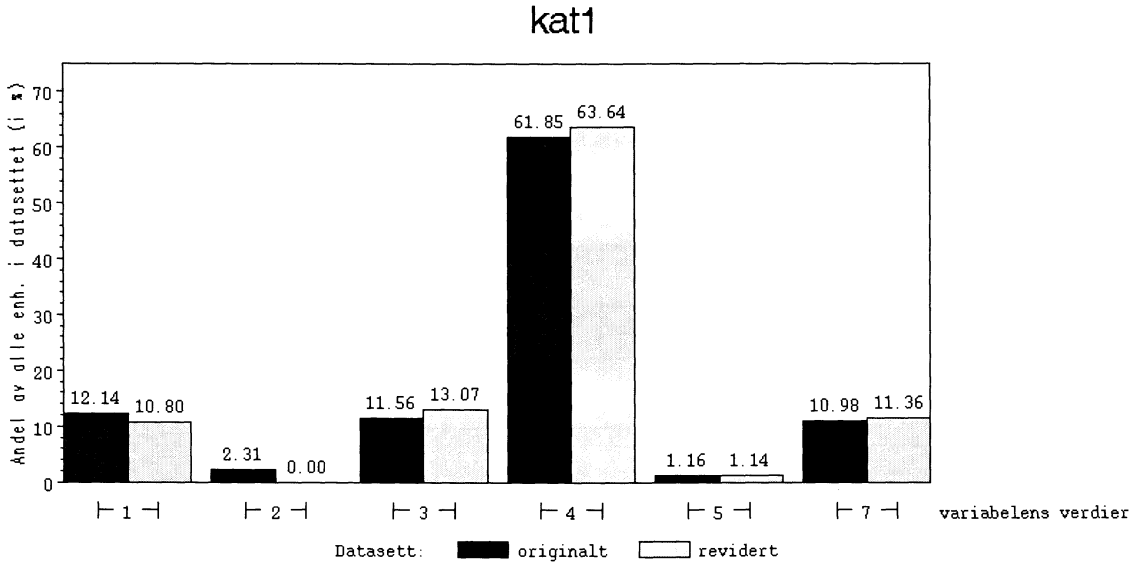
Programmet perVariabel.sas
Mens programmet perEnhet.sas teller opp antall endringer per enhet, teller dette programmet antall endringer per variabel. Dvs. programmet teller opp hvor mange enheter som har fått endret en bestemt variabel. Det opprettes et datasett "perVariabel" med disse tallene. Programmet lager også et stolpediagram som viser antall endringer til de 18 variablene med flest endringer. I figur 10.1.2 vises et slikt stolpediagram fra småviltjakten.

Figur 10.1.2. Eks. på stolpediagrammet som lages av programmet perVariabel.sas.



Programmet kategoriskVariabel.sas
Dette programmet gir en grafisk fremstilling av verdiene til en kategorisk variabel før og etter revidering. Programmet beregner andelen enheter som har de forskjellige verdiene før revidering, og andelen enheter som har de forskjellige verdiene etter revidering. Disse andelene presenteres så i et stolpediagram. I figur 10.1.3 vises stolpediagrammet for en variabel som heter kat1. Variabelen kat1 er en kategorisk variabel som kan ha verdiene 1, 2, 3, 4, 5, 6 og 7. Disse verdiene er markert langs den horisontale aksen i stolpediagrammet, og for hver av verdiene er det to stolper. Den venstre av de to stolpene viser andelen enheter som har den aktuelle verdien før revidering, mens den høyre stolpen viser den tilsvarende andelen etter revidering.

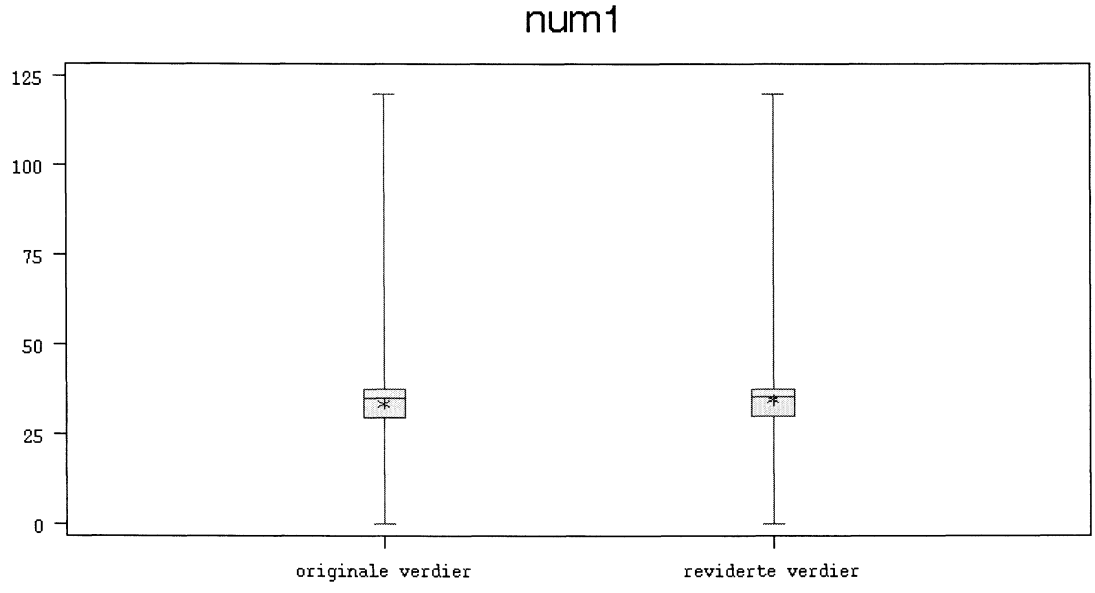
Figur 10.1.3. Illustrasjon av stolpediagrammet som lages av programmet kategoriskVariabel.sas



Programmet numeriskVariabel.sas
Dette programmet gir en grafisk fremstilling av reviderte og ureviderte verdier for en numerisk variabel. Følgende blir presentert:
- Et boksplott over ureviderte verdier og et boksplott over reviderte verdier
- Et xy-plott av reviderte verdier mot ureviderte verdier

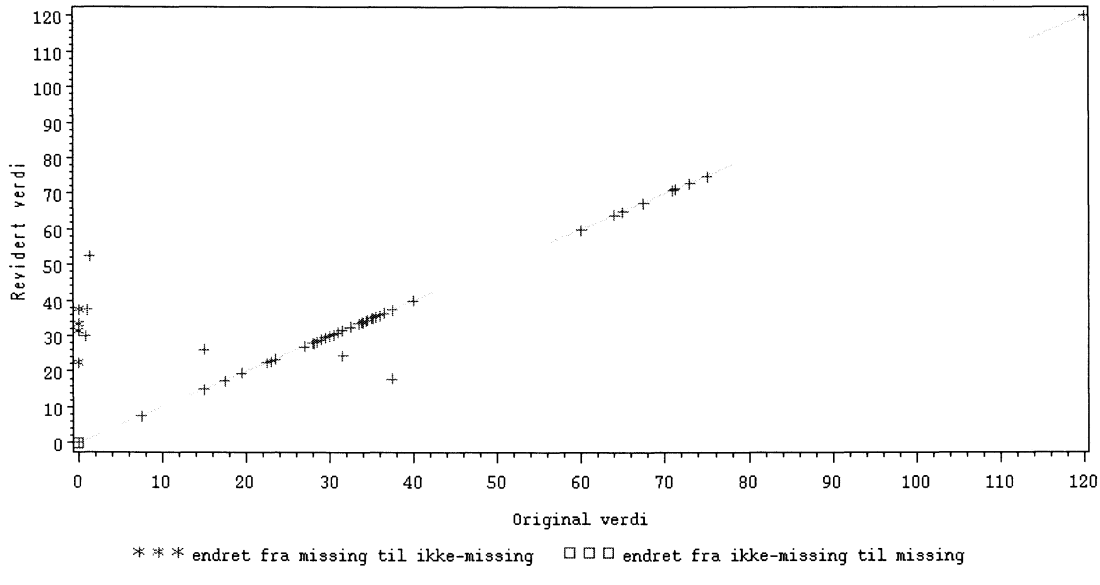
I figur 10.1.4 vises boksplottene for en variabel som heter num1. I boksplottene er følgende markert: Største og minste verdi, øvre og nedre kvartil, gjennomsnittsverdi (markert som stjerne) og median (markert som horisontal strek i den grå boksen).

Figur 10.1.4. Illustrasjon av boksplottet som lages av programmet numeriskVariabel.sas



Figur 10.1.5 viser xy-plott av originale og reviderte verdier for samme variabel. Originale verdiene vises langs x-aksen (horisontale aksen) og de reviderte verdiene langs y-aksen (vertikale aksen). Den heltrukne linjen indikerer hvor revidert verdi er lik original verdi, dvs. at enheter som er markert langs denne linjen er uendret (for den aktuelle variabelen).

Figur 10.1.5. Illustrasjon av xy-plottet som lages av programmet numeriskVariabel.sas



10.2. Kort presentasjon av Hidiroglou-Berthelot metoden

Anta at variabelen $X_i(t)$ skal kontrolleres, der *i* er objekt og *t* er periode. Gitt data fra to påfølgende perioder:

$(X_i(t-1), X_i(t)) \quad i = 1,2,...,n$

Da er forandrings-/endringskvoten gitt ved:

$$R_i = \frac{X_i(t)}{X_i(t-1)}$$

HB-metoden innebærer at testvariabelen transformeres to ganger, og at et akseptintervall for den endelig transformerte testvariabelen beregnes ut fra de data som skal granskes. Robuste parametere som median og kvartilavstand sammen med transformeringen gjør at akseptintervallet påvirkes lite av feilaktige data og den naturlige variasjonen i forandringskvoter.

Symmetritransformasjonen

For at det skal være samme mulighet for å avsløre økning og reduksjon og behandle på samme måte, utføres følgende symmetritransformasjon:

$$S_i = \begin{cases} 1 - R_{median} / R_i, & 0 < R_i < R_{median} \\ R_i / R_{median} - 1, & R_i \geq R_{median} \end{cases},$$

der R_{median} er medianverdien til R_i.

Halvparten av S-ene blir positive, mens den andre halvparten blir negative.

(I KPI benyttes en enklere symmetritransformasjon: $S_i = \begin{cases} 1 - 1/R_i, & 0 < R_i < 1 \\ R_i - 1, & R_i \geq 1 \end{cases}$. Hvis $R_{median} = 1$, er de to variantene identiske.)

Størrelsestransformasjon

$$E_i = S_i * (MAX(X_i(t-1), X_i(t)))^U, \quad 0 \leq U \leq 1$$

Denne transformasjonen gjør akseptintervallet mer følsomt for numerisk store verdier på variabelen. Følsomheten er avhengig av størrelsen på *U*. Dersom *U=0*, vil termen $(MAX(X_i(t-1), X_i(t)))^U$ bli lik 1, og dermed legges ingen vekt på verdien av variabelen. *U=1* vil derimot gjøre at det legges maksimal vekt på verdien, mens man med en verdi mellom *0* og *1* legger noe mindre vekt på verdien.

Akseptgrenser

HB-metoden foreskriver følgende akseptgrenser:

$$\begin{aligned} D_{Q1} &= MAX(E_{median} - E_{Q1}, |A * E_{median}|) \\ D_{Q3} &= MAX(E_{Q3} - E_{median}, |A * E_{median}|) \end{aligned},$$

der indeks Q1 og Q3 står for nedre og øvre kvartil, og A er en konstant som normalt settes til 0,05. Grunnen til at termen $A * E_{median}$ inkluderes, er at en ønsker å unngå vanskeligheter når kvartilavstandene er veldig små, dvs. når E_i er konsentrert rundt en verdi. Dette kan nemlig innebære at selv små avvik blir klassifisert som utliggere.

Nedre grense: $E_{median} - C * D_{Q1}$

Øvre grense: $E_{median} + C * D_{Q3}$

C er en konstant, som sammen med parameteren U bestemmes ut fra tester på innsamlede data.

Det er verdt å merke seg at parameteren A vil bli satt ut av kraft hvis $E_{median} = 0$. Dette vil alltid skje når medianversjonen av symmetritransformasjonen benyttes for et observasjonssett med et odde antall observasjoner, men kan også forekomme for et like antall observasjoner og for den enklere versjonen av symmetritransformasjonen. En måte å unngå dette på er å erstatte $|A * E_{median}|$ med A i uttrykkene for D_{Q1} og D_{Q3}, tilsvarende beskrivelsen ovenfor av kvartilmetoden. Parametrene A må fastsettes særskilt for hvert enkelt produkt.

10.3. Poengfunktionen DIFF

Poengfunksjonen DIFF er utviklet av Latouche och Berthelot (1992). Metoden brukes av Statistics Canada m.fl.

La

$y_{i,k,t}$ være verdien for enhet i (i = 1, ... ,I) og variabel k (k = 1, ... ,K) ved tidspunkt t

$y'_{i,k,t-1}$ være verdien for enhet i (i = 1, ... ,I) og variabel k (k = 1, ... ,K) ved tidspunkt t-1

$w_{i,t}$ være vekten for enhet i (i = 1, ... ,I) ved tidspunkt t

P_k være relativ viktighet for variabel k

$$Z_{i,k,t} = \begin{cases} 1 & \text{hvis variabel k feilmarkeres av en eller flere kontroller} \\ 0 & \text{ellers} \end{cases}$$

$\hat{Y}_{d,k,t-1}$ være et estimat av totalen for gruppe d (der undersøkningsenhet i inngår) for variabel k, ved tid t-1.

Poengfunksjonen DIFF blir da:

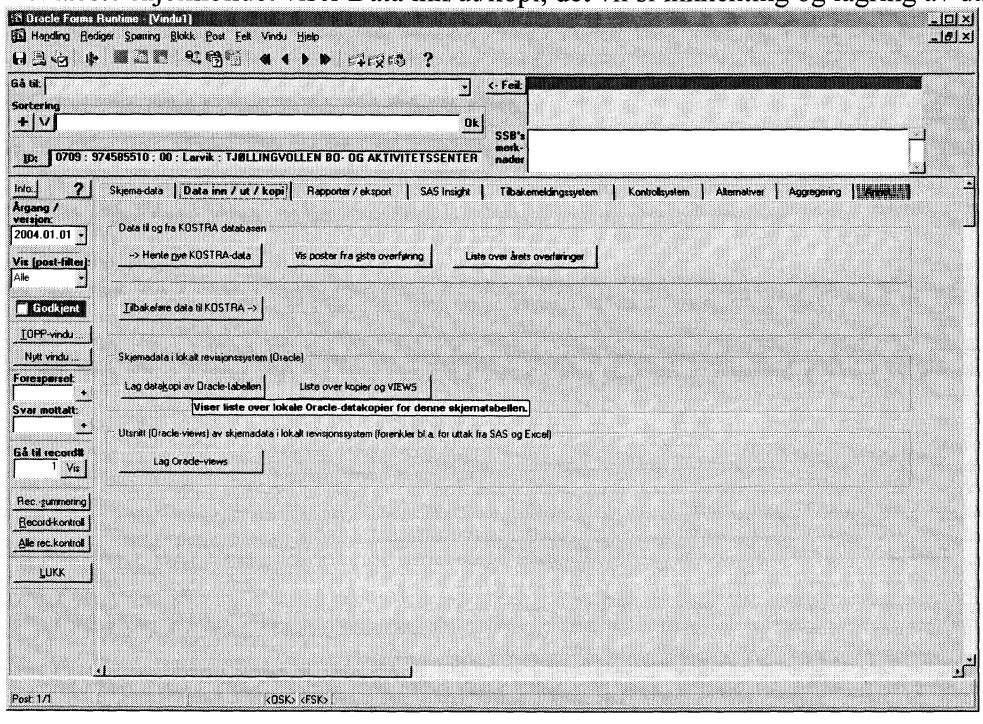
$$DIFF_{i,.,t} = \sum_{k=1}^{K} \frac{w_{i,t} |y_{i,k,t} - y'_{i,k,t-1}| Z_{i,k,t} P_k}{\hat{Y}_{d,k,t-1}}$$

Om $DIFF_{i,,,t}$ er større eller lik en kritisk verdi, utløses videre kontroll/retting for enhet i.

10.4. Skjermbilder fra GenRev

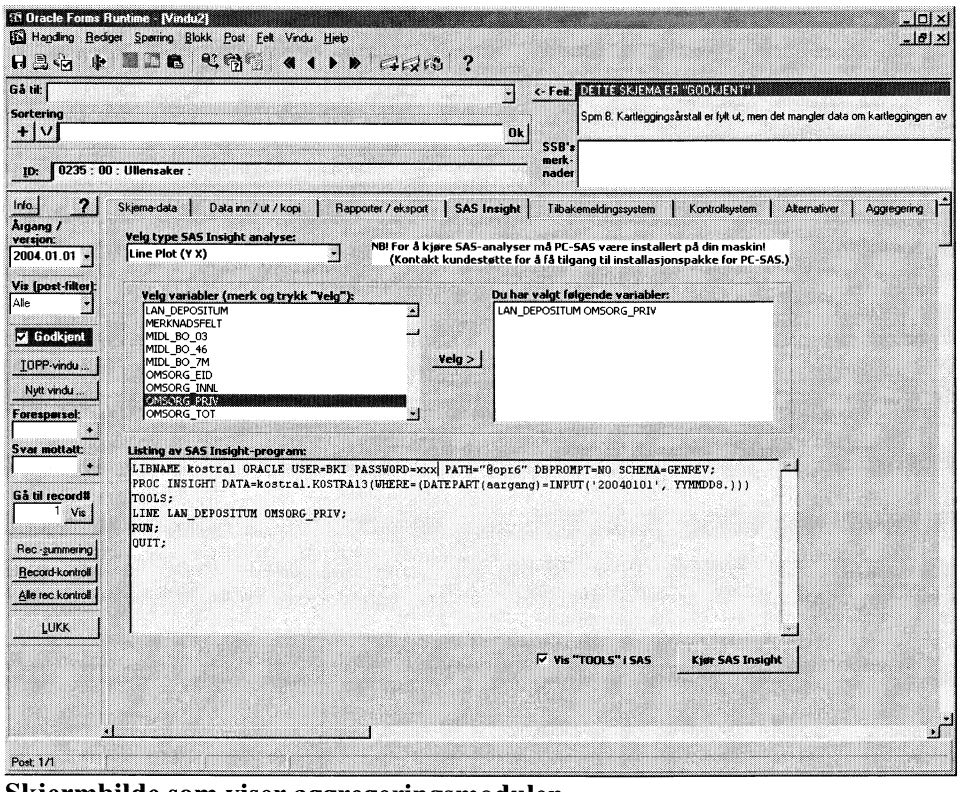
Her presenteres eksempler på skjermbilde fra den generelle revisjonsapplikasjonen i KOSTRA.

Det første skjermbildet viser Data inn/ut/kopi, det vil si innhenting og lagring av data.



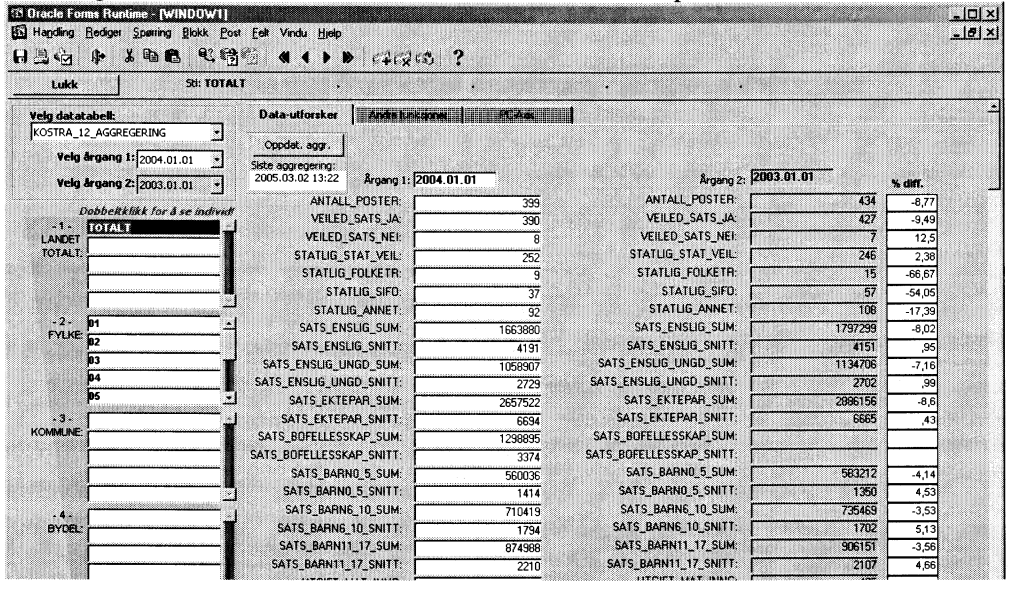
Det er tilgang til applikasjonene for skjemadata, rapport/eksport, SAS-insight, tilbakemeldingssystem, kontrollsystem, alternativer, aggregering og annet ved å velge andre faner i dette skjermbildet. Noen av disse applikasjonene vises nedenfor.

Skjermbilde som viser muligheten for analyse av variabler gjennom SAS-insight



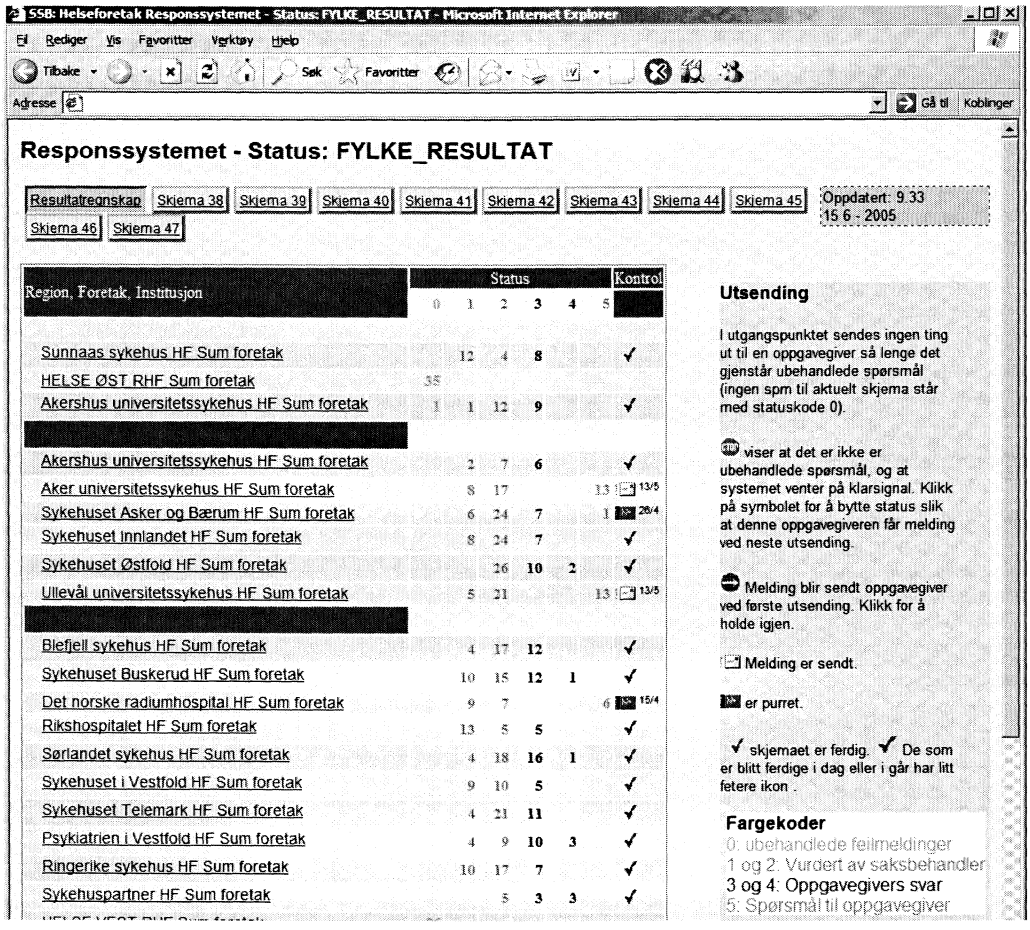
Skjermbilde som viser aggregeringsmodulen

I venstre kolonne velges aggregeringsnivå, det vil si bydel, kommune, fylke eller landet totalt.
I andre kolonne vises aggregerte verdier for det siste året og i tredje kolonnen vises aggregerte verdier for foregående år. Den siste kolonnen viser differansen i prosent mellom de to siste årene.

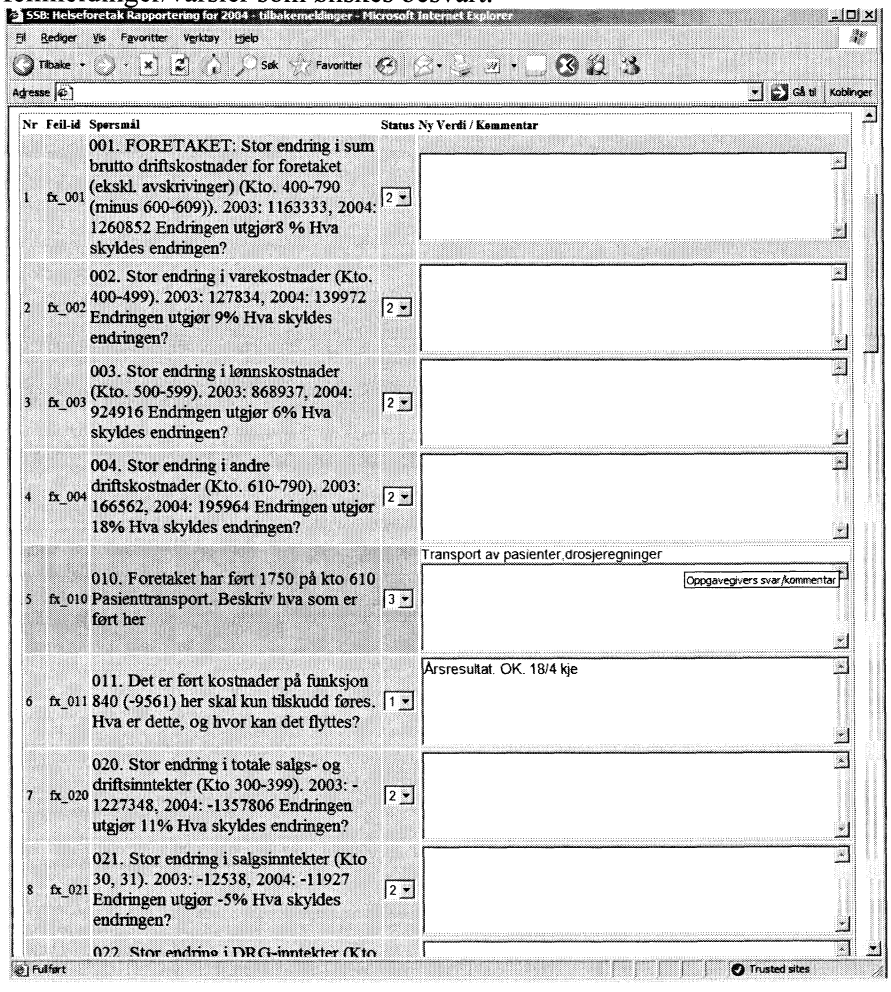


Tilbakemeldingsmodulen til oppgavegiver

Dette skjermbildet viser status i revisjonsprosessen, hvilke tilbakemeldinger som er sendt til oppgavegiver og status på eventuelle svar.


SSB: Helseforetak Responssystemet - Status: FYLKE_RESULTAT - Microsoft Internet Explorer
Fil Rediger Vis Favoritter Verktøy Hjelp
Tilbake Søk Favoritter
Adresse
Gå til Koblinger
Responssystemet - Status: FYLKE_RESULTAT
Resultatregnskap Skjema 38 Skjema 39 Skjema 40 Skjema 41 Skjema 42 Skjema 43 Skjema 44 Skjema 45 Skjema 46 Skjema 47
Oppdatert: 9.33 15.6.2005
Region, Foretak, Institusjon
Status
Kontroll
0 1 2 3 4 5
Sunnaas sykehus HF Sum foretak 12 4 8 ✓
HELSE ØST RHF Sum foretak 35
Akershus universitetssykehus HF Sum foretak 1 1 12 9 ✓
Akershus universitetssykehus HF Sum foretak 2 7 6 ✓
Aker universitetssykehus HF Sum foretak 8 17 13 13/5
Sykehuset Asker og Bærum HF Sum foretak 6 24 7 1 26/4
Sykehuset Innlandet HF Sum foretak 8 24 7 ✓
Sykehuset Østfold HF Sum foretak 26 10 2 ✓
Ullevål universitetssykehus HF Sum foretak 5 21 13 13/5
Blefjell sykehus HF Sum foretak 4 17 12 ✓
Sykehuset Buskerud HF Sum foretak 10 15 12 1 ✓
Det norske radiumhospital HF Sum foretak 9 7 6 15/4
Rikshospitalet HF Sum foretak 13 5 5 ✓
Sørlandet sykehus HF Sum foretak 4 18 16 1 ✓
Sykehuset i Vestfold HF Sum foretak 9 10 5 ✓
Sykehuset Telemark HF Sum foretak 4 21 11 ✓
Psykiatrien i Vestfold HF Sum foretak 4 9 10 3 ✓
Ringerike sykehus HF Sum foretak 10 17 7 ✓
Sykehuspartner HF Sum foretak 5 3 3 ✓
Utsending
I utgangspunktet sendes ingen ting ut til en oppgavegiver så lenge det gjenstår ubehandlede spørsmål (ingen spm til aktuelt skjema står med statuskode 0).
viser at det er ikke er ubehandlede spørsmål, og at systemet venter på klarsignal. Klikk på symbolet for å bytte status slik at denne oppgavegiveren får melding ved neste utsending.
Melding blir sendt oppgavegiver ved første utsending. Klikk for å holde igjen.
Melding er sendt.
er purret.
✓ skjemaet er ferdig. ✓ De som er blitt ferdige i dag eller i går har litt fetere ikon .
Fargekoder
0: ubehandlede feilmeldinger
1 og 2: Vurdert av saksbehandler
3 og 4: Oppgavegivers svar
5: Spørsmål til oppgavegiver

Detalj som viser skjermbilde for tilbakemelding til oppgavegiver, med utlisting av hvilke feilmeldinger/varsler som ønskes besvart.



10.5. Ordliste

Absolutt kontroll	Logisk kontroll. Kontroll som identifiserer verdier i datasettet som med sikkerhet er feil, for eksempel ugyldig verdi eller sumfeil
Administrative data	Data innhentet av andre etater for administrative formål
Aggregert metode	Generell feilsøkingsmetode i to trinn: Først makrokontroll på aggregerte data for å identifisere mistenkelige verdier, dernest mikrokontroll av de mistenkelige verdiene
Automatisk retting	Maskinelle korreksjoner, uten innblanding av saksbehandler, foretatt på grunnlag av på forhånd fastlagte regler. Dette kan enten være logiske sammenhenger i skjema, basert på data fra samme enhet i foregående undersøkelse eller fra andre enheter i samme undersøkelse; for eksempel gjennomsnittverdier
Cold-deck imputering	Beregner manglende verdier på grunnlag av data for samme enhet fra foregående undersøkelse. Den enkleste form for cold-deck imputering er ren kopiering av for eksempel priser fra forrige undersøkelse, men det kan legges inn mer avanserte beregninger.
Deduktiv imputering	Manglende verdi(er) for en enhet fastsettes fra andre data for samme enhet på grunnlag av logiske regler. For eksempel kan en sumpost beregnes som summen av underpostene (hvis disse er oppgitt). Hvis det går fram av oppgaven at en person er mor til barn, kan manglende avkryssing for kjønn settes til kvinne.
Ekstremverdi	Dataverdi som avviker betydelig fra andre verdier i datasettet og som kan mistenkes for å være feil. Det må kontrolleres om verdien er riktig (utligger) eller feil og eventuelt må rettes.
Endringsfrekvens	Andelen endrede verdier av totalt antall verdier som er mulig å endre.
Estimering	Beregne verdien for en ukjent størrelse med data fra et utvalg.
Flagging	Merking av enheter eller variabelverdier som forteller hvordan dataene er vurdert og behandlet i revisjonsprosessen
Frafall	*Enhetsfrafall:* Enheter i undersøkelsesbestanden som mangler fullstendig av en eller annen grunn (nekting, opphør mv.) *Partielt frafall:* Enheter med i undersøkelsesbestanden, men enkelte opplysninger mangler (ufullstendig oppgave)
Grafisk revisjon	Bruk av grafiske presentasjoner for å få oversikt over datamaterialet, bl.a. avvikende verdier. Slike diagrammer kan være interaktive slik at en både kan identifisere og rette enheter og automatisk få fram ny presentasjon av datamaterialet etter oppretting.
Grenseverdier	Øvre og nedre grense en verdi må ligge innenfor for ikke å indikere mulig feil i en intervallkontroll. Grenseverdier kan fastsettes manuelt eller maskinelt.

Hidiroglou-Berthelot-metoden	Maskinell statistisk feilsøkingsprosedyre basert på egenskapene til dataene. Tar hensyn til både relativ og absolutt endring i dataverdi fra foregående periode. Utviklet av Statistics Canada
Hot-deck imputering	Beregner manglende verdier på grunnlag av data for en annen enhet i samme undersøkelse; en enhet som likner mest på «mottakende» enhet, f.eks. foretak i samme bransje eller personer i samme aldersgruppe.
Identifikasjonsdata	Variable som identifiserer enheten, f.eks. fødselsnummer eller organsisasjonsnummer
Imputering	Imputering vil si å sette inn verdier for manglende opplysninger. Man forsøker å utnytte annen informasjon til å finne rimelige verdier for de manglende variablene.
Indikator	Mål for effekten av en prosess, f.eks. andel enheter markert for feil
Innligger (inlier)	Dataverdi som ikke avviker særlig fra normale verdier, men som er feil. Kan ha betydelig påvirkning på totalresultatet, og kan være vanskelig å oppdage
Intervallkontroll	Kontrollerer om en verdi ligger innenfor visse fastsatte grenseverdier
Iterativ prosess	Prosess som gjentas og som brukes ved kontroll av samlet materiale: Oppretting av de viktigste feilene først, dernest nye kontroll- og retterunder inntil datamaterialet er akseptabelt
Kategorisk variabel	Variabel som beskriver egenskaper ved en enhet. Egenskapene er kategorisert i klasser. Klassene kan ha numeriske verdier, men det har ingen mening å regne på disse verdiene, f.eks. fylkeskode.
Kjennemerke	Et annet ord for variabel.
Koding	Gruppering av data etter en kodeliste (standard) og påføring av gyldig verdi
Konsistenskontroll	Kontroll av logiske sammenhenger mellom to eller flere variable
Korreksjon	Endring av verdi som er feil eller antatt å være feil til en riktig verdi eller antatt riktigere verdi.
Makrodata	Mikrodata aggregert (summert, slått sammen), f.eks. til celler i tabeller for publisering
Makrokontroll	Kontroll på samlet materiale for å identifisere individuelle feil
Manuell retting	Saksbehandler setter selv inn verdi for feil eller manglende data på grunnlag av regler og egen vurdering, eventuelt i samråd med oppgavegiver.
Markeringsfrekvens	Andelen verdier markert som (mulig) feil i en kontroll i forhold til totalt antall verdier som er kontrollert.
Metadata	Informasjon om dataene, som f.eks. når datasettet ble laget, hvor det ligger lagret, formater, forklarende tekst til variablene og lignende
Mikrodata	Informasjon for den minste statistiske enheten, som regel person/husholdning eller bedrift/foretak. Opplysningene om enheten kan foreligge med full identifikasjon (navn, adresse), avidentifisert eller anonymisert

Mikrokontroll	Kontroll av individuelle enheter enkeltvis, ofte kjennemerke for kjennemerke
Numerisk variabel	Variabel som er tellbare, f.eks. antall rom og omsetning. Det gir mening å regne på numeriske verdier, f.eks. gjennomsnittlig antall rom eller total omsetning i en bransje.
Nøkkelvariable	Variable som er sentrale i publisert statistikk
Observasjon	Den enkelte enhet med variabelverdier
Optisk lesing	Skjema blir maskinelt "fotografert" og dataene lagt direkte på fil
Reviderte data	Data som er ferdig kontrollert og eventuelt korrigert. Benevnes også statistikkdata
Revisjon	Gransking, kontroll og endring av data. I hovedsak brukt om regnskapsdata, men i denne publikasjonen brukt om gransking av alle typer data.
Revisor	Person som utfører gransking (vanligvis brukt i økonomisk statistikk, f.eks. regnskap)
Rådata	Opprinnelige data mottatt fra oppgavegiver, før kontroll og oppretting starter
Selektiv revisjon	Makrokontrollrutine som identifiserer verdier/feil som har stor innvirkning på totalresultatet. Revisjonen blir konsentrert om disse feilmeldingene
Statistikkdata	Reviderte data
Statistisk enhet	Observasjonsenheten er den enheten vi skal kartlegge egenskaper ved, svært ofte person eller bedrift. Rapporteringsenheten kan være forskjellig fra observasjonsenheten, f.eks. kan et foretak rapportere for alle bedriftene i foretaket og person kan rapportere for husholdningen samlet
Systematisk feil	Gjennomgående feil som trekker i en retning og forekommer for mange enheter i undersøkelsen. Det kan ofte skyldes misforståelse eller svake variabeldefinisjoner
Top-down metode	Makrokontrollrutine der en kontrollerer de viktigste enhetene først; f.eks. de største enhetene i absolutt verdi eller de enhetene som har størst endring
Treffsikkerhet av kontrollen	Andelen endrede verdier av totalt antall verdier som er markert som mulig feil.
Utligger (outlier)	Ekstremverdi som er korrekt
Variabel	Egenskap ved de statistiske enhetene i en undersøkelse, f.eks. alder eller omsetning.
Vekt	Hver observasjon/enhet blir tillagt vekt slik at nettoutvalget blir representativt for hele populasjonen
Vurderingskontroll	Kontroll av verdier eller sammenhenger mellom verdier som virker mistenkelige, men som kan være riktige

10.6. Litteratur og referanseliste

Abrahamsen, A. S. (2005). Analyse av revisjon - Feilkoder og endringer i utenrikshandelsstatistikken. Notater 2005/10. Statistisk sentralbyrå.

Abrahamsen, A. S. og Seierstad, A. (2004). Analyse av revisjon. Kostra kommunehelse. Notater 2004/72. Statistisk sentralbyrå.

Andersen,T. (1998). Klargjøring, Feilidentifisering, kontroll og imputering. Et støttesystem for revisor. Dataeditering. Begreper og metoder. Upublisert notat.

Brørs A. S., Dybendal K., Foss A. H. og Jakobsen T. (2000). Dokumentasjon av BESYS-befolkningsstatistikksystemet. Notater 2000/24. Statistisk sentralbyrå.

Foss A. H. (2003): Grafisk revisjon av nøkkeltallene i Kostra. Notater 2003/75. Statistisk sentralbyrå.

Granquist,L., Arvidson,G., Elffors, C., Norberg, A., Lundell, L.-G. (2002). *Guide til granskning.* Statistics Sweden, CBM 2002:1.

Håndbok i datarevisjon (1998). Statistisk sentralbyrås håndbøker 66.

Latouche , M. and Berthelot, J.-M. (1992). Use of a score function to prioritize and limit recontacts in business surveys. *Journal of Official Statistics,* Vol 8, 389 - 400.

Mevik, A-K. (2005). Revisjon av Strukturstatistikk for industrien. Et forslag til selektiv revisjon. Notater 2005/46. Statistisk sentralbyrå.

Roll-Hansen D., Ferstad S., Stålnacke M., Tuhus P. og Nøtnæs T. (2002). En spørreskjemametodisk gjennomgang av datainnsamling gjennom Grunnskolen informasjonssystem (GSI). Notater 2002/23. Statistisk sentralbyrå.

*Statistical Data Editing. Methods and Techniques. Volume No.*1 (1992). United Nations Statistical Commission and Economic Commission for Europe. United Nations, New York.

Vedø, A. (2005). Analyse av revisjon. Lønn i bygge- og anleggsvirksomhet. Notater 2005/29.

De sist utgitte publikasjonene i serien Statistisk sentralbyrås håndbøker

45 Håndbok i datasikkerhet og fysisk sikring. Revidert utgave, november 1998. 1998. 83s.

46 Telefonkatalog. 1998. 89s.

47 EØS-avtalen. Det statistiske samarbeid og konsekvenser for Statistisk sentralbyrås statistikkproduksjon. 1994. 55s.

48 Håndbok i tilsettingssaker. 1994. 32s.

49 Oppgaveplikt og tvangsmulkt. 1995. 55s.

50 Emneinndeling 1995. 1995. 43s.

51 Intervju: EDB-arbeidsbok. 1995.

52 Intervju: EDB-oppslagsbok. 1995.

53 Intervju: Opplæring og administrasjon. 1995.

54 Internkontroll: Revidert utgave 1997. 25s.

55 Nordisk statistikk på CD-ROM: Veiledning. 20s.

56 PC-Axis versjon 2.2: Brukerhåndbok. 69s.

57 Produktregister versjon 4.0: Brukerveiledning. 49s.

58 Håndbok i prosjektstyring. 20s.

59 Personalreglement for Statistisk sentralbyrå. 22s.

60 Produktnummerkatalog pr. 28.02.1996. 55s.

61 Innkjøpshåndbok. 1996.

62 Timeplan versjon 3.0: Brukerveiledning. 16s.

63 Håndbok i EDB-metode. 52s.

64 Publiseringshåndbok: Regler og retningslinjer for publisering i Statistisk sentralbyrå. 93s.

65 Håndbok i utvikling av statistikksystemer: Med vekt på IT-metode. 52s.

66 Håndbok i datarevisjon. 48s.

67 Arkivnøkkel for Statistisk sentralbyrå. 76s.

68 Rapporteringshåndbok for KOSTRA-regnskap 1999: Oppslagshefte til hjelp ved filuttrekk for KOSTRA-rapportering. 52s.

69 Yrkeskatalog for innrapportering av yrke til arbeidstakerregisteret. 86s.

70 Håndbok for KOSTRA-rapportering 2000. Oppslagshefte til hjelp ved filuttrekk for KOSTRA-rapportering, regnskap. Revidert utgave oktober 2002. 73s.

71 Håndbok i SAS. Del 2: Oppslag. 243s.

72 Yrkeskatalog pr. november 2002. Korrigert utgave. 170s.

73 Håndbok i SAS. Del 1: Innføring. 65s.

74 Håndbok i datalagring på Unix i Statistisk sentralbyrå. 4. utgave. 73s.

75 The EFTA/EU Statistical Co-operation outside and within the EEA Framework - Legal Basis, Practical Experiences and Guidelines. 55s.

76 Intervju: Intervjupermen.

77 Intervju: Arbeidsbok.

78 Håndbok i rapportering av regnskapsdata for helseforetak og regionale helseforetak 2002. Oppslagshefte til hjelp ved filuttrekk. 43s.

79 Håndbok for rapportering av regnskapsdata for helseforetak og regionale helseforetak. 2003. Oppslagshefte til hjelp ved filuttrekk. 42s.

80 Prosjekthåndboka - slik gjør vi det i SSB. 34s.

81 Råd for utvikling og utforming av webskjema. Verson 1.1. 76s.

82 Håndbok for kirkelige fellesråd - rapportering 2004. Statistisk sentralbnyrås håndbøker. Oppslagshefte til hjelp ved filuttrekk for Elektroniskrapportering, regnskap . November 2004

Statistisk sentralbyrå
Statistics Norway

Design: Siri Boquist / Foto: Photos.com