Det personstatistiske registersystem

Seminar
10.-13. januar 1994

DANMARKS
STATISTIK

Forord

Siden oprettelsen af CPR i 1968, hvor man samtidig introducerede personnummeridentifikationen, er der sket en voldsom udbygning af det administrative registersystem. Foruden at tjene administrative formål er der gennem dette system skabt mulighed for at indhente data til en lang række statistikformål. Der har fra starten været megen opmærksomhed omkring dette, jf. Lov om Danmarks Statistik, idet man på denne måde begrænsede omkostningerne ved indsamling af oplysninger, reducerede belastningen af borgerne med udfyldelse af spørgeskemaer meget betydeligt, opnåede muligheder for ny statistik og udbygning af eksisterende statistik samt hævede statistikkens kvalitet gennem, i mange tilfælde, mere præcise oplysninger og uden den bortfaldsproblematik, som ofte er forbundet med den traditionelle spørgeskemabaserede dataindsamling. Dertil kommer, at man ikke i samme grad som tidligere er henvist til stikprøvebaseret statistik, hvilket i sig selv øger kvaliteten.

Der rejser sig imidlertid en række spørgsmål i forbindelse med anvendelsen af administrative registre som statistikgrundlag. Kan vi opnå den information, der er behov for? Hvilke metoder kan anvendes til udbedring af fejl og mangler? Hvordan organiseres de store datamængder, der er tale om, på en hensigtsmæssig måde? Er der sikkerhed for individuel anonymitet? Disse og beslægtede spørgsmål er blevet belyst og diskuteret på et seminar afholdt af Danmarks Statistik i dagene 10. - 13. januar 1994.

Seminarets indlæg vil efter redaktionelle tilpasninger og eventuelle tilføjelser på udækkede områder blive udgivet i bogform i såvel en dansk som en engelsk udgave. Det vil ske i sommeren 1994.

Da der er en betydelig interesse for hurtigere at gøre sig bekendt med seminarets indhold, har Danmarks Statistik valgt at udsende denne dokumentsamling i begrænset oplag. Samlingen indeholder de papirer, som blev forelagt på seminaret, idet forfatterne dog efterfølgende har haft lejlighed til at revidere indholdet under hensyntagen til de synspunkter og bemærkninger, der fremkom under seminaret.

Der har været lagt afgørende vægt på, at udsendelsen af denne dokumentsamling skete meget hurtigt. Den redaktionelle bearbejdelse har af den grund været meget begrænset, hvorfor der, når disse oprindeligt enkeltstående dokumenter ses i sammenhæng, kan forekomme uhensigtsmæssig disponering af stoffet. Det vil naturligvis blive rettet op i den endelige udgave.

Redaktionsgruppen

Indholdsfortegnelse:

Kravene til den officielle statistik

Lars Thygesen

Indledning

Hvordan kan man bedst tilrettelægge en officiel statistik baseret på administrative kilder, og er det i det hele taget fornuftigt? En diskussion af disse spørgsmål må tage udgangspunkt i den opgave, statistikken skal løse. Opgaven er meget sammensat, fordi den officielle statistik er en del af et lands *generelle* infrastruktur, og den benyttes til en lang række forskellige formål.

Brugerne af statistikken

Den officielle statistik skal give en bred og dybtgående beskrivelse af samfundet. Den skal i hvert fald dække flg.: Befolkningsforhold, sociale forhold, helbred, erhvervsforhold, offentlig økonomi og den nationale økonomi[1]. Statistikproduktionen skal indrettes, så den til enhver tid belyser de væsentlige træk i samfundsforhold og -problemer.

Offentligheden

Brugerne af denne information er i første række *offentligheden*. Statistikken skal give et godt grundlag for den generelle samfundsdebat, som den finder sted i hjemmene, på arbejdspladserne, i de politiske organisationer o.s.v. Til denne brugergruppe kan man også henregne undervisningen i skoler og på højere læreanstalter. Den vigtigste informationskanal for statistikken til offentligheden er nyhedsmedierne: Aviser, radio og TV. Derfor er det vigtigt for Danmarks Statistik at offentliggøre statistikker, der bliver citeret og viderebragt af medierne. Det stiller både krav til indholdet og til den måde, nyhederne bringes på. En del af offentligheden vil dog også være direkte brugere af de officielle statistiske publikationer

Lovgivningsmagten

En gruppe af brugere af den officielle statistik, som altid har været anerkendt som uhyre vigtig, er *lovgivningsmagten*. I nogle lande har det endda været god latin, at kun de krav til statistikken, som blev stillet af regering og parlament, blev tillagt større betydning, men Lov om Danmarks Statistik gør det klart, at dette ikke er tilfældet i Danmark. Men lovgivningen stiller på mange områder krav om meget fyldige og detaljerede statistiske data, som skal gøre det muligt at bedømme, hvilke behov for lovgivning der er, og hvordan påtænkte love vil virke. Disse krav bliver mere og mere udtalte, efterhånden som lovgivningen på det skattemæssige og det sociale område bliver stadig mere kompliceret.

Den offentlige administration

Den officielle statistik bruges også til andet end lovgivningsformål af *den offentlige administration:* Staten, kommunerne og amtskommunerne. Statistikken tjener som basis for planlægning af virksomheden og for opfølgning og resultatvurdering.

Erhvervslivet

Et tilsvarende behov findes i *det private erhvervsliv*, som har brug for at få belyst markedsmæssige muligheder, konkurrenceforhold m.m.

---

[1] Jf. Danmarks Statistik (1991)

Forskningen

Endelig er *samfundsforskningen* en vigtig brugergruppe. Den omfatter såvel de sociale videnskaber som grænseområder inden for medicin og naturvidenskaber.

Internationale organisationer

Alle de nævnte grupper er velkendte på det nationale informationsmarked, men hertil kommer, at også *internationale organisationer* har brug for statistik. De internationale organisationer skal til syvende og sidst tilfredsstille de samme typer af behov, som allerede er nævnt, idet de skal betjene offentligheden, virksomhederne m.fl., som ønsker at kunne foretage sammenligninger af forholdene fra land til land. De internationale behov er omfattende og bliver i disse år i Europa stadig mere påtrængende. EUs forordninger og direktiver på statistikområdet er bindende for medlemslandene, og de krav, der følger heraf, skal altså under alle omstændigheder opfyldes.

Hvem bestemmer indholdet?

Der er altså en række grupper, der har brug for den officielle statistik, og som derfor stiller krav til den. Men der er også andre interessenter: Dem der skal finansiere statistikken, og dem der skal levere oplysninger til den.

Styrelsen bestemmer...

I Danmark er ordningen den, at man har Danmarks Statistik som den centrale statistikmyndighed, jf Lov om Danmarks Statistik, §1. Institutionen ledes af en uafhængig styrelse, som fastlægger arbejdsplanen inden for de budgetmæssige rammer, som Finansloven giver. Dette er en prioriteringsopgave, hvor de forskellige brugergruppers behov skal afvejes over for hinanden, idet de konkurrerer om ressourcerne. Juridisk bindende krav til statistikken skal naturligvis forlods opfyldes, herunder de krav, som EU stiller. I prioriteringen skal der også tages hensyn til respondenterne, som normalt må antages at føle det som en byrde, at der skal afgives oplysninger til statistikken. Der ligger en særlig forpligtelse for Danmarks Statistik til at minimere *respondentbyrden* og til at maksimere *effektiviteten*.

... og får hjælp fra rådgivende udvalg

Til at hjælpe med prioriteringsprocessen har styrelsen nedsat en række rådgivende udvalg med deltagelse af både brugere af statistikken og oplysningsgivere. Et udvalg beskæftiger sig typisk med et bestemt statistik-tema, fx socialstatistikken. Enige indstillinger fra disse udvalg har naturligvis stor vægt, men det er styrelsens opgave at påse, at programmet kan rummes i den økonomiske ramme.

Indsamlingsmetoder

Surveys

I de fleste lande indsamles data til den officielle statistik helt overvejende ved at man spørger de enheder, som statistikken omhandler: Personer, virksomheder, kommuner. Normalt bruger man mere eller mindre ensartede skemaer, som respondenterne enten selv skal udfylde og indsende, eller som udfyldes af en interviewer. Man kan spørge alle de enheder, der skal indgå i statistikken (totaltælling), eller man kan nøjes med en del af dem, som så skal udvælges i overensstemmelse med forskellige videnskabelige principper for

at man kan sikre sig, at de repræsenterer hele populationen (stikprøveundersøgelser). Alle disse typer af undersøgelser går under fællesbetegnelsen *surveys*, og der findes en meget omfattende litteratur om, hvordan de kan eller bør tilrettelægges, og om deres fejlkilder og pålidelighed.

Problemer ved surveys

Det er et fælles træk ved surveys, at de belaster de medvirkende, og de giver i mange lande anledning til stor modvilje. I nogle lande (fx USA) har man haft parlamentariske udvalg, der har haft til opgave at stille forslag om, hvordan man kunne nedbringe respondentbyrden. Både som følge af arbejdet med at give oplysninger og på grund af respondenternes frygt for misbrug af data forekommer der altid et vist bortfald i surveys, således at man ikke får svar fra alle de adspurgte. Dette er en af de mest alvorlige fejlkilder ved metoden, idet erfaringen viser, at de, der ikke svarer, altid adskiller sig systematisk fra de mere medgørlige, fra hvem man opnår svar. I næsten alle undersøgelser, hvor man spørger personer, er der fx langt større bortfald blandt arbejdsløse end blandt andre og større blandt byboer end på landet[2].

Registre

En anden måde at fremskaffe statistikkens grunddata på er at udnytte allerede indsamlede oplysninger. En meget rig kilde for den officielle statistik er de data, som den offentlige administration indsamler om borgere og virksomheder. Det er ikke noget nyt at udnytte disse oplysninger til statistik. Nogle af de ældste statistikker fremkom på denne måde. Det gælder således de tidligste danske statistikker om befolkningsudviklingen, som byggede på kirkebøgerne.

I de sidste 25 år er der imidlertid sket en dramatisk ændring i statistikproduktionen, idet en meget stor del af statistikken efterhånden baseres på edb-registre, der føres af den offentlige administration. Dette hænger naturligvis sammen med, at administrationen også har undergået voldsomme ændringer, og der er siden etableringen af CPR i 1968 opstået en lang række registre, som alle benytter nogle få fælles identifikationer for de objekter, administrationen retter sig imod: Borgere, virksomheder, boliger og ejendomme. På denne måde kan de forskellige registre kombineres og skabe et nuanceret billede af samfundet.

Når det har været muligt for Danmarks Statistik at udnytte de nye muligheder i meget høj grad, skyldes det, at Lov om Danmarks Statistik i §6 indeholder en stærk hjemmel til at indhente oplysningerne fra de offentlige myndigheder. Da loven blev udformet i 1966, var forberedelsen af de nye registre begyndt, og lovgiverne så klart de kommende registres muligheder til statistikformål. Dette fremgår af lovens bemærkninger, som omtaler mulighederne for at nedbringe respondentbyrden og endda komme så vidt, at en folketælling i fremtiden måske kunne gennemføres helt uden spørgeskemaer, udelukkende gennem brug af registrene. Lovgiverne har her været meget fremsynede, idet disse tanker først kunne realiseres i 1981.

Registerstatistikken i internationalt perspektiv

Interessen for den registerbaserede statistik har naturligvis ikke kun gjort sig gældende i Danmark. Den har været livligt diskuteret i det internationale

[2] Jf fx ...

samarbejde siden slutningen af 1960'erne. Blandt de førende lande har fra tidlig tid været Norge[3] og Sverige, og registerstatistikken har været fast punkt på nordiske statistikermøder og chefstatistikermøder. Interessen har været stor i USA[4]. Også i det europæiske samarbejde i EF har der været megen diskussion om spørgsmålet, navnlig i forbindelse med overvejelser om at basere folke- og boligtællinger på registre[5]

Diskussionen har afspejlet en blanding af optimistiske forventninger og skepsis, en skepsis som har ført til, at det kun er ganske få lande i verden, der er gået så vidt som Danmark i retning af at basere den officielle statistik på registre.

Indvendinger mod registerstatistikken

De indvendinger, der især rejses mod den registerbaserede statistik, og som vil blive diskuteret indgående i denne bog, er:

- Krænkelse af privatlivets fred. Registerstatistikken forudsætter, at statistikbureauet opbevarer en stor mængde oplysninger om personer og virksomheder, hvilket skaber en risiko for misbrug, dvs. brug af oplysningerne til andre formål end fremstilling af anonyme statistikker.

- Folkelig modvilje, som hænger sammen med ovennævnte punkt. Uanset om der reelt er risiko for misbrug, kan modviljen være så stærk, at data ikke kan fremskaffes. Dette og det foregående punkt diskuteres i kap. 7.

- Datakvalitet. Der argumenteres for, at registrene ikke indeholder lige præcis de begreber, der er *relevante* for statistikken, og at data ikke er *pålidelige*. Grunden er, at statistikerne ikke kan styre indholdet i registrene, som de er vant til fra deres egne undersøgelser. Diskuteres i kap. 4.

Den danske registerstrategi

I Danmark har disse indvendinger naturligvis også været gjort gældende, men holdningen har gennemgående været langt mindre pessimistisk. I midten af 1970'erne besluttede Danmarks Statistik således, at det fremover var en strategisk målsætning at udvikle et sammenhængende system for personstatistikken baseret på data fra administrative registre. Strategien omfattede, at man skulle arbejde for at få udbygget de administrative registre med nogle få oplysninger, der var vigtige for statistikken. Et succesmål for strategien var, at folke-og boligtællingen 1981 skulle gennemføres med fuldt informationsindhold og uden brug af spørgeskemaer. Dette mål blev nået. Senere har også Finland i 1990 gennemført en registerbaseret folke- og boligtælling.

Den danske strategi har betydet, at man i hvert fald på det personstatistiske område har udviklet et sammenhængende system, hvor rygraden er oplysninger fra registrene. Alle databehov kan ikke imødekommes af denne vej. Fx er der ikke mulighed for at fremskaffe 'bløde' data om holdninger ad denne vej,

---

[3] En banebrydende artikel er således *Norbotten (1967)*

[4] Se fx *National Center for Health Statistics (1980)* og *US Department of Commerce (1980b)*

[5] *Redfern (1987)*

og disse statistikker må indhentes vha. surveys, som også indgår i systemet, men betragtes som et supplement.

Det har været afgørende for udviklingen i Danmark, at lovgivningsmagten som nævnt fra starten har stillet sig meget positivt til, ja ligefrem forudsat en sådan udvikling. De stramme økonomiske vilkår for Danmarks Statistik har yderligere tilskyndet til valget af strategi[6]

[6] *Nordic Statistical Secretariat 1981, kap. 4*

Referencer

Danmarks Statistik (1991). *Om Danmarks Statistiks målsætning.*

Harala, R.: *Evaluation of the Results of the Register Based Population and Housing Census 1990 in Finland.* Bulletin of the ISI, Contributed Papers, 49th Session, Firenze 1993

National Center for Health Statistics (1980): *The Person-Number Systems of Sweden, Norway, Denmark, and Israel.* Data Evaluation and Methods Research, Series 2, No. 84. Hyattsville, Maryland

Norbotten, S. (1967). *Om arkivstatistiske systemer.* Bilag 4 i Statistical Reports of the Nordic Countries, vol. 14. København 1967

Nordic Statistical Secretariat (1981). *The Meeting of the Chief Statisticians of the Nordic Countries in Copenhagen 1979.* København 1981

Redfern, P. (1987). *A Study of the Future of the Census of Population: Alternative Approaches.* Eurostat Theme 3 Ser. C. Luxembourg

US Department of Commerce (1980b). *Report on Statistical uses of Administrative Records.* Statistical Working Paper No. 6. Washington DC

Statistikkens grundlag: Administrative registre i Danmark

Carsten Torpe

1. Om begrebet register

1.1 Generelle betragtninger

Betegnelsen register anvendes i mange forskellige betydninger. I bredeste forstand omfatter registerbegrebet enhver logisk afgrænset samling af data, herunder også datasæt beregnet til rent midlertidigt brug. Ordet register defineres dog ofte langt snævrere. Fx opererer registerlovgivningen med et begreb, hvor det afgørende er, om oplysningerne kan henføres til bestemte personer.

I denne fremstilling vil vi som hovedregel opfatte et *register* som et systematisk sæt af informationer om en bestemt gruppe af enheder (fx personer, firmaer), hvor enhederne er afgrænset af et præcist regelsæt (fx "bopæl i Danmark"), og hvor indholdet ajourføres i takt med de ændringer, objektgruppen undergår.

Et register må altså omfatte et system til indrapportering af dataændringer. Ajourføringen kan finde sted med større eller mindre forsinkelse i forhold til den hændelse, der forårsager ændringen. I nogle tilfælde kan indrapporteringen ikke ske i tilslutning til hændelsen, således at man i stedet må nøjes med at konstatere de korrekte data med visse tidsintervaller; det gælder således nogle af de data, der findes i bygnings- og boligregistret, BBR. Data, der slet ikke kan ajourføres systematisk, betragtes ikke som egentlige registerdata.

Den anførte definition af registre refererer udelukkende til indholdet af de datasamlinger, der er tale om. Registrenes oplysninger kan opbevares på forskellig måde. Et register kan eventuelt opdeles i flere delregistre af praktiske årsager, og der kan anvendes manuelle og/eller maskinelle processer ved vedligeholdelsen, men disse forhold spiller ingen rolle for registerbegrebets indhold.

Det må dog fremhæves, at de mere omfattende administrative registre, som gennem de seneste 10 - 20 år har været fundamentet for statistikproduktionen i Danmark, alle drives med anvendelse af elektronisk databehandling, og de erfaringer, som skal fremdrages her, vedrører derfor næsten udelukkende edb-registre.

Af særlig interesse for statistikken er registre, som drives med en mangesidet anvendelse som formål, såkaldte basisregistre. Folkeregistersystemet, samlet i det centrale personregister (CPR), er det mest oplagte eksempel på et basisregister, idet formålet med dette system er at levere almindelige personoplysninger til brug overalt i den offentlige administration.

Et andet basisregister er erhvervsregistret, som registrerer grundoplysninger om selvstændige erhvervsdrivende og arbejdsgivere til brug for den offentlige forvaltning og andre. Tilsvarende må bygnings- og boligregistret betragtes som et basisregister.

De fleste registre oprettes imidlertid til brug for én bestemt forvaltningsopgave, fx skatteligning eller beregning og udbetaling af sociale ydelser.

1.2 Statistikregistre og administrative registre

Sondringen mellem administrative registre og statistikregistre er af basal betydning for forståelsen af de registerstatistiske problemstillinger, som tages op i denne bog.

I Danmarks Statistiks planlægningshåndbog, OPUS, defineres et *statistikregister* som et eller flere afgrænsede, færdigbehandlede, og færdigredigerede datasæt, som udgør fundamentet for tabellering og analyser på et bestemt statistikområde. Formålet er mao. at fremstille statistiske opgørelser, som ikke tillader identifikation af den enkelte enhed. De specificerede oplysninger i et statistikregister må under ingen omstændigheder benyttes som grundlag for administrative afgørelser i enkeltsager. På den anden side er det et væsentligt formål med statistik, at den skal kunne danne grundlag for overvejelser om udformning af love og regler, som har betydning for den enkelte.

Modsat står de *administrative registre*, hvis data indsamles som led i den administrative proces med det formål at danne grundlag for individuelle afgørelser angående den enkelte borger, familie, virksomhed osv.

Et statistikregister kan oprettes på grundlag af traditionel statistisk data-indsamling, fx en folketælling eller en survey-undersøgelse, eller statistikregistret kan være bygget på kopier af eller uddrag fra administrative registre, som er underkastet en særlig statistisk bearbejdning. Det er navnlig den sidstnævnte produktionsmetode, som behandles i denne bog.

En række af de omhandlede problemer - fx tidsafgrænsning, fejlretning og komplettering af datamangler - er i princippet fælles for statistikregistre og administrative registre, men de metoder og løsninger, som kan benyttes i forbindelse med oparbejdning af statistikregistre (fx imputering og maskinel fejlretning på grundlag af mekaniske beslutningsregler), kan ikke uden videre overføres til de administrative registre. Dette skyldes, at de to typer af registre, som følge af deres forskellige formål, har forskellige datakrav.

Registre, der udelukkende anvendes i statistisk eller videnskabeligt øjemed, er i henhold til Lov om offentlige myndigheders registre underkastet særlige regler. Der er således stærkt begrænsede muligheder for videregivelse af

oplysninger, og de registrerede har ikke ret til egen-access. For samkøring af registre til statistiske formål gælder ligeledes særlige regler, idet adgangen hertil er friere end for administrative registre.

2. Opbygningen af administrative registre

2.1 Personregistrering

Danmark er blandt de lande i verden, som i dag har de mest udbyggede, integrerede og velfungerende administrative edb-systemer. Udviklingen startede tidligt på personregistreringsområdet, og med udgangspunkt heri er der gennem 1970'erne og 1980'erne foregået en bevidst og omfattende opbygning af landsdækkende, administrative systemer.

2.1.1 Folkeregistrene og CPR

I løbet af 1800-tallet indførte mange vestlige lande en løbende og systematisk registrering af borgernes bopæls- og civilstandsforhold i såkaldte personregistre. Sådanne registre blev ikke dengang oprettet i Danmark, men i begyndelsen af det 20. århundrede opstod der som følge af væksten i den offentlige administration - og i de kommunale og statslige udgifter - et øget behov for pålidelige oplysninger om borgernes opholdssted. Således viste det sig vanskeligt for kommunerne at inddrive alimentationsbidrag og at opkræve skatter, fordi skatteyderne ikke kunne findes, og der var desuden et mere specielt behov for at etablere en effektiv administration af rationeringsordningerne efter 1. verdenskrig. Man blev også opmærksom på, at personregistre kunne tjene andre formål. Således ville de være egnede ved politiets eftersøgning af personer, som grundlag for udskrivning af valglister og som grundlag for befolkningsstatistik.

I 1924 vedtog Folketinget derfor "Lov om Folkeregistre", som bestemte, at hver af landets kommuner samme år skulle oprette et folkeregister, dvs. et kartotek indeholdende oplysning om alle personer, der havde bopæl i kommunen, uanset om de måtte være midlertidigt fraværende.

Folkeregistrene skulle indeholde identifikationsoplysninger som stilling og navn, fødselsdag og -sted, og herudover var de vigtigste oplysninger bopælsadresse, familieforhold og statsborgerforhold. Disse grundoplysninger blev indhentet gennem folketællingen i 1924. Kommunerne skulle holde kartotekerne løbende ajourførte, idet de skulle modtage oplysninger om fødsler, dødsfald, vielser, skilsmisser m.m. fra forskellige myndigheder, mens borgerne selv blev forpligtet til at indberette flytninger samt ind- og udvandringer direkte til folkeregistret.

I den debat, der gik forud for folkeregisterlovens vedtagelse, indgik også forslag om et landsregister, et manuelt drevet CPR i form af et kæmpekartotek, som skulle være en sammensmeltning af de kommunale folkeregistre omfattende hele Danmarks bosiddende befolkning. Dette forslag nød heldigvis ikke fremme, da man kunne forudse effekterne i form af en enorm administrativ byrde.

Folkeregistrene blev i de følgende år gradvist udbygget som hjælpemiddel for den kommunale og den øvrige offentlige administration. Der blev indført formularsæt til bl.a. skattemyndigheder og sygekasser, således at disse myndigheder kunne føre egne personregistre. Efterhånden anvendtes hulkortteknik ved løsning af periodevise masseopgaver, såsom skatteberegning og udskrivning af valglister. Til løsning af disse opgaver oprettedes kommunalt ejede, regionale hulkortcentraler. Anvendelsen af folkeregistrene var imidlertid i begyndelsen af 1960'erne efterhånden udbygget så kraftigt, at systemet var ved at nå en kapacitetsgrænse, som nødvendiggjorde en strukturel og teknisk ændring.

Med folkeregistreringsloven af 10/6 1968 gennemførtes en afgørende reform, som havde til formål at forenkle og effektivisere den samlede offentlige personregistrering ved anvendelse af edb-teknik. Ved siden af de kommunale folkeregistre, som fortsatte med at fungere, oprettedes det centrale personregister, CPR, et landsdækkende magnetbåndsregister over den danske befolkning.

En afgørende del af reformen var indførelsen af et fast identifikationsnummer for hver enkelt borger, *personnummeret*. Dette nummer betragtedes som en praktisk nødvendighed for driften af det centrale personregister (*). Nummeret skulle tillige indføres overalt i den offentlige administration og således afløse de talrige nummersystemer, som tidligere havde været anvendt af forvaltningsgrenene.

I forbindelse med etableringen af CPR blev en række andre standarder og koder fastlagt, fx kommunekoder, sognekoder, stillingskoder og statsborgerskabskoder. Væsentligst var *adressekoden*, som entydigt identificerer den enkelte adresse, en meget central del af CPR-systemet.

Oprettelsen af CPR skyldtes først og fremmest ønsket om at undgå dobbeltregistrering og det ekstra ressourceforbrug, som fulgte heraf. Medvirkende har også været udsigten til en skattereform med indførelsen af kildeskat, som vanskeligt kunne gennemføres uden et meget sikkert system til identifikation af landets borgere.

CPR blev fra starten vedligeholdt via de fælleskommunale edb-centraler gennem løbende indberetninger fra kommunerne, som igen fik indberetninger fra borgerne og en lang række offentlige myndigheder. Der er naturligvis gennem årene gennemført adskillige forbedringer og moderniseringer. I dag sker opdateringen dagligt gennem online terminaladgang fra kommunerne. CPR består af flere dele, de vigtigste er:

- *Personregistret* med aktuelle og historiske oplysninger om hver person, herunder personnummer, henvisninger til forældre og evt. børn og ægtefælle, adresse, navn, civilstand, statsborgerskab.

- *Boligregistret*, som for alle adresser (bopæle) rummer oplysninger om boligen (fra BBR) og dens lokalitet.

• *Vejregistret* med oplysninger om alle veje og bynavne i landet. Hver vej (vejstykke) har en vejkode, et navn og oplysning om vejens (eller vejstykkets, angivet ved husnr.-interval) placering i fx. sogn, postdistrikt, politikreds eller i kommunalt fastsatte inddelinger som skoledistrikt mv.

Alle offentlige myndigheder har adgang til at få oplysninger fra CPR. CPR leverer dagligt eller ugentligt status- og/eller ændringsudtræk til godt 75 abonnerende myndigheder. Desuden har en lang række myndigheder terminaladgang til at foretage landsdækkende søgning og opslag i hele CPR's personkreds, dvs. personer som bor eller har boet i Danmark efter 1. april 1968.

Private virksomheder kan også rekvirere oplysninger fra CPR, forudsat at virksomheden selv, i henhold til lov om private registre, kan levere de pågældende personnumre. Det drejer sig i praksis alene om virksomheder i den finansielle sektor.

2.1.2 Andre administrative personregistreringer

Med oprettelsen af CPR i 1968 var Danmark det første land, der indførte entydige personnumre og adressekoder. Tidspunktet var overmåde velvalgt, idet CPR stod klar *før* udviklingen på edb-området for alvor var slået igennem i administrationen - altså *før* der var sat fart i opbygningen af store, landsdækkende systemer.

CPR's sikre identifikationssystem og basale personoplysninger blev således helt fra starten føderegister for næsten alle andre større, offentlige administrative systemer, og også for mange private systemer, især i den finansielle sektor. I modsætning til næsten alle andre lande undgik Danmark på denne måde, at de forskellige administrative registre "opfandt" hvert sit ID-system, og der var derfor i Danmark ingen tekniske hindringer for at indføre den nuværende effektive og ressourcebesparende edb-infrastruktur.

Det fælles identifikationssystem er forudsætningen for, at man næsten overalt - skattevæsen, bankvæsen, sociale sager, vægtafgifter, radio- og TV-licenser osv. - baserer administrationen på løbende oplysninger fra CPR om navne, adresser, ægteskabelig stilling mv. kombineret med de specifikke sagsdata på det pågældende område.

Det er indenfor rammerne af denne bog ikke muligt at komme nærmere ind på indholdet i de administrative systemer. Registertilsynet har i dag registreret forskrifter for ca. 2.800 personregistre, hvoraf langt hovedparten er administrative registre mens et mindre antal er registre oprettet i statistik- eller forskningsøjemed. Registrene fordeler sig således (omtrentlige tal):

- 1050 statslige registre (heraf 60 statistikregistre i Danmarks Statistik)
- 80 fælleskommunale registre (på Kommunedata I/S)
- 370 enkelt-amtslige registre
- 1300 enkelt-kommunale registre

En betydelig del af disse registre omfatter kun en begrænset personkreds og tjener ganske afgrænsede og/eller lokale formål. Hovedinteressen, når det gælder en statistisk anvendelse, samler sig naturligvis om de administrative, landsdækkende statslige og fælleskommunale systemer, som findes på næsten alle områder, herunder:

- Skatteregistre
- Arbejdsmarkedsregistre
- Sociale registre
- Sundhedsregistre
- Hospitalsregistre
- Registre over uddannelsessøgende
- Registre på erhvervsområdet

De fleste store administrative fællesopgaver løses ved hjælp af landsdækkende, ensartede, centrale systemer samlet på Kommunedata og Datacentralen, hvilket i betydelig grad har lettet de statistiske anvendelsesmuligheder. Der har imidlertid i de seneste år været en tendens til at vælge decentrale kommunale løsninger, fx gælder det på daginstitutionsområdet og aktiveringsområdet, hvilket ikke i sig selv hindrer en statistisk anvendelse, men øger besværet med dataindsamling, dækningsgrad og harmonisering af uensartede datasæt.

2.2 Ejendomsregistrering

Ejendomsregistreringen her i landet kan siges at være startet med oprettelsen af matriklen i 1662, hvis formål var jordbeskatningen. Opmåling af de registrerede arealer og deres grænser blev senere foretaget og indført i registret. Siden begyndelsen af dette århundrede har matriklen mistet sin direkte tilknytning til ejendomsbeskatningen, som så blev foretaget med udgangspunkt i særlige vurderingsregistre.

I 1972 oprettedes et landsdækkende edb-ejendomsregister til varetagelse af ejendomsbeskatning og ejendomsvurdering. Registret føres af kommunerne, og grundlaget for registreringen er oplysninger i matriklen og i tingbøgerne. Foruden matrikulære oplysninger mv. registreres forskellige data af betydning for vurderingen.

I 1977 indsamledes i forbindelse med ejendomsvurderingen oplysninger til bygnings- og boligregistret, BBR, som ligeledes føres kommunalt, og som rent teknisk er en udbygning af ejendomsregistret. Formålet med denne udbygning var, at registerkomplekset skulle udgøre et egentligt basisregister for ejendomme, bygninger og boliger. Oplysningerne i BBR vedrører bygningernes og boligernes alder, størrelse, installationer og anvendelse. Registret ajourføres med meddelelser om ændringer, nybygninger og nedrivninger, idet disse oplysninger indhentes i forbindelse med kommunernes byggesagsbehandling.

Såvel i CPR som i BBR indgår adresseoplysningen. Ved løbende kommunikation mellem de to basisregistre sikres, at adresserne er identiske, således at oplysningerne i registrene kan sammenkobles.

I 1980'erne påbegyndtes etableringen af de kommunale planregistre over offentlighedsretlige rådighedsindskrænkninger såsom kommune- og lokalplaner, byfornyelses- og varmeplaner. Planregistret indeholder således oplysninger om den planlagte (og i en række tilfælde tillige den faktiske) arealanvendelse.

Et særligt register er krydsreferenceregistret - et nøgleregister, hvor alle ændringer og ajourføringer af adresser, matrikelbetegnelser, ejendomsnumre, planbetegnelser og tilhørende geokoder (koordinater i form af en geografisk reference) edb-mæssigt samles. Herved er de forskellige ejendomsregistre bragt til at spille sammen, ligesom samspillet med personregistreringerne er sikret.

2.3 Erhvervsregistrering

Opbygningen i 1960'erne af administrative erhvervs- og virksomhedsregistre til brug for bl.a. ATP, toldvæsenet og skattevæsenet medførte indsamling af erhvervsdata, som også Danmarks Statistik kunne benytte til statistiske formål, herunder ajourføring af statistiske erhvervsregistre. De generelle erhvervstællinger, som Danmarks Statistik tidligere havde gennemført med tiårige intervaller, kunne på denne baggrund ophøre.

De administrative registres registrering af erhvervsvirksomhederne og arbejdsgiverne var imidlertid i den periode, registeropbygningen fandt sted, kun i mindre omfang koordineret. Der var mange, men ofte små forskelle med hensyn til de forskellige registres enhedsdefinitioner, dataindhold mv. Endvidere anvendte hvert register sit specielle nummersystem til identifikation af enhederne.

Med henblik på en forenkling af det offentliges erhvervsregistrering gennemførtes i 1975 en lov om oprettelse af et centralt erhvervsregister.

Registret, der administreres af Danmarks Statistik, optager oplysninger om selvstændige erhvervsdrivende og arbejdsgivere, som i henhold til anden lov er berettiget eller forpligtet til registrering hos offentlige myndigheder eller institutioner. Registret optager endvidere oplysninger om filialer eller afdelinger under de oven for nævnte objekter.

Erhvervsregistrets oplysninger om firmaernes navn, adresse og indehavere registreres og ajourføres hovedsageligt ved direkte overførsel af oplysninger fra Told og Skattestyrelsens registre. Danmarks Statistik tilfører oplysninger om brancher og særskilt beliggende afdelinger, filialer mv. (arbejdssteder) samt supplerende oplysninger om registerobjekterne.

I loven om erhvervsregistret forudsættes indført en entydig fælles nummerering af registerenhederne til brug for det offentliges erhvervsregistre.

Mulighederne for indførelse af fælles registreringsregler og enhedsnumre blev forbedret væsentligt, da toldvæsenet og skattevæsenet i 1990 blev samlet til Told og Skattestyrelsen under Skatteministeriet. Told og Skattestyrelsen har oprettet et fælles stamregister (SE-registret), der sikrer, at enheder, der er registreringspligtige for moms og kildeskat mv. får samme nummer i de forskellige delsystemer. Der er dog endnu nogle problemer ved opbygningen af stamregistret, men disse forventes at være overvundet omkring årsskiftet 1995/96. Når stamregistret er fuldt udbygget kan det herved sikres, at erhvervsregistret kan fremtræde som en bestand af entydigt afgrænsede enheder med et fælles identifikationsnummersystem.

3. Administrative registre som statistikkilde

3.1 Udviklingen i hovedtræk

Selv om folkeregistrene i 1924 blev oprettet med administrative opgaver som det helt klare hovedformål, var man fra begyndelsen klar over, at registrene tillige skulle anvendes til forbedring af befolkningsstatistikken. Folkeregistrene blev derfor pålagt at indsende flyttebeviser for samtlige flyttede personer hvert kvartal til Det Statistiske Departement. Samtidig skulle sendes en særlig statistikindberetning, som udviste totaltal for antal personer i registret ved kvartalets begyndelse og slutning, hvortil kom en optælling af kvartalets befolkningsbevægelser af kategorierne: fødte, døde, indenlandske til- og fraflytninger samt vandringer til- og fra udlandet.

På grundlag af flyttebeviserne blev det muligt at fremstille en detaljeret statistik over flytningerne mellem kommuner og over ind- og udvandringer. Statistikindberetningerne benyttedes til en summarisk opgørelse af "registerbefolkningen og dens bevægelser". Denne opgørelse måtte naturligvis være bundet af indberetningsformen og kunne fx ikke opdeles mht. alder og køn. Hovedparten af den detaljerede befolkningsstatistik hvilede fortsat på andre kilder.

Oprettelsen af CPR i 1968 medførte en gennemgribende omlægning af statistikproduktionen. Det blev nu folkeregistrene, der via CPR kom til at danne basis for alle grene af befolkningsstatistikken. Nogle opgørelser henter fortsat oplysninger fra andre kilder, men disse oplysninger opfattes som supplement til data fra CPR.

Indførelsen af den nye produktionsteknik gav i de første år anledning til betydelige vanskeligheder. Allerede i 1969 startede man forsøgsvis en befolkningsstatistik baseret på et CPR-materiale, som skulle beskrive befolkningens sammensætning og alle typer af befolkningsbevægelser. CPR-statistikken for 1969 og 1970 blev imidlertid særdeles mangelfuld, og det var først fra og med 1973, at det samlede statistiksystem kunne betragtes som oprettet.

Vanskelighederne i starten skyldtes ikke alene mangler og fejl i CPR's dataindhold, men også tolkningsproblemer som følge af upræcis og utilstrækkelig kommunikation mellem statistikproducent og registermyndighed.

Det befolkningsstatistiske system har nu gennem mange år fungeret næsten problemfrit. Og i denne periode har registerbaseret statistikproduktion over en bred front indtaget en dominerende rolle. Som omtalt har Danmarks Statistik nu ansvaret for 60 personorienterede statistikregistre, som i hovedsagen hviler på løbende dataleverencer fra administrative edb-systemer. Dog baseres enkelte statistikregistre fortsat på hel eller delvis manuel dataindsamling, og nogle af registrene er dannet ved horisontal integration, dvs. ved tværgående anvendelse af eksisterende statistikregistre; et godt eksempel på sidstnævnte er den sammenhængende socialstatistik, jf afsnit 3.2, men nævnes kan også visse registre oprettet i forbindelse med forskningsopgaver. En oversigt over hovedpunkter i udviklingen siden 1968 ses i afsnit 3.3.

Det er klart, at denne udvikling kun har været mulig som følge af den tidligere omtalte hurtige udbygning af det offentliges personregistrering, som er enestående for Danmark, og den lidt senere etablerede centrale registrering af virksomheder, ejendomme mv.

Udvidelsen af registreringsaktiviteten er ofte sket udelukkende med administrative formål, men i andre tilfælde er hensynet til statistikproduktionen indgået med stor vægt i begrundelsen for oprettelse af nye registre eller udvidelse af bestående registres indhold. Det var således af stor betydning for vedtagelsen i 1976 af loven om det nye bygnings- og boligregister BBR, at registret kunne danne grundlag for en bolig- og beboerstatistik; i 1974 var det nemlig blevet besluttet ikke at afholde den traditionelle folke- og boligtælling året efter under den forudsætning, at boligstatistik på lidt længere sigt kunne fremskaffes på registergrundlag. Ved opbygningen i 1960'erne af administrative erhvervs- og virksomhedsregistre var de statistiske formål ligeledes med i de forudgående overvejelser.

3.2 Eksempel: Det sociale område

På det sociale område kom overgangen til registerbaseret statistikproduktion først for alvor i gang i 1980'erne, men baggrunden, udviklingsmønsteret og implikationerne kan genfindes indenfor andre personstatistikområder.

Der er en lang tradition for, at Danmarks Statistik står for indsamling og offentliggørelse af socialstatistiske oplysninger. Før 1980 foregik dataindsamlingen overvejende via blanketter med summariske data suppleret med tabeludskrifter fra de fælleskommunale edb-centraler. På dette grundlag - og med en betydelig manuel indsats - produceredes den løbende socialstatistik.

Dataene kunne imidlertid ikke imødekomme det stigende behov for flexible og tværgående opgørelser med baggrundsdata fra den øvrige personstatistik, herunder lovforberedende analyser (Lovmodellen). På denne baggrund etab-

leredes i første halvdel af 1980'erne et stort antal personorienterede social-statistikregistre baseret på årlige udtræk fra de fælleskommunale kommunale udbetalingssystemer.

I dag er der på praktisk taget alle sociale områder, fx sociale pensioner, sygedagpenge, sygesikring, boligstøtte og børnefamilieydelse, oprettet personregistre - 14 i alt - med 100% dækningsgrad. Statistikken om sociale ressourcer, dvs. om institutioner, personale og indskrevne, er dog kun delvis baseret på personregistre, da der på dette område endnu ikke findes tilstrækkeligt dækkende administrative systemer. Det samme gælder statistikken om bistandsydelser til børn og unge, hvor personoplysningerne stadig (i dag en undtagelse) indsamles via manuelt udfyldte blanketter.

Det omfattende og detaljerede personorienterede datagrundlag forøgede radikalt mulighederne for at tilgodese de forskellige statistikbehov. Den løbende statistik omfatter både statusopgørelser om antal modtagere ved årsskiftet og årsbaserede opgørelser om personer berørt af ydelser i kalenderåret. Der indgår desuden baggrundsdata om familie, bolig, indkomst og beskæftigelse mv. hentet fra eksisterende personstatistikregistre. Foruden den almindelige publicering leveres efter aftale særlige tabelsæt til ministerier, styrelser og kommuner. Dataene indgår desuden i databankerne, og i Lovmodellen benyttes anonymiserede persondata på udsnitsbasis.

En egentlig *tværgående socialstatistik* blev etableret i 1992, men statistikken rummer data fra og med 1984. "Den sammenhængende Socialstatistik" er et integreret statistikregister over modtagere af indkomsterstattende ydelser, samkørt (på personnummer) ved udtræk fra følgende områder: Arbejdsløshed, sygedagpenge, kontanthjælp, folkepension, førtidspension, efterløn og tjenestemandspension. De enkelte udtræk fra delregistrene harmoniseres og samkøres til det endelige register, som suppleres med oplysninger om indkomst- og familieforhold. Det færdige register rummer ca. 1,7 mio. records (svarende til antal berørte personer på årsbasis). Registret anvendes bl.a. i forbindelse med forskningsopgaver og lovforberedende analyser.

Det nye datagrundlag gav desuden mulighed for at oprette et nyt modul i KÅS - Kommune Års Service - som giver kommunerne mulighed for at købe statistik fordelt på vilkårlige geografiske delområder. Der findes i dag seks KÅS-moduler: befolkning, bolig, beskæftigelse, indkomst, pendling og - det nyeste - social. *KÅS-social* blev søsat i 1991. KÅS-social rummer oplysninger om samtlige modtagere og udbetalinger af sociale ydelser fordelt efter modtagernes bopæl. Ydelserne belyses dels enkeltvis - social pension, efterløn, tjenestemandspension, kontanthjælp, dagpenge ved sygdom, graviditet og fødsel, boligstøtte, børnetilskud og daginstitutionsbenyttelse - dels i form af tværgående opgørelser, incl. arbejdsløshedsdagpenge, hentet fra den sammenhængende socialstatistik. Fordelinger foretages efter ydelsens art, demografiske forhold, familie- og husstandstype, erhverv og boligforhold

3.3 Oversigt: Milepæle i registerstatistikken siden 1968

Danmarks Statistiks virksomhed samt de vigtigste ændringer og planer omtales i Danmarks Statistiks arbejdsplan, som udgives årligt. I det følgende omtales nogle hovedpunkter i udviklingen, specielt hvad angår anvendelsen af edb-teknik og registermetoder:

1. Med virkning fra 1968 etableredes en statistik vedrørende omsætning mv. i samtlige erhverv, baseret på oplysninger fra momsadministrationen kombineret med data fra Danmarks Statistiks erhvervsregister.

   Fra 1971 iværksattes en tilsvarende statistik vedrørende erhvervenes beskæftigelse og fra 1974 tillige om udbetalt lønsum baseret på arbejdsgivernes indberetninger til ATP-ordningen henholdsvis skattemyndighederne.

   Efter etableringen af disse opgørelser er udarbejdelse af generelle erhvervstællinger, der blev udarbejdet for 1925, 1935, 1948 og 1958, ophørt.

2. Pr. 1. maj 1970 udarbejdedes første gang en totalopgørelse af befolkningen i de enkelte kommuner med fordeling på køn, alder og ægteskabelig stilling. I perioden 1970-73 gennemførtes en omlægning af den øvrige befolkningsstatistik til CPR-basis, således at manuelle indberetninger fra kommunerne kunne ophøre.

3. I 1971 anskaffede Danmarks Statistik eget edb-anlæg.

4. Fra 1970 omlagdes indkomststatistikken i forbindelse med kildeskattens indførelse til registerbasis. Fra 1976 rekonstrueredes statistikken væsentligt, bl.a. ved at indkomstoplysningerne samkørtes med oplysninger fra et arbejdsklassifikationssystem, som også anvendes til andre formål, jf. punkt 7.

5. I 1971 etablerede Danmarks Statistik et "udsnitsarkiv" til brug for udviklingen af registerstatistiske metoder, herunder vedrørende samkøring, og for udarbejdelse af analyser og ad hoc opgørelser.

6. Efter beslutning i 1971 overtog Danmarks Statistik i 1973 fra Undervisningsministeriet statistikken vedrørende elever og studerende. De individuelle oplysninger, som undervisningsinstitutionerne indberetter, indgår i et uddannelsesstatistisk register, således at institutionernes løbende indberetninger kun skal omfatte ændringer i forhold til året i forvejen.

7. I 1974 besluttede økonomiministeren, at der ikke skulle gennemføres en traditionel, skemabaseret folketælling i folketællingsterminen 1975/76, jf. lov om Danmarks Statistik. I stedet gennemførtes for 1976 en registerbaseret folketælling på basis af CPR-oplysninger kombineret med en række forskellige andre registeroplysninger vedrørende erhvervsforhold mv.

8. Fra 1977 etableredes en bistandslovsstatistik baseret på personorienterede oplysninger fra det fælleskommunale økonomisystem suppleret med skemaindberetninger fra de ikke-tilsluttede kommuner og kombineret med familieoplysninger fra befolkningsstatistikken.

9. Pr. 1.4.1977 gennemførtes den første bygnings- og boligopgørelse på basis af det nyoprettede Bygnings- og boligregister, BBR. Pr. 1. januar 1980 gennemførtes den første registerbaserede boligtælling, som kombinerer oplysninger om boligerne med oplysninger om beboerne.

10. I 1979 besluttedes det at oprette et Arbejdspladsstatistikregister baseret på eksisterende registeroplysninger og en begrænset supplerende dataindsamling med bistand af skattemyndighederne og arbejdsgiverne. Dette indebærer bl.a., at totale dataindsamlinger fra befolkningen i form af traditionelle folketællinger definitivt er ophørt.

11. Med virkning fra 1979 blev statistikken om strafferetlige afgørelser, som før var summarisk og skemabaseret, omlagt til at basere sig på årlige indberetninger fra Rigspolitiets centrale kriminalregister.

12. Pr. 1.1.1981 gennemførtes den første egentlige folke- og boligtælling på registerbasis.

13. I løbet af 1982-84 blev oprettet en stribe sociale statistikregistre, som bygger på udtræk fra de landsdækkende fælleskommunale udbetalingssystemer. En betydelig manuel arbejdsbyrde ophørte samtidig med, at de statistiske anvendelser blev drastisk forøget (jf. afsnit 3.2).

14. I 1990 besluttedes en helt ny edb-strategi med hovedvægten lagt på decentralisering, som blev implementeret i 1991/92. Hver medarbejder fik PC-adgang opkoblet til netværk, som igen er koblet til den centrale computer. En række nye programmer, værktøjer og decentrale arbejdsgange blev indført som standard.

15. I 1992 oprettedes motorkøretøjsstatistikregistret baseret på udtræk fra Rigspolitiets centrale motorregister. Registret danner grundlag for opgørelser om ejerforhold og anvendelse af motorkøretøjer.

16. I 1992 oprettedes en ny sundhedsstatistik baseret på udtræk fra Landspatientregistret kombineret med en række baggrundsoplysninger fra eksisterende statistikregistre om bl.a. befolknings-, social-, indkomst-, beskæftigelses-, uddannelses- og boligforhold.

17. I 1994 etableres en udvidet statistik om personer i aktiveringsordninger og tilbagetrækningsordninger baseret på personorienterede indberetninger fra de administrative myndigheder.

18. I 1994 etableres, efter flere års omfattende udviklingsarbejde, et samlet, moderniseret erhvervsregistersystem med en række indholdsmæssige og tekniske forbedringer.

19. I 1994-95 gennemføres en omfattende lønstatistisk reform, som vil føre til samlede og konsistente personorienterede opgørelser for alle lønmodtagere i både den private og den offentlige sektor.

Litteratur:

Betænkning afgivet af folkeregisterkommissionen af 1920.
København 1922.

Betænkning om folkeregistrenes medvirkende ved indførelse
af elektronisk databehandling i den offentlige forvaltning mv.
København 1963.

Delbetænkning om offentlige registre.
Betænkning nr. 767. København 1976.

Lov om offentlige myndigheders registre.
Lov nr. 294 af 8. juni 1978.

Lov om private registre mv.
Lov nr. 293 af 8. juni 1979.

Registre indenfor sundhedsområdet.
Rapport udgivet af DIKE, København november 1982.

Helsetjenesteforskning og registre.
Rapport fra Udvalget vedrørende Helsetjenesteforskning
og Medicinsk Teknologivurdering. København marts 1986.

Fortegnelse over offentlige myndigheders registre 1987.
Registertilsynet, København 1987.

Intern registerfortegnelse.
Registertilsynet, maj 1993.

Oversigt over forskrifter for fælleskommunale registre.
Finansministeriet, Administrations- og Personaledepartementet, 1993.

CPR, Danmarks folkeregister.
Notat, Henrik Nielsen, Indenrigsministeriet, marts 1991.

CPR-systemet i 1993.
Notat fra Indenrigsministeriet, CPR-kontoret, 1993.

OPUS, Om Planlægning og Udvikling af Statistiksystemer.
Danmarks Statistik, planlægningshåndbog, udgivet 1990, opdateres løbende.

Registeroplysninger til statistikformål

Vøgg Løwe Nielsen

Indledning.

Afsnittet omhandler de krav, som de grundlæggende registre bør kunne opfylde for at tilfredsstille registerstatistikkens behov. Det er klart, at kravene ikke kan forventes opfyldt helt i praksis, og der omtales nedenfor en del eksempler på, hvilke problemer sådanne mangler kan medføre for registerstatistikken.
Herudover indeholder afsnittet en kort diskussion af de indholdsmæssige forskelle ved henholdsvis register- og surveystatistik.

Der skal gøres opmærksom på, at store dele af afsnittet er identisk med eller er en let revideret udgave af dele af kapitel 5 i "personstatistik på registergrundlag". Afsnit 5.2 om "registret som statistisk model" er dog udeladt med henblik på behandling i forbindelse med OPUS-problematikken.

Statistikkens krav kan opdeles i indholdsmæssige og systemmæssige krav.

Indholdsmæssige krav.

De indholdsmæssige krav er det vanskeligt at behandle generelt. Det drejer sig altså om at afgøre, hvilke oplysninger der skal findes registreret, for at man på grundlag af registrene kan få en "rimelig" dækning af statistiske behov.
Der har været udbredt enighed om, at det navnlig er af betydning for den samfundsbelysende statistik, at der i det administrative system findes følgende registreringer:

a) Et personregistersystem med en bred vifte af oplysninger:
Demografiske, beskæftigelsesmæssige, sociale, uddannelsesmæssige mv.

b) Et boligregistersystem, som også omfatter de lokaliteter i bygningsmassen, der benyttes til erhvervsformål

c) Et erhvervsregistersystem, som omfatter både de juridiske enheder og de lokale enheder, arbejdsstederne.

Betegnelsen registersystem skal her opfattes meget bredt, idet fx personregistersystemet kan bestå af en række registre, der føres for forskellige myndigheder. For at hvert af disse registersystemer kan betragtes som en helhed, er det nødvendigt, at der findes et fælles identifikationssystem for hver af de tre typer af enheder, hvilket også er forudsætningen for, at rekombination af registeroplysninger fra forskellige myndigheder kan finde sted. Det er endvi-

dere af afgørende betydning, at der kan etableres forbindelser mellem de tre registersystemer og deres identifikationssystemer indbyrdes:

d) En nøgle mellem personer og de boliger, de bebor.

e) En nøgle mellem personer og de virksomheder (arbejdsstederne), hvor de er beskæftiget.

f) En nøgle mellem arbejdsstederne og de lokaliteter, de optager.

Er disse indholdskrav opfyldt helt eller tilnærmelsesvis, har man redskaberne til at opbygge et registerstatistisk system, som på grundlag af de administrative data kan give opgørelser af folke- og boligtællingstyper, og som tillige kan ligge til grund for væsentlige dele af den øvrige statistik vedrørende personer, boliger og lokale erhvervsenheders forhold.

Forskellige typer registerdata

I det følgende vil det blive illustreret, hvordan registerdata kan inddeles i nogle kategorier efter det formål, det enkelte datum tjener i registret. Det har nemlig vist sig, at nogle af de krav, som den statistiske anvendelse stiller til registrene, knytter sig til bestemte typer af registerdata. Desuden må visse datatyper betegnes som særligt problematiske i relation til statistikken.
Et bestemt datum kan optræde i forskellige roller i forskellige registre; fx er boligens adressekode identifikation i bygnings- og boligregistret, mens den i CPR er et basisdatum.

Identifikationsdata

En grundlæggende forudsætning for effektiv registerdrift er, at identifikationssystemet er entydigt. Dette forudsætter igen, at to betingelser er opfyldt, nemlig (a) at enhederne er entydigt defineret, og (b) at identifikationssystemet er præcist og hensigtsmæssigt udformet samt "robust" i forhold til almindeligt forekommende fejlmuligheder.
Hvad det første krav angår, er mulighederne for at imødekomme det stærkt varierende fra enhedstype til enhedstype. I nogle tilfælde (personer, biler, skibe, fly) er der efter sagens natur intet - eller næsten intet - problem, i andre tilfælde (ejendomme, erhvervsenheder, bygninger osv.) må der forsøges opstillet operationelle definitioner eller retsregler, der på den ene side tager udgangspunkt i objektivt konstaterbare forhold, på den anden side bedst muligt modsvarer de til registreringen knyttede formål.
Kravet til identifikationssystemets udformning er af mere teknisk natur, og de styringsprocesser, som skal sikre mod dobbeltnummerering og mod fejl i anvendte numre (checkcifre) er velkendt. Spørgsmålet om, hvorvidt identifikationen skal være informationsløs eller informationsbærende, har også været debatteret i forbindelse med indretningen af de fleste nummersystemer. Det her i landet vigtigste identifikationsnummer - personnummeret - indeholder som bekendt visse informationer (fødselsdato og køn), men da disse er (næsten) uforanderlige, volder dette ikke styringsmæssige problemer.

Tværtimod kan man sige, at netop en fødselsdato er en hensigtsmæssig oplysning at lade indgå i identifikationen, bl.a. for at gøre det let at huske, men det kan desværre ikke praktiseres for alle typer af enheder.
Det samlede registervæsen bygger på nogle få generelle identifikationssystemer: Personnummer, adressekode (for boliger mv.), materikelnr. for ejendomme, virksomheds- eller firmanummer. Dette er afgørende for de faktiske integrationsmuligheder. Hver af disse identifikationssystemer må af hensyn til en effektiv drift styres ét sted, og de registre, som varetager denne funktion, får derfor en særlig betydning.
Disse styrende, primære identifikationer ligger til grund for de enkelte registres drift og for samspillet imellem disse, men de dækker ikke alle identifikationsbehov. Der anvendes fx en række udadvendte identifikationer, hvoraf navn og adresse for personer selvsagt er vigtige. Oplysning om fødselsregistreringssted er et andet vigtigt element i en personidentifikation, der bl.a. etablerer forbindelsen mellem folkeregistersystemet (inkl. CPR) og "civilstandsregistrene" (kirkebøgerne).
Identifikationerne tjener også til at udtrykke enhedsrelationer. For eksempel registreres i CPR relationen "person A er gift med person B" ved, at B's personnummer er anbragt i et særligt felt i person A's CPR-record (og omvendt).
En registreret enhedsrelation kan samtidig være af betydning som sekundær identifikation, som tilfældet fx er, når ejeren til et firma registreres i dettes record i erhvervsregistrene.

Tidsreferencer

Tiden er en afgørende dimension i vor daglige opfattelse og tolkning af fænomener i verden omkring os.
Da det statistiske register skal opbygges som en model til beskrivelse af en større eller mindre del af denne virkelighed, må tiden nødvendigvis indgå som en vigtig del af registret.
I det følgende skal der peges på forskellige former for tidsreferencer, som man kan håbe at finde i et administrativt register, idet der drages en hovedsondring mellem dateringer, der tidsfæster fænomener i virkelighedens verden og dateringer, der tidsfæster vor måling eller registrering af disse.
Med hensyn til dateringer, der tidsfæster forhold af den førstnævnte art, kan man sondre mellem dateringer knyttet til selve enhedens eksistens, og dateringer, der knytter sig til bestemte variable.
Dateringer, som knytter sig til enhedens eksistens, er de samme som afgrænser enhedens levetid, dvs. fødsels- og dødstidspunkt. Disse dateringer vil typisk være centrale i basisregistre, men kan også være vigtige i andre statistiske registre, hvis disse dækker så lang en periode, at der sker en tilgang og afgang af enheder.
Dateringer, som knytter sig til andre variable, må opdeles efter disse variables karakter.
Det er her vigtigt at sondre mellem variable, hvis værdi kun kan fastlægges i en periode (strømvariable) og variable, hvis værdi kan fastlægges på et tidspunkt (beholdningsvariable).
Som eksempel på strømvariable kan nævnes en persons bruttoindkomst, og det må i et statistisk register, der indeholder strømvariable, være vigtigt at datere den periode, variabelværdien refererer til.

Som eksempel på beholdningsvariable kan nævnes civilstand eller uddannelse, og man kan her sondre mellem to forskellige former for datering af disse.
Dels kan man tale om en datering i diskret tid, hvor man kun kender de pågældende variables værdi på et (eller flere) givne tidspunkt (er), dels kan man tale om en datering i kontinuert tid, hvor man kender den periode, inden for hvilken de pågældende variabelværdier gælder. Ved datering i kontinuert tid får man samtidig en datering af variabelværdiændringer (hændelser), som vil være afgørende for, om der kan udarbejdes en egentlig forløbsstatistik.
Endelig kan det som nævnt være vigtigt at datere, hvomår målingen eller registreringen af et bestemt fænomen er foretaget, og der går her en hovedsondring mellem retrospektiv og simultan måling.
Ved en retrospektiv måling foretages målingen en kortere eller længere tid efter fænomenets forekomst, og måske foretages der flere retrospektive målinger af det samme forhold. I disse tilfælde er det vigtigt at kunne holde resultatet af de forskellige målinger ude fra hinanden.
Ved en simultan måling vil der normalt ikke være det samme behov for en eksplicit datering af målingen, idet denne med god tilnærmelse vil falde sammen med de dateringer, der knytter sig til de forskellige variable i registret.

Basisdata

Ved registermæssige basisdata forstås i det følgende sådanne registeroplysninger, som er beregnet til en mangesidet anvendelse, herunder til fremtidige, måske endnu ukendte formål. Begrebet svarer på dataplanet til det tidligere indførte begreb basisregister.
Det mest udprægede eksempel på sådanne oplysninger er hovedparten af CPR's dataindhold og særligt den del, som er "folkeregisterdata", fx bopæl, civilstand, nationalitet.
Udsondringen af disse gennemgående data fra øvrige registerdata har betydning i en række henseender:

a) Lovgivningen vil hyppigt benytte basisdata som parametre (tildelingskriterier mv.), men det anses som regel ikke for nødvendigt at fastlægge det begrebsmæssige indhold gennem lovgivningen, fordi der er tale om almindeligt anerkendte begreber.

b) Datasikkerhedsmæssigt set har de en særstilling, eftersom de efter sagens natur må få en betydelig udbredelse. Hertil kommer, at de på grund af deres almene karakter, som regel ikke objektivt set er særligt følsomme (jf. at enhver kan få oplysninger om en række af erhvervsregistrets data).

c) Oplysningeme må begrebsmæssigt set være "objektive" og "neutrale" og mindst mulig præget af forhold på specielle administrative områder.

d) Datakvalitet og ajourføringsgrad må være høj, idet manglende nøjagtighed på anvendelsestidspunktet kan give anledning til

administrativt besvær og ubehagelige konsekvenser for de registrerede.

Hvis oplysningerne skal opbevares i flere registre, er konsekvensen, at ajourføringsoplysninger må overføres mellem registersystemerne. Hvis de derimod skal opbevares ét og kun ét sted, er konsekvensen, at andre systemer skal hente deres data her, når de skal anvendes. I begge tilfælde vil en form for samkøring være en uundgåelig konsekvens.

Specialdata

Ved specialdata forstås i denne sammenhæng data, som indsamles direkte med henblik på funktionelt at indgå i bestemte administrative processer eller som er et resultat af sådanne. De mest typiske eksempler herpå skal formentlig hentes i kildeskattesystemets registre (lønmodtagerfradrag, lejeværdi af egen bolig) og i de sociale systemer (dagpenge, sygesikring osv.).
Selv om der i forhold til basisoplysningerne er mange grænsetilfælde, er der klare principielle skillelinier til disse i en række henseender:

a) Det begrebsmæssige indhold af specialdata vil oftest være fastlagt gennem specielle lovregler og administrative forskrifter.

b) Datasikkerhedsmæssigt set vil de ofte være mere "følsomme" end basisdata, men også genstand for mindre cirkulation.

c) Oplysningerne må nødvendigvis være defineret i forhold til formålene og underkastet de begrebsmæssige ændringer, som følger af lovændringer mv.

d) Datakvalitet og ajourføringsgrad bestemmes ud fra de konkrete anvendelsesbehov.

e) Registerfunktionelt må de "fødes" i den pågældende administrationsgrens regi.

Baggrundsdata

Foruden de ovenfor omtalte datatyper, som altid er direkte nødvendige for den administrative proces, kan myndighederne have behov for nogle supplerende oplysninger. Når disse oplysninger indsamles systematisk, benævnes de baggrundsdata. Oplysningerne edb-registreres ikke nødvendigvis, men må alligevel siges at udgøre et led i det samlede registersystem.
Hvis brugen af baggrundsoplysningerne i konkrete tilfælde er skønsmæssig og relativ sjælden, vil det ikke være nødvendigt systematisk at kontrollere og komplettere oplysningerne i tilknytning til indsamlingen, idet det kan tænkes at kræve mindre arbejde alene at rette manglerne op under sagsbehandlingen i de tilfælde, hvor oplysningerne har betydning for afgørelsen af et bestemt spørgsmål. Baggrundsdata anvendes normalt ikke uden for det forvaltningsområde, hvor de er indsamlede som direkte grundlag for administrative pro-

cesser, og er ikke umiddelbart velegnede til statistiske formål. Før disse oplysninger kan bruges i statistikken, må de i det mindste underkastes særlig grundig kontrol.
Som eksempler på sådanne baggrundsdata kan nævnes de specifikationer og oplysninger på selvangivelserne, som ikke dataregistreres hos ligningsmyndighederne, og en række af oplysninger på de tidligere vurderingsskemaer.

Notatdata

Disse oplysninger er af endnu løsere karakter end baggrundsdata. Det drejer sig om oplysninger, der ikke systematisk indsamles, men som fx er tilvejebragt under en sagsbehandling. Det kan være hensigtsmæssigt at registrere sådanne oplysninger, hvor de har eller kan tænkes at få fremtidig betydning i relation til den pågældende sagsbehandler enhed.
Oplysningerne kan være edb-registrerede og da under en systematisk ramme, men vil oftere have karakter af noteringer på kartotekskort eller lignende. Grænsen til egentlige sagsoplysninger er naturligvis ikke skarp.
Som eksempel kan nævnes en række notatfelter i bygnings- og boligregistret BBR, hvor den enkelte kommune helt på egen hånd bestemmer, hvad der skal indføres i registret og i hvilken form, det skal ske. Tilsvarende har arbejdsformidlingskontorerne registreret en række notatoplysninger i AF-match. Oplysningerne vedrører fx jobønsker og kortere uddannelser.
Oplysninger af denne karakter vil som regel kun være meningsfyldte for den pågældende sagsbehandlende enhed.

Statistikdata

De under 3.1.d. - 3.1.g. nævnte data vil som hovedregel være teknisk egnede og ofte også relevante til statistikformål, og det samme kan i sjældnere tilfælde gælde for baggrundsdata. Bortset fra de former for statistik, som knytter sig snævert til et givet administrationsområde, er det imidlertid ikke sikkert, at oplysningerne frembringes i de kombinationer, der er relevante for statistiske formål. Da der kan rettes op på dette ved samkøring, får de reelle muligheder for sådan samkøring generelt set stor betydning i relation til statistiske formål. Dette indebærer igen, at identifikationssystemerne, hvis effektivitet er afgørende for samkøringsmulighederne, spiller en særlig rolle for statistikproduktionen. Kombinationsmulighederne kan udstrækkes til at omfatte traditionelle statistikdata derved, at de generelle identifikationssystemer også anvendes i forbindelse med indsamling af oplysninger gennem postenquete eller interviews.
En særlig form for statistikdata er data, som er optaget i administrative registre, men udelukkende med statistiske formål. Denne form for dataindsamling er blevet kaldt integreret dataindsamling og den omtales nærmere i afsnit 4.3.
I den administrative proces kan de statistikdata, der indsamles på denne måde, sidestilles med baggrunds- og notatdata i den forstand, at deres kvalitet og fuldstændighed ikke har direkte og afgørende betydning for administrationen. Pålideligheden af denne type oplysninger kan derfor blive forringet.
Denne situation er ikke acceptabel for statistikproduktionen, idet de data, det drejer sig om, typisk vil være at betragte som meget centrale for statistikken,

- i modsat fald ville man ikke have ønsket dem indført i registrene. Problemet er imidlertid vanskeligt at tackle.

Som eksempel på et statistikdatum af den omtalte art skal nævnes arbejdsløshedsårsag, der angives ved afkrydsning på dagpengekortet. Koden skal vise om arbejdsløsheden fx skyldes hjemsendelse på grund af dårligt vejr, arbejdsfordeling o.l.

Mens kortets øvrige oplysninger skal anvendes ved den administrative behandling af udbetalingen af arbejdsløshedsdagpenge er årsagskoden udelukkende til statistisk brug. Det fremgår direkte af dagpengekortet, at årsagskoden er "til statistisk brug", og at den ikke er omfattet af den "tro og love erklæring", som den ledige skal underskrive vedrørende kortets øvrige oplysninger.

Resultatet har været, at besvarelsen af årsagskoden klart har været undladt i mange tilfælde. Der er således A-kasser, hvis medlemmer har en del udendørs vinterarbejde, og således ifølge sagens natur af og til må opleve arbejdsløshed på grund af dårligt vejr, men hvor denne årsagstype aldrig forekommer i statistikken. Efter arbejde i flere forskellige udvalg med repræsentanter for arbejdsmarkedsmyndigheder, A-kasser og arbejdsmarkedets parter er det opgivet at forbedre oplysningen om arbejdsløshedsårsag via dagpengekortet.

I stedet har udvalgsarbejdet resulteret i ønske om en interviewundersøgelse om bl.a. arbejdsløshedsårsager. Resultaterne skal efterfølgende samkøres med CRAM-statistikken, således at årsagsoplysningen hermed kan kombineres med den løbende arbejdsløshedsstatistik

Som et andet eksempel skal nævnes stillingskoden, som efter ønske fra Danmarks Statistik blev indføjet i CPR og i skatteregistrene. Stillingsbetegnelsen blev anført af borgerne på selvangivelsen, og de kommunale skattemyndigheder skulle sørge for, at stillingsændringer blev registreret.

Resultatet af denne indsats var ikke alt for opmuntrende. Man har derfor benyttet andre veje til at forbedre stillingsoplysningerne, idet man har suppleret dem med en række oplysninger fra andre kilder, som kan belyse arbejdsforholdene for større eller mindre grupper af de beskæftigede.

Trods de beskrevne vanskeligheder står det klart, at integreret dataindsamling på flere områder er en forudsætning for en udbygning af registerstatistikken. De nyeste erfaringer tyder da også på, at sådan indsamling kan gennemføres med et godt resultat, hvis man har mulighed for at ofre de nødvendige ressourcer på en hurtig og effektiv fejlkontrol og -opretning.

Således har det vist sig muligt i arbejdspladsprojektet at lukke nogle væsentlige huller i statistiksystemet gennem begrænsede justeringer af de administrative rutiner. De registreringer, arbejdspladsprojektet tilvejebringer, er dels systematiske oplysninger om, hvilke lokale arbejdssteder der eksisterer, dels oplysning om, hvem der er beskæftiget på disse arbejdssteder. Oplysningerne er afgørende for mulighederne for at fremstille geografisk beskæftigelsesstatistik, pendlingsstatistik og opgørelser om de lokale erhvervsenheders forhold.

Oplysningerne om de eksisterende arbejdssteder indsamles i det store og hele gennem traditionelle metoder, nemlig ved henvendelse til større private firmaer på basis af erhvervsregistret, til kommuner og amtskommuner samt til statslige myndigheder.

For de private arbejdsgiveres vedkommende indhentes nøglen mellem arbejdstager og arbejdssted gennem den årlige oplysningsseddel til skattevæsenet, som i den anledning er udvidet med et felt for arbejdssted; det er naturligvis kun arbejdsgivere med flere arbejdssteder, der skal udfylde dette felt. De foreliggende oplysninger viser, at også mange af de større private arbejdsgivere allerede har disse oplysninger inde i deres lønsystemer under en eller anden form. Danmarks Statistik udnytter i det omfang, der er mulighed for det, registreringerne i den form, hvori de foreligger hos arbejdsgiveren, idet det må antages, at datakvaliteten på denne måde bliver den bedste mulige. Det betyder, at der efter indsamlingen må udføres et harmoniseringsarbejde som led i statistikproduktionen.
Danmarks Statistik overtager selv kontrollen med oplysningerne umiddelbart efter, at de er indsamlet og registreret. Det er en følge af, at skattemyndighederne ikke hidtil har haft brug for arbejdsstedskoden i det administrative arbejde.
For den offentlige sektors vedkommende findes oplysningen i stort omfang allerede registreret i de statslige og kommunale edb-lønanvisningssystemer. I projektet benytter man derfor oplysninger, som hentes direkte fra disse systemer, som dog forinden må kompletteres og systematiseres.

Samkøring af registre og OPUS begreber

Claus Ib Olsen

Samkøringer

Synopsis:
1. Indledning om samkøring
2. Om Begreber i OPUS
3. Om Registre
4. Horisontal integration
5. Vertikal integration
6. Syntese: Samkøring mhp generering af ny enheder

Indledning

Om dette indlæg

Dette indlæg handler om samkøring af registre, formuleret efter centrale begreber i OPUS-håndbogen. Der beskrives centrale problemstillinger i forbindelse med brug af personregistre.

Jeg skal takke Lisbeth Laursen for, at jeg har fået lov til flittigt at bruge manus til brochuren om 'Introduktion til Sygehusbenyttelsesstatistik' i forbindelse med nærværende indlæg.

Samkøring

Ved en 'samkøring' af registre forstår vi: den proces eller metode, hvorved informationer fra forskellige statistikregistre om en bestemt statistisk **enhed** bindes sammen, eller hvorved informationer om forskellige **enheder**, der indgår i bestemte relationer bindes sammen

Formålet med denne sammenbinding af informationer, er at bringe os ny viden om de **enheder** (genstande) vi er interesserede i (eller rettere sagt : afdække og synliggøre denne viden; for informationen eksisterer allerede i det enkelte register, vi har bare ikke bragt informationen frem i lyset endnu. Ved samkøringen sker der en bearbejdelse af den eksisterende information, hvorved værdien af informationen forøges).

For en ordens skyld skal anføres, at en samkøring kun er mulig, hvis den relevante statistiske **enhed** i de kilderegistre, der skal integreres, har fælles identifikation/nøgle (fx person-nr).

Det siger sig selv, at denne metode til dataindsamling har overordentlig stor praktisk betydning, idet vi genudnytter information, der allerede er indsamlet, til statistikbrug. Herved vil omkostninger til dataindsamling, registrering og fejlsøgning være betydelig mindre end ved traditionel dataindsamling, respondentbyrden falder stort set væk, det færdige statistikprodukt vil komme langt hurtigere på gaden, ligesom vi i visse tilfælde har mulighed for, at skaffe statistik om forhold, vi normalt ville anse for praktisk og økonomisk

uoverkommelige. En af ulemperne er, at vi må skal holde os til de data der nu een gang er tilgængelig i det administrative register; selvom vi har mulighed for - via forhandlinger og i begrænset omfang - at få indført registrering af særlige statistikrelevante oplysninger i administrative registre.

Samkøring af aggregerede data

Man kan tænke sig, at det til visse formål vil være tilstrækkeligt, at der kan ske integration på et ikke-individniveau (aggregeret niveau), fx på kommune niveau, men derved kan egenskaber for den enkelte individ og ændringer i disse ikke registreres, men kun de gruppevis målte nettoændringer. Herved mistes den fulde styrke af samkøring, og megen information vil gå tabt. Denne integrationsmåde vil ikke blive nærmere berørt.

Typer af samkøring

Vi kan nævne forskellige typer af samkøring efter deres art og formål:

Samkøringsart:

Ved **horisontal integration** forstår vi den proces, hvorved to forskellige registre med informationer(data) om samme **enhed,** integreres med henblik på, at danne en foreningsmængde af **egenskaber** fra de to registre om de pågældende enheder, eller integration af to registre mhp at danne en foreningsmængde af **enheder** for de to registre (ex: samkøres lønregistre fra den private sektor og den offentlige sektor, for at danne et samlet lønregister). Typiske formål er 1) supplering af egenskaber (fx ønsker vi at finde adresser fra befolkningsstatistik-registeret til brug for uddannelsesklassifikationsmodulet), 2) kontrolformål eller fejlsøgning (fx tjek af valide personnumre) og 3) afgrænsning af populationer (dvs: ved samkørsel ønsker vi at lokalisere de **enheder** (personer) i det ene register, der har bestemte **egenskaber** i det andet register).

Ved **vertikal integration** forstår vi den proces, hvorved et register i to eller flere tidsmæssigt forskudte versioner sammenlignes. Typiske formål er: konstatering af tilgang og afgang af **enheder** (fx hvilke personer går ud og ind af uddannelsessystemet) og ændring i enhedens **egenskaber** (fx ændring af en persons ægteskabelige status over tid)

Ved **Enhedssyntese** forstås den proces hvorved registre bearbejdes mhp dannelse af ny **enheder.** Typiske formål er fx dannelse af familier og husstande udfra henvisninger i befolkningsstatistikregisteret.

I nærværende indlæg beskrives de forskellige integrationsformer hver for sig; i praksis vil de forskellige typer og formål af integrationer imidlertid ofte optræde sammen. Fx vil afgrænsning af en population og kontrol af identifikationer skulle foretages i mere eller mindre samme procesforløb.

Fejl i identifikationer

Samkøring af statistikregistre har en overordentlig høj nytteværdi som alternativ til traditionel dataindsamling. For at samkøring skal være vellykket kræver det imidlertid, at identifikation af personer og andre enheder i de pågældende kilderegistres udtræk er entydige og veldefinerede. Dette er desværre ikke altid tilfældet og der kan ligge en meget stor arbejdsindsats i, at få klarlagt og oprettet fejlagtige identifikationer og mangler i identifikationerne.

Om OPUS-begreber

Lad os kort gøre et OPUS-ologisk sidespring og ridse et par begreber op:

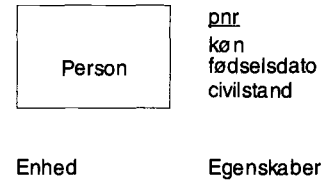
OPUS

OPUS står for: 'Om Planlægning og Udvikling af Statistiksystemer' og er de principper, man i Danmarks Statistik har vedtaget at følge i forbindelse med planlægning af statistikprocessen. Der er udgivet en håndbog : OPUS-håndbogen, der vedligeholdes af planlægningsstøtten i brugerservice-kontoret.

Nedenfor gennemgås nogle vigtige begreber fra Indholdsanalysen i OPUS, som er særdeles nyttige i forbindelse med beskrivelser af registre og samkøring af disse. Det, vi i OPUS-sammenhæng kalder Indholdsanalyse, betegnes også som: informationsanalyse, entitets-relationsmodellen eller Chen-notation.

Enhed

Enheder er de personer eller genstande, som man vil samle statistik om, dvs tællingsenhederne. I OPUS anføres **enheder** grafisk ved et rektangel :



Egenskaber

Egenskaber (eller **variable**) er det, man beskriver **enheden** med. Fx. alder eller størrelse. Vi skelner mellem de(n) egenskab(er), der identificerer den enkelte enhed: **nøglen** eller identifikationen - fx. personnummer eller BBR-nummer - og de egenskaber der i øvrigt beskriver enheden: de beskrivende egenskaber. Egenskaber kan være kvalitative, som fx køn (mand, kvinde) eller kvantitative, som fx indkomst(200.234 kr.) og alder (29 år).

Nøglen anføres på grafen med understreget tekst (fx pnr) og de beskrivende egenskaber med almindelig tekst.

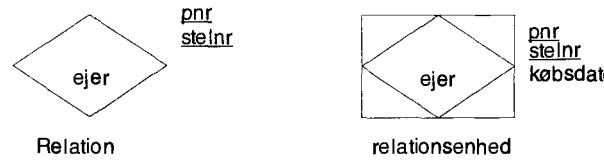
Hvis man ønsker at sammenholde oplysninger om en enhed fra flere forskellige kilder er det **nøglen** man bruger til sammenkoblingen.

En samling af enheder med bestemte egenskaber betegnes **enhedsgruppe** eller **populationen**.

Hvis man ønsker at undersøge forbindelserne mellem forskellige enheder er det nøglerne fra de indgående enheder man bruger, og herved er vi ovre i relationer.

Relationer

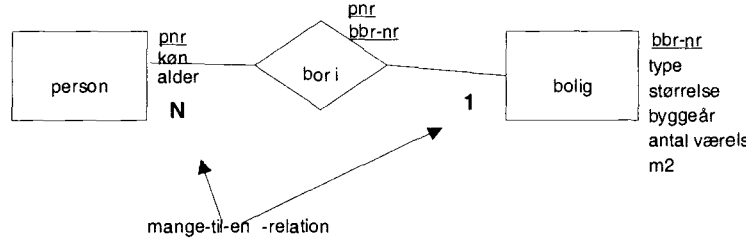
En **relation** karakteriserer et forhold mellem 2 eller flere **enheder**, fx. 'en person bor i en bolig' eller 'en person ejer en cykel'. En relation identificeres ved kombinationen af de indgående enheders nøgler, Fx: 'person: PNR ejer cykel: STELNR'. Grafisk anføres relation som en diamant.



En relation kan - udover nøglerne - også indeholde specifik information om relationens karakter. I eksemplet ovenfor til højre, indeholder relationen oplysning om købsdato for cyklen. Hvis en relation indeholder oplysninger, ud over de indgående nøgler, betegnes den som **relationsenhed** og anføres grafisk, som et rektangel med en diamant inden i.

Enhedsgrafen

Den logiske sammenhæng mellem **enheder**, **egenskaber** og **relationer** kan afbildes grafisk :



Eksempel på enhedsgraf

Relationerne mellem enhederne opdeles ofte i typerne 'en til en', 'en til mange' og 'mange til mange', hvilket angiver hvordan enhederne 'står i forhold til hinanden': **en** person kan højst have **en** ægtefælle; medens **en** person kan have flere elskerinder eller eje **flere** cykler; **mange** gymnasieskoler udbyder **mange** gymnasiale retninger.

På grafen angiver ved betegnelserne **1** og **N**, hhv en og mange forholdene.

Grafen, der viser sammenhængen mellem personer og boliger, læses således: Personer bor i boliger; der kan bo **mange** personer i **een** bolig og **een** person kan kun bo i **een** bolig; de konkrete personer, der bor i konkrete boliger bestemmes af pnr og bbr-nr.

Om Registre

Næsten alle personrelaterede **statistikregistre** i Danmarks Statistik er produceret på baggrund af administrative **kilderegistre** i statslig eller kommunal regi (eks: Indenrigsministeriets centrale personregister CPR), evt suppleret

med manuelt indsamlede oplysninger fra andre kilder (eks: Statistik over levendefødte: CPR og fødselsanmeldelser).

Kilderegister

Det administrative kilderegistre kan være egentlige dagligdags brugerregistre, som er dynamiske registre i den forstand, at de løbende (dagligt/ugentligt/månedligt) modtager **transaktioner** (=opdateringer) af informationer på baggrund af **hændelser** vedr. de **enheder** der findes i registeret, således at registeret til en hver tid afspejler den administrative virkelighed. Kilderegisteret kan dog også være centrale integrerede registre, der periodisk modtager oplysninger fra de egentlige brugsregistre (eks: Landspatientregisteret er et register i Sundhedsstyrelsen, der een gang årligt modtager udtræk fra de enkelte sygehuses patientregistre).

Statistikregister

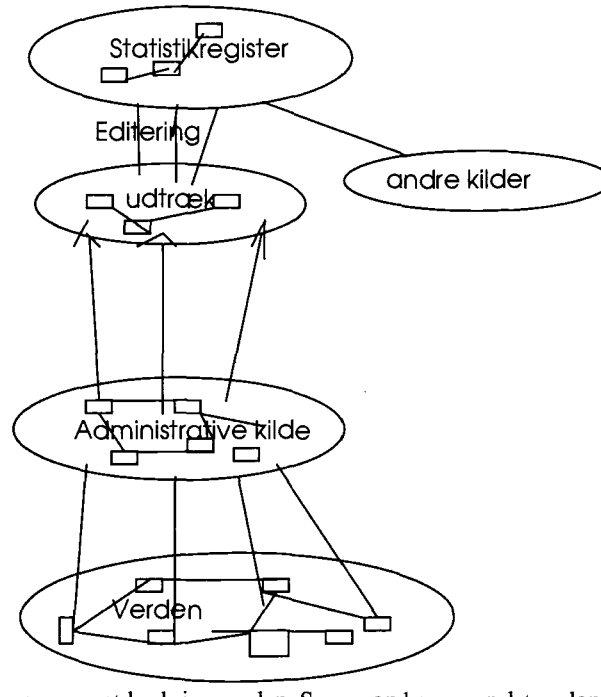
Fra kilderegisteret skal Danmarks Statistik modtage udtræk til statistisk bearbejdelse og dette udtræk vil ofte (bl.a. af praktiske grunde) kun afspejle kilderegisterets tilstand på et nærmere defineret tidsafsnit. Til statistisk brug vil Danmarks Statistik typisk modtage et udtræk een eller få gange om året fra det administrative kilderegister, der indeholder de oplysninger, som er relevante og udtrækket vil altså kun kunne afbilde kilderegisteret på det pågældende tidspunkt eller for den pågældende periode. Der kan opridses tre hovedtyper af udtræk

- Udtræk **vedr. status for populationen pr. en given dato** (eks: personer med social pension og pensionens størrelse pr. januar måned)

- Udtræk vedr. status for population pr. en given dato, suppleret med oplysninger **om udviklingen i centrale egenskaber i en periode** (eks: Bistandslovsstatistik året 1992)

- Udtræk med **transaktioner for alle de relevante enheder** (eks sygehusbenyttelse, cpr?)

Udtrækket vil nu blive editeret og formodentlig sammenholdt med andre oplysninger og registre, for tilsidst at blive til det færdige **statistikregister.** Denne fase af processen er omfattende



Det er opgaven er at beskrive verden. Som man kan se er det en lang snor fra verden til statistikregisteret og problemerne mange.

Kilde vs statistik

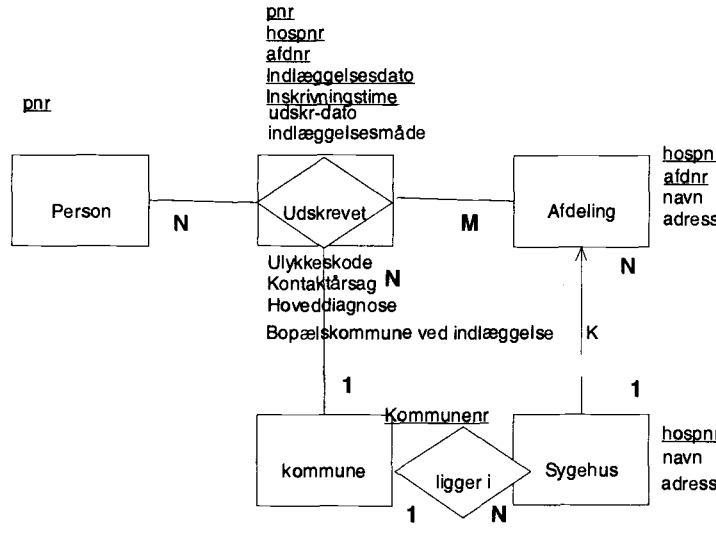
Nogle af de registerstatistiske problemstillinger er opstået fordi vi 'kun' har oplysninger pr givne perioder:

- at vi måske skal genere forløb udfra status udtræk. (ex: tilgang og afgang fra to udtræk).
- at vi skal sammenligne status data fra et tidspunkt med forløbsdata over en periode

Eksempel SBR

Sygehusbenyttelsesregisteret indeholder oplysninger om de personer, der i et kalenderår udskrives fra somatiske heldøgnsafdelinger. Kilderegisteret er Sundhedsstyrelsens landspatientregister, der een gang om året modtager oplysninger fra de enkelte sygehuse. Registeret indeholder ikke oplysninger fra psykiatriske afdelinger. Oplysninger fra skadestuer og ambulatorier er kun delvist indberette på nuværende tidspunkt.

Fra landspatientregisteret foretages et årligt udtræk. Oplysningerne i vedrører de personer, der i årets løb er udskrevet fra en somatisk døgn-afdeling, medens udskrivninger fra skadestuer og ambulatorier endnu ikke tages med, da oplysningerne først fra 1995 vil være fuldt indberettet. Enhederne er: **person** og **hospitalsafdeling;** mellem person og afdeling eksisterer relationen: **udskrivning** (og naturligvis indlæggelse). Se flg. graf eksempel 3.2.1.



Eks. 3.2.1: Enhedsgraf for udtræk fra Landspatientsregisteret

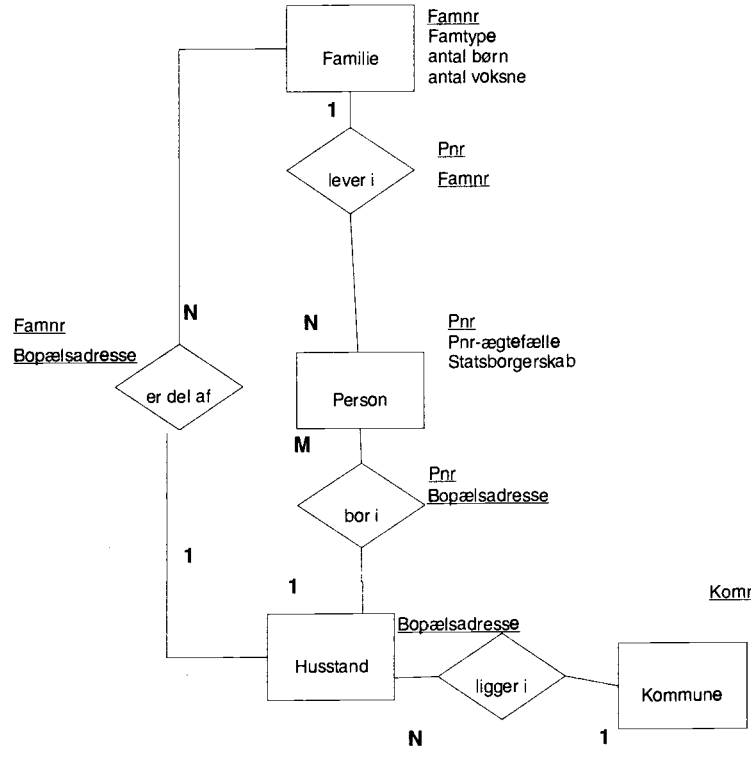
Aflæsning af enhedsgrafen:
Enheder:
Der findes personer (ca. 700.000)
Der findes sygehuse (ca. 90)
Der findes sygehusafdelinger (500)

Relationer:
Et sygehus kan have mange afdelinger
Personer udskrives fra en hospitalsafdeling (1.048.012 udskrivninger)

- en person kan være udskrevet/indlagt flere gange og på forskellige sygehusafdelinger
- en sygehusafdeling kan have mange personer udskrevet/indlagt

Eksempel befolkningsstatistik statusregister

Befolkningsstatistikregisteret består af et statusregister over befolkningen samt et bevægelsesregister. Oplysningerne i statusregisteret baseres på udtræk fra CPR modtaget pr. 1. januar. Udtrækket indeholder de faktisk tilstedeværende personer og inaktive personer. Udtrækket fejlsøges og korrigeres med oplysninger fra bevægelsesregisteret, der indeholder oplysninger om fødsler, dødsfald mv. Fra udtrækket kan der via egenskaber for personen dannes familier og husstande. Nedenstående graf, eksempel 3.2.2, viser de væsentlige enheder og relationer for befolkningsstatus:



Eksempel 3.2.2: Befolkningen d. 1. januar

Aflæsning af enhedsgrafen med status januar 1992:
Enheder:
Der findes personer (ca. 5.150.000)
Der findes familier (ca. 2.800.000)
Der findes husstande (2.300.000)

Relationer:
En famille kan bestå af flere personer
En husstand kan bestå af flere personer
En husstand kan bestå af flere familier
En husstand har en adresse i en kommune

Horisontal integration

Som allerede omtalt, forstår vi ved horisontal integration en sammenkøring af to forskellige registre, der har til formål :

1. at supplere egenskaber fra eet register med nye egenskaber fra et andet:

2. at kontrollere validiteten af oplysninger i et register mod oplysninger i et andet, eller

3. at afgrænse en population.

Supplering

Hvis enhederne er entydigt definerede og identifikationen af disse (nøglerne) er korrekte, volder samkøring af to registre ingen problemer. Der, hvor problemerne opstår er: hvor enhederne ikke er entydigt definerede og/eller nøglerne ikke er korrekte. Eksempler på fejl er tilfælde, hvor det administrative kilderegister misbruges til registrering af fremmede enheder (ex: maskiner får personnumre i et kommunalt lønregister) eller 'administrativt dannede enheder', eller registre hvor primæridentifikation af den administrative transaktion måske ikke har så stor betydning for systemet.

En anden væsentlig faktor, der kan vanskeliggøre integration, er forskelle i tidskomponenten i de to kilderegistre. Eksempelvis når to registre har væsentlig forskellige opgørelses-tidspunkter, og/eller hvor det ene register er et status-registre, med opgørelse pr. en given dato og det andet et register, der indeholder forløbsdata for de samme enheder over en periode (en meget almindelig situation).

Et eksempel på samkøring er dannelse af Sygehusbenyttelsesregisteret, hvor et udtræk fra landspatientregisteret om personers udskrivninger fra somatiske sengeafsnit (se eksempel 3.2.1 ovenfor) samkøres med en række baggrundsregistre i Danmarks Statistik bl.a. Befolkningsstatistikregisteret (statusregisteret se eksempel 3.2.2), det boligstatistiske register , uddannelsesklassifikationsmodulet, mm. Ved samkøringen får vi integreret oplysninger om personers sygehusbenyttelse med deres familiemæssige, boligmæssige, uddannelsesmæssige forhold osv, og kan på denne måde belyse en række levevilkårsbetingede sygelighedsforskelle for den danske befolkning i den pågældende periode. Samtidig får vi illustreret nogle af problemerne i samkøring.

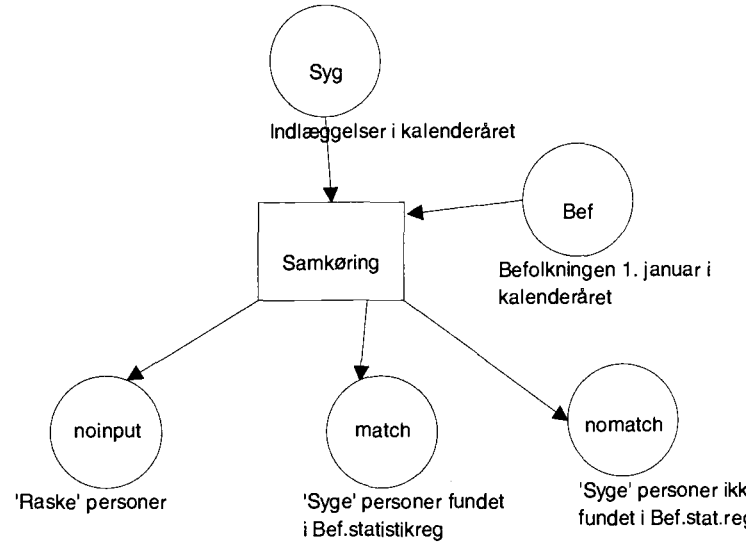
Figur 3.2.3 viser et overordnet systemdiagram for samkøring mellem udtræk fra landspatientregisteret og befolkningsstatistikregisteret.

Cirklerne i systemdiagrammet repræsenterer datasæt. Ved sammenkørslen benævnes sygehusudtrækket som **input** og befolkning som **match.** Der dannes tre datasæt:

- Noinput: 'Raske'. Personer i Befolkningsstatistikregisteret pr 1. januar, der ikke findes i udtrækket fra Landspatientregisteret, dvs personer pr. 1. januar, der ikke har været udskrevet fra et sygehus i årets forløb.

- Match: 'Syge' personer, der findes i begge registre

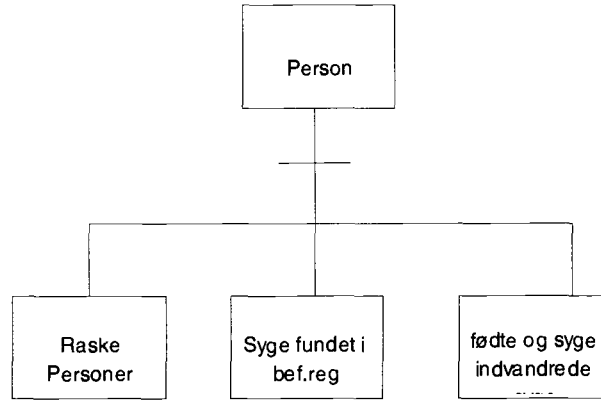
- Nomatch: 'Syge' personer i sygehusudtrækket, der ikke findes i Befolkningen pr. januar. Denne gruppe består af personer, der er indvandret eller født siden 1. januar eller fejl. (Det skal oplyses at i årsudgaven 1990 var der ca. 64.400 'sygehuspersoner' der ikke havde bopæl i Danmark. Langt hovedparten, nemlig 61.700 var født i årets løb.)



Eksempel 3.2.3: Systemdiagram for samkøring mellem Sygehusindlæggelser i kalenderåret og Befolkning 1. januar.

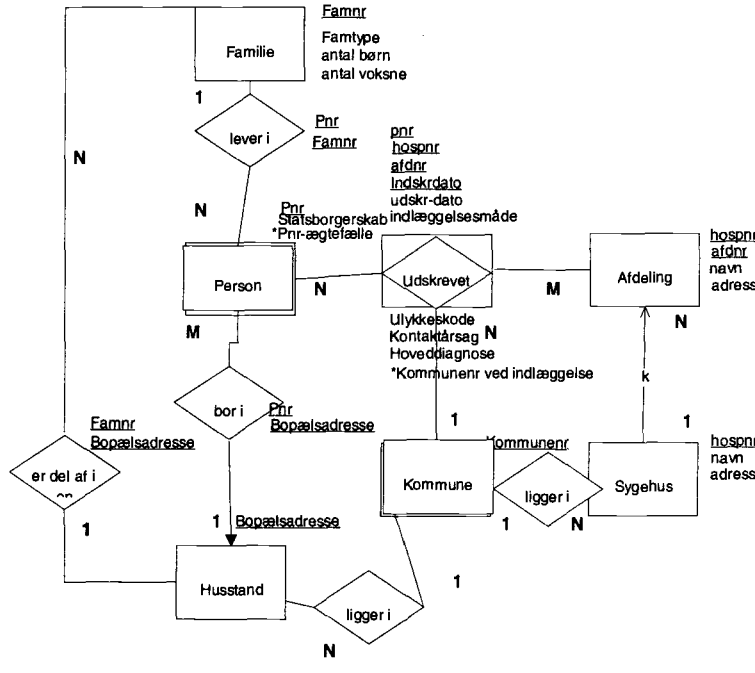
I match-gruppen, findes også personer pr. 1. januar, der dør eller forlader landet i årets løb. For at finde disse må man foretage integration med næste års statusregister, eller supplere med oplysninger fra den løbende befolkningsstatistik samme år.

Resultatet af samkørslen kan beskrives indholdsanalytisk på flere måder. Man kan fx vælge at formulere resultatet som subtyper af persongrupper. Tilsammen udgør de tre grupper hele befolkningen . For hver subtype kan der dannes enhedsgraf ved at kombinere enhedsgraferne eksempel 1 og 2



Eksempel 3.2.4: Tre subtyper af personer i samkørslen mellem sygehusbenyttere og befolkningen

Den mest fuldstændige findes for subtypen 'syge' fundet i befolkningsregisterert som er følgende:



Eksempel 3.2.5 Enhedsgraf, der viser integration mellem sygehusbenyttere og befolkningsstatusregisteret.

Som det fremgår af enhedsgrafen 3.2.5, får vi ved integrationen dannet forbindelseslinier mellem de hidtil adskilte enheder via de fælles nøgler. Herved bliver de to registres informationer fælles gods og helt nye og spændende sammenhænge kan afdækkes. (Fx kan vi undersøge sammenhængen mellem hospitalsindlæggelser og familie/husstandsmønstre, via relationen fra personenhed til hustandsenhed og familieenhed). I sygehusbenyttelsesregisteret er udtrækket fra landspatientregisteret - udover befolkningstatusregisteret - integreret med arbejdsklassifikationsmodulet, det boligstatistiske register, det medicinske fødsels- og dødsfaldsregister, uddannelsesklassifikationsmodulet, sygesikringsstatistikregisteret og registeret over indkomsterstattende ydelser. Det siger sig selv, at denne omfattende integration af oplysninger giver mulighed for at belyse et ufattelig stort antal problemstillinger.

Kontrol af validitet

Ved at samkøre to registre, hvor vi a priori ved at visse oplysninger i det ene (input-registeret) skal genfindes eller kunne beregnes i det andet (match-registeret), giver os mulighed for at checke validiteten af de pågældende oplysninger. I ovennævnte eksempel vedr. dannelse af sygehusbenyttelsesregisteret, ved vi, at sygehusbenytterne skal findes i Befolkningsstatistikregisteret - dog ekskl. personer, der er indrejst eller født siden 1. januar. Ved sammenkøringen får vi altså tjekket personidentifikation, for de fødte ved vi jo de er født, så tilbage er der kun de indrejste som kan kontrolleres af befolkningsstatistikkens bevægelsesregister.

Afgrænsning af population

Ved samkøring af to registre vil der kunne ske en afgrænsning af population, dvs. ved at kombinere egenskaber i de to registre vil nye populationstyper kunne bestemmes , jvf ovenstående eksempel, hvor vi får inddelt befolkningen i 2 grupper: hospitalsbrugere og andre; herved kan forskelle i gruppen belyses.

Vertikal integration

Ved vertikal integration forstås samkøring af 'samme' register på forskellige tidspunkter, hvorved ændringer i værdien af egenskaber for de pågældende enheder belyses. Af typiske formål kan nævnes:

1. Registrering af tilgang og afgang af enheder over tid

2. Registrering af ændringer i enhedernes egenskaber over tid

Tilgang og afgang

Ved at sammenligne et register i to tidsmæssigt forskudte versioner, kan man få opgjort netto tilgange og afgange af personer for den pågældende periode. Når det er afgørende at få et billede af bevægelserne på de enkelte personer og man ikke har mulighed for at få adgang til de egentlige transaktioner, er vertikal integration et godt værktøj til måling af ændringer over tid. Her sammenligner man tidsmæssigt forskudte versioner af det samme register. På denne måde får vi, med en vis præcision belyst tilgang og afgang som kan fordeles efter de relevante baggrundsklassificeringer. Begrænsningen ligger naturligvis i det forhold at bruttobevægelserne for den enkelte enhed (person) i den mellemliggende periode ikke belyses.

Ændringer i egenskaber

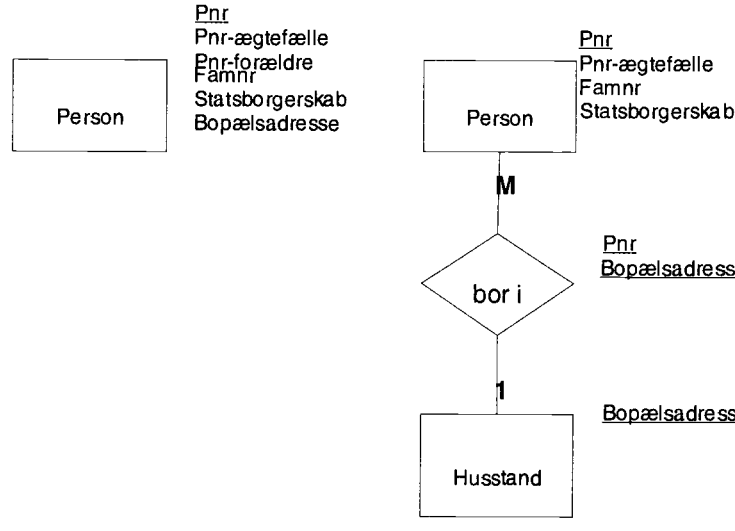
Ved at sammenligne to eller flere tidsmæssigt adskilte versioner af samme register kan man få opgjort (netto)-ændringerne i enhedernes beskrivende egenskaber. Formålet hermed er, at følge udviklingen i en eller flere egenskaber for bestemte enhedsgrupper og dermed konstatere et forløb for den enkelte enheder. Dette har både interesse i forbindelse med måling af korttidsvirkninger og i forbindelse med longitudinale studier, som fx 'Dødelighed og erhverv'.

Enhedssyntese

Et tredie hovedtype ad 'samkøring' er, at generere en ny **enhed**sgruppe, ved hjælp af en syntetisering af eksisterende oplysninger. (Eksempel: dannelse af **husstand** i befolkningsstatistikregisteret udfra **person**oplysninger). For at kunne danne en enhedssyntese kræver det at der i datamaterialet findes de nødvendige egenskaber, der angiver relationen mellem basisenheden og den syntetiske enhed.

Husstand og familie

Dannelse af husstand og familie i Befolkningsstatistikregisteret:



Før syntese

Efter syntese: dannelse af husstand

Eksempel 3.2.6 Syntese af relation til enhed: Dannelse af husstande.

I datamaterialet fra det Centrale Personstatistikregister findes der for hver person**enhed** oplysning om adresse og oplysning om forældres, børns og ægtefælle/registreret partnerskabs personnr. Ud fra disse oplysninger er muligt ved syntese at generere husstande og familier. En husstand defineres som: de personer, der på et givet tidspunkt, findes på samme adresse, medens definition af en familie følger mere komplekse regler, der er redegjort for andet steds.

Ovenstående enhedsgraf beskriver syntesen af husstande. (Det skal dog bemærkes, at der knytter sig specielle problemer til definition af fælleshusholdninger, så som plejehjem, døgninstitutioner, skolehjem, kostskoler o. lign.). Før syntesen har vi for hver person på et givet tidspunkt oplysning om adresse. Ved samkøringen tildeles personer med samme adresse, samme entydige identifikation: Bopælsadresse.

Der kan nu dannes en ny enhed: Husstand og grafen viser at der - på et givet tidspunkt - findes husstande og at en husstand kan have mange personer, medens en person ikke kan bo i flere husstande.

På tilsvarende vis - dog efter noget mere komplicerede principper - kan der dannes familier, som igen kan bo i husstande, jvf tidligere eksempel 3. 2 .2

Statistiksystemet

Finn Spieker

1. Formålet med anvendelse af registerdata

Genanvendelsen af data, som er indsamlet til administrative formål, har været et grundlæggende princip i den moderne danske statistikproduktion. Lov om Danmarks Statistik indeholder en række bestemmelser, om adgangen til at anvende administrative data til statistik og udgør derfor de nødvendige forudsætninger for en strategi, som indebærer anvendelsen af foreliggende registeroplysning i størst muligt omfang.

Etableringen af de administrative registersystemer startede i 1968 med oprettelsen af CPR, hvor personnummeret blev introduceret. Siden er systemer udviklet på en lang række områder med personnummeret som generel identifikation. Data fra disse forskellige områder kan derfor kædes sammen på individniveau, hvorved oplysninger, der ellers skulle indsamles ved direkte henvendelse til borgerne, helt eller delvist kan udledes fra registrene. Disse muligheder for gennem samkøring af registre at samle oplysninger om enkeltpersoner kan virke forskrækkende og dermed medføre en vis modvilje i befolkningen, så det er af afgørende betydning, at behandlingen af registeroplysninger sker under iagttagelse af strenge sikkerhedsbestemmelser, så der skabes sikkerhed for og tro på, at disse statistikdata ikke anvendes til andre formål end statistik.

Oplysningernes egnethed som statistikgrundlag har været meget diskuteret, men der er efterhånden en voksende erkendelse af de fordele, som er forbundet med denne fremgangsmåde. Dette gælder ikke mindst i international sammenhæng. Gennem anvendelse af registerdata som grundlag for statistiske opgørelser opnås en række fordele i forhold til at basere statistikken på data indsamlet ved direkte henvendelse til hele befolkningen eller dele af denne. Man belaster ikke borgerne med afgivelse af oplysninger til en offentlig myndighed, som man allerede har afgivet til en anden offentlig myndighed. Noget sådant kan blive oplevet som en unødig og irriterende belastning og dermed påvirke kvaliteten af de afgivne oplysninger. Omkostningerne ved indsamling og behandling af data vil blive minimeret, når indsamlingen kun finder sted én gang, og man vil ikke i samme grad være henvist at benytte udsnit af omkostningsbestemt størrelse, når foreliggende oplysninger kan anvendes. Oplysningernes kvalitet vil ofte være høj, idet afgivelsen af oplysningerne oftest er lovbunden, og den efterfølgende administrative anvendelse sikrer i hvert fald til en vis grad, at oplysningerne er korrekte i forhold til den administrative anvendelse. Endelig undgår man den bortfaldsproblematik, som ofte er forbundet med de traditionelle indsamlingsmetoder.

Oplysningerne i de administrative registre afspejler de regler, som skal administreres. Det betyder, at der kan forekomme konkrete statistikopgaver, hvor man må konstatere, at datamulighederne umiddelbart er utilstrækkelige. En supplerende dataindsamling vil da være en mulighed med henblik på at supplere de foreliggende registeroplysninger. På denne måde kan registergrundlaget udbygges på områder, hvor den administrative dækning er mangelfuld, og registrene kan i den forbindelse være grundlag for at sikre en hensigtsmæssig stratifikation ved udtræk af stikprøver.

Danmarks Statistik har siden starten af CPR fulgt den strategi, som loven lægger op til. Det har betydet en gennem årene stærkt stigende anvendelse af registeroplysninger. Den tunge og omkostningskrævende indsamlingsmetode med anvendelse af interviews eller postspørgeskemaer benyttes kun i de tilfælde, hvor udarbejdelsen af en meget højt prioriteret statistik ikke er mulig på registergrundlag.

2. Det administrative datagrundlag

Grundlaget for statistiksystemet er en lang række statslige og kommunale administrative registre og andet administrativt materiale. Et særligt led i anvendelsen af de offentlige registre er den integrerede dataindsamling, hvor man indhenter statistikdata uden betydning for den administrative proces samtidig med andre oplysninger til det administrative register. Disse statistikdata behandles ikke af den administrative myndighed, men videresendes sammen med de andre oplysninger til Danmarks Statistik. I visse tilfælde suppleres registerbaserede oplysninger med indberetninger fra private virksomheder og organisationer eller direkte fra den enkelte person.

Kernen i datagrundlaget er Det Centrale Personregister (CPR), som sætter de ydre rammer for populationsafgrænsningen i personstatistikken, og registrets personnumre er den generelt anvendte identifikation i de øvrige administrative registre, der indeholder personoplysninger. Dette personnummer er således nøglen til samkøring af oplysninger fra forskellige kilder på personniveau. Fra CPR indhentes basale oplysninger om familieforhold, bopæl, statsborgerforhold mv. samt de hændelser, som fører til ændringer i disse. De øvrige administrative kilder omfatter oplysninger om beskæftigelse, ledighed, uddannelse, indkomst, sociale forhold, sundhed og kriminalitet. Foruden CPR er de vigtigste administrative kilder registre i Told- og Skattestyrelsen, Arbejdsmarkedsstyrelsen, det fælleskommunale system og de offentlige lønsystemer samt indberetninger fra uddannelsesinstitutionerne og private lønsystemer.

Foruden de egentlige personoplysninger indgår der oplysninger om ejendomme og boliger. Specielt for sidstnævnte gælder det, at Bygnings- og Boligregistret (BBR) under Kort- og Matrikelstyrelsen indeholder en præcis adressekode til identifikation af den enkelte bolig-/erhvervsenhed. Den samme adressekode er i CPR knyttet til den enkelte person, således at den fungerer som nøgle til at knytte boligoplysninger til personstatistikken. En

tilsvarende korrespondance tilstræbes i forhold til Det Centrale Erhvervsregister for så vidt angår erhvervsenheder.

I systemet indgår også en nøgle til at forbinde den enkelte person med eventuelle ansættelsessteder. I Det Centrale Erhvervsregister vedligeholdes en kode til identifikation af et enkelte arbejdssted. Tilsvarende koder indhentes fra arbejdsgivere i forbindelse med indberetning af løn mv. for de enkelte ansatte til Told- og Skattestyrelsen. Sammenhængen mellem personnummer og arbejdsstedsidentifikation vil således være registreret i Det Centrale Oplysningsseddelregister (COR).

3. Systemets opbygning

Grundoplysningerne i form af uddrag fra de administrative kilderegistre eller andet indsamlet materiale bearbejdes og organiseres i en række statistikregistre samt i enkelte klassifikationsmoduler. Det sker efter en emne-/formålsbestemt strategi. Denne strategi er bestemt af, hvem der er ansvarlig for og har indsigt i det enkelte statistikområde, hvem der har kontakt med dataleverandørerne, og adgangsbestemmelserne som fastlægger, hvem der kan få adgang til hvad.

Det enkelte statistikregister har traditionelt været emneafgrænset i den forstand, at det skal være grundlag for statistikken inden for et bestemt område. Det betyder, at der opereres med indholdsmæssigt klart afgrænsede registre, hvor der er tydelig sammenhæng mellem kilden med de for det pågældende emne centrale oplysninger og det enkelte statistikregister. Modellen er som hovedregel den, at man indhenter de emnespecifikke statistikdata fra en enkelt kilde og til disse føjer en række baggrundsoplysninger, ofte fra andre statistikregistre, og disse sammenstillede oplysninger udgør herefter registret for det pågældende statistikområde.

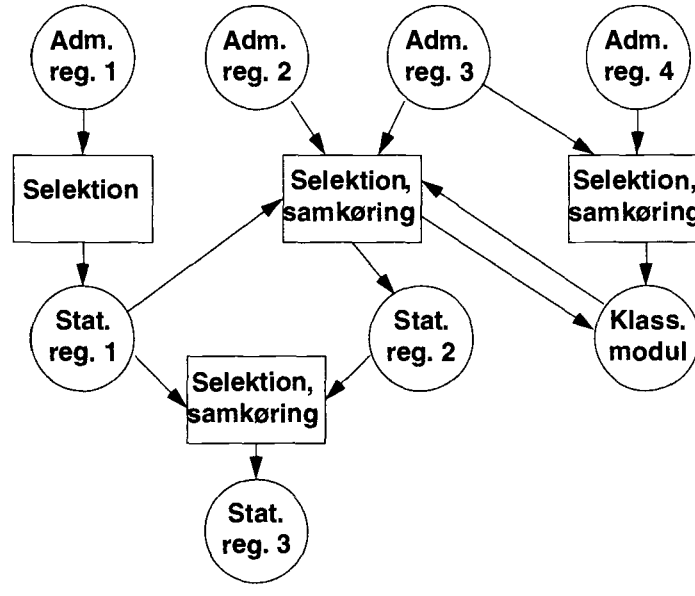
Denne klare linje mellem grundregister og statistikregister kan fortsat ses på en række områder, men behovet for at anskue tingene i en bredere sammenhæng (horisontal integration) har gjort, at nyere statistikregistre i stigende grad er et produkt af samkørte statistikregistre. Man kan fortsat tale om emneafgrænsede statistikregistre, men emnerne har fået en lidt anden karakter. På tilsvarende måde har interessen for forløbsstatistik bevirket, at forskellige års-versioner af samme register er grundlag for dannelsen af et nyt register (vertikal integration).

I visse tilfælde bearbejdes grundoplysningerne centralt og opbevares i et klassifikationsmodul. Disse moduler er til rådighed i forbindelse med dannelse af statistikregistre, og de kan indgå i samkøring med statistikregistre med henblik på dannelse af anonyme datasæt til enkeltstående specialopgaver.

Statistikregistrene og modulerne udgør samlet Danmarks Statistiks personstatistiske registersystem. Systemet er meget skematisk illustreret i figur 1, hvor statistikregister 1 er den enkle og meget direkte form for dannelse af et statistikregister, og statistikregister 2 fremtræder med et mere komplekst

sammensat kildegrundlag i et delvist integreret forløb med dannelsen af et klassifikationsmodul, eksempelvis indkomstregistret og AKM, medens statistikregister 3 er et integrationsregister baseret på andre statistikregistre. De enkelte elementer kan anvendes isoleret og i sammenhæng. Forbindelsen mellem dem skabes som hovedregel gennem personnummeret og i enkelte tilfælde gennem adressekoden.

Figur 1. Personstatistiksystemet



I personstatistiksystemet indgår ejejendomsstatikregistret som et af systemets elementer. Det indeholder bl.a. en nøgle mellem ejerpersonnummer og ejendomsidentifikation, således at man via denne nøgle kan trække på oplysninger om fast ejendom til anvendelse i personstatistikken. Endvidere er der gennem arbejdspladsstatistikregistret forbindelse til erhvervsstatistikken gennem arbejdstedsidentifikationen. Ved anvendelse af denne identifikation kan erhvervsregistrets oplysninger inddrages i personstatistikken som et led i klassificeringen af personer efter beskæftigelsesforhold.

4. Klassifikationsmoduler

Et klassifikationsmodul er et register, der har til formål at supplerer andre statistikregistre og statistikopgaver i øvrigt med visse relevante og generelt anvendte baggrundsoplysninger. Det kan ske ved direkte overførsel af moduloplysninger til et statistikregister, eller det kan ske ved at behandle fore-

liggende statistikregisteroplysninger efter bestemte regler, typisk med anvendelse af omsætningsnøgler. Hvis oplysningerne blot skal samles i et anonymt statistikprodukt, så vil det pågældende modul indgå i den samkøringsproces, som skal danne grundlaget for produktet.

Der indgår følgende 6 klassifikationsmoduler i statistiksystemet:

Befolkning/datomodulet afgrænser befolkningen på et givet tidspunkt.

Områdekodningsmodulet afgrænser geografiske områder på grundlag af adressekoder.

Familie/husstandsmodulet grupperer personer i familier og husstande efter de gældende definitioner.

Uddannelsesklassifikationsmodulet (UKM) anvendes til klassificering af personer efter højest afsluttede almene uddannelse, evt. afsluttet erhvervsuddannelse og evt. igangværende uddannelse.

Arbejdsstedsmodulet identificerer det enkelte lønmodtageransættelsesforhold ved personnummer og arbejdsstedsidentifikation og muliggør derved klassificering af personer efter virksomhedsforhold og virksomhedsklassificering efter personalesammensætning. Endvidere kan arbejdsstedets adressekode sammenholdes med bopælsadresse, hvorved der skabes grundlag for pendlingsstatistik.

Arbejdsklassifikationsmodulet (AKM) klassificerer den enkelte person efter væsentligste beskæftigelse (beskæftigelsesstatus, stilling og branche for vigtigste arbejdssted) samt mål for omfanget af beskæftigelse og ledighed i løbet af et kalenderår.

Klassifikationsmodulets funktion som generelt grundlag for hyppigt anvendte oplysninger til forskellige statistikområder indebærer en række fordele. De kan kort opregnes i følgende punkter:

højere datakvalitet
større effektivitet
konsistens mellem forskellige opgørelser

Statistikkens begreber skal være i størst mulige overensstemmelse med den verden, som statistikken beskriver. Kvaliteten af de anvendte statistikoplysninger vurderes i forhold hertil. De oplysninger, som indhentes fra de administrative systemer, kan hver for sig være utilstrækkelige som statistikdata i forhold til konkrete behov. Gennem nærmere analyser af de foreliggende administrative data udsøges de mest egnede og især kombinationer af disse, som ved oprettelsen af modulet indgår i et regelsæt for behandling af de ofte forskelligartede data fra mange kilder. Dette gør oplysningerne anvendelige på et højt kvalitetsniveau. I vedligeholdelsen af modulet tages højde for påtvungne ændringer i de administrative grunddata, og de nødvendige undersø-

gelser gennemføres med henblik på en tilpasning af regelsættet for databehandling, som så vidt muligt fastholder begrebsdefinitioner i forhold til tidligere år, forudsat naturligvis, at de fortsat er relevante.

Ved oprettelse og vedligeholdelse af modulet gennemføres de procedurer, som er nødvendige for at danne de ønskede baggrundsvariable èn gang for alle, hvilket betyder større effektivitet i statistikproduktionen. Den ofte meget komplicerede og ressourcekrævende databehandling, der er forbundet med disse procedurer, skal ikke gennemføres inden for de enkelte statistikområder, hvor der er behov for disse variable.

Ved at anvende modulernes standardklassifikationer vil der være sikkerhed for konsistens mellem forskellige statistiske opgørelser, hvor klassificeringer af pågældende art indgår. Definition, anvendte regler for dannelse og tidspunkt for opdatering efter ændringer vil være det samme. Det betyder også, at dokumentationen lettes, idet der kan refereres til modulet. Endvidere vil statistikbrugerne i højere grad blive fortrolig med de anvendte begreber.

Klassifikationsmodulerne skal være indrettet således, at de giver den bredest mulige dækning med hensyn til klassifikationernes anvendelse. Det betyder, at der skal være fleksibilitet i modulernes klassificeringer, således at relevante tilpasninger til den konkrete opgave kan foretages. På den måde kan det til en vis grad sikres, at brugen af modulerne ikke fører til et stift system, hvor man føler sig tvunget til at bruge klassifikationer, som måske ikke helt passer til formålet. Kravet om fleksibiliteten kan forekomme uforeneligt med hensynet til konsistens mellem statistiske opgørelser, men der er en lang række områder, hvor en valgt standardklassifikation både kan og bør benyttes. Hvor dette ikke er relevant, skal der være mulighed for anvendelse af modulets elementer så langt, det er foreneligt med opgørelsens formål. Som eksempel kan nævnes arbejdstyrkestatistikkens brug af AKM. I førstnævnte afgrænses lønmodtagere efter regler, der knytter sig til forhold på et givet tidspunkt, medens AKM's standardklassificering afgrænser denne gruppe efter en helårsbetragtning under hensyntagen til, hvad der har været væsentligst i løbet af et år. Denne forskel i afgrænsningen af lønmodtagere udelukker ikke, at AKM's stillingsklassifikationer kan overføres til og anvendes i arbejdsstyrkestatistikken. Stillingsklassifikationerne er dannet for alle personer, hvor grundmaterialet har givet mulighed for det uden hensyn til beskæftigelsesstatus. Disse oplysninger er derfor til rådighed også for alternativt afgrænsede lønmodtagere.

5. Statistikregistrene

Kilderne til statistikregistrenes dataindhold kan være administrative registre, klassifikationsmoduler, andre statistikregistre og data indsamlet direkte gennem interviews eller postspørgeskemaer.

Det enkelte statistikregister er som tidligere nævnt afgrænset i forhold en bestemt emnekreds eller et statistikområde, således at registrets indhold og populationsafgrænsning er foretaget under hensyntagen til, hvad der er relevant

inden for netop dette område. Der er til en vis grad ensartethed mellem emneafgrænsningen af de administrative registres indhold og opdelingen af det samlede statistikgrundlag i statistikregistre, hvilket betyder, at der for mange statistikregistres vedkommende vil være én hovedkilde og én eller flere andre kilder med supplerende oplysninger. De emne- eller sagstypeafgrænsede registre i det fælleskommunale system og deres anvendelse især inden for socialstatistikkens forskellige områder er eksempler på dette.

Den tilsyneladende klare forbindelse mellem hovedkilde og statistikregister betyder dog ikke, at de administrative oplysninger kan anvendes umiddelbart. Der kan være problemer med at tidsfæste oplysninger. I den administrative praksis kan tidsreferencen være bestemt af administrative handlinger, som ligger senere end tidspunktet for den hændelse eller observation, som skal indgå i statistikken. Endvidere kan begrebsdefinitioner i det administrative system afvige fra de statistisk relevante begreber på en måde, som umiddelbart vil forringe statistikkens kvalitet, i visse tilfælde til det uacceptable, hvis de administrative oplysninger bruges i deres rå form. Endelig forekommer utilstrækkelig ajourføring i visse tilfælde med den virkning, at der forekommer fejl og mangler i kildeoplysningerne. Det kan forekomme i administrative systemer med blandet automatisk og manuel indberetning, eller hvor fejlagtige angivelser ikke har administrative konsekvenser. Eksempel på sidstnævnte er indkomstskattesystemet, hvor det hidtil har været uden administrativ betydning, hvordan en A-indkomst er specificeret. Skatteberegningen vil give samme resultat. De manuelt indsendte hovedtal til et system, der er baseret på automatisk indberetning, kendes også fra skattesystemet.

De nævnte forhold har ført til, at der i forbindelse med oprettelse og ajourføring af et statistikregister ofte vil være behov for at trække på flere kilder for at sikre korrekt tidsafgrænsning, opnå de mest relevante begrebsdefinitioner samt, så vidt det er muligt, kompensere for fejl og mangler.

Ved direkte indsamling af statistikoplysningerne vil spørgeskemamaterialet typisk udgøre hovedgrundlaget, men man vil ved tilrettelæggelsen af den pågældende statistik inddrage og benytte sig af de nødvendige oplysninger, der allerede foreligger og derfor kan indhentes fra andre statistikregistre. Der indgår naturligvis i den forbindelse en vurdering af registeroplysningernes kvalitet, idet formålet med den direkte henvendelse kan være at kompensere for utilstrækkelige registeroplysninger.

Behovet for at undersøge årsagssammenhænge mellem forskellige fænomener har ført til, at der oprettes registre med et tværgående indhold. Den emneorienterede opbygning af registersystemet har betydet, at nye samkøringer skal foretages, hvis der opstår konkrete behov for at se emner i sammenhæng og dermed datasammenstillinger ud over, hvad der foreligger i det enkelte register. Opgaver af denne karakter er blevet og bliver fortsat gennemført på grundlag af samkøringsbaserede anonyme datasæt, men mulighederne for at analysere forløb over en vis tid, som systemet nu lægger op til, har skabt stor interesse især blandt forskere, og det har ført til, at der dannes såkaldte integrationsregistre. Registre af denne karakter er dels et led i det almindelige

statistikberedskab, og dels er det en konsekvens af databehov i forbindelse med konkrete serviceopgaver.

En særlig form for grunddata i det statistiske registersystem er indleverede datasæt. I forbindelse med forskningsprojekter gennemført uden for Danmarks Statistik kan der være indsamlet data, som ønskes samkørt med oplysninger i Danmarks Statistiks registre for at opnå en bredere belysning af det omhandlede emne. Hvis personnumre indgår i primærmaterialet er en sådan samkøring mulig, og forskningsmaterialet kan på denne måde tilføres supplerende oplysninger eller klassifikationer. Det samkørte produkt er dog ikke et led i registersystemet, da det vil være anonymiseret, så det kan kun bruges isoleret til analyser i forbindelse med det konkrete forskningsprojekt i Danmarks Statistik eller til tabelkørsler.

Strukturen med en række emneorienterede eller formålsbestemte statistikregistre kan diskuteres. Den umiddelbare konsekvens af dette er en ret betydelig redundans i systemet. Den samme oplysning om en given enhed vil forekomme i flere og i enkelte tilfælde en del registre. Hensigtsmæssigheden af dette må anskues ud fra forskellige synsvinkler. Det teknisk/økonomiske synspunkt tilsiger umiddelbart, at en given oplysning kun skal være registreret ét sted. Pladskravene ville derved blive reduceret, og dokumentation og vedligeholdelse ville blive forenklet. Beredskabet i forholdet til anvendelsen er lidt vanskeligere at vurdere, da det vil afhænge af den valgte tekniske løsning og adgangsbestemmelserne. Netop sidstnævnte er et væsentligt led i argumentationen for det nuværende system. Med opdelingen i en række registre fastlægger man for hvert af disse, hvad det må indeholde, hvad det må bruges til, og hvem der må bruge det. Specielt ved fastlæggelsen af adgangen til et register sikres det, at kredsen er lille, og at samme person ikke har adgang til mange registre. Disse forhold gør det meget gennemskueligt, hvad der foregår, når de mange oplysninger anvendes i statistisk sammenhæng, og det betyder, at den utryghed, der forekomme, vil være forsvindende. Endvidere er det enkelte register umiddelbart klar til brug inden for det område, som det dækker, så beredskabet er på et højt niveau. Når spørgsmålet tages op, så skyldes det bl.a., at der i stigende grad dannes integrationsregistre, som på en vis måde er et brud med den gældende registersystemorganisation, og så øger de yderligere redundansen i systemet.

I bilaget er de enkelte statistikregistre anført med en kort omtale af deres indhold.

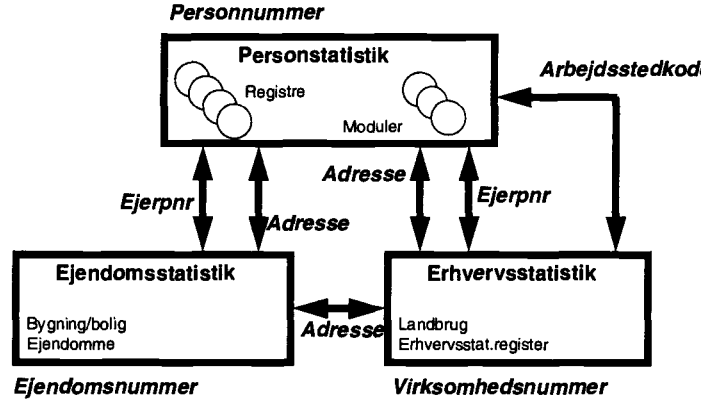
6. Strukturen i det samlede statistiksystem

Det grundlæggende element i statistiksystemet er personnummeret, som er den fælles identifikation af statistikenheder i personstatistikregistrene. Ved anvendelse af dette nummer kan i princippet alle oplysninger kombineres. I praksis sker det ikke, da de foretagne samkøringer af oplysninger kun omfatter data, som er relevante i forhold til konkrete statistikopgaver. For opgaver af den karakter vil registrene udgøre en helhed, som er til rådighed for statistiske opgørelser, og en angivelse af de faktiske muligheder for, hvad der kan

laves af statistik, vil derfor i princippet være en kombineret dataliste for samtlige registre. Der sættes grænser for, hvad der må kombineres, idet kombinationer af oplysninger skal være sagligt begrundede, og det må ikke være muligt at genkende enkeltpersoner.

Som tidligere omtalt omfatter systemet også nøgler til ejendoms-/boligenheder samt arbejdssteder via adressekode og arbejdsstedskode. Der er derfor mulighed for at overføre bolig- og erhvervsstatistiske oplysninger til personstatistikken, og der kan overføres personstatistiske oplysninger til boligstatistikken og erhvervsstatistikken. I nedenstående figur er dette sammenhængende statistiksystem skitseret med angivelse af de identifikationer, som knytter systemets elementer sammen.

Figur 2. Det samlede statstiksystem



De tre systemer er forbundet gennem adressekoder. Mellem ejendomsstatistikken og personstatistikken er der en personnummersammenhæng for så vidt angår ejere af fast ejendom. Tilsvarende gælder mellem erhvervsstatistikken og personstatistikken for virksomhedsindehavere. Endelig er der en sammenhæng mellem lønmodtagere og virksomheder gennem arbejdsstedskoden, så oplysninger om den ansatte person kan forbindes med oplysninger om ansættelsessted.

Bilag

Registre i Danmarks Statistik

Nedenfor gives en oversigt statistikregistrene i Danmarks Statistik pr. 1. januar 1994. For hvert register er navnet og hovedformålet anført. En nærmere omtale af det enkelte registers indhold og anvendelse mv. frem går af registerforskrifterne for de pågældende registre. Registre vedrørende Grønland er udeladt af oversigten, da de forventes overført til Hjemmestyret, således at den grønlandske befolkning ikke vil være omfattet af registersystemet.

Statistikregistrene er rubriceret i to hovedgrupper: a) personstatistikområdet og b) erhvervsstatistikområdet.

A. Statistikregistre inden for personstatistikområdet

1. **Befolkningsstatistikregistret**
   Hovedformål: At danne grundlag for befolkningsstatistikken for Danmark, omfattende opgørelser af den samlede befolkning på bestemte tidspunkter samt af befolkningsbevægelser.

2. **Medicinsk fødsels- og dødsfaldsstatistikregister**
   Hovedformål: At danne grundlag for statistik angående fødsler (herunder dødfødsler) og dødsfald.

3. **Folke- og boligtællingsregistret**
   Hovedformål: At danne grundlag for generel og specialiseret statistik over befolkningens størrelse og sammensætning med hensyn til demografiske kendetegn, beskæftigelses- og erhvervsforhold, uddannelsesforhold, familieforhold, boligforhold og pendling på givne tællingstidspunkter.

5. **Erhverv/dødelighedsstatistikregistret**
   Hovedformål: At danne grundlag for undersøgelser af den talmæssige sammenhæng mellem beskæftigelsesforhold og dødsrisiko i den danske befolkning.

6. **Indkomststatistikregistret**
   Hovedformål: At danne grundlag for statistiske opgørelser vedrørende befolkningens indkomst- og formueforhold.

8. **Beskæftigelsesstatistikregistret**
   Hovedformål: At danne grundlag for statistiske oplysninger om befolkningens beskæftigelsesforhold.

9. **Beskæftigelsesundersøgelsesregistret**
   Hovedformål: At danne grundlag for statistiske opgørelser over arbejdsstyrkens størrelse og sammensætning.

10. Ungdomsbeskæftigelsesundersøgelsesregistret
Hovedformål: At danne grundlag for statistiske opgørelser over unges beskæftigelses- og ledighedsforhold.

11. Arbejdsstyrkeundersøgelsesregistret
Hovedformål: At danne grundlag for statistiske opgørelser over arbejdsstyrkens størrelse og sammensætning samt befolkningens beskæftigelses- og ledighedsforhold.

12. Løn- og personalestatistikregistret for den offentlige sektor
Hovedformål: At danne grundlag for statistiske oplysninger om løn- og beskæftigelsesforhold for ansatte i stat og kommuner mv.

13. Arbejdsklassifikationsmodulet
Hovedformål: At danne grundlag for beskæftigelsesoplysninger til brug ved personstatistiske opgørelser.

14. Arbejdspladsstatistikregistret
Hovedformål: At give statistiske oplysninger om befolkningens beskæftigelsesforhold og geografiske fordeling efter arbejdssted og om lokale erhvervsenheders fordeling efter placering, branche og personaleforhold.

15. Arbejdsløshedsstatistikregistret
Hovedformål: At danne grundlag for statistiske oplysninger til belysning af ledighedens struktur og udvikling.

16. Selvangivelsesstatistikregistret
Hovedformål: At danne grundlag for en hurtig, detaljeret statistik over befolkningens indkomst-, fradrags- og formueforhold.

17. Uddannelsesstatistikregistret
Hovedformål: At danne grundlag for statistik om uddannelsessøgendes bevægelser gennem uddannelsessystemet.

18. Statistikregister vedrørende langvarigt uddannedes beskæftigelse
Hovedformål: At danne grundlag for statistik om nyuddannedes forhold på arbejdsmarkedet.

19. Uddannelsesklassifikationsmodulet
Hovedformål: At danne grundlag for statistik om befolkningens skole- og erhvervsuddannelser.

20. Bistandslovsstatistikregistret
Hovedformål: At danne grundlag for statistik over modtagere af visse ydelser efter bistandsloven.

21. Kriminalstatistikregistret
Hovedformål: At danne grundlag for statistik om personer, der pålægges en strafferetlig sanktion.

22. Færdselsuheldsstatistikregistret
Hovedformål: At danne grundlag for statistik over færdselsuheld og for trafiksikkerhedsforskning i øvrigt.

23. Huslejepanelets register
Hovedformål: At danne grundlag for de halvårlige huslejeundersøgelser, hvis resultater indgår ved beregning af indeks for boligudgift i reguleringspristallet og i forbrugerprisindekset.

24. Udsnitsarkivet
Hovedformål: At danne grundlag for at udnytte allerede indsamlet datamateriale i nye sammenhænge eller efter nye metoder og udgøre beredskab for tidsrække- og tidsforløbsanalyser.

26. Omnibusundersøgelsesregistret
Hovedformål: At danne grundlag for statistiske og videnskabelige undersøgelser, der baseres på interview med et mindre udsnit af den danske befolkning.

27. Forbrugsundersøgelsesregistret
Hovedformål: At give statistiske oplysninger om befolkningens forbrug, indkomst- og opsparingsforhold, herunder at danne grundlag for vægtsammensætningen i reguleringspristallet og for analyser af offentlige ydelser.

28. Pensionsstatistikregistret
Hovedformål: At danne grundlag for statistiske opgørelser vedrørende udbetaling af pensioner i Danmark.

29. Børnetilskudsstatistikregistret
Hovedformål: at danne grundlag for statistiske opgørelser vedrørende udbetaling af børnetilskud og ungdomsydelser i Danmark.

30. Boligstøttestatistikregistret
Hovedformål: at danne grundlag for statistiske opgørelser vedrørende udbetaling af boligsikring og boligydelse i Danmark.

31. Erhverv/cancerstatistikregistret
Hovedformål: At danne grundlag for statistiske undersøgelser af den talmæssige sammenhæng mellem cancerforekomst og erhvervs- og andre samfundsmæssige forhold i Danmark.

32. Statistikregistret vedrørende arbejdskraftmobilitet
Hovedformål: At danne grundlag for statistiske undersøgelser vedrørende arbejdskraftefterspørgsel og -mobilitet.

33. Sygedagpengestatistikregistret
Hovedformål: At danne grundlag for statistiske opgørelser vedrørende udbetaling af sygedagpenge i Danmark.

34. Daginstitutionsstatistikregistret
Hovedformål: At danne grundlag for statistiske opgørelser vedrørende børn og unges benyttelse af daginstitutioner og dagpleje, der drives i henhold til bistandsloven.

35. Ejendomsstatistikregistret
Hovedformål: At danne grundlag for statistik vedrørende fast ejendom i Danmark.

36. Sygesikringsstatistikregistret
Hovedformål: At danne grundlag for statistiske opgørelser om ydelser i henhold til sygesikringsordningen i Danmark.

38. Statistikregistre vedrørende enkeltfagskursister
Hovedformål: At danne grundlag for statistiske undersøgelser vedrørende enkeltfagskursisters sociale rekruttering og mobilitet.

40. Indkomstforløbsregistret
Hovedformål: At danne grundlag for uddrag af data til en flerårig modelbefolkning med henblik på konsekvensberegninger samt analyser af den økonomiske udvikling.

41. Ressourceundersøgelsesregistret
Hovedformål: At danne grundlag for en undersøgelse af de ressourcefattige gruppers situation i 1988 samt det samspil af forhold, som ligger forud.

42. Statistikregistret vedrørende kvinder i mandefag
Hovedformål: At belyse arbejdsmarkedsforholdene for kvinder i faglærte og ikke-faglærte mandefag.

43. Statistikregistret vedrørende indkomsterstattende ydelser
Hovedformål: At danne grundlag for statistiske opgørelser vedr. udbetaling af indkomsterstattende ydelser.

44. Statistikregistret vedrørende tidsanvendelsen 1950-58
Hovedformål: At belyse befolkningens tidsanvendelsesmønster.

45. Flytteundersøgelsesregistret
Hovedformål: At danne grundlag for detailtabeller til analyse af flyttestrømme over en flerårig periode.

46. Statistikregistret vedrørende ledige med kompetencegivende uddannelse
Hovedformål: At belyse tilknytningen til arbejdsmarkedet efter uddannelsens afslutning.

47. Motorkøretøjsstatistikregistret
Hovedformål: At danne grundlag for statistiske opgørelser vedrørende ejerforhold og anvendelse af motorkøretøjer.

48. Statistikregistret vedrørende erhvervsprogligt uddannede
Hovedformål: At belyse de erhvervsprogligt uddannedes faglige organisationsforhold og tilknytning til arbejdsmarkedet.

49. Valgstatistikregistret
Hovedformål: At danne grundlag for valgstatistikken vedr. valg til amtskommunale og kommunale råd.

50. Statistikregistret for arbejdsmarkedsforskning
Hovedformål: At danne grundlag for person- og virksomhedsorienterede statistiske undersøgelser af arbejdsmarkedsforhold.

51. Velstandsstatistikregistret
Hovedformål: At danne grundlag for statistiske opgørelser vedr. befolkningens velstand.

52. Lønstatistikregistret for den private sektor
Hovedformål: At give oplysninger til statistiske opgørelser vedr. lønforhold inden for den private sektor.

53. Erhvervsindlæggelsesregistret
Hovedformål: At danne grundlag for statistiske undersøgelser af den talmæssige sammenhæng mellem sygdomsforekomst og erhvervs- og andre samfundsmæssige forhold i Danmark.

55. Sygehusbenyttelsesstatistikregistret
Hovedformål: At danne grundlag for statistiske opgørelser vedr. personer, der har været indlagt på somatiske sygehusafdelinger i Danmark.

56. Statistikregistret vedrørende tilpasningsstrategien mellem arbejdsliv og familieliv
Hovedformål: At belyse familielivets påvirkning af arbejdslivet og arbejdspladsernes tilpasning til familielivet.

57. Social- og arbejdsforløbsregistret
Hovedformål: At danne grundlag for dannelse af modeldata for en flerårig periode til analyser af sammenhænge mellem beskæftigelsesforhold og modtagelse af sociale ydelser i befolkningen.

59. Statistikregistret vedrørende børns opvækstvilkår
Hovedformål: At belyse den nuværende situation for unge, der har været anbragt uden for hjemmet, sammenlignet med en gruppe jævnaldrende, hvor familien har fået langvarig bistandshjælp, og en gruppe af tilfældigt udvalgte jævnaldrende.

60. Statistikregistret for fertilitetsforskning
Hovedformål: At danne grundlag for personorienterede statistiske undersøgelser af fertilitetsforhold.

61. **Regionalt forløbsregister**
Hovedformål: At danne grundlag for dannelse af modeldata for en flerårig periode til analyser af regionale sammenhænge mellem sociale, arbejdsmarkeds-, indkomst- og boligforhold i befolkningen.

62. **Statistikregistret vedrørende den kommunale beskæftigelsesindsats**
Hovedformål: At belyse situationen for personer, der har været omfattet af kommunale beskæftigelsesforanstaltninger samt effekten af disse.

63. **Statistikregistret vedrørende arbejdsmarkedspolitiske foranstaltninger**
Hovedformål: At danne grundlag for forskningsprojekter til belysning og vurdering af arbejdsmarkedspolitiske foranstaltninger.

64. **Statistikregistret vedrørende sociale processer og boligforhold**
Hovedformål: At danne grundlag for dannelse af modeldata for en flerårig periode til analyser af sammenhænge mellem udviklingen i arbejdsmarkedstilknytning, indkomstsammensætning, uddannelsesforhold, familie- og husstandsforhold og boligforhold.

B. Statistikregistre inden for erhvervsstatistikområdet

1. **Det erhvervsstatistiske registersystem**
Hovedformål: At danne grundlag for indhentning af oplysninger om erhvervsmæssige forhold samt danne grundlag for statistiske opgørelser vedrørende samme.

2. **Landbrugsstatistikregistret**
Hovedformål: At danne grundlag for indhentning af oplysninger vedrørende landbrug og gartneri samt danne grundlag for statistiske opgørelser vedrørende samme.

I den endelige version af bogen "Det personstatistiske registersystem" vil der ingå en mere detaljeret beskrivelse af registrenes indhold.

Administrative data som statistikdata

Lars Thygesen

1. Indledning

I dette kapitel diskuteres det, hvilke krav der skal være opfyldt, for at man med held kan basere den officielle, samfundsbelysende statistik på administrative registre. Hvilke typer af data, der må anses for helt centrale, og hvordan forskellige datatyper kan indgå i grundlaget. Desuden omtales nogle praktiske problemer i forbindelse med tilrettelæggelsen af en registerbaseret statistik, og hvordan de kan løses.

2. Statistikkens krav til data

De krav, der må stilles til datagrundlaget, dvs. til de administrative registre, har at gøre dels med datas indholdsmæssige dækning, dels med kvalitative forhold. Det er umuligt at opstille helt præcise krav, som under alle omstændigheder skal opfyldes, før det er interessant at producere statistik på grundlag af administrative registre. Det er da også klart, at selv isolerede, enkeltstående administrative registre kan anvendes og bliver anvendt over alt i verden som statistikgrundlag, og nogle af de ældste statistikker i Danmark har været baseret på optegnelser som fx kirkebøgerne.

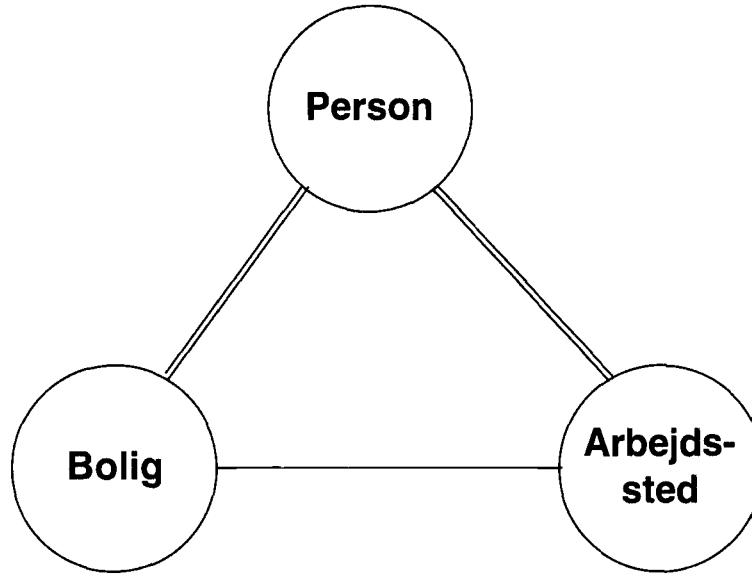
I det følgende forsøger jeg at opstille nogle krav, som må opfyldes, for at det skal være muligt at skabe et *sammenhængende, dækkende personstatistisk system*, som i hovedsagen er baseret på registrene, dvs. et system af samme karakter som det danske.

2.1. Indhold

Først og fremmest må registrene overhovedet indeholde data, der dækker de vigtigste emner i den samfundsbelysende statistik.

Enheder

Der er mange forskellige enheder, man ønsker at kunne belyse i statistikken, men der er tre typer enheder, der er særlig vigtige: Personer, virksomheder og boliger. I denne fremstilling beskæftiger vi os hovedsagelig med personstatistikken, og i den forbindelse er personerne selvfølgelig i centrum, men også de to andre enheder har stor betydning i personstatistikken. Det fremgår fx af, at en traditionel folke- og boligtælling også knytter an til de boliger, hvor personerne bor, og de virksomheder (arbejdssteder), hvor de arbejder. En stor del af statistikbrugerne, både i den offentlige og den private sektor, har brug for netop disse relationer.



Der er naturligvis mange andre enheder, det er vigtigt at have registreringer om. Nogle af disse enheder er *hændelser*, som vedrører personerne, fx de klassiske demografiske hændelser (fødsler, dødsfald osv.) eller sociale begivenheder af forskellig slags. Andre enheder af betydning kan også nævnes, såsom biler.

Præcis definition af enheder

For den statistiske anvendelse er det vigtigt, at enhederne er veldefinerede. Dette er ikke noget problem for "naturlige" enheder som personer eller biler, hvor definitionen og afgrænsningen over for andre enheder så at sige giver sig af sig selv. Også begrebet bolig er nogenlunde let at håndtere, selv om der er nogle grænsetilfælde, hvor eksistensen eller afgrænsningen af en bolig kun kan fastslås efter nærmere fastlagte regler. Nogle enheder fremkommer som kombinationer af "naturlige" enheder. Det gælder familier og husstande, og problemer i den forbindelse er omtalt i et supplerende papir. Det bliver hurtigt mere problematisk, når man ser på andre typer enheder, og her er erhvervsenheder et godt eksempel. En virksomhed eller et arbejdssted er et abstrakt begreb, og eksistens og afgrænsning kan kun konstateres efter meget nøje fastlagte regler. Uklarhed og uenighed om disse definitioner har gennem årtier i høj grad vanskeliggjort udnyttelsen af registre i erhvervsstatistikken og har i øvrigt også ødelagt sammenhæng og konsistens i statistikken.

Dækning

Registrene skal gerne indeholde oplysning om alle enheder i landet af de nævnte typer, dvs. de skal være totaldækkende.

Som bekendt er det også muligt ved hjælp af stikprøvemetoder at lave statistik baseret på materialer, der ikke er totaldækkende. Det forudsætter imidlertid, at den stikprøve, der indgår i grundlaget, er udvalgt, så den er repræsentativ. Der kendes næppe nogen eksempler på, at administrative registre er indrettet, så de dækker en repræsentativ stikprøve af befolkningen.

Totaldækning er værdifuld

En af de store styrker ved den registerbaserede statistik i forhold til den traditionelle, spørgeskemabaserede statistik er netop, at man ofte har mulighed for at få totaldækning, eller meget tæt på dette. I spørgeundersøgelser må man dels som regel nøjes med en repræsentativ stikprøve, dels acceptere et uundgåeligt bortfald, der som regel er skævt, således at de, der indgår i undersøgelsen, adskiller sig systematisk fra dem, der ikke er med. Bortfaldet har ofte en betydelig størrelse, typisk fra 15 til 50%, når det drejer sig om frivillige undersøgelser blandt privatpersoner. I nogle undersøgelser i udlandet ser man endog bortfald på op til 80%, hvilket i de fleste sammenhænge gør resultaterne helt værdiløse [Van Bastelaer (1993)].

Bredde

Det er ønskeligt, at dataindholdet er bredt og omfatter mange variable om de tre typer enheder.

2.2. Identifikationer

Identifikationer spiller en meget stor rolle, både for den administrative registerdrift (jf. nedenfor) og for den statistiske udnyttelse. For statistikken er det især vigtigt, at fælles identifikationer anvendt i flere registre kan bruges til at sammenknytte oplysninger fra de forskellige kilder om én enhed. Da registrene ikke er dannet til statistiske formål, skulle det være sært, om de netop indeholdt data i de kombinationer, der er ønskelige i statistikken. Det er også sjældent, at dette forekommer i praksis. Derfor er det en vigtig del af dannelsen af næsten enhver registerbaseret statistik, at data skal samkøres ved hjælp af fælles nøgler.

Den vigtigste nøgle i denne forbindelse er naturligvis *personnummeret*, som vedligeholdes i CPR, og hvis formål netop er at være en fælles identifikation, der bruges af hele den offentlige sektor.

To andre identifikationer, der har stor betydning, er *boligens adresse*, som i Danmark er en helt præcis identifikation for hver bolig (og for bygninger og til en vis grad for erhvervsenheder), og *arbejdsstedsnummer*, som identificerer de lokale erhvervsenheder. Det er disse identifikationer, der skal bruges til at fastholde relationerne mellem personer, boliger og arbejdssteder, jf. figuren ovenfor.

2.3. Tidsreferencer

Tidsdimensionen spiller en ganske særlig rolle i den samfundsbelysende statistik. Inden for alle emneområder er det nødvendigt, at man kan foretage sammenligninger over tiden. Hvis man ser i statistiske publikationer eller databanker, er tidsdimensionen stort set altid med. Derfor er det helt nødvendigt for den statistiske anvendelighed, at forskellige datoer er godt registreret i registrene.

Hændelsesdatoer nødvendige...

Først og fremmest må datoen for ændringer findes. Det gælder for mange forskellige slags hændelser. Nogle af de vigtigste er datoerne for enhedernes "fødsel" og "død", men også tiden for andre ændringer vedr. enhederne har

betydning. Der tænkes her på det reale tidspunkt, hvor en hændelse har fundet sted, der ændrer ved et datum, fx dato for en flytning eller dato for et brancheskift for en virksomhed.

...men ikke altid muligt at tidsfæste

Det sidstnævnte eksempel henleder opmærksomheden på det forhold, at det i praksis kan være overordentlig svært at få fastslået den reale hændelsesdato: Hvornår skifter en virksomhed branche? Her er der tale om en variabel, som så at sige ændres gradvist, efterhånden som virksomhedens fokus skifter, og det er måske ikke muligt at fastslå, hvornår den nye aktivitet er den dominerende. Men også for mange andre variable, som i princippet kan observeres på ethvert tidspunkt, og hvor en ændring derfor kan tidsfæstes ganske nøje, er der i praksis problemer med registreringen. Det kan skyldes, at det ikke nødvendigvis er interessant for den administrative brug af data, hvornår en dataværdi er trådt i kraft, blot man på ethvert tidspunkt har adgang til den mest muligt aktuelle værdi. I andre tilfælde kan problemerne skyldes, at det ikke er muligt eller praktisk gennemførligt at få registreret hændelserne i real tid, men at man i stedet må nøjes med at få kendskab til dataændringer ved tilbagevendende forespørgsler, fx en gang om året i forbindelse med en selvangivelse.

Hvis der ikke er mulighed for at få registreret nøjagtige datoer, er tilnærmelser selvfølgelig anvendelige og langt bedre end ingenting.

Registreringsdato næste lige så vigtig

I tillæg til hændelsesdatoer eller ændringsdatoer er der brug for *registreringsdatoer,* dvs. angivelse af hvornår den pågældende dataværdi er blevet optaget i registret. Dette har sin afgørende betydning, når registret skal opfattes som en *statistisk model* af den virkelighed, som statistikken skal beskrive, jf. næste afsnit.

Det ideelle er derfor, at ethvert datum i det administrative register er ledsaget af to datoer: En ikrafttrædelsesdato og en registreringsdato. Virkelighedens registre adskiller sig ofte stærkt fra dette ideal.

Foreløbige og endelige tal

Forsinkelser i opdateringen af de administrative registre fører i mange tilfælde til, at man må publicere flere udgaver af tallene: *Foreløbige tal, der* er nødvendige af hensyn til behovet for aktuelle tal, men som mangler præcision på grund af manglende opdateringer, og *endelige tal,* som er mere fuldkomne, men som først fremkommer på et senere tidspunkt. Flere mellemversioner kan forekomme. Eksistensen af flere versioner af tallene kan imidlertid forårsage stor forvirring og ulempe hos brugerne og bør af den grund undgås, hvis det er muligt.

2.4. Registret som statistisk model

Når et registersystem skal bruges som grundlag for statistik, betyder det, at registret skal kunne opfattes som en model af den gruppe enheder, statistikken skal beskrive.

I nogle tilfælde må det yderligere forudsættes, at de dataændringer, der indføres i registret, kan opfattes som model af hændelser, der beskrives i stati-

stikken. Det gælder fx befolkningsstatistikken, hvor opgørelser af hændelser i befolkningen (flytninger, vielser etc.) skal kunne udledes af dataændringer i CPR. Registrene må derfor være indrettet sådan, at opgørelser af registrets bestand og af bevægelserne heri tidsmæssigt kan afgrænses i overensstemmelse med de tidspunkter, hvor hændelserne faktisk har fundet sted.

Korrektioner

Til bevægelsesopgørelser må det kræves, at det er muligt i registret at skelne mellem virkelige hændelser og korrektioner og mere teknisk betonede dataændringer. Indholdet af dette sidste krav kan bedst belyses gennem nogle problemer, som rent faktisk har gjort sig gældende ved befolkningsstatistikkens anvendelse af CPR, - selvom netop CPR må siges at være det registersystem, der på nuværende tidspunkt bedst imødekommer statistikproduktionens behov.

Opgørelserne over indenlandske flytninger tager udgangspunkt i oplysninger om de personer, som i en given periode får ændret adresseangivelse i CPR. I denne gruppe indgår de personer, som faktisk er flyttet i perioden, men der indgår tillige en række persongrupper, som ikke har interesse for statistikken, nemlig bl.a.:

1. Personer, der er omfattet af grænsereguleringer mellem kommunerne eller kommunesammenlægninger.

2. Personer, hvis adresse ændres som følge af en vejnavneændring eller en omnummerering eller omkodning af vejen.

3. Korrektion af forkerte adressekoder.

Navnlig adresseændringer af type 2 har haft et betydeligt omfang, som frem til slutningen af 1970'erne blev anslået til ca. 1/4 mio. personer årligt. Selvom kommunerne efter reglerne skulle give indberetningerne af disse adresseændringer en særlig markering, er det hidtil kun sket i knap halvdelen af tilfældene, og af denne grund var det ikke muligt at fremstille en brugbar generel statistik over samtlige indenlandske flytninger, men kun over den del af flytningerne, der passerer kommunegrænser.

Disse problemer blev stort set løst ved en teknisk omlægning af CPR, som fandt sted i 1978.

De egentlige korrektioner af forkert indberettede data frembyder imidlertid et problem for statistikken, selv om man er i stand til at kende dem fra de virkelige hændelser. Korrektioner må nemlig i mange tilfælde betragtes som rettelser til tidligere indberettede hændelser: En adressekorrektion kan være en rettelse til en indberetning om flytning. En korrektion kan også annullere en tidligere indberettet hændelse, hvilket i CPR-systemet navnlig finder sted, når en person, efter at have meldt udvandring til folkeregistret, ombestemmer sig og bliver i landet.

Korrektioner til hændelser er vanskelige at behandle, fordi de i realiteten kan blive indberettet med vilkårlig lang forsinkelse. I befolkningsstatistikken har

man valgt at holde dem helt uden for opgørelserne af bevægelserne, men denne fremgangsmåde er kun brugbar, fordi korrektionernes antal er ringe, sammenlignet med hændelsernes antal.

2.5. Stabilitet

Som nævnt er det en vigtig egenskab ved statistikken, at den kan beskrive tidsforløb. Hvordan udvikler en bestemt størrelse sig fra måned til måned, fra år til år. Derfor er det også meget ønskeligt, at begreberne i de administrative registre ligger fast over længere perioder. I modsat fald opstår der ofte store problemer med at opnå sammenlignelige tal fra den ene periode til den næste. I nogle tilfælde, men langt fra alle, har man dog mulighed for at korrigere mere eller mindre præcist for ændringerne. Der er også eksempler på, at man har måttet opgive at offentliggøre en statistik i en periode, fordi det har været umuligt at lappe på konsekvenserne af ændringer i registrene, jf. senere i dette afsnit.

2.6. Kvalitet

De kvalitetskrav, den statistiske anvendelse stiller, falder i nogen grad sammen med de krav, som også må være opfyldt af hensyn til registrenes primære formål.

Relevans

Det første kvalitetskrav, man må stille til grunddata, er, at de skal være *relevante* for statistikken. D.v.s. at data skal handle om de begreber, der ønskes belyst i statistikken. Det er langt lettere (men alligevel ikke let!) at kontrollere relevansen i traditionel statistik, som baseres på data, der er specielt indsamlet til formålet. Når statistikken skal baseres på data, der er indsamlet til andet formål, må man som oftest lade sig nøje med det, som er nødvendigt eller ønskeligt til det administrative formål.

I gunstige tilfælde kan statistikerne dog have held til at påvirke indholdet i registrene, jf. afsnit 5 nedenfor.

Der er også mulighed for at lade forskellige datakilder supplere hinanden gennem samkøring, således at man danner den ønskede oplysning (fx arbejdsstilling) ved at sammenholde en række oplysninger fra forskellige registre, der mere eller mindre perfekt belyser forholdet (i eksemplet: stilling fra CPR, arbejdsfunktion fra lønregistre, arbejdsløshedskasse fra CRAM, osv.). Statistikvariablen bliver på denne måde dannet ved *estimation*, jf. kap. 4.2.

Risikoen for et skævt verdensbillede

I forbindelse med diskussionen om relevans må man være opmærksom på risikoen for, at et statistisk system, der udelukkende eller i meget høj grad er baseret på offentlige myndigheders registre, kan komme til at give et verdensbillede, der er præget af myndighedernes briller. Begreber eller persongrupper, som ikke eksisterer i skattevæsenets eller socialforvaltningens referenceramme, kommer måske slet ikke med i statistikken. Det kunne fx være hjemløse personer.

For at imødegå denne risiko er det vigtigt, at der eksisterer nogle basisregistre som fx CPR, som har til opgave at registrere alle enheder uden at tænke på noget specifikt administrativt formål. Endvidere er det en sikkerhed, at mange registreringer fra forskellige myndigheder kan samkøres, så man kan få en statistik, der ikke er bundet til en enkelt administrativ synsvinkel. Alligevel er det vigtigt at være opmærksom på, at registrene ikke kan eller skal give os hele vores samfundsbillede, og at der er behov for at indsamle data specielt til at belyse de sider, som ikke dækkes i registrene. Disse ekstra data kan indsamles gennem interview eller spørgeskema, og de bør ideelt have identifikationer, som gør det muligt at sammenholde dem med registerdata. Dette omtales mere detaljeret i kap. 4.5.

Pålidelighed

Det næste krav, som må stilles til statistikkens grunddata, er kravet om pålidelighed. Det gælder uanset om data kommer fra spørgeskemaer eller registre. Der må være en høj grad af sikkerhed for, at de registrerede data belyser de faktiske forhold. Hvis det fremgår, at en person er maler, og han faktisk er rigsstatistiker, risikerer man fejlagtige konklusioner af den færdige statistik.

Det kan ganske vist indvendes, at så længe fejlene ikke er systematiske, vil de ikke resultere i skæve randfordelinger: Fejl i den ene og i den modsatte retning vil have en tendens til at ophæve hinanden. Men når man, som det ofte er nødvendigt, ser på *statistiske sammenhænge* mellem flere variable, bliver det et problem med sådanne fejl, idet de kan forstyrre bedømmelsen af sammenhængene.

Hvis fejlene systematisk går i en bestemt retning, bliver problemerne endnu mere voldsomme.

Præcision

Endelig er det vigtigt, at data registreres med en præcision, der passer til statistikkens behov, d.v.s. at den skala, der benyttes, er tilstrækkelig fin. Som eksempel herpå kan nævnes, at specifikationsgraden af stillingsoplysning skal være temmelig fin, hvis den skal kunne bruges til epidemiologiske undersøgelser, hvor sygdomsrisiko skal ses i forhold til arbejdsspecifikke påvirkninger.

3. Forskellige typer administrative registre

De administrative registre kan grupperes efter deres formål i basisregistre og målrettede registre

Basisregistre

Ved basisregistre forstås registre, der drives med det formål at ligge til grund for den offentlige administration i almindelighed eller i hvert fald for flere forskellige administrationsgrene. Disse registre har typisk til opgave at holde styr på bestanden af enheder, dvs. registrere enhedernes fødsel og død og dermed også hvilke enheder, der til enhver tid er til stede i populationen. Desuden skal de vedligeholde identifikationsoplysningerne, som bruges af de andre administrative registre, og de skal indeholde nogle få *basisdata*, som er af fælles interesse for forvaltningerne.

Basisregistrene skal være tilgængelige for flere forvaltninger. Adgangen kan ske på flere måder, fx ved at opdateringer vedr. de enheder, der er relevante for en forvaltning, automatisk videregives til forvaltningens registre fra basisregistret. I de sidste 10 år er det blevet almindeligt, at forvaltningerne har online-adgang til basisregistrene. Det er også typisk sådan, at opdateringer til basisregistrene kommer fra en lang række forskellige myndigheder, hvorved datakvaliteten kan blive usædvanlig høj. Hertil kommer, at basisregistrene ideelt set ikke er påvirket af én bestemt forvaltnings synsvinkel; de forsøger i stedet at tegne et neutralt billede af enhederne. Basisregistrene har derfor også den allerstørste betydning for statistikken.

Som typiske eksempler på basisregistre kan nævnes CPR, BBR og Det Centrale Erhvervsregister.

Organisatorisk kan basisregistrene være placeret på forskellig måde: Enten hos en "uafhængig" myndighed, som er indrettet med det formål at føre registret, eller hos en myndighed, der er den primære bruger. Set fra statistikkens synspunkt er det ideelle den uafhængige placering, som kendes fra CPR, hvor det centrale ansvar er placeret i et kontor i Indenrigsministeriet. CPR ajourføres i meget høj grad af kommunerne, og her har de fleste kommuner et selvstændigt folkeregister til denne opgave. Administrationen på det lokale niveau af BBR sker derimod oftest i teknisk forvaltning, som er en stor databruger.

Målrettede registre

Til forskel fra basisregistrene tjener de målrettede registre kun ét eller en velafgrænset gruppe af formål. Registrene føres af den myndighed, som skal bruge oplysningerne, og dataindholdet farves naturligt heraf. De målrettede registre modtager ofte oplysninger om populationen og om basisdata fra et basisregister, men fører selv andre data til.

4. Ændringer i de administrative data

Som omtalt kan det give anledning til store problemer for statistikken, når der sker ændringer i dataindholdet i de administrative registre som følge af lov- eller regelændringer. Dels kan det blive vanskeligt eller umuligt at bedømme den tidsmæssige udvikling i en bestemt størrelse, hvis der pludselig bruges andre definitioner i grundlaget, dels kan der opstå problemer med at fastslå, hvilke ændringer i dataværdier, der skal opfattes som udtryk for hændelser i virkeligheden, og hvilke, der kun afspejler de nye begreber eller definitioner.

Lovændringer

Ændringerne i de administrative registre kan skyldes ændringer i den gældende lovgivning på området. Konsekvenserne for statistikken afhænger af, hvilken type statistik der er tale om. Hvis der netop er tale om en *lovbelysende statistik,* som skal vise hvordan loven administreres i forhold til fx borgere, må statistikken naturligvis blot følge med og adoptere den nye lovgivnings begreber. Statistikken om kontanthjælp må således til enhver tid afspejle de regler, der gælder på området, og det kan så være vanskeligt at bedømme fx adfærdsmæssige ændringer.

Hvis det derimod drejer sig om en mere *generel statistik*, kan det ikke accepteres, at statistikkens begreber fra før og efter lovændringen ikke kan sammenlignes. Den generelle statistik forsøger at belyse nogle begreber, der ikke er fastlagt i lovgivningen, fx arbejdsløshed, som er defineret som begreb i en international konvention. Hvis statistikken bygger på udbetaling af forskellige ydelser, og reglerne for disse ændres, kan det være vanskeligt eller umuligt at kompensere for det i statistikken. Man må så i det mindste forsøge at opstille skøn over, hvor meget ændringerne har betydet, så en vis sammenkædning af tidsserierne kan blive mulig.

Et eksempel: ATP-statistikken

Et eksempel herpå er den konjunkturbelysende statistik over beskæftigelsesudviklingen baseret på arbejdsgivernes samlede ATP-indbetalinger. Denne statistik væsentligste formål er at beskrive korttidsudviklingen, og det er derfor afgørende at eliminere eller reducere betydningen af regelændringer på målingen af beskæftigelsen.

I 1977 udvidedes ATP-medlemskredsen bl. a. med værnepligtige og 16-17 årige lønmodtagere. For at kunne vurdere beskæftigelsesudviklingen fra kvartalerne i 1977 til 1978, blev det nødvendigt at korrigere tallene; grundlaget for denne korrektion var resultater fra Arbejdsstyrkeundersøgelserne. Et andet og mere subtilt eksempel på betydningen af regelændringer i denne statistik er den lovmæssige sammenkobling af ATP-indbetaling og kompensation for indbetalte arbejdsmarkedsbidrag i 1988. Der syntes i perioden indtil 1988 at være en vis tilbøjelighed hos nogle arbejdsgivere til at udskyde en større del af den samlede årlige indbetaling til 4. kvartalerne end der relativt skulle forventes; denne udskydelse havde ingen konsekvenser i form af "bøder", renter eller lignende. Med indførelse af kompensationen for indbetalte arbejdsmarkedsbidrag til arbejdsgiverne på basis af indbetalte ATP-bidrag blev det en fordel at indbetale rettidigt. Der skete derfor en ændring af indbetalingsmønstret, som gjorde det problematisk at vurdere beskæftigelsesudviklingen. På basis af en række undersøgelser og vurderinger blev det derfor nødvendigt at korrigere tallene på kvartalsbasis.

I nogle få tilfælde har det ikke været muligt at vurdere betydningen af ændringerne tilstrækkeligt præcist, og det har - bl. a. i forbindelse med omfattende regelændringer i 1993 og 1994 - været nødvendigt midlertidigt at indstille beskæftigelsesstatistikken på grundlag af ATP-indbetalingerne.

Andre regelændringer

Der kan naturligvis også være tale om ændringer i andre administrative regler end love, hvilket kan give anledning til ganske lignende problemer. Eksempelvis har det i ATP-statistikken i flere kvartaler været nødvendigt at tage forbehold og foretage korrektioner som følge af antallet af lønudbetalings*tidspunkter*; ATP-beløb bliver i de fleste tilfælde beregnet i forbindelse med lønudbetalinger, og antallet af kvartalsvise lønudbetalingstidspunkter kan for ugelønnede variere mellem 12 og 14. Et andet eksempel på andre ændringer end lovmæssige med betydning for ATP-indbetalingerne er overenskomstmæssige ændringer; fx betød ferieforlængelsen fra 4 til 5 uger en nedsættelse af det årlige ATP-bidrag for ugelønnede men ikke for månedslønnede. Da det ikke i forbindelse med ATP-indbetalingerne er

muligt at skelne mellem den ene og den anden type lønmodtagere, betød denne ændring alt andet lige en "nedgang" i beskæftigelsen. Lignende forhold har gjort sig gældende ved arbejdstidsnedsættelser kombineret med timegrænserne for ATP- indbetaling.

Systemomlægninger

Endelig kendes der eksempler på, at de administrative registre ændrer sig, uden at det skyldes ændringer i de grundlæggende regler. Typisk drejer det sig om, at man af tekniske grunde eller for at opnå en rationalisering omlægger systemerne. Det kan betyde problemer for statistikken, fx hvis man beslutter, at et bestemt datum ikke længere er nødvendigt.

Systemomlægninger kan også have en mere lokal karakter, som fx når man ændrer på en kommunegrænse eller på husnummereringen på en vej, hvorved et vist antal personer får en ny adresse, uden at der er tale om nogen flytning. Her vil det være meget nyttigt, hvis ændringerne i det mindste kan identificeres som korrektioner.

Konsekvenser for statistikken

Det er klart, at de omtalte ændringer er alvorlige for statistikken. De kan gennemføres med en større eller mindre grad af hensyntagen til statistikken, og det er vigtigt, at statistikerne er med i forberedelsen af ændringerne, så man i det mindste kan blive opmærksom på konsekvenserne af ændringer, inden de foretages.

5. Statistikernes mulighed for at påvirke data

Lige siden den registerbaserede statistiks opblomstring fra 1970 har det været anerkendt som et stort problem, at man ikke havde den kontrol over indholdet i grunddata, som man var vant til - eller troede man havde - i den skemabaserede statistik. Man kan ikke være sikker på, at registrene dækker lige præcis de enheder, der er relevante, eller at data er defineret i overensstemmelse med brugernes behov. Og ekstra problemer kommer som nævnt til, når registrenes indhold ændrer sig - det er endda også et problem, hvis de bliver mere pålidelige, for også i det tilfælde får man et databrud.

Det er naturligvis meget ønskeligt, at statistikerne kan have en vis indflydelse på dataindholdet, men samtidig må det erindres, at registrene føres med ganske bestemte administrative formål for øje, og registerejerne har det som deres opgave at løse disse opgaver mest effektivt. De kan derfor ikke lytte alt for meget efter ønsker, der "kun" er til statistiske formål.

I Danmark er der imidlertid en lovbestemmelse i Lov om Danmarks Statistik, § 1, som siger, at den, som planlægger at indrette et register, skal drøfte planerne med Danmarks Statistik, for at registret kan blive mest muligt egnet som grundlag for statistikken. Der er da også et ganske godt og tæt samarbejde mellem Danmarks Statistik og registerførerne på mange områder. Danmarks Statistik sidder som medlem af brugerkomitéer for både statslige og fælleskommunale systemer, og planer om registrenes udvikling diskuteres normalt også i de rådgivende udvalg, som Styrelsen har nedsat på mange statistikområder. Herigennem bør det i hvert fald kunne sikres, at man kender statistikkens ønsker, før man foretager ændringer, og at man ved opbygning

af nye systemer får drøftet, om der kan opnås statistiske fordele gennem mindre justeringer.

Beskedenhed en dyd

Man må imidlertid ikke forvente, at statistikerne kan stille - og få opfyldt - meget betydelige krav om ekstra data, andre definitioner eller lignende. Registerførerne må se nøje på deres egen effektivitet og ressourceindsats. Statistikken må derfor være meget beskeden og kun fremføre større ønsker, når der er en meget stor samfundsmæssig gevinst at hente ved en tilpasning af registrene. Der kendes kun nogle få eksempler på, at ekstra data indsamles gennem registrene udelukkende af hensyn til statistikken, jf. papir 4.3. om integreret dataindsamling.

6. Dannelse af statistikregistre

Inden de administrative grunddata kan benyttes til statistik, skal de underkastes en statistisk databehandling, som skal gøre dem velegnede til formålet. Som resultat af denne proces indordnes data i nogle *statistikregistre*, som er datasamlinger med identifikationsnøgler (altså registre), hvor populationen og dataindholdet i hvert register er afstemt, så det passer med ét statistikområdes behov, fx befolkningsstatistikkens.

Data i et statistikregister kan komme fra et eller flere administrative registre. Alle registerdata, som kan yde et væsentligt bidrag til at belyse statistikkens hjørne af virkeligheden, bør inddrages. Dannelsen af statistikregistret indebærer derfor ofte samkøring på enhedsniveau af grunddata. På basis af de forskellige data skal man forsøge at finde de bedste estimatorer.

Tidsreferencerne skal bearbejdes, så man får den bedst mulige statistiske model af enhedsbestanden og dens data på de ønskede tidspunkter eller i de ønskede perioder. De administrative registre forsøger normalt ikke at efterligne en sådan model, idet de ofte har til formål at give de nyeste tilgængelige data på ethvert område. Det betyder, at nogle data er helt nye, mens andre er noget ældre.

Kontrol af registeroplysningerne

En af de vigtigste og mest ressourcekrævende processer i registerdannelsen er en kontrol af de indgående grunddata og rettelse af evt. fejl. Statistikregistret skal ideelt set være "renset" og konsistent, så der ikke findes indbyrdes modstrid mellem forskellige data. Usandsynlige oplysninger skal undersøges. I denne proces er det ikke usædvanligt, at der afsløres systemfejl i de administrative registre eller i de udtræk, dataleverandøren har foretaget. I et supplerende indlæg er forskellige fejltyper og deres behandling omtalt.

Eksempel på fejl: Personnummeret

En af de alvorligste fejl, der kan forekomme, er fejl i identifikationen. Der findes som nævnt særlige sikkerhedsforanstaltninger, som modvirker, at der registreres fejlagtige identifikationer. Men det forekommer, at enhederne tildeles fejlagtige identifikationer, når de skal registreres i basisregistrene. Således har det vist sig, at der i få tilfælde registreres forkerte personnumre. Når dette kan forekomme, skyldes det, at personnummeret forbryder sig mod det ideale krav om, at identer ikke må indeholde information. Derfor er det nød-

vendigt at rette i personnummer, hvis det viser sig, at en person har fået registreret forkert fødselsdato eller køn.

I mange år anså man dette problem for marginalt, men i forbindelse med epidemiologiske forløbsanalyser viste det sig efterhånden, at det blev et problem, man var nødt til at gøre noget ved. I *Lynge & Thygesen (1990)* beskrives en analyse af cancer og erhverv, hvor man i en årrække følger en kohorte af personer, som var levende 9.11.70, og konstaterer senere diagnosticerede kræfttilfælde fra Cancerregisteret. Når personnummeret ændrer sig, mislykkes koblingen til cancerregistreringerne, og det vil synes som om personernes risiko for cancer er nul. Blandt 270.000 cancertilfælde i 1970-80 konstaterede og rettede man denne fejl i 286 tilfælde. Senere har man foretaget lignende rettelser i forbindelse med analyserne af dødelighed og erhverv, hvor fejlene ellers ville resultere i, at nogle personer syntes at leve evigt.

7. Afslutning

Brug af registre ikke uden problemer...

Som det er fremgået af det foregående, er der en del problemer forbundet med at basere statistikken på administrative registre. Det kan undertiden være vanskeligt at få de *relevante* oplysninger til statistikken, fordi statistikerne ikke har kontrollen over data, og i den forbindelse kan man risikere at få et *skævt verdensbillede*. Data kan være *fejlbehæftede*, hvilket der dog undertiden kan rådes bod på ved omhyggelig fejlsøgning og editering. *Tidsreferencerne* kan være mangelfulde, og korrektioner kan være vanskelige eller umulige at skelne fra oplysninger om virkelige hændelser. Endelig kan det giver store *kontinuitetsproblemer*, når registrene eller reglerne bag dem ændres.

...men også store fordele

På den anden side kan registerstatistikken give store fordele, hvilket de danske erfaringer også har vist. Det gælder ikke mindst de *økonomiske fordele*, som kan opnås, når oplysningerne allerede er indsamlet til andet formål, og samtidig kan man *nedbringe befolkningens indberetningsbyrde*. Statistikkens *hurtighed* kan i nogle tilfælde forbedres væsentligt. Kvalitetsmæssigt er registrenes *totale populationsdækning* er stor gevinst, og data er ofte *mere pålidelige* i registrene. Endelig skal det nævnes, at registerdata giver helt enestående muligheder for at gennemføre *forløbsundersøgelser*.

Referencer

Danmarks Statistik (1982). *Personstatistik på registergrundlag.* København 1982

Danmarks Statistik (1991). *Om Danmarks Statistiks målsætning.* Ikke publiceret.

Lynge, E. & Thygesen, L (1990): *Occupational cancer in Denmark.* Scandinavian Journal of Work, Environment & Health, vol. 16, supplement 2, 1990. Helsinki

Van Bastelaer, A. (1993). *Differences in the designs of the labour force survey in the European Community and some consequences.* Netherlands Central Bureau of Statistics, Heerlen. Unpublished

Familiestatistik på grundlag af status-udtræk fra CPR

Anna Qvist

Familiestatistikkens grundlag

Danmarks Statistik modtager pr. 1. januar hvert år et **statusudtræk fra CPR**. Det består af en record for hver person, der den 1. januar er tilmeldt CPR, og indholdet beskriver den aktuelle tilstand for personen. På grundlag heraf fremstilles statistik over befolkningen fordelt på køn, alder, bopælskommune, statsborgerskab og civilstand. Disse variable ligger i udtrækket for hver person klar til brug, eller de udledes af personnummeret.

Udtrækket viderebearbejdes i **husstands- og familiemodulet**, hvilket resulterer i en husstandsfil, en familiefil, og en personfil, alle med husstands- og familieoplysninger. På grundlag af disse filer udarbejdes 1. kontors årlige husstands- og familiestatistik. Men de dannede filer benyttes også af andre kontorer som baggrunds- og klassifikationsoplysninger i forbindelse med en lang række andre statistikker. Som eksempler kan nævnes statistik vedrørende sygesikring, benyttelse af daginstitutioner og sygehuse, samt dagpengestatistik.

Gammel og ny familiedefinition

Dannelsen af familieoplysninger på registergrundlag foregår efter nogle regler, der i første omgang blev udformet i forbindelse med den registerbaserede folke- og boligtælling i 1981. Familiedefinitionen, der dengang blev fastsat, blev senere taget i brug i en årlig familiestatistik. Der er offentliggjort statistik med denne definition for den 1. januar fra 1980 til 1991. Den 1. januar 1991 indførtes en ny familiedefinition, der siden er anvendt i statistikken for 1992 og 1993. Pr. 1. januar 1991 findes der således opgørelser efter begge familiedefinitioner. Dette tidspunkt er derfor velegnet til belysning af følgerne af ændringerne. Se oversigt 1. Et vigtigt formål med den nye definition var at få en bedre belysning af de ikke-gifte par.

I det følgende beskrives både den gamle og den nye definition. Selvom den gamle definition ikke benyttes mere, kan det være nyttigt at inddrage den som baggrund og sammenligning ved beskrivelsen af den nye. Det illustrerer endvidere de problemer, der må overvejes ved fastlæggelse af en familiedefinition på grundlag af et register, der ikke direkte indeholder familieafgrænsninger.

De benyttede data

Familiedannelsen foregår, nu som før, på grundlag af følgende variable fra CPR-udtrækket: For det første adressen, for det andet køn, alder og civilstand, og for det tredie henvisningsnumre til seneste ægtefælle eller registreret partner, samt til forældre.

Grundregler for familieafgrænsningen

Følgende grundregler er fælles for den gamle og den nye definition:

- En husstand omfatter alle personer på en adresse. Den består af en eller flere familier, og en familie består af en eller flere personer.

- Alle personer er enten børn, enlige eller personer i et par. Enlige og personer i par kan have børn, der indgår i familien.
- En familie består af én eller to generationer. I to-generationsfamilier består yngste generation af hjemmeboende børn, der skal være under en bestemt alder.
- Kun personer, der ikke selv har børn, aldrig har indgået ægteskab, og ikke indgår i et par, regnes med til forældrenes familie, hvis alderskravet er opfyldt.
- Søskende, der ikke bor hos deres forældre, regnes ikke til samme familie.
- Der opereres med en juridisk familie, A-familien, der ikke offentliggøres statistik for. I statistikken benyttes en familietype, der tilstræber at give en bedre beskrivelse af de faktiske familieforhold end A-familien. I den gamle og den nye definition hedder de henholdsvis B-familien og C-familien.

Pardannelse i den gamle familiedefinition

Princippet bag den gamle definitionen var, at personer, der blev regnet for et par, skulle have sikre familiemæssige relationer i form af enten ægteskab eller, hvis der ikke forelå ægteskab, i form af fællesbarn/børn. Der opereredes kun med disse to partyper, kaldet ægtepar og papirløse par. I den sidste gruppe skulle der være mindst ét hjemmeboende fællesbarn under 26 år, som var aldersgrænsen for børn, der regnedes med til forældrenes familier. Personer, der ikke indgik i en af disse to partyper blev betragtet som enlige, hvis de ikke var børn i en familie. Der var således tale om en minimumsopgørelse af de ikke-gifte par. Ikke-gifte par uden fællesbørn kom ikke med, men blev regnet for enlige i familiestatistikken, og disse par udgjorde flertallet blandt de ikke-gifte par, der boede sammen. Det fremgik af omnibus-undersøgelserne, der som supplement til familiestatistikken løbende målte antallet af papirløse samlivsforhold på grundlag af interviewing af forholdsvis små stikprøver. Disse opgørelser, der er offentliggjort for årene 1977 til 1984, indgik ikke som en integreret del af familiestatistikken. Resultaterne herfra kunne udvise ret store udsving, som var vanskelige at fortolke. Dette var grunden til, at offentliggørelsen af statistik over papirløse samlivsforhold på grundlag af omnibusundersøgelserne indstilledes, mens en mere tilfredsstillende løsning blev planlagt.

Ændringsbehov

Den mangelfulde statistiske dækning af ikke-gifte par, samt 26 årsaldersgrænsen for hjemmeboende børn, der i mange sammenhænge virkede uforståelig høj, medførte et stigende behov for en revision af familiedefinitionen. Beslutningen om den nye familiedefinition blev taget efter et udvalgsarbejde med deltagere fra brugerkontorer i Danmarks Statistik. Definitionen må på mange punkter betragtes som resultatet af afvejning af forskellige forhold mod hinanden. De to mest betydende ændringer var for det første sænkningen af den maksimale alder for hjemmeboende børn, og for det andet dannelsen af en ny partype, samboende par, af ikke-gifte par uden fællesbørn. Det blev betragtet som væsentligt at kunne opdele de ikke-gifte par i par med og par uden fællesbørn. Begrundelsen er, at dette vil muliggøre en skelnen mellem, på den ene side personer, om hvem vi ved, at der reelt er tale om samlivsforhold, hvorfor vi kalder dem **samlevende par**, og på den anden side personer, som vi kun kan gætte på udgør par, men om hvem vi i virkeligheden kun ved, at de bor på samme adresse, hvorfor de kaldes **samboende par.**

Aldersgrænsen for hjemmeboende børn

Aldersgrænsen for hjemmeboende børn er nu 18 år, svarende til myndighedsalderen og ophør af forældrenes forsørgelsespligt, hvilket har haft betydning for valget af denne grænse. På den ene side kan aldersgrænsen på 26 år virke uforståelig høj og uden sammenhæng med andre forhold, og på den anden side kan man synes, at 18 år er en meget lav grænse, når man tænker på, at de fleste i praksis indgår i deres forældres familier i et par år efter denne alder. Men da denne aldersgrænse endvidere benyttes i mange internationale sammenhænge, blev den valgt. Omkring 208.000 unge på 18 - 25 år ændrede ved definitionsændringen familiestatus fra børn til enlige, hvilket bevirkede en øgning af familieantallet på denne størrelse. Dog besluttedes det at definere en familietype, kaldet D-familien, for hvilken der ingen aldersgrænse var for børn hørende med til forældrenes familier (ud over kravet om at de skal være ugifte og ikke have børn). Der fremstilles ingen statistik på denne familietype. Det er blot indlagt som et vist beredskab i husstands- og familiemodulet.

Samboende par

Den anden store ændring var dannelse af en ny partype, samboende par, der omfatter par bestående af en mand og en kvinde, der ikke er gift med hinanden, og som ikke har børn sammen. I praksis må definitionen indrettes efter registrets muligheder, hvilket har medført det definitoriske krav, at der ikke måtte være andre voksne på adressen. Endvidere bør der ikke indgå søskendepar i denne gruppe. Ældre søskendepar vil det dog ikke være muligt at få sorteret væk, da de mangler forældrehenvisningsnumre. Se nedenfor vedrørende datakvalitet. Som parallel til det forhold, at unge under 18 år kan få børn eller gifte sig og på disse måder blive regnet for voksne i statistikken, tillader definitionen af samboende par, at unge på ned til 16 år danner samboende par, hvis de øvrige betingelser er opfyldt. Det samlede antal af mindreårige voksne er under 1000 personer.

For samboende par kræves det endvidere, at aldersforskellen er mindre end 15 år. Dette er begrundet med, at vi herved udelukker par bestående af en person og dennes far eller mor. Dette kan ikke sikres gennem forældre-henvisningsnumrene, da disse så godt som ikke er til stede for personer, der er født i begyndelsen af 1950-erne eller tidligere. Samtidig antages det, at i en stor del af de tilfælde, hvor der bor en enlig sammen med en logerende, vil aldersforskellen være større end 15 år, så disse heller ikke indgår som samboende par. Femtenårsgrænsen har været meget diskuteret. Der er selvfølgelig tilfælde, hvor et reelt samboende par har en større aldersforskel. Men betydningen af dette er forholdsvis ringe. Det kan nævnes, at under 2 pct. af de øvrige to-kønnede par har en forskel på 15 år eller mere. (For de registrerede partnerskaber er andelen 13 pct.) Den 1. januar 1991 var der 370.000 personer, der efter den gamle definition var enlige, men efter den nye blev samboende, hvilket reducerede familieantallet med det halve af denne størrelse.

Registrerede partnerskaber

Ved omlægningen af familiestatistikken benyttede man sig af lejligheden til at indføre en fjerde partype, de registrerede partnerskaber. Loven om registrerede partnerskaber trådte i kraft den 1. oktober 1989. Denne partype forekommer i lighed med de øvrige både med og uden børn, der i disse tilfælde altid er særbørn. Udskillelsen af de registrerede partnerskaber volder ingen

problemer, idet der i CPR-registret er indført tre nye civilstandskoder. Kodeværdien P bruges for personer, der er i registreret partnerskab, O for personer, der har fået opløst partnerskabet, og L anvendes for personer, der har overlevet deres partner. De øvrige koder er U for ugift, G for gift eller separeret, F for fraskilt og E for enkestand. Talmæssigt betyder de registrerede partnerskaber ikke meget, men da et registreret partnerskab juridisk er stillet som et ægteskab, anses det for vigtigt at få dem med som familietype. Derimod er det forbundet med alt for store problemer og af ringe interesse at forsøge at få tal på ikke-registrerede parforhold mellem personer af samme køn.

Samlevende par

Den gamle familietype, papirløse par, svarer ret nøje til den nye gruppe af samlevende par. Som papirløse par regnedes kun par, der havde fælles hjemmeboende børn under 26 år på statustidspunktet. Flyttede eller døde børnene, eller fyldte de 26 år, var disse personer ved en senere opgørelse ikke længere papirløse par. Dette uheldige forhold er der søgt at råde bod på ved, at man i det nye husstands- og familiemodul opsamler de samlevende par fra år til år, således at et par, der på ét tidspunkt har været klassificeret som samlevende par, hvilket de bliver ved tilstedeværelsen af fællesbørn på adressen uanset børnenes alder, vedbliver med at være dette ved næste opgørelse, såfremt de stadig bor sammen, uanset om børnene fortsat er til stede. Derved opstår den umiddelbart lidt uforståelige gruppe af samlevende uden børn, dvs uden hjemmeboende børn under 18 år. Denne gruppe kan forventes at stige forholdsvis kraftigt i en årrække på grund af opsamligen, der kun går tilbage til 1. januar 1990.

Sammenhængen mellem civilstand og familietype

Civilstandens rolle som kriterium ved afgørelsen af en persons familiestatus begrænser sig til tre typer af tilfælde. Kun ugifte personer kan regnes som børn, og civilstanden gift eller registreret partner er en betingelse for at have familiestatus som henholdsvis gift eller registreret partner. Alle øvrige kombinationer af civilstand og familiestatus kan i teorien forekomme, hvilket de også gør bortset fra nogle få vedrørende forhenværende registrerede partnere, som talmæssigt er få. Fx. består gruppen af enlige af personer fra alle syv civilstandsgrupper (tabel 1). Kun 95 pct. af personerne med civilstanden gift, har familiestatus som gift. Man skal i denne forbindelse huske, at personer med civilstanden gift også omfatter separerede. Den tilsvarende procent for registrerede partnere er på kun 89 pct. Enlige blandt personer, der er fraskilte eller i enkestand, udgør henholdsvis 73 pct. og 94 pct. For de tilsvarende grupper af forhenværende registrerede partnere er procenten af enlige henholdsvis 98 pct. og 97 pct. Blandt de ugifte er kun 68 pct. enlige i henhold til familiedefinitionen.

Tabel 1 **Voksne personer fordelt efter civilstand og familiestatus pr. 1. januar 1993**

Familiestatus	Civilstand							Voksne i alt
	Ugift	Gift	Fraskilt	Enke-stand	Regi-streret partner	Opløst partner-skab	Længst-levende partner	
I alt	**1252271**	**2126391**	**352722**	**362079**	**2228**	**105**	**108**	**4095904**
Enlig	848126	90851	259247	342074	246	103	105	1540752
Gift	.	2026838	.	.	.	.	.	2026838
Registreret partner	.	.	.	.	1972	.	.	1972
Samlevende par	150683	953	22750	762	.	.	.	175148
Samboende par	253462	7749	70725	19243	10	2	3	351194

Flere regnes nu som par

Mens der med den gamle familiedefinition kun dannedes par i tilfælde, hvor der var stor vished for rimeligheden i det, og det samlede antal parfamilier derfor blev sat lavt, bliver der med den nye familiedefinition dannet parfamilier af nogle personer, der reelt burde betragtes som enlige, ligesom der er personer, der regnes for enlige, mens de i virkeligheder udgør et par. Det sidste sker antagelig mest i tilfælde, hvor der bor andre voksne på adressen. Men alt i alt er der grund til at tro, at den nye familiedefinition giver en bedre beskrivelse af befolkningens pardannelser. I hvert fald har man opnået den fordel, at ændringer i antallet af forskellige familietyper skyldes reelle ændringer i befolkningen og ikke stikprøveusikkerhed eller valget af opregningsmetoder. De antalsmæssige ændringer, der er sket siden indførelsen af den nye familiedefinition pr. 1. januar 1991, kan ses af oversigt 2.

Høj datakvalitet for de fleste data...

De grundlæggende data fra CPR-registret, som familiestatistikken bygger på, har for de flestes vedkommende meget høj kvalitet. De omfatter adresse, køn, alder og civilstand, samt ægtefællehenvisningsnummer.

Adressen afgrænser husstanden og dermed også familien. Adresseoplysningerne i CPR er ret præcise. Da de for langt den overvejende del af befolkningen også er udtryk for, hvor de virkelig bor, er adresseoplysningerne velegnede som grundlæggende kriterium ved afgrænsningen af familier. I de antagelig meget få tilfælde, hvor en person ikke er tilmeldt CPR på den korrekte adresse, bliver familietyperne forkerte på både CPR-adressen og på den virkelige adresse. Det er umuligt at komme bag om denne fejlkilde.

Køn, alder og civilstand beskriver den enkelte person og indgår som kriterier ved fastlæggelsen af familieforhold. Oplysningerne fra CPR-nummeret (køn og alder) kan anses for fuldstændig korrekte, og civilstandsoplysningerne kommer tæt på denne tilstand.

Henvisningsnumre til seneste ægtefælle eller registrerede partner benyttes til afgørelsen af, om et ægtepar (eller to registrerede partnere) bor sam-

men, og derfor skal regnes for en par, eller om dette ikke er tilfældet. Denne oplysning er bl. a. på grund af betydningen for skatteligningen af meget høj kvalitet.

...men mange forældrehenvisningsnumre mangler

Henvisningsnumre til forældre benyttes til identifikation af familierelationer. Disse numre henviser til de juridiske forældre, hvilket vil sige, at vi ikke kan skelne biologiske forældre fra adoptivforældre. Desværre er hyppigheden af disse numres tilstedeværelse faldende med stigende alder, hvilket har betydning for muligheden for at udskille søskendepar fra samboende par. Personer under 18 år har meget fin dækning med hensyn til henvisninger til forældre, kun 0,15 pct af dem mangler begge henvisningsnumre. Opdelingen af børn i hjemmeboende og ikke-hjemmeboende foregår derfor på ret sikkert grundlag. Personer født i 1960 mangler for 7 pct.'s vedkommende begge forældrehenvisninger, og for 1950-årgangen drejer det sig om 96 pct. (tabel 2). Disse henvisningsnumre slettes ikke i CPR-registret , når de én gang er indført. Derfor forbedres befolkningens dækning i denne henseende gradvist år for år. Dette resulterer i, at antallet af voksne, der bor hos deres forældre, udviser årlig stigning, for tiden mest for personer, der er i 40'erne (tabel 3). De forbedrede forældrehenvisningsnumre vil i fremtiden give mulighed for bl.a. opgørelse af børnefamilier, der bor sammen med bedsteforældregenerationen.

Tabel 2 **Henvisningsnumre til forældre pr. 1. januar 1993**

Fødsels-år	Alder (år)	Procentandel med henvisningsnummer til				Total
		begge forældre	kun mor	kun far	ingen	
1990	2	98,0	1,9	0,0	0,1	100,0
1985	7	98,3	1,5	0,1	0,0	100,0
1980	12	98,4	1,4	0,1	0,1	100,0
1975	17	97,8	1,2	0,2	0,7	100,0
1970	22	95,1	1,0	0,2	3,6	100,0
1965	27	90,5	3,7	0,6	5,2	100,0
1960	32	86,6	5,3	1,1	7,1	100,0
1955	37	60,7	4,7	1,3	33,2	100,0
1950	42	4,0	0,3	0,1	95,5	100,0
1945	47	1,8	0,2	0,1	97,9	100,0
1940	52	1,0	0,2	0,0	98,8	100,0
1935	57	0,7	0,2	0,1	99,1	100,0
1930	62	0,5	0,2	0,1	99,3	100,0
1925	67	0,3	0,2	0,1	99,5	100,0
1920	72	0,1	0,1	0,0	99,8	100,0
1915	77	0,0	0,1	0,0	99,9	100,0
1910	82	0,0	0,0	0,0	100,0	100,0

Tabel 3 **Personer med forældrehenvisningsnumre til personer på samme adresse pr. 1. januar 1993**

Alder	1991	1992	1993
20 - 29 år	123 524	124 119	126 654
30 - 39 år	17 973	19 170	19 864
40 - 49 år	852	1 005	1 827
50 - 59 år	197	196	206
60 - 69 år	56	58	56
70 - 79 år	6	5	4

Oversigt 1

Forskydninger i familietyper som følge af ny familiedefinition pr. 1. januar 1991

(benævnelserne svarer til den nye familiedefinition)

	Gammel definition	Ny definition	Ændring i antal
Personer i alt	5 146 469	5 146 469	0
Familier i alt	2 762 853	2 800 349	37 496
Husstande i alt	2 287 592	2 287 592	0
Familietyper:			
Enlige kvinder uden børn	767 054	705 238	- 61 816
Enlige kvinder med børn	141 720	101 872	- 39 848
Enlige mænd uden børn	726 154	690 726	- 35 428
Enlige mænd med børn	27 393	16 129	- 11 264
Ægtepar uden børn[2]	501 456	580 733	79 277
Ægtepar med børn[2]	523 280	438 654	- 84 626
Registrerede partnerskaber uden børn[4]	-	636	-
Registrerede partnerskaber med børn[4]	-	27	-
Samlevende par uden børn[1, 5]	-	1 270	-
Samlevende par med børn[3]	75 796	77 132	1 336
Samboende par uden børn[1]	-	151 481	-
Samboende par med børn[1]	-	21 460	-
Ikke-hjemmeboende børn under 18 år[4]	-	14 991	-
Enlige i alt	1 662 321	1 513 965	- 148 356
heraf:			
Enlige mænd i alt	753 547	706 855	- 46 692
Enlige kvinder i alt	908 774	807 110	- 101 664
Enlige uden børn i alt	1 493 208	1 395 964	- 97 244
Enlige med børn i alt	169 113	118 001	- 51 112
Parfamilier i alt	1 100 532	1 271 393	170 861
heraf:			
Ægtepar	1 024 736	1 019 387	- 5 349
Registrerede partnerskaber[4]	-	663	
Samlevende par	75 796	78 402	2 606
Samboende par[1]	-	172 941	
Samlevende og samboende i alt	75 796	251 343	175 547
Parfamilier uden børn i alt	501 456	734 120	232 664
Parfamilier med børn i alt	599 076	537 273	- 61 803
Familier uden børn i alt	1 994 664	2 130 084	135 420
Familier med børn i alt	768 189	655 274	- 112 915
Hjemmeboende børn	1 283 083	1 074 727	- 208 356

[1] Ny gruppe, personeme kommer hovedsagelig fra gruppen af enlige.

[2] Det samlede antal ægtepar mindskedes med 5.349 ved definitionsændringen. Dette skyldes fraskilte par, der igen bor sammen. De blev før regnet for ægtepar, men er nu samlevende, samboende eller enlige, alt efter hvilke betingelser de opfylder.

[3] Udgør den familietype, der kommer nærmest den gamle familiedefinitions papirløse par. Øgningen skyldes mest fraskilte ægtepar, der bor sammen.

[4] Ny gruppe, blev tidligere medregnet under enlige.

[5] Samlevende par uden børn har haft hjemmeboende fællesbørn (uden aldersgrænse) den 1. januar 1990 eller ved et senere årsskifte.

Oversigt 2

Udviklingen i familietyperne siden indførelsen af den nye familiedefinition

	1.1.1991	1.1.1992	1.1.1993	%-ændring 1991 - 93
Personer i alt	5 146 469	5 162 126	5 180 614	0,7
Familier i alt	2 800 349	2 815 723	2 832 553	1,1
Husstande i alt	2 287 592	2 309 177	2 324 865	1,6
Familier pr. husstand	1,2	1,2	1,2	-0,5
Personer pr. familie	1,8	1,8	1,8	-0,5
Personer pr. husstand	2,2	2,2	2,2	-1,0
Familietyper:				
Enlige kvinder uden børn	705 238	710 303	715 950	1,5
Enlige kvinder med børn	101 872	102 327	103 695	1,8
Enlige mænd uden børn	690 726	697 312	705 581	2,2
Enlige mænd med børn	16 129	15 745	15 526	-3,7
Ægtepar uden børn	580 733	586 218	591 045	1,8
Ægtepar med børn	438 654	430 216	422 374	-3,7
Registrerede partnerskaber uden børn	636	810	940	47,8
Registrerede partnerskaber med børn	27	41	46	70,4
Samlevende par uden børn[1]	1 270	1 789	2 300	81,1
Samlevende par med børn	77 132	81 598	85 274	10,6
Samboende par uden børn	151481	153392	154574	2,0
Samboende par med børn	21 460	21 422	21 023	-2,0
Ikke-hjemmeboende børn under 18 år	14 991	14 550	14 225	-5,1
Enlige i alt	1 513 965	1 525 687	1 540 752	1,8
heraf:				
Enlige mænd i alt	706 855	713 057	721 107	2,0
Enlige kvinder i alt	807 110	812 630	819 645	1,6
Enlige uden børn i alt	1 395 964	1 407 615	1 421 531	1,8
Enlige med børn i alt	118 001	118 072	119 221	1,0
Parfamilier i alt	1 271 393	1 275 486	1 277 576	0,5
heraf:				
Ægtepar	1 019 387	1 016 434	1 013 419	-0,6
Registrerede partnerskaber	663	851	986	48,7
Samlevende par	78 402	83 387	87 574	11,7
Samboende par	172 941	174 814	175 597	1,5
Samlevende og samboende i alt	251 343	258 201	263 171	4,7
Parfamilier uden børn i alt	734120	742 209	748 859	2,0
Parfamilier med børn i alt	537 273	533 277	528 717	-1,6
Familier uden børn i alt	2 130 084	2 149 824	2 170 390	1,9
Familier med børn i alt	655 274	651 349	647 938	-1,1
Hjemmeboende børn	1 074 727	1 070 917	1 070 485	-0,4

[1] Samlevende par uden børn har haft hjemmeboende fællesbørn (uden aldersgrænse) den 1. januar 1990 eller ved et senere årsskifte.

Administrative data som statistikdata

Søren Hostrup-Pedersen

Fejlsøgning og gode administrative basisregistre giver god statistik

Administrative oplysninger må fejlsøges

I en række statistiksystemer anvendes registeroplysningerne ikke i alle tilfælde direkte som statistikgrundlag. Der foretages først en fejlsøgning af materialet, før statistikproduktionen kan gå igang. Fejlsøgningen indrettes således, at enkeltoplysninger i de grundliggende administrative data undersøges for eventuelle fejl, og disse rettes, før materialet anvendes som statistikgrundlag.

De formelle krav er ikke opfyldt

Fejlene kan bestå i, at oplysningerne ikke opfylder de formelle krav for den pågældende oplysningstype. I sådanne tilfælde, som ofte optræder gruppevis, vil der skulle træffes en afgørelse af, om den pågældende værdi skal ændres, så oplysningen falder inden for de tilladte intervaller, eller træffes en afgørelse om, at de pågældende oplysninger ikke kan indgå i statistikgrundlaget, da en korrektion ikke vil kunne foretages på en meningsfyldt måde.

Enkeltoplysninger rettes

Fejl kan også bestå i, at det konstateres, ved direkte opslag i supplerende kilderegistre, at en given individoplysning i et administrativt register ikke er korrekt. I disse tilfælde rettes enkeltoplysningen, før den egentlige tabellering finder sted.

Ingen henvendelse til den administrative myndighed om enkelttilfælde

Da der her er tale om udnyttelse af administrative oplysninger, der ikke er indhentet med det direkte formål at foretage statistiske opgørelser, må der efter registerlovens bestemmelser i almindelighed ikke rettes henvendelse til den administrative myndighed om den pågældende fejl, da dette vil kunne få indflydelse på den administrative afgørelse for enkeltpersonen. Den statistiske opgørelse vil derfor ikke afspejle alle detaljer i det administrative materiale, som iøvrigt har dannet grundlag for opgørelsen.

- men samarbejde om rettelse af fejltyper i basisregistrene.

Modsat bør man imidlertid påpege, at det er afgørende at eventuelle grundlæggende fejltyper i de administrative grundregistre, hvorfra der udtrækkes data til statistiske formål, bliver rettet. Dette kan kun ske ved en indgående kontakt med de registeransvarlige myndigheder. Der er mange eksempler på, at statistikerne har opdaget fejltyper, som senere er blevet rettet i basisregistrene, til glæde både for basisregisteret og den endelige statistik.

Opslag i administrative basisregistre

I statistikarbejdet indgår administrative registre ofte som grundlag for ajourføring af oplysninger i statistikgrundlaget. Eksempler herpå er CPR-registret, hvor Danmarks Statistik nu har adgang til versioner, der hele tiden er ajourførte. CPR-registret er i mange detaljer omtalt andetsteds og vil derfor ikke blive omtalt yderligere her. Et andet eksempel er Erhvervsregisteret, der er et

administrativt register over samtlige erhvervsenheder. Erhvervsregisteret danner grundlaget for erhvervsstatistiske opgørelser og giver erhvervsoplysninger i personstatistikken, hvor sådanne oplysninger er relevante. Enkeltopslag i Erhvervsregisteret benyttes løbende i arbejdet, og en oplysning som erhvervsenhedens branche bliver meget ofte ændret, når der modtages information direkte fra virksomhederne eller på anden måde. I Erhvervsregisteret indgår også arbejdsstederne, som ajourføres dels gennem løbende statistikker og dels gennem arbejdspladsprojektet.

Arbejdspladsprojektet er indgående omtalt i gennemgangen af den integrerede dataindsamling, hvor administrative og statistiske oplysninger indsamles i sammenhæng. Pointen er her, at de administrative og statistiske oplysninger er sammenhængende, og en høj kvalitet nås kun gennem et intensivt fejlsøgningsarbejde.

Gode ajourførte basisregistre giver god statistik

Udnyttelsen af administrative oplysninger til statistikformål kræver intensiv kontrol og fejlsøgning af oplysningerne. Til dette formål er løbende ajourførte , administrative basisregistre af stor værdi.

Anvendelse af flere kilder

Gunvor Højberg

I den samfundsbeskrivende statistik anvendes i stor udstrækning administrative variable direkte, fordi de er en del af den sociale virkelighed og betydningen af dem er velkendt. Inddeling efter køn, alder og civilstand behøver således ingen nærmere forklaring. Administrative begreber må dog som oftest beskrives, fordi man ikke kan gå ud fra, at brugerne af statistikken er fortrolig med den administrative praksis. I visse situationer kan man være nødt til at omdefinere det administrative begreb, fx. hvis en opgørelse skal bruges til international sammenligning og den anvendte definition er afvigende fra dansk praksis.

Imidlertid anvendes i den samfundsbeskrivende statistik også inddelinger, som ikke kendes i administrationen. Det drejer sig for det første om inddelinger, hvor enheden i statistikken i virkeligheden står i en flertydig situation, fx at en person på samme tid er pensionist og lønmodtager, og hvor man i de statistiske opgørelser ønsker et mere forenklet billede af virkeligheden, altså at personen betragtes enten som pensionist eller som lønmodtager. For det andet drejer det sig om inddelinger, hvor klassificeringen ikke i alle tilfælde kan ske på grundlag af en enkelt oplysning, fordi klassifikationen inddrager mange aspekter i bedømmelsen. Et eksempel herpå er fagklassifikationen, som af administrationen kun anvendes af arbejdsformidlingen, men kun for arbejdssøgende og af arbejdsskadestyrelsen, men kun for tilskadekomne. Den øvrige administration har ingen information, der tager sigte på, at angive den enkeltes arbejdsfunktion(er).
I den første situation benytter man sig i den registerbaserede statistik af, at de forskellige nødvendige oplysninger findes i forskellige registre. Disse uddrages og sammenholdes og en konklusion m.h.t. klassifikationen træffes gennem et maskinelt opererbart regelsæt. I den anden situation anvender man sig i det valgte eksempel af en erstatningsoplysning, nemlig stillingsbetegnelsen. I en del tilfælde er oplysningen ikke tilstrækkelig til at placere personen efter den anvendte klassifikation og man benytter derfor andre oplysninger, der muliggør en sandsynlig placering.

Udover en beskrivelse af fremgangsmåden i registerbaserede tællinger, beskrives i det følgende også fremgangsmåden i en spørgebaseret tælling, nemlig i Folke- og boligtællingen 1970, hvor man er ude i den samme problemstilling: at man søger en statistisk standard, som ikke er umiddelbart kendt.

Beskæftigelsesstatus/Arbejdsstilling i folke- og boligtællingen 1970.
I den skemabaserede folketælling i 1970 stillede man med henblik på at kunne klassificere personerne efter begreber som beskæftigelsesstatus eller arbejdsstilling to spørgsmål:
a) Beskæftigelsesforhold på tællingsdagen
b) Ansættelsesforhold

Til det første spørgsmål er givet eksempler på svar dels for 'Personer, der normalt er erhvervsmæssigt beskæftiget' og dels for 'Personer som er uden erhvervsmæssig beskæftigelse'.
Til det andet spørgsmål er anført de mulige svar så som 'Selvstændig med ansatte', 'Selvstændig uden ansatte', 'Funktionær', 'Arbejder', 'Medhjælper', 'Elev' o.s.v.
Bortset fra at det af eksemplerne fremgår, at arbejdsløse, personer på barselsorlov, på ferie, på kursus m.m. hører til de erhvervsmæssigt beskæftigede er der kun givet den yderligere vejledning, at deltidsansatte, medhjælpende hustruer og studerende med erhvervsarbejde betragtes som beskæftigede.
Det er således i en skemabaseret folketælling som den danske i 1970, hvor skemaerne i almindelighed udfyldes uden hjælp fra tællingskommissærer, de adspurgte, der træffer afgørelsen m. h t. klassificeringen i grænsetilfælde. I kodeprocessen er der dog mulighed for at foretage rettelser, fx. er skolesøgende børn med erhvervsmæssigt arbejde placeret som personer uden for arbejdsstyrken. Der er ikke foretaget en validitetsundersøgelse af besvarelserne og jeg må derfor nøjes med at antage, at den mest betydningsfulde svaghed ved den skemabaserede tælling er manglende besvarelser. Omfanget af manglende besvarelser kendes heller ikke, idet man bl. a. har anvendt folkeregisterets stillingsbetegnelse, hvor kommunen slet ikke har haft kontakt med pågældende i forbindelse med tællingen.

Beskæftigelsesstatus i Indkomststatistikregistret

Med beskæftigelsesstatus ønsker man at skelne mellem:

Personer i erhverv:
- Selvstændig næringsdrivende
- Medhjælpende ægtefæller
- Lønmodtagere

Personer ude af erhverv:
- Pensionister
- Andre ude af erhverv

Beskæftigelsesstatus dannes i Indkomststatistikregistret (og dermed i Arbejdsklassifikationsmodulet) ved at sammenkæde oplysninger fra Erhvervsregisteret, Det centrale Skatteyder-register og skattevæsenets Oplysningsseddelregister, idet der sigtes mod at klassificere efter, hvad der har været gældende for året som helhed.

En person betragtes i almindelighed som værende i erhverv, såfremt han/hun har været til rådighed for arbejdsmarkedet i et vist minimum af tid. En sådan oplysning findes ikke i de administrative registre for selvstændige og medhjælpende ægtefæller og kun på sekundær måde for lønmodtagere i form af oplysning om ATP-beløbets størrelse. I stedet for kræves for en lønmodtager,

at lønindkomsten overstiger et grænsebeløb, der for 1991 udgør 36903 kr. og for selvstændige, at de er indehavere af en momsbetalende virksomhed eller er arbejdsgiver eller har overskud af egen virksomhed (fx. som forfatter, læge m.m.).
Personer uden for arbejdsstyrken klassificeres som pensionist, såfremt han/hun modtager en offentlig pension, hvilket fremgår af Oplysningsseddel-registeret.
Da en person på samme tid kan tilhøre flere af de nævnte kategorier foretages den endelige placering efter forholdet mellem forskellige indkomstarter og størrelsen af forskellige indkomstbeløb.

I regelsættet, der klassificerer de skattepligtige m.h.t. beskæftigelsesstatus skelnes mellem

Eneindehavere
Medindehavere
Regnskabspligtige
Personer i øvrigt med egen virksomhed

Generelt kan siges, at klassifikation som selvstændig for eneindehavere kræver, at overskuddet af selvstændig virksomhed er større end lønindkomsten eller større end pensionsindkomsten. For andre selvstændige kræves, at overskuddet er større end A-indkomsten samtidig med at A-indkomsten ikke overstiger et givet minimum (der arbejdes med beløbsstørrelser på 18.000 kr., 20.000 kr. og 30.000 kr.).

Som medhjælpende ægtefælle klassificeres personer, hvis indkomst som medhjælpende ægtefælle er større end lønindkomsten. Det fremgår af Skatteyderregisteret, hvilket beløb, der er overført fra en selvstændig til ægtefællen som medhjælpende i virksomheden. Det maksimale beløb, der kan overføres, fastsættes af skattemyndighederne og beløbsstørrelsen reguleres årligt.

Gruppen 'Lønmodtagere' omfatter personer, der ikke er klassificeret som selvstændig eller medhjælpende ægtefælle, og som har en lønindkomst - incl. arbejdsløshedsunderstøttelse, sygedagpenge og vederlag - der overstiger det før omtalte grænsebeløb og som ikke har modtaget et endnu større beløb i offentlig pension.

Gruppen 'Pensionister' omfatter personer, der modtager en offentlig pension og som ikke kan klassificeres som person i erhverv. Efterløn betragtes ikke som pensionsindtægt og efterlønsmodtagere klassificeres således som 'Andre ude af erhverv'.

De forskellige registre, der anvendes, opdateres løbende og er aldrig endeligt opdaterede på det tidspunkt, hvor oplysningerne tages ud til statistikregistrene. De ændringer, der kommer efter at statistikuddragene er foretaget, drejer sig imidlertid i almindelighed ikke om ændringer i udbetalte lønbeløb, men om ændringer i skattemæssige fradrag, størrelsen af overskuddet af egen virksomhed og andet, der kræver indsigt i skattelovgivningen. Det kan ændre klassifikationen af den enkelte, men kun i grænsetilfælde, hvor den pågæl-

dende både er selvstændig og pensionist eller selvstændig og lønmodtager o.s.v., og det antages, at disse sene ændringer ikke er af betydning for de statistiske resultater.

Større indflydelse har muligvis ændringer i det regelsæt, der fører til klassificeringen -eller rettere sagt manglen på ændringer. Det må have betydning for grænsedragningen mellem grupperne, at visse beløbsstørrelser ændres fra år til år medens andre ligger fast. Reglen, at en person klassificeres som selvstændig, hvis han/hun er regnskabspligtig som selvstændig og har en A-indkomst, der er mindre end 30.000 kr., har været gældende siden 1980 med et uændret grænsebeløb, hvorimod grænsebeløbet for lønindkomst, der er afgørende for om en lønmodtager klassificeres som sådan reguleres årligt, således at det næsten er fordoblet i løbet af et tiår (1980: 20.000 kr., 1986: 30817 kr., 1991: 36.903 kr.). Regelsættet har således bevirket, at det gennem årene er 'blevet lettere', at blive betragtet som selvstændig.
Det skal også bemærkes, at ved at anvende det samme grænsebeløb for alle lønmodtagere klassificeres lønmodtagere med samme timetal forskelligt alt efter om de er højt - eller lavtlønnet, i arbejde eller får arbejdsløshedsunderstøttelse fra en A-kasse. .

Arbejdsstilling i den registerbaserede arbejdsstyrketælling.

Med arbejdsstilling i den registerbaserede arbejdsstyrkestatistik (RAS) ønsker man at belyse tilknytningen til arbejdsmarkedet på et givet tidspunkt -i praksis ultimo november - gennem en inddeling i:

Personer i arbejdsstyrken
- Selvstændige
- Medhjælpende ægtefælle
- Beskæftiget lønmodtager
- Arbejdsløs

Personer uden for arbejdsstyrken
- Uddannelsessøgende og børn
- Pensionister
- Efterlønsmodtagere
- Øvrige uden for arbejdsstyrken

Man anvender har de samme typer oplysninger fra de administrative registre som i Arbejdsklassifikationsmodulet med tilføjelse af to meget centrale oplysninger, nemlig oplysningen i Oplysningsseddelregistret om hvorvidt et arbejdsforhold har eksisteret den sidste hverdag i november måned og oplysningen i Arbejdsløshedsstatistikregistret om hvorvidt en person var arbejdsledig og arbejdssøgende den sidste uge i november.
I processen dannes først bruttobestande af forskellige statusgrupper som

A. Personer, der i Arbejdsdirektoratets centrale register for arbejdsmarkedsstatistik (CRAM) var registreret som fuldt arbejdsledige i den sidste uge i november - omfattende både forsikrede og ikke forsikrede arbejdssøgende

B. Personer, der modtog efterløn ved udgangen af året og ikke var berørt af ledighed i den sidste uge af november
C. Personer, der var lønmodtager ultimo november, påvist gennem oplysning om

a. Medlemskab af en arbejdsløshedskasse for lønmodtagere ultimo året
b. Ansættelse som lønmodtager ultimo november iflg Oplysningsseddelregistret
c. ATP-bidragets størrelse

For at indgå i gruppen stilles krav om, at personen har en lønindkomst i årets løb på et givet mindstebeløb svarende til 80 timers beskæftigelse

Lønmodtagere opdeles i heltids- og deltidsansatte på grundlag af oplysning om hvorvidt de er heltidsforsikrede i en arbejdsløshedskasse eller på grundlag af oplysning om ATP-beløbets størrelse i forhold til antal ansættelsesdage eller evt. på grundlag af lønstørrelsen i forhold til antal ansættelsesdage.

D. Selvstændige, sammensat af:

a. Arbejdsgivere (Har mindst 1 ansat af lønmodtagerne)
b. Momsbetalere (Aktive i årets løb og ikke afmeldt ultimo november
c. Forsikrede i en arbejdsløshedskasse for selvstændige
d. Personer med overskud af egen virksomhed iflg. AKM

E Medhjælpende ægtefæller iflg. AKM

For personer, der indgår i flere bruttobestande prioriteres derefter i rækkefølgen:

1. Arbejdsløse
2. Efterlønsmodtagere
3. Arbejdsgivere
4. Heltidsbeskæftigede lønmodtagere
5. Momsbetalere
6. Arbejdsløshedsforsikrede selvstændige
7. Øvrige lønmodtagere
8. Øvrige selvstændige
9. Medhjælpende ægtefæller.

Prioriteringen er valgt ud fra en vurdering af oplysningernes pålidelighed.
I kapitel 4.5 sammenlignes resultatet af inddelingen med en tilsvarende inddeling i en spørgebaseret undersøgelse TASU).

Fagvariablen i Arbejdsklassifikationsmodulet

Inddeling efter fag er en statistisk standard, som tager sigte på at gruppere personerne efter det arbejde, de udfører som aktive på arbejdsmarkedet.

Målet er at henføre personer, der i en given periode - et år eller en uge - har haft samme slags arbejde til samme gruppe uanset formålet med virksomhedens aktiviteter, uanset hvilken arbejdsstilling personen indtager og uanset hvilken uddannelse personen har gennemgået. Det vil sige, at det er selve arbejdsfunktionen, der skal klassificeres. En persons arbejdsfunktion består imidlertid ofte af flere forskelligartede funktioner eller funktionen har forskellige sider, man kan lægge vægt på - og derved komme til forskellige resultater. Dette er en fælles problemstilling ved alle indsamlingsmetoder. Men muligheden for at få de nødvendige informationer er forskellige.

Ved folketællingen i 1970 stilledes i alt 4 spørgsmål til de erhvervsaktive med henblik på at kunne placere dem arbejdsmæssigt, nemlig dels de tidligere nævnte spørgsmål om beskæftigelsen på tællingsdagen og om arbejdsstilling og dels spørgsmål om fag- eller stillingsbetegnelse og om virksomhedens art. Af forskellige årsager er det ikke let at sammenligne den skemabaserede folketælling i 1970 med de registerbaserede tællinger i årene efter 1980. For det første kan man sige, at faginddelingen i 1970 gik så langt som spørgeteknikken tillod, idet inddelingen blev defineret ved - for hver enkelt kode - at opremse en række stillingsbetegnelser, som henførtes til koden. For det andet fjerner vi os mere og mere fra den situation, at en stillingsbetegnelse dækker over et bestemt arbejde og at en uddannelse fører til et bestemt arbejde. Fra ILO's side, der har udarbejdet den fagnomenklatur, som Danmarks Statistik nu har taget i brug under navnet DISCO 88, er det da også sagt, at det ikke er nok, at få oplyst en stillingsbetegnelse, men at der også må indhentes oplysning om, hvori den vigtigste arbejdsopgave består. Om et sådant spørgsmål vil blive besvaret hensigtsmæssigt i en tælling som den danske folketælling 1970 uden generel brug af tællingskommissærer, får stå hen i det uvisse. Problemerne med at klassificere de erhvervsaktive efter fag i den registerbaserede statistik må ses i lyset heraf.

Klassificering af selvstændig næringsdrivende

Helt i overensstemmelse med DISCO-88 nomenklaturen klassificeres i Arbejdsklassifikationsmodulet selvstændige på grundlag af oplysning om virksomhedens branche og antal beskæftigede. Begge oplysninger findes i Indehaverregisteret, der er en del af Det erhvervsstatistiske Registersystem.
Indehaverregisteret omfatter ejere af personligt ejede virksomheder, der betaler meromsætningsafgift eller som har ansatte.
Andre selvstændige klassificeres efter stillingsbetegnelsen i Det centrale Personregister, dog forudsat at stillingsbetegnelsen giver udtryk for, at vedkommende er eller kan være selvstændig næringsdrivende, fx. 'Læge', 'Forfatter', '.....indehaver', '.....ejer', '.....forpagter' o.s.v.. Derimod anvendes ikke stillingsbetegnelser som Kontorassistent, Falckredder, Fuldmægtig, der er meningsløse i den forbindelse.

M.h.t. Erhvervsregisteret kan der være en vis afstand mellem en begivenheds indtræden og noteringen heraf i registeret. Registeret opdateres dog løbende, bl. a. på grundlag af meddelelser til Told- og Skattestyrelsen.
Der må regnes med en større fejlplacering blandt selvstændige, der ikke findes i Indehaverregisteret, idet stillingsbetegnelsen i folkeregisteret - og dermed i Det centrale Personregister - kun rettes på initiativ af den enkelte borger.
Det skal bemærkes, at DISCO-88 kræver oplysning om antal beskæftigede i virksomheden for at kunne placere selvstændige på 2-ciffer niveau og at en sådan oplysning ikke blev indhentet i de skemabaserede folketællinger.

Klassificering af medhjælpende ægtefæller

Medhjælpende ægtefæller skal -i lighed med andre erhvervsaktive - klassificeres efter arten af det arbejde, de udfører i virksomheden. Den eneste oplysning, vi har i de administrative registre, er oplysningen om branchen. Kun i sjældne tilfælde er der grund til at antage, at den medhjælpende ægtefælles uddannelse har relation til arbejdet og der er ikke foretaget undersøgelser af om det vil give væsentlig bedre resultater at inddrage uddannelsesoplysningen i klassifikationen af medhjælpende ægtefæller.
Medhjælpende ægtefæller i landbruget klassificeres som ufaglærte arbejdere i landbrug, medhjælpende ægtefæller i detailhandelen som ekspedienter medens øvrige klassificeres som kontormedhjælpere, men kun på 1-ciffer niveau.
I denne situation kan argumenteres for, at det vil være en fordel for brugerne, at man undlod at foretage denne klassificering, men i stedet gav en opdeling efter branche, således som det er gjort hidtil såvel i de skemabaserede som i de registerbaserede tællinger.

Klassificering af lønmodtagere

De datatyper i de administrative registre, der giver mest information om, hvilket arbejde en lønmodtager udfører, er stillingsbetegnelser eller angivelse af stillingsgruppe. Dernæst kan oplysninger om elev-ansættelsesforhold placere personer under praktikuddannelse. Endelig suppleres med oplysninger om arbejdsstedets branche, om antal ansatte i virksomheden, om tilhørsforhold til en arbejdsløshedskasse og om højeste afsluttede uddannelse

Stillingsbetegnelserne i Arbejdsklassifikationsmodulet hentes i Det centrale Personregister, som dækker hele befolkningen og dermed alle lønmodtagere. Men oplysningen er af flere årsager særdeles mangelfuld. Den kan for det første mangle helt eller være uaktuel. Den manglende opdatering skyldes, at oplysningen ingen eller næsten ingen betydning har for de administrative myndigheder. Initiativet til at få indført en stillingsbetegnelse i folkeregisteret eller til at rette den, skal i reglen komme fra den enkelte borger. Mange læg-

ger imidlertid ikke vægt på, hvorledes de bliver tituleret og siden skattevæsenet er ophørt med at bruge den ved adresseringen, er det også sjældent, at den enkelte borger bliver præsenteret for den.
For det andet giver stillingsbetegnelsen ikke altid oplysning om arbejdets art men fx. om, hvilken social position personen indtager (i samfundet eller i virksomheden), hvilket i det højeste kan udelukke en del fagkoder eller giver oplysning om, hvilken uddannelse personen har. Eksemplerne er talrige: 'Kontorchef' siger ikke om arbejdet er af administrativ eller faglig karakter, men angiver inden for det offentlige en lønklasse og i det private erhvervsliv en relativ lønposition. 'Konsulent' kan være sælger, rådgiver inden for mange forskellige discipliner, sagsbehandler, 'Sekretær' kan være kontorassistent, men også faglig medarbejder, betegnelserne 'Specialarbejder' og 'Arbejder' er ikke engang tilstrækkelig til at placere personen på 1-ciffer niveau i DISCO-klassifikationen, idet der skelnes mellem maskinoperatører (herunder proces-operatører og chauffører og lign.) og andre arbejdere. Endelig har nogle betegnelser slet ingen relation til nomenklaturen, fx. hr., fru, komtesse eller er næsten uden informationsværdi i relation til nomenklaturen, fx. assistent, medhjælper, elev.

Stillingsgrupperingskoder hentes fra lønstatistikregistrene. Registrene dækker på den ene side hver især kun en speciel del af arbejdsstyrken men lapper på den anden side over hinanden. Disse registre er:

Lønstatistikregisteret for ansatte i staten m.v.
Lønstatistikregisteret for ansatte i kommunerne m.v.
Lønstatistikregisteret for funktionærer i den private sektor

Blandt stillingsgrupperne findes i alle lønstatistikregistrene opsamlingsgrupper af typen 'Andre faglærte', 'Andre assistenter' og stillingsgrupper der mere angiver en lønklasse end et bestemt arbejde og følgelig ikke kan føre til en bestemt fagkode. Eksempelvis findes i lønstatistikken for statsansatte gruppen 'Inspektør med angivet arbejdsområde', der bl. a. omfatter skoleinspektør, lodsinspektør, landinspektør, fiskeriinspektør, og gruppen 'Instruktør', der bl. a. omfatter lokomotivinstruktør og sceneinstruktør. Fra funktionærlønsstatistikken kan anføres eksemplet 'Afdelingsleder', der kan være butiksleder, værkfører, byggeleder m.m.

Den grundlæggende enhed i lønstatistikken for ansatte i statsligt regi er et ansættelsesforhold i løbet af året, og der medtages alle ansættelsesforhold. I AKM anvendes ansættelsesforholdet med det største lønbeløb. Tilsvarende gælder for ansatte i kommunalt regi. Hvis en person har været ansat både i statsligt og kommunalt regi, anvendes ansættelsesforholdet med det største lønbeløb, dog kun såfremt det fører til et positivt resultat, d.v.s. til en fagkode på mindst 1-ciffer niveau.

Indberetningerne til lønstatistikken for funktionærer i den private sektor vedrører ansættelsesforhold i september måned. Der medtages kun ansættelsesforhold, der drejer sig om en beskæftigelse på mindst 15 timer ugentlig. Såfremt en person har mere end et ansættelsesforhold i september måned i de

firmaer, som indberetter til lønstatistikken, anvendes i AKM forholdet med det største lønbeløb.

Da lønstatistikregistrene refererer til aktuelle beskæftigelser, prioriteres oplysningerne herfra i forhold til oplysningen om stillingsbetegnelsen i CPR. Indbyrdes prioriteres de efter højeste lønbeløb, dog kun såfremt det fører til et positivt resultat. Om ansættelsesforholdet med det største lønbeløb i løbet af året er en ansættelse i statsligt regi , kommunalt regi eller i den private sektor konstateres ved hjælp af Oplysningsseddelregisteret. Det kan medføre, at en person, der i størstedelen af året har været beskæftiget hos en privat arbejdsgiver, der ikke indberetter til lønstatistikken får en fagkode fra et ansættelsesforhold, han/hun kun har haft i september måned eller kun har haft som bijob.

Personer, der ikke kan klassificeres efter stillingsangivelserne og som er under uddannelse på et uddannelsestrin og i en - retning, der gør det sandsynligt, at de på arbejdsmarkedet er ansat som elev, klassificeres efter denne oplysning i Uddannelsesklassifikationsmodulet

Til fagklassificeringen af lønmodtagere anvendes endelig oplysning om medlemskab af en arbejdsløshedskasse, om arbejdsstedets branche og om højeste uddannelse. Oplysningerne anvendes kun i kombination med hinanden og kun såfremt den pågældende ikke er klassificeret på grundlag af stillingsoplysninger eller på grundlag af oplysning om et elevforhold, der kræver fuldtidsbeskæftigelse på arbejdsmarkedet.
Således placeres medlemmer af S.I.D., der beskæftiget inden for landbruget, som ufaglært landbrugsarbejder, medlemmer, der er beskæftiget i bygge - og anlægsbranchen som ufaglærte byggearbejdere, medlemmer af H.K., der er beskæftiget i engroshandel klassificeres som handelsassistent, hvis de er uddannet som sådan, medlemmer af H.K., der er beskæftiget i detailhandel klassificeres som ekspedient medens øvrige medlemmer af H.K. klassificeres på 1-ciffemiveau i hovedgruppen 'Kontorarbejde'. Personer, der er uddannet som jurist og som er beskæftiget i et advokatfirma klassificeres som advokatfuldmægtig, ingeniører, der er beskæftiget i et rådgivende ingeniørfirma klassificeres som ingeniør o.s.v.

Medlemskab af en arbejdsløshedskasse refererer til situationen ved udgangen af året. Det er problematisk i hvor høj grad oplysningen er af værdi for fagklassificereingen. Dels har de fleste arbejdsløshedskasser en noget sammensat medlemsskare, dels finder et skifte ikke sted samtidigt med et skifte i arbejdet.

Arbejdsstedets branche refererer til virksomheden, der har udbetalt den højeste lønsum i løbet af året. Løn fra evt. flere arbejdsperioder i samme virksomhed er sammentalt til et lønbeløb. Har virksomheden flere arbejdssteder er i reglen anført det seneste arbejdssted, vedkommende har været beskæftiget på.

Oplysningen om højeste afsluttede uddannelse refererer til situationen i oktober måned. En uddannelse fører ikke nødvendigvis til samme arbejde og den enkelte erhverver sig i tidens løb andre kvalifikationer, som ikke registreres. Oplysningen anvendes derfor kun, hvor den 'passer' med brancheoplysningen og må alligevel betragtes som en klassificering med et noget ringe kvalitetsstempel.

Fagklassifikation som lønmodtager gennemføres for alle personer i Indkomststatistikregisteret og dermed i Arbejdsklassifikationsmodulet uanset beskæftigelsesstatuskode, ligesom fagklassifikation som selvstændig gennemføres for alle, der findes i Indkomststatistikregisteret og tillige i Indehaverregisteret eller som har en stillingsbetegnelse i Det centrale Personregister, der angiver, at der er tale om en selvstændig næringsdrivende. På denne måde kan fagklassifikationen anvendes i tællinger, hvor man når frem til en anden lønmodtagergruppe end i Arbejdsklassifikationsmodulet, fx i den registerbaserede arbejdsstyrketælling.

Registrenes forskellige opdateringstidspunkter og prioriteringsregler kan bevirke, at der kan være inkonsistens mellem de oplysninger, der anvendes i statistikken. Specielt vil jeg pege på, at brancheoplysningen refererer til den vigtigste branche for året som helhed medens andre oplysninger som oftest refererer til et tidspunkt i slutningen af året og at oplysningen fra lønstatistikregistrene, som foretrækkes, kan vedrøre en bibeskæftigelse.

Fremtidige ændringer i registeroplysningerne.

De oplysninger Arbejdsklassifikationsmodulet modtager fra lønstatistikregistrene vil i de kommende år ændre sig på flere måder.
For det første vil lønstatistikken for den private sektor ikke kun omfatte funktionærer men alle ansatte.
For det andet udvides kredsen af virksomheder, som indgår i undersøgelsen, idet der rettes henvendelse til alle virksomheder med mere end 10 ansatte. Indtil da var grænsen 20 ansatte i servicevirksomheder og 50 ansatte i industrivirksomheder.
For det tredie skal virksomhederne i stedet for den nuværende stillingsgruppe -kode angive den 4-cifrede DISCO-kode
Endelig vil også stillingsgruppekoderne i lønstatistikken for offentlig ansatte blive erstattet af DISCO-koden.

Hermed vil antallet af personer, der klassificeres på grundlag af stillingsoplysningen i Det centrale Personregister blive yderligere reduceret og klassificeringen vil for personer, der indgår i lønstatistikken ske på grundlag af en vurdering af arbejdsopgaverne, der skal løses af den enkelte og ikke på grundlag af en stillingsbetegnelse.

Socioøkonomisk klassificering

Formålet med en socioøkonomisk klassificering er at opdele befolkningen i grupper, der adskiller sig fra hinanden i henseende til social, økonomisk og kulturel adfærd. Denne adfærd anses i meget høj grad at have sammenhæng med en persons erhvervsmæssige arbejde som et vigtigt grundlag for den indflydelse og anseelse, som den pågældende opnår. Andre faktorer som fx. uddannelse, evner, opvækstvilkår fysiske levevilkår, indkomst- og formueforhold o.s.v. har også indflydelse på adfærdsmønstret. I de fleste undersøgelser, der bruger en socioøkonomisk gruppering, vil det imidlertid ikke være muligt at inddrage så mange faktorer for at fastlægge klassificeringen. Grundlaget for klassificeringen er derfor i praksis ofte udelukkende oplysninger, der vedrører den erhvervsmæssige beskæftigelse så som beskæftigelsesstatus, type af arbejde (fx manuelt contra ikke manuelt), evt. branche (især landbruget og liberale erhverv placeres i egne grupper) og for selvstændige evt. antal ansatte. Dette er også tilfældet med den socioøkonomiske gruppering, som fra FN's side er opstillet som fælles referenceramme i forbindelse med folketællingsrunden i 1990. Grupperingen er en variabel, der direkte kan afledes af den anbefalede statusinddeling, faginddeling og brancheinddeling.

Det er ligeledes tilfældet med den socioøkonomiske inddeling, som Det statistiske Departement introducerede i forbindelse med folketællingen 1960 med henvisning til anbefalingerne fra FN.
Med overgangen i 1980 til en ny tællingsteknik og til en ny fagkode (DFK) indførtes for lønmodtagere en selvstændig 1-cifret kode for socioøkonomisk gruppe, der har rødder i den gamle inddeling, idet den oprindelige skelnen mellem funktionærer, faglærte arbejdere og andre arbejdere blev opretholdt ligesom funktionærgruppen blev opdelt i tre, der indholdsmæssigt med nogle få undtagelser svarer til den oprindelige inddeling

Med den registerbaserede statistik vil det være muligt at inddrage flere forhold ved den socioøkonomiske klassificering. Man vil således kunne differentiere selvstændige ud fra oplysningen om antal beskæftigede og lønmodtagere ud fra oplysning om antal ansatte på arbejdsstedet. Dette vil måske især have interesse fordi der i den nye lønstatistik også vil være oplysning om, hvorvidt lønmodtageren har en lederfunktion i forhold til andre med samme fagkode, eller er elev o. l. eller har et væsentligt selvstændigt ansvar m.m.. Det vil også som det er ønsket i den fællesnordiske socioøkonomiske standard at udskille langvarigt arbejdsløse som en særlig gruppe i klassifikationen
Endelig skal nævnes, at man med det registerbaserede statistiksystem som det danske også vil have mulighed for at opdele personer ude af erhverv i flere socioøkonomiske grupper, fx på grundlag af deres tidligere erhvervsmæssige beskæftigelse og/eller uddannelse og/eller formueforhold og mulighed for at klassificere familier efter flere personers individuelle placering.

Integreret dataindsamling - et eksempel

Søren Hostrup-Pedersen

1. Definition.

Når en administrativ myndighed indsamler oplysninger for en anden myndighed samtidig med indsamling af oplysninger til egen administrativ brug kaldes indsamlingsformen "Integreret dataindsamling". Det ligger i definitionen, at den indsamlende myndighed hverken kan eller skal anvende de oplysninger, de indsamler for den anden myndighed.. Oplysningerne indsamles af den administrative myndighed integreret med indsamling af andre oplysninger, som myndigheden selv har brug for til løsning af sine administrative opgaver.

Ved denne definition skal der ved integreret dataindsamling således forstås, at en administrativ myndighed, fra virksomheder, institutioner el. lign., dels indsamler de oplysninger, som de selv skal benytte til løsning af deres opgaver, dels indsamler den administrative myndighed også oplysninger til en anden myndighed - fx. Danmarks Statistik - på samme skema og i samme arbejdsgang, som de indsamler oplysningerne til eget brug.

De oplysninger, myndigheden indsamler for Danmarks Statistik har myndigheden ikke selv behov for i sin administration. Der eksisterer bestemte regler for, hvilken brug den administrative myndighed herefter må gøre af de oplysninger, som er indsamlet for Danmarks Statistik og udelukkende til brug for Danmarks Statistik.

Der spares penge

Formålet med den integrerede dataindsamling er at spare omkostninger i forbindelse med dataindsamlingen. De omkostninger der spares på, er de samlede samfundsmæssige omkostninger, idet det er langt billigere at indsamle oplysninger i en samlet proces end i flere processer.

Oplysningerne bliver sammenhængende

Hertil kommer, at den integrerede dataindsamling bidrager til, at de indsamlede oplysninger er sammenhængende, og dækker de samme enheder (virksomheder og institutioner), mens dette ikke nødvendigvis ville være tilfældet, hvis indsamlingen blev foretaget af to institutioner i flere processer.

Den integrerede dataindsamling medvirker også til, at rent statistiske oplysninger kan indgå i sammenhæng med administrative data, som Danmarks Statistik også vil skulle gøre anvendelse af, i den eller de statistikker, som oplysningerne i den sidste ende skal anvendes til.

2. Lovgrundlaget

Lov om Danmarks Statistik

Lov om Danmarks Statistik indeholder bestemmelser om hvilke data, der kan indsamles af institutionen. Loven giver ingen retningslinier for indsamlings-

metoderne. Det er således i loven fastsat hvilke oplysninger eller hvilken type oplysninger, der kan indsamles fra erhvervsvirksomheder, institutioner m.v. , men der er ingen bestemmelser om oplysningerne skal indsamles ved hjælp af spørgeskema, ved direkte videregivelse af oplysninger, som institutionerne eller virksomhederne er i besiddelse af, ved maskinel dataoverførsel til Danmarks Statistik (efter aftalte retningslinier) - eller om oplysningerne indsamles af en anden myndighed ved integreret dataindsamling, eller eventuelt ved en helt femte metode.

Lov om offentlige myndigheders registre

Lov om offentlige myndigheders registre indeholder bestemmelser om behandlingen af personoplysninger på elektronisk medium. Lov om offentlige myndigheders registre skal sikre, at det offentliges oprettelse og brug af edb-registre med personoplysninger sker på en sådan måde, at den enkelte borgers retsbeskyttelse og integritet ikke krænkes. Efter loven må registre, der føres for en statslig myndighed, kun oprettes efter godkendelse af en minister. Forinden et register tages i brug, skal der være fastsat forskrifter for registerets opbygning og drift, der opfylder lovens bestemmelser. For registre, der føres af en kommunal myndighed, skal beslutning om oprettelse og forskrifternes fastsættelse som hovedregel træffes af vedkommende kommunalbestyrelse.

Hvem må bruge oplysningerne

Kapitel 3 i lov om offentlige myndigheders registre indeholder bestemmelser om registrering og opbevaring af oplysninger. §9 stk. 1. " Der må kun registreres oplysninger, der klart er af betydning for varetagelsen af vedkommende myndigheds opgaver. Der må endvidere registreres oplysninger, der klart er af betydning for varetagelsen af en anden myndigheds opgaver, såfremt registeret indrettes således, at oplysningerne kun kan benyttes af den anden myndighed ". Bestemmelsen i stk. 1, andet punktum , har til formål at sikre, at en myndigheds registre indrettes således, at der ud over de oplysninger, myndigheden selv har brug for, også kan registreres oplysninger, som klart er af betydning for en anden myndighed, når det findes praktisk. Det vil typisk være, hvor en myndighed i forvejen indhenter en række oplysninger fra borgerne (eller virksomhederne om f. eks. de ansatte) til brug for egen sagsbehandling, medens den anden myndighed skal anvende en oplysning, der knytter sig snævert hertil.

I en række tilfælde vil Danmarks Statistik have behov for både de oplysninger, der indsamles direkte og udelukkende til brug for Danmarks Statistik (og som derfor ikke må benyttes af den indsamlende myndighed) og de oplysninger, som den indsamlende administrative myndighed indsamler til eget brug. I disse tilfælde vil det samlede sæt af oplysninger indgå i de data, som oversendes til Danmarks Statistik, og indgå i de relevante statistikregistre i Danmarks Statistik.

3. Arbejdspladsprojektet

Arbejdspladsprojektet har til formål, at give en fortegnelse over alle arbejdssteder og fordele alle beskæftigelsesforhold på de korrekte arbejdssteder. Oplysningerne til arbejdspladsprojektet indhentes ved integreret dataindsam-

ling af Told og Skattestyrelsen fra de oplysningssedler, som arbejdsgiverne hvert år skal aflevere til Told og Skattestyrelsen for hver enkelt ansat. Oplysningssedlerne indeholder oplysninger om løn, pension, honorarer, uddelinger fra fonde m.v. for den enkelte ansatte eller modtager. Denne er identificeret ved personnummer, mens arbejdsgiveren er identificeret ved SE-nummer samt kode for ansættelsessted, i de tilfælde, hvor virksomheden har flere arbejdssteder. For både personen og virksomheden er der også oplysning om navn og adresse i klarskrift. Oplysningssedlen indeholder desuden oplysninger om ansættelsesperiode, samlet A-indkomst og A-skat, samt andre ydelser mv., hvor der ikke er indeholdt A-skat. Endelig indeholder oplysningssedlen i særlige felter oplysninger om, hvilke typer ydelser, der er ydet i det pågældende kalenderår.

Én gang om året - i et samarbejde

Arbejdspladsprojektet gennemføres en gang årligt i samarbejde mellem Told og Skattestyrelsen, Danmarks Statistik og Datacentralen, der fungerer som edb-bureau for Told og Skattestyrelsen. Told og Skattestyrelsen udsender i november/december de rå oplysningssedler samt vejledninger til virksomhederne om årets indberetninger. I denne udsendelse medsendes arbejdsstedsfortegnelser til virksomheder med flere arbejdssteder. Arbejdsstedsfortegnelserne er udarbejdet af Danmarks Statistik. De indeholder en fortegnelse over de arbejdssteder, som Danmarks Statistik har registreret for den pågældende virksomhed, således som det fremgår af Danmarks Statistiks Erhvervsregister. Fortegnelserne indeholder navn, adresse, og branche (branchenavn og kode) på hver enkelt arbejdssted. Endelig indeholder arbejdsstedsfortegnelserne den kode for arbejdsstedet, som skal påføres oplysningssedlen for hver enkelt ansat. Arbejdsgiveren skal ajourføre arbejdsstedsfortegnelsen med hensyn til oplysningerne om det enkelte arbejdssted. Arbejdsgiveren skal endvidere komplettere arbejdsstedsfortegnelsen med nye og nedlagte arbejdssteder, således at fortegnelsen, efter eventuelle korrektioner, indeholder en komplet liste over alle arbejdssteder, som har været aktive i årets løb.

Arbejdsstedsfortegnelsen indsendes til Danmarks Statistik. Såfremt der er tilkommet nye arbejdssteder, nedlagt arbejdssteder eller ændring i oplysningerne for registrerede arbejdssteder, påføres ændringerne, således at det sikres, at der er rigtige oplysninger for arbejdsstederne, at der er koder på de enkelte arbejdssteder, og at Danmarks Statistik er bekendt med de koder, som virksomheden anvender på oplysningssedlerne for det pågældende kalenderår.

4. Hvilke oplysninger indhentes i den integrerede proces.

4.1 Oplysninger til Danmarks Statistik

Arbejdsstedskoder og ansættelsesperiode til Danmarks Statistik

Arbejdsstedskoderne indhentes til brug for Danmarks Statistik. Desuden indhentes oplysninger om den ansattes ansættelsesperiode i det pågældende år, udelukkende til Danmarks Statistiks brug.

Ved oplysning om ansættelsesperiode udfyldes et af fire felter:
Enten: Ansat hele året - 1/1-31/12,
Eller: Ansat i en sammenhængende periode. Her angives fra dag, måned (hver med to cifre) til dag, måned (også med to cifre hver),
Eller: Ansat i flere perioder. Her angives kun om vedkommende var ansat sidste arbejdsdag i november måned, 30/11, (i 1993), der sættes kryds.
Eller: Ikke ansat sidste arbejdsdag i november måned. I så fald sættes et kryds.

Oplysningerne anvendes bl.a. til statistikker, der opgør beskæftigelsen ultimo november, og dette er årsagen til , at der i tilfælde af flere periodeansættelser i løbet af året skal gives en markering af, om vedkommende var beskæftiget den 30/11. For helårsansættelse vil personen blive betragtet som beskæftiget 30/11 og er man beskæftiget i en sammenhængende periode, men ikke hele året, vil det fremgå af periodeangivelsen, om man er beskæftiget 30/11.

Èn oplysningsseddel pr. ansættelsesforhold

Der skal kun afleveres en oplysningsseddel pr. ansat, også i de tilfælde, hvor en person i løbet af året er ansat på forskellige af virksomhedens arbejdssteder. Dette fremgår ikke eksplicit af vejledningen til oplysningssedlen, men i de instrukser, der gives af Danmarks Statistik til virksomhederne, når de spørger herom, siger vi, at det, der bør anføres som arbejdssted, er det seneste aktuelle (i de fleste tilfælde i slutningen af året).

Periodeangivelserne ikke fejlfri

Periodeangivelserne volder problemer; De er ikke i alle tilfælde korrekte. Det bør imidlertid indledningsvis slås fast, at hovedparten af oplysningssedlerne er korrekt udfyldte i periodeangivelsesfelterne. Omkring to trediedele af de ansatte er nemlig ansat hele året i samme virksomhed. Det viser opgørelsen over bruttobevægelserne i to på hinanden følgende år. Der er dog formentlig for mange oplysningssedler, der er udfyldt med markeringen: "ansat hele året". Mange lønsystemer er indrettet således, at personer ikke slettes af lønsystemet, før året er gået, selvom en person er afgået inden årets udgang. Hvis han har været ansat fra årets begyndelse, vil han derfor stå som ansat hele året. I tilfælde, hvor ansættelsen er sket i løbet af året vil han (eller hun) stå med en periodeangivelse, der har rigtig startdato, men med afgangsdato 31/12, fordi det derved sikres, at han kommer med i oplysningsseddeludtrækket.

Det er klart, at Danmarks Statistik forsøger at få rettet sådanne fejl - ved henvendelse til virksomhederne og til lønsystemerne. Vi er dog formentlig ikke kommet helt til bunds i løsningen af dette problem.

Ansat eller ej ansat, that is the Question

I visse tilfælde kan det være meget vanskeligt at afgøre, hvorvidt en person, der er ansat i flere perioder, eller som arbejder ind imellem, skal anføres som ansat 30/11 eller ikke ansat 30/11. Som eksempler herpå, er personer, i uopsagt stilling, der arbejder i perioder, og for hvem det kan være vanskeligt at afgøre, om der faktisk blev udøvet aktivitet ved udgangen af november. Et eksempel herpå er censorer. Disse står i nogle tilfælde anført med helårsansættelse og i andre tilfælde med en markering for ansat eller ej ansat ultimo november.

4.2 Oplysninger til Told og Skat

Skat, kontrol og ligning

Lønoplysningerne indhentes af Told og Skat til brug for deres indkrævning og kontrolfunktion, og til brug for ligningsarbejdet, som foregår i et samspil med de kommunale ligningsmyndigheder.

I relation til arbejdsgiverne individualiserer oplysningssedlerne den indeholdte A-skat. Summeres for samtlige ansatte i virksomheden, fremkommer den samlede indeholdte A-skat, som indgår i virksomhedens samlede mellemværende med skattemyndighederne.

Oplysningssedlen indeholder som nævnt også andre ydelser, som der ikke trækkes A-skat for, men som indgår som indtægt for pågældende modtager i ligningen i forbindelse med selvangivelsen for det pågældende indkomstår. Visse af disse ydelser indgår i Danmarks Statistiks indkomstbegreber som en del af den personlige indkomst.

Kopi af alle oplysninger til Danmarks Statistik

Danmarks Statistik modtager en kopi af samtlige oplysninger på oplysningssedlen, dels de oplysninger, der er indsamlet af Told og Skat direkte til Danmarks Statistiks brug, og dels de oplysninger, der indsamles til Told- og Skat's egen administrative brug.

For Danmarks Statistik er oplysningerne en helhed. Skulle Danmarks Statistik selv indsamle oplysningerne, ville de administrative oplysninger være lige så betydningsfulde som arbejdsstedsoplysningerne og datomarkeringerne, og de to sæt oplysninger skal for Danmarks Statistik ses i sammenhæng, idet den tolkning man kan give de for Danmarks Statistik indhentede oplysninger, afhænger af værdierne på de øvrige oplysninger.

Oplysningerne hænger sammen

På oplysningssedlen kvalificerer de data, der indsamles til Danmarks Statistik (arbejdsstedskoderne og datomarkeringerne) de data, der indsamles til direkte brug af den administrative myndighed (Told og Skat). Således giver datomarkeringerne en tidsperiode for den løn, der iøvrigt fremgår af oplysningssedlen. Herved kan oplysningerne bl.a. benyttes til skønsmæssige beregninger over den indtjente tidløn, opgjort i timer. Dette nyttiggør oplysningen til andre formål, end de formål, der ville kunne tilgodeses ved udelukkende at udnytte oplysningssedlens indtjeningsoplysninger.

5. Indsamlingsmetoden

Den integrerede dataindsamling foretages under een myndigheds hovedansvar. Virksomhederne er ansvarlige over for denne myndighed for aflevering af de pågældende oplysninger. Således kan man se på det, for så vidt angår selve oplysningssedlen. For denne blanket (som jo i øvrigt i de fleste tilfælde afleveres på maskinelt medium) er der således én myndighed , som virksomheden har ansvar overfor.

Oplysningssedler og arbejdsstedsfortegnelse

Det samlede sæt af oplysninger fra virksomhederne til brug for den statistiske anvendelse til arbejdspladsprojektet, indeholder dog mere end selve oplys-

ningssedlen. De tilhørende arbejdsstedsfortegnelser, der er udleveret til virksomhederne sammen med oplysningsseddelmaterialet, indsendes ikke til Told og Skat, men direkte til Danmarks Statistik, som herefter kontrollerer oplysningerne - i første omgang, at alle de oplysninger, der bør være på arbejdsstedsfortegnelsen, er der. Dette gælder f.eks. navn på arbejdsstedet, adresse, branche og arbejdsstedskode. I tilfælde af mangler, kontaktes virksomheden af Danmarks Statistik.

Tidsfrister

Tidsfristerne for indsendelse af Arbejdsstedsfortegnelserne er fastlagt således, at alle arbejdsstedskoder i princippet er aftalt, når selve oplysningsseddelmaterialet modtages i Danmarks Statistik. Det må understreges, at dette er det principielle og ideelle. I mange tilfælde viser den konkrete gennemgang af oplysningsseddelmaterialet, at virksomhederne har anvendt koder for arbejdssteder, som ikke fremgår af arbejdsstedsfortegnelserne. I andre tilfælde er der ikke ført arbejdsstedskoder på oplysningssedlerne, selvom det fremgår af arbejdsstedsfortegnelsen for den pågældende virksomhed, at der er flere arbejdssteder - og at koder for disse er aftalt mellem virksomheden og Danmarks Statistik.

Når bøgen springer ud går vi rigtigt igang

Ultimo april modtager Danmarks Statistik fra Told- og Skat en første version af oplysningsseddelregistret, der de seneste år indeholder godt 97 % af det endelige materiale, som modtages senere på året. Denne første version danner grundlaget for den konkrete gennemgang af oplysningssedlerne fra de enkelte virksomheder. Ved første gennemgang kontrolleres, at arbejdsstedskoderne svarer til koderne på arbejdsstedsfortegnelserne. Som nævnt tidligere, er dette dog ikke altid tilfældet. For halvdelen af de virksomheder, som har flere arbejdssteder er materialet formelt i orden, når det modtages i Danmarks Statistik. Der kan dog senere i processen vise sig at være fejlagtige oplysninger. For de virksomheder, hvor der mangler koder for alle, eller for nogle ansatte, kontaktes virksomheden af Danmarks Statistik for at få påført koder, hvor de mangler.

Kontakt med virksomhederne

Kontakten til virksomhederne sker skriftligt eller telefonisk eller en kombination heraf. I denne proces er det nødvendigt at rette henvendelse til ca. 4.000 virksomheder, hvor der for nogle ansatte mangler koder på oplysningssedlerne. Det samlede antal oplysningssedler, der dækker et lønmodtageransættelsesforhold udgør 5.003.000. Af disse skal 1.233.000 (1992) have arbejdsstedskoder, nemlig samtlige ansættelsesforhold i 8.200 private virksomheder med flere arbejdssteder, i alt 34.000 arbejdssteder. Den resterende del af lønoplysningssedlerne for den private sektor, omfatter ansættelsesforhold hos 168.000 arbejdsgivere med kun et enkelt arbejdssted, og hvor der derfor ikke skal påføres arbejdsstedskoder på de i alt 2.202.000 oplysningssedler i denne gruppe. For den offentlige sektor udgør antallet af oplysningssedler, der dækker et lønmodtageransættelsesforhold 1.567.000. For den offentlige sektor skal der imidlertid ikke påføres arbejdsstedskoder, idet arbejdsstedsoplysningerne indhentes i forbindelse med et særligt udtræk fra de offentlige lønsystemer, der giver kombinerede løn- og arbejdsstedsoplysninger.

Mange arbejdsstedsrettelser

Materialet gennemgår en lang række fejlsøgninger, blandt andet sammenlignes med forrige års materiale. På denne baggrund foretages også kontakt med

virksomhederne, og en årsproduktion giver anledning til rettelse af **ca.** 8.000 arbejdsstedsoplysninger, 4.000 arbejdsstedsrettelser vedr. nye arbejdssteder, 2.000 nedlæggelser og 2.000 ændringer i oplysninger på eksisterende arbejdssteder.

Virksomheder som ikke er omfattet af arbejdsstedsfortegnelserne, kan godt vise sig at have flere arbejdssteder , selv om de i Erhvervsregisteret kun er opført med et arbejdssted, og virksomheder med flere arbejdssteder kan godt have flere arbejdssteder, end der er angivet på arbejdsstedsfortegnelsen.

Bopæl og arbejds-sted - pendling

For at undersøge om der mangler arbejdssteder gennemgår oplysningsseddelmaterialet bl.a. en "pendlingskontrol", der sammenholder arbejdsstedsadressen med bopælsadressen. Arbejdsstedsadressen fremgår af Erhvervsregisteret og af arbejdsstedsfortegnelsen, mens bopælsadressen for den enkelte ansatte fremgår af CPR's oplysninger for den pågældende person, idet PNR identificerer personen i det enkelte ansættelsesforhold. Ved at sammenholde de to sæt adresser på kommuneniveau, kan det for hvert enkelt ansættelsesforhold afgøres, om afstanden mellem hjem og arbejde synes rimelig, idet der, som grundlag herfor, er indlagt en afstandsmatrice, der måler kilometerafstanden fra kommunecentrum til kommunecentrum, dvs. en matrice på 275 gange 275. Hvis der viser sig at være afstande, som overstiger "det rimelige", markeres det, og det giver anledning til, at der opdages arbejdssteder, som ikke fremgår af koderne eller af arbejdsstedsfortegnelsen. Denne kontrol gennemføres både for flere arbejdsstedsarbejdsgivere og for virksomheder, der kun er opført med et enkelt arbejdssted. I begge tilfælde retter Danmarks Statistik henvendelse til virksomheden, hvis pendlingsafstanden overstiger de grænser, som Danmarks Statistik har fastsat i fejlsøgningsprocessen.

Ved pendlingskontrollen fanges ikke alle arbejdssteder, idet der blandt andet kan være problemer de steder, hvor afstanden imellem arbejdsstederne er så lille, at pendlingskontrollen ikke er tilstrækkelig fintmasket. Der arbejdes imidlertid med opbygning af alternative metoder til løsning af dette problem.

6. Den statistiske anvendelse af oplysningerne

Indgår i personstatistikken som klassifikation

Arbejdspladsprojektets oplysninger anvendes af Danmarks Statistik til statistiske opgørelser. Arbejdspladsoplysningerne danner i første omgang grundlaget for individoplysninger i det personstatistiske system, der dækker arbejdsmarkedsoplysninger for en årsperiode samt for en tidspunktsopgørelse (ultimo november). Oplysningerne indgår således i det personstatistiske systems klassifikationsgrundlag.

Bindeled mellem personstatistik og erhvervsstatistik

Oplysningerne indeholder imidlertid også det afgørende bindeled mellem de personstatistiske opgørelser og erhvervsopgørelserne, idet SE-nummeret, med tilhørende arbejdsstedskode, giver oplysninger om personernes tilknytning til konkrete virksomheder og arbejdssteder. Dette giver et grundlag for opdatering af Erhversregisterets enheder, og personsummationerne opdaterer Erhvervsregisterets beskæftigelsesoplysninger på arbejdsstedsniveau.

Oplysningerne på personniveau giver således mulighed for erhvervsklassificeringer på individniveau, og erhvervsoplysninger giver gennem de tilknyttede personnumre mulighed for at knytte personvariable fra det personstatistiske registersystem til erhvervsenhederne.

RAS giver tilknytningen til arbejdsmarkedet

Den registerbaserede arbejdsstyrkestatistik,.. (RAS), inddrager arbejdspladsoplysningerne. RAS giver en beskrivelse af befolkningens tilknytning til arbejdsmarkedet på et givet tidspunkt i løbet af året, nemlig slutningen af november, svarende til datooplysningerne fra oplysningssedlen. Datagrundlaget udgøres af arbejdspladsprojektets grundoplysninger samt et uddrag fra en række administrative og statistiske registre, hvis oplysninger er bearbejdet med henblik på at belyse arbejdsmarkedstilknytningen i den sidste uge af november det pågældende år, for alle personer med bopæl i Danmark den 1. januar i det efterfølgende år.

RAS klassificerer

RAS er en statistik, der klassificerer alle personer i den danske befolkning efter deres tilknytning til arbejdsmarkedet. I RAS-statistikken indgår en primær og en sekundær klassifikation af befolkningen. Hovedopgørelsen vedrører den primære klassifikation, og det er herudfra, at hovedbegreberne "beskæftigede", "arbejdsløse", "i arbejdsstyrken" og "uden for arbejdsstyrken", hentes. De beskæftigede opdeles i selvstændige, medhjælpende ægtefæller og lønmodtagere, som kan underopdeles i stillingsgrupper. Der er tilknyttet uddannelsesoplysninger, således at man både for personer i arbejdsstyrken og uden for arbejdsstyrken kan opgøre, om de er uddannelsessøgende. For de færdiguddannede indgår endvidere oplysninger om, hvilken uddannelse, de har gennemgået. Der er herudover oplysninger om køn, alder og bopæl. Erhvervsoplysningerne vedrører det juridiske ejerforhold for firmaet samt branche og geografisk beliggenhed. Branchen opgøres på arbejdsstedsniveau, ligesom den geografiske beliggenhed opgøres på det konkrete arbejdssted.

Erhvervsbeskæftigelsen - koordineret med RAS

Materialet fra arbejdspladsprojektet og oplysningssedlerne anvendes også til udarbejdelse af erhvervsbeskæftigelsesstatistikken. Erhvervsbeskæftigelsen blev udarbejdet første gang for året 1990. Der anvendes samme definitioner i Erhvervsbeskæftigelsen som i RAS, og de to opgørelser er koordinerede.. Der er opbygget et basisregister, hvorfra der foretages udtræk til både den registerbaserede arbejdsstyrkestatistik og erhvervsbekæftigelsesstatistikken. I sidstnævnte opgøres både den primære og den sekundære beskæftigelse.

- på enhedsniveau

Koordinationen mellem de to statistikker er foretaget på enhedsniveau. Dette indebærer, at hver erhvervsenhed indeholder oplysninger, om de personer, der er tilknyttet virksomheden, og for hver person eksisterer entydig oplysning om den erhvervsenhed personen er tilknyttet. I erhvervsbeskæftigelsen kan man således opgøre virksomhedernes personelle sammensætning efter antal ansatte og disses fordeling på køn, alder, uddannelse og stillingsgruppe.

Man kan udtrykke det på den måde, at den registerbaserede arbejdsstyrkestatistik opgør arbejdsmarkedstilknytningen og beskæftigelsen set fra per-

sonsiden, mens erhvervsbeskæftigelsen opgør beskæftigelsen set fra erhvervssiden.

Uddrag af statistikregistrene anvendes endvidere som baggrundsdata i andre dele af det personstatistiske registersystem.

7. Fordelene ved integreret dataindsamling

Fordelene ved den integrerede dataindsamling kan resumeres som følger:

Mange fordele

- Der spares omkostninger i forbindelse med dataindsamlingen.
- Det sikres, at de indsamlede oplysninger er sammenhængende
- Statistiske oplysninger kan indgå i sammenhæng med administrative data
- Administrative og statistiske oplysninger bliver til en helhed
- Administrative data kan blive kvalificeret af de data, der indsamles udelukkende til Danmarks Statistik
- Virksomhederne ulejliges kun en enkelt gang

- ikke mindst økonomiske

Rækkefølgen ovenfor er ikke helt tilfældig. Omkostningerne ved den integrerede dataindsamling ligger væsentligt lavere end de omkostninger, der skulle have været afholdt, såfremt de supplerende statistiske oplysninger skulle have været indhentet af Danmarks Statistik ved henvendelse til virksomhederne. Der er tale om omkostninger af meget betydelig størrelse, så store, at det formentlig i praksis ville have været økonomisk ufremkommeligt at gennemføre en sådan supplerende dataindsamling. Arbejdspladsprojektets oplysninger, som her er eksemplet på integreret dataindsamling, er et uundværligt led i det personstatistiske system. De danner bindeledet mellem person- og erhvervsstatistiske opgørelser og er desuden et betydningsfuldt grundlag for den generelle erhvervsstatistik.

Det forhold, at statistiske oplysninger kan indgå i sammenhæng med administrative data medfører, at de administrative data bliver statistisk berigede og kvalificerede. Oplysningerne bliver sammenhængende og de kan derfor bedre dække de statistiske behov. Den integrerede dataindsamling sikrer således også, at nogle af de traditionelle svagheder ved brugen af administrative data til statistiske formål bliver reduceret. Man plejer at sige, at de administrative data er vanskelige at udnytte som grundlag for statistiske opgørelser, fordi de er underlagt ændringer i lovgivning og i følge sagens natur ændringer i administrative rutiner og procedurer. Denne ulempe kan således til en vis grad reduceres ved udformning af de statistiske data, der supplerer de administrative data i den integrerede dataindsamling.

Nemmere for virksomhederne

Respondentbyrden for virksomhederne reduceres. Selv om den samlede datamængde er den samme, hvad enten man sender forespørgsel til virksomhederne i en eller flere omgange, er ulejligheden for virksomheden betydeligt mindre ved en enkelt koordineret henvendelse.

8. Ulemperne ved integreret dataindsamling.

Nogle ulemper

- De oplysninger, der indsamles til brug for Danmarks Statistik kan være fremmede for den administrative myndighed.
- Kontrollen med den korrekte indberetning af oplysningerne til Danmarks Statistik kan blive nedprioriteret af den indsamlende administrative myndighed.

- men de kan reduceres ved godt samarbejde

Det kan volde problemer at indsamle oplysninger for en anden myndighed, da det administrative personale ikke er bekendte med de oplysninger, der indsamles for en anden myndighed. Instrukser og vejledninger vil kunne afhjælpe problemet, men fortolkningen af tvivlstilfælde kan være vanskelig at håndtere for den indsamlende myndighed. I arbejdspladsprojektets tilfælde kanaliseres nogle af disse problemer til Danmarks Statistik, idet en del af de til oplysningssedlen knyttede oplysninger om arbejdssteder afleveres direkte til Danmarks Statistik.

For den indsamlende administrative myndighed er det naturligt, at man prioriterer egne oplysninger højest. Dem er man bekendte med, og de er af afgørende betydning for hele institutionens arbejde, mens man ikke har direkte interesse i oplysningerne til Danmarks Statistik. Opfylder de data, der indsamles til eget brug, ikke de formelle og legale krav må man kontakte virksomheden og bede om nye og korrekte oplysninger. Fejlsøgningsrutiner og -processer er indarbejdede og kendte, mens dette ikke gælder for de øvrige oplysninger. I arbejdspladsprojektet har man derfor været ude for, at kontrollen med statistikdataene ikke i alle tilfælde er tilstrækkelig effektiv. Det er derfor af afgørende betydning, at så mange procesdetaljer som muligt aftales mellem den administrative myndighed og Danmarks Statistik, for at sikre, at indberettede fejl opdages og rettes så hurtigt og effektivt som muligt.

9. Konklusion

Samarbejde er nøgleordet

Fordelene ved den integrerede dataindsamling overstiger langt ulemperne. Denne konklusion er ikke overraskende - ellers ville man næppe have valgt denne dataindsamlingsmetode. Afgørende for succes'en er imidlertid, at der etableres et godt samarbejde mellem den administrative myndighed og Danmarks Statistik omkring indsamlingen af oplysningerne. Det kan sikre en god kvalitet og en effektiv udnyttelse af den administrative myndigheds professionalisme i indsamling af oplysninger hos virksomheder eller borgere. Kombineres det med professionalisme hos Danmarks Statistik er den integrerede dataindsamling en effektiv metode til at opnå data af høj kvalitet.

Imputering

Lone Solbjerghøj

Indledning

I statistikproduktionen støder man hyppigt på problemet, at ikke alle ønskede og relevante data er oplyst, eller der er ikke data for alle enheder i statistikken. Et problem man kender fra statistik produceret på basis af administrative registre såvel som, når dataindsamlingen er sket via surveys.

Manglerne kan forekomme i mange forskellige grader. Lige fra det tilfælde, hvor en relevant variabel slet ikke er belyst i registret, til det tilfælde, hvor det drejer sig om mangel af en enkel ud af flere elementaroplysninger, som skal stykkes sammen og fastlægge værdien af en bestemt variabel for en enhed - således som det fx. sker med variablen arbejdsstilling i arbejdsklassifikationsmodulet.

De manglende oplysninger kan i nogle tilfælde indhentes ved at kontakte datadonorerne igen. Men selvom dette er mulig, vil det ofte kræve mange ressourcer og belaster datadonorerne yderligere. Derfor foretrækker man ofte at imputere de manglende oplysninger, dvs. at generere de pågældende data v.hj.a. nærmere fastsatte regler, der knytter sig til andre kendte størrelser (Danmarks statistik, 1981).

Når oplysninger mangler helt

Mangler der information om den relevante variabel i statistikgrundlaget, kan den eventuelt indhentes ved samkørsel med andre registre eller ved at udvide indberetningerne til registret således, som det er beskrevet i afsnittene 4.2 og 4.3. I nogle tilfælde vil der kun være mulighed for at indhente oplysningerne ved en selvstændig dataindsamling. Det er ofte meget bekosteligt at indhente data for hele registerpopulationen. Dette kan afbødes ved, såfremt det er muligt og forsvarligt, at indhente oplysningerne for et udsnit af populationen, hvorefter der efterfølgende opregnes til hele statistikpopulationen. Nedenfor er beskrevet et eksempel, hvor denne fremgangsmåde har været anvendt, og hvor imputeringen således sker i forbindelse med opregningen.

Eksempel 1:

Arbejdsløshedsstatistikken bygger på CRAM-registret, som modtager indberetninger fra kommuner og arbejdsløshedskasser om alle personer, der i en periode har været berørt af ledighed, og som fik udbetalt arbejdsløshedsdagpenge eller kontanthjælp som ledige. Man ønskede at supplere statistikken med oplysninger om årsagerne til ledighed - altså årsagen til, at personen påbegynder en ledighedsperiode og dermed optræder i registret. Disse oplysninger kunne ikke indhentes ved samkørsler med andre registre og en udvidelse af indberetningerne til

registret totalt ville ikke kunne sættes i værk så billigt og hurtigt, som det var ønsket.

Man har derfor valgt at indhente oplysningerne om forekomsten af ledighed og årsagerne hertil ved interviews, men kun for et udsnit af befolkningen. For hver interviewperiode (et kvartal) sammenholder man interviewpersonerne med arbejdsløshedsregistret dækkende samme periode og finder de interviewpersoner, der optræder i registret. Man regner derefter svarene angående ledighedsårsager for disse interviewpersoner op til den samlede registerpopulation.

Kritik: fordele og ulemper

Metoden er velegnet, såfremt man ønsker at udbygge et Statistikregister, der som CRAM-registret er et omfattende register med høj datakvalitet men uden den store variabelrigdom.

Kombination af registerdata med data indsamlet ved interviews er velegnet, hvis man ønsker at udbygge statistikgrundlaget med oplysninger af mere kvalitativ karakter, som ikke altid egner sig til at blive indhentet fra administrative registre.

Supplement med interviewdata kan dog blive ressourcekrævende, idet hele organisationen omkring afvikling af en survey skal stilles på benene. I forbindelse med den beskrevne undersøgelse af ledighedsårsager kunne omkostningerne holdes nede, idet det ikke var nødvendigt at opbygge en selvstændig survey. Der var mulighed for at koble sig på den løbende dataindsamling, der finder sted til den interviewbaserede arbejdskraftundersøgelse.

Imputering når oplysninger mangler delvis

Hvorfor imputere

I andre tilfælde er de relevante variable belyst i datagrundlaget, men der mangler oplysning om værdien af en eller flere variable for enkelte af enhederne. Der kan også være tale om, at kvaliteten er så ringe, at eventuelle oplysninger er uanvendelige som statistikgrundlag.

I en statistik, der bygger på data indsamlet ved survey, vil man kunne tage hensyn til bortfaldet ved opregningen, idet man der antager, at bortfaldet har samme fordeling, som man finder for de kendte enheder eller for en bestemt del af dem.

Men statistikregistre indgår ofte ved samkørsler med andre registre i sammenhængende systemer og danner grundlag for flere forskellige statistikker. Man skulle derfor foretage nye opregninger hver gang, registret indgik i en specialstatistik.

Da dette ikke er hensigtsmæssigt, vil man foretrække at generere en værdi for den manglende variabel gennem anvendelse af et sæt regler knyttet til kendte størrelser. De pågældende værdier kan være forkerte i forhold til den pågældende enhed, men vil i den aggregerede statistik fremtræde med en fordeling, som med stor tilnærmelse giver en korrekt belysning af den pågældende variabel (Danmarks statistik, 1981).

Imputeringsmetoder

Fastlæggelse af reglerne afhænger af manglernes karakter og de oplysninger, man har til rådighed, og der findes en lang række imputeringstekniker, der dog kan grupperes i 3 hovedtyper. For en mere detaljeret gennemgang af metoderne se Ferguson (1993) og ABS (1993):

1. Den imputerede værdi fastlægges p.grl.a. anden information om enheden

I nogle tilfælde kan den manglende information udledes deduktivt dvs. korrekt eller med stor sikkerhed fra redundant information. Dette sker fx. i den registerbaserede beskæftigelsesstatistik, hvor man bl.a. tæller personer, der er beskæftiget på en bestemt dato (sidste arbejdsdag i november). I nogle tilfælde mangler oplysningen om ansættelsesperiode, og dermed om ansættelsen var gældende i november, men er der indbetalt et bidrag til Arbejdsmarkedets Tillægspension svarende til et årsbidrag, udleder man heraf, at personen har været beskæftiget i november.

Men oftest indgår et skøn i imputeringen. På grundlag af andre oplysninger om enheden fx. uddannelse og branche vil man kunne imputere eksempelvis en arbejdsstilling. For en løbende statistik vil oplysningen om variabelværdien fra tidligere opgørelser evt. kunne anvendes. Se i øvrigt nærmere i afsnit 4.4.3.1.

2. Den imputerede værdi fastlægges ved brug af modeller

Her kan der være tale om at anvende viden om gennemsnitsværdi eller medianværdi for den variabel, der skal imputeres for, eller man inddrager viden om variablens værdi i forhold til værdien for andre variable, som det sker ved ratio- og regressionsimputering.

Er der fx. behov for at imputere et lønbeløb, kan en gammel værdi for enheden fra året før blive fremskrevet med den forudsætning, at udviklingen i det seneste år for enheden har været den samme, som blev fundet for de fuldt oplyste enheder af samme type.

3. Den imputerede værdi fastlægges på grundlag af information om andre enheder.

Mens man ved brugen af modeller ved tildeling af imputeringsværdier anvender viden om, hvorledes andre lignede enheder opfører sig, vil man her imputere en konkret værdi fra en af de andre lignende enheder.

Ved såkaldt cold-deck imputering anvender man data fra en tilsvarende, tidligere gennemført tælling, mens man ved hot-deck imputeringsmetoderne anvender data fra den aktuelle tælling.

Som centrale elementer i metoderne indgår en donorenhed og en modtager- eller værtsenhed. En nærmere beskrivelse af forskellige teknikker til udvælgelse af den mest velegnede donorværdi findes i ABS, (1993). I eksempel 3 nedenfor er anvendt såkaldt hierarkisk hot-deck imputering, hvor man gentagne gange sorterer og matcher donorer og modtagere på et faldende antal variable, indtil man opnår match.

Ofte indgår flere imputeringsteknikker sammen - således også i de eksempler, der nedenfor er beskrevet nærmere.

Editeringsteknik

Udvælgelsen af en imputeringsværdi kan ske manuelt eller automatisk. Ved manuel behandling er det en person med kendskab til de sammenhænge, der ligger i data, der foretager valget af den imputerede værdi. Ved en automatisk imputering sker valget af værdier maskinelt efter fastlagte regler.

Ved brug af automatiske dataediteringsprogrammer, der omfatter automatiske kodnings- og fejlsøgningssystemer, kan også imputeringsregler være tilkoblet således, at behandlingen af uoplyste variable eksempelvis af skoleuddannelse sker integreret med en automatisk kodning og fejlsøgning. I de tilfælde, hvor skoleuddannelsen mangler, men en senere erhvervsuddannelse er oplyst, kunne afsluttet skoleuddannelse evt. - efter nærmere fastlagte regler - genereres i en del tilfælde.

Der er adskillige fordele ved maskinel imputering:

- Automatisk imputering er billig, forudsat at antallet af uoplyste enheder har et vist omfang, for som det også vil fremgå af eksemplerne nedenfor, er der forbundet et vist arbejde med at udvikle imputeringsprogrammer.
- Metoden sikrer, at alle enheder bliver behandlet ens - efter de samme imputeringsregler.
- Og endelig sikrer maskinel imputeringsteknik, at alle har (eller nemt kan få) kendskab til de regler, der anvendes.

Men som det også vil fremgå af eksempel 2, bør man ikke fuldstændigt forlade sig på automatisk dataeditering, idet man jo kun ved den manuelle datarevision får kontrolleret de regler, imputeringen er bygget op omkring.

Imputering på grundlag af anden information om enheden

Eksempel 2:

Eksemplet er hentet fra den registerbaserede beskæftigelsesstatistik (Egmose, 1991). Alle beskæftigede personer knyttes her til et arbejdssted, og har arbejdsgiveren flere lokale arbejdssteder, skal hver enkelt ansat placeres på den filial, hvor han eller hun arbejder. Personerne kan så tildeles en række variable vedrørende deres erhvervsforhold fx. branche. Oplysningen om arbejdssted findes i langt de fleste tilfælde, men der er tilfælde, hvor der mangler oplysning om hvilken filial, personen arbejder på, og det ikke har været muligt at fastlægge dette i den manuelle behandling. Ved imputering sker der så en placering på de lokale arbejdssteder af de ufordelte personer.

Imputeringen foretages af to trin. Først undersøges, om det er muligt at henføre personen til samme arbejdssted som sidste år. Dette kræver dog, at arbejdsstedet var oplyst for personen året før, at arbejdsstedet stadig eksisterer, at personen er ansat hos samme arbejdsgiver i år, og at personen ikke har ændret bopæl (over større afstande).

Når imputering til sidste års arbejdssted ikke kan foretages, sker der i næste trin en imputering til nærmeste arbejdssted i forhold til bopælen. Har flere arbejdssteder samme afstand til bopælen, vælges det største arbejdssted (målt på lønsum).

Afstandsmatricen, der anvendes til fastlæggelse af nærmeste arbejdssted, opererer kun med afstande mellem kommunecentre. Har arbejdsgiveren flere arbejdssteder i samme kommune, kan kun ét indgå i imputeringen, og det største arbejdssted målt på lønsum udvælges. Har man ingen oplysninger om arbejdsstedernes størrelse (fx. når ingen af de ansatte hos arbejdsgiveren er fordelt ud på filialerne), udvælges det arbejdssted i kommunen, der har den laveste identifikationskode.

Resultaterne af denne maskinelle imputering til nærmeste arbejdssted checkes manuelt. Har et arbejdssted ved imputeringen fået tildelt et større antal ansættelsesforhold, checkes det, om imputeringen er acceptabel, eller om der skal ske korrektioner. Specielt er en manuel efterkontrol nødvendig i de tilfælde, hvor arbejdsgiveren har flere arbejdssteder i en kommune - eksempelvis en detailhandelskæde. Den maskinelle imputering henfører jo alle personer med manglende filialoplysninger, som bor i eller nær den pågældende kommune, til det ene arbejdssted, der er udtaget til at indgå i imputeringen.

Korrektioner foretages ved at fastlægge, hvilken andel af ansættelsesforholdene, der skal flyttes til de forskellige arbejdssteder. Udvælgelsen af hvilke personer, der skal tilknyttes hvilke arbejdssteder, sker maskinelt på grundlag af de 2 tilfældige tal i personnumrene.

Kritik: fordele og ulemper

Ovenstående eksempel beskriver et relativt kompliceret imputeringsforløb, der inddrager en lang række variable og beregninger. Inden imputeringsmetode vælges bør den valgte metode testes. I eksemplet beskrevet ovenfor blev metoden således først testet på et fuldstændigt oplyst materiale. Man udtog en gruppe enheder med fuldt oplyste og valide arbejdsstedskoder, undertrykte derpå oplysningen om arbejdsstedskode og gennemførte en imputeringsproces, hvorefter resultatet heraf kunne sammenlignes med de korrekte resultater, og metoden blev justeret, til den tilnærmelsesvis gav de samme resultater som ved brug af korrekte data.

Man bør løbende følge omfanget af enheder og se, om de typer af enheder, der henvises til imputering, ændres.

Ligeledes skal man løbende følge udviklingen og resultaterne af de foretagne imputeringer for at sikre sig de mest hensigtsmæssige regler til fastlæggelse af værdier.

Da de anvendte regler jo bygger på tidligere erfaringer, kan imputering få en konserverende virkning på statistikken.

Den skønnede værdi, der tildeles ved imputeringen, kan være forkert i forhold til den pågældende enhed. Og dette kan være et problem, når enheden indgår i flerdimensionale opgørelser. Man bør derfor altid markere, at en værdi er imputeret.

Imputering på grundlag af information om andre enheder

Et specialtilfælde af imputering skal kort omtales her. Metoden anvendes, når variabelværdien ikke fastsættes på grundlag af andre oplysninger om enhederne, men på grundlag af værdier for andre - men lignende - enheder.

I processen indgår en værtsfil A, som indeholder information om variabelsættene (X,Y) og en donorfil B, som indeholder variabelsættene (X,Z). Man ønsker en samlet belysning af (X,Y,Z) eller (Y,Z). Variabelsættet X identificerer enhederne i de 2 filer - kan fx. bestå af demografiske variable.

Eksempel 3:

I arbejdskraftundersøgelserne indhentes for interviewpersonerne bl.a. en lang række oplysninger om beskæftigelsesforhold, jobsøgning og deltagelse i uddannelsesaktiviteter.

Interviewdata indhentes for enkeltpersoner, men man ønsker også målvariablene belyst for hele familien.

I befolkningsstatistikregistret finder man de øvrige personer i familien. Disse personer skal tildeles værdier på målvariablene fra de allerede indhentede interviewdata.

Dette sker på følgende måde: der samkøres først med flere registre, således at man får hver enkelt af interviewpersonens familiemedlemmer beskrevet med køn, alder, bopæl, familietype, uddannelse, arbejdsløshedsberøring etc. Blandt de gennemførte interviews afgrænser man derefter de interviews, der er gennemført med personer, der matcher den type personer, der er i familien, hvorefter man for hvert familiemedlem udvælger (tilfældigt) en passende interviewperson. Herefter kan interviewpersonens familiemedlemmerne få tildelt værdier på målvariablene udfra hvordan, de matchende personer har svaret.
Lykkes det ikke at finde en donorrecord i første omgang, reduceres antallet af variable successivt, indtil match opnås.

Kritik: fordele og ulemper

Der er her tale om en metode, hvor man med et begrænset ressourceforbrug får suppleret en statistik med manglende oplysninger, idet man undgå den dyre proces med en selvstændig dataindsamling.

Endvidere kan metoden være velegnet i forbindelse med registerstatistik, da man her arbejder med de store datasæt, der er nødvendige for at opnå en sikker match. Anvendes metoden ved mindre datasæt, kan man få problemer med at finde en donorrecord, der helt svarer til variabelsættet X. Og jo mere specificerende X er, jo hyppige vil man komme i den situation, at man må svække på matchningskravene for at kunne finde en donorrecord. Eventuelt vil matchningsgraden derfor være forskellig for de enkelte (Y,Z) records i datasættet.

Når metoden anvendes, skal man være opmærksom på, at flere forhold kan få betydning for det opnåede resultat. Det skal fastlægges, om Z-værdierne skal tildeles med tilbagelægning, eller om en donorrecord kun kan bruges en

gang. Metoden bygger på antagelsen af uafhængighed mellem datasættene Y og Z. En forudsætning, man skal testes inden anvendelsen. Og et problem for en sikker match kan være, at de identificerende variable X (køn, alder,....) har høj korrelation med nogle af variablene i Z, men ikke med dem alle.

For en nærmere gennemgang og kritik af metoden se Rodgers (1984) og Singh,Mantal,Kinack og Rowe (1993).

Litteraturliste:

Australien Bureau of Statistics (May 1993): Dataediting.
Internt papir

Danmarks Statistik (1981): Det personstatistiske registersystem

Egmose, Sven (1991): Imputeringen 1989
Internt papir, Danmarks Statistik

Ferguson, Dania P. (1993): Reviews of Methods and Software used in Dataediting.
US Dept. of Agriculture.

Rodgers, Willard L. (1984): An Evaluation of Statistical Matching
Journal of Business & Economic Statistics, Vol2, no 1, pp 91-102

Singh, A. C., Mantal, H. J., Kinack, M. D. og Rowe, G. (1993): Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption.
Survey Methodology, vol 19, no 1, pp. 59-79. Statistics Canada.

Surveys og registre

Marius Ejby Poulsen

Hovedkilden for de oplysninger der anvendes i personstatistikken er administrative registre. I nogle tilfælde er disse, som det tidligere er fremgået, utilstrækkelige, hvorfor man kan se sig nødsaget til at indsamle de nødvendige oplysninger direkte hos de omhandlede personer eller et udsnit af disse, dvs. foretage en survey.

De oplysninger man indsamler i en survey, kan dels ses som en statistik i sig selv og dels som en udbygning af den eksisterende statistik. Beskæftigelsesstatistikken i Danmarks Statistik er et eksempel på dette, med en registerbaseret del, der bla. omfatter "Registerbaseret Arbejdsstyrkestatistik" (RAS) og en surveybaseret del "Arbejdsstyrkeundersøgelsen" (TASU). De to dele er hver for sig grundlag til belysning af arbejdsmarkedsforhold i Danmark, men surveyoplysningerne kan også ses som supplerende information, der kan integreres i RAS, med henblik på en udbygning af visse områder med nye eller omdefinerede variable. De sammenfaldende begreber i de to typer af oplysninger kan samtidig danne grundlag for en evaluering af datakvaliteten, hvilket er temaet i dette papir.

Ved at benytte data, som er knyttet til de til enhver tid gældende regler, vil der, når reglerne ændres, være en risiko for, at de anvendte begreber ikke er i overensstemmelse med de oprindelige definitioner eller at betydningen af valgte kriteriebetingelser kan vise sig at være anderledes end forventet. Derfor er der behov for en analyse af registerstatistikbegrebernes faktiske indhold, for at sikre, at man belyser det tilsigtede. På surveysiden formuleres en række spørgsmål, som tilsammen repræsenterer en form for "idealdefinition" af de begreber det tilsigtes at belyse. Man kan ikke altid forvente, at de indsamlede og behandlede oplysninger, er i overensstemmelse med denne idealdefinition, men en sammenstilling af oplysningerne herfra med tilsvarende oplysninger fra registerstatistikken, giver alligevel mulighed for at undersøge, om der er utilsigtede virkninger af den anvendte fremgangsmåde, i behandlingen af data fra de to kilder.

I efteråret 1993 påbegyndtes et projekt - registerkontrolprojektet. Den første opgave under registerkontrolprojektet drejer sig om at sammenligne data fra arbejdstyrkeundersøgelsen og den registerbaserede arbejdsstyrkestatistik. Denne opgave danner basis for beskrivelsen af de generelle forskelle mellem surveys og registre nedenfor, ligesom de udvalgte resultater stammer herfra.

Hvis man skulle anføre nogle væsentlige forskelle på surveys og registre indenfor personstatistikken, kan dette være antallet eller omfanget af personer, kravet om ajourføring[1] samt måden hvorpå oplysningerne indsamles.

[1] Jf. Danmarks Statistik 1982, s. 10.

Forskellen med hensyn til antal og omfang af personer eksisterer blandt andet af to grunde. For det første er der i en survey typisk tale om, at man indsamler oplysninger for et repræsentativt udsnit af en større gruppe personer (fx fra et register), for eventuelt senere at bearbejde disse oplysninger ved opregningsprocedurer, til en størrelsesorden, der svarer til alle personer. For det andet er der spørgsmålet om bortfaldsproblematikken, da man i princippet kunne foretage en survey for alle personer, men på grund af bortfald, vil de ønskede oplysninger kun være til stede for de personer, der deltager aktivt i den givne survey.

Med hensyn til kravet om ajourføring, er der for nogle surveys vedkommende tale om, at der foretages dataindsamling for de samme personer på flere tidspunkter. Dette er dog i surveys ikke et krav, som for registrenes vedkommende, men typisk en konsekvens af formålet.

Hvad angår måden hvorpå oplysningerne tilvejebringes i de to indsamlingsmetoder og de konsekvenser der følger heraf, skal følgende bemærkes. På registersiden skal man skelne mellem administrative registre, der typisk føres udfra administrative formål og statistikregistre, der anvendes til statistiske formål og som i mange tilfælde har karakter af bearbejdede administrative registre. I forbindelse med statistikregistre indsamles oplysninger om personer typisk, i den form, som de administrative myndigheder har dem liggende, hvorefter de eventuelt bearbejdes, med det formål at foretage (register)statistiske opgørelser.

I surveys er de oplysninger der indsamles typisk personers svar på givne spørgsmål, hvilket i mange tilfælde kan være mere holdningsprægede spørgsmål. Dermed opstår der direkte en usikkerhed med hensyn til troværdighed, dvs. det forhold, at interviewpersonernes svar ikke altid stemmer overens med de faktiske forhold (forhold der eventuelt belyses i registrene). Surveys har dog den fordel fremfor registre, at man er mere frit stillet overfor hvilke typer af oplysninger man kan uddrage.

Generelt kan man tale om indholdsmæssige argumenter for at vælge den ene indsamlingsmetode fremfor den anden. Der kan i nogle tilfælde ligeledes være tale om økonomiske/ressourcemæssige argumenter. Begge argumenter indgår i forklaringen af den kraftige udbygning af administrative registre og tradition for anvendelse af disse til statistiske formål, der kendetegner Danmark.

At der er forskelle i de to former for indsamlingsmetoder betyder ikke nødvendigvis, at der er forskel i de oplysninger der indsamles. Således er det oplagt at sammenligne den information man har om en gruppe personer i registerstatistikken, med tilsvarende information fra en survey.

Der kan være flere formål med at foretage en sådan sammenligning. Et formål kan være at se på hvorledes de to former for dataindsamling karakteriserer enkeltenheder, hvilket i denne sammenhæng vil sige personer. Et andet

formål kan være at se på, hvad de to dataindsamlingsmetoder siger om aggregerede størrelser, såsom arbejdsløshedens størrelse og sammensætning, beskæftigelsesstrukturen, mm. Det er førstnævnte problematik, der behandles i dette papir.

Sammenligning af RAS-91 og TASU-91

Indtil videre er nogle overordnede overvejelser mht. sammenligning af surveybaserede og registerbaserede opgørelser blevet nævnt. I resten af afsnittet vil et par af disse blive uddybet og resultater fra konkrete sammenligninger vil blive præsenteret[2] .

På surveysiden anvendes:
- Telefoninterviewbaseret ArbejdsStyrkeUndersøgelse fra 1991 (**TASU-91**)[3] .

På registersiden anvendes:
- Registerbaseret ArbejdsstyrkeStatistik fra 1991 (**RAS-91**)[4] .
- Ledighedsoplysninger fra Arbejdsløshedsstatistikregistret fra 1991 (**LEDR-91**)[5]
- Bearbejdede oplysninger fra det centrale oplysningsseddelregister (COR) fra 1991 (**CON-91**).

Uanset hvilken fremgangsmåde man vælger ved sammenligning af surveys og registre, med henblik på evaluering, bør følgende forudsætninger være opfyldt:

Tids-konsistens: De oplysninger der sammenstilles skal tidsmæssigt stemme overens. Man kan dog forestille sig, at andre oplysninger kan agere "bindeled" mellem tidsmæssigt adskilte oplysninger.
Begrebs-konsistens: De begreber eller variable der sammenlignes skal begrebsmæssigt stemme overens. Man kan dog i nogle tilfælde godt anvende variable, der baseres på forskellige klassifikationer, typisk på et overordnet niveau, ligesom man kan forestille sig variable der anvendes som indikator for evaluering af andre.

Med hensyn til forudsætningen om den tidsmæssige konsistens, er denne ikke opfyldt ved sammenligning af oplysninger fra TASU-91 og RAS-91. Referenceperioden i TASU-91 er fra 11. marts til 28. april 1991, mens RAS-91 refererer til sidste uge i november 1991. Dette skaber umiddelbart et problem, idet uoverensstemmelser mellem oplysninger, blot kan skyldes, at personerne har skiftet status i den mellemliggende periode.

For at tage højde for ovenstående problem er oplysninger fra CON-91 inddraget. CON-91's rolle i analysen er at agere tidsmæssigt "bindeled" mellem

---

[2] I gennemgangen af eksemplerne anvendes de relevante variabelnavne og i nogle tabeller de tilhørende variabelværdier. For sammenhængen mellem variabelnavne,- værdier og tekster henvises til Bilag 1 og 2.
[3] Se Danmarks Statistik, 1992:20.
[4] Se Danmarks Statistik, 1993:17.
[5] Se Danmarks Statistik, 1992:24.

RAS-91 og TASU-91. CON-91 indeholder bearbejdede oplysninger om alle de ansættelsesforhold, der er registreret via de oplysningssedler, som private arbejdsgivere, offentlige myndigheder, m.fl. hvert år indsender til Told- og Skattestyrelsen. På baggrund af disse er det til en vis grad muligt at tidsfæste ansættelsesforhold, og derved kan fx personer med ansættelsesforhold, der strækker sig over perioden fra 11. marts til sidste uge i november 1991, afgrænses.

Det er dog forbundet med visse problemer at anvende disse oplysninger. Det er flere gange vist, at specielt oplysningerne om ansættelsesperiode på oplysningssedlerne, ikke stemmer helt overens med virkeligheden.

I en rapport fra forskningsgruppen vedr. Arbejdsmarkedspolitikkens Tilrettelæggelse og Administration, ATA-projektet (se Nielsen, P., 1987), blev der foretaget en analyse af reliabiliteten af oplysningerne i COR-83 for Århus og Ålborg kommune. Analysen gik blandt andet ud på at vurdere, hvor stor en del af en given mængde ansættelsesforhold, der kunne antages at være stabile, dvs. at oplysningerne om ansættelsesforhold, efter en grundig revisionsprocedure, der blandt andet inddrog tidligere årgange af COR- og RAS-oplysninger samt oplysninger om ATP-bidrag, løn og ledighedsgrad, blev "godkendt", som værende i overensstemmelse med de faktiske ansættelsesforhold.

I analysen opererede man med de personer, der opfyldte følgende kriterier i COR-83:

1. Markering for helårsansættelse
2. Alternativt markering for ansættelse fra 1/1 i året til 31/12 i året.

Blandt disse personer var der efter revisionsproceduren 75 % tilbage, for hvem oplysningerne om ansættelsesforhold i hele året, med stor sandsynlighed måtte antages at være i overensstemmelse med de faktiske ansættelsesforhold.

I en anden analyse (se Danmarks Statistik, 1991, s. 30) siges det endvidere: "Det synes nærliggende at antage, at oplysningssedlerne i et vist omfang ikke er udfyldt fuldt korrekt. I forhold til den faktiske ansættelsestid må det formodes, at der sker en overvurdering af ansættelsesperioden, fordi det er enklere blot at angive 'helårsbeskæftigelse' end at skulle finde og angive de korrekte datoer for start og afslutning på et ansættelsesforhold.".

På trods af de ovenfor beskrevne problemer, er det fundet nødvendigt at anvende oplysningerne fra oplysningssedlerne (CON-91) i analysen, for at kunne skabe en form for tidskonsistens. De eksempler på sammenligninger der præsenteres i det følgende, skal ses i lyset af disse problemer.

Et alternativ til anvendelse af oplysningssedlerne, er at anvende flere årsversioner af de registre og surveys man ønsker at sammenligne. Dette er ikke gjort i nærværende analyse, men vil indgå i det fremtidige arbejde i registerkontrolprojektet.

Udover ovennævnte problem, ved forsøg på at skabe konsistens i tid, skal beregbskonsistensen også uddybes. Som udgangspunkt er der væsentlig forskel i på den ene side at anvende et formaliseret begrebs- og klassifikationsapparat, som det gøres i det meste registerstatistik og på den anden side at "begrebsdefinere" og klassificere udfra svar i en survey. Ved evaluering af samme forhold (fx personers stilling eller branche), er der tale om dels de metoder der anvendes til at klassificere , dels troværdigheden i de interviewede personers svar samt kodningen af de oplysninger der gives.

Evaluering af branche- og stillingsoplysninger

I RAS-91 findes bla. oplysninger om arbejdsstyrkens branchetilhørsforhold samt hvilken stilling de har. Disse oplysninger, bliver der ligeledes spurgt om i TASU-91.

CON-91 anvendes som nævnt ovenfor og ved at supplere med oplysninger herfra, er man i stand til at afgrænse grupper af personer, for hvem der med større sandsynlighed burde være enslydende branche- og stillingsoplysninger i RAS-91 og TASU-91.

Til brug for analysen er alle personer, der indgik i TASU-91 blevet udtaget. Denne population er herefter opdelt på hvorvidt de deltog aktivt eller ej i undersøgelsen, hvor førstnævnte gruppe benævnes de aktive TASU-personer. Det er udelukkende denne gruppe, der indgår i analysen[6] .

Brancheoplysninger

Det første der skal undersøges er, i hvor stort omfang brancheoplysningerne i RAS-91 og TASU-91, for det vi kan kalde de stabile personer, stemmer overens. Med stabile personer menes personer, der indgår i såvel TASU-91 som RAS-91, og som har ét ansættelsesforhold (i flg. CON-91), der strækker sig over minimum den periode der ligger mellem disse (10. marts - 1. december 1991)[7] .

8707 personer opfylder stabilitetskravet. For disse personer er deres branchetilhørsforhold i flg. TASU-91 og RAS-91 udtrukket. I TASU-91 drejer det sig om spørgsmålet: "Hvilken slags virksomhed er De ansat i/arbejder De i?" og i RAS-91 om formålsbranchekoden FORMALBR. Hvad angår TASU-91 bliver interviewpersonens svar omkodet til en 5-cifret branchekode, som repræsenteres i variablen VIRK10A (samme klassifikation anvendes i RAS-91)[8] .

---

[6] I analysen inddrages bortfaldsproblematikken således ikke .

[7] Det skal bemærkes, at der ikke er sat som betingelse, at personerne skal være beskæftigede i flg. RAS-91.

[8] Det skal bemærkes, at i de tilfælde hvor personen indgik i TASU-90, er der mulighed for at markere for hvorvidt den branche, der blev oplyst her, stadig gælder.

Resultatet af sammenligningen på den 5-cifrede branchekode er, at for 5936 personer (68,2 %) stemmer branchekoden overens og for 2771 personer (31,8 %) er der tale om uoverensstemmende branchekoder. Sidstnævnte gruppe er så stor, at den bør analyseres nærmere.

I branchegrupperingen refererer 1. ciffer til hovedbrancheområde, hvorfor man kunne opstille den hypotese, at en del af forskellen i den 5-cifrede branchekode, blot er et udtryk for, at det kan være vanskeligt at placere personer på et så detaljeret niveau, mens en placering på 1. ciffer skulle passe. Derfor er 1. ciffer i de to branchevariable krydset for de 2771 personer.

Tabel 1: Sammenhængen mellem 1. ciffer i branchekoden i TASU-91 og RAS-91, antal personer. (Pct. er den procentvise andel af personer, for hvem der er overensstemmelse).

	1. ciffer i VIRK10A											
1. ciffer i FORMALBR	0	1	2	3	4	5	6	7	8	9	I alt	Pct.
0	-	9	-	36	1	10	23	2	7	23	111	-
1	-	**13**	-	9	-	5	3	2	2	16	50	26
2	-	-	-	-	-	3	1	1	-	1	6	-
3	-	10	2	**629**	1	27	57	4	18	28	776	81
4	-	-	-	2	**2**	5	4	-	-	8	21	10
5	-	5	-	32	-	**42**	8	4	8	12	111	38
6	-	11	-	148	1	15	**289**	13	35	75	587	49
7	-	2	1	8	-	9	9	**68**	3	16	116	59
8	-	5	-	38	-	7	23	6	**103**	31	213	48
9	-	25	-	43	4	13	33	8	42	**612**	780	78
I alt	-	80	3	945	9	136	450	108	218	822	2771	
Pct.	-	16	-	67	22	31	64	63	47	74		

Ved at sammenligne på branchekodens 1. ciffer ses det, at uoverensstemmelsen reduceres til 11,6 % (1013/8707), hvilket til dels støtter ovenstående hypotese. For næsten alle branchers vedkommende, ses ligeledes en koncentration af personer på diagonalen.

Der kan være flere årsager til den generelle uoverensstemmelse.
1. Manuel kodning af svar i TASU.
2. CON-oplysningernes kvalitet.
3. De interviewedes "viden" om deres branchetilhørsforhold.
4. Andelen af interviews hvor interviewpersonen ikke selv svarer.
5. Kvaliteten af brancheoplysningerne i Erhvervsregistret.
6. Andre årsager

Det er vanskeligt at estimere hvor meget de forskellige faktorer betyder.
Hvad angår pkt. 1. foregår der i TASU en kodning af interviewpersonernes svar. Kodningen går til dels ud på at omsætte en branchetekst til en defineret branchekode. En mangelfuld branchetekst kan således give anledning til en forkert branchekode. Desuden skal man være opmærksom på, at det er fors-

kellige personer, der foretager kodningen, og der vil erfaringsmæssigt opstå uoverensstemmelser på baggrund af dette. Ligeledes kan der være tale om enkelte fejlkodninger.

CON-oplysningernes kvalitet mht. periodeangivelser er allerede nævnt, men måske kunne man yderligere inddrage det resultat fra den interne brancheenquette, at der viste sig at være fejl i ca. 20 % af branchekoderne. Det skal dog bemærkes, at der for hovedpartens vedkommende var tale om små virksomheder.

Med hensyn til pkt. 4., bliver der i TASU spurgt om hvem i husstanden der er blevet interviewet. En overrepræsentation af interviews, hvor andre end interviewpersonen selv er interviewet, kan tænkes at være en indikator for usikkerheden i de anførte svar. Ud af de 2771 personer (interviews) viser det sig, at i 2196 tilfælde er det interviewpersonen selv der er interviewet, i 482 tilfælde er det interviewpersonens ægtefælle/samlever og i 93 tilfælde er der tale om andre personer. En lignende fordeling gør sig gældende for de 5936 personer for hvem branchekoderne stemmer overens. Der synes således ikke at være nogen væsentlig indikation af, at de uoverensstemmende branchekoder skyldes, at interviewpersonen ikke selv er interviewet.

Stillingsoplysninger

Med hensyn til kvaliteten af stillingsoplysningerne, er de 8707 personer fra analysen omkring brancheoplysninger opdelt efter deres svar i TASU-91 på spørgsmålet: "Er de lønmodtager, medhjælpende ægtefælle eller selvstændig ?", hvilket svarer til variablen STAT5. Af disse personer svarede 8407, at de var lønmodtagere, 47 svarede at de var medhjælpende ægtefælle og 253 at de var selvstændige.

De 8407 "lønmodtagere" er herefter opdelt på deres svar på spørgsmålet: "Hvad er Deres stilling mere præcist?". Svarene bliver kodet til en 3-cifret stillingskode, hvilket i TASU-91 svarer til variablen STIL9K. Det 1. ciffer i denne kode er sammenlignelig med den i RAS benyttede stillingsklassifikation, repræsenteret i variablen ARBSTILL, således at 1. ciffer i STIL9K, i klassifikationsmæssig henseende, svarer til 2. ciffer i ARBSTILL for værdierne 31 til 37.

For 70 %'s vedkommende stemmer 1. ciffer i STIL9K og ARBSTILL overens. Indenfor gruppen af funktionærer kunne man antage, at det kan være forbundet med vanskeligheder at afgøre, hvorvidt personerne er overordnede funktionærer, ledende funktionærer eller funktionærer i øvrigt. Hvis man slår disse tre grupper sammen øges overensstemmelsen for denne gruppes vedkommende til godt 90 %, hvilket til dels kan støtte antagelsen. Dog er der stadig en vis spredning over lønmodtagergrupperne. Således ses det blandt andet, at 146 ud af 1592 personer, der i flg. RAS-91 er ledende funktionærer, i TASU-91 svarer at de er ufaglærte arbejdere.

Tabel 2: Sammenhængen mellem 1. ciffer i stillingskoden fra TASU-91 og stillingskoden i RAS-91, for lønmodtagere i flg. TASU-91, antal personer. (Pct. er den procentvise andel af lønmodtagere, for hvem der er overensstemmelse).

	1. ciffer i STIL9K								
ARBSTILL	1	2	3	4	5	6	7	I alt	Pct.
11	1	4	5	5	3	1	-	19	-
12	1	3	9	12	1	5	-	31	-
31	**31**	4	1	-	1	11	1	49	63
32	22	**610**	170	52	1	7	-	862	71
33	10	138	**1047**	214	32	146	5	1592	66
34	14	74	212	**2124**	39	167	14	2644	80
35	3	5	62	38	**764**	136	3	1011	76
36	1	8	69	163	86	**1317**	3	1647	80
37	3	7	25	105	30	275	**5**	450	1
40	-	1	5	13	9	36	-	64	-
50	-	-	-	2	3	6	1	12	-
60	-	-	-	1	-	9	-	10	-
90	-	-	1	5	-	10	-	16	-
I alt	86	854	1606	2734	969	2126	32	8407	
Pct.	36	71	65	78	79	62	16		

En interessant uoverensstemmelse i tabel, er de 102 personer, der i flg. CON-91 skulle være helårsansatte, men som i RAS-91 optræder som enten fuldt ledige i uge 48, efterlønsmodtagere, pensionister eller øvrige udenfor arbejdsstyrken og som i TASU-91 svarer, at de er lønmodtagere. Forklaringen på tilstedeværelsen af de 102 personer peger blandt andet på de tidligere nævnte problemer i periodeoplysningerne på oplysningssedlerne.

Med hensyn til de 50 personer (19+31), der i flg. RAS-91 er selvstændige eller momsbetalere, men som i flg. TASU-91 er lønmodtagere af forskellig slags, består populationen, som tidligere nævnt, udelukkende af de personer, der i flg. TASU-91 har svaret at de var lønmodtagere. Der kan således både være tale om fejl i surveyen, som utilstrækkelige metoder til skelnen mellem lønmodtagere og selvstændige i registerstatistikken.

Resultaterne ovenfor illustrerer, at der er nogle relativt store uoverensstemmelser mellem stillingsoplysningerne i RAS-91 og TASU-91. I det følgende ses der på "sammenhængen" mellem disse uoverensstemmelser og de tidligere beskrevne uoverensstemmelser i brancheoplysningerne. I tabel 4 og 5 nedenfor er de 8407 lønmodtagere, der indgik i tabel 3, opdelt på hhv. de personer for hvem brancheoplysningerne ikke stemmer overens og de lønmodtagere for hvem brancheoplysningerne er ens.

Tabel 3: Sammenhængen mellem 1. ciffer i stillingskoden fra TASU-91 og stillingskoden i RAS-91, for lønmodtagere i flg. TASU-91, med FORMALBR ≠ VIRK10A, antal personer. (Pct. er den procentvise andel af lønmodtagere, for hvem der er overensstemmelse).

	1. ciffer i STIL9K								
ARBSTILL	1	2	3	4	5	6	7	I alt	Pct.
11	1	3	5	5	3	1	-	18	-
12	1	2	7	12	1	5	-	28	-
31	**15**	2	-	-	-	3	1	21	71
32	12	**171**	51	16	-	3	-	253	68
33	3	46	**242**	52	13	31	1	388	62
34	4	35	71	**466**	12	67	10	665	70
35	1	3	25	10	**309**	51	3	402	77
36	-	3	28	39	35	**481**	3	589	82
37	2	4	10	40	14	106	**2**	178	1
40	-	1	5	13	9	36	-	64	-
50	-	-	-	2	3	6	1	12	-
60	-	-	-	1	-	9	-	10	-
90	-	-	1	5	-	10	-	16	-
I alt	39	270	445	661	399	809	21	2644	
Pct.	38	63	54	70	77	59	10		

Ud af 2644 lønmodtagere for hvem brancheoplysningerne ikke stemte overens, er der 64 % for hvem stillingsoplysningerne er overensstemmende. Blandt gruppen af funktionærer, bliver antagelsen fra tabel 3 omkring det vanskelige ved at opdele på de tre funktionærgrupper delvist støttet, med en overensstemmelse på godt 85 %.

Af tabel 5 ses det, at den samlede overensstemmelse mht. stillingsoplysninger er 73 %. Dette er således væsentligt bedre, end for gruppen af lønmodtagere, for hvem brancheoplysningerne ikke stemte overens. For den samlede gruppe af funktionærer ses det samme billede, som i de foregående to tabeller, nemlig at overensstemmelsen øges betydeligt, i dette tilfælde til godt 92 %.

Tabel 4: Sammenhængen mellem 1. ciffer i stillingskoden fra TASU-91 og stillingskoden i RAS-91, for lønmodtagere i flg. TASU-91, med FORMALBR = 1. ciffer i VIRK10A, antal personer. (Pct. er den procentvise andel af lønmodtagere, for hvem der er overensstemmelse).

	1. ciffer i STIL9K								
ARBSTILL	1	2	3	4	5	6	7	I alt	Pct.
11	-	1	-	-	-	-	-	1	-
12	-	1	2	-	-	-	-	3	-
31	**16**	2	1	-	1	8	-	28	57
32	10	**439**	119	36	1	4	-	609	72
33	7	92	**805**	162	19	115	4	1204	67
34	10	39	141	**1658**	27	100	4	1979	84
35	2	2	37	28	**455**	85	-	609	75
36	1	5	41	124	51	**836**	-	1058	79
37	1	3	15	65	16	169	**3**	272	1
40	-	-	-	-	-	-	-	-	-
50	-	-	-	-	-	-	-	-	-
60	-	-	-	-	-	-	-	-	-
90	-	-	-	-	-	-	-	-	-
I alt	47	584	1161	2073	570	1317	11	5763	
Pct.	34	75	69	80	80	63	27		

Evaluering af ledighedsoplysninger

På baggrund af et uddrag fra Arbejdsmarkedsstyrelsens centrale register for arbejdsmarkedsstatistik (CRAM), ses der i det følgende på ledighedsoplysningerne for de personer, der i løbet af 1991, har været berørt af ledighed (LEDR-91) og som samtidig har deltaget aktivt i TASU-91

Med hensyn til ledighedsoplysninger i LEDR-91, optræder tidskonsistensproblemet ikke, da oplysningerne genfindes i præcis de uger, hvor personerne er blevet spurgt i TASU-91 - den såkaldte referenceuge.

I alt er der 3976 aktive TASU-personer, der optræder i LEDR-91. Disse er opdelt på ledighedsgrader i referenceugerne for TASU-91, dvs. de uger hvor personerne rent faktisk bliver interviewet. En yderligere opdeling er foretaget mht. deres svar i TASU-91 på spørgsmålet: "Hvad er deres nuværende stilling ?", repræsenteret i variablen HBESK1.

Tabel 6: Aktive TASU-personer der optræder i LEDR-91, fordelt på ledighedsgrad samt HBESK1 fra TASU-91.

	Ledighedsgrad i referenceuge i TASU-91				
	1	[0.5-1[	]0-0.5[	0	I alt
HBESK1					
1. Erhvervsarbejde	213	77	123	1757	2170
2. Værnepligtig	1	-	-	20	21
3. Arbejdsløs	1087	29	31	236	1383
4. Hjemmearb. uden erhvervsarbejde	9	-	-	16	25
5. Pensionist, efterløn, rentenyder mm.	10	-	-	13	23
6. Bistandshjælp, revalidering, langtidssyg	37	-	-	51	88
7. Skoleelev, studerende	25	-	-	234	259
8. Ude af erhverv i øvrigt	2	-	-	5	7
I alt	1384	106	154	2332	3976

Specielt 1. og 4. søjle er interessante i ovenstående tabel. Ud af 1384 personer, der i flg. LEDR-91 var fuldtidsledige i den uge hvor de blev interviewet, er der 213 personer (15,4 %), der svarer, at de har erhvervsarbejde. En mulig forklaring kunne være, at folk ikke bryder sig om at sige, at de er arbejdsløse. En anden kunne være, at de er midlertidigt arbejdsløse, men at de for det meste er i beskæftigelse.

Ud af 2332 personer, der i flg. LEDR-91 ingen ledighed har i den uge hvor de bliver interviewet, svarer 236 personer (10,1 %), at de er arbejdsløse. En forklaring kunne være, at personerne ikke opfylder de betingelser der gælder i arbejdsløshedsstatistikken for at være arbejdsløse, men at de selv opfatter sig som sådan. Her tænkes blandt andet på betingelsen om at være aktivt arbejdssøgende samt hvorvidt der modtages arbejdsløshedsdagpenge eller bistandsydelse.

Afslutningsvis er der forsøgt set på, om der er væsentlig forskel på kvaliteten af oplysningerne i TASU-91 og LEDR-91 (som illustreret i tabel 6), som følge af, for det første, hvem i husstanden der er interviewet og for det andet om interviewet er foretaget via telefoninterview eller postspørgeskema. Med hensyn til førstnævnte er der ingen forskel. Med hensyn til sidstnævnte forhold, er der en tendens til bedre overensstemmelse, når interviewet foregår via telefon.

Det fremtidige arbejde

De resultater der er fremlagt i dette papir, repræsenterer de første tiltag i registerkontrolprojektet. Med hensyn til arbejdsmarkedsstatistiske opgørelser på såvel register- som surveyform, er der stadig mange interessante aspekter at tage fat på. Dog er arbejdsmarkedsstatistikken kun et hjørne af den rigdom af informationer, der forefindes i Danmarks Statistik

I den nærmeste fremtid, vil det arbejdsmarkedsstatistiske område dog præge det arbejde, der udføres under registerkontrolprojektet og nedenfor beskrives kort hvilke opgaver man kunne tage fat på henholdsvis gøre lidt mere ud af.

Et aspekt der ikke er berørt, men som vil blive taget op i det fremtidige arbejde, drejer sig om anvendelsen af oplysninger i registre og surveys til at opnå supplerende information.

Som nævnt tidligere kan evaluering af registre via surveys i højere grad ses som en evaluering af de metoder, hvormed oplysningerne i registrene bliver dannet, end som en evaluering af selve oplysningerne. En mere indgående evaluering af disse metoders betydning for datakvaliteten, vil der i det fremtidige arbejde blive lagt mere og mere vægt på.

Afslutningsvis er det værd at nævne den nye struktur i arbejdsstyrkeundersøgelsen, hvor interviewing foregår løbende over hele året. Når resultater fra denne foreligger, og den registerbaserede arbejdsstyrkestatistik er opdateret til 1993, er det oplagt at tage sammenligninger af den karakter, der er præsenteret i dette oplæg, op igen. Dette ligger formentlig lidt længere ude i fremtiden end de ovenstående opgaver.

Sammenfatning

I dette afsnit er nogle eksempler på sammenligning af arbejdsstyrkeundersøgelsen 1991 på surveysiden og hhv. den registerbaserede arbejdsstyrkestatistik og arbejdsløshedsstatistikken fra 1991 på registersiden blevet præsenteret.

De oplysninger der er sammenlignet drejer sig om branche-, stillings- og ledighedsoplysninger. Specielt hvad angår branche- og stillingsoplysningerne fremgik det, at der her var tale om ikke uvæsentlige uoverensstemmelser. For ledighedsoplysningernes vedkommende var uoverensstemmelsen ikke så markant, men dette skal dog ses i lyset af den større grad af tidskonsistens der eksisterer på dette område.

Der er formentlig mange forklaringer på de uoverensstemmelser der fremgik af analysen, men det er vanskeligt at udpege de mest betydende. Dog har en forklaringsfaktor været fremhævet, nemlig betydningen af kvaliteten i de op-

lysninger om ansættelsesforhold - specielt periodeangivelser - på de oplysningssedler, der hvert år indsendes til Told- og skattestyrelsen.

Litteraturliste

Nielsen, P., 1987: "Markedsdynamik og arbejdsformidling", bilag 2 og 3, ATA-projektet, rapport nr. 7.

Danmarks Statistik, 1992:20: "Statistiske efterretninger, Arbejdsmarked, 1992:20".
Danmarks Statistik, 1992:24: "Statistiske efterretninger, Arbejdsmarked, 1992:24".
Danmarks Statistik, 1993:17: "Statistiske efterretninger, Arbejdsmarked, 1993:17".

Danmarks Statistik, 1991: "Anvendelse af oplysningssedlerne i IDA".
Danmarks Statistik, 1982: "Personstatistik på registergrundlag".

Danmarks Statistik, 1993: "Brancheomkodning af enhederne i erhvervsregistret", internt papir.

Bilag 1: Sammenhæng mellem variabelnavne og spørgsmål i TASU-91

HBESK1

1. Hvad er Deres nuværende stilling?

Erhvervsarbejde (selvstændig, medhj. ægtefælle, lønmodtager, elev el. lærling)	1	Gå til spm. 3
Værnepligtig	2	
Arbejdsløs	3	
Hjemmearbejdende uden erhvervsarbejde	4	
Folkepensionist, pensionist, efterløn, invalidepensionist, renteyder o.l.	5	Gå til spm. 2
På bistandshjælp, revalidering, langtidssyg	6	
Skoleelev, studerende	7	
Ude af erhverv i øvrigt	8	

IARB2

2. Har De arbejde ved siden af?

Ja	1	Gå til spm. 3
Nej	2	Gå til spm. 19

IARB3

3. Var De på arbejde i sidste uge?
(jf. referenceperiode på adressekortet)

Ja, arbejdede mindst 1 time	01	
Var midlertidig fraværende i hele ugen pga.:		
Dårligt vejr	02	
Stille perioder, arbejdsfordeling	03	
Strejke, lockout	04	
Uddannelse, kursus uden for arbejdspladsen	05	Gå til spm. 5
Sygdom, ulykke	06	
Barselsorlov	07	
Ferie	08	
Andre grunde (fx arbejde hver anden uge):	09	
Nej	10	Gå til spm. 4

STAT5

5. Er DE

Lønmodtager	1	Gå til spm. 7
Medhjælpende ægtefælle	2	Gå til spm. 9
Selvstændig	3	Gå til spm. 6

TIDBEGR7

7. Er De ansat i et tidsbegrænset job?

Ja	1	Gå til spm. 8
Nej	2	Gå til spm. 9

STIL9K

9. Hvad er Deres stilling mere præcist?
(Nøjagtig angivelse, fx. smedesvend ikke blot smed, kontorchef i skattevæsenet ikke blot kontorchef)

_____

_____

Gå til spm. 10

Hvis stillingen er den samme som på adressekortet, skriv 1 her:

VIRK10A

10. Hvilken slags virksomhed er De?
ansat i/arbejder De i?
(Fx. bryggeri, supermarked, cafeteria, el-installation, rederi, revisionsfirma, folkeskole, politi)

_____

_____

Gå til spm. 11

Hvis virksomheden er den samme som på adressekortet, skriv 1 her:

Evt. navn og adresse.

EJER11

11. Hvad laver man på virksomheden?
(Det væsentligste i virksomhedens arbejde, fx. fremstiller haveredskaber, Engroshandel med trikotage, edb-arbejde)

_____

_____

Gå til spm. 12

HELDEL12

12. Er De beskæftiget på:

Heltid	1	Gå til spm. 14
Deltid	2	
Korttid	3	Gå til spm 13.

RESP53

53. Hvem i husstanden er interviewet?

IP selv	1
IP's ægtefælle/samlever	2
Andre	3

Bilag 2: Oversigt over samhørende variabelnavne,- værdier og tekster.

1. ciffer i FORMALBR og 1. ciffer i VIRK10A

0 Brancheoplysning mangler/brancher uden indhold
1 Landbrug, jagt, skovbrug og fiskeri
2 Råstofudvinding
3 Fremstillingsvirksomhed
4 El-, gas-, varme- og vandforsyning
5 Bygge- og anlægsvirksomhed
6 Handel, restaurations- og hotelvirksomhed
7 Transportvirksomhed mm.
8 Bankvirksomhed, finansieringsvirksomhed, forsikringsvirksomhed, ejendomshandel og -administration, forretningsservice.
9 Tjenesteydelser

ARBSTILL (det fremhævede 2. ciffer svarer til 1. ciffer i STIL9K)

11 Arbejdsgivere
12 Momsbetalere
13 Arbejdsløshedsforsikret selvstændig
14 Årsafgrænset selvstændig
19 Anden selvstændig (uden ansatte)
20 Medhjælpende ægtefælle
31 Direktør
32 Overordnet funktionær
33 Ledende funktionær
34 Funktionær i øvrigt
35 Faglært arbejder
36 Ikke-faglært arbejder
37 Beskæftiget lønmodtager uden nærmere angivelse
40 Arbejdsløs (fuldt ledig i uge 48)
50 Efterlønsmodtager
60 Pensionist
90 Øvrige udenfor arbejdsstyrken

STAT5

1 Lønmodtager
2 Medhjælpende ægtefælle
3 Selvstændig

Surveys og registre - mulighederne for at integrere de 2 datakilder

Bo Møller

Der har måske i for lang tid været en tendens til at betragte statistik byggende på administrative registre og statistik byggende på surveys som 2 helt adskilte verdener.
Der bør derfor i fremtiden lægges mere vægt på at analysere ikke kun styrke og svagheder ved de 2 dataindsamlingsformer isoleret betragtet, men også på hvilken måde, de effektivt kan spille sammen og berige hinanden.

Fordele og ulemper ved de 2 typer af statistikkilder

Fordelen ved surveydata er især, at man meget direkte får mulighed for at styre dataindholdet, idet man i spørgeskema mv. kan stille spørgsmål om netop de emner, man ønsker belyst. Hertil kommer, at surveydata - især hvis moderne teknik i form af CAPI, CATI mv. anvendes - normalt vil kunne bearbejdes og offentliggøres meget hurtigt i forhold til registerdata, hvor man er afhængig af måske årlige registerdannelser.
De største ulemper ved survey-baseret statistik er, at det er en ret omkostningskrævende dataindsamling, at man uvægerligt støder ind i stikprøveskævheder og frafaldsproblemer, samt at de interviewede husstande, personer mv. måske ikke altid er i stand eller ønsker at give korrekte svar.

Fordelen ved registerbaseret statistik er især, at den normalt er stort set totaldækkende, og at omkostningerne er relativt små. Hertil kommer, at registerdata er behagelige at arbejde med i den forstand, at sammenkoblingen mellem forskellige registre sker forholdsvis let ud fra veldefinerede nøgler (CPR-nummer mv.)
Ulempen ved registerstatistik er derimod, at man er helt bundet op på de definitioner, administrativ praksis m.v., der anvendes af de registeransvarlige myndigheder. Hertil kommer, at Danmarks Statistik selv kommer forholdsvis langt væk fra dataindsamlingen, hvilken kan medføre manglende viden i institutionen om det præcise dataindhold, om datakvaliteten etc.

Begge typer af statistik har altså helt klare fortrin - det forekommer derfor oplagt, at man - ved på en fornuftig vis at udnytte begge formers stærke sider - vil kunne opnå et godt samspil, der vil kunne give endnu bedre statistik.

Muligheder for samspil mellem de 2 dataindsamlingsmetoder

På forhånd kan 2 former for 'samspil' eller integration tænkes:

1. Surveydata kan suppleres med registerdata.
2. Registerdata kan suppleres med surveydata.

Den første form for samspil har med stort held været anvendt i mange sammenhænge, mens den anden form - så vidt vides - ikke har været dyrket i Danmarks Statistik - hvilket forfatteren opfatter som en oplagt mangel.

Surveydata suppleres med registerdata

På dette område har vi i Danmarks Statistik adskillige erfaringer, hvor jeg vil holde mig til forbrugsundersøgelsen.
I forbrugsundersøgelsen fra 1987 samt i den nye forbrugsundersøgelse anvendes registerdata i vidt omfang som supplement til surveydata indhentet ved besøgsinterviews hos husstandene.
De registerdata, der anvendes, er hovedsagelig indkomstdata, oplysninger om erhverv mv., om uddannelse samt oplysninger om boligen (BBR).

Selve anvendelsen af registerdata er for så vidt teknisk simpel, idet forbrugsundersøgelsen er bekendt med CPR-numre og bopælskoder for alle de personer og adresser, der indgår i undersøgelsen. Det er derfor enkelt at foretage et direkte match med eksisterende registeroplysninger for herved at importere registerbaserede oplysninger til forbrugsundersøgelsen.
Når dette sker i et ret omfattende omfang er årsagen især, at det for det første sparer tid i selve interviewsituationen (og tid er i den sammenhæng en direkte omkostningsfaktor). For det andet vil vi på denne måde i en del tilfælde nok få mere præcise oplysninger - fx kan oplysninger om modtagen løn nok normalt fås bedre fra registrene end ved at spørge personerne selv.
På den anden side er det klart, at vi for en del emneområder ikke kan nøjes med at benytte registerdata, men er nødt til at stille supplerende interviewspørgsmål - enten fordi registerdata ikke findes, eller fordi de ikke kan specificeres i den fornødne grad.

I forbrugsundersøgelserne er det fx fundet vigtigt at opgøre husstandenes samlede indkomster og andre økonomiske tilgange. De store beløb i form af lønindkomst, offentlige pensioner og virksomhedsoverskud mv. hentes via registre, men en del indkomster kan ikke findes i de administrative registre. I interviewet med husstandene indhentes derfor oplysninger om mere 'perifere' indkomstarter, herunder bl.a. støtte fra familie, forsikringsudbetalinger, gevinster og indkomst ved sort arbejde mv. Disse indkomstarter tegner sig for alle husstande under ét for relativt beskedne beløb, men for enkelte husstande kan de være af afgørende betydning som forklaringsvariabel i relation til forbrug, levestandard mv.

2 problemer rejser sig i den sammenhæng: problemet omkring den præcise afgræsning af indkomstbegreberne i registre og survey samt periodiseringsproblemet. Umiddelbart er det ikke altid lige nemt at afgøre, om et indkomstbeløb faktisk allerede er dækket af registeroplysningerne eller ej. Det er ikke engang sikkert, at husstanden eller personen selv er ganske klar over det. Risikoen for, at der begås dobbeltregninger - enten pga. af fejlagtigt formulerede interviewspørgsmål eller pga. manglende viden hos respondenterne - er derfor altid til stede. Specielt har dette problem været følt omkring forskellige former for pensionsudbetalinger, hvor det er ganske svært for husstandene præcist at redegøre for, hvilken type udbetaling, der faktisk er tale om, og herunder om den er underkastet almindelig indkomstbeskatning (og derfor indgår i registrene) eller ej.
Hertil kommer periodiseringsproblemer, idet skattemyndighederne ikke altid anvender samme periodisering, som husstande eller personer finder naturligt.
På trods af disse problemer, findes metoden med at supplere interview med registeroplysninger dog helt klart som den mest effektive.

Det kan nævnes, at ved den senest gennemføre 1987-forbrugsundersøgelse er den samlede husstandsindkomst i sin mest detaljerede opgørelse underopdelt i 55 enkeltkomponenter eller indkomstarter. 18 af disse hentes direkte fra registre (indkomststatistikregister og bistandslovsstatistikregister). De registerbaserede indkomstarter dækker over 93 pct. af den samlede husstandsindkomst, mens de resterende 7 pct. af indkomsten - svarende til knap 19.000 kr. pr. husstand - er indhentet i selve forbrugsundersøgelsen. Men som nævnt, er dette beløb yderst ulige fordelt blandt de enkelte husstande.

Registerdata suppleres med surveydata

Så vidt vides, er det i Danmarks Statistik aldrig direkte forsøgt at opbygge en statistik, som grundlæggende bygger på registerdata, men hvor supplerende oplysninger fra surveys (eller andre kilder) direkte inddrages - jævnfør dog omtalen af den nye arbejdsløshedsstatistik i papiret "Imputering" af Lone Solbjerghøj. Jeg mener, at der på dette område er en del uudnyttede muligheder.

Inddragelsen af supplerende oplysninger kan ske på 2 måder:
På makroniveau ved i tabeller mv. at give et 'tillæg' byggende på supplerende surveydata, eller ved på mikroniveau at importere supplerende oplysninger til registrene.

Som et illustrerende eksempel til belysning af problemstillingen kan vi betragte indkomststatistikken, der jo lider af den åbenlyse mangel, at fx sorte indkomster (og andre ikke-skattemæssigt registrerede indkomster) ikke indgår. Ingen ved præcis, hvor stort beløb disse indkomster tegner sig for, eller hvordan de er fordelt - men noget ved vi jo alligevel. Samtidig er der ikke megen tvivl om, at disse ikke-registrerede indkomster for en del husstande har et ret betydeligt omfang, hvorfor en anvendelse af indkomststatistikken til fx belysning af velfærd er begrænset, når sådanne indkomster ikke inddrages.
For mig at se er det derfor en statistisk pligt at prøve at gøre noget ved problemet - også selv om løsningen aldrig kan blive ideel.

Fra andre undersøgelser (fx forbrugsundersøgelsen eller Rockwool-projektets undersøgelser) har man en beskrivelse af de sorte indkomster og deres fordeling blandt forskellige typer af husstande. Beskrivelsen er nok langt fra dækkende, men er alligevel værdifuld. Det vil derfor være naturligt og rimeligt at prøve at inddrage disse supplerende oplysninger om de sorte indkomster mv. i Danmarks Statistiks generaliserede indkomststatistik.

På makroniveau vil dette kunne ske ved 'blot' at tillægge et skøn over de sorte indkomster i alle eller udvalgte tabeller, idet skønnene vil skulle foretages for alle de kategorier af husstande og personer, som indgår i tabellerne. Det bemærkes, at disse skøn kan være ret komplicerede at foretage i det omfang, tabelkategorierne ikke svarer til de kategorier, der anvendes i de supplerende surveys.

Det vil derfor efter min mening nok være såvel lettere som mere 'elegant' at foretage imputationer på mikroniveau. Samtidig vil man ved den efterføl-

gende anvendelse af materialet ikke være bundet til på forhånd valgte grupperinger af husstande og personer mv., men vil kunne bruge materialet fleksibelt ved fx at kunne ændre gruppering i indkomstintervaller etc.
Et sådant projekt vil altså bestå i, ud fra informationer i de pågældende surveys, at importere skøn over sorte indkomster til enkeltindivider i registrene. Hertil må anvendes specielle metoder - fx en form for statistisk match. Proceduren kunne fx være følgende:
Surveymaterialet opdeles i et større antal relevante strata. Indenfor hvert stratum opstilles en fordelingsfunktion over forekomsten af og størrelsen af de sorte indkomster. Registerpopulationen opdeles i tilsvarende definerede strata. På statistisk tilfældig vis tilregnes nu et antal individer indenfor hvert stratum sorte indkomster på en sådan måde, at fordelingsfunktionen efter denne import er den samme i registerpopulationen som i surveypopulationen.
Resultatet heraf vil være, at man stadig på makroniveau vil have de bedst mulige tal for de sorte indkomster. Samtidig vil omgrupperinger af husstande og personer kunne ske fleksibelt, idet resultaternes kvalitet dog vil være helt afhængige af, hvor 'heldig' man har været ved den gennemførte stratificering - dvs. af hvor entydig og stabil fordelingen af de sorte indkomster kan bestemmes ud fra de tilgængelige baggrundsoplysninger. I det omfang, stratificeringen er direkte korreleret med forekomsten af de sorte indkomster, vil resultatet blive godt.

Som et andet eksempel på, hvorledes supplerende oplysninger kunne tænkes at berige registeroplysninger, kan nævnes indkomstopgørelsen for selvstændige. Her vil det sandsynligvis være muligt gennem en inddragelse af eksisterende regnskabsoplysninger at kunne opstille en mere fyldestgørende statistik.

Man kan diskutere, om resultatet af sådanne øvelser giver egentlig ny erkendelse, eller om der 'blot' er tale om en for de interesserede læsere mere hensigtsmæssig præsentationsform.
Det er klart, at opgørelsen af fx det sorte arbejde byggende på surveydata aldrig vil kunne blive bedre end surveymaterialet i sig selv tillader. Man skal derfor være ganske forsigtig ved anvendelsen af sådanne 'syntetiske' totaltal. Hvor detaljeret man på faglig forsvarlig vis kan tillade sig at offentliggøre sådanne syntetiske tal må afhænge af en vurdering af kvaliteten af den pågældende survey samt af en vurdering af hvor præcis modellen eller fordelingsfunktionen for de sorte indkomster er.

Konklusion

Det er min opfattelse, at fremtidige statistikudviklinger bør bygge på et intensivt samspil mellem forskellige datakilder. For at dette skal kunne ske effektivt skal en metodeudvikling gennemføres såvel på det indsamlingsmæssige område som på udviklingen af hensigtsmæssige metoder til statistisk match mv.
Samtidig - hvilket måske er nok så vigtigt - må der ske et vist holdningsskift. Det må erkendes, at registerstatistik har sine begrænsninger, at vi ikke blot skal 'lære at leve' med disse begrænsninger, samt at en omend kun delvis overvindelse af begrænsningerne (gennem inddragelsen af supplerende informationer, fx fra surveys) er bedre end intet at gøre.

Sammenhængende socialstatistik (et eksempel på et horisontalt integreret statistiksystem)

Jørn Daugård Pedersen

1. Statistikkens baggrund

Udviklingen af en sammenhængende socialstatistik skal ses dels på baggrund af samfundsudviklingen indenfor det sociale område i 80erne og dels på baggrund af den øgede anvendelse af administrative registre i statistikproduktionen.

Samfundsudviklingen

I takt med den øgede arbejdsløshed er der fra politisk og administrativt hold sat stadig større fokus på udviklingen i de sociale udgifter og det sociale apparats evne til at forhindre, at stadig større grupper bliver udelukket fra arbejdsmarkedet og primært forsørget ved hjælp af offentlige midler.

Forskellige undersøgelser fra midt i 80erne har påvist, at der syntes at være tale om en ophobning af langtidsforsørgede indenfor visse dele af det sociale system, samtidig med at de enkelte sociale delsystemer - i kraft af forskellige lovgivninger - i ringe grad var i stand til at koordinere de sociale opgaver. Der opstod et behov for at kunne analysere det samlede sociale ydelsessystem indenfor en sammenhængende referenceramme.

Statistikproduktionen

Danmarks Statistik har fra starten af 80erne etableret en række personorienterede statistikregistre på det sociale område, baseret på oplysninger i administrative, fælleskommunale afregningssystemer. Med overgang til registerbaseret socialstatistik - med personnummeret som nøgle, voksede mulighederne for at inddrage baggrundsoplysninger fra den øvrige personstatistik (f.eks. familie-, bolig- og beskæftigelsesforhold), og i arbejdet med de administrative registre udvikledes en særlig kompetence i behandling og produktion af pålidelige registerbaserede socialdata.

I 1983 iværksattes et forsøgsprojekt med tværgående opgørelser over modtagere af indkomsterstattende ydelser og de følgende år gennemførtes flere forskningsprojekter på servicebasis, der inkluderede såvel horisontale som vertikale registersamkøringer, ligesom der blev foretaget afprøvninger af tværgående sammenhænge mellem forskellige sociale hovedområder.

Ultimo 1986 kom den sammenhængende socialstatistik på arbejdsplanen og første offentliggørelse i NYT udkom januar 1992 - med data fra 1989. Sammenhængende data er oparbejdet fra 1984.

2. Statistikkens afgrænsning og formål.

Den sammenhængende socialstatistik er et *individbaseret socialstatistikregister* over modtagere af *indkomsterstattende ydelser*. En indkomsterstattende ydelse har til formål at sikre personens/familiens forsørgelsesgrundlag, når der optræder et indkomstbortfald som følge af ledighed, sygdom, invaliditet, alderdom eller anden social situation, der antaster personens/familiens forsørgelsesgrundlag. De indkomsterstattende ydelser er omfattet af en række lovgivningsbestemte ordninger der vedrører personer, der er helt eller delvist udelukket fra arbejdsmarkedet.

Formål

Statistiksystemet har til formål at belyse personens/familiens kontakt med de indkomsterstattende ordninger, målt på:

1. Ordning art. (Det administrative lovområde).
2. Ydelsens art. (Den lovgivningsbestemte ydelse).
3. Ordningens/Ydelsens omfang:
   3.1 Varigheden (Målt i dage - start til slut.)
   3.2 Størrelse. (Målt i kroner.)
4. Bevægelsen mellem de enkelte ordninger/ydelser. (Målt brutto/netto på 1.-3.)
5. Til- og afgang.

samt at beskrive personen/familien socialt, målt på:

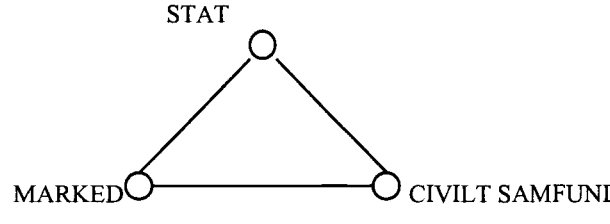
1. Demografiske variable. (Alder, køn, statsborger skab, civilstand, familiestatus.)
2. Markedsvariable. (Forsikringskategori, stillingskategori, indkomst.)

3. Teoretisk koncept.

3.1 Forsørgelse

Nøglebegreber er begreberne *indkomsterstatning* og *forsørgelse* idet begge begreber knytter sig til det overordnede begreb *indkomst*. Følgende figur illustrerer begrebssammenhængen:

Figur 1 **Forsørgelsestrekanten**



Marked

Figuren er en model over samfundets forskellige forsørgelsesarenaer. Den primære arena for forsørgelse finder sted indenfor *markedet*, hovedsagelig indenfor *arbejdsmarkedet*, hvor man som lønmodtager modtager *indkomst* til forsørgelse i form af *løn*, eller som selvstændig i form af *overskud af egen virksomhed.* Under ét er der her tale om indkomst, der traditionelt kategoriserе som *Primær indkomst.* En anden side af markedet omfatter *kapitalmarkedet*, hvor indkomster hentes i form af udbytte eller afkast af kapitalinvesteringer. Der er her tale om indkomster, der traditionelt kategoriseres som *Formueindkomst.*

Stat

Staten, her forstået som centralstaten, amter og kommuner under et, udgør en anden vigtig arena for forsørgelse. En del af statens indtægter i form af skatter og afgifter tilbageføres som sociale ydelse til borgerne efter specifikke lovbestemte ordninger. Herved får disse ydelser karakter af indkomst i forhold til modtageren. Denne type af indkomster kategoriseres traditionel som *Overførselsindkomst.* Visse former for overførselsindkomst udbetales som hel eller delvis erstatning for anden indkomst, *indkomsterstattende ordninger*, mens andre former har karakter af et supplement til anden indkomst, *tilskudsordninger.*

Civilt samfund

Det civile samfund, familie, slægt og lokalt miljø, spiller i dag stort set ingen rolle som arena for indkomst og forsørgelse. Tidligere har det civile samfund spillet hovedrollen i forsørgelsen af personer eller familier, der har været ramt af midlertidigt eller varigt indkomstsvigt. Der har f.eks. været oprettet frivillige, sociale hjælpekasser, og i landsbysamfundet har de rige gårdmænd stået for underholdet af sognets fattige, mens ældre borgere har kunnet indgå aftægtskontrakter. På det seneste har der fra politisk hold været sat større fokus på muligheder af, at det civile samfund overtog visse af statens sociale opgaver, hvorfor denne forsørgelsesarena stadig må indgå i det teoretiske koncept.

3.2 Indkomsterstattende ydelse.

Udfra ovenstående betragtninger kan vi afgrænse de indkomsterstattende ydelser til først og fremmest at omfatte den del af de statslige overførselsindkomster, der vedrører de indkomsterstattende ordninger. På markedet findes der imidlertid også en række ordninger, der tager sigte på indkomsterstatning, nemlig de private og kollektive pensionsforsikringsordninger. Det er her tale om individuelle, kontraktmæssige ordninger mellem personen og det enkelte selskab, idet ordningerne dog er underlagt visse overordnede, generelle regler, udstukket af Finanstilsynet, herunder regler vedrørende udbetalingsforhold. Disse ordninger vil som løbende pensionsudbetalinger være omfattet af begrebet indkomsterstattende ydelser. Andre former for formueindkomst vil ikke være omfattet.

3.3 Indkomsterstattende ordning.

De enkelte ordninger

Det er herefter muligt at identificere følgende indkomsterstattende ordninger:

1. Markedsfinansierede pensionsordninger.

samt følgende overvejende statsfinansierede ordninger:

1. Dagpengeordninger vedr. arbejdsløshed, sygdom og barsel.
2. Aktiverings- og uddannelsesordninger.
3. Bistandslovens forsørgelsesordninger (Kontanthjælp eller revalidering).
4. Sociale pensionsordninger.
5. Efterlønsordningen.
6. Tjenestemandspensionsordningen.

For flere af ordningerne gælder det, at en person på et givet tidspunkt kan være tilknyttet andre ordninger samtidigt, og over tid kan personen vandre mellem flere forskellige ordninger.

Sammenfald mellem ordningerne

Der er i første tilfælde tale om ordninger, som principielt ikke udelukker hinanden. F.eks. kan modtagere af tjenestemandspension også samtidigt oppebære social pension, eller personer over 67 år kan være i beskæftigelse og ved sygdom modtage dagpenge, samtidig med at de modtager reduceret folkepension (frem til det fyldte 70 år).

Der kan også være tale om ordninger, der principielt udelukker hinanden, men som administrativt optræder som sammenfaldende ordninger. Ikke-forsikrede ledige vil dels være registreret som ledig og dels modtage kontanthjælpsydelser. Personer, for hvem der er indledt en førtidspensionssag, vil under sagens behandling være tilknyttet den ordning, hvorfra sagen indledes, idet der herefter vil ske en administrativ omkontering af det udbetalte beløb til førtidspensionssystemet, hvis sagen afsluttes med en tilkendelse af førtidspension.

Vandring mellem ordningerne

Der er tale om en udbredt vandring mellem de enkelte ordninger - og indenfor den enkelte ordning, en vandring mellem specifikke ydelser. Visse vandringer opstår som logisk følge af lovgivningen, andre som følge af ændringer i ydelsesmodtagerens sociale situation og endnu andre er en følge af administrative forhold.

4. Virkelighedsmodel

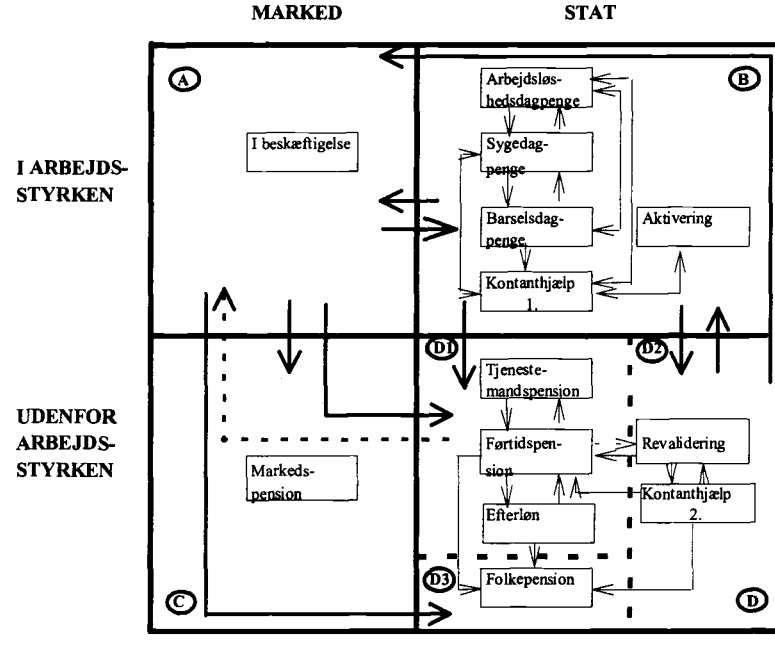
På tværs af ordningernes finansieringsgrundlag, er adgangen til de enkelte ordninger bestemt af personens markedsplacering.

I nedenstående figur 2 er de enkelte indkomstordninger ordnet i et skema, der dels placerer ordningen indenfor den forsørgelsesarena - MARKED (A, C) eller STAT (B,D) - hvor ordningen finansieres, og dels efter personens tilknytning til arbejdsmarkedet, I ARBEJDSSTYRKEN (A,B) hhv. UDENFOR ARBEJDSSTYRKEN (C,D).

Vandringerne på tværs af forsørgelsesarenaerne er illustreret ved de vandrette, fede pile, mens vandringerne ind og ud af arbejdsstyrken er illustreret

ved de lodrette fede pile. Vandringerne mellem de enkelte ordninger er illustreret ved de tynde pile. De stiplede pile angiver teoretisk gyldige vandringer. Mulige sammenfald af ordninger fremgår ikke af figuren.

Figur 2 **Samfundsmæssig placering af de indkomsterstattende ordninger**



Modellen beskriver fire samfundsmæssige *platforme* for forsørgelse.

A omfatter arbejdsmarkedet, hvor forsørgelsesgrundlaget er primær indkomst.

B omfatter indkomsterstattende ordninger, der som forudsætning bygger på forekomst af et *midlertidigt* indkomstbortfald indenfor arbejdsmarkedet som følge af arbejdsløshed, sygdom eller barsel. Indkomsterstatning i form af kontanthjælp omfatter her ydelser til ikke-forsikrede og registrerede arbejdsløse, her benævnt Kontanthjælp 1. Indenfor B befinder sig også de forskellige former for midlertidige indkomsterstatninger, der følger af aktiveringsordninger.

C omfatter de markedsfinansierede *varige* indkomsterstattende ordninger, først og fremmest de kollektive pensionsordninger, der kommer til udbetaling i forbindelse afgang fra arbejdsmarkedet som følge af alderdom, eller under særlige forhold ved udbetaling som en førtidspensionslignende ydelse.

D omfatter såvel *varige* som *midlertidige* indkomsterstattende ordninger. D1 omfatter pensioner, der optræder som principielle *varige* indkomsterstattende ordninger. Ordningerne betragtes som varige i den forstand, at der indenfor en given ordning primært vil foregå vandring til andre former for varige ord-

ninger. D2 omfatter *midlertidige* indkomsterstattende ordninger. Her er dels tale om indkomsterstatning i form af revalidering, hvis primære sigte er rekvalificering af ydelsesmodtageren, så denne senere vil blive i stand til at forsørge sig selv på markedet. Dels er der tale om indkomsterstatning for personer, der akut er ramt af et forsørgelsessvigt, og som ikke er omfattet af andre indkomsterstattende ordninger. Ordningen er her betegnet kontanthjælp 2. Denne omfatter desuden personer, der står overfor en umiddelbar overgang til andre indkomsterstattende ordninger. Endelig omfatter D3 folkepension, der som aldersbetinget *varig* ordning omfatter alle over 66 år. En vandring ud af denne ordning vil alene finde sted som følge af dødsfald eller udvandring.

6. Registermodel

Et flertal af de her nævnte indkomsterstattende ordninger findes som allerede etablerede, individorienterede statistikregistre, indsamlet og konstrueret med henblik på en belysning af de specifikke, isolerede lovområder og ydelsesordninger: Hvem og hvor mange får ydelser efter hvilke lovparagraffer ? I hvor lang tid modtages en given ydelse ? Hvor meget udbetales til hvem efter en given ordning ? etc.

Som det er redegjort for i det foranstående, er der indenfor det samlede område for indkomsterstatning imidlertid ikke tale om et simpelt 1 til 1 forhold mellem personen og den enkelte ordning, men derimod om et ofte særdeles kompliceret samspil mellem personen og de enkelte ordninger og ydelser, hvadenten dette samspil betragtes på et givet tidspunkt eller over tid. Det vil derfor alene udfra de enkelte statistikker, enkeltstående eller samlet, være vanskeligt at opnå noget totalt billede. Løsningen på dette problem vil være en integration af enkeltstatistikkerne i et sammenhængende socialstatistikregister.

Det sammenhængende statistiske koncept

Et integreret og sammenhængende socialstatistikregister over modtagere af indkomsterstattende ydelser må i princippet indeholde ensartede og sammenlignelige statistiske oplysninger indenfor samtlige ordninger, der vedrører begrebet indkomsterstatning. Den centrale enhed bør være personen, med cprnr. som nøgle, idet der herved åbnes mulighed for samling af personer i familieenheder, der er samfundets centrale økonomiske enhed.

Udover den statistiske belysning af det enkelte lovområde, bør det integrerede register kunne belyse de totale populationsmængder, foreningsmængder og fællesmængder samt vandringen mellem de enkelte områder, brutto- såvel som nettovandringer. Det integrerede register bør desuden kunne måle ressourcemængden, målt på varighed og udbetaling, og målingen bør kunne relateres til den enkelte indkomsterstattende ordning samt til det totale indkomsterstattende system, brutto såvel som netto.

Endelig bør det integrerede socialregister kunne relatere samtlige målinger til den specifikke platform, hvor ordningen hører hjemme.

Modifikationer

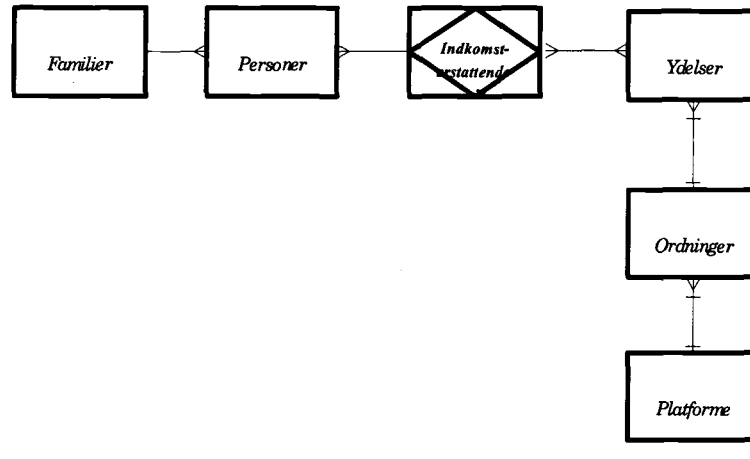
Oprindeligt har udviklingen af en registermodel for den sammenhængende socialstatistik bevæget sig omkring en mere pragmatisk tilgang til begrebet indkomsterstattende ydelser. Begrebet har hovedsagelig været bestemt med baggrund i , hvad der har kunnet belyses ved integration af eksisterende statistikregistre, dvs. man har først og fremmest kunnet inddrage overførselsindkomsterne. Spørgsmålet om platform for ydelsen har herved kunnet reduceres til et spørgsmål om ydelsens *totalitet* i forhold til modtageren. Ved totalitet forstås her ydelsens omfang, set i forhold til modtagerens mulighed for i fremtiden at kunne forsørge sig selv på markedet. En logisk klassifikation har derfor været en opdeling i *midlertidige* og *Varige* ordninger.

Platform i den sammenhængende socialstatistik består således i en opdeling af den indkomsterstattende ordning i *midlertidige* og *varige* ordninger. Der er derfor heller ikke tale om nogen opdeling af kontanthjælpen i forhold til modtagerens markedstilknytning (en sådan opdeling kan dog udledes af registret), og ordningerne kontanthjælp og revalidering tilhører derfor begge platformen for de midlertidige ordninger.

Med henvisning til figur 2 omfatter *midlertidige* ordninger B og D2, mens *varige* ordninger omfatter D1 og D3.

Dette forhold vil blive nærmere diskuteret i det afsluttende afsnit.

Figur 3 **Enhedsgrafen**



Enhedsgrafen angiver relationerne mellem registrets enheder, som indbyrdes er forbundne med nøgler. Der er tale om følgende enheder: Familier, personer , ydelser, ordninger og platforme.

Flere personer kan optræde i samme familie, som på sin side kan modtage flere ydelser. De enkelte ydelser knytter sig i første omgang til personen, idet flere personer hver især kan modtage flere ydelser. De enkelte ydelser er underlagt er underlagt en specifik ordning, idet den enkelte ordning kan bestå af flere ydelser. Endelig knytter de enkelte ordninger sig til en given platform, der på sin side kan rumme flere ordninger.

7. Registerdannelsen

Datagrundlag

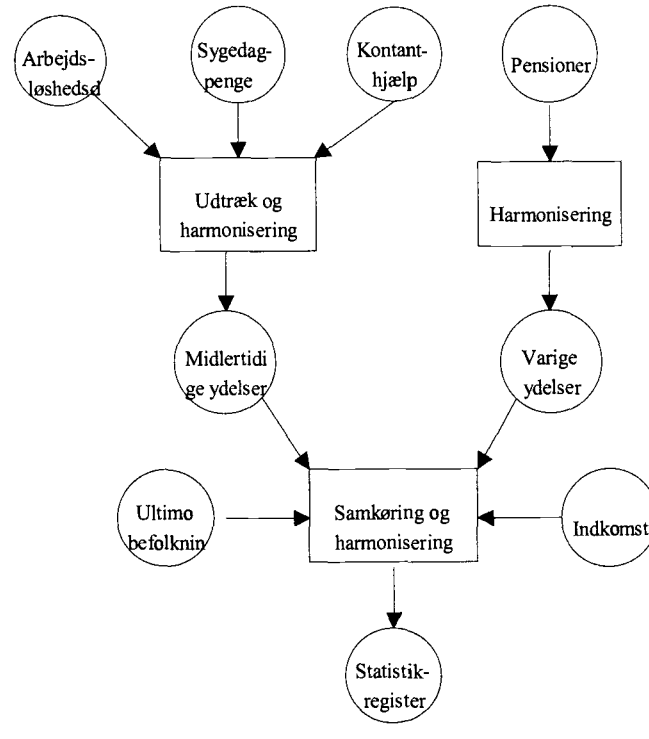
Datagrundlaget er eksisterende statistikregistre, der forbehandles og integreres i et sammenhængende registersystem. Med undtagelse af Aktiveringsordningerne og Markedspensionsordningerne findes der i Danmarks statistik personorienterede statistikregistre, der omfatter samtlige de indkomsterstattende ordninger, der er afbildet i figur 2. Hvad angår Markedspensionsordningerne er det med baggrund i Indkomststatistikregistret muligt af afgrænse en persongruppe, der modtager markedspensioner, men testkørsler har vist, at det er forbundet med stor usikkerhed at bringe data på en form, så de kan indgå med samme detaljeringsgrad som de øvrige indkomsterstattende ordninger. I nær fremtid forventes det imidlertid, at data kan hentes direkte i form af administrative registerdata.

Danmarks statistik råder p.t. heller ikke over et statistikregister over aktiveringsordninger. Et sådant forventes imidlertid etableret med virkning pr. 1/1 1994.

Dataprocesser

For at kunne danne de tværgående grunddata, har det været nødvendigt at etablere et omfattende produktionssystem, som består i en trinvis *samkøring, harmonisering og datareduktion* af de indgående registre. Herigennem dannes først to delregistre, hhv. for varige og midlertidige indkomsterstattende ydelser, som så udgør grundlaget for at danne det samlede datasæt.

Figur 4 **Procesdiagram**



Beregning af ydelsernes varighed

Opnåelse af en *fælles måleenhed for ydelsernes varighed* forudsætter en del omregninger, da de indgående registre på dette punkt i særlig grad er uensartede: For arbejdsløse er oplyst en *ledighedsgrad pr. uge*, for sygedagpenge kendes de faktiske *dage* (datoer), for kontanthjælp og sociale pensioner er der tale om *hele måneder*, for efterløn og tjenestemandspension er informationerne umiddelbart indskrænket til *hele året*. Ved en række omformninger og opretninger, som delvis baseres på beslutningsregler udfra sandsynlighedsprincipper, og hvor oplysninger fra andre statistikregistre tages til hjælp, er det lykkedes at få alle varighedsdata bragt på formen: *antal dage pr. måned*. Ved registrets anvendelse tages der højde for, at summen af antal dage med ydelser for en person *netto* højst kan udgøre 30 dage pr. måned (360 dage/år). *Bruttovarighederne* kan derimod være betydeligt længere. Dette skyldes de før omtalte muligheder for ydelsessammenfald.

En mere detaljeret gennemgang af databehandlingen gennemgås i det følgende.

7.1 Arbejdsløshedsdagpenge.

Kildemateriale

Oplysningerne er hentet fra arbejdsløshedsstatistikregistret, som bygger på CRAM-registret. CRAM omfatter alle personer i arbejdsstyrken, der i årets løb har været registreret ledig i én eller flere perioder indenfor året. Der skelnes i CRAM mellem forsikrede ledige, der modtager dagpenge, og ikke-forsikrede ledige, der under ledighedsperioden modtager kontanthjælp. Ledighedsoplysningerne findes som ugeoplysninger, der bl.a. giver oplysninger om den lediges forsikringskategori (heltids-, deltids- eller ikke-forsikret), årsagen til ledigheden samt ledighedsgraden. Ledighedsgraden er et tal mellem 0 og 1, der angiver ugeledigheden som forholdet mellem ugens ledige timer og antallet af forsikrede (mulige) arbejdstimer. Ikke-forsikrede ledige behandles som heltidsforsikrede og får ledighedsgraden 1 for alle uger med fuld ledighed.

Afgrænsning

I forbindelse med udtrækket til den sammenhængende socialstatistik fraselekteres de personer, som i årets løb alene har modtaget feriedagpenge, uden anden ledighed indenfor året iøvrigt.

CRAM-registrets årsafgrænsning følger ikke et fuldt kalenderår, men et administrativt bestemt CRAM-ugeår, således at CRAM-året starter med de sidste uger af foregående år og tilsvarende slutter før udgangen af årets sidste uger.

Ved at bortskære ugeelementer, der tilhører det foregående år, og ved - gennem samkøring med følgende år - at tilføje ugeelementer, der tilhører kalenderåret, konverteres udtrækket til et "normalt" kalenderår. Konverteringen foregår ved anvendelse af en speciel konstrueret CRAM-kalenderalgoritme.

Harmonisering

Varigheden i dage beregnes på baggrund af ugeledighedsgraden, idet en ledighedsgrad på 1 fastlægges til 7 dages ledighed. Månedsledigheden beregnes ved at samle ugeelementer, der tilhører samme kalendermåned. Ugeledigheden i de uger, der spænder over to måneder, fordeles proportionalt.

Ved samkøring med indkomststatistikregistret overføres oplysninger om udbetaling af arbejdsløshedsdagpenge, idet beløbsdata ikke indgår i CRAM-registret.

Personkredsen er, med nævnte undtagelser, alle der indenfor året har været registreret ledige i en del af året eller hele året.

7.2 Syge-/barselsdagpenge.

Kildemateriale

Sygedagpengestatistikken hviler på årlige personoplysninger fra Det fælleskommunale sygedagpengesystem. Personkredsen omfatter alle, der som følge af indtægtstab ved uarbejdsdygtighed grundet sygdom, graviditet, fødsel eller adoption, har modtaget ydelser via dagpengesystemet.

En del af dagpengeperioden finansieres af arbejdsgiveren, mens resten finansieres af staten. Statistikken omfatter alene den statsfinansierede del, dvs. kun personer, der modtager ydelsen over dagpengesystemet, mens der ikke indgår oplysninger vedrørende den såkaldte arbejdsgiverperiode.

Arbejdsgiverperiodens længde - den første del af fraværsperioden - har været fastlagt forskelligt for offentligt og privat ansatte, og periodens længde har varieret. Det betyder, at den persongruppe, der er omfattet af statistikken de enkelte år, har skiftet, hvilket påvirker tallenes sammenlignelighed over tid. Den mest radikale ændring fandt sted pr. 1. april 1990, hvorefter arbejdsgiverperioden for offentligt ansatte dagpengemodtagere kom til at omfatte hele fraværsperioden, hvilket betyder, at offentligt ansatte personer fra denne dato ikke indgår i statistikkens opgørelser. (Arbejdsløse, tidligere offentlig ansatte, der bliver syge, indgår dog stadig).

Afgrænsning

I det personorienterede udtræk fra dagpengeregistret er foretaget en opdeling i to hovedsagsarter, sygdom hhv. barsel, fødsel eller adoption. Personer, registreret med udbetalingsperiode i året, men med fraværsperiode det foregående år, indgår ikke i opgørelserne, ligesom personer, der er registreret som modtager af efterbetaling eller som tilbagebetaler af for meget udbetalt ydelse, heller ikke indgår.

Harmonisering

Varighedsmålet, antal dage med sygedagpenge indenfor måneden, beregnes udfra oplysninger om start- og slutdato for udbetaling af ydelse indenfor en given sagsart, idet en måned som nævnt er fastlagt til 30 dage. Der foretages en optælling af varigheden for hver af de to hovedsagsarter: Sygdom hhv. barsel, fødsel eller adoption.

7.3 Kontanthjælp/revalidering.

Kildemateriale

Kontanthjælpsstatistikregistret omfatter persongruppen, der har modtaget ydelser efter bistandslovens kap. 9 og 10. Enheden er familier. Som selvstændige familier regnes ægtepar, ugifte i alderen 18 år og derover samt ikke hjemmeboende børn i alderen 15-17 år.

Afgrænsning

Fra kontanthjælpsregistret udtrækkes ydelsesmodtagere, der har modtaget ydelser til direkte forsørgelse, som erstatning for manglende eller reduceret indkomst. De enkelte ydelser har skiftet indhold fra år til år.

Modtagere af ydelser som tillæg til almindelig indkomst, erhvervet ved lønarbejde eller selvstændig virksomhed, f.eks. Hjælp til arbejdsmaskiner og Igangsætningsydelse, indgår ikke i udtrækket.

Harmonisering

Ydelserne under bistandslovsstatistikken konteres månedsvis, dvs. der findes oplysning om den måned, hvori en given ydelse administrativt konteres for en familie, men ingen oplysninger om ydelsens varighed. En konteringsmåned fastlægges til 30 dage, og der sondres mellem *kontanthjælp* i form af forsørgelsesydelser, og *revalidering* i form af revalidering eller uddannelse.

Det familiebaserede ydelsessystem betyder, at sager vedrørende ægtepar rent administrativt er registreret under én af ægtefællerne (den såkaldte "hovedperson"). Der er således ikke oplysninger om, hvem i familien, de forskellige ydelser er "beregnet" for.

I den sammenhængende socialstatistik er *begge ægtefæller regnet som modtagere* med samme varighed og ydelsesart. De udbetalte ydelser er fordelt ligeligt mellem ægtefællerne.

7.4 Pensioner.

Kildemateriale

Oplysning om pensionsydelser stammer fra pensionsstatistikregistret, der rummer personoplysninger om *sociale pensioner*, førtidspension og folkepension, samt oplysninger om modtagere af *efterløn* og *tjenestemandspensioner*. Registret rummer oplysninger om alle personer, der i løbet af året har modtaget én eller flere pensionsydelser. Grundoplysningerne om sociale pensioner stammer fra det fælleskommunale pensionssystem og er meget detaljerede. Derimod er oplysningerne om modtagere af efterløn og tjenestemandspension mere summariske, og dataene stammer fra andre kilder. Beskrivelsen er derfor opdelt i to underafsnit:

7.4.1 Sociale pensioner.

Afgrænsning og harmonisering

Samtlige personer, der er omfattet af pensionsregistret, indgår i udtrækket. Ud fra oplysninger om start- og slutdato for modtagelse af pensionsydelse, beregnes ydelsens varighed. En måned er definitorisk sat til 30 dage.

Fra og med måneden efter en person fylder 67 år, henføres ydelsen til folkepension, mens ydelser før dette tidspunkt henføres til førtidspension.

I pensionsregistret findes alene oplysninger om det samlede årlige beløb for hver ydelsesart. For personer, der overgår fra førtidspension til folkepension, foretages en fordeling af udbetalingen på de to pensionsformer. Grundbeløbet videreføres uændret, mens der sker en fordeling af tillæg i de, der alene kan udbetales i forbindelse med førtidspension, og de, der alene kan udbetales i forbindelse med folkepension. Tillæg der kan udbetales i forbindelse med begge pensionsformer, fordeles proportionalt med varigheden på de to pensionsformer.

7.4.2 Efterløn og tjenestemandspensioner.

Afgrænsning

Efterlønsmodtagerne er afgrænset ved samkøring af CRAM og indkomststatistikregistret. I CRAM findes oplysning om overgang fra arbejdsløshed til efterløn, og i indkomststatistikregistret findes oplysning om den udbetalte efterløn.

Tjenestemandspensionistgruppen er alene afgrænset på grundlag af indkomststatistikregistrets oplysninger. Som følge af vanskeligheder med at afgrænse denne gruppe, har tallene gennem årene været noget svingende. En mere præcis afgrænsning gælder fra og med 1988.

Harmonisering

For efterløn og tjenestemandspension er oplysningerne om ydelsernes start- og sluttidspunkter og om den samlede varighed delvis konstrueret på basis af mekaniske beslutningsregler, og disse data er derfor underkastet en vis usikkerhed.

Start- og slutdato for gruppen af tjenestemandspensionister fastlægges med udgangspunkt i indkomststatistikregistrets oplysninger om udbetaling af tjenestemandspension. Personer, der har modtaget tjenestemandspension året før, gives startdato 1. januar. For øvrige modtagere af tjenestemandspension undersøges det, om der findes andre pensionsydelser med start i året. I disse tilfælde sættes startdato for tjenestemandspension lig startdato for anden pensionsydelse. Restgruppen fordeles proportionalt i forhold til udbetalingens størrelse, med startdatoer hen over årets måneder. Ved samkøring med registret over året døde sættes slutdato lig dato for død, mens øvrige får status som løbende modtagere ved årets slutning, dvs. får slutdatoen 31. december.

For gruppen af efterlønsmodtagere, der i årets løb overgår til efterløn fra arbejdsløshed, findes startdatoen med baggrund i CRAM-registrets oplysning om ugenr. for overgang til efterløn. Hvis personen modtog efterløn året før, sættes startdato til 1. januar. Restgruppen får startdato den 1. i måneden, hvor personen fylder 60 år. Slutdato sættes til udgangen af den måned, personen fylder 67 år. Hvis personen ifølge dødsfaldsstatistikregistret er død inden det fyldte 67. år, sættes slutdato lig dato for død.

7.5 Det samlede datagrundlag

De harmoniserede data om samtlige ydelser/ordninger, varigheder og udbetalte beløb mv. er kombineret i det samlede register om modtagere af indkomsterstattende ydelser.

Ved samkøring med befolkningsstatistikregistret er overført *individ- og familieoplysninger* opgjort som *ultimodata*, dvs. pr. 31/12 i året. Fra indkomststatistikregistret er overført data om *arbejdsstilling* og *samlet indkomst*.

8. Fejlsøgning og vedligeholdelse

Fejlsøgning

Som det er fremgået af det foranstående, synes stort set alle sammenhænge mulige, hvilket gør det vanskeligt at etablere nogen mekanisk fejlsøgningsprocedure. Visse sammenhænge, der på forhånd synes ugyldige, viser sig under specielle omstændigheder alligevel at være gyldige. Fx synes det på forhånd udelukket, at personer over 66 år, som automatisk registreres som folkepensionister, kan modtage kontanthjælp. Men det er faktisk muligt. Personer over 66 år, med en yngre ægtefælle der modtager kontanthjælp, vil i kraft af den mekaniske fordeling af kontanthjælpsydelse mellem ægtefæller, i registret optræde som kontanthjælpsmodtager. Eller fremmede statsborgere (familiesammenførte) over 67 år, der ikke er pensionsberettiget, kan under visse omstændigheder modtage kontanthjælp.

Fejlsøgningen består hovedsagelig i sammenligning af et antal kørselslogger, hvor hovedstrømme kortlægges under afviklingen af de enkelte step i dataprocessen, fra de første udtræk af grundregistrene frem til den endelige registerdannelse. Når registret er færdigdannet, produceres et antal analysetabeller, der danner grundlag for en analyse af fordelingerne i forhold til de tilsvarende fordelinger indenfor grundstatistikkerne. Fejl i grunddata vil det dog sjældent være muligt at korrigere for, og det vil ofte være tilfældet, at eventuelle fejl og unøjagtigheder i et grundregister først kommer til syne i forbindelse med integreringen med de andre registre.

Vedligeholdelse

Den registerinterne data-vedligeholdelse er ukompliceret, idet systemet er tilsluttet TIMES. En mere kompliceret problemstilling består i at indfange alle betydningsfulde ændringer i lovgivning, personkreds og data indenfor de enkelte indkomsterstattende ordninger. Studier af de direkte lovinitiativer kan give anledning til, at opmærksomheden rettes mod en bestemt problemstilling, men via kendskabet til et lovinitiativ vil det ikke altid være muligt at se, hvad indflydelse det vil få på det specifikke administrative register. I forbindelse med bestillingen af udtræk, der foregår på en særlig blanket, anmodes den registeransvarlige om at notere ændringer i forhold til tidligere år - evt. systematiske fejl - samt beskrive eventuelle konsekvenser af ændringen. Et sådant system er ikke ideelt, da de aktuelle årsdata i den sammenhængende statistik som regel vil være et år gamle i forhold til det aktuelle grundstatistikområde. Desuden vil det ofte være tilfældet, at der i perioden er kommet ny tællingsansvarlig for grundregistret, og denne har ikke nødvendigvis en

viden om tidligere ændringer. Systemet er således stærkt afhængig af kvaliteten og omfanget af den registerdokumentation der findes indenfor de enkelte grundområder. Den bedste metode ville bestå i at den ansvarlige for det integrerede register selv forestod etableringen af grundstatistikkerne. Den næstbedste, og mere praktisk realistisk, at have en løbende og tæt kontakt med personerne på de enkelte områder.

9. Behov for videreudvikling

En anden form for vedligeholdelse gælder registerstrukturen, og registrets muligheder for at producere statistikker og data, der lever op til de krav administration og samfundsforskere stiller.

Debatten om de sociale ordninger

I takt med den almindelige samfundsudvikling, og specielt som følge af den eksplosive vækst i overførselsindkomsterne, har der fra politisk hold været fremsat forskellige synspunkter omkring en nødvendig ændring af den sociale lovgivning. Et fællestræk har været, at der i debatten har været lagt vægt på en øget markedsrelatering af de sociale overførsler.

De forskellige udvalg og kommissioner der har været nedsat til at kulegrave området for indkomsterstattende ydelser, lægger i deres konklusioner op til en differentieret indsats overfor personer, der har opnået en markedstilknytning, og de personer der står udenfor eller kun har ringe markedstilknytning. Zeuthen-udvalget lægger op til et decentralt og behovsorienteret aktiveringssystem, der tilgodeser de regionale behov, hvor arbejdsmarkedets parter i højere grad står for finansieringen af ydelserne. Den tidligere firkløverregering har fremlagt et idékatalog, hvor der er tale om en todeling af dagpengeydelserne: En grundydelse der finansieres solidarisk og er ens for alle medlemmer, der bliver arbejdsløse, og en tillægsordning der bygger på en frivillig, forsikringsmæssig finansieringsordning. Denne todeling af ydelserne skulle ligeledes gælde pensionsordningerne. Her tales der om en grundydelse, tilstrækkelig stor til at dække de basale forsørgelsesbehov, mens en yderligere dækning må hvile på private eller kollektive forsikringsordninger.

Aktuelt har lovgivningen om orlovsordningerne yderligere skabt turbulens omkring begrebet *indkomsterstattende ordninger*. Er det den person som går på orlov, der er omfattet af en indkomsterstattende ordning, eller er det den arbejdsløse, der midlertidigt ansættes i en orlovsstilling, der er omfattet af en indkomsterstattende ordning? En eventuel lovgivning, der nærmer sig "Skraldemandsmodellen" vil yderligere problematisere begrebet *indkomsterstatning*.

Justering af registergrundlaget

Udfra en pragmatisk betragtning er det statistiske problem til at overse: *Persongruppen der modtager indkomsterstattende ydelser omfatter alle, der er omfattet af en indkomsterstattende ordning.* En sådan forenklet betragtningsmåde kan være tilstrækkelig, når det drejer sig om statistisk belysning af afgrænsede og enkeltstående lovområder, men når det drejer sig om integrerede registre som den sammenhængende socialstatistik, kompliceres forholdet.

Som det er beskrevet ovenfor, vil opbygningen af et integreret register bestå i en sammenkædning af administrative data, der knytter sig til hvert sit lovområde. De enkelte lovområder vil imidlertid ikke nødvendigvis kunne sammenfattes indenfor rammerne af et logisk sammenhængende lovkompleks, tværtimod. I konstruktionen af registret over modtagere af indkomsterstattende ydelser, har ordningens *platform* - de *midlertidige* hhv. *varige* ordninger - tjent som de logisk samlende enheder blandt rækken af særdeles umage ordninger. Der er her lagt vægt på ordningens grad af totalitet.

De tanker der har været fremme i diskussionen af en nødvendig revision af de sociale ordninger, har ikke taget sit udgangspunkt i ordningens grad af totalitet, men derimod i ordningens *finansieringsgrundlag*. Disse forhold og ønsket om komplettere den sammenhængende socialstatistik med andre og nye former for indkomsterstattende ordninger, kan pege på nødvendigheden af en bredere tilgang i definitionen og afgrænsningen af begrebet *platform*, end den i dag er defineret.

Et eksempel på en sådan mere omfattende definition, er behandlet her. Det er ikke nødvendigvis noget "guldæg", men behovet for en fleksibel registeropbygning, der kan opfange væsentlige ændringer i samfundet, er særlig stor når det gælder et integreret register som den sammenhængende socialstatistik.

Horisontal Integration

Lene Skotte

Den sidste spørgeskemabaserede folke- og boligtælling i Danmark blev gennemført med tællingsdato 9. november 1970.

De administrative, landsdækkende registre, der var opbygget i 1970'erne, udnyttedes i 1976 til gennemførelse af den første registerbaserede folketælling i Danmark . 1976-tællingen indeholder ikke boligoplysninger, da bygnings- og boligregistret først blev etableret efter denne tælling.

1976-tællingens oplysninger er dannet ved sammenkobling af befolkningsstatistikregistrets personoplysninger med oplysninger fra statistikregistrene baseret på kildeskattesystemet og erhvervsregistret. Nøglerne, der gjorde sammenkoblingen mulig, var personnummeret og CIR-nummeret.

Folke- og boligtællingen 1981 var den første tælling i verden, der udelukkende byggede på oplysninger fra administrative registre. Dens indhold svarede stort set til , hvad der indsamledes på skemaerne i 1970-tællingen.

Ved 1981-tællingen sammenkobledes op mod en halv snes grundlæggende statistikregistre og -moduler. De vigtigste registre i denne sammenhæng var:

-Det befolkningsstatistiske register
-Det boligstatistiske register
-Arbejdsklassifikationsmodulet
-Beskæftigelsesstatistikregistret
-Uddannelsesklassifikationsmodulet
-Byklassifikationsmodulet
-Fælleshusholdningsmodulet

Sammenkoblingsnøglerne var personnummeret, bopælsadressen og CIR-nummeret, suppleret med en eventuel arbejdsstedskode.

Alle oplysningerne i 1981-tællingen samledes i folke- og boligtællingsregistret.

Efter publiceringen af 1981-tællingen var der på næsten alle de områder, der traditionelt dækkes af en folke- og boligtælling, etableret årlige strukturbelysende statistikker. Behovet for folke- og boligtællinger hvert 5. år var således faldet bort.

EF-direktivet om harmonisering og synkronisering af de almindelige folketællinger forpligter medlemslandene til at gennemføre en tælling i 1990 eller i 1991. Den danske tabelsamling er blevet udarbejdet på samme måde som tællingen i 1981.

Vertikal integration

Otto Andersen, Lisbeth B. Knudsen og Søren Leth-Sørensen

Definition

Ved vertikal integration sammenbringes oplysninger fra forskellige tidsversioner af det samme register med henblik på at belyse forløb eller årsagssammenhænge over tid. Det er næsten altid således, at der anvendes flere registre, hvorfor den vertikale og den horisontale integration er snævert knyttet sammen.

Udviklingen

Ved starten af registerepoken i Danmarks Statistik har det naturligvis været afgørende at få dannet det enkelte årsregister mhp. statistikproduktionen for det pågældende år og det longitudinelle aspekt har ikke haft så stor tyngde. Det enkelte årsregister hviler på en måde i sig selv. Det er givet et synspunkt, der stadig er fremherskende.

EDB-teknikken har i starten sat ret snævre grænser for dels hvor store registrene rent praktisk kunne være (tænk blot på den tid det tidligere har taget at sortere et register omfattende hele den danske befolkning), så sammenkoblinger over tid har i starten været problematisk alene af tekniske grunde. Udviklingen er som bekendt gået meget stærkt og det er i dag muligt at opbygge meget omfattende longitudinelle registre (jf. Fertilitetsdatabasen og IDA), der uden alt for store vanskeligheder kan udnyttes effektivt.

De tekniske muligheder har støttet interessen især blandt forskere for at udvikle såkaldte multivariate modeller, som tillader inddragelsen af mange variable (horisontalt såvel som vertikalt) i analyserne. Alle blot nogenlunde omfattende statistikpakker, der i dag tilbydes (SAS, SPSS, BMDP mv.) har disse muligheder indbygget. I den kontakt Danmarks Statistik har med forskere er der næsten altid et ønske om datasammenstillinger til anvendelse i longitudinelle, multivariate analyser. Det er næsten, som om almindelige to/tredimensionale tabeller er gået helt af mode.

Problemerne

Nedenfor vil blive givet eksempler på de problemer, der rent faktisk opstår ved vertikal integration.

Indledningsvis skal der peges på nogle generelle problemer.

Årsregistrene lever som omtalt så at sige deres eget liv. Det er en betydningsfuld opgave at gøre det enkelte årsregister så perfekt som muligt, og gerne bedre end sidste år. Et eksempel herpå er udviklingen af Arbejdsklassifikationsmodulet (AKM), der fra sin spæde start i begyndelsen af 1980'erne har været under stadig udbygning. Gruppen "lønmodtagere uden nærmere angivelse" er bevidst søgt reduceret, hvilket er godt, men rejser naturligvis et problem, når det er den tidsmæssige udvikling, der er i fokus. AKM er naturligvis ikke det eneste eksempel. Der kan iøvrigt sikkert også findes eksempler på forringelser fra år til år.

Det for al statistik så klassiske problem med databrud eksisterer naturligvis også i den vertikale registerintegration. Klassifikationer, opgørelsesmetoder og det administrative datagrundlag ændres over tid og giver registereksperten store problemer. Et eksempel, som snart bliver aktuelt, er de nye stillings- og branchekoder, koder som er et integreret led i nogle af IDA's fundamentale definitioner. Sådanne ændringer medfører en stor arbejdsindsats for at bevare de longitudinelle registres værdi.

De longitudinelle registres datamængder har tilbøjelighed til at blive meget omfattende. Det er ofte vanskeligt (specielt for forskere) at forestille sig, at der kan undværes noget fra grundregistrene i den longitudinelle version. Synspunket er logisk nok, at man først ved om en variabel er væsentlig, når det er afprøvet i praksis. Det er imidlertid en erfaring, Danmarks Statistik har gjort, at variabelvalget *skal* være (stærkt) begrænset af hensyn til overskueligheden og *bør* dikteres af den teori, der er fremherskende på det statistik- eller forskningsområde, der er tale om. Man kunne i princippet tænke sig, at alle Danmarks Statistiks registre med personnr. blev sat sammen til ét stort register over tid. Enhver, der har arbejdet med registre ved, hvor urealistisk dette synspunkt er.

Endelig skal kravet til dokumentationen præciseres. Denne er altid vigtig, men set over tid stilles der særligt store krav. Ikke blot skal det enkelte års dokumentation samles sammen fra mange formentligt meget forskelligt opbyggede dokumentationssamlinger i de enkelte fagkontorer, men databrud, kodeændringer, skal nøje nedskrives. Ved dannelsen af IDA og Fertilitetsdatabasen har TIMES være en betydelig hjælp.

I det efterfølgende belyses problemstillinger nærmere ved to eksempler, nemlig de to forskningsdatabaser Danmarks Statistik har etableret i de senere år. Det drejer sig om Fertilitetsdatabasen og om IDA (Integreret Database for Arbejdsmarkedsforskning).

Fertilitetsdatabasen

Fertilitetsdatabasen er etableret i Danmarks Statistik i årene 1990-1992 med støtte fra Statens Samfundsvidenskabelige Forskningsråd. Ved etableringen af Fertilitetsdatabasen er dannelsen af populationen og specielt fastlæggelsen af forældre-barn relationen, foregået på en for Fertilitetsdatabasen specifik måde. Der blev derfor foretaget en bearbejdning af de henvisninger til forældre, der eksisterede i Befolkningsstatistikregisteret.

Det blev tidligt besluttet, som princip, at foretage grundig kontrol og nødvendige korrektioner i forbindelse med etableringen af populationerne, herunder forældrerelationen, mens oplysningerne i de "ydre" registre, d.v.s. de registre vedr. f.eks. uddannelse og arbejdsmarkedstilknytning, der blev leveret i færdig form fra andre kontorer i Danmarks Statistik, er taget for givet. Disse registre er således anvendt uden kontrol, idet der er afstået fra f.eks. en

konsistenscheck mellem forskellige årgange af samme register eller mellem forskellige registre, dækkende det samme kalenderår.

Problemer relateret til vertikal integration er derfor (indtil videre) kun identificeret i forbindelse med etableringen af populationen i Fertilitetsdatabasen.

Fertilitetsdatabasens opbygning og indhold

Fertilitetsdatabasen indeholder demografiske og sociale oplysninger om samtlige kvinder og mænd i fertil alder (defineret som 12-49 år for kvinder og 12-64 år for mænd), der har haft bopæl i Danmark i løbet af 1980'erne. Desuden er der oplysninger om, hvor mange børn, disse kvinder og mænd har fået, samt hvornår. Antalsmæssigt omfatter de årlige populationer ca. 1,3 mill kvinder og 1,8 mill mænd. Det antal (levendefødte) børn, de enkelte voksne havde på et givet tidspunkt og dermed ved en given alder, blev optalt ud fra henvisningerne fra børn til forældre i Befolkningsstatistikregisteret i Danmarks Statistik.

De demografiske oplysninger om de voksne er indhentet fra årligt færdigdannede registre og omfatter bl.a. køn, alder, civilstand, bopælskommune og statsborgerskab. *De sociale oplysninger*, som også er indhentet årligt, karakteriserer kvinders og mænds uddannelse, erhvervsarbejde og socioøkonomiske placering, indkomst, familie- og boligforhold samt omfanget af evt. sociale ydelser. Mens de demografiske oplysninger om kvinder og mænd findes for hver 1. januar i årene 1980 - 1989 findes de sociale oplysninger for perioden 1981-1988 [1].

Der er således tale om en omfattende samling af data, både med hensyn til antal omfattede personer og datamængden om hver person. Efter gennemførelsen af den første analyse er der i 1993 påbegyndt en opdatering til og med 1992. Herefter er det hensigten, at Fertilitetsdatabasen skal opdateres årligt.

Etableringen af en database af det omfang, som Fertilitetsdatabasen har, både hvad angår antal personer, antal år, antal registre og dermed antal variable og relationer mellem personer, indebærer store muligheder for så vidt angår udnyttelsen, men gav også en del problemer undervejs i etableringen.

Populationen

Populationen i Fertilitetsdatabasen består af en række årlige bestandsopgørelser (pr. 1.1.) af kvinder hhv. mænd i fertil alder i Danmark, uanset om de har børn eller ej. Den hidtil foretagne deskriptive analyse er således baseret på en række tværsnitsopgørelser og -beregninger [2].

For hver årlig population er der påført en række oplysninger, f.eks. om vedkommende person genfindes næste år, udvandrer eller dør. Indvandrere er indført i populationen pr. 1.1. i det år, de indvandrer, da de har mulighed for at føde et barn i landet det pågældende. år.

Karakteristika ved kvinder og mænd (de voksne) i Fertilitetsdatabasen kan således beskrives hvert kalenderår i 1980'erne. For børnenes vedkommende

[1] Fertilitetsudviklingen i Danmark i 1980'erne (1993)
[2] Knudsen (1993a)

er derimod kun medtaget "stam"-oplysninger vedrørende karakteristika ved barnet og om forældrene på barnets fødselstidspunkt.

Forældre relation

Basis for fastlæggelse af forældrerelationen har været de forældrehenvisninger, der findes på børnenes records i Befolkningsstatistikregistret. Børn, som kun har haft henvisning til en mor og/eller far, der ikke indgik i en af de årlige populationer, indgår ikke i Fertilitetsdatabasen.

Opgørelser ud fra BSR 1.1.1989 viste, at andelen af de enkelte fødselsårgange, der havde henvisning til mor (far) steg kraftigt for årgangene fra 1950'erne. Fra omkring årgang 1960-62 har mere end 90% henvisning til mor og/eller far.

Endvidere viser det sig, at andelen, der mangler henvisning til en mor er omkring 75% blandt personer, der er født udenfor Danmark, men næsten 0 blandt dem, der er født i Danmark (fra 1960 og frem).

De personer, der især mangler forældrehenvisning, er altså i stor udstrækning personer, der er indvandret som voksne.

Vertikal integration i Fertilitetsdatabasen

I etableringen af Fertilitetsdatabasen opstod der to hovedgrupper af problemer. Den samme persons personnummer kunne være forandret over tid og der var uforudset uoverensstemmelse mellem dokumentation og dataindhold i de ældste årgange af Befolkningsstatistikregisteret.

Rettelse af CPR-numre

Under etableringen af populationen opdagede vi, at der fra år til år kunne være foretaget nogle rettelser af personnummeret på samme person. Det er dog kun ændringer af børnenes personnumre, der er konstateret og bearbejdet.

Det blev nemlig tidligt besluttet at beholde de af forældrenes personnumre, der var gældende på et givent tidspunkt, dvs. som indgik i en given registerårgang. Det var nødvendigt at beholde det historisk gældende personnummer af hensyn til sammenkobling med de forskellige ydre registre fra samme år, hvorfra oplysningerne om den enkeltes sociale karakteristika blev hentet. Der er derfor slet ikke set efter evt. forandringer i de voksnes personnumre.

Der har dog formentlig alligevel været et vist tab af information, idet nogle personer kan have ændret personnummer mellem de to datoer, der er opgørelsestidspunkt for to forskellige registre.

Optælling af paritet

Når det har været så vigtigt at kontrollere børnenes personnumre, skyldes det, at antallet af børn, hver voksen har fået, den såkaldte *paritet,* er en væsentlig variabel til analyse af både fertilitetsmønster og -udvikling.

For hver voksen i populationerne er der (pr. hver 1.1. i perioden 1980-1989) optalt det antal børn, der er identificeret gennem Befolkningsstatistikregistret (1979->), suppleret med oplysninger fra Det medicinske Fødsels- og døds-

faldsregister (1973->). Dette tal er derefter anvendt som **paritet,** antallet af levendefødte børn.

Det var i forbindelse med denne optælling, problemet med ændrede personnumre viste sig.

Fremgangsmåden var følgende: materialet blev nu "vendt" fra børnene til de voksne. For hver kvinde i den samlede population, dvs. enhver kvinde, der blot forekom på ét tidspunkt i perioden, blev herefter dannet en record for hver gang, hun var registreret som mor til et barn. For de kvinder, der ingen børn havde, bestod recorden blot af deres eget personnummer.

Listen af records for hver kvinde blev dernæst kontrolleret for inkonsistens i børnenes fødselsdatoer. Hvert tilfælde, hvor en kvinde havde to enkeltfødte børn, der var født med mindre end 5 måneders interval, blev kontrolleret.

En stor del af de korrektioner, der herefter blev foretaget, hang sammen med at der var to records for det samme barn, hvoraf den ene var tidligere og mangelfuld i forhold til den seneste. Andre opstod pga. kombinationen af oplysninger fra de to registre: Befolkningsstatistikregistret og Det medicinske Fødsels- og dødsfaldsregister, hvorved man f.eks. kunne se, at den første fødselsdato, som var registreret, senere var blevet korrigeret (skrive eller tastefejl). Andre igen skyldtes korrektioner af indvandrede børns fødselsdatoer, som kunne finde sted op til flere år efter indvandringen og omfatte samtlige en kvindes børn. En tredie type var de rettelser, der fandt sted blandt adoptivbørn, der ikke ønskede at beholde et cpr. nr., der indeholdt information om at der var tale om en adoption, således som det var tilfældet i en årrække.

Ved at registrere fra hvilket original-register, den pågældende oplysning stammede fra, var det muligt at tidsfastlægge de enkelte records og derefter prioritere korrektheden. Med hensyn til oplysninger fra Statusbåndene (Befolkningsstatistikregisteret) blev den seneste (den nyeste) registrering altid valgt som den mest korrekte.

Da man principielt ikke kan indlægge en kontrol for afstanden mellem en mands børn, er rettelserne udelukkende gennemført i forbindelse med kvinderne, hvorefter de rettede oplysninger på barnets record er overført til mandepopulationerne.

Da rettelser kun har berørt børnenes personnumre, svarer de årlige populationer af voksne til tidligere kendte tal. Men rettelserne påvirker f.eks. opgørelser af kvindernes (og mændenes) alder ved fødslen af deres børn og dermed kalenderårsopgørelser over f.eks. aldersfordeling. Men ændringerne udgør dog ikke så stort et antal, at det vil være mere end minimale afvigelser.

Dataproblemer

De ældste bånd, der blev anvendt til at etablere populationerne, var Det medicinske fødsels- og dødsfaldsregister fra 1973 og Befolkningsstatistikregisteret fra 1979. Det førstnævnte register var veldokumenteret, mens der var problemer i dokumentationen for Befolkningsstatistikregisterets første årgange. Desuden viste det sig, at CPR-systemet, som dette register er baseret

på, indeholdt mange fejl i de første år - fejl, som først blev opdaget, når populationen blev etableret.

Den fejl, der især blev fundet, var, at der i mange tilfælde var byttet om på moderens og faderens personnummer (henvisning) på recorden for en person, der var "barn" i vores betydning (dvs. født efter 1941). Det var disse henvisninger, der skulle bruges til optælling af pariteten. Fejl af denne art kan opdages ved en simpel kontrol, pga. den forskellige opbygning af mænds og kvinders personnummer. Men selvom det er en "simpel" fejl, skal man - for at finde den - på forhånd være opmærksom på netop denne mulighed.

For Fertilitetsdatabasen betød opdagelsen af dette, at der måtte vælges en anden fremgangsmåde ved etableringen end den først planlagte, idet en række påtænkte maskinelle rutiner ikke kunne gennemføres som først planlagt.

Konsekvenser af de foretagne korrektioner

De foretagne korrektioner indebærer tidsmæssige besparelser for fremtidige brugere af Fertilitetsdatabasen og for personer, der har brug for sikre forældrerelationer. En sådan korrektion er ikke til at gennemføre uden et godt kendskab til registrene. Det er dog også nødvendigt at vurdere kvaliteten af det endelige materiale. Den beregnede paritet i Fertilitetsdatabasen er derfor sammenlignet med en opgørelse, alene baseret på oplysninger i det medicinske fødselsregister [3] .

Umiddelbart er der to store fordele i relation til fremtidig forskningsmæssig anvendelse ved etableringen af en database som Fertilitetsdatabasen: *For det første* den måske mest indlysende fordel, at så mange registre allerede *er* koblet sammen. *For det andet* den beskrevne opretning af populationerne, som sikrer en bedre kvalitet i materialet.

Der resterer dog stadig et problem, hvis man vil foretage en forløbsanalyse, hvor det er vigtigt, at de samme personer indgår år for år. De forandringer, der kan være foregået i personnumrene for de voksne, er som nævnt ikke korrigeret. Det ville være en god ide, hvis en eller anden form for korrigeret befolkningsregister blev dannet, indeholdende oplysninger om art af og tidspunkt for rettelser.

Konsistensen i oplysningerne fra de ydre registre er der som sagt ikke gjort noget ud af. Da Fertilitetsdatabasen endnu er relativt nyetableret har der været forholdsvis få serviceudtræk og da kun i en form, hvor evt. konsistens over tid i de forskellige sociale oplysninger ikke har spillet nogen væsentlig rolle og ikke har kunnet identificeres i de pågældende udtræk.

[3] Knudsen (1993b)

IDA - databasen

I det følgende omtales først enkelte centrale problemer ved konstruktion af en virksomhedsdatabase på grundlag af registerdata. Dernæst behandles erfaringer med lagring af oplysningerne i en relationsdatabase. Til slut påpeges med udgangspunkt i et konkret eksempel nogle af de særlige problemer, der opstår ved analyse af longitudinelle data.

Hvad er IDA-databasen?

IDA-databasen

I Danmarks Statistik er der i perioden 1988-1990 blevet opbygget en database på grundlag af registeroplysninger til brug for arbejdsmarkedsforskningen. Denne samling af data kaldes for IDA-databasen, en forkortelse af "Integreret Database for Arbejdsmarkedsforskning". Personer, ansættelser og virksomheder kan her følges i Danmark i perioden 1980-89. Formålet med oprettelsen af databasen har været at kunne foretage analyser af personers mobilitet mellem forskellige virksomheder og endvidere at kunne belyse virksomheders livsforløb (Danmarks Statistik 1991).

Enheder i databasen

Oplysningerne i databasen omfatter samtlige personer i befolkningen, samtlige ansættelser og virksomheder med lønnet beskæftigelse i perioden 1980-1989. Databasen opfatter således ca. 5 mill. personer, ca. 2,5 mill. ansættelser og ca. 200.000 virksomheder (arbejdssteder). Oplysningerne vedrører fortrinsvis situationen i slutningen af november måned i de enkelte år. Enhederne kan karakteriseres ved hjælp af en lang række variable. Der indgår i alt godt 300 variable.

Udgangspunkt vedr. data

Kildematerialet til IDA-databasen består af oplysninger fra en række forskellige administrative registre. Disse oplysninger er bearbejdet med henblik på anvendelse i en statistisk sammenhæng.

Variable i databasen

Nedenfor er givet nogle eksempler på typer af variable, der findes i databasen:

Personer:
køn, alder, civilstand, uddannelse, beskæftigelse, arbejdsløshed og indkomst.
Ansættelser:
Stilling, timeløn, anciennitet, af- og tilgang.
Arbejdssteder:
Branche, ansatte og lønniveau, identitet over tid.

For en oversigt over nogle af konkrete anvendelserne af IDA-databasen henvises til fx Arbejdsmarkedspolitisk årbog 1993, udsendt af Arbejdsministeriet, der som tema har: *Rundt om IDA - nye snitflader i arbejdsmarkedsforskningen.*

Identitetsproblemet

Et grundlæggende problem, der har skullet løses i forbindelse med opbygning af databasen, har været at afgøre, hvornår en virksomhed er den samme over tid (fra et år til det næste). På personsiden er det teknisk set uden problemer at følge personerne over tid ved at anvende det officielle personnummer som nøgle. For virksomhederne findes på tilsvarende måde et nummersystem af hensyn til bl.a. skat- og momsbetaling, men virksomhederne kan skifte numre, uden at der er tale om andre ændringer. Hvis man baserer en opgørelse af af- og tilgang af virksomheder i mellem 2 år på disse numre, vil man overvurdere antallet af virksomhedsnedlæggelser og -oprettelser.

Egenskaber ved virksomheder

At en virksomhed stadig er den "samme" fra det ene år til det næste indebærer, at der er visse kendetegn, der i den konkrete sammenhæng anses for centrale, der (stort set) er uforandrede fra det ene år til det næste. Imidlertid kan en virksomhed karakteriseres ved en række forskellige kendetegn eller egenskaber, der kan indgå ved fastlæggelsen af om en virksomhed stadig eksisterer i det følgende år. Problemet er derfor at vælge, hvad der er afgørende for, at der er tale om en "identisk" virksomhed.

Definition af "ny" virksomhed

Problemet kan eksemplificeres ved at tage udgangspunkt i, hvornår der foreligger en "ny" virksomhed. Hvis interessen består i at opgøre antallet af nye jobs, kan man lægge vægt på, at der tale om en virksomhed, der ikke eksisterede året før. En andet mulighed kunne være, at man var interesseret i at opgøre omfanget af nye ejere for at vurdere konkurrenceforholdene inden for bestemte brancher. En tredje mulighed kunne være, at man var interesseret i virksomhedernes beliggenhed, og herunder om der inden for et givet område var kommet nye virksomheder, der evt. kunne være tilflyttet fra et andet område. Dvs. at fastlæggelsen af, hvornår man vil hævde, at der foreligger en ny virksomhed i høj grad anhænger af den sammenhæng, som denne oplysning skal anvendes i (Baldwin, Dupuy & Penner 1992).

Fortolkningsproblemer i den virkelige verden

Dette problem med at fastlægge, hvad der definerer en (identisk) virksomhed, har også betydning uden for den statistiske verden. Fx skal virksomheder ved afskedigelser af et større antal ansatte, gives meddelelse til de lokale arbejdsmarkedsnævn. Problemet har her bl.a. været, hvad der skal forstås ved "større" afskedigelser i de tilfælde, hvor indskrænkningen skal ske på et lokalt arbejdssted, der indgår i et større firma (Arbejdsministeriet 1977). Tilsvarende kan der opstå fortolkningsproblemer i forbindelse med om overenskomster er gældende ved fx salg af virksomheder, hvor fagforeningerne ønsker, at overenskomster ikke kan ophøre på grund af fx proforma salg.

Virksomheden kan skifte ejer og flytte

I forbindelse med konstruktionen af IDA-databasen har hensigten som nævnt bl.a. været at kunne belyse arbejdspladsmobiliteten og virksomhederne livsforløb - forstået som nye arbejdspladser - og fastlæggelsen af kriterierne for om en virksomhed er den "samme" afspejles derfor heraf. Begge formål har medført, at en virksomhed både bør kunne skifte ejer og flytte, uden at der dermed er tale om en "ny" virksomhed.

Det operationelle plan

Fastlæggelsen af virksomhedernes identitet over tid er baseret på de lokale enheder - arbejdssteder - der kan indgå i et større firma. Identitetsproblemet er bl.a. blevet løst ved at inddrage oplysninger om sammenfaldet af ansatte på det enkelte arbejdssted fra det ene år til det næste[4] . En forudsætning for at anvende disse oplysninger har været anvendelse af EDB. Sammen med andre oplysninger om ejerskab, branche mv. indgår også dette personsammenfald ved afgørelsen af, om et arbejdssted er oprettet, fortsat eksisterende eller nedlagt.

På det operationelle plan indgår følgende oplysninger fra 2 på hinanden følgende år ved fastlæggelsen af virksomhedernes identitet over tid:

- ejerskab
- branche
- beliggenhed
- arbejdsstyrke

Den endelige regelsæt vedr. virksomhedernes identitet over tid er udformet som en trinvis procedure (se nærmere i Danmarks Statistik 1991).

Klassifikation af arbejdssteder

I forbindelse med den operationelle fastlæggelse af identiteten over tid opstod muligheden for at klassificere ændringer over tid på forskellige måde. Ved at anvende personsammenfaldet viste det sig, at en del "nye" virksomheder var kendetegnet ved at en større andel af de ansatte kom fra samme arbejdssted året før. Denne type oprettelse er blevet klassificeret som "oprettet ved udskilning fra et arbejdssted", hvis en række betingelser er opfyldt. Tilsvarende gør sig gældende for nedlagte virksomheder, der kan være "nedlagt via opsugning i et andet arbejdssted". Disse 2 kategorier gør det muligt at foretage analyser, hvor der kan tages hensyn til, om ændringer af denne type skal medtages eller ej. Dette er bl.a. af interesse i forbindelse med opgørelser af jobskabelse, som jeg vender tilbage til.

Konklusion

Fastlæggelse af virksomheders identitet over tid er et tilbagevendende problem, der ikke kan løses "én gang for alle", men den konkret problemstilling må være afgørende. Det er endvidere væsentligt at påpege den nære forbindelse mellem fastlæggelse af begreber og selve det EDB-mæssige udviklingsarbejde. Ved at operere med differentierede begreber er der skabt mulighed for at analyser med forskellige udgangspunkter. Dette betyder desuden, at forskere ikke uden videre kan arbejde med data af denne type, men at det kræver et vist arbejde at sætte sig ind i, hvorledes data er konstrueret. Der stilles derfor større krav til brugerne af data, og det er væsentligt, at der er et tæt samarbejde mellem den eller dem, der administrerer disse data og brugerne. Et fornuftigt samarbejde forudsætter også, at databaseadministrationen er i stand til at forstå de problemstillinger og data-ønsker, som brugerne kommer med. Det vil sige, at de, der står for databaseadministrationen, bør følge den teoretiske udvikling på et generelt niveau og at der er mulighed for at få kendskab til den metodiske og statistiske udvikling på området.

[4] I Canada har man anvendt samme grundlæggende ide, jf. Baldwin, Dupuy & Penner 1992.

Lagring af data: Database-struktur for IDA-databasen

Relationsdatabase

Oplysningerne i IDA-databasen er organiseret med data for hvert af årene i perioden 1980 til 1989 for sig. Endvidere er oplysningerne for det enkelte år opdelt i adskilte datasæt, der udgør en relationsdatabase[5] . Det indebærer at oplysninger refererer til forskellige typer af enheder, og at der til det givne datasæt i princippet kun findes oplysninger om denne type enhed. De væsentligste oplysninger i IDA-databasen vedrører følgende enheder:

- Personer
- Ansættelser
- Arbejdssteder

Fordele og ulemper

Fordelen ved at lagre oplysningerne på denne måde er for det første, at man undgår redundant information. Oplysningerne om arbejdsstederne findes fx kun i dette datasæt, og hvis man vil karakterisere personerne ved hjælp af oplysninger om arbejdsstedet skal disse oplysninger først findes i datasættet vedr. arbejdssteder. For det andet har denne organisering den klare fordel, at det er nemmere at orientere sig i det store udvalg af oplysninger. Man bliver ikke så let i tvivl om, hvorvidt en given variabel knytter sig til den ene eller den anden type enhed. Ulempen ved at lagre oplysningerne i denne form fremkommer, når man skal foretage et udtræk af data til en konkret problemstilling. Her vil man normalt skulle anvende oplysninger for at alle 3 typer enheder, og dette forudsætter derfor en række samkøringer, inden de ønskede data er blevet dannet. Her har anvendelse af programpakken SAS vist sig at være yderst hensigtsmæssig, fordi der kan foretages samkøringer af flere datasæt i en enkelt kørsel, og at der findes en Makro facilitet.

Variables konsistens over tid

En af de væsentlige fordele ved at etablere en database er, at man derved sikrer en høj grad af konsistens med hensyn til udvalget af variable og deres indhold. Det gøres ved så vidt muligt at bringe data i en fælles form inden de lægges ind i databasen. For brugerne af data er det afgørende, at der kan etableres en sammenhængende række data for flere år uden større besvær.

TIMES og dokumentation

Endelig skal det påpeges, at det er vigtigt for anvendelsen, at databasen og dens indhold er veldokumenteret. Her har dokumentationssystemet TIMES vist sig at være uundværlig på mange forskellige måder. Det ville være en nærliggende ide at udbygge samarbejdet med andre TIMES-brugere inden for afgrænsede områder som fx arbejdsmarkedsstatistik, hvor de forskellige tællingsansvarlige også blev ansvarlige for at vedligeholde dokumentationen af centrale variable over for andre ansatte i Danmarks Statistik.

Analyser over tid

Opgørelserne i IDA-databasen er baseret på oplysninger om situationen i slutningen af november måned i de enkelte år for såvel personer (Det økonomiske Råd 1992; Leth-Sørensen 1993) som virksomheder (Vejrup-Hansen

[5] Det er interessant, at Hamermesh (1990), der er professor i økonomi ved Michigan State University, i en fremtidsvision har beskrevet, hvordan der i løbet af 1990'erne blev etableret datamaterialer, hvor det var muligt at knytte oplysninger om personer og virksomheder sammen på tilsvarende måde som i IDA-databasen.

1993). Ændringer kan derfor opgøres ved at sammenligne oplysningerne for to tidspunkter. Der er dermed ikke tale om en kontinuert registrering, men om en sammenligning af to statusopgørelser.

Prospektive og retrospektive analyser

At personer og virksomheder følges over tid giver mulighed for foretage både prospektive og retrospektive analyser. Dvs. at der kan besvares spørgsmål eller problemstillinger, der typisk indebærer, at man ser "fremad": Hvor stor en andel af ufaglærte har været opadgående social mobile (Rohwer & Leth-Sørensen 1993)? Tilsvarende kan spørgsmål, der fx vedrører tilgangen af en bestemt gruppe personer - hvor man ser "tilbage" - besvares: Hvilken erhvervskarriere har personer, der deltager i voksenuddannelses-kurser i 1989? Endvidere kan der foretages opgørelser over en række år svarende til de 10 år, som databasen dækker.

Et eksempel fra virkelighedens verden: Job-creation and -destruction

Et område der har påkaldt sig stor interesse er opgørelser af det, man med begreber på engelsk kalder *job-creation* og *job-destruction* (Leonard 1987; OECD 1987,1993; Finansministeriet 1992; Ministeriet for Erhvervspolitisk Samordning 1993). Når man sammenligner beskæftigelsesniveauet i et land fra et år til det næste, er der normalt kun tale om relativt beskedne ændringer. Som supplement hertil er man begyndt at se på omfanget af nye jobs i nyetablerede virksomheder og i nettovæksten af jobs på virksomhedsniveau i allerede eksisterende virksomheder. Tilsvarende for virksomheder der forsvinder eller indskrænker. Det viser sig nu typisk, at de beskedne ændringer på makro-planet dækker over en meget betydelig jobomsætning på virksomhedsniveau opgjort på denne måde. Det skal bemærkes, at opgørelsen af jobomsætningen på denne måde er en minimumsbetragtning over den årlige jobomsætning. Da IDA-databasen netop indeholder årlige oplysninger om af- og tilgang af virksomheder og om antallet af beskæftigede, har det været nærliggende at anvende data herfra til opgørelse af jobskabelse og -nedlæggelse.

Jobomsætningen i den private sektor 1988-89

I det følgende skal nogle af analyseproblemerne ved anvendelse af longitudinelle data belyses med udgangspunkt i problemerne ved at opgøre jobomsætningen ved hjælp af IDA-databasen. Konkret drejer det sig om jobomsætningen i den private sektor, og som eksempel anvendes oplysninger fra 1988-89. Resultaterne fremgår af tabel 1.

Tabel 1. Job-creation and destruction at establishment level 1988-1989. Number of jobs. The private sector

			Establishment Status					
			Expansions	Births	Contractions	Deaths	Other	ALL
	'Employed 1988'	'Employed 1989'	Change no. employed	Change no. employed	Change no. employed	Change no. employed	Change no. employed	Change no. employed
Establishment Size								
1-19	693638	708363	80941	41109	-67881	-50197	10753	14725
20-99	541952	533579	34486	9038	-37258	-10003	-4636	-8373
100-499	315574	304125	13392	3651	-17805	-4553	-6134	-11449
500+	176327	166369	3569	2100	-4805	-2489	-8333	-9958
ALL	1727491	1712436	132388	55898	-127749	-67242	-8350	-15055

Enheder der ikke skal med

En første selektion af arbejdssteder sker ved kun at udtrække arbejdssteder i den private sektor i hvert af årene. Her viser det sig, at et givet arbejdssted kan være klassificeret forskelligt med hensyn til sektor i de 2 år. Der er her valgt at se bort fra disse tilfælde.

Enheder primo eller ultimo?

Ikke mindst i en longitudinal sammenhæng er det væsentligt at klargøre hvilken population, en analyse skal omfatte. Man kan tage udgangspunkt i en primo-bestand af virksomheder afgrænset på en given måde og så følge disse virksomheder over tid, fx med hensyn til opgørelser over overlevelseschancerne for forskellige typer virksomheder. En anden mulighed kunne være at anvende en ultimo-bestand: Hvorfra kommer de virksomheder, der findes på opgørelsestidspunktet?

Fælles- eller foreningsmængden?

Endvidere kan man anvende fælles- eller foreningsmængden af virksomheder på de to tidspunkter. I forbindelse med job-creation og -destruction er det netop foreningsmængden, der er interessant. Arbejdsstederne kan således opdeles i 3 overordnede kategorier:

- Tilgang af arbejdssteder
- Fortsat eksisterende arbejdssteder
- Afgang af arbejdssteder

Hvad med udskilning eller opsugning?

Et særligt problem knytter sig til, at arbejdsstederne kan afgå via "opsugning" eller tilgå via "udskilning" fra andre arbejdssteder. "Opsugning" betyder, at et arbejdssted er blevet nedlagt, og at de ansatte året efter sammen indgår i et andet arbejdssted, og tilsvarende gør sig gældende med hensyn til "udskilning". For det bevarede arbejdssted, der indgår i "opsugning" eller

"udskilning" vil der på tilsvarende måde kunne ske relativt store ændringer i beskæftigelsen. Samtidigt findes der i IDA-databasen oplysninger om, hvorvidt en opsugning/udskilning sker inden for samme firma. Hvis dette er tilfældet, er der i denne sammenhæng valgt at betragte disse typer for ændringer, som en residual form for oprettelse/nedlæggelse. Disse ændringer tæller derfor ikke som nye eller forsvundne jobs, men kun *nettovirkningen* af opsugningen/udskilningen for de berørte arbejdssteder medregnes.

Fastlæggelse af variabelværdier

I forbindelse med longitudinelle data af den her anvendte type viser der sig en mærkværdighed med hensyn til, hvordan en enhed kan beskrives. Hvis vi betragter et arbejdssted, der eksisterer i 1988, kan man på samme tid om denne enhed sige:

1. at enheden er tilgået (født)
2. at enheden er afgået (død)

Dette tilsyneladende paradoks skyldes, at man fra det givne år som tidligere nævnt kan se både "fremad" og "bagud" Hvis man i 1988 ser "bagud" kan man hævde, at enheden er tilgået, da den ikke fandtes året før. Tilsvarende hvis man i 1988 ser "fremad". Enheden kan da være forsvundet, og man vil kunne sige, at den er afgået.

På tilsvarende måde kan et arbejdssted karakteriseres på forskellig måde med hensyn til en given variabel afhængigt af, om man ser på situationen i det første eller det andet år. I forbindelse med job-creation og -destruction er det naturligt at karakterisere virksomhederne ved hjælp af oplysningerne fra det første af årene.

Resultat ved start i det efterfølgende år

Valget af at karakterisere arbejdsstederne ved situationen i det første af årene får imidlertid en anden uønsket konsekvens. Når man fx har opgjort ændringskomponenterne fra det ene år til det næste, får man naturligvis også et antal jobs i en given kategori i det følgende år. Men hvis man nu ser på data for det efterfølgende par af år, dvs. i det konkret tilfælde ændringerne fra 1989 til 1990, så vil antallet af jobs i 1989 i en given kategori *ikke* svare til ultimo-antallet i den forudgående periode. Dette skyldes netop, at der ved beregningen af 1988-89 ændringerne er taget udgangspunkt i oplysninger om arbejdsstederne i 1988. Men en vis del af arbejdsstederne er netop skiftet til en anden kategori i 1989.

Beregning af ændringsstørrelser

Det er normalt, at omfanget af ændringer i jobs sættes i forhold til det samlede antal jobs. Men her dukker igen problemet om tidsfastsættelse op. Det må dog atter være mest naturligt at sætte ændringer i forhold til antallet af jobs på det første tidspunkt.

Konklusion

Analysen af longitudinelle data rejser en række nye problemer med bl.a. populations afgrænsning og den tidsmæssige reference for variable, selv hvor der kun er tale om data fra 2 perioder.

Litteratur

Fertilitetsafsnittet

Fertilitetsudviklingen i Danmark i 1980'erne (1993). Befolkning og valg 1993:12. Statistiske Efterretninger, Danmarks Statistik, 1993.

Knudsen, Lisbeth B. (1993a) *Fertility Trends in Denmark in the 1980s. A Register Based Socio-demographic Analysis of Fertility Trends.* København: Danmarks Statistik.

Knudsen, Lisbeth B. (1993b) *Paritetsoplysningen i Sundhedsstyrelsens medicinske fødselsregister.* Validering ved hjælp af ny fertilitetsdatabase i Danmarks Statistik. Ugeskr Laeger 1993;155:2525-9

Afsnittet om IDA-databasen

Arbejdsministeriet 1977. Bekendtgørelse om virksomhedsbegreb og om antal lønmodtagere i forbindelse med foretagelse af afskedigelser af større omfang. Arbejdsministeriets bekendtgørelse nr. 74 af 4. marts 1977

Arbejdsministeriet 1993. Arbejdsmarkedspolitisk Årbog.

Baldwin, J., Dupuy, R. & Penner, W. 1992. Development of longitudinal panel data form business registers: Canadian experience. Statistical Journal of the United Nations vol. 9

Danmarks Statistik 1991. IDA - en integreret database for arbejdsmarkedsforskning. Hovedrapport.

Ministeriet for Erhvervspolitisk Samordning 1993. Erhvervsredegørelse.

Finansministeriet 1992. Vejen til fuld beskæftigelse. Analyser.

Hamermesh, D.S. 1990. A data user's look back from 2015. Monthly Labor Review. April 1990.

Leonard, J.S. 1987. In The Wrong Place at the Wrong Time: The Extent of Frictional and Structural Unemployment. I Lang, K. & Leonard, J.S. (Eds). Unemployment and the Structure of Labor Markets. New York: Basil Blackwell Inc.

Leth-Sørensen, S. 1993. Jobmobilitet belyst ved hjælp af IDA-databasen. Dansk Sociologi 1993; 4, no.1

OECD 1987. The Process of Job Creation and Job Destruction.
Employment outlook September 1987

OECD 1993. Employment/Unemployment Study. Interim Report by the Secretary-General

Rohwer, G. & Leth-Sørensen, S. 1993. Upward Mobility in the Danish Labour Market. I Jesper Lund (red.): Symposium i anvendt statistik

Vejrup-Hansen, P. 1993. Virksomhedsdemografi: Overlevelse og vækst i nye virksomheder. Samfundsøkonomen marts 1993

Det økonomiske Råd 1992. Dansk Økonomi, november 1992

Vertikal integration: Det Integrerede elevregister

Leo Elmbirk Jensen

Det integrerede elevregister er et kumuleret register, hvor alle *uddannelsesforhold* indsamlet fra 1973/74 og fremefter er organiseret på en sådan måde, at man kan følge den enkelte persons *uddannelseskarriere*, som den udvikler sig over tid.

Data om de enkelte uddannelsesforhold

Et uddannelsesforhold er en *persons* deltagelse i en *uddannelse* på en *institution.*

En uddannelse betragtes i første række som en aktivitet, der er tidsmæssigt afgrænset med et *påbegyndelsestidspunkt* og et *afgangstidspunkt.*

Uddannelsesaktiviteten resulterer, hvis den gennemføres med succes, i en *kompetence,* der med andre ord er aktivitetens resultat.

Den uddannelsesmodel, der bruges, indeholder en kode for aktiviteten, hvortil der kan knyttes en eller flere koder for kompetencer, idet der kan være mulighed for at standse aktiviteten på forskellige uddannelsesniveauer (f.eks. cand. phil./cand.mag.), eller der kan være tale om kvalitativt forskellige kompetencer (retninger, f.eks. kemiingeniør, bygningsingeniør mv.) på samme uddannelsesniveau.

De registrerede uddannelsesforhold består således af følgende data (nøgle/ident i kursiv):

Personnummer
Uddannelse
Påbegyndelsestidspunkt
Afgangstidspunkt
Afsluttende uddannelse (kompetence)
Institution (indberettende)

Datafangst

Institutionerne indberetter årligt til Danmarks Statistik. Alle elever/studerende, der har været på den enkelte institution i tællingsperioden, dvs. det forløbne år forud for tællingstidspunktet d. 1. okt., skal indberettes.

Modtageområder/ tællinger

Institutionerne er fordelt på 4 modtageområder:
- Skoler (folke-/gymnasieskoler),
- Erhvervsskoler mv.,
- Videregående uddannelsesinstitutioner,
- Frie skoler (ungdoms-, landbrugs-, søfartsskoler mv.).

Indberetningsmedier/ -kanaler

Der findes en række indberetningsmedier og kanaler. Det er således muligt at indberette på skema, diskette eller en form for magnetbånd. Der kan indberettes direkte fra den enkelte institution eller via et for flere institutioner fælles elevadministrationssystem. Det største system er ESAS, som dækker alle erhvervsskoler (handels- og tekniske skoler). På alle modtageområderne er der flere indberetningsformer, der skal integreres i en årstælling.

Formålet med en årstælling er:
- at få ajourført *elevbestanden*, dvs. de elever, der på forrige tællingstidspunkt var i gang med en uddannelse,
- at få indberettet *tilgangen* af elever i løbet af tællingsperioden.

Det ajourførte årsregister på et område viser, hvilke uddannelsesforhold der er:
- i gang på tællingstidspunktet,
- fuldført med en kompetence i tællingsperioden,
- afbrudt (uden kompetence) i tællingsperioden.

Dataintegration

Når 'årets høst' er i hus, skal data indlægges i det *Integrerede elevregister.*

Dette register har p.t. registreret ca. 7.000.000 uddannelsesforhold. De fleste af disse registreringer ligger som afsluttede uddannelsesforhold, der udgør den store uforanderlige historiske del af registret. En mindre del af registreringerne udgøres af *de igangværende uddannelsesforhold.* For disse gælder, at vi skal have at vide, om de stadig er i gang, eller om de er blevet afsluttet og i givet fald, hvordan?

Ajourfør bestand og indlæg tilgang

For at afklare dette matches årsregistrene op mod den aktive del af det Integrerede elevregister og de enkelte uddannelsesforhold ajourføres. En del af registreringerne i årsregistrene vil ikke matche, nemlig nytilgangen. Denne indlægges hermed i det Integrerede elevregister.

Billedligt kan man sammenligne det Integrerede elevregister med et træ, hvor der hvert år tillægges en ny årring udad til i form af tilgang og afsættes fast ved indad til i form at afsluttede uddannelsesforhold.

Ren maskinel indlæggelse

Tidligere var indlæggelsen en meget tidskrævende proces, idet man behandlede alle tilfælde af mismatch manuelt. I dag er indlæggelsen en ren maskinel proces med mulighed for manuel overvågning af de maskinelle processer.

Foreløbig/endelig version

Da indlæggelsen ressourcemæssigt er overkommelig har vi valgt at lave en foreløbig version på basis af indberetninger, der er 'aflusede' men ikke fejlsøgt til bunds. Den *foreløbige* version forventes at kunne være klar ca. 8 mdr. og den *endelige* ca. 12 mdr. efter tællingstidspunktet.

Uddannelseskarrierer

I det Integrerede elevregister er personernes uddannelsesforhold lagt i kronologisk orden, dvs. de kan læses som personlige *uddannelseskarrierer, der udfolder sig i tid.* Da registret dækker (næsten) alle uddannelser, har vi med dette register enestående muligheder for at give en tilstandsbeskrivelse i uddannelsessystemet på et hvilket som helst tidspunkt samt belyse de samlede bevægelser i en given periode.

Krav til data

Det første krav til data er et *fuldstændighedskrav.* Vi skal så vidt muligt dække alle offentligt anerkendte uddannelser. Det kritiske begreb i denne sammenhæng er 'restgruppen', dvs. de unge, der ikke kommer i gang med en erhvervsrettet uddannelse. Denne gruppe skal vi helst kunne afgrænse så præcist som muligt. Vi har et problem med uddannelser, hvor det er svært at finde en indberettende instans (visse etatsuddannelser og tidligere uddannelserne i den finansielle sektor). Et andet problem er, at en stigende del af uddannelsesaktiviteterne foregår som deltidsuddannelser i voksenuddannelsessystemet.

På længere sigt er vi nødt til at 'importere' resultaterne af voksenuddannelsesaktiviteterne i det Integrerede elevregister for at kunne give en fyldestgørende beskrivelse af uddannelseskarriererne. Jeg vil ikke sige mere om denne problemstilling, men vende mig mod *kvalitetskravene* til de data vi faktisk har.

Entydighed

Til en uddannelseskarriere stiller vi det krav, at den skal være *entydig*, dvs. at en person kun må være i gang med 1 uddannelse ad gangen. Dette hænger sammen med, at vi (på nær visse undtagelser) kun har fuldtidsuddannelser i registret. En person kan derfor pr. definition ikke være i gang med to samtidige uddannelser.

I praksis viser det sig at være svært umiddelbart at leve op til dette krav af to grunde:

1. Der findes konkurrerende indberetninger, hvor forskellige institutioner gør krav på at have den samme person som studerende
2. Visse uddannelser er bygget op af elementer, der indberettes hver for sig og som kan forløbe parallelt (f.eks. to overbygningsuddannelser på RUC).

Tidsmæssigt overlap

I det første tilfælde er der overvejende tale om at indberetternes registre ikke er ajourført frem til indberetningstidspunktet, dvs. en afmelding ikke er kommet på plads. Der er her tale om det, man kunne kalde 'ulovligt' tidsmæssigt overlap mellem to uddannelsesforhold. Overlapsproblemer af denne type fjernes gennem tidsmæssige justeringer/sletninger af records efter visse regler.

Parallelle uddannelser

I det andet tilfælde er der derimod tale om et helt igennem lovligt tidsmæssigt overlap, som man ikke uden videre kan fjerne. Der er med andre ord ikke tale om en fejl, men det har dog den ubehagelige konsekvens, at

princippet om entydighed i uddannelseskarriererne brydes. Det vanskeliggør f.eks. opgørelser over antal studerende på et givet tidspunkt, idet vi har et dobbelttællingsproblem.

Institutionsskift

Et andet forhold som vanskeliggør statistikken er institutionsskift. Da vi får indberetningerne fra institutionerne, vil institutionsskift betyde, at den første institution må indberette en afbrudt uddannelse og den anden en påbegyndt uddannelse. Dette er naturligvis korrekt, men gør det vanskeligt at opgøre, hvor mange der i en given periode *netto* har påbegyndt, hhv. afbrudt en uddannelse.

Det Komprimerede elevregister

I praksis har det vist sig at det er ganske vanskeligt at arbejde med det Integrerede register af de ovennævnte grunde. For at kunne leve op til kravet om entydighed og samtidig afhjælpe ulemperne lader vi registret gennemgå en bearbejdning, hvor vi samler uddannelsesdele, der hører sammen, i en record. Det betyder at vi:

- samler forskellige institutioners indberetning vedrørende den samme uddannelse i 1 record og lader den seneste institution bære indberetningen
- samler deluddannelser (især på de videregående uddannelsers område) til 1 record.

For universitetsuddannelsernes vedkommende er dette en vanskelig sag. Her vælter basisuddannelser, grunduddannelser hovedfag, bifag, sidefag og suppleringsfag rundt mellem hinanden. Disse skal stykkes sammen til hele uddannelser, hvorved vi får noget der svarer til den almindelig opfattelse af en universitetsuddannelse, nemlig at den er en helhed med et start- og et sluttidspunkt. De uddannelsesenheder, der hermed fremkommer, er nemmere at forstå for lægmand end de teknisk betonede deluddannelser.

Vi har med det Komprimerede register den behagelige situation at alle opgørelser nu er enkle:

- man kan lave bestandsopgørelser uden at skulle korrigere for dobbelttælling
- man kan lave tilgangs-/afgangsopgørelser, som uden videre giver mening
- overgange mellem uddannelser er reelle overgange og ikke interne overgange mellem uddannelsesdele.

Til-/afgangs-formen

Overgangene inde for uddannelseskarriererne er repræsenteret fysisk i registrets organisation, men ikke på en form der direkte kan udnyttes. For at gøre overgangene direkte tilgængelige for vore programmer bringer vi det Komprimerede register på Til-/afgangsformen.

Fremadrettede og bagudrettede overgange

Når vi bringer registeret på denne form forsyner vi hver eneste record med oplysninger om det umiddelbart foregående og det umiddelbart efterfølgende uddannelsesforhold. Vi har således en *centeruddannelse* der er omgivet af en *tidligere* og en *efterfølgende* uddannelse.

Uddannelses-resultat/-forud-sætning

Samtidig med at vi arbejder os ned gennem uddannelseskarrieren opdaterer vi to 'resultattavler' med den højeste almenuddannelse hhv. den højeste erhvervsrettede uddannelse. Vi har således undervejs et udtryk for personens højeste uddannelse, dvs. de uddannelsessmæssige forudsætningerne for at starte den næste uddannelse i rækken.

Det Komprimerede elevregister i til-/afgangsformen er vort vigtigste statistikgrundlag og udgangspunkt for en række anvendelser.

Statistik anvendelserne

Periodeudtræk

Fra samtlige uddannelseskarrierer udtrækkes de uddannelser der ligger inden for en tællingsperiode. På grundlag af dette udtræk kan der laves opgørelser over elevbestand på tællingstidspunktet og bevægelserne i tællingsperioden: tilgang og afgang fordelt på afbrudt/fuldført.

Tværsnit

På grundlag af de fremadrettede overgange inden for en given periode beskrives strømmene mellem hovedgrupper af uddannelser.

Uddannelses-prognose

På grundlag af de fremadrettede aldersbetingede overgange inden for en given periode og et aldersfordelt befolkningsinput i bunden af modellen foretages fremskrivninger af uddannelsespopulationer.

Nettoafgang

En defineret årgangsdelt afgangspopulation følges over en årrække. På tællingstidspunkterne opgøres uddannelsesstatus. Det er således muligt at følge f.eks. en række afgangsårgange fra gymnasiet/hf og sammenligne deres uddannelsesstatus på forskellige måletidspunkter

Lokale strømme

På baggrund af de bagudrettede og fremadrettede overgange er det muligt at følge tilstrømningen til og afstrømningen fra et defineret område i uddannelsessystemet. Eksempelvis vil man for en uddannelse kunne belyse tilgangens uddannelsesmæssige baggrund og belyse om de, der afbryder uddannelsen, afviger fra resten samt følge afbryderne i deres videre færd inden for uddannelsessystemet.

UKM

På baggrund af de ajourførte 'resultattavler' (højeste almenuddannelse, højeste erhvervsrettede uddannelse) og evt. igangværende uddannelse samt uddannelsesoplysninger fra FOB 70 dannes UKM, dvs. opgørelsen over befolkningen uddannelse på seneste tællingstidspunkt.

Personstatistikkens samspil med andre statistikområder

Poul Jensen

1. Indledning

Hvad er "områder"?

At inddele statistik i områder er vanskeligere end det lyder og der gives næppe noget entydigt princip for adskillelse af de forskellige statistikområder fra hverandre. Et blik på statistiske centralbureauers organisationsplaner vil illustrere dette. I registerstatistisk sammenhæng er det imidlertid naturligt at definere statistikområderne med udgangspunkt i den type af registrerede enheder eller "tællingsenheder", som de pågældende opgørelser bygger på. Det skal i parentes bemærkes, at dette begreb ikke er ganske det samme som de edb-prægede fænomener som kaldes "entiteter", "enheder" (i OPUS) eller "objekter" (anvendt i bl.a. i "Personstatistik på registerbasis").

Enhederne

For denne fremstillings formål anvendes ordet "enheder" som betegnelse for registerstatistikkens kernefænomener

- personer, familier, husstande (personstatistikken, CPR)
- virksomhedsenheder af forskellige typer (erhvervsstatistikken, CER)
- ejendomme, bygninger, boliger (areal-, bygnings- og boligstatistik, BBR)

Man kunne tilføje andre enhedstyper - fx skibe, der også kan opfattes som en slags erhvervsenheder eller motorkøretøjer, jf. det ny motorkøretøjsregister.

Et paradoks?

Med dette udgangspunkt for *adskillelsen* mellem statistikområderne er *relationerne* mellem disse enhedstyper bestemmende for forbindelsen mellem dem.

I det følgende behandles forholdet mellem personstatistikken og andre statistikker ud fra denne tankegang.

Afsnit 2 indeholder en kort gennemgang vedrørende "the state of the art" med hensyn til opbygning og anvendelse af centrale enhedsrelationer.

Afsnit 3 opridser betydningen af anvendelse af fælles data på forskellige statistikområder, og afsnit 4 omtaler de anvendte enheders validitet og betydningen heraf.

Endelig indeholder afsnit 5 et enkelt perspektiv angående de fremtidige muligheder i registerstatistikken.

2. Status for enhedsrelationer

Under den registerstatistiske udvikling i Danmark siden slutningen af 70'erne er to fundamentale enhedsrelationer blevet opdyrket, én har været i konstant venteposition og en fjerde kunne anvendes mere systematisk end tilfældet er for øjeblikket. Hver af disse relationer omtales i det følgende afsnit 2.1-2.4.

2.1. Person/bolig

Hovedelement

Dette der er mest afgørende element i den teknik der tillader, at der foretages folke- og boligtællingsopgørelser på registerbasis her i landet. Muligheden blev skabt i forbindelse med BBR's oprettelse, idet der da blev etableret fælles adresseidentifikationer for de enkelte lejligheder i BBR og CPR. En oprindelig tanke om til formålet at anvende fysiske lejlighedsidentifikationer i form af lejlighedsnumre på entredørene nød ikke fremme.

Præcisionen

Erfaringerne har stort set vist, at den valgte, mere begrænsede løsning har været holdbar - uden dog at være ideel. Skarpheden og præcisionen i de anvendte specifikationer er næppe så stor som ønskeligt.

2.2. Person/arbejdssted

Arbejdsplanen

Etableringen af person/boligrelationen var én forudsætning for at der kunne udarbejdes "syntetiske" folke- og boligtællinger. En anden var, at en relation kunne etableres mellem beskæftigede personer og "deres" erhvervsenheder. Det var nærmere betegnet den lokal-faglige variant man skulle bruge. Relationen skulle nemlig indgå i opgørelse dels af "erhvervspendlingen", dels af erhvervsfordelt beskæftigelse. Dette kræver, at enhederne kan henføres til et geografisk sted og branchegrupperes nogenlunde præcist.

Oplysningssedler

Hovedinstrumentet til at konstatere relationen var en lille udvidelse af informationen på skattevæsenets oplysningssedler, der i forvejen viste de beskæftigedes tilknytning til overordnede erhvervsenheder (af typen firma eller juridiske enheder). Danmarks Statistik opnåede denne udvidelse ved stor imødekommenhed fra det daværende Statsskattedirektorat. Ideen var, at arbejdsgivere med flere lokal-faglige enheder (fra da af kaldet "arbejdssteder") på oplysningssedlerne for hver af deres beskæftigede skulle anføre et forud aftalt løbenummer, der svarede til arbejdsstedet. Den statistiske effekt af denne strategiske detalje var stor, og ordningen har i al væsentligt levet op til forventningerne, selv om de praktiske vanskeligheder - som andetsteds omtalt - også har haft en anselig størrelse. Dette har bl.a. bidraget til en lang bearbejdningstid, som det ville være særdeles ønskeligt at kunne afkorte.

2.3. Arbejdssted/erhvervsarealenhed

Begrebet

En erhvervsarealenhed kan defineres som det ejendoms- eller bygningsareal som et arbejdssted beslaglægger til sine aktiviteter. En sådan enhed går noget

på tværs af de i BBR-systemet anvendte enheder (ejendomme, bygninger, boliger) og erhvervsarealenheder er kun i visse tilfælde afgrænset i BBR.

Kobling?

Kunne man imidlertid systematisk afgrænse erhvervsarealenheder i ejendoms - og BBR-systemerne og kunne de identificeres i disse systemer på samme måde, som i erhvervsregistersystemet, kunne der etableres en kobling svarende til de under 2.1. og 2.2. nævnte.

- men alligevel

Man kan sige, at denne problemstilling ikke har så meget med personstatistik at gøre, men den må alligevel omtales på grund af sin strategiske betydning for registerstatistikken.

Lang historie

Problemet har været under observation i en del år, men umiddelbart synes der ikke at være nogen "nem" løsning af en karakter, som svarer til de to forannævnte tilfælde.

Problem

Mindst to delproblemer skulle i givet fald klares.

- Det ene er, at der må autoriseres et fælles - for alle landets ejendomme og bygninger fælles - adressereferencesystem. Der findes to gode udgangspunkter herfor: CPR's adresse- og vejkodesystem og "krydsreference-registeret" i boligministeriets regi. Det vil imidlertid være ønskeligt, at der blev etableret integreret og autoriseret system, som formentlig i forhold til de nuværende skulle udbygges noget, bl.a. med umatrikulerede arealer, som vist nok ikke er komplet registreret i noget af de nuværende systemer.
- Mere vanskeligt vil det være at etablere et system til afgrænsning af erhvervsarealenheder, der er kongruente med arbejdssteder, (eller evt. et andet lokalt virksomhedsbegreb).

Men noget sker

Der er forskellige relevante aktiviteter i gang på området: (1) I forbindelse med etablering af det nye landbrugsregister i landbrugsministeriets regi skabes en forbindelse mellem "landbrugsejendomme" og "landbrugsbedrifter". (2) Der gennemføres for tiden forsøg på kobling til brug for arealstatistik ejendomsregister og "erhvervsregister" i fem udvalgte kommuner.

Ny ide

Endelig bør en ny tanke formentlig afprøves nærmere. Den går ud på, at forsøge at foretage en kobling fra erhvervssiden ved at indsamle oplysninger om hvilke ejendoms/bygningselementer den pågældende erhvervsenhed bruger i virksomhedsudøvelsen.

2.4. Ejendomme/personer - erhvervsenheder

Noget andet og lettere

En mere fremkommelig koblingsmulighed, hvis grundlag til en vis grad allerede eksisterer i dag, er at forbinde oplysninger fra ejendomsregistret med person og/eller erhvervsoplysninger via de i ejendomsregistret opførte *ejeroplysninger* identificeret ved person eller virksomhedsnumre. Denne koblingsmulighed vil blive yderligere styrket, når et standarderhvervsnummer (CVR-nummer jf. nedenfor) bliver etableret. På et sådant grundlag ville man

kunne udvikle en økonomisk ejendomsstatistik med størrelser som ejendomsvurderingsværdier, salgspriser, omsætningsfrekvenser garneret med person- og/eller virksomhedsoplysninger som baggrundsvariable. Overvejelser om en sådan statistik er endnu på et foreløbigt "skitseplansniveau".

3. Anvendelse af fælles data

Transmissionseffekter

Enhedsrelationerne har den egenskab, at oplysninger fra et statistikområde kan trækkes over i et andet og vice versa. Det betyder igen, at forskellig typer statistik kan udarbejdes på samme grundlag, men i forskellige rammer.

Hovedeksemplet herpå er "trillingestatistikkerne"

- registerbaseret arbejdsstyrke statistik-"RAS"
- erhvervsbeskæftigelsesstatistikken
- pendlingsstatistikken

Samme indhold

Alle disse statistikker bygger på nogle få stort set fælles strukturelle grundoplysninger, herunder den strategiske kobling mellem beskæftigede og arbejdssted, der bl.a. skaber mulighed for dobbeltlokalisering af de beskæftigede efter deres bopæl og deres arbejdssteds beliggenhed. De grundlæggende informationer er målingsteknisk de samme, medens begrebernes afgrænsning og anvendelse er forskellig. Det fælles produktionssystem tillader at forskelle i resultaterne kan forklares. Fx er forskellen mellem antallet af beskæftigede personer i arbejdsstyrken og antallet af "jobs" i erhvervsbeskæftigelsesstatistikken forklaret ved antallet af personer med bierhverv.

4. Begrebernes validitet

De faktiske forhold

På godt og ondt er registerstatistikken bundet til de begreber, der faktisk findes i registrene. Disses egenskaber bliver derfor afgørende for statistikkens kvalitet. Det gælder selvfølgelig også, når oplysninger fra ét område ved hjælp af enhedsrelationerne trækkes over i et andet. Denne operation medfører desuden ofte, at der bliver sat kritisk focus på de begrebsmæssige egenskaber fx eksempel ud fra sammenligning med traditionel statistik. Når overgangs- og indkøringsperioden er overstået er grundbegrebernes kvalitative egenskaber på deres egne præmisser imidlertid afgørende - også for samspillet mellem statistikområderne. I det følgende skal kort omtales de mest centrale oplysningers karakteristika.

Luksusklassen

"Personoplysningerne" er kvalitativt en klasse for sig. Begrebet "en person" er så veldefineret, at det slet ikke behøver definition og personnummeret udgør en unik identifikation, der følger personen på tværs i informationssystemerne og på langs over tiden.

På tværs men ikke på langs

Noget vanskeligere er sagen med hensyn til de begreber, der bygger på grupperinger af personer: *familie og husstand.* Disse begreber kan forholdsvis let afgrænses på et givet tidspunkt ved hjælp af personrelationer og adressefæl-

lesskab. Med hensyn til identifikation over tid er sagen vanskeligere - også fordi den foran omtalte adressekode kan forandre sig i forbindelse med vejnummeromlægning o.l.

ditto

Adressens manglende stabilitet spiller også en rolle, hvis man vil følge det iøvrigt ret veldefinerede *boligbegreb* over tid. Dette løses ikke af, at ejendomme og bygninger som fysiske fænomener er forholdsvis veldefinerede.

Et knudeproblem

Det er imidlertid inden for *erhvervsområdet* at de i særklasse vanskeligste problemer med hensyn til begrebsafgrænsningen foreligger.

Tre typer

I den danske erhvervsorienterede statistik arbejder vi med tre grundlæggende enhedstyper. De kan imidlertid sammensættes til andre enhedstyper, såfremt behovet herfor skulle opstå fx i forhold til EF-statistik.

De tre enhedstyper er

- den *juridiske* enhed, som udgør en definitorisk platform og er et praktisk udgangspunkt for de andre enhedskategorier, dels fordi det er et veldefineret fænomen, dels fordi vore donorregistre - om end hidtil kun i princippet - bygger på dette hovedbegreb
- den økonomiske enhed *(firmaet)* er det samme som den juridiske enhed, undtagen i visse, forholdsvis få tilfælde hvor en opdeling i flere juridiske enheder ikke afspejler en reel arbejdsdeling
- den lokalfaglige enhed *(arbejdsstedet)* der er en lokalt og fagligt afgrænset del af et firma.

Dobbeltfunktion

Danmarks Statistik erhvervsregistersystem, der betjener både "Det Centrale Erhvervsregister" og erhvervsstatistikken, har i længere tid lidt under mangelen på systematik i de juridiske enheders afgrænsning og nummerering. Et andet problem har været ufuldstændig/langsom ajourføring af arbejdsstederne og mangelen på selvstændig nummerering af disse på en sådan måde, at de kan følges over tid.

E-reformen

Det er på denne baggrund et hovedpunkt i den igangværende reform af den generelle erhvervsstatistik at skabe et registersystem, hvori disse mangler i væsentlig grad er udbedret.

Det nye CVR

Imidlertid er der under udviklingsforløbet truffet beslutning om oprettelse af et nyt centralt basisregister - CVR - i Told•Skats regi. Dette ventes at træde i virksomhed i løbet af et par år. Det er en selvfølge, at der i forbindelse med oprettelse af dette register sker en kraftig sanering af juridiske enheder, herunder en tiltrængt oprydning i numrene. Disse skal i øvrigt bygge på de såkaldte SE-numre, der også skal afløse de nuværende aktieselskabsnumre.

De helt rigtige arbejdssteder

Danmarks Statistik har ud fra egne og andre potentielle CVR-brugeres interesser set det som opgave at argumentere for, at det nye register systematisk kommer til at indeholde arbejdsstedsniveau, at det bliver overladt Danmarks Statistik at "producere" disse enheder i samarbejde sammen med andre ho-

vedbrugere, og at disse enheder forsynes med "eget" uafhængigt nummersystem således, at de kan følges over tid uden påvirkning af ændringerne i de numre, som identificerer den juridiske enhed. Sammen med arbejdsministeriet, der er administrativ storbruger af sådanne enheder har Danmarks Statistik fremsat et forslag herom, der tilgodeser begge institutioners interesser, og som i alt væsentligt bygger på Danmarks Statistik hidtil anvendte arbejdsstedsbegreb men nu kaldet *"produktionsenheder"*.

Nærliggende fejlslutning

At registerbegrebernes egenskaber er centralt for statistikkens betydning indebærer ikke nødvendigvis, at man skal videreføre de fra den traditionelle statistik kendte begreber i uændret form. Det er også vigtigt at begreberne er stabile, velafgrænsede og egnede til administrative formål, således at oplysninger fra forskellige administrationer kan indpasses i samme helhed.

Duer register-statistikken

I den internationale diskussion anføres ofte som et hovedargument mod registerstatistik, at disse hensyn er vanskelige at forene med de krav som anerkende internationale definitioner stiller til begreberne. Dette synspunkt kan selvfølgelig ikke afvises, men med udgangspunkt i, at grundregistrene og den generelle statistik i nogen grad har fælles interesser med hensyn til de grundlæggende begreber kan man alligevel nå et stykke vej, såfremt problemet prioriteres ved den grundlæggende planlægning.

Danmarks Statistiks rolle

På dette område har Danmarks Statistik med sit omfattende erfaringsgrundlag - og en tilgang til begrebsproblematikken, som ikke er farvet af bestemte administrative hensyn - en betydningsfuld rolle at spille - en rolle som den fremsynede lovgivning har forudset i § 1, stk. 1 nr. 3 i lov om Danmarks Statistik.

5. Perspektiv

Fremskridt

Allerede nu indebærer den registerstatistiske udvikling på personstatistikkens område et stort spring fremad, blandt andet fordi den muliggør, at man opgør "bruttoforskydninger" i stedet for nettobevægelser. Det er en gammel, men stadig gyldig sandhed, at "ingen hidtil har set en nettovandrer". Noget andet er, at den fulde udnyttelse af registerstatistikkens potentiale næppe er opnået endnu.

Samspilsstatistik

Med den rette fastlæggelse af begreberne inden for erhvervsområdet opstår der tilsvarende muligheder, hvor "erhvervsdemografiske opgørelser" er stærkt savnet. Mere relevant i dette papirs sammenhæng er "dobbeltdemografiske opgørelser" til belysning af det komplicerede, men virkelighedsnære samspil mellem forskellige komponenter i beskæftigelsesudviklingen: beskæftigede i ophørte enheder - beskæftigede i nye enheder - bruttotilgang og bruttoafgang af beskæftigede i bestående erhvervsenheder. Det er de enkelte af disse komponenter - ikke "nettovandreren" - der kan påvirkes af erhvervspolitiske og arbejdsmarkedspolitiske tiltag.

Registerlovgivning og datapolitik

Finn Spieker

1. Registerloven

Danmarks Statistiks personstatistikregistre er omfattet af Lov om offentlige myndigheders registre, jf. Lovbekendtgørelse 1991-09-20 nr. 654. Denne lov gælder for edb-registre, der føres for den offentlige forvaltning, og som indeholder personoplysninger. Edb-registre defineres som registre eller andre systematiske fortegnelser, hvor der gøres brug af elektronisk databehandling, og ved personoplysninger forstås oplysninger, som kan henføres til bestemte personer, selv om det forudsætter kendskab til personnummer, registreringsnummer eller lignende særlige identifikationer.

Loven indeholder en række bestemmelser om oprettelse af registre, om registrering og opbevaring af oplysninger, om registrerede personers adgang til oplysninger om sig selv, om videregivelse, om Registertilsynet samt om straf.

For statistik og forskning gælder enkelte undtagelser i forhold til de almindelige regler. Det drejer sig om adgangen til at videregive oplysninger (§18 og § 21, stk. 5), adgangen til egenindsigt (§ 13 , stk. 5) samt adgangen til at samkøre registre (§4, stk 3 og § 8d).

Som hovedregel må videregivelse af personoplysninger afgivet til statistikformål ikke videregives. Loven åbner dog mulighed for, at sådanne oplysninger kan videregives til andre statistiske eller videnskabelige formål med tilladelse fra Registertilsynet. Endvidere kan videregivelse ske, hvor der foreligger en særlig hjemmel. EF-forordningen (nr. 1588/90) om fremsendelse af fortrolige oplysninger til De Europæiske Fællesskabers Statistiske Kontor er eneste eksempel på en sådan hjemmel for så vidt angår personstatistikregistrene. Forordningen bemyndiger statistikbureauerne til at fremsende fortrolige statistiske oplysninger til EUROSTAT. Fra det erhvervsstatistiske registersystem finder en begrænset videregivelse sted til Det Centrale Erhvervsregister med hjemmel i lov om Det Centrale Erhvervsregister.

Videregivelsesbestemmelserne er en væsentlig forudsætning for de to øvrige undtagelser. Med disse bestemmelser sikres det, at statistikoplysninger ikke får konsekvenser for den administrative behandling af personoplysninger. Hvis Danmarks Statistik ved behandlingen af statistikoplysningerne konstaterer fejl eller andre uregelmæssigheder om enkeltpersoner, så kan disse ikke meddeles til den pågældende administrative myndighed. Kun, hvor det drejer sig om overordnede systemmæssige fejl i forbindelse med udtræk og levering af oplysninger til Danmarks Statistik, kan Danmarks Statistik gøre opmærksom på fejlen, men det sker uden henvisning til enkeltpersoner. I forbindelse med Befolkningsstatistikregistret har Danmarks Statistik adgang til at rette henvendelse til folkeregistre med henblik på at rette konkrete fejl i statistikoplysningerne. Dette gennemføres nu i praksis

ved, at Danmarks Statistik gennem direkte terminaladgang til Det Centrale Personregister (CPR) selv kan udsøge de relevante oplysninger.

Reglerne om videregivelse bestemmer på en vis måde, at personstatistikregistrene i princippet ikke indeholder egentlige personoplysninger. Indholdet må kun være grundlag for anonyme statistiske opgørelser. Den enkelte person vil således ikke i nogen sammenhæng blive konfronteret med oplysninger fra et statistikregister og har derfor ingen egeninteresse i at kende indholdet af et sådant register. Det er baggrunden for, at loven i § 13, stk. 5 indeholder en undtagelsesbestemmelse med hensyn til den enkelte persons adgang til at få oplyst, hvad der er registreret om den pågældende selv i et statistikregister.

Den sidste undtagelse gælder som nævnt adgangen til at samkøre registre. Efter de almindelige regler skal der oprettes forskrifter for en sådan, men det gælder ikke, hvis formålet er statistik eller forskning. Begrundelsen er igen, at det samkørte produkt ikke kan danne grundlag for afgørelser vedrørende enkeltpersoner og derfor ikke kan skade sådanne. Det skal dog tilføjes, at undtagelsen kun gælder, hvis samkøringen resulterer i et anonymt datasæt, ellers vil der være tale om et nyt register, for hvilket der skal foreligge forskrifter.

Videregivelsen af oplysninger fra de administrative registre til Danmarks Statistik er omfattet af reglerne i Lov om offentlige myndigheders registre § 21, stk. 2 og 3. For særligt følsomme data gælder den hovedregel, at videregivelse ikke må finde sted, § 21, stk. 1. Blandt undtagelserne fra denne regel er lovhjemlet adgang til videregivelse, eller hvis formålet med videregivelsen er statistik eller forskning. Lov om Danmarks Statistik er således tilstrækkeligt grundlag for indsamling af oplysninger til statistik. Den generelle statistikbestemmelse gør det muligt, i forbindelse med serviceopgaver, at anvende supplerende oplysninger fra administrative registre til samkøring i Danmarks Statistik på brugerens foranledning. Tilsvarende bestemmelser findes i Lov om private registre § 4, stk. 1 og 3.

2. Registerforskrifter

Registerloven indeholder en bestemmelse om, at der skal fastsættes forskrifter for hvert register. Der er undtagelser fra denne regel, men de er uden betydning for statistikregistrene. Registerforskrifterne skal indeholde de regler som gælder for driften af det enkelte register. Indholdet omfatter følgende punkter:

Formål med oprettelsen af registret
Registrets indhold
Opdatering
Anvendelsen af registret
Sikkerhedsforanstaltninger
Evt. videregivelsesbestemmelser
Opbevaring

I bilag til forskrifterne gives en specificeret beskrivelse af, hvilke typer af oplysninger der indgår i det pågældende register, og hvor de kommer fra. I et andet bilag beskrives anvendelsen af registret. Endelig er sikkerhedsforanstaltningerne beskrevet i Danmarks Statistiks Datasikkerhedsreglement, der ligeledes er bilag til forskrifterne.

Videregivelsesbestemmelserne er fastsat sådan, at det for de fleste registre gælder, at videregivelse ikke kan finde sted. Danmarks Statistiks datapolitik har på dette område været mere restriktiv end loven, og det gælder også for registre, hvor man i konkrete sager har afvist anmodninger om videregivelse til statistikformål, selv om forskriftens regler ikke var til hinder for det. Det er derfor kun i meget få tilfælde, at videregivelse har fundet sted, og det er kun sket, hvor videregivelsen har været åbenbart rimelig og uden sikkerhedsrisiko.

Forskrifterne sætter også regler for, hvor længe oplysningerne må opbevares. Det gælder således, at de må bevares så længe, det er nødvendigt for at tjene det formål, hvortil de er indsamlet, dog er der fastsat en bestemt frist for, hvornår et register eller en registerversion skal slettes eller arkiveres i Rigsarkivet. Danmarks Statistik har traditionelt anvendt en hovedregel om 5 års opbevaring med enkelte undtagelser op til 20 år. Registerversioner, der opbevares i Rigsarkivet kan efter tilladelse fra Registertilsynet stilles til rådighed for konkrete statistikopgaver ud over den fastsatte frist.

3. Datasikkerhedsorganisationen

Rigsstatistikeren er øverste ansvarlige for datasikkerheden i Danmarks Statistik. Til støtte for ham er nedsat et datasikkerhedsudvalg, som løbende behandler spørgsmål i forbindelse med datasikkerheden og kommer med indstillinger om eventuelle ændringer. Udvalget skal mindst én gang om året gennemgå sikkerhedsforanstaltninger og afgive beretning til Rigsstatistikeren. Endvidere forestår udvalget udarbejdelsen af en datasikkerhedshåndbog, som beskriver alle regler og foranstaltninger i forbindelse med datasikkerheden. Udvalget skal sikre, at en ajourført datasikkerhedshåndbog altid er let tilgængelig for alle ansatte i Danmarks Statistik.

4. Danmarks Statistiks diskretionspolitik

De mange muligheder for videregående analyser og forskningsprojekter, som statistikregistersystemets omfattende dataindhold og datasammenhænge giver, har skabt behov for, at der stilles detaljerede oplysninger til rådighed for eksterne brugere. Det er Danmarks Statistiks opgave at fremskaffe de efterspurgte informationer til brugerne i en efter formålet hensigtsmæssig form. Samtidig skal Danmarks Statistik sikre, at man ikke af informationernes indhold kan genkende enkeltpersoner, så der opstår en risiko for misbrug af personoplysninger. Det sidste hensyn er ufravigeligt, men det kan være vanskeligt at afgøre, hvornår der opstår en risiko. Brugerens ønsker kan være meget

detaljerede oplysninger, der kan være sagligt velbegrundede i forhold til den anvendelse, som han agter at gøre af dem. Der opstår derfor nemt en konflikt mellem de to hensyn, og Danmarks Statistik har på den baggrund formuleret en diskretionspolitik, som giver retningslinier for, hvornår en opgave må afvises.

a. Lovlighed og mulighed

En opgave skal være lovlig, hvilket betyder, at evt. supplerende oplysninger, som indberettes af en opgavestiller, ikke kan indgå, hvis det er i strid med registerlovgivningen. Endvidere skal den være mulig i den forstand, at de for opgaven relevante oplysninger skal være til rådighed i et omfang, som sikrer en tilfredsstillende behandling af emnet. Endelige skal som nævnt enkeltpersoners anonymitet sikres.

b. Emnet

I almindelighed er det Danmarks Statistiks styrelse, som gennem den årlige arbejdsplan fastlægger, hvilke emner der skal behandles. Specielt i forbindelse med serviceopgaver vil det imidlertid være opgavestilleren, der bestemmer emnet. Et sådant emne kan være af en karakter, som vil være generende for visse grupper. Særlige interesser kunne være truet, hvis resultaterne fører til uønskede beslutninger, eller resultaterne kunne vise koncentration af uheldige egenskaber i en særlig mindre gruppe af befolkningen. Behandlingen af sådanne emner kunne medføre utilfredshed med Danmarks Statistik i visse kredse. Trods dette, at resultaterne på den måde kan være ubekvemme for Danmarks Statistik, så må det ikke i sig selv føre til, at en opgave afvises. Til gengæld forbeholder Danmarks Statistik sig ret til at kritisere undersøgelsens konklusioner.

c. Tabeller

Den i almindelighed mest velkendte form for statistik er tabellerne, som offentliggøres i Danmarks Statistiks publikationer. Der vil normalt ikke være problemer forbundet med de aggregerede oplysninger, som fremgår af disse tabeller. Anderledes stiller det sig med tabeller, der fremstilles efter særlig aftale. Her kan kravene til detaljeringsgrad være sådan, at der opstår en risiko for genkendelse. Det gælder specielt, hvis der indgår opdeling i mindre geografiske områder. Derfor må der stilles krav om en mindstegrænse for størrelsen af et sådant område. Den er målt i antal indbyggere som hovedregel sat til 1000, men den må klart vurderes i forhold til specifikationsgraden for de øvrige oplysninger i tabellen.

d. Modeldata

Hovedreglen for, hvad der kan videregives af oplysninger fra Danmarks Statistik er, at de skulle kunne offentliggøres uden risiko for genkendelse af en-

keltpersoner. Det betyder ikke nødvendigvis, at offentliggørelse faktisk finder sted. Undtagelser fra denne regel er mulig på særlige betingelser. Modeldatasæt, dvs. individualoplysninger uden identifikation, kan stilles til rådighed for statistik eller forskningsprojekter. Det kan ske gennem levering af et datasæt til brugeren, der så kan behandle oplysningerne på eget edb-anlæg, eller det kan ske ved at give brugeren adgang til at behandle datasættet på Danmarks Statistiks edb-anlæg. Sidstnævnte kan ske gennem en fast opkoblet linje eller fra en arbejdsstation i Danmarks Statistik på såkaldt forskerplaceringsvilkår, hvilket bl.a. indebærer opsyn med behandlingen af oplysningerne.

Levering af modeldata er omfattet af det overordnede princip, at der ikke må være en reel risiko for, at modtageren kan identificere enkeltpersoner og på den måde opnå ny viden om sådanne. For at sikre dette, er der en række regler for, under hvilke omstændigheder modeldata kan stilles til rådighed:

> Hvis et modeldatasæt omfatter identificerede personoplysninger, der er indleveret til samkøring med oplysninger i Danmarks Statistik, så kan det kun stilles til rådighed under forskerplaceringsordningen i Danmarks Statistik.
>
> Dataindholdet i et modeldatasæt må kun indeholde de oplysninger, som er nødvendige for at opfylde formålet med dannelsen af datasættet. Der kan kompenseres for forekomsten af mange data gennem en reduktion af stikprøvestørrelsen eller ved en reducere specifikationsgraden for den enkelte variabel.
>
> Der må ikke forekomme oplysninger med høj identifikationsgrad med mindre, der er tale om tilbagelevering af oplysningerne til den, som oprindeligt har tilvejebragt alle personoplysningerne, eller oplysningerne skal indgå i et udviklingsprojekt om hvilket, der er indgået en aftale om samarbejde med Danmarks Statistik som deltager.
>
> Modeldata kan som hovedregel kun stilles til rådighed under forskerplaceringsordningen eller forskerpostkassen. Kun hvis der er tale om en stikprøve med en meget lille udvalgsandel eller en forholdsvis lav specifikationsgrad, kan et modeldatasæt leveres til behandling på brugerens eget edb-anlæg.

I forbindelse med leveringen af modeldatasæt aftales de nærmere vilkår for behandlingen af oplysningerne. En sådan aftale skal omfatte en specifikation af formålet, en angivelse af, hvem der har adgang til at bruge datasættet, samt en præcisering af, hvor databehandlingen må finde sted. Endvidere skal det indgå, at videregivelse af oplysninger kun må ske i en bearbejdet form, der er i overensstemmelse med kravene til, hvad der kan publiceres. Endelig forbeholder Danmarks Statistik sig ret til efterkritik, da oplysningerne ikke er offentligt tilgængelige, og der derfor ikke er mulighed for at andre kan vurdere brugerens anvendelse af de pågældende oplysninger.

e. Lovmodel

En særlig form for modeldata er den såkaldte Lovmodel. Den er organiseret i flere "modelbefolkninger" og gennem disse har brugerne adgang til meget detaljerede oplysninger fra mange forskellige registre. Adgangen sker fra arbejdsstationer hos brugeren selv over fast opkoblede linjer. Danmarks Statistik har lagt følgende klausuler på benyttelsen af disse modeldata:

> at de udelukkende vil blive anvendt til lovmodelberegninger, herunder afprøvning af sådanne beregninger,
>
> at kun autoriserede lovmodelbrugere får adgang til oplysningerne, og
>
> at brugerne ikke er berettiget til at offentliggøre egentlige statistiske opgørelser, der måtte blive fremstillet på grundlag af modelbefolkningerne, uden forudgående aftale med Danmarks Statistik.

Risikoen i forbindelse med lovmodellen er, at man af hensyn til udefinerede behov ophober alt for mange data. Der gælder derfor følgende begrænsende regler:

> Der må ikke leveres data om stikprøver større end 3 pct. af hele befolkningen.
>
> Der må normalt ikke leveres forløbsdata om enkeltindivider. Kun i særlige tilfælde, hvor der kan påvises specifikke behov i forbindelse med lovforberedende arbejde, vil det være muligt.
>
> Der må ikke indlægges oplysninger, der ikke kan bruges på grund af kvalitetsproblemer.

f. Identificerede oplysninger

Levering af oplysninger om enkeltpersoner med formel identifikation er normalt ikke tilladt. Undtaget fra denne regel er levering af simpelt tilfældigt udvalgte stikprøver til Socialforskningsinstituttet. Stikprøverne kan benyttes til undersøgelser, der udføres af instituttet. Stikprøverne må kun afgrænses efter køn, alder og bopæl. Hvis en stikprøve ønskes stratificeret efter kendetegn fra andre registre end befolkningsstatistikregistret, så må databehandlingen foregå i Danmarks Statistik.

En anden undtagelse gælder levering af identificerede personoplysninger til det grønlandske hjemmestyres statistiske kontor. Oplysningerne vedrørende befolkningen i Grønland er således blevet overgivet i forbindelse med, at Hjemmestyret har overtaget ansvaret for den fremtidige statistik.

5. Reaktioner fra statistikbrugere

Reaktioner på Danmarks Statistiks diskretionspolitik kommer især fra brugere af modeldatasæt. De kan have svært ved at forstå nødvendigheden af den restriktive politik. Set fra en forskersynsvinkel kan det være af væsentlig betydning af have detaljerede oplysninger til rådighed. Det kan være bestemmende for, hvilke analysemetoder der kan anvendes, og i visse tilfælde kan det betyde, at projekter må begrænses eller helt opgives.

Adgangen til at anvende modeldata er blevet lempet gennem de senere år. Den tidligere omtalte forskerplaceringsordning har således åbnet mulighed for at anvende meget detaljerede oplysninger, men det opleves ikke af alle som tilstrækkeligt. Det føles af nogle som en betydelig ulempe, at databehandlingen skal foregå fra en arbejdsstation i Danmarks Statistik, og det må da også erkendes, at den geografisk afstand i nogle tilfælde kan være en væsentlig hindring for at benytte ordningen.

For at afhjælpe det afstandsproblem er den såkaldte postkasseordning indført. Gennem denne ordning har brugeren mulighed for af lave sine edb-programmer hjemme og derefter indsende dem til afvikling af kørsler på aftalte datasæt i Danmarks Statistik. Brugeren har ikke direkte adgang til data, og det kan betyde en lidt tung kommunikation, men det skal så afvejes mod forskerplaceringsordningen eller begrænsninger i det datamateriale, som kan stilles til rådighed.

I almindelighed vil forskerne af forståelige grunde foretrække at få detaljerede datasæt hjem. De støttes af registerlovens videregivelsesbestemmelser, hvorefter noget sådant er muligt med tilladelse fra Registertilsynet. Danmarks Statistik på sin side ser det som et meget væsentligt led i institutionens virksomhed at stille oplysninger til rådighed, og man bestræber sig på at finde acceptable løsninger på de problemer, som den førte diskretionspolitik rejser. Der er imidlertid et overordnet og meget tungtvejende hensyn, som er baggrund for denne politik, og det gælder sikringen af, at der i befolkningen er tillid til, at oplysninger i Danmarks Statistik ikke bliver misbrugt. Det er simpelthen af afgørende betydning for Danmarks Statistiks fortsatte virksomhed og dermed også for forskernes muligheder for at få datamaterialer fremover.

6. Eventuelle virkninger af EF-direktivet

Med henblik på af skabe grundlag for informationers frie bevægelighed i EU har Kommissionen udarbejdet et forslag til "Rådets direktiv om beskyttelse af fysiske personer i forbindelse med behandlingen af personoplysninger og om fri udveksling af sådanne oplysninger". Forslaget er gennem længere tid blevet behandlet dels i en arbejdsgruppe under Rådet, og dels i de enkelte medlemslande.

Forslagets indhold vil, hvis det gennemføres i den foreliggende form, betyde væsentlige begrænsninger i mulighederne for statistisk anvendelse af personoplysninger. Problemerne er især forbundet med adgangen til at genbruge administrative data til andre formål end det oprindelige, fx. statistikformål, men det handler også om adgangen til egenindsigt, form og varighed for opbevaring af statistikoplysninger m.m. Fra dansk og fra anden side er der gjort opmærksom på konsekvenserne for statistik og forskning, hvis det foreliggende forslag gennemføres, og der synes da også at være bevægelse i retning af at tilgodese statistikernes synspunkter. det er imidlertid for tidligt at sige noget om, hvordan det ender.

Beredskab og formidling til forsknings- og udredningsopgaver

Otto Andersen

Baggrund

Tidligere tiders store undersøgelser som f.eks. folketællingerne opbevares for "evigt" i Rigsarkivet, men udnyttelsen sker kun til slægtsforskning m.v. Der er blandt mange historikere interesse for en overførsel af skemaerne til et edb-medium, idet dette sammen med kirkebogsmateriale mv. ville give historikere enestående muligheder for at gennemføre demografiske og socialhistoriske analyser. Det er selv uden større undersøgelser indlysende, at et projekt af den nævnte karakter er ressourcemæssigt og økonomisk af en næsten uoverkommelig opgave.

EDB-registrenes udvikling og anvendelse i statistikproduktionen har stillet nye krav til beredskab og formidling. Primærmaterialets karakter af datasæt opbevaret som edb-filer fremfor som papir giver mulighed for genanvendelse i langt større udstrækning, end det tidligere har været praktisk muligt. Registrenes mange data kan efterfølgende relativt let anvendes til nye analyser, til nye projekter og til helt nye sammenstillinger af data enten alene eller i forbindelse med andre registre. Dette giver Danmarks Statistik en stor forpligtelse til nøje at overveje og søge at påvirke, hvad der skal gemmes for eftertiden på såvel kort som lang sigt.

I dette notat beskrives beredskabsprincipper og metoder til formidling af data, først og fremmest i forbindelse med forsknings- og udredningsprojekter. Danmarks Statistik serviceberedskab iøvrigt er kun behandlet sporadisk. Bagved mange af betragtningerne i det følgende ligger Danmarks Statistiks diskretionsprincipper og dokumentationsforpligtelsen, som imidlertid bliver behandlet særskilt i andre notater til seminaret. Disse to overordentligt vigtige områder betragtes derfor mere eller mindre som kendte.

Beredskabet

Datagrundlaget for registrene i Danmarks Statistik er enten Danmarks Statistiks egne indsamlede data f.eks. på spørgeskemaer (eks. den sociale ressourceundersøgelse) eller andre edb-registre fra administrative myndigheder (eks. indkomststatistikregisteret).

Skal primærdata arkiveres?

Det første eksempel adskiller sig ikke fra tidligere tiders dataindsamling på anden måde, end at data er overført til et edb-medium. I det omfang at man stoler på, at data er overført korrekt til et edb-medium, og at der ikke af en eller anden grund er udeladt nogle informationer i overførselen, har man til-

syneladende ikke mere brug for papirversionen. Denne kan makuleres. Under alle omstændigheder får man næppe nogensinde ressourcer til at vende tilbage til grunddata i større udstrækning.

Helt kan det dog ikke afvises. Ved etableringen af et register baseret på et spørgeskema foretages der ofte en kodning, som aggregerer en række informationer. Et eksempel har vi i Folke- og boligtællingen i 1970 (FOB70), hvor hver person på husstandsspørgeskemaet står opført med personens egen fag- og erhvervsbetegnelse, som af kodepersonalet er "fortolket" og omsat til en af de 218 fagkoder og 245 erhvervskoder. I et specialprojekt med Cancerregisteret har det ikke været nok med disse koder, idet det var nødvendigt med en præcisering (f.eks. skulle personer beskæftiget i renserier udskilles), og dette kunne ske ved at gå ned i primærmaterialet. Det var derfor overordentligt vigtigt, at edb-registeret indeholdt et skemanr, som muliggjorde en fremfinding af de oprindelige skemaer i Rigsarkivet.

Konklusionen bør således være, at registerversionerne i de tilfælde, hvor der er foretaget en omfattende kodning, skal muliggøre en fremfinding af grundmaterialet, og at dette bør arkiveres. Hvor der blot er sket en ren og skær overførsel af oplysninger (f.eks. forud fastlagte koder i lukkede spørgsmål), er behovet ikke det samme.

For de mange registre, der er etableret på grundlag af administrative registre, gælder det, at der i regelen foretages omfattende bearbejdninger, så det færdige registerprodukt i Danmarks Statistik selvfølgelig afspejler primærdata, men har sin helt selvstændige status. Det er derfor vigtigt at diskutere, hvad der skal arkiveres, primærdata og eller registerversionen i Danmarks Statistik. Vi kommer hermed ind på spørgsmålet om forældelse af registre og arkivering i Rigsarkivet.

Forældelsesfrister

De enkelte registerforskrifter indeholder bestemmelser om forældelse, det være sig 5, 10 eller 20 år efter datas indførelse i registeret. Indtil den pågældende forældelsesdato skal Danmarks Statistik sikre en pålidelig teknisk opbevaring og dokumentation. Efter forældelsesfristens udløb skal der ske en sletning eller en arkivering i Rigsarkivet. Kravet til sletning eller arkivering af data kan muligvis fraviges af Registertilsynet, dersom der er gode argumenter for det, så som værdien af at opretholde longitudinelle data for så lang en periode som muligt.

Er sletningen betryggende

Rigsarkivet er bekendt med registerforskrifterne og tager stilling til, hvad der ved forældelsesfristernes indtræden skal arkiveres. Det er et godt spørgsmål om beslutninger om sletning af data træffes på et grundlag, der er tilfredsstillende for forskere. Som eksempel kan nævnes indkomsstatistikregisteret, som Rigsarkivet giver tilladelse til at slette, idet primærinformationerne fra Told og Skat opbevares.

Argumentet for kun at opbevare den oprindelige version er naturligvis, at informationeme kun bør findes et sted. Bekymringen kan derimod være, at der

i registerversionen i Danmarks Statistik er foregået en række bearbejdninger samt sammenkobling med andre kilder, f.eks. befolkningsstatistikken, som alt i alt vil bevirke, at genanvendelsen vil være langt mere fremkommelig ved Danmarks Statistiks registerversion end ved Told og Skats primærmateriale.

Spørgsmålet om forældelse og arkivering bør overvejes nøje i lyset af, at registerstatistikken nu har nået en så moden alder, at det virkelig spiller en rolle. Vi ser det tydeligt i de mange forskningsprojekter, der ønsker at anvende data longitudinelt. Det er med stadig større hyppighed arkivversionerne i Rigsarkivet, der skal anvendes.

Konklusion

Konklusionen (til diskussion) må være:

1. hvordan skal processen om arkivering indrettes, således at fremtidens (specielt forskningens) interesser tilgodeses?
2. hvad skal arkiveres, primærdata og/eller registerdata?
3. skal der oprettes et "råd", der indstiller datasamlinger til arkivering?

Databaseberedskab

Den vertikale integration af registre, jf. seminarets sektion 5.2 vil som tiden går blive mere og mere relevant. I dette afsnit skal der gøres rede for den tankegang, der ligger bag oprettelsen af forskningsdatabaser i Danmarks Statistik ud fra et beredskabssynspunkt.

Halvfabrikata

Danmarks Statistik har siden 1988 oprettet to databaser beregnet til forskningsformål, nemlig **IDA-databasen** og **fertilitetsdatabasen**. Baggrunden for at etablere databaserne har været et ønske om at bearbejde de omfattende datasamlinger i Danmarks Statistik til et produkt, en slags halvfabrikata, som skulle gøre det langt lettere for enkelte forskere indenfor bestemte forskningsområder at anvende data.

Danmarks Statistiks datasamlinger er som hovedregel emnemæssigt afgrænsede, f.eks. befolkningsstatistikregisteret og indkomststatistikregisteret. Forskere har som oftest et ønske om at kombinere data fra forskellige registre, hvilket i regelen kan lade sig gøre som ad hoc opgaver. Der er i forskellige projekter ofte et fælles element af baggrundsoplysninger, det være sig arbejdsmarkedsdata, uddannelsesdata mv. I ad hoc projekterne starter den enkelte forsker forfra hver gang, med et ressourcespild som resultat.

Fundamentale problemer løses

Da tankerne om at danne IDA (Integreret Database for Arbejdsmarkedsforskning) opstod , var baggrunden endvidere den, at der indenfor arbejdsmarkedsforskningen var nogle helt fundamentale dataproblemer, som burde løses "en gang for alle" og ikke af hver enkelt forsker i det enkelte projekt, hvilket iøvrigt næsten uden undtagelse ville være praktisk og økonomisk uoverkommeligt.

Ideen med IDA var at danne et datasæt, hvor personer og virksomheder blev knyttet sammen over tid, en problemkreds, som var (og er) meget vanskelig. Personer har et personnummer fra fødsel til død, men virksomheder kan

skifte virksomhedsnr. af mange årsager. Det var et oplagt forskningsprojekt at danne den omtalte forbindelse. Det tog tre personer 3 år at danne databasens grundsubstans.

For fertilitetdatabasens vedkommende var problemstillingen nogenlunde den samme, idet det her drejede sig om "en gang for alle" at finde forbindelsen datamæssigt mellem børn og forældre, således at nogle stærkt savnede analyser af fertiliteten i Danmark kunne gennemføres.

Ressourcetungt

Danmarks Statistik er gået ind i disse projekter med støtte fra Statens Samfundsvidenskabelige Forskningsråd, men med et betydeligt eget bidrag. Efter etableringen af databaserne har Danmarks Statistik ansat 2 akademikere til at servicere og vedligeholde data, ligesom der er knyttet 2 programmører til projekterne.

Databaserne anvendes ikke til egentlig statistikproduktion, men det er dog tanken at udnytte dele af fertilitetsdatabasen i de årlige offentliggørelser. I og med at formålet er forskning, er Danmarks Statistiks mulighed for at opretholde og opdatere data betinget af en tilstrækkelig interesse fra forskningsmiljøerne. Data tilbydes på Danmarks Statistiks almindelige servicevilkår.

Andre databaser?

Der kan tænkes databaser på en række forskningsområder. Nærliggende kunne være indenfor sundhedsforskning eller socialforskning, men det er nok vigtigt at holde sig for øje, at det ikke lader sig gøre på nogen rimelig måde at etablere en altomfavnende database. Man er nødt til at holde sig anvendelsen for øje, ellers bliver datasamlingerne helt uoverskuelige og opdateringen næppe realistisk.

Baggrundsdatabase

På et enkelt område kunne det dog være en tanke værd, at oprette en slags database, som ikke var til anvendelse på et specifikt forskningsområde. Det drejer sig om baggrundsdata for forskningsprojekter. Når man har arbejdet dagligt med vejledning af forskere, støder man på næsten ensartede ønsker om data for enkeltpersoner, såsom civilstand, statsborgerskab, boligforhold, fag og erhverv, uddannelse, indkomstforhold mv. Det er egentlig de samme oplysninger, som blev dannet ved den første (og hidtil eneste) registerfolketælling. Uden at argumentere for at gentage registerfolketællingen kunne det overvejes, om et langt mindre datasæt med baggrundsoplysninger kunne dannes uden alt for store omkostninger. For forskerne ville det betyde en begrænsning i valget af baggrundsdata, men også en betydelig besparelse, idet datasættet var "fælleseje".

Serviceberedskabet

Beredskabet i de enkelte fagkontorer kan bestå i specielle versioner af registeret, som er specielt designet til serviceopgaver, jf. f.eks. notatet om sygehusbenyttelsesregisteret. Det kan bestå i særligt bearbejdede versioner (sumdata), som gør standardtabulering lettere. I dødelighed- og erhvervsprojektet er der foretaget en række grundlæggende grupperinger af fag- og erhvervsgrupper samt af dødsårsagerne, og der er udviklet et specielt tabelsystem, der generer nogle standardtabeller, der indeholder den for forskere nødvendige information. Det har dog vist sig, at forskere (og ganske velbegrundet) ønsker andre grupperinger and standardklassifikationerne. Man er

ofte på jagt efter at finde overdødelighedsmønsteret for specielle faggrupper, om hvem man ved, at de har været særligt udsat for arbejdsmiljøpåvirkninger. Som følge deraf ønsker man også specielle dødsårsager specificeret. Det har ført til, at der næsten hver gang dannes datasæt på basis af grundregisteret.

Når det drejer sig om beredskab af gamle data, er det helt indlysende, at der er to aspekter, der nøje må overvåges.

Teknikken

Det ene aspekt er den tekniske mulighed for at bearbejde data, der er arkiveret. Det skal ikke behandles her (forfatteren ved ikke noget om det), men det må være et ubønhørligt krav, at data altid findes på et medium, der er læsbart på det relevante tidspunkt.

Dokumentationen

Det andet aspekt er dokumentationen. Dette behandles andetsteds i seminaret, men en det er selvfølgelig indlysende, at uden en tilstrækkelig dokumentation, er beredskab og arkivering uden nogen mening.

Formidling

Den del af registerinformationen, der finder vej ud i Danmarks Statistiks publikationsserier eller i databankerne skal ikke behandles her, men alene den del af formidlingen, som er går ud på at udnytte registrenes individoplysninger i udredningsopgaver eller forskningsprojekter.

Registerlov og diskretion

Gennemgangen af de principper, der benyttes, er stærkt præget af registerlovgivningen og Danmarks Statistiks diskretionsprincipper. Disse er behandlet i seminarets pkt. 7, men det er et helt afgørende element, at Danmarks Statistik ikke må videregive registerinformationer om individer og virksomheder, således at disse kan identificeres. Formidlingen tager sigte på at overholde disse regler samtidig med, at anvendelsen af data ønskes så intensiv som mulig.

Teknikken er afgørende

Der er forskellige forhold, der har skærpet problemstillingen om registeranvendelse til udredning og forskning. EDB-teknikken har gennemgået en så kraftig udvikling, at det selv på PC er muligt at behandle store datamængder effektivt. I kølvandet på denne udvikling er det statistiske software forbedret ikke blot mht. kapacitet i databehandlingen, men tillige på metodesiden. Der er vel næsten ingen forskningsprojekter, som ikke anvender mulitivariate statistiske analyser under en eller anden form. Det kan være multipel kontingenstabelanalyse, multipel regressionsanalyse m.v.

Non-stop pres

Udviklingen har medført et næsten non-stop pres på Danmarks Statistik for at få adgang til individdata.

Under disse forhold er de gode gamle tabelleverancer stort set uddøde, og det må også indrømmes, at det er uhyre besværligt at kommunikere med f.eks. forskere mhp. at fastlægge de tabeller, der er ønskelige til et projekt. Ressourcerne er ikke tilstrækkelige til at omsætte ønsker til tabeller, idet det er en "trial and error" proces. Det eneste konsekvente er, at forskeren selv

prøver sig frem. Problemet er imidlertid, hvorledes dette kan ske på en efter Danmarks Statistiks synsvinkel sikker måde.

Afledt af disse problemer har Danmarks Statistik i de seneste år udviklet en række serviceordninger specielt for forskere, men også flittigt anvendt af ansatte i ministerier, som er i gang med en udredningsopgave.

Metoderne består af:

1. Modeldata
2. Forskerplaceringsordningen
3. Den elektroniske postkasse

Modeldata

Modeldata er afidentificerede datasæt med ringe identifikationskraft. Hvis datasættet omfatter mange oplysninger om personer eller virksomheder, skal antallet af records være lille (f.eks. 100 records). Omvendt, hvis der er relativt få oplysninger, kan antallet af records være større (f.eks. 1000). Grænsen beror på en afvejning, hvor specielt karakteren af data overvejes. Visse data vil ikke blive udleveret som modeldata, det gælder f.eks. oplysninger fra kriminalstatistikregisteret og Danmarks Statistik oplysninger i forbindelse med data som f.eks. en forsker selv har indleveret til Danmarks Statistik. I det sidste tilfælde kender forskeren individerne, hvorfor det helt klart er en videregivelse af data, dersom Danmarks Statistik tilføjer og tilbageleverer egne data til et sådant modeldatasæt.

Formålet med modeldata er udelukkende at give brugeren en mulighed for at orientere sig i data og f.eks. udarbejde edb-programmer, som kan afvikles på et stort datasæt. Der er en indbygget konflikt i dette, idet det er erfaringen, at enhver forhandling med f.eks. forskere om udlevering af modeldatasæt altid indebærer et ønske fra forskeren om at få et datasæt, som er så omfattende, at der kan udledes egentlige resultater.

Uanset bestræbelserne på at modeldatasæt ikke må kunne føre til identifikation, er der selvfølgelig altid en mulighed for at dette kunne ske. Der underskrives en aftale om datasættets brug, som bl.a. erklærer data fortrolige. En aftaleskitse er vedlagt som bilag 1.

Forskerplaceringsord -ningen

Efterhånden som Danmarks Statistik har følt et stadig stigende pres for udlevering af individdata til forskningsprojekter stod det klart, at der måtte ske en nytænkning af mulighederne for forskeres adgang til data. Omkring 1987 etableredes forskerplaceringsordningen, som gav forskere adgang til en arbejdsplads i Danmarks Statistik.

Tanken var den, at forskerne ved at sidde fysisk i Danmarks Statistik kunne få adgang til ganske detaljerede (omend afidentificerede) data, men at grunddata ikke på noget tidspunkt forlod Danmarks Statistik. Forskeren underskriver en aftale, som ligesom i modeldataordningen præciserer fortrolighed mv. En aftaleskitse er vedlagt som bilag 2.

Ordningen har været benyttet og benyttes rimeligt flittigt ikke blot af forskere men også af ministeriers medarbejdere i forbindelse med udvalgs- eller kommissionsarbejde.

For Danmarks Statistik er der indlysende fordele ved ordningen. Ikke blot løses diskretionsproblemerne på en forsvarlig måde, men kontakten til forskerne er meget tilfredsstillende. Der har dog været en klar tilbøjelighed til, at den enkelte forsker ansætter en studerende til at afvikle edb-kørsler mv., hvilket er ret naturligt, men det bør ikke føre til, at forskerne forsøger at styre projektet fra sin egen arbejdsplads. For forskerne er fordelen den, at de sidder tæt på de medarbejdere i Danmarks Statistik, som har forstand på detaljerne i data. Forskernes problemer består i, at de må arbejde i fremmede omgivelser med et edb-system, de ikke kender på forhånd.

Fra midten af 1993 er der taget nye lokaler i brug, som har forbedret de fysiske rammer for ordningen betydeligt.

Ordningen garanterer ikke mod misbrug, for ved at sidde internt i Danmarks Statistik, bliver der naturligvis mulighed for forskeren at "bortføre" edb-udskrifter, som kunne anvendes til at krænke fortroligheden gennem bagvejsidentifikation. Risikoen må imidlertid anses for beskeden.

Forskernes udgifter til ordningen består i betaling for dannelsen af det grundliggende datasæt, i udgifter til leje af arbejdsplads, konsulentbistand ydet af Danmarks Statistik (i regelen programmørbistand) og i betaling af CPU-tid. Denne sidste omkostning er den, der plager forskerne mest, idet det på forhånd er meget vanskeligt at skønne over forbruget. Der er nok behov for at få udformet en model for betaling af forbruget, som giver forskerne tryghed for ikke at blive præsenteret for ruinerende krav fra Danmarks Statistik, men samtidig sætter nogle rammer op for maksimumforbruget, som Danmarks Statistik kan leve med. I de tilfælde, hvor forskerens penge slipper op, før projektet har nået en forsvarlig afslutning, opstår der et klart dilemma.

Den elektroniske postkasse

Forskerplaceringsordningen har fra forskerside være kritiseret for at være umulig at bruge for forskere udenfor København. Det har dog også vist sig, at selv forskere i København har haft en vis modvilje mod at sidde rent fysisk i Danmarks Statistik.

Inspireret af denne kritik har Danmarks Statistik i 1992/1993 udviklet den elektroniske postkasse. Bortset fra elektronikken minder den meget om tidligere tider arbejdsform på universiteter mv. i hulkorttiden.

Fremgangsmåden er følgende:

1. Danmarks Statistik udarbejder et afidentificeret datasæt til et forskningsprojekt på sædvanlig måde.

2. Forskeren modtager fra Danmarks Statistik et "setup", som definerer datasæt mv. Det svarer til de såkaldte job-kort, der starter enhver edb-kørsel.

3. Forskeren udarbejder sine edb-programmer (p.t. SAS-programmer) på egen PC, og sender disse til en PC i Danmarks Statistik. Denne PC er opstillet i Danmarks Statistiks driftsekpedition, og er en ganske almindelig stand-alone PC, dvs. ikke koblet op til edb-anlægget eller netværket i Danmarks Statistik. Forbindelsen etableres via modem og et BBS (Bulletin Board System).

4. Danmarks Statistik "tømmer" PC'en - foreløbig to gange daglig - og afvikler kørslerne fra en anden PC.

5. Output lægges tilbage på den første PC, og forskeren kan trække dette output hjem til sin egen maskine.

Alt logges

Ikke blot er forskeren aldrig i on-line forbindelse med datasættet i Danmarks Statistik, men alt output "logges", dvs. at Danmarks Statistik har fuld indsigt i, hvad forskeren foretager sig.

Hvilke projekter?

Ordningens tilfredsstillende udnyttelse kræver en særlig arbejdsform og det er givet ikke alle projekter, der er egnede. Projekter med omfattende tabuleringer, f.eks. på spørgeskemadata er oplagte for ordningen, medens projekter med mange kørsler efter "trial and error" metoden, måske er mindre velegnede, idet der ikke svares mere end et par gange om dagen.

En aftaleskitse er vedlagt som bilag 3.

Udviklingsmuligheder?

Sammen med registerudviklingen er der sket en omfattende udvikling i formidlingsmetoderne, som imidlertid kun har skærpet forskernes interesse for at få yderligere adgang til data. Diskussionen er den samme hele tiden. Forskerne ønsker principiel helt fri og uhindret adgang til data og henviser til deres egen interesse i ikke at kompromittere Danmarks Statistiks registersystem. Danmarks Statistik vil i den forbindelse altid blive opfattet som den, der søger at bremse forskningsprojekter, og finder kun megen ringe forståelse for sine synsvinkler.

Tiden (og teknikken) må vise, om der kan udvikles metoder, der kan supplere eller erstatte de ordninger, der allerede eksisterer.

BILAG 1

AFTALE
OM UDLEVERING AF MODELDATA
TIL FORSKNINGSPROJEKT

Mellem Danmarks Statistik og

(forskerens navn)

er indgået følgende aftale om udlevering og brug af modeldata til forskningsprojektet:

(projektets titel og indhold)

For aftalen gælder følgende:

1. (navn) er ansvarlig for de leverede data og for deres anvendelse

2. Data skal opbevares på et edb-anlæg under sådanne sikkerhedsforhold, at kun har adgang til data. har dog ret til at give andre projektdeltagere adgang til data. Denne regel gælder også for opbevaring på PC.

3. Datasættets grundoplysninger skal betragtes som fortrolige oplysninger, jf. Forvaltningslovens § 27, stk. 3 og Straffelovens § 152.

4. Resultater af bearbejdningen af grunddata på papir eller andre medier skal være af en sådan karakter, at det enkelte individ eller den enkelte virksomhed ikke kan identificeres.

5. Skriftlige arbejder, der er baseret på data fra projektet, skal inden offentliggørelse forevises Danmarks Statistik

6. Modeldatasættet tilbageleveres til Danmarks Statistik senest ved projektets afslutning, dog senest den . Der må ikke efter tilbageleveringen forefindes kopier af datasættet udenfor Danmarks Statistik. Det samme gælder for afledte datasæt med oplysninger på individ- eller virksomhedsniveau.

7. Aftalen træder i kraft den.....

BILAG 2

AFTALE OM
FORSKERPLACERING I DANMARKS STATISTIK

Mellem Danmarks Statistik og
(forskerens navn)
er indgået nedenstående aftale om forskerplacering i Danmarks Statistik.

Aftalen vedrører projektet:
(projektets titel og indhold)

For aftalen gælder følgende:

1. Danmarks Statistik giver adgang til at arbejde i Danmarks Statistiks lokaler med de aftalte oplysninger, alt under hensyntagen til Danmarks Statistiks diskretionsprincipper, datasikkerhedsregler og ressourcemæssige situation.

2. skal følge de i nærværende aftale fastsatte bestemmelser samt Danmarks Statistiks datasikkerhedsreglement, som er bilag til denne aftale. Projektets grundoplysninger skal betragtes som fortrolige oplysninger, jf. Forvaltningslovens § 27, stk. 3 og Straffelovens § 152.

3. må kun bearbejde de nævnte oplysninger i Danmarks Statistiks lokaler. Resultater af bearbejdningen på papir eller andre medier må kun medtages ud fra Danmarks Statistiks lokaler efter tilladelse fra forskningskonsulenten eller dennes stedfortræder.

4. Skriftlige arbejder, der er baseret på data fra projektet skal inden offentliggørelse forvisse Danmarks Statistik. Der må ikke offentliggøres oplysninger, hvor det enkelte individ kan identificeres.

5. Danmarks Statistik stiller arbejdsplads til rådighed i et lokale med kontorinventar, telefon og skærmterminal.

6. Projektets omkostninger betales efter de af Danmarks Statistik til enhver tid fastsatte servicetakster.

7. Aftalen træder i kraft den

BILAG 3

AFTALE OM ANVENDELSE
AF DEN "ELEKTRONISKE POSTKASSE"

Mellem Danmarks Statistik og

(forskerens navn)

er indgået nedenstående aftale om adgang til data til forskningsprojektet:

(projektets titel)

1. Datasættet beror i Danmarks Statistik og har adgang til data via Danmarks Statistiks PC-ordning for forskere.

2. Projektets grundoplysninger skal betragtes som fortrolige oplysninger, jf. Forvaltningslovens § 27, stk. 3 og Straffelovens § 152.

3. Der må ikke udskrives individoplysninger (enkelte records) fra grunddata uden tilladelse fra Danmarks Statistik.

4. Skriftlig arbejde, der er baseret på data fra projektet, skal inden offentliggørelse forevises Danmarks Statistik.

   Der må ikke offentliggøres oplysninger, hvor det enkelte individ kan identificeres.

5. Aftalen træder i kraft den

Sygehusbenyttelsesregistret - et eksempel

Lisbeth Laursen

Indhold

Kort fortalt indeholder SBR oplysninger om alle de personer, der har været indlagt på et sygehus i et kalenderår. Oplysningerne modtages fra Sundhedsstyrelsens Landspatientregister. For disse personer er der hentet en lang række baggrundsoplysninger fra øvrige af Danmarks Statistiks registre.

Baggrund

Danmarks Statistik havde i årene forud for registrets etablering været involveret i dannelsen af en række datasæt, hvor Danmarks Statistiks rolle have været at forsyne udtræk, som eksterne rekvirenter medbragte fra Landspatientregistret, med en række baggrundsoplysninger fra Danmarks Statistiks personstatistiske registre. Da det er forholdsvist dyrt og tidskrævende, at foretage sådanne samkøringer ved løsningen af den enkelte opgave, var et af formålene med etableringen af sygehusbenyttelsesregistret, at udvikle et beredskab for en lettere og hurtige dannelse af forskningsdatasæt til brug for belysning af befolkningens forbrug af sundhedsydelser.

Beredskab

Beredskabet på SBR udgøres dels af den store mængde af baggrundsoplysninger, der findes i registret og dels af et standardservicesystem:

- mange data

I SBR findes en lang række oplysninger om de personer, der har været indlagt på et sygehus: familie- og husstandsforhold, boligforhold, uddannelse, beskæftigelse- og indkomstforhold, evt. dødsfaldsoplysninger og oplysninger om modtagelse af indkomsterstattende ydelser og af sygesikringsydelser. Da der i SBR-systemet desuden findes tilsvarende oplysninger om de personer, der *ikke* har været indlagt, er der skabt et datagrundlag, der giver rige muligheder for analyser af sammenhængen mellem sygehusforbrug og sociale forhold.

- standardsystem

Til løsning af mindre serviceopgaver i form af tabeller, er der udviklet et standardservicesystem. Systemet er bygget op omkring DAKOTA, der er et slags menustyret tabelleringssystem udviklet i Danmarks Statistik. I systemet er der defineret en række faste tabelhoveder, der kan vælges imellem og som kan sammensættes med en vilkårlig forspaltevariabel ud fra det definerede record-layout for den fil, der arbejdes på.

Servicesystemet indeholder også muligheder for dannelsen af sumdatasæt, hvor op til 5 brudvariable kan defineres.

Informationshæfte

Med henblik på at udbrede kendskabet til SBR og de muligheder, der ligger i registret er der udarbejdet et informationshæfte, der grundigt redegør for de data, der findes i registret og hvor standardservicesystemet præsenteres med en oversigt over de standardtabelhoveder og forspaltevariable, der kan kombineres. Det er forhåbningen, at disse oversigter kan fungere som inspiration og vejledning for interesserede brugere.

Standardisering

Sygehusforbruget varierer stærkt med køn og alder. Når opgørelser fra SBR skal vurderes vil det derfor ofte være relevant at foretage en køns- og aldersstandardisering af et givent materiale.

Servicesystemet på SBR indeholder ikke på nuværende tidspunkt mulighed for på en enkel måde, at foretage standardiseringer. En udvikling af et sådant system ville være en væsentlig forbedring af servicesystemet og også en facilitet, som brugere af systemet med rimelighed kan forvente.

Dokumentation

Søren Netterstrøm

I de foregående kapitler, er det beskrevet, hvordan der personstatistiske registersystem har opsamlet store mængder af data, og hvordan disse data gennem horisontal og longitudinel integration kan udgøre basis for mange statistiske undersøgelser.

For at data skal kunne anvendes, er det imidertid en forudsætning, at de er dokumenterede. Dokumentation må bestå af både en teknisk og en indholdsmæssig dokumentation. Den tekniske dokumentation omfatter information om lagringsmedie, organisationsform og recordslayout. Den indholdmæssige dokumentation omfatter feltbeskrivelser, beskrivelser af værdisæt mv.

Status

På nuværende tidspunkt, er langt den overvejende del af det personstatistiske registersystem dokumenteret med anvendelse af TIMES. Dokumentationen er tilgængelig som udskrifter fra TIMES, der omfatter både teknisk og indholdmæssig dokumentation. Deruder findes der dokumentation, der beskriver, hvordan de enkelte delregistre er blevet dannet udfra de administrative registre. Der findes ikke en samlet dokumentation af systemet, men de enkelte delregistre er dokumenteret hver for sig.

Selvom den foreliggende dokumentation kan siges at leve op til de formelle krav, således som de fremgår af bl.a. OPUS, er der en række svagheder.

- Dokumentationen er spredt
  Dokumentationen er ikke samlet for hele register-systemet, men dokumentationen for det enkelte delregister forefindes seperat.
- Ændringer over tid, er svære at spore
  Dokumentationen er ofte opdelt på enkeltversioner af registeret. Der findes ikke en oversigt over ændringer fra en version til den næste.
- Kvalitetsdata opsamles ikke.
  Dokumentationen er normalt et produkt af planlægningen af statistikregisteret. Der findes ikke nogen måde, hvorpå erfaringsmateriale og andre former for kvalitetsdata automatisk opsamles.

Et nyt dokumentations-system.

Det er planen i løbet af de kommende år, at opbygge et nyt dokumentations-system til afløsning af TIMES. Projektet, der stadig er i støbeskeen, har fået navnet TIMES-2000. Som for TIMES, vil der være tale om et system, der skal dække Danmarks Statistiks samlede dokumentationsbehov, altså ikke en system der udelukkende retter sig imod det personstatistiske registersystem.

Data i centrum

I forbindelse med udarbejdelse af Edb-systemer, indtager dokumentation en central rolle. Alligevel er det, som om analyse- og programmerings-metoder

til udvikling af systemer har været de elementer i udviklingen, der har tiltrukket sig hele opmærksomheden i metodearbejdet, herunder i de forskellige 'skoler' der gennem årene har behersket debatter, struktureret analyse, struktureret programmering, Yordun, Jackson osv. I denne sammenhæng, har det været antaget, at dokumentationen i stort omfang var lig output fra analysefasen, som så evt. kunne suppleres med driftshåndbøger og brugervejledninger. Resultatet heraf har blandt andet været, at dokumentationen af dataflow og processer har haft en tendens til at blive det centrale, mens dokumentationen af data er reduceret til en teknisk beskrivelse. Med relationsdatabasens fremkomst og anvendelse af tre-skema arkitekturen, er der dog sket en hvis ændring, hvor dokumentationen af data bliver en selvstændig disciplin.
For den registerbaserede statistik udgør dette imidlertid problem. Problemet er, at registeret fra at være et administrativt instrument ændrer karakter og bliver basis for information. Jeg skal i det efterfølgende argumentere for, at dette betyder, at der stilles helt andre krav til dokumentationen af registeret. Jeg skal forsøge at argumentere for, at vi må stille data (eller rettere dokumentationaf data) i centrum.

Hvis vi ikke brugte registre

I traditionelle statistiske undersøgelser, baseret på interviews og spørgeskemaer, er dokumentationen af data i centrum fra første øjeblik. Hele tilrettelæggelsen af undersøgelsen tager sit udgangspunkt i, hvilke spørgsmål der ønskes belyst, hvilket informationsindhold de indsamlede data skal rumme. Udfra dette tilrettelægges interview/spørgeskema, populationen afgrænses osv. Der holdes styr på bortfald osv., således at hele dataindsamlingsfasen resultater i et veldokumenteret statistikregister, der efterfølgende kan gøres til genstand for analyser af forskellig karakter. En detalje der er værd at fremhæve i denne sammenhæng er, at interview- eller spørge-skemaet direkte indgår i dokumentationen, præcist hvordan er de enkelte spørgsmål formuleret.

Vi anvender administrative registre.

Når vi i stedet for anvender administrative registre, står vi imidlertid i en anden situation. Data er ikke indsamlet med statistik som primært formål og data er dokumenteret fra en administrativ synsvinkel. I praksis betyder det ofte, at dokumentationen består at en teknisk dokumentation (recordlayouts) og en kortfattet beskrivelse af de enkelte felter og deres værdisæt. Denne dokumentation er som hovedregel taget som udpluk fra den samlede systemdokumentation (programmeringsgrundlag), evt. følger den samlede systemdokumentation med. Problemet med denne dokumentationsform er, at den er systemcentreret. Formålet med den er, at danne grundlag for konstruktionen af et Edb-system og i denne sammenhæng sikre, at de nødvendige informationer er til stede. Derimod er informationsværdien af de enkelte data kun beskrevet indirekte.
Målsætningen må være, at skabe en dokumentation, der kvalitetsmæssigt og indholdsmæssigt, kan leve op til de krav, vi ville stille hvis vi gennemførte en traditionel survey-undersøgelse.

Dokumentationen må konverteres

Når et administrativt register skal udnyttes som basis for statistik, vil det derfor være nødvendigt, at der laves en ny dokumentation af registeret (eller det afledte statistikregister). Denne nye dokumentation må selvfølgelig tage

sit udgangspunkt i den foreliggende dokumentation, men fokus ændrer sig, idet det centrale bliver dokumentation af registeret og dets informationsindhold, kvalitet osv.

Resten af dette papir vil handle om, hvilke krav der må stilles til denne dokumentation, hvorfor der skal dokumenteres, hvornår, hvem skal bruge dokumentationen og hvad skal dokumenteres.

Hvorfor dokumentation.

Det enkle svar på dette spørgsmål er, at hvis der ikke foreligger dokumentation for et Edb-baseret statistikregister, kan man ikke anvende registeret.
Dette svar kan siges at tage udgangspunkt i den rent data-tekniske side af sagen. Det er klart, at hvis vi ikke kender navnet på det datasæt, der indeholder et bestemt register (en bestemt registerversion), kan vi ikke få fat i data. Hvis vi ikke kender strukturen af de records, datasættet består af, er det heller ikke muligt at hente information ud af data, ligesom vi i den sidste ende må kende betydningen af de koder, som det enkelte felt indeholder, for meningsfyldt at kunne anvende data i registeret. Alt dette er for så vidt banalt.

Man kan imidlertid også sige, at dokumentationen skal foreligge, fordi det er dokumentationen, der gør det muligt at omsætte data til information. Med dette udgangspunkt lægges vægten i højere grad på at dokumentere den indholdsmæssige side af data (informations-potentialet). Som eksempel kan nævnes oplysninger om, hvorledes populationen i registeret er afgrænset i f.eks. tid og rum, en væsentlig oplysning ved fortolkningen af data i registeret. Denne tankegang vil senere blive fuldt op i afsnittet om dokumentationens indhold.

Et vigtigt aspekt ved dokumentationen er, at registre er en fælles ressource. Mange forskellige projekter (og personer) kan trække på det samme register, og anvendelse kan foregå over et ganske langt tidsrum. Kun hvis der foreligger en god og uddybende dokumentation af data, både hvad angår teknik og informationsindhold, er dette muligt. Specielt er det vigtigt at huske på, at ikke alle anvendelser af registeret ud i fremtiden er kendt, dvs. dokumentationen kan ikke begrænses til at dække de øjeblikkelige behov.

Hvornår skal dokumentationen skabes.

Helt generelt kan man sige, at dokumentationen skal skabes så tidligt som muligt i forløbet.
Den ideelle situation ville være, at dokumentationen for informations-indholdet af et register, var udgangspunktet for al udvikling af det Edb-system, der skaber registeret. I praksis er det dog som oftest en analyse af de administrative processer, som registeret indgår i, der er udgangspunktet, og som derfor bestemmer registerets indhold og opbygning, samt dokumentationens form og indhold.
Ved omformningen af et administrativt register til et statistikregister, er det derimod muligt at starte med at dokumentere informationsindholdet i statistikregisteret. Hvilke spørgsmål ønsker vi at kunne besvare, hvilke tabeller og tidsserier ønsker vi at kunne skabe? Hvilken information er nødvendigt for at kunne gøre dette? Vi vender altså processen om, således at vi tager

udgangspunkt i registerets informationsmodel, og så herudfra bestemmer den nødvendige datamodel, under hensyntagen til de tilgængelige datakilder, så kan vi sikre os, at vi har netop det informationsindhold vi ønsker os, samtidig med at vi faktisk har grundlaget for de transformationer af rådata, der udgør en væsentlig del af det system som skal danne statistikregisteret udfra de administrative registre.

Hvem anvender dokumentationen.

I forbindelse med udarbejdelse af dokumentation, er det, som ved alle andre dokumenter der laves, vigtigt at gøre sig klart hvem målgruppen er, hvem der skal anvende informationen.
I forbindelse med udviklingen af det system, der producerer registeret, kommer den første anvendelse af dokumentationen. Foreløbige versioner kan anvendes til at teste, at det påtænkte register indeholder den ønskede informationsmængde. Der kan på basis at sådanne foreløbige versioner ske en dialog mellem de der planlægger registeret og forskellige interessenter (fremtidige brugere).
Når registeret er endeligt fastlagt, er dokumentationen af dette et væsentligt dokument, for dem der skal udvikle de applikationer, der skal danne registeret. Selv om der skal suppleres med andre dokumenter, vil det være af stor betydning, at alle hele tiden kan checke mod definitionen af registeret.
Den primære målgruppe er anvenderne af registeret, dvs. statistikere, statistikbrugere og forskere, der ønsker at skabe information ud fra data i registeret. At tilgodese deres behov for information om registeret og dets informationspotentiale er det primære mål med dokumentationen, på samme måde som det at skabe grundlag for sådanne anvendelser er registerets eget formål. Nogle af disse brugere vil selvfølgelig have været med, allerede under udviklingen af registeret, men andre vil først komme til senere. I denne sammenhæng skal man gøre sig klart, at registeret faktisk er et arkivmateriale, og dokumentation må tilrettelægges herefter.
Endelig er der en ofte overset gruppe, nemlig dem, der anvende resultatateme af statistiske undersøgelser, baseret på registeret. Her bør dokumentation af de bagved liggende data være en tilgængelig kilde til nærmere studier af, hvad det egentlig er, der ligger til grund for en tabel eller tidsserie. Noget vil selvfølge blive formidlet i forbindelse med offentliggørelse, men i forbindlese dybere analyse, er dokumentation af statistikregisteret en vigtig kilde.

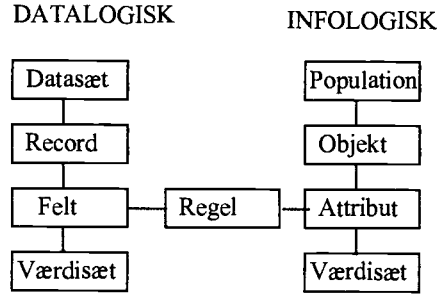
Hvad skal dokumenteres.

Hvad der skal dokumenteres, eller indholdet af dokumentationen, må selvfølgelig ses i lyset af hvem, der skal anvende dokumentationen, eller rettere hvad dokumentationen skal anvendes til. Der kan her fremdrages to forskellige synsvinkler, en datalogisk synsvinkel og en infologisk synsvinkel.
Den datalogiske synsvinkel handler om den fysiske opbygning af registeret. Er det flade filer eller en database. Hvad hedder filerne/tabellerne, hvad er opbygningen af records/tabeller og formatet af de enkelte felter/kolonner. Denne type information er nødvendig for at vi kan skrive programmer, der håndterer registeret.
Den infologiske synsvinkel handler om informationsindholdet i registeret. Hvad er populationen, tidsafgrænsningen. Hvilke objekter er beskrevet. Hvilken information findes om det enkelte objekt. Hvordan er denne information repræsenteret, dvs. hvilke kodesystemer er anvendt til at beskrive de enkelte informationsdele. Hvad er kvaliteten af data, både registreret som

helhed og enkeltinformationer. Denne type af information er nødvendig for at kunne tilrettelægge en statistisk undersøgelse med udgangspunkt i et register. Generelt kan man sige, at enhver udnyttelse af data til at skabe information, skal ske udfra den infologiske model.

Skal der laves to sæt dokumentation?

Spørgsmålet er derfor, om det er nødvendigt at lave to forskellige dokumentationssystemer, et der anskuer registeret udfra en datalogisk synsvinkel, og et der anskuer registeret udfra en infologisk synsvinkel? Fordelen ved at bygge to systemer er indlysende, idet systemerne umiddelbart vil kunne tilrettelægges efter de behov der er. Ulempen er, at der skal vedligeholdes to systemer og at der er stor risiko for, at systemerne ikke hænger sammen, at ændringer i det ene system ikke slår igennem i det andet.
Imidlertid er der en tæt sammenhæng mellem de to systemer. Et system, der beskriver den infologiske model, må indeholde en beskrivelse af, hvordan et infologisk begreb udledes af de fysiske data. Hvis vi f.eks. har et fysisk data som CPRNR, vil vi have de infologiske begreber KØN og ALDER pr. dato, der dannes fra CPR ved hjælp af simple regler. Tilsvarende kan vi have et fysisk data, INDKOMST (i hele kroner), og udfra dette danne et infologisk begreb INDKOMSTGRUPPE udfra en simpel regel. Et forsøg på at illustrere denne tankegang kan være nedstående skema.



Skemaet er et forsøg på lave et hieraki for henholdsvis det datalogiske og det infologiske syn på et register.
Forbindelsen mellem de to sider er som antydet, de regler, der omformer data til information.
Det er imidlertid værd at bemærke, den ensartethed der er mellem de sider i deres opbygning. Først og fremmest er der værdisæt, der optræder i begge hierakier, som det laveste niveau. Ikke bare navnet vil være det samme, også selve informationen vilhave samme opbygnnig, enten som et tilladt værdiområde (numerisk, positiv eller nul, mellem 0 og 100 etc) eller som en værdiliste med tilhørende tekst (1 = Mand, 2 = Kvinde). I de tilfælde, hvor det infologiske begreb er en direkte kopi af data, f.ek.s INDKOMST (infologisk) = INDKOMST (datalogisk) eller KOMMUNE (Infologisk) = KOMMUNE (datalogisk) vil indholdet også være det samme. En del kunne derfor tale for at ændre modellen, således at værdisæt kun optræder en gang.

Ser vi på de lag, der ligger øverst i modellen, vil situationen ofte være, at der er helt banale sammenhænge mellem objekt og record, population og datasæt. Helt banalt, men nok ganske almindeligt, er at der er et datasæt = en population der indeholder en recordtype, en type af objekter. Fordelen ved

modellen er imidlertid, at dette ikke er påkrævet. Det vil være muligt at sammenstille informationen om objekter fra flere forskellige records fra flere forskellige datasæts, ja det vil endog være muligt at udlede en enkelt atrribut på basis af flere felter beliggende i forskellige records i forskellige datasæt.

Undersøgelser der bygger på flere registre.

Det vil her være nødvendigt at komme ind på den situation, at en undersøgelse kombinerer data fra flere statistikregistre. Der er her to mulige fremgangsmåder. Enten at danne et nyt statistikregister (datasæt) ved at samkøre de oprindelige registre. Men en anden mulighed vil være, at foretage en infologisk sammenkørsel, dvs. beskrive det infologiske indhold på baggrund af de oprindelige registre. Den sidste fremgangsmåde vil efterhånden som vi flytter vores registre fra flade filer til databaser blive langt mere naturlig, selvom der i teorien ikke er noget i vejen for at gøre sådan allerede med den nuværende teknologi.
Udgangspunktet vil altså være, at danne en ny infologisk population udfra de oprindelige, og udfra dette skabe nye objektdefinitioner og attributter til disse. I praksis vil i hvert fald attributterne i stort omfang kunne kopieres direkte fra de oprindelige definitioner, men ofte vil også objekterne være de samme (da sammenkobling eller kan være vanskelig at foretage, ligesom populationen må være enten en delmængde af en af de oprindelige populationer eller foreningsmængden af disse.

Præcis hvordan dette skal tackles, er ikke afklaret på nuværende tidspunkt. Noget kunne tale for, at beskrive sammenhængen mellem objekter og records, hvor en sådan findes. Dette vil så betyde, at attributter til et objekt naturligt udledes af felter i den eller de records, objektet er knyttet til, og at selve sammenhængen mellem flere registre beskrives på objekt/record niveau. Man kan argumentere for, at udvide den datalogiske side med begrebet VIEWS, sådan som det kendes fra relationsdatabaser. Et VIEW vil arve de felter fra de underliggende records/tabeller, der omfattes at VIEWet, men samtidig kunne knyttes til objektbegrebet på en eentydig måde.

Tidsdimensionen

Et statistikregister vil som regel blive dannet i en række versioner, der hver for sig beskriver den samme population på forskellige tidspunkter eller dækkende begivenheder i et bestemt tidsrum. Der kan være tale om at de enkelte versioner dannes helt uafhængigt af hinanden (Fødte xte kvartal 19xx), eller at senere versioner dannes som opdateringer af den foregående (uddannelsesregisteret).

Databrud

I forbindelse med dokumentationen er der her særlige problemer, der må takles, både på den datalogske og infologiske side. Problemerne opstår, fordi det ikke er muligt at fastholde alle definitioner over et længere tidsforløb. Tager vi som eksempel uddannelses-statistikken, vil der opstå nye uddanelser og andre vil ophøre, ting der må afspejle sig i værdisættet. Registre vedrørende skatteforhold vil ændre sig i takt med ændrede skatteregler osv.
Af hensyn til mulighederne for at lave tidsserier og anden statistik med sammenhæng over tid, er det imidlertid af stor betydning at kunne fastholde de infologiske begreber/attributter over så lange perioder som muligt. Her hjælper den overfor opstillede opsplitning i en datalogisk og infologisk del imidlertid. Ofte vil et infologisk begreb kunne fastholdes på den måde, at

definition og værdisæt er uforandret, det det ændres er de regler der anvendes til at danne informationen.
Det vil ved opbygningen af et informationssystem imidlertid være vigtigt, at systemet kan håndtere denne dimension på en sådan måde, at det er muligt at fastholde begreber over tid og samtidig få overblik over, hvilke ændringer der er sket i de underliggende regler og data.

Livsforløb.

I forbindelse med dokumentation af statistikregistre, kan tidsdimensionen også anskues fra en anden vinkel. Der har i litteraturen om EDB-systemer ofte været fokuseret på en systemets livsforløb (life-cycle). Hvad der måske er overset i denne sammenhæng er, at udover systemernes livsforløb, kan der opstilles et data livsforløb, der i hvert fald fra en statistik synsvinkel er mere interessant. Vi vil ofte se, at det system der danner et register ændrer sig, f.eks. omlægges til ny teknologi, uden at dette grundlæggende ændrer på de data, registeret indeholder. Vi vil på den anden side kunne opleve, at data ændrer sig uden at dette fra en systemsynsvinkel er andet end en mindre justering. Der kan derfor være grund til at adskille system- og data-dokumentation fra hinanden, at opfatte disse to livsforløb som uafhængige. Desuden kan der så opstilles et tredie livsforløb, nemlig det infologiske. Selvom data ændrer sig, behøver dette ikke grundlæggende at ændre det infologiske syn på data, som beskrevet ovenfor.
En anden side af denne sag er, at dokumentationen af data og specielt den infologiske dokumentation, har interesse langt udover levetiden for de systemer, der danner data, ja muligvis udover levetiden for dataene. Denne dokumentation er jo også en del, at dokumentationen for de resultater der formidles, f.eks. i form af tidsserier og i databanker.

Datakvalitet, informationskvalitet.

Et vigtigt aspekt ved dokumentationen af statistikregistre, er at give information om kvaliteten af de data (informationer), registeret indeholder.
Kvaliteten af data kan omhandle registeret som helhed. F.eks. kan der være problemer med samtidighed, dvs. om data i registeret er fuldt opdateret i forhold til den periode/dato registeret omhandler eller problemer med fuldstændighed. Desuden bør informationer om registerets administrative anvendelse medtages her, da denne information er væsentlig for at afdække, om bestemte problemstillinger med rimelighed kan belyses gennem anvendelse af registeret. Disse ting vedrører normalt registeret på tværs af registerversioner (tid). Der bør derudover findes en log over de enkelte registerversioners tilblivelse, hvor evt. problemer med indflydelse på datakvaliteten registreres. Loggen bør indeholde information om alle registerversioner, dvs. også eksplicit indeholde information om, at der ikke er noget at bemærke.
Kvalitetsdata vil derudover typisk findes på felt / attribut niveau. Her vil der normalt være tale om tidsuafhængige bemærkninger. Der vil her være tale om generelle bemærkinger, f.eks. at kvaliteten af data er lav, fordi data ikke har betydning for den administrative proces, eller at data er af svingende kvalitet, fordi data indberettes decentralt og instrukserne til indsamlerne ikke er tilstrækkelig klare eller ikke efterleves. Derimod vil bemærkninger på dette niveau sjældent handle om enkelte registerversioner.

Hvis der har været særlige kilder, f.eks. stikprøveundersøgelser, der har været anvendt til at vurdere datakvaliteten, er det vigtigt at henvise direkte til disse og de opnåede resultater.

Registerets anvendelse.

Det er allerede nævnt, at registerdokumentation er en vigtig informationskilde, hvis man vil gå tættere på de resultater, der publiceres gennem anvendelse af registerdata.
I denne forbindelse spiller det infologiske syn på registeret den afgørende rolle. Det er udfra den infologiske information, at alle tabeller og analyser af registeret må foretages. Omvendt må den infologiske information dække alle de begreber, grupperinger osv. der faktisk anvendes i forskellige anvendelser af registeret.
Det er vigigt at fastholde disse krav, fordi de sikrer, at resultater fra forskellige analyser baseret på registeret anvender de samme definitioner på forskellige begreber. En konsekvens er, at den infologiske information om registeret udvides efterhånden som registeret anvendes, idet nye infologioske begreber kan tilføjes.

I forbindelse med registerets anvendelse, bør der også føres log over denne, således at der i register-dokumentation er en henvisning til databanker og publicerede tabeller, der dannes på basis af registeret, med information om registerversionen og de attributter, der er anvendt. Denne dokumentation er knyttet til den infologiske model.
I det omfang registeret sammen med andre registre er anvendt i andre infologiske modeller, dvs. undersøgelser der bygger på mere end et register, bør der også være henvisninger til disse.

Times 2000 i forhold til det personstatistiske registersystem.

I det foregående har der været tale om en generel diskussion af dokumentationsproblemetikken for statistiksystemer. Imidlertid gælder disse betragtninger også på det personstatistike registersystem. Der er imidlertid i forbindelse med dette en række særlige forhold, som er værd at knytte nogle bemærkninger til.

Først er fremmest er det værd at hæfte sig ved, systemet med sin opbygning af basisregistre og klassifikationsmoduler er en oplagt kandidat for den opdeling i infologisk og datalogisk synsvinkel, som der tidligere er argumenteret for, idet den infologiske synsvinkel typisk vil inddrage data fra et eller flere klassifikationsmoduler sammen med et enkelt basisregister. Da både basisregistre og klassifikationsmoduler bygger på CPR-nummeret som nøgle, må muligheden for en infologisk sammenkobling ligge inden for rækkevidde. Dette gælder iøvrigt også evtuel sammenkobling mellem flere basisregistre.

Desuden er det karakteristisk for systemet, at der i stort omfang vil være 1 til 1 relationer mellem datasæt og record på den ene side og population og objekt på den anden side, idet det er op til de systemer, der danner statistik-registrene udfra de administrative data, at skabe disse 1 til 1 relationer.

Udfra de hidtige anvendelse af systemet er det også klart, at det er vigtigt at dokumentationssystemet kan holde styr på forandringer i definitioner over tid, specielt den type, hvor det infologiske begreb ikke ændrer sig, men hvor de bagved liggende data har ændret sig. En særlig variant her vil være, at der heller ikke er en ændring på datalogisk niveau, men at den faktiske ændring ligger i det system der danner statistikregisteret. Her vil det være vigtigt at dokumemtere sådanne 'ændringer', da de kan have betydning for tolkningen af samme data over tid.

Hvordan får vi skabt metadata.

Et problem med alle metadatasystemer synes at være, hvorledes man får data ind i systemet og sikrer at data er vedligeholdt. Det synes at være et problem, specielt hvis de der leverer metadata ikke er de samme, som dem der anvender metadata.
I forbindelse med TIMES-2000, vil dette problem blive forsøgt tacklet på forskellige måde.
Data-indholdet kan opdeles i tp klasser, hårde data og bløde data. Hårde data er recordlayouts, værdisæt på listeform og regler, der er skrevet i en særlig syntax. Karakterisk for disse data er, at de kan behandles og anvendes maskinelt. Bløde data er beskirvelser, kvalitetsdata, bemærkninger om databrud osv. Disse data kan opbevares maskinelt, men kan ikke gøres til genstand for en maskinel ananlyse.

Erfaringerne fra TIMES (og DAKOTA) tyder på, at et vigtigt incitament for leverandørerne vil være, hvis data umiddelbart kan genbruges i forbindelse med edb-opgaver mod registeret, dvs. hvis metadata direkte kan anvendes i forbindelse med tabellering, listning og andre processer mod registeret. Det vil da også være et væsentlig mål for TIMES-2000, at skabe sådanne muligheder, dels gennem forbehandlere til f.eks. SAS, dels gennem at vores egne standardsystemer direkte skal kunne trække på metadata fra TIMES-2000. Målet må være, at det er muligt at bestille tabeller mv. udfra de infologiske begreber som metadatabasen indeholder.

Desværre dækker ovenstående kun de hårde data. Hvad angår de bløde data, vil det i praksis blive vanskeligt, at få samlet alle data op. De hårde data udgør imidlertid en godt skelet for systemet. Den overordnede struktur med datasæt, record, felt, population, objekt, attribut og værdisæt skal være på plads, for at de hårde data kan beskrives.
Alle bløde data i systemet er karakteriseret ved, at de knytter an til denne struktur, idet de handler om eller beskriver et bestemt element i denne struktur, hvadenten det er et datasæt eller en attribut.
Det er tanken, at de bløde data i systemet skal ligge som notatfelter, knyttet til det enkelte element, idet der til hvert element kan være en vilkårligt antal notater, evt. opdelt i forskellige klasser af notater.
Mange bløde data, vil på et eller andet tidspunkt opstå i form af tekst i et eller ander notat eller ander tekstbehandlingsdokument. Der må skabes en mulighed for, fra Words at kunne overføre et dokument, eller dele heraf, direkte til TIMES-2000. I praksis kan man tænke sig en parallel til

postsystemet, idet forskellen er, at modtageren er det element i TIMES-2000, som noten skal knyttes til. Man kan sige, at TIMES-2000 i denne sammenhæng skal være det naturlige arkivsystem eller journalsystem for sådanne noter.

Det er imidlertid lige så vigtigt, at der aktivt sættes en kampagne ind for, at sådanne data skal skabes og at en sådan kampagne følges op. Det vil stille krav til ledelsen på alle niveauer, aktivt at arbejde for, at det at data er veldokumenteret er en vigtig del af vores samlede aktivitet i Danmarks Statistik.

Bilag a: Program for seminar over emnet "Det personstatistiske registersystem", 10.-13. januar 1994 på Gentofte Hotel.

Program:

(*O = organisator, H = hovedindlæg, D = diskutant, X = supplerende indlæg*)

Mandag den 10. januar

kl. 09.30 - 10.30 1. Kravene til den officielle statistik

O: Finn Spieker
H: Lars Thygesen
D: Hans E. Zeuthen

kl. 10.45 - 12.00 2. De administrative registre i Danmark

O: Lars Borchsenius
H: Carsten Torpe
D: Finn Spieker

kl. 13.30 - 21.00 3. Statistikken og registrene

kl. 13.30 - 15.30 **3.1. Registeroplysninger til statistikformål**

O: Anitta Lange
H: Vøgg Løwe Nielsen
D: Lars Thygesen

kl. 16.00 - 17.30 **3.2. Samkøringer**

O: Lars Thygesen
H: Claus Ib Olsen
D: Sven Egmose

kl. 19.30 - 21.00 **3.3. Statistiksystemet**

O: Otto Andersen
H: Finn Spieker
D: Carsten Torpe

Tirsdag den 11. januar

kl. 09.00 - 22.00 4. Statistikgrundlaget

kl. 09.00 - 10.30 **4.1. Administrative data som statistikdata**
O: Bjarne Simonsen
H: Lars Thygesen
D: Anita Lange
X: Anna Qvist
X: Søren Hostrup-Pedersen

kl. 10.45 - 12.00 **4.2. Anvendelse af flere kilder**
O: Otto Andersen
H: Gunvor Højberg
D: Anna Qvist

kl. 13.30 - 15.30 **4.3. Integreret dataindsamling**
O: Finn Spieker
H: Søren Hostrup-Pedersen
D: Finn Spieker

kl. 16.00 - 17.30 **4.4. Imputering**
O: Palle Qvist
H: Lone Solbjerghøj
D: Lars Borchsenius

kl. 19.30 - 22.00 **4.5. Surveys og registre**
O: Sven Egmose
H: Marius Ejby Poulsen
D: Olaf Ingerslev
X: Bo Møller

Onsdag den 12. januar

kl. 09.00 - 15.30 5. Integrationsregistre

kl. 09.00 - 12.00 **5.1. Horisontal integration**
O: Lars Borchsenius
H: Jørn Daugård Pedersen
D: Finn Spieker
X: Lene Skotte

kl. 13.30 - 15.30 **5.2. Vertikal integration**
O: Carsten Torpe
H: Otto Andersen
D: Britta Kyvsgård
X: Leo Jensen
X: Lisbeth B. Knudsen
X: Søren Leth-Sørensen

kl. 16.00 - 18.00 6. Personstatistikkens samspil med andre statistikområder

O: Vøgg Løwe Nielsen
H: Poul Jensen
D: Peter Maskell

Torsdag den 13. januar

kl. 09.00 - 10.30 7. Registerlovgivning og datapolitik

O: Lars Thygesen
H: Finn Spieker
D: Henrik Waaben

kl. 10.45 - 12.00 8. Beredskab og formidling

O: Søren Hostrup-Pedersen
H: Otto Andersen
D: Peter Maskell
X. Lisbeth Laursen

kl. 13.30 - 15.00 9. Dokumentation

O: Lisbeth Laursen
H: Søren Netterstrøm
D: Kirsten Wismer

kl. 15.00 - 15.30 Afslutning

Bilag b: Deltagerliste for seminar over emnet "Det personstatistiske registersystem", 10.-13. januar 1994 på Gentofte Hotel.

Deltagerliste

F Fuld deltagelse B Begrænset deltagelse

	Fra Danmarks Statistik	Mandag	Tirsdag	Onsdag	Torsdag
F	Lars Borchsenius	X	X	X	X
F	Anita Lange	X	X	X	X
F	Anna Qvist	X	X	X	X
F	Leo Jensen	X	X	X	X
F	Vøgg Løwe Nielsen	X	X	X	X
F	Bjarne Simonsen	X	X	X	X
F	Lone Solbjerghøj	X	X	X	X
B	Gert Schmidt		X		
B	Kaj Kammer Madsen		X		
F	Kirsten Wismer	X	X	X	X
F	Palle Qvist	X	X	X	X
F	Søren Netterstrøm	X	X	X	X
B	Anne Nærvig Petersen				X
B	Sten Mogensen		X		
F	Carsten Torpe	X	X	X	X
F	Lisbeth Laursen	X	X	X	X
F	Jørn Daugård Pedersen	X	X	X	X
B	Poul Henning Larsen			X	
F	Søren Hostrup-Pedersen	X	X	X	X
B	Thorkild Wedebye		X		
B	Leif Nielsen		X		
B	Annette Jerlach			X	
F	Sven Egmose	X	X	X	X
B	Bo Møller		X	X	
F	Finn Spieker	X	X	X	X
F	Marius Ejby Poulsen	X	X	X	X
F	Gunvor Højberg	X	X	X	X
F	Otto Andersen	X	X	X	X
F	Lisbeth B. Knudsen	X	X	X	X
B	Søren Leth-Sørensen			X	
F	Hans E. Zeuthen	X	X	X	X
F	Lars Thygesen	X	X	X	X
	Eksterne deltagere				
F	Olaf Ingerslev, AKF	X	X	X	X
B	Britta Kyvsgård, Kriminalistisk Institut, Københavns Universitet			X	
B	Peter Maskell, Handelshøjskolen i København			X	X
B	Henrik Waaben, Registertilsynet				X