

## Om arkivstatistiske systemer

av

*Svein Nordbotten*

### Innhold

0. Innledning.
1. Behov for statistiske informasjonsarkiver.
2. Produksjonsteoretiske betraktninger.
3. Sentrale registre.
4. Data-innsamling.
5. Data-arkiver.
6. Statistikk-produksjon.
7. Statistikkarkiver.
8. Avslutning.

### 0. Indledning

På det Nordiske Statistiker Møte i Helsingfors i 1960 ble det under betegnelsen et *arkivstatistisk system* lagt fram en del synspunkter på statistikkproduksjon. I dette innlegg tar en sikte på å legge fram en oversikt over videre synspunkter som knytter seg til et slikt system og hvordan vi i Norge tar sikte på å sette enkelte idéer ut i livet. Framstillingen bygger på mange impulser fra statistikere som en har hatt høve til å drøfte synspunktene med, uten at disse på noen måte kan lastes for svakheter i resonnement.

Betegnelsen arkiv-statistisk system kan gi inntrykk av at det her er tale om revolusjonerende nye idéer. Dette er ikke riktig. Det system som skal drøftes, bygger i stor utstrekning på gamle idéer som det nå er teknisk mulig å realisere.

#### 1. Behovet for statistiske informasjonsarkiver

Grunntanken bak det arkiv-statistiske system er at enhver ny observasjon eller registrering representerer økt kunnskap og har en informasjonsverdi under forutsetning av at den kan gjøres praktisk tilgjengelig. Tidligere har kostnadene ved å gjøre en observasjon tilgjengelig ofte overstegte observasjonens informasjonsver-

di. Relasjonen mellom kostnad og informasjonsverdi er nå i ferd med å endres både fordi lagringskostnader pr. informasjonsenhet reduseres og fordi integrasjonsmulighetene gjør informasjonsverdien pr. informasjonsenhet vesentlig større. Dessuten øker informasjonsverdien fordi etterspørselen etter statistikk generelt øker.

De økende krav til statistisk informasjon kommer delvis til uttrykk i ønsker om spesialbearbeidinger utover det en finner grunnlag for å gi i standard tabellverk. Ofte krever slike spesialprodukter at data for de samme enheter, men fra forskjellige kilder knyttes sammen, eller at data fra forskjellige tidspunkter eller perioder for de enkelte enheter knyttes sammen.

Dette kommer klart fram i problemstillinger hos stadig flere samfunnsforskere. De er ikke lenger tilfredse med å basere sine analyser på aggregater for grupper av enheter, fordi aggregatene bare gir kunnskap om variasjonen mellom gruppene, mens variasjonen i gruppene ikke kommer fram, og fordi en rekke modeller om de enkelte enheters tekniske eller adferdsmessige funksjoner ikke lar seg aggregere til tilsvarende makro-modeller. Alternativet til makro-beregninger på aggregerte størrelser fra det statistiske standardprodukt er derfor spesielle mikro-beregninger med etterfølgende aggregering av resultatene.

For å møte slike behov vil det kreves omfattende informasjonsarkiver hvor observasjoner og registreringer som gjøres på forskjellige steder samles, systematiseres og holdes i beredskap.

De moderne datamaskiner kan bidra til realisering av en slik beredskap på to måter. For det første ved at de representerer et teknisk, effektivt arkiveringshjelpemiddel som gjør det

mulig å arbeide med mange millioner tall i arkivene. For det andre bidrar anvendelsen av datamaskiner i stadig flere administrative institusjoner til at data som er innhentet for ikke-statistiske formål kan kopieres for statistisk bruk både raskere, i større omfang og for mindre kostnader enn tidligere.

Liknende synspunkter er i den senere tid også kommet til uttrykk i flere land. I Sverige er det i Statistiska Centralbyråen etablert en utredningskomité for arkivstatistiske systemer med en rekke underutvalg som arbeider med forskjellige sider av systemet. I U. S. A. har en komité, Ruggles-komitéen, oppnevnt av Social Science Research Council for å vurdere statistikk-produksjonen i U. S. A., nylig pekt på at det — for vitenskapelig utnyttelse — er behov for å opprette et nasjonalt data-arkiv til systematisk oppbevaring av data som blir samlet inn av forskjellige organer. Tanken ble tatt opp av Bureau of the Budget som har latt utarbeide en detaljert rapport om behovet for et slikt data-arkiv.

I det følgende vil det bli gjort rede for de prinsipielle synspunkter en har i Norge på dette felt og for de framstøt en gjør eller har planer om å gjøre for å sette de arkiv-statistiske idéer ut i livet. Den koordinering og planlegging systemet forutsetter foregår i et langtidsplanleggingssystem med årlig rullering.

## 2. Produksjonsteoretiske betraktninger

Under drøfting av arkiv-statistiske systemer er det nyttig å formalisere betraktningene ut fra produksjonsteoretisk synsvinkel. Utgangspunktet er at samfunnsproduksjonen målt på en eller annen måte er avhengig av den masse av erfaring og kunnskap samfunnet forvalter. Det er her verdt å merke seg at kunnskapen ikke har karakter av et gode som forbrukes, men at den har større likhet med en kapitalbeholdning. En del av denne kunnskap er av statistisk natur som atskiller seg fra annen kunnskap ved at den gjelder grupper av enheter eller fenomener, ikke de enkelte enheter. Vi skal her bare drøfte den kunnskap som stammer fra det statistiske system.

Tilførselen av statistisk informasjon i samfunnet skjer ved mangfoldiggjøring og spredning av resultatene fra statistiske beregninger.

La oss betegne denne prosessen med

$$(1) \quad I = I(M, S)$$

hvor  $I$  er informasjonstilførselen pr. tidsenhet,  $S$  beholdningen av statistiske størrelser beregnet og  $M$  en mangfoldiggjøringsfaktor. Vi vil kalle  $S$  for *statistikk-kapitalen* fordi den spiller samme rolle i produksjonen av statistisk informasjon som realkapitalen i en vanlig produksjonsprosess, og kan, på samme måte som realkapitalen nyttes, om igjen. Informasjons-spredningen kan derfor foregå uten at det skjer noen endring i statistikk-kapitalen.

Statistikk-kapitalen er produsert ved beregninger dels fra data for statistiske enheter, og dels ved en reproduksjon fra statistikk-kapitalen selv. Statistikk-kapitalen består av aggregerte størrelser som totaler, gjennomsnitt, etc. for grupper av to eller flere statistiske enheter. *Investering* i statistikk-kapitalen skjer ved statistikkproduksjon, og investeringen pr. tidsenhet betegnes med:

$$(2) \quad \dot{S} = S(V, U, S, D)$$

hvor

$$(3) \quad \frac{dS}{dt} = \dot{S}$$

$D$  er statistikk-systemets beholdninger av data som vi her kaller *data-kapitalen*,  $V$  og  $U$  står for utnyttingsgraden av kapasiteten av henholdsvis  $S$  og  $D$ . Vi kan derfor ha statistikkproduksjon uten at det samtidig skjer noen endring i datakapitalen. Data-kapitalen består, i motsetning til statistikk-kapitalen, av data for de enkelte statistiske enheter.

*Investering* pr. tidsenhet i datakapital,

$$(4) \quad \frac{dD}{dt} = \dot{D}$$

skjer gjennom data-innsamling.

Til dette statistiske system knytter det seg kostnader både til informasjonsspredning,  $I$ , statistikkproduksjon,  $S$ , datainnsamling,  $D$ , og til oppbevaring av statistikk-kapital,  $S$ , og data-kapital,  $D$ . Kostnadene pr. tidsenhet kan vi derfor betegne med:

$$(5) \quad C = C(I, \dot{S}, \dot{D}, \dot{S}, \dot{D})$$

Alle variable i systemer (1)–(5) er tidsfunksjoner.

Statistikk-systemets målsetning vil kunne tenkes å være å finne de tidsfunksjoner for  $\bar{D}$ ,  $V$ ,  $U$  og  $M$  som ut fra gitte initialbetingelser for  $S$  og  $D$  og under bibetingelse av (1)–(5) maksimerer en eller annen funksjonal

$$(6) \quad W = W \left( \begin{bmatrix} I \\ O \end{bmatrix} T, \begin{bmatrix} C \\ O \end{bmatrix} T \right)$$

som representerer et typisk dynamisk programmeringsproblem over en tidsperiode på  $T$  tidsenheter.

I en mer detaljert modell vil det være nødvendig å sondre mellom forskjellige typer og aldersklasser av statistikk- og datakapital med ulik produksjonskostnad, produktivitet og informasjonsverdi. Det kan tenkes situasjoner hvor produktiviteten av f. eks. eldre data-kapital blir så lav at den må utrangeres, og det kan da være hensiktsmessig å innføre begrepet kapital-slit.

Det som skiller det statistiske system, som er skissert ovenfor, fra den mer tradisjonelle betraktningmåte, er at  $S$  og  $D$  er innført. Dersom disse to størrelser ikke tas med, får vi i stedet et statistisk system, som fører til et statistisk programmeringsproblem. Sagt med andre ord, i det system vi her har drøftet vil situasjonen på ethvert tidspunkt være avhengig av det tidligere tidsforløp for  $\bar{S}$  og  $\bar{D}$ , fordi allerede beregnet statistikk og innsamlede data kan brukes om igjen og vil påvirke framtidige avgjørelser. I et system hvor en ikke tar i betraktning statistikk- og datakapital, vil situasjonen på ethvert tidspunkt være uavhengig av tidligere statistikkproduksjon og datainnsamling.

Betraktningene ovenfor kan ha en rekke konsekvenser for både datainnsamling, statistikkproduksjon og informasjonsspredning. Mens vi hittil f. eks. har nøyd oss med å la opplegget for en innsamling i stor utstrekning bli bestemt av en eneste etterfølgende bearbeiding, innebærer tankegangen ovenfor at vi også må ta i betraktning mulige framtidige beregninger hvor data vil bli knyttet sammen med data som senere vil bli innhentet. Dette kan kreve et innsamlingsopplegg hvor en jevnere innsamling over tiden blir mer framtreddende. Et opplegg for produksjon av statistikk må f. eks. ta i betraktning muligheten av å utnytte tidligere innsamlede data i større utstrekning, og dette

vil kunne kreve nye estimeringsmetoder og beregningsteknikk.

Betraktningene ovenfor viser at data-kapital og statistikk-kapital er sentrale begreper i systemet. Forutsetningen er imidlertid at de representerer data og statistikk som er organisert slik at de kan utnyttes. Kravene til organisasjon kan illustreres ved en *databoks* med en rekke små rom for oppbevaring av observasjoner. Observasjonene identifiseres ved at hver statistisk enhet får sin faste plass langt boksens første akse, hvert kjennemerke får sin bestemte plass langs den andre akse, mens hver tidsperiode ordnes langs den tredje aksens. Innholdet i observasjonsrommene som skjæres av et snitt på tvers av boksens tidsakse vil gi et data-bilde av situasjonen på et tidspunkt. Et snitt på tvers av kjennemerkeaksen vil gi utvikling på et bestemt område, mens snittet normalt på enhetsaksen vil gi oss den registrerte livshistorie for en enhet.

Tilsvarende gjelder også for statistikk-kapitalen med den forskjell at i stedet for enheter, må hver statistiske gruppe få sin faste plass langs boksens første akse.

Betingelsene for organisasjon av data i databoksen er at en har permanente identifikatorer for alle statistiske enheter og kodestandard for alle kjennemerker. I statistikk-kapitalen gir nasjonalregnskapssystemet eksempler på identifikasjoner og standard-koder, som bidrar til å gjøre statistikken konsistent og sammenliknbar over tiden. For data-kapitalen har vi ikke ennå i samme utstrekning utviklet tilsvarende identifikator-systemer.

### 3. Sentrale registre

#### 3.1. Alminnelige betraktninger.

Et register er en liste over de statistiske enheter i en masse. Oppretting av en slik liste krever at en har funnet fram til en *enhets*-definisjon og en definisjon av *massen*. Både enhets- og masse-definisjon fører til en rekke problemer som vi ikke har høve til å ta opp her. Registeret slik en her skal drøfte det, representerer en kryssreferanse mellom to typer identifikatorer, den *interne* og den *eksterne* identifikator. Den interne identifikator som skal følge data inn og ut av databoksen et stort antall ganger, bør konstrueres slik at det blir

minst mulig plasskrevende. Sagt på en annen måte, det bør være et identifikatorsystem som er mest mulig kompakt slik at databoksen ikke blir unødig bred p. g. a. ubenyttede identifikasjonsplasser langs dens første akse. På denne måte reduseres kapasitetsbehov og operasjonstider til det absolutt nødvendige. Den interne identifikator kan f. eks. være et tall eller en tegnkombinasjon. Den eksterne identifikator blir nytt til å finne fram til den statistiske enhet når en har behov for å få opplysninger direkte fra den og kan f. eks. være navn, gateadresse og poststed. I enkelte situasjoner kan samme identifikator nyttes både eksternt og internt, men oftest vil den eksterne identifikator være ubruktbar til intern identifikasjon fordi den både er unødig lang og dessuten ustabil.

Kravet om permanent identifikasjon medfører at en videre må ta standpunkt til definisjoner av *fødsel*, *flytting* og *død* for de forskjellige typer av enheter. Dette fører til problemer en tidligere ofte gikk utenom. Nu må de imidlertid løses dersom registrene skal holdes vedlike og kravet om permanent identifikasjon tilfredsstilles.

Det er svært viktig at flest mulig av de organer som samler inn data fra samme masse av enheter nytter samme register. Det vil gjøre det mulig å overlata data til statistikkssystemet som kan ordne disse entydig i databokser. En oppnår dessuten en betydelig forenkling i vedlikeholdet. I mange tilfelle er det naturlig at ansvaret for slike sentrale registre overlates statistikkssystemet som vil sitte med de mest omfattende opplysninger om de enkelte enheter.

I Norge har statistisk Sentralbyrå flere sentrale registre som nå legges opp med sikte på å tilfredsstille de krav et arkiv-statistisk system stiller.

### 3.2. Sentralt personregister.

Folkeregistreringen i Norge er fastlagt ved lov av 15. nov. 1946, og Statistisk Sentralbyrå er tillagt funksjonen som Sentralkontoret for folkeregistreringen. I dag har en i Norge 460 lokale folkeregistre. I tillegg til de forskriftsbestemte registre er det nå ved alle folkeregistre også opprettet hullkortkopier av registrene. For-

uten det statistiske formål skal folkeregistrene betjene en rekke offentlige registerbehov.

I forbindelse med folketellingen i 1960 ble mulighetene for et sentralt personregister drøftet spesielt med utgangspunkt i ønskeligheten av et landsomfattende permanent personnummer-system. Byrået utredet saken og fikk høve til å ta opp arbeidet med et sentralregister i 1963.

Den permanente personidentifikasjon som ble bygd opp omfatter et 11-sifret nummer. De seks første siffer, kalt fødselsdata, er fødselsdag, -måned, og de to siste siffer i fødselsåret. De tre etterfølgende siffer er nytt til å skille mellom personer med ens fødselsdata, mellom menn og kvinner, etc. De to siste siffer er kontrollsiffer og beregnet etter modulus 11-systemet. En har anslått at av feil som forekommer i identifikasjonsnummeret vil bare 1 av 100 000 slippe gjennom uoppdaget. Dette identifikasjonsnummer vil bli brukt som intern identifikator, men det er – som en forstår – ikke ideelt konstruert for dette formål.

Opprettelsen av det sentrale register foregikk ved at navn, adresse, kjønn og fødselsdato for alle personer ble punchet i hullkort fra folketellingslistene. Deretter ble disse kort automatisk tildelt identifikasjonsnummer av en automatisk maskinrutine som selv holder rede på hvilke nummer som er disponible på en hvilken som helst dag. Disse nye kortene ble sendt ut til de lokale registre for kontroll og for å gjøre de lokale registre kjent med de nummer som heretter skal brukes i folkeregistreringen. Tilgang av fødte og innvandrede meldes løpende til sentralregisteret som tildeler nummer. På grunn av tidsforskjellen mellom folketellingstidspunkt og kontrolltidspunkt som ble fastsatt til 1. oktober 1964, har kontroll- og suppleringsarbeidet fått et meget stort omfang. Sentralregisteret er lagt opp på magnetbånd, og en versjon av registeret opptar i dag ca. 30 magnetbåndruller. Under oppbygningsarbeidet nyttes imidlertid mange hundre bånd for dette formål.

Det sentrale personregister vil bli holdt ajour gjennom de ordinære registreringskanaler for fødsler, vigsler, flyttinger og dødsfall, og vil etter hvert overta de større serviceoppgaver fra de lokale folkeregistrene. Allerede i år vil identifikasjonsnummersystemet bli tatt i bruk i skatsektoren og i trygdesektoren. Både politiet

og helsevesenet vil trolig ta det i bruk, og vi regner med at det vil komme inn i undervisningssektoren allerede fra grunnskolen.

I løpet av 1967 har en planer om å bygge registeret videre ut til et registersystem som omfatter både et personregister og et husholdningsregister. Den lokale registrering gir allerede i dag grunnlag for utbygging og vedlikehold av et slikt system.

### 3.3. Sentralt bedrifts- og foretaksregister.

I tilknytning til bedriftstelingen i 1953 ble det opprettet et sentralt bedrifts- og foretaksregister som i virkeligheten er to registre som er integrert i et system. Fram til 1965 ble registeret holdt på hullkort, men er nå overført til magnetbånd. Det omfatter de aller fleste næringer, bortsett fra Jordbruk, Skogbruk, Fangst og Fiske. I enkelte næringer er enheter hvor eieren arbeider alene, ikke med. I alt omfatter registeret ca. 110 000 foretak og 130 000 bedrifter. Den eksterne identifikasjon i registeret er i alminnelighet navn og postadresse. Registerne opptar bare noen få magnetbåndruller.

For å tilfredsstille behovet for permanente interne enhetsidentifikatorer ble det i 1965 innført uavhengige bedrifts- og foretaksnummer som, i samsvar med tankegangen ovenfor, er laget så kompakte som mulig, dvs. de har ingen informasjon innebygget. Dette har gitt oss nummer med 6 siffer pluss 1 kontrollsiffer beregnet etter modulus 11-systemet for både foretaksnummer og bedriftsnummer. Det vil slippe gjennom ett par prosent av feilene i identifikasjonsnummeret som antas å være tilstrekkelig på grunn av lav sannsynlighet for feil fordi identifikasjonsnummeret blir automatisk pre-kodet på skjemaene allerede ved utsending i de fleste anvendelser. Integrasjonen av registrene skjer ved at foretaksnumrene nyttes som kjennemerker i bedriftsregisteret. Innføringen av permanente identifikasjoner har tvunget fram en drøfting av definisjoner for „fødsler“, „flytting“ og „dødsfall“ for både foretak og bedrifter. Foretak og bedrifter kan ha forskjellig levetid, og dette var årsaken til at vi fant det nødvendig å innføre uavhengige nummer for foretak og bedrifter, slik at betingelsen om permanent identifikasjon blir oppfylt.

Bedrifts- og foretaksregistrene ble opprettet med statistiske undersøkelser for øye. Det har senere vist seg et meget stort behov for tjenester fra registrene, ikke minst for næringslivet. I sin nåværende form kan de imidlertid ikke dekke viktige offentlige registerbehov som for eksempel de behov skattemyndighetene har.

Et vanskelig problem med hensyn til bedrifts- og foretaksregisteret er tilgang av nye enheter. En har her måttet nytte mange kilder hvorav den viktigste har vært trygdevesenet. Annet vedlikehold dekkes ved hjelp av såkalte Navnekort som sendes ut med jevne mellomrom til registerenhetene. Dette er enkle skjemaer, til dels automatisk utfylt fra Byrået med opplysninger fra registeret som mottakeren bes korrigerer og supplere om nødvendig.

I forbindelse med kommende jordbruksteling drøfter en muligheten av å utvide registeret også til denne næring.

### 3.4. Arbeidsgiver-register.

Skattemyndigheter og trygdevesen har i forbindelse med overføring av likningsarbeidet til staten og innføring av folketrygden, et behov for et landsomfattende register over arbeidsgivere til bruk ved direkte og indirekte beskatning, innkreving av trygdeavgift etc. Enhetene vil dels være personlige arbeidsgivere og dels ikke-personlige firmaer eller selskaper. Det er ønskelig at arbeidsgiverne i dette register til en viss grad kan karakteriseres ved den næringsgruppe hvor de hører hjemme. Byråets foretaksregister har derfor vært vurderet som en mulig utgangspunkt, og det foreligger nå forslag om at Byrået gis i oppdrag å etablere et arbeidsgiverregister og holde det vedlike på grunnlag av meldinger fra de lokale skatte- og trygdeorganer.

Fra statistisk synspunkt er et slikt register av spesiell interesse fordi det vil kunne supplere Byråets foretaksregister med enheter som det ikke er praktisk mulig å få med i den nåværende situasjon. Vedlikeholdsmeldingene vil dessuten kunne løse problemet med registrering av tilgangen bedre enn det er mulig i dag.

Byrået har derfor skissert oppbygging av et arbeidsgiver-register på grunnlag av person- og foretaksregistrene i et felles registersystem for personer og foretak. En forutsetter at personregisteret utvides til også å omfatte alle juri-

diske, ikke-fysiske personer som får sine permanente identifikasjonsnummer. I stedet for de nåværende foretaksnummer som ble omtalt ovenfor, vil foretakene få eierens nummer som identifikasjon. På denne måte vil en oppnå en naturlig og effektiv koordinering av identifikasjonene for fysiske personer, juridiske personer, foretak og avledede enheter som arbeidstakere og arbeidsgivere, skattytere etc. Med andre ord, ethvert foretak vil som foretaksnummer få samme nummer som eieren — enten han er en fysisk eller ikke-fysisk person — er tildelt i personregisteret, enhver skatteyter — enten han er person eller selskap — vil få samme nummer som han er tildelt i personregisteret, osv.

Det er ennå ikke tatt noe standpunkt til når en slik utbygging og koordinering skal skje og fra hvilket tidspunkt en vil starte vedlikeholdet av arbeidsgiverregisteret. En regner imidlertid med at arbeidet muligens vil ta til allerede i inneværende år.

### 3.5. Sentralt jordregister.

Utenfor Byrået arbeides det med å samle inn materiale for å etablere et sentralt register over jordstykker som bl. a. vil bli koordinatbestemt, bonitets-karakterisert, størrelsesmålt og knyttet sammen med det gamle gårds- og bruksnummersystem. Det er interesse for å få Byrået til å påta seg det tekniske vedlikehold av dette sentrale jordregister.

## 4. Data-innsamling

Investering i data-kapital skjer ved *innsamling* av data enten direkte fra den enkelte enhet eller — samlet for en gruppe av enheter — fra en institusjon som har hentet inn data for administrative formål. Administrative data som tilfredsstillende de krav statistikken stiller, vil i alminnelighet være rimeligere å kopiere enn å gå til en direkte innsamling av tilsvarende data.

Ved å arbeide systematisk for å introdusere våre sentrale registre og kjennemerkedefinisjoner i flest mulig administrative prosesser kan vi gjøre en stadig større del av de administrativt innhentede data brukbare for statistikken. I Norge synes personnummersystemet nå å ha slått meget godt an. Stadig flere institusjoner

tar sikte på å gå over til dette. Vi har også et berettiget håp om at det samme vil skje om vi får bygd opp et arbeidsgiverregister slik vi mener det bør være. Vi satses således relativt meget på å bygge opp og gi register-service. Det vi håper å få igjen for innsatsen er en stadig økende masse av brukbare data fra administrative kilder.

Denne utvikling vil trolig erstatte noe av den direkte innsamling vi hittil har foretatt. Det er f. eks. allerede nå klart at det ikke vil være nødvendig å gi de store tellinger den samme bredde som hittil. Sannsynligvis vil de få karakter av undersøkelser med sikte på å kontrollere registrenes fullstendighet og å skaffe opplysninger om forhold som ikke kan skaffes på annen måte.

De sentrale registre og data-arkiver vil begrense den direkte innsamlingen bare til de data en tidligere ikke har registrert. En Fiskertelling som nå planlegges i Norge for gjennomføring i slutten av året, vil f. eks. bare omfatte data som ikke er kjent fra Fiskertellingen i 1960 og senere folkeregistreringer. De data som nå hentes inn vil gjennom personregisteret bli knyttet sammen med data fra 1960 Fiskertelling m. m., og denne kombinerte datamasse vil utgjøre det en ellers måtte samle inn ved en fiskertelling.

## 5. Data-arkiver

### 5.1. Arkiveringsmåter.

Det konkrete uttrykk for det teoretiske begrep datakapital er *dataarkivene*. De kan være ordnet på mange måter, men det arkiv-statistiske systems betingelser er at hver observasjonsverdi må være ledsaget av en permanent enhetsidentifikasjon og av kjennemerke- og tidsspesifikasjon. Vi skal her begrense oss til å drøfte de data-arkiver som tilfredsstillende denne betingelse. Arkivene kan ordnes etter *enhet*, *art* eller *tid* alt etter hvilken av de tre typer identifikatorer som har dannet grunnlaget for arkivenes hovedsortering. Den optimale ordning vil avhenge både av den ordning data kommer inn i og den ordning de kreves i av statistikkproduksjonen.

I Statistisk Sentralbyrå blir data-arkivene holdt på magnetbånd som ennå synes å være den mest effektive lagringsmåte. I første om-

gang blir data nå arkivert på art/tid-ordnede bånd, dvs. data fra hver undersøkelse eller administrativ kilde blir lagret på egne bånd for hver årgang. Innenfor hvert bånd vil data være ordnet etter enhets- og kjennemerke-identifikatorer.

### 5.2. Persondata-arkiver.

Det sentrale personregisteret gir befolkningsmessige data for hver person som kan følges fra folketellingen 1960. En får også relasjoner mellom foreldre og barn registrert. Slike relasjonskjennemerker gjør det mulig å knytte en rekke indirekte kjennemerker til de enkelte enhetene. Barn kan f. eks. ved hjelp av foreldrenes kjennemerker karakteriseres ved en rekke miljøforhold. I forbindelse med registrering av fødsler og dødsfall, vil en også få en rekke medisinske data. Allerede de arkiv-data en nå får inn, vil gi nye og meget vide muligheter til å analysere fødsels-, ekteskaps- og dødsforhold. Like ens vil flyttinger nå kunne studeres med langt større intensitet enn forholdene tidligere har tillatt. I 1970 vil det derfor neppe bli behov for en folketelling etter gammelt mønster.

Fra inntektsåret 1967 vil vi få skatte- og inntektsdata for alle personlige skattytere med personidentifikasjon som vil gjøre det mulig å knytte disse sammen med data fra de befolkningsmessige data-arkiver, og følge inntekts- og formuesutviklingen på enhetsbasis. Det vil gi grunnlag for mer intensive inntekts- og formuesanalyser enn de vi i dag har høve til å utføre og til å studere virkning på de enkelte skatte-enheter av tiltak som blir satt i verk. Det vil også gi muligheter for å studere endringer i ekteskaps- og fødselshyppigheter ut fra en økonomisk bakgrunn.

Senere vil en kunne føye både retts-, sosial-, helse- og undervisningsstatistikken inn i systemet. En får her ikke bare bedre grunnlag for sosiologiske analyser, men ved å knytte disse data til befolkningsmessige og økonomiske data for de samme enheter åpnes et nytt og viktig område for sosio-økonomiske analyser med sikte på samspillvirkningen mellom de sosiale og økonomiske forhold. Av konkrete prosjekter som her har vært drøftet er dødelighetsanalyser etter livs- og miljøforhold, og utdannelsesmodeller etter inntekts- og miljøbakgrunn.

Om det sentrale arbeidsgiverregister blir etablert slik som skissert ovenfor, vil det gjennom oppgaver over trekk av skatt som alle arbeidsgivere plikter å sende inn om sine ansatte, foreligge et materiale som inneholder relasjoner mellom arbeidstaker og arbeidsgiver. Denne relasjon kan utnyttes til å karakterisere arbeidstakere med arbeidsgiverens kjennemerker, og gjør det mulig å trekke inn forholdene i foretaket som medvirkende forklaringsfaktorer for beslutninger og handlinger som personene foretar, f. eks. sosiologiske studier med arbeidstedets miljø som bakgrunnsfaktorer.

Jordregisteret og til dels personregisteret omfatter begge de nå brukte gårds- og bruksnummer for henholdsvis beliggenhet og bosted. Det første registrerer dessuten geografiske koordinater. Gjennom gårds- og bruksnummeret har vi her mulighet for indirekte å karakterisere personer med den geografiske koordinat for beliggenheten av deres bosted. Dette vil gjøre det mulig i personstatistikk å definere geografiske områder, uavhengig av og mindre enn de administrative inndelinger, og det antas å kunne bli av stor betydning for næringsgeografiske studier og regionalplanlegging.

Administrative kilder vil i stor utstrekning kunne forsyne arkivene med data om personer. Et kjennemerke som imidlertid foreløpig synes å være vanskelig å få registrert tilfredsstillende uten direkte observasjon, er yrke.

### 5.3. Bedrifts- og foretaksdata-arkiver.

Innføring av permanente identifikasjonsnummer i bedrifts- og foretaksregistre tillater oss å lage data-arkiver for bedrifts- og foretaksstatistikk fra og med året 1964. Disse arkivene vil som utgangspunkt få data fra bedriftstelingen for 1963. For de aller største foretak, ca. 600 foretak og 1 300 bedrifter, er Byrået nå i ferd med å knytte sammenhengen tilbake til 1959.

Disse data-arkivene gir allerede i dag en beredskap med hensyn til optellinger av enheter, som utnyttes i såkalt registerstatistikk. Etter hvert som en får flere permanent identifiserte årganger i arkivene, vil disse gi grunnlag for studium av tilgangs- og avgangsårsaker, bedriftenes og foretakenes livs-syklus, etc. Studier som allerede er foretatt peker i retning av at f. eks. aldersfordeling av bedrifter kan bli

en langt nyttigere klassifikasjon enn vi hittil har trodd. Et annet område som trolig vil bli av stor interesse er beregning av produktfunksjoner med tids-lag, analyser av investeringsadferd, lagerhold og andre adferdsfunksjoner på grunnlag av bedrifts- og foretaksdata.

Stort sett vil en rekke data om bedriftene fortsatt måtte hentes inn direkte etter som det ikke finnes noe administrativt organ som er primært interessert i dem. Men blir det sentrale arbeidsgiverregister koordinert med foretaksregisteret, vil vi for foretakene få en rekke identifiserte data fra administrative kilder. I første omgang vil dette bli fra skattevesenet. I tillegg til de personlige skattytere vil vi få identifiserte inntekts- og skattedata for selskaper, og begge typer kan knyttes sammen med andre foretaksdata. Dessuten vil det foreligge regnskaps- og oppgjørsoppgaver som foretak i en del næringer plikter å sende inn til likningsvesenet. Disse oppgaver kan bli et nyttig supplement eller kanskje erstatning for direkte innhentet materiale for regnskapsstatistikk. Et koordinert foretaks-arbeidsgiverregister vil føre til at detaljomsetningsoppgaver som innhentes administrativt i forbindelse med omsetningsavgiften kan kombineres med andre data om foretakene og vil kunne gi et bedre grunnlag for korttidsstatistikk for og analyser av detaljomsetningen.

Opgavene over skattetrekk som ble nevnt under personstatistikk, vil også kunne bli brukt til å knytte f. eks. personlige kjennemerker for foretakslederne som indirekte kjennemerker til foretakene. Her vil en kunne studere foretakenes utvikling ut fra f. eks. ledernes utdannelses- og erfaringsbakgrunn.

Generelt vil data-arkivene trolig bidra til at virkningen av endringer i fordelinger av forskjellig slag vil komme langt sterkere fram i problemstillinger og analyser.

#### 6. Statistikk-produksjonen

Utnytting av data- og statistikk-arkivene til beregning av statistiske størrelser for grupper av enheter er her kalt statistikk-produksjon. Hittil har som nevnt statistikk-produksjonen i vesentlig grad vært innstilt på at data rasjonelt bare kunne bearbeides en gang, og bearbeidingen måtte derfor være preget av å gi flest mulig generelt anvendelige resultater.

Analytikere har dessuten stort sett vært tvunget til å foreta sine beregninger på aggregerte størrelser. Når data-arkivene nå vil gjøre gjentagne bearbeidinger av samme data-masse praktisk mulig, vil ikke motivene for å produsere generelle resultater være så sterke lenger. I stedet vil innsatsen kunne konsentreres om flere beregninger direkte for spesielle analytiske behov.

Det kreves stadig økt og bedre kjennskap til den makønisme som er bestemmende for samfunnsutviklingen. Hittil har modellbygningen vært basert på relativt få og summariske aggregater. Data-arkivene vil her gi et langt mer omfattende materiale for presis estimering av koeffisienter, og tillate at det i modellene tas med relasjoner som makrodata ikke gir grunnlag for å estimere.

I Norge har vi ennå ikke nådd så langt at vi kan vise til omfattende erfaringer på dette felt, men vi vil i når framtid ta opp spørsmålet om planer for utnytting av data-arkivene for nye undersøkelser og analyser.

#### 7. Statistikk-arkivene

Statistikk-kapitalen oppbevares i *statistikk-arkivene* hvor statistiske beregningsresultater identifiseres med gjennomgående gruppe-, kjennemerke- og tidsidentifikasjoner. Slike statistikk-arkiver har i den senere tid ofte blitt betegnet som statistikk-banker og blir holdt på hullkort og magnetbånd for raskt å kunne kjøre fram den statistikk en bruker har behov for dersom den hører til systemets faste produksjonsopplegg eller allerede er produsert som en spesiell bearbeiding.

På dette felt er det hittil for statistikken over utenrikshandelen og for nasjonalregnskapet en har kommet lengst i Norge. For statistikken over utenrikshandelen blir det løpende holdt et statistikk-arkiv som er mer detaljert enn det som mangfoldiggjøres månedlig. Foreløpig blir dette statistikk-arkiv utnyttet i en abonnementsordning som tillater brukere på en effektiv måte å få statistisk informasjon utover det som publiseres. Når det gjelder nasjonalregnskapet, blir hovedboken holdt på hullkort og gir et beredskap for videre beregninger på grunnlag av nasjonalregnskapsoppstillinger.

Statistikk-arkivene kan teknisk holdes i be-

redskap i et data-maskinsystem. Med en eller flere tilkoblede spørre-stasjoner kan en hvilken som helst arkivert opplysning hentes fram uten ventetid. Dette vil kunne få stor betydning i framtiden for effektiv informasjonsspredning fra statistikk-systemet.

#### 8. Afslutning

I det foregående er det pekt på en del problemer som reiser seg ved gjennomføring av de idéer som er kalt et arkiv-statistik system. Det er imidlertid en rekke andre problemer som det ikke har vært høve til å ta opp. Av disse bør imidlertid to nevnes.

Det er her understreket den store betydning sentrale og permanente enhetsregistre må ha i et arkiv-statistisk system. Like viktig er imidlertid standardisering av kjennemerke-definisjoner

på en slik måte at de data en har om den enkelte enhet er konsistente på samme måte som en har søkt å gjøre de aggregerte statistiske størrelser konsistente gjennom nasjonalregnskapssystemer etc. Et stykke på veg er en kommet ved utarbeiding av statistiske klassifikasjonsstandarder, men det gjenstår trolig meget både med hensyn til å få disse gjennomført i administrative prosesser og med hensyn til å etablere begrepsstandarder og -systemer på områder hvor det ikke er utarbeidd noe ennå.

Det andre spørsmål som bør nevnes er den fare for eventuelt å gjøre skade ved misbruk av de data som etter hvert samles opp i arkivene. En positiv reaksjon fra publikum på utviklingen i retning av et arkiv-statistisk system vil sannsynligvis i høy grad avhenge av at systemet utvikles med sikte på at risikoen for misbruk blir så liten som mulig.