



Registerstatistik

– administrativa data för statistiska syften

Statistics Sweden

Statistiska centralbyrån

**Befolknings-
registret**

Uppgiftslämnarens namn och postadress

Uppgiftslämnarens person-/organisationsnr

Rättelsedatum om rättad uppgift sänds in (ex. 0 301)

Delägare m.fl. i firmans företag

Arbetsställenummer från SCB

Skatt

Avdragen A-skatt

Tjänst

Kontant lön m.m.

03 Kontant bruttolön m.m.

Kontrolluppgifter från arbetsgivaren m.fl. för inkomståret 2004 **KU 10**

Inkomsttagarens person-/organisationsnr

08 Fr.o.m. 09 T.o.m. 570 Specifikationsnr

Del av år (ex 05–11)

Inkomsttagarens fullständiga namn och postadress

Förmåner m.m.

41 Bostads småhus 42 Kost 43 Bostads ej småhus 44 Ränta

45 Parkering 47 Annan förmån 48 Förmån har justerats

Kod för förmånsbil

Siffrorna i vänsterkanten motsvarar de siffror som finns i inkomstdeklarationen under avsnitt Tjänst

**Företags-
registret**

Anders Wallgren & Britt Wallgren

INLEDNING

TILL

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.

Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.

Föregångare:

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Efterföljare:

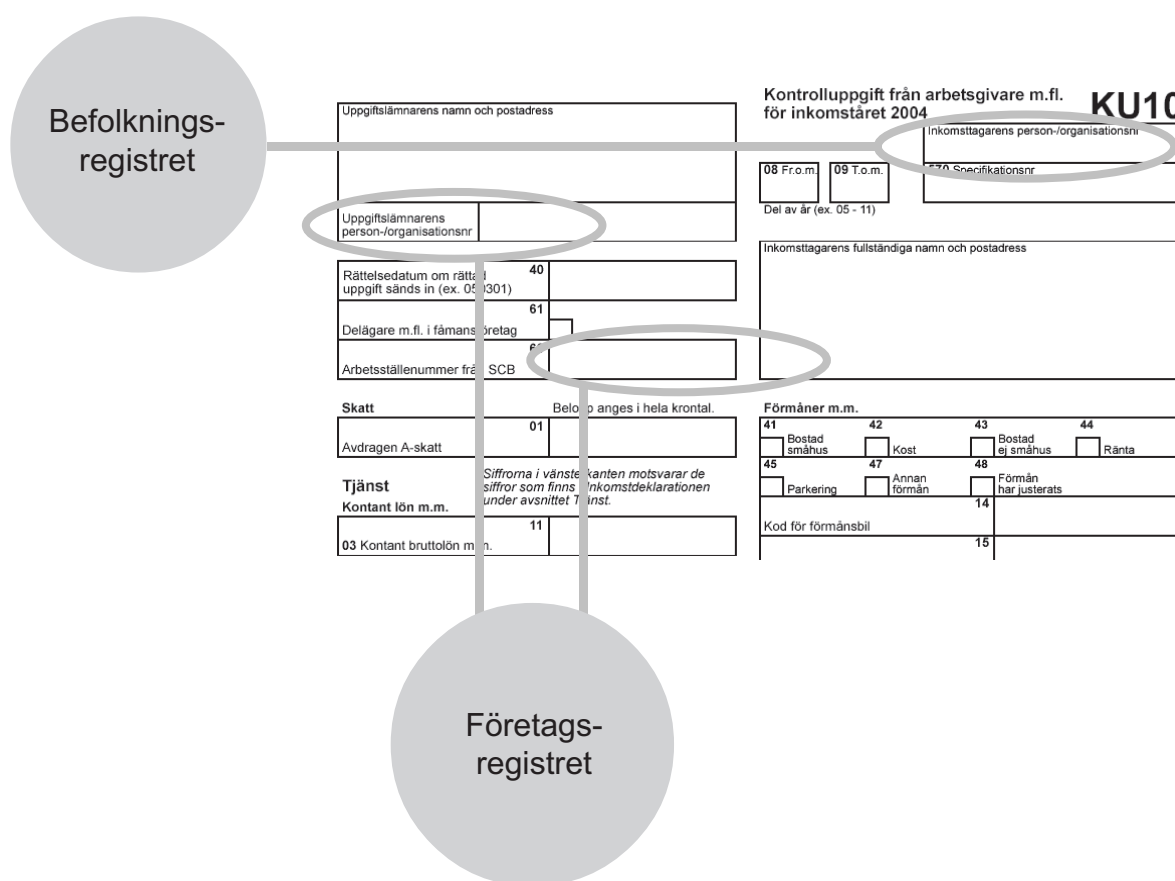
Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report 2004:2. Registerstatistik – administrativa data för statistiska syften / Anders Wallgren; Britt Wallgren.

Digitalt skapad fil, anpassad efter de digitaliserade delarna i serien. Statistiska centralbyrån (SCB) 2016.

Registerstatistik

– administrativa data för statistiska syften



R&D Report 2004:2

Research – Methods – Development

Register statistics

– administrative data for statistical purposes

Producent Statistiska centralbyrån, Statistics Sweden, Metodenheten
Box 24300, SE-104 51 STOCKHOLM

Förfrågningar Anders Wallgren och Britt Wallgren
anders.wallgren@scb.se
britt.wallgren@scb.se
Telefon: 019-17 61 98

Om du citerar ur denna publikation, var god uppgi: Källa: SCB, + publikationens namn

©2004, Statistiska centralbyrån
ISSN 0283-8680

Printed in Sweden

SCB-Tryck, Örebro 2004.09  MILJÖMÄRKT Trycksak 341590

Förord

Registerstatistik – administrativa data för statistiska syften är resultatet av ett arbete som inleddes efter att Registerprojektet vid SCB avslutades. Registerprojektet startade 1996 som ett verksprojekt med följande huvuduppgifter:

”SCB:s olika register som bygger på administrativa källor skall utvecklas till ett väl fungerande registersystem. Systemets olika delar skall vara klart definierade och genom ökad samverkan skall systemet effektiviseras så att nya integrationsregister kan skapas för att möta nya krav från användarna. Det registerstatistiska metodarbetet skall synliggöras och utvecklas så att registerstatistikens kvalitet kan beskrivas och förbättras.”

Projektgruppen redovisade en slutrapport som först remissbehandlades inom SCB. Därefter diskuterades den av verksdirektionen för att slutligen ligga till grund för ett GD-beslut i maj 2001. GD-beslutet innebar bland annat att Utvecklingsavdelningen fick i uppdrag att skriva en handbok i registerstatistik. Detta uppdrag anförtroddes Anders och Britt Wallgren.

Svensk officiell statistik bygger sedan länge på en omfattande användning av administrativa data för att bygga upp olika statistiska register. Dessa register används flitigt både som bas för våra urvalsundersökningar men också för att ta fram totalräknad registerstatistik av olika slag.

Under slutet av 1960-talet och början av 1970-talet diskuterades på SCB ett arkivstatistiskt system och en omorganisation av SCB med standardiserade input- och outputprocesser kopplade till ett centralt datalager. Av olika skäl genomfördes dock inte dessa förslag.

Under slutet av 1980-talet framförde personer inom Finansdepartementet synpunkter på att SCB:s register borde utnyttjas bättre. Inom SCB sågs emellertid registren som avdelningarnas och enheternas angelägenheter och registren betraktades ofta ur rent IT-perspektiv. 1993 började man inom Utvecklingsavdelningen att inventera vilka metodstudier som gjorts inom den registerbaserade statistiken. I dessa metodstudier beskrivs endast problem inom respektive register utan att hänvisa till ett gemensamt registersystem.

Danmarks Statistik publicerade 1994 *”Personstatistik i Danmark - Et registerbaseret statistiksystem”*. I samband med att denna bok kom ut, ordnades ett nordiskt möte där registerstatistikens situation i de olika länderna diskuterades.

Den danska boken och diskussionerna under mötet inspirerade de svenska deltagarna till att undersöka hur SCB:s registersystem såg ut och hur det fungerade. Resultatet av den första inventeringen var att det saknades en helhetssyn på registersystemet. Detta medförde att de strategiska möjligheter, som ett effektivt system med samverkande register skulle kunna innebära, inte utnyttjades.

Med den danska modellen som bakgrund utvecklades inom projektet en modell som beskriver förhållandena inom SCB. Det faktiska registersystemet jämfördes med det teoretiskt lämpliga systemet och man kunde då se vilka utvecklingsinsatser som borde göras.

Till Registerprojektet knöts en grupp med ca 20 deltagare från olika register inom SCB. Denna kontaktgrupp var mycket betydelsefull för projektarbetet. De som arbetade inom projektet kunde via kontaktgruppen diskutera olika registerproblem med personer som hade mångårig praktisk erfarenhet av registerstatistik. Genom att låta deltagare i kontaktgruppen och andra inbjudna berätta om sina register och erfarenheter spreds kunskaper om olika register av allmänt intresse.

I Registerprojektets slutrapport föreslogs som redan nämnts att en handbok i registerstatistik skulle tas fram. Trots erfarenheterna från Registerprojektet visade sig arbetet med boken svårare än väntat. Huvudorsaken till detta är att en gemensam etablerad teori för registerstatistik saknas. En gemensam terminologi som beskriver vad som är specifikt för området saknas även internationellt. Som en följd av detta är det knapphändigt med skriftlig dokumentation av de metoder som används. Dessutom är många metoder lokala ad hoc-lösningar och praxis och synsätt kan också variera mellan SCB:s olika enheter.

Det har därför inte varit möjligt att skriva en traditionell handbok som beskriver utprovade och allmänt vedertagna registermetoder utan istället har två böcker/rapporter tagits fram. *Registerstatistik – en introduktion*, ges ut i Utvecklingsavdelningens handboksserie och är av mer beskrivande karaktär. *Registerstatistik – administrativa data för statistiska syften* har en högre ambitionsnivå och lägger fram idéer, metoder och förslag till ett effektivt registerstatistiskt system.

I denna bok ges många exempel från dagens SCB för att illustrera registerstatistiska metodproblem. Betoningen ligger hela tiden på metodarbete och systemtänkande.

Ett syfte är att beskriva registersystemet och diskutera hur det bör fungera och vilka möjligheter ett väl fungerande registersystem kan ge. Ett annat syfte är att strukturera och beskriva det registerstatistiska metodarbete som ligger till grund för skapandet av systemet. I flera fall föreslår författarna en ny terminologi.

Boken kan ses som en början till ett mer systematiskt arbete kring registerstatistiska frågor. Min förhoppning är att den skall stimulera till vidareutveckling inom detta viktiga område samt till översyn och förbättringar av nuvarande arbetssätt vid SCB. Den vidareutveckling som behövs, bland annat för få bättre konsistens mellan olika registerbaserade undersökningar, är en flerårig uppgift.

Boken är i första hand skriven för dem som arbetar med att framställa samhällsstatistik som baseras på administrativa register. Då många av de principer, som beskrivs i boken, är av generell natur har den också ett vidare intresse.

De tre första kapitlen i *Registerstatistik – administrativa data för statistiska syften* behandlar det system av statistiska register som är grunden för huvuddelen av den samhällsstatistik som produceras i Sverige. De övriga kapitlen behandlar den statistiska metodik som används, eller kan användas, när statistiska register skapas och registerstatistik framställs.

Som redan nämnts är Anders och Britt Wallgren författare till boken och ansvariga för dess innehåll. Lars-Göran Lundell har skrivit kapitel 9 och avsnitt 11.2 och har också deltagit aktivt i diskussionerna av flera andra avsnitt. SCB:s chefsjurist Birgitta Pettersson har granskat avsnittet om imputering i kapitel 6, samt kapitel 10 som behandlar integritetsskydd.

Professor Carl-Erik Särndal har varit en mycket viktig samtalspartner under författarnas arbete med boken. Hans breda erfarenhet från statistikbyråer i olika länder och hans utgångspunkt som specialist på urvalsundersökningar har varit till stor nytta. Därutöver har ca 50 personer inom SCB läst och kommenterat olika delar av boken. Flera personer har dessutom intervjuats för att ge material till olika exempel och metodavsnitt. De synpunkter och den uppmuntran som alla dessa personer givit har betytt mycket för arbetet med boken.

De statistiska centralbyråerna i Norden och Nederländerna har sedan 2001 haft ett samarbete kring registerstatistiska frågor. Vissa delar av innehållet i denna bok (kapitel 1-3, 6 och 7) har presenterats och diskuterats vid seminarier med detta registerstatistiska nätverk.

Det är min förhoppning att *Registerstatistik – administrativa data för statistiska syften* med dess förslag kommer att stimulera diskussionerna om registerstatistik, främst inom SCB men också utanför verket. Jag hoppas också att den skall bidra till ett ökat systemtänkande inom registerområdet.

Ulf Jorner

Chef för Utvecklingsavdelningen

Innehåll

Kapitel 1. Vad är registerstatistik?	10
1.1 Behövs en registerstatistisk teori?	10
1.2 Vad är en statistisk undersökning?	11
1.3 Vad är ett register och vad är registerstatistik?	12
1.3.1 Registerstatistik – ett exempel	13
1.3.2 Hur gjordes denna registerbaserade undersökning?	15
1.4 En modell av SCB:s statistikproduktion	17
1.5 Hur kan man använda SCB:s registersystem?	18
1.5.1 Systemet används för registerstatistik	18
1.5.2 Registersystemet används för urvalsundersökningar	18
1.5.3 Register och urval kompletterar varandra	19
1.6 Varför skall statistik baseras på administrativa data?	19
1.6.1 För- och nackdelar med surveystatistik och registerstatistik	20
1.6.2 Kostnadsaspekten – är registerstatistik billigt?	21
1.6.3 Register, integritet och allmänna opinionen	22
1.7 Administrativa och statistiska informationssystem	22
1.7.1 Olika syften	23
1.7.2 Olika roller för enskilda objekt	23
1.7.3 Synen på fel	23
Kapitel 2. Hur strukturera ett registersystem?	24
2.1 Objekttyper och relationer	24
2.2 Systemet av basregister	28
2.2.1 Befolkningsregistret	29
2.2.2 Aktivitetsregistret	29
2.2.3 Företagsregistret	31
2.2.4 Fastighetsregistret	33
2.3 Grundläggande variabler i registersystemet	34
2.3.1 Länkar	34
2.3.2 Klassifikationer	34
2.4 Hela registersystemet	35
2.5 Statistiska registersystem utanför SCB	37
2.5.1 Registerstatistik i andra länder	37
2.5.2 Statistiska register hos andra statistikansvariga myndigheter	38
2.5.3 Statistiska register inom företag	38
Kapitel 3. Hur skapas ett registersystem?	39
3.1 Administrativa register omvandlas till statistiska register	39
3.2 Systemet får data från många administrativa källor	42
3.3 Helheten skapas genom samordning och samarbete	44
3.4 Skapa ytterligare register och användningar	46
3.4.1 Personstatistik baserad på SCB:s registersystem	46
3.4.2 Varför är denna del av systemet så bra?	47
3.5 SCB:s och övriga SAM:s register samverkar	48
Kapitel 4. Olika synsätt – olika språkbruk	50
4.1 Olika synsätt inom statistikvetenskapen	50
4.1.1 Urvalsundersökningar – registerbaserade undersökningar	51
4.1.2 Även granskning, bortfall, mätfel och presentation skiljer sig	52
4.2 Olika synsätt inom en statistikbyrå	53
4.2.1 Mitt register – registersystemet	53
4.2.2 Administrativa data skall bearbetas för statistiska syften	53

4.3	Termer – olika språkbruk	54
4.3.1	<i>Begrepp, termer och tidens tand</i>	54
4.3.2	<i>Termen statistisk undersökning</i>	54
4.4	Registertermer	55
4.4.1	<i>Register, datamatrix och tabell som statistiska termer</i>	55
4.4.2	<i>Termerna databas och databastabell</i>	56
4.4.3	<i>Termer för olika slag av register</i>	58
4.4.4	<i>Register och tidsdimensionen</i>	59
4.4.5	<i>Termen objekt</i>	61
4.5	Termer för olika slag av variabler	61
4.5.1	<i>Allmänt om variabler inom statistikvetenskapen</i>	62
4.5.2	<i>Allmänt om variabler inom systemvetenskapen</i>	62
4.5.3	<i>Härledda variabler</i>	63
4.5.4	<i>Variabler i registersystemet</i>	64
Kapitel 5.	Hur skapas ett statistiskt register?	68
5.1	Hur strukturera registerbaserade undersökningar?	68
5.1.1	<i>Metodarbete på olika nivåer</i>	69
5.1.2	<i>Arbetsgången då ett statistiskt register skapas</i>	70
5.2	Problemformuleringsfasen	71
5.3	Inventeringsfasen – att inventera olika källor	73
5.4	Att avgränsa registrets objektmängd	73
5.4.1	<i>Att definiera en population</i>	74
5.4.2	<i>Att definiera en population – primära statistiska register</i>	75
5.4.3	<i>Att definiera en population – integrationsregister</i>	76
5.4.4	<i>Basregistren skall användas då populationer avgränsas</i>	77
5.4.5	<i>Kraven på ett basregister</i>	78
5.4.6	<i>Alla skall stötta basregistren!</i>	80
5.5	Att avgränsa och härleda objekt	82
5.5.1	<i>Administrativa objekttyper</i>	83
5.5.2	<i>Objekttyper skapade i samarbete med administrativa myndigheter</i>	83
5.5.3	<i>Statistiska objekttyper skapade av SCB via datainsamling</i>	83
5.5.4	<i>Statistiska objekttyper skapade av SCB via härledning</i>	84
5.5.5	<i>Objekt och identiteter</i>	84
5.5.6	<i>Primära statistiska register</i>	85
5.5.7	<i>Kontroll av objektmängder och objektavgränsningar</i>	85
5.5.8	<i>Att skapa härledda objekt</i>	86
5.6	Att bestämma registrets variabelinnehåll	87
5.6.1	<i>Variabeldefinitioner</i>	87
5.6.2	<i>Arbetsgång när registrets variabler bildas</i>	89
5.7	Granska och åtgärda grunddata	89
5.7.1	<i>Granskningsarbetet inom Inkomst- och förmögenhetsregistret</i>	90
5.7.2	<i>Granskningsarbetet inom Kontrolluppgiftsregistret</i>	92
5.7.3	<i>Granskningsarbetet inom Företagsstatistiken - SRU</i>	94
5.7.4	<i>Granskning av SRU-data för FRIDA-registret</i>	95
5.7.5	<i>Sammanfattning – granskningsprocessen i registersystemet</i>	97
5.7.6	<i>Vad lär vi oss mer av dessa exempel?</i>	99
5.8	Att bilda härledda variabler med modeller	101
5.8.1	<i>Exakt beräkning av värden för en härledd variabel med en regel</i>	102
5.8.2	<i>Skatta värden för en härledd variabel med en regel</i>	102
5.8.3	<i>Skatta värden för en härledd variabel med en sambandsmodell</i>	104
5.8.4	<i>Härledda variabler och imputerade variabelvärden</i>	107
5.8.5	<i>Att bilda variabler genom kodning</i>	107
5.9	Att skapa longitudinella register	108

Kapitel 6. Beräkningsmetoder	110
6.1 Skattningsmetoder i survey- och registerstatistik	111
6.1.1 Skattningsmetoder som använder vikter	111
6.1.2 Skattningsmetoder för registerstatistik	111
6.1.3 Från datamatrix till statistiska tabeller	112
6.2 Skattningsmetoder vid flervärda variabler	114
6.2.1 Yrke i Aktivitets- och Yrkesregistret	115
6.2.2 Näringsgren i Företagsregistret	118
6.2.3 Transformation av vikter	120
6.2.4 Importera flervärda variabler – många flervärda variabler samtidigt	121
6.2.5 Konsistens mellan olika variabler	124
6.2.6 Urvalsundersökningar med flervärda registervariabler	125
6.2.7 Konsistens mellan skattingar från olika register	126
6.2.8 Flervärda variabler – hur gör vi i praktiken?	128
6.2.9 Objekttyperna i Företagsregistret	130
6.2.10 Skattningsmetoder för årgångsregister	132
6.3 Skattningsmetoder för att korrigera för övertäckning	133
6.3.1 Täckningsproblem vid undersökningar med egen datainsamling	133
6.3.2 Täckningsproblem vid registerbaserade undersökningar	134
6.3.3 Övertäckning i Befolkningsregistret	135
6.3.4 Kalibrering av vikter	136
6.4 Partiellt bortfall – värden saknas	137
6.4.1 Det partiella bortfallet varierar över tiden	137
6.4.2 Två traditioner för att åtgärda bortfall	138
6.4.3 Bortfallskorrektion med vikter	139
6.4.4 Bortfallskorrektion med imputering	141
6.4.5 Bortfallskorrektion i ett system av register	144
6.5 Skattningsmetoder för att korrigera för tidsseriebrott	147
6.5.1 Olika sätt att redovisa tidsseriebrott	147
6.5.2 Skatta effekter av tidsseriebrott – länkning på makronivå	148
6.5.3 Länkning av tidsserier på mikronivå med vikter	150
6.5.4 Länkning av tidsserier på mikronivå med kombinationsobjekt	152
Kapitel 7. Registerstatistikens kvalitet	156
7.1 Vad är specifikt för registerstatistik?	158
7.1.1 Undersökningar med egen datainsamling	159
7.1.2 Statistiska register	159
7.1.3 Användarnas och producenternas syn på kvalitet	160
7.2 Kvalitetssäkring – detaljkunskaper om registrets egenskaper	161
7.2.1 Problemformuleringsfasen – hur har den påverkat registret?	162
7.2.2 Inventeringsfasen – hur har den påverkat registret?	162
7.2.3 Planeras förändringar?	162
7.2.4 Leverantörskontakter och granskning	162
7.2.5 Bearbetningsfasen – hur har objektmängden skapats?	163
7.2.6 Bearbetningsfasen – hur har objekten avgränsats?	164
7.2.7 Bearbetningsfasen – hur har variablerna skapats	165
7.2.8 Dokumentation som del av kvalitetssäkringen	166
7.2.9 Nivåer för kunskap om tillförlitligheten	167
7.2.10 Vad säger dessa indikatorer om registrets kvalitet?	167
7.3 Helhetsbedömning av kvaliteten	168
7.3.1 En helhetsbedömning av ett registers kvalitet	168
7.3.2 En helhetsbedömning av en registerbaserad undersöknings kvalitet	169

Kapitel 8. Dokumentation och metadata	170
8.1 Primära statistiska register – behovet av metadata	170
8.1.1 Dokumentation av administrativa källor	171
8.1.2 Dokumentation av källor inom SCB:s registersystem	171
8.1.3 Dokumentation av det nya registret	172
8.2 Förändringar över tiden – behovet av metadata	172
8.3 Integrationsregister – behovet av metadata	173
8.4 Klassifikations- och definitionsdatabas	174
8.4.1 Klassifikationsdatabas	174
8.4.2 Definitionsdatabas och härledda variabler	174
8.5 Sammanfattning – behovet av metadata för register	175
Kapitel 9. Produktionsprocessen och IT-verktyg	176
9.1 Projekt	176
9.2 Analys	177
9.3 Design	178
9.3.1 Teknikval	178
9.3.2 Databasutformning	179
9.3.3 Applikationsarkitektur	182
9.4 Fysisk realisering	182
9.4.1 Fysisk optimering för registersystemet	182
9.4.2 Indexering	183
9.5 Övriga faser i systemutvecklingen	184
9.6 Användning av register	184
9.6.1 Stora datamängder	184
9.6.2 Matchning	184
9.6.3 Rätt verktyg för uppgiften	187
Kapitel 10. Integritetsskydd	188
10.1 Förekomst av klartext	188
10.2 Förekomst av identiteter	189
10.3 Røjanderisk i tabeller	190
10.3.1 50/90-regeln för summor och medelvärden	190
10.3.2 Basen skall vara minst tre observationer	191
10.3.3 Røjanderisken får inte vara för hög	191
10.4 Røjanderisk i mikrodata	193
Kapitel 11. Samordning och sam anvä ndbarhet	194
11.1 Organisatorisk samordning	194
11.1.1 En gemensam enhet med övergripande ansvar för systemet	195
11.1.2 Samgranskning	195
11.1.3 Integrationsregister och uppdrag	196
11.2 Teknisk samordning	196
11.3 Innehållsmässig samordning	197
11.3.1 Samordning och standardisering av populationer	197
11.3.2 Samordning och standardisering av variabler	199
11.3.3 Samordning och standardisering av grupperingar	199
11.4 Samordning av statistiska metoder	201
11.5 Sam anvä ndbarhet	201
Kapitel 12. Visionen om ett effektivt system	206
Bilaga 1. Riktlinjer för statistiska register på SCB	209
Bilaga 2. Referenser	210
Bilaga 3. Registerstatistiska termer	213

Kapitel 1. Vad är registerstatistik?

I detta och kommande kapitel introducerar vi en rad begrepp och principer som kommer att användas i denna bok när vi diskuterar *registerstatistik*, dvs. statistik som bygger på användningen av administrativa data. Dessa begrepp och principer är grunderna till en teori för denna typ av statistik.

1.1 Behövs en registerstatistisk teori?

Inom de nationella statistiska byråerna framställs tre olika slag av statistik – statistik baserad på urvalsundersökningar, statistik baserad på totalundersökningar och statistik baserad på administrativa register. Det vanliga är att man endast brukar skilja på urvals- och totalundersökningar, där statistikbyrån ansvarar för datainsamlingen. Dessa två undersökningstyper domineras av arbetet med att samla in data.

Denna bok handlar om det tredje slaget av statistik, dvs. statistik baserad på administrativa register, där man i stället för att samla in egna data bearbetar administrativa register från olika källor.

Urvalsundersökningar baseras på metoder som bygger på en etablerad teori – *surveyteorin*. Denna teori har utvecklats både inom universitetsvärlden och inom de statistiska byråerna och består av termer och principer som är allmänt kända. Detta innebär att det finns vetenskaplig litteratur och tidskrifter, där man utvecklar och sprider metodiken för urvals- och estimationsmetoder. Genom att termer och principer är allmänt kända kan personer som arbetar med urvalsundersökningar lätt kommunicera och utbyta erfarenheter.

Totalundersökningar med egen datainsamling bygger på en lång tradition av folkräkningar och insamling av data från kommuner, skolor och olika typer av företag. Mätproblem, utformning av frågeformulär och problem med bortfall är metodfrågor som är gemensamma med urvalsundersökningar. Total- och urvalsundersökningar är metodmässigt närbesläktade – totalundersökningar betraktas ofta som specialfall där urvalet är hela populationen.

Statistik baserad på administrativa register kommer vi i fortsättningen att kalla för *registerstatistik*. Vi kommer i denna bok att skilja på registerstatistik och surveystatistik, där surveystatistik baseras på urvals- eller totalundersökningar. Det engelska begreppet *survey* har i statistiska sammanhang betydelsen *icke-experimentell undersökning*. Men termen *survey* förknippas ofta med urvalsundersökningar: "A survey gathers information from a sample."

Trots att registerstatistiken är den äldsta och vanligaste formen av statistik har den ingen väletablerad teori. Det saknas välkända termer och principer vilket försvårar utvecklingen av både registerstatistik och registerstatistisk metodik. En viktig orsak till denna brist är att ämnesområdet registerstatistik inte behandlas inom universitetsämnet statistik. Det som uppfattas som statistisk teori inom statistikvetenskapen består av *sannolikhetsteori* och *inferensteori*. Surveyteorin hör hemma inom detta teoretiska tänkande, men däremot inte den totalräknade registerstatistiken.

Detta citat kommer
från American
Statistical Association
(1996)

Tyvärr har statistikvetenskapen hittills inte innehållit någon teori om statistiska system. Inom statistiska byråer, större företag och organisationer görs inte enstaka undersökningar så ofta, däremot bygger man upp statistiska informationssystem som ständigt genererar nya data. En statistisk teori behövs som beskriver generella principer och utvecklar en begreppsapparat för sådana statistiska system och registerstatistiken skulle höra hemma inom denna teori.

Danmarks Statistik publicerade 1994 "*Personstatistik i Danmark - Et registerbaserat statistiksystem*". Den danska boken ger en systematisk genomgång av registerstatistiskt arbete och hur ett genomtänkt registersystem kan utformas.

I den här boken om registerstatistik bygger vi vidare på detta danska arbete. Vi introducerar i kommande kapitel ett antal registerstatistiska begrepp och principer. I bilaga 3 finns alla begrepp och termer samlade. Syftet är att termerna skall kunna användas allmänt av alla de som vill utveckla registerstatistiken. Vi formulerar fyra principer för hur administrativa register skall utnyttjas:

- en statistikbyrå skall ha tillgång till många administrativa register
- dessa administrativa register skall transformeras till statistiska register
- alla statistiska register skall ingå i ett samordnat registersystem
- med detta system kan befintliga data användas effektivt och flexibelt för att kunna möta nuvarande och framtida behov.

I boken kommer vi att använda dessa principer och gradvis föra in de registerstatistiska termer som behövs för diskussionen.

1.2 Vad är en statistisk undersökning?

Utgångspunkten för en undersökning är ett antal frågor med anknytning till ett visst problemområde och för att besvara dessa frågor görs en undersökning. Undersökningsprocessen kan beskrivas mer eller mindre detaljerat. Enligt en översiktlig beskrivning består arbetet av följande tre faser:

1. problemformulering och planläggning av undersökningen
2. införskaffande av data samt bearbetning av data
3. skattningar bildas, analys och presentation av undersökningens resultat.

Görs undersökningen med statistiska metoder talar vi om en statistisk undersökning. Statistisk metodik är en sammanhängande strategi för att samla in, bearbeta och analysera data. Även i icke-statistiska undersökningar samlar man in, bearbetar och analyserar data, men i statistiska undersökningar samlas data in för att kunna bearbetas och analyseras med *kvantitativa* metoder.

Inom SCB arbetar man som regel med löpande undersökningar – samma undersökning upprepas varje år, kvartal eller månad. Vid sådana undersökningar arbetar man i huvudsak med fas 2 och 3. Men även dessa löpande undersökningar har haft en problemformuleringsfas, som kan ligga långt tillbaka i tiden.

En enstaka undersökning kan vara ett uppdrag där man på SCB får göra hela undersökningen, och då arbetar man med alla tre faserna. Men i många uppdrag gör kunden fas 1 och 3, och på SCB arbetar man endast med fas 2.

En mer utförlig beskrivning finns i SCB (2001a)

Undersökningens fas 2, införskaffande av data, kan göras på olika sätt:

Makrodata

- a. Redan befintliga makrodata används vid en *skrivbordsundersökning*. Makrodata avser sammanfattande uppgifter för grupper av objekt och brukar presenteras som innehållet i statistiska tabeller.
Exempel: Detta är den vanligaste typen av undersökning som SCB:s användare gör när de utnyttjar tabellerna i SCB:s publikationer eller makrodatabaser. Även Nationalräkenskaperna och Miljöräkenskaperna är exempel på undersökningar som bygger på befintliga makrodata.

Registerbaserad undersökning

Mikrodata

- b. Redan befintliga mikrodata används vid en *registerbaserad undersökning*. Mikrodata avser uppgifter för enskilda objekt. Befintliga administrativa eller statistiska register med uppgifter som t.ex. avser enskilda personer eller företag används för den registerbaserade undersökningens syften.
Exempel: Ungdomars övergång från studier till arbetsmarknaden. Detta exempel beskrivs i avsnitt 1.3.1 nedan.
- c. Med egen datainsamling vid en *totalundersökning*.
Exempel: Folk- och bostadsräkningen 1990, där alla hushåll och fastighetsägare fick fylla i blanketter som sedan bearbetades av SCB. Eftersom totalundersökningar resulterar i att register skapas, kommer även mikrodata från totalundersökningar att ingå i SCB:s registersystem och därigenom vara underlag för registerbaserade undersökningar.
- d. Med egen datainsamling vid en *urvalsundersökning*.
Exempel: Arbetskraftsundersökningen med nytt urval, ny datainsamling och rapportering månadsvis.
- e. Dessutom kan man göra egen datainsamling vid ett *experiment*.
Exempel: Denna insamlingsmetod används inom SCB för att testa olika mätmetoder vid intervjuer eller enkäter.

Eftersom de fyra typerna av icke-experimentella undersökningar (a – d ovan) skiljer sig åt metodmässigt är det lämpligt att begreppsmässigt skilja dem åt. Registerbaserade undersökningar, total- och urvalsundersökningar är de undersökningstyper som är viktigast inom en statistisk byrå.

1.3 Vad är ett register och vad är registerstatistik?

Ett *register* avses vara en *fullständig förteckning* över objekten i en viss objekt-mängd eller population, men på grund av kvalitetsbrister kan uppgifter om vissa objekt saknas. Uppgifter om objektens identitet skall finnas så att registret kan uppdateras och utökas med nya variabelvärden för respektive objekt.

Termen register används i denna bok för register inom SCB:s registersystem. Sådana register baseras antingen på en totalundersökning genomförd av SCB eller på administrativa register från myndigheter och organisationer utanför SCB.

Datainsamlingen i en urvalsundersökning ger inte upphov till ett register i denna mening, då mikrodata avseende urvalet endast innehåller en liten del av den undersökta populationens objektmängd.

Termerna *observationsregister* och *slutligt observationsregister* används både för urvals-, total- och registerbaserade undersökningar. De används bl.a. inom SCB:s dokumentationssystem SCBDOK. I bilden nedan jämförs de tre slag av undersökningar som dominerar inom statistikbyråer.

Bild 1.1 Jämförelse mellan de tre typerna av undersökningar

Urvalsundersökning	Totalundersökning	Registerbaserad undersökning
Ingår inte i registersystemet	Ingår i registersystemet	
Egen datainsamling		Använder befintliga register

Med termen *registerstatistik* avses statistik som bygger på registerbaserade undersökningar. När vi diskuterar registersystem, som i kapitel 2 och 3, skiljer vi inte mellan totalundersökningar och registerbaserade undersökningar. Men när vi diskuterar metodfrågor, så avser termen registerstatistik enbart registerbaserade undersökningar. Termen *surveystatistik* avser statistik baserad både på total- och urvalsundersökningar.

1.3.1 Registerstatistik – ett exempel

Unga personers möte med arbetsmarknaden efter avslutad utbildning är numera ett viktigt område för olika undersökningar. Sådana studier bör göras som *longitudinella undersökningar*, där grupper av personer följs under en följd av år. Om man gör dessa studier som urvalsundersökningar görs ett urval varje år bland personer som avslutar viss utbildning och varje urval intervjuas eller besvarar enkäter en gång per år under t.ex. sju år.

Denna undersökningsmetod har nackdelar. Dels blir uppgiftslämnarbördan tung – de utvalda måste besvara en stor uppsättning frågor varje år, dels kommer bortfallet att gradvis öka under perioden. Har man dessutom inte tillgång till ett bra register över avslutade utbildningar måste man först göra ett stort urval av personer inom vissa åldersgrupper för att hitta de som avslutat gymnasie- respektive högskolestudier.

En alternativ undersökningsmetod är att basera studien enbart på befintliga register. SCB publicerar sådan registerstatistik som beskriver övergången till arbetsmarknaden. Denna statistik är baserad på administrativa källor men för att producera denna statistik är det inte tillräckligt med tillgång till de administrativa källorna. *Det är också nödvändigt ha tillgång till ett fungerande system av statistiska register.* Exemplet används här för att illustrera viktiga egenskaper hos registerstatistik och registersystem.

I tabellen och diagrammen nedan kan vi jämföra personer med lägre och högre utbildning med avseende på deras övergång till arbetsmarknaden. Sex årgångar (*kohorter*) som består av alla personer som slutat gymnasiet 1987-1992 följs under åren 1988-1993 och deras övergång till förvärvsarbete kan jämföras med motsvarande sex kohorter av dem som tagit grundexamen på högskola. Denna tidsperiod är särskilt intressant då arbetsmarknaden förändrades dramatiskt. Alla de personer som tillhör dessa tolv kohorter undersöks genom att man bildar longitudinella register som sedan analyseras.

Kohort: en grupp av objekt som har sitt ursprung från en viss period (år)

Varje kolumn i tabellen nedan beskriver utvecklingen för en sådan kohort.

Tabell 1A. Andel förvärvsarbetande 1988-1993 efter avslutad utbildning 1987-1992

	<i>Avgångna från gymnasiet</i>						<i>Efter högskoleexamen</i>					
	År då utbildningen avslutades:						År då utbildningen avslutades:					
	1987	1988	1989	1990	1991	1992	1987	1988	1989	1990	1991	1992
1988	74,0						85,1					
1989	72,6	73,5					86,2	85,1				
1990	72,1	70,6	70,5				87,4	86,3	85,8			
1991	70,0	67,6	64,5	60,9			86,8	87,0	86,6	84,2		
1992	66,5	63,5	58,5	50,4	43,9		86,8	85,7	86,0	84,3	81,1	
1993	63,5	59,5	54,0	46,0	38,0	33,5	83,8	83,4	82,7	81,0	78,6	73,2
Antal personer i resp. kohort	88 955	88 772	89 407	86 187	88 923	91 987	22 738	22 521	23 335	21 964	23 539	24 850

Anm: Förvärvsarbetande avser endast förvärvsarbetande, de som kombinerar studier och förvärvsarbete ingår inte

Cirklarna i bild 1.2 och 1.3 nedan representerar andelen förvärvsarbetande *ett* år efter avslutad utbildning. Kurvorna visar utvecklingen av andelen förvärvsarbetande inom respektive kohort.

Andel förvärvsarbetande efter avslutad utbildning 1987-1992

Bild 1.2 Avgångna från gymnasiet

Procent

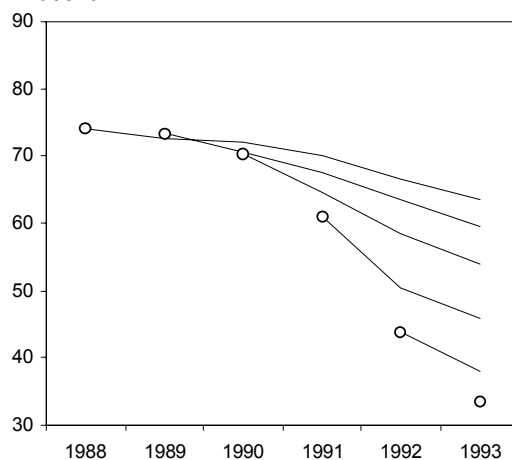
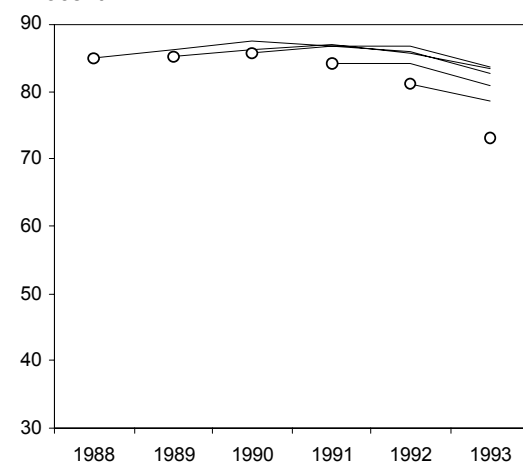


Bild 1.3 Efter högskoleexamen

Procent



I början av 1990-talet inträffade den allvarligaste krisen på arbetsmarknaden sedan 1930-talet. Bild 1.2 och 1.3 visar konjunkturedgången i början av 1990-talet som förändrade unga personers möjligheter att komma ut på arbetsmarknaden. Andelen som förvärvsarbetade ett år efter avslutade gymnasiestudier har förändrats dramatiskt från 74 % till 34 % under dessa år. Enligt andra uppgifter från undersökningen ökade andelen som varken studerade eller förvärvsarbetade från 14 % till 42 % under dessa år bland dem som avslutat gymnasiestudierna. För de som har examen från högskola har utvecklingen inte varit lika dystert – det verkar som att högskolestudier lönar sig.

Detta exempel är baserat på longitudinella data. Varje år är varje person klassificerad som förvärvsarbetande, högskolestuderande eller varken förvärvsarbetande eller studerande. För de förvärvsarbetande är den årliga arbetsinkomsten

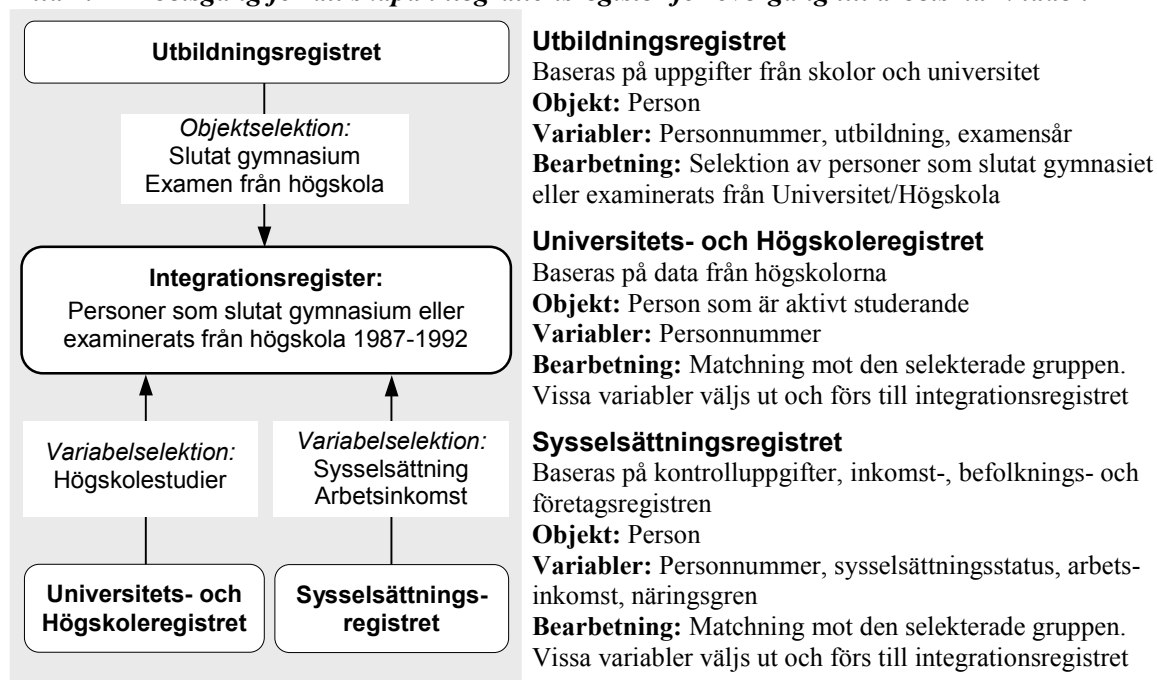
registrerad och också inom vilken näringsgren varje person arbetar. Personerna är dessutom klassificerade efter kurs/studieprogram, kön och region vilket medger detaljerade redovisningar. Årskostnaden för denna produkt som finns redovisad i Statistiska meddelanden U 81 SM 9501 var 0,4 miljoner kronor varav hälften utgjordes av tryckkostnader.

1.3.2 Hur gjordes denna registerbaserade undersökning?

Diagrammen ovan visar unga personers övergång från studier till arbetsmarknaden. För att skapa det longitudinella register som diagrammen bygger på har man gjort på följande sätt:

- Genom att kombinera information i tre SCB-register för åren 1987-1993 (dvs. totalt 21 olika register) skapas ett nytt *integrationsregister*, som markeras med kraftigare ram i bild 1.4 nedan.
- Objekten i det nya registret skapas genom *selektion av vissa objekt* i Utbildningsregistret.
- Genom att *matcha* objekten i det nya registret med motsvarande objekt i Utbildnings-, Universitets- och Högskoleregistret samt Sysselsättningsregistret hämtas variabelvärden till det nya registret. Arbetsgången illustreras i diagrammet nedan.

Bild 1.4 Arbetsgång för att skapa integrationsregister för övergång till arbetsmarknaden



Vi har här fört in en rad viktiga termer. Genom att sammanföra eller integrera information från flera register skapas ett *integrationsregister*. Vi skiljer mellan *objektselektion* där vissa objekt väljs ut och *variabelselektion* där vissa variabler väljs ut. *Matchning* innebär att länkarna i två eller flera register jämförs, resultatet blir träff eller icke-träff.

Integrationsregister
Objektselektion
Variabelselektion
Matchning

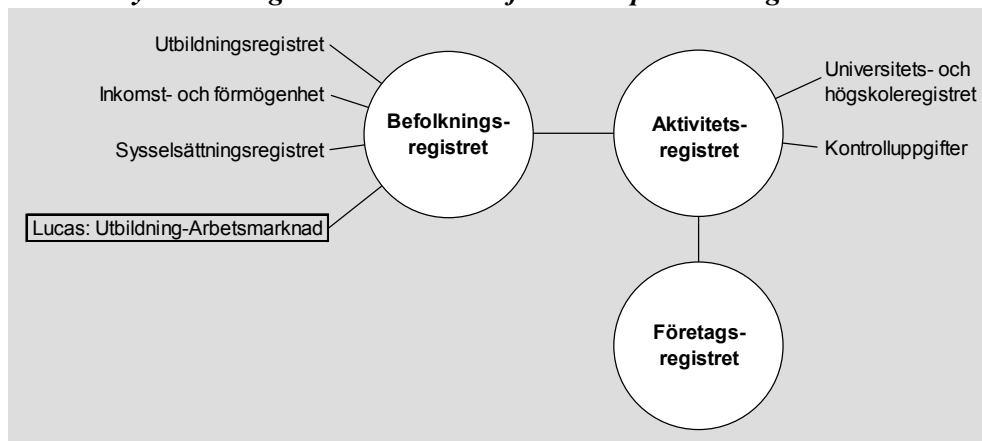
Lucas
Longitudinellt register
för utbildnings- och
arbetsmarknads-
statistik

För att skapa det nya integrationsregistret, *Lucas*, behövs ingen ny datainsamling, utan registret bygger på ökad användning av redan befintliga register. Även Sysselsättningsregistret bygger på redan befintliga register i SCB:s registersystem. Däremot bygger Utbildnings-, Universitets- och Högskoleregistret på att SCB får administrativa data från myndigheter utanför SCB.

För att skapa och samordna dessa register fordras dessutom omfattande bearbetningar av registerdata i flera register. Detta innebär att tillgången på administrativa källor inte räcker för att kunna framställa denna statistik. Utan SCB:s samordnade system av statistiska register skulle det vara omöjligt att producera denna statistik.

Åtta register används för att skapa integrationsregistret Lucas i vårt exempel med övergång från studier till arbetsmarknaden. Detta illustreras i bilden nedan.

Bild 1.5 System av register som används för att skapa Lucasregistret



Personnummer är länken mellan Befolknings-, Utbildnings-, Sysselsättnings-, Inkomst-, Lucas-, Aktivitetsregistret och Universitets- och högskoleregistren. *Arbetsställenummer* är länken mellan Aktivitets- och Företagsregistret. De två kopplingsvariablerna *Personnummer* och *arbetsställenummer* är tillsammans länken mellan Aktivitetsregistret och Kontrolluppgiftsregistret. Med hjälp av länkarna kan Lucasregistret påföras variabler från alla övriga register i registersystemet.

Länk
Kopplingsvariabel

Vi har här infört termerna länk och kopplingsvariabel. En *länk* mellan två register består av en eller flera kopplingsvariabler som finns i båda registren. En *kopplingsvariabel* består av information som identifierar enskilda objekt eller grupper av objekt.

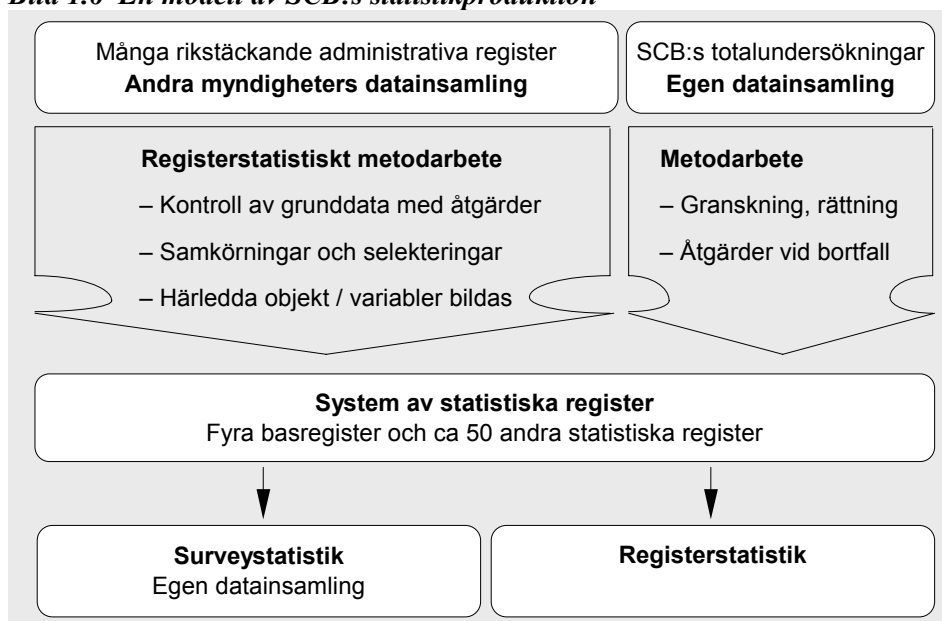
De register som används i detta exempel är delar av det registersystem som kommer att presenteras mer ingående i kapitel 2 och 3. Tre av systemets basregister används, Befolknings-, Aktivitets- och Företagsregistret. Dessutom måste det finnas fungerande länkar mellan de inblandade registren för att det skall vara möjligt att skapa Lucasregistret. Den viktiga kopplingsvariabeln personnummer finns i alla registren i bild 1.5 utom Företagsregistret. Länken mellan Aktivitets- och Företagsregistret är företagets organisationsnummer (PeOrgNr) eller arbetsställets identitet (CFAR-nr).

1.4 En modell av SCB:s statistikproduktion

Vi skiljer mellan de *administrativa register*, som andra myndigheter lämnar till SCB, och de *statistiska register* som SCB har skapat av dessa, efter i många fall omfattande *registerstatistiska bearbetningar*.

Systemet av statistiska register är grunden för all statistik som SCB producerar. I alla urvals- och totalundersökningar används register som urvalsram, registervariabler används i stället för att fråga intervjupersonerna samt som hjälpvariabler vid estimationen då hjälpvariabler används för att minska slumpmässiga och systematiska fel. Förutom denna statistik baserad på egen datainsamling producerar SCB också *registerstatistik* som direkt baseras på bearbetningar av uppgifter i SCB:s register.

Bild 1.6 En modell av SCB:s statistikproduktion



Administrativa register

Statistiska register

Registerstatistiska bearbetningar

Ett register inom SCB kan vara baserat på en totalundersökning genomförd av SCB eller baserad på administrativa register som skapats utanför SCB

Registerstatistiken är ca hälften av det ekonomiska värdet av SCB:s produktion

Ett statistiskt registersystem består av en rad olika komponenter, förutom ett samordnat system av register krävs också etablerade metoder samt ett fungerande dokumentationssystem:

- *basregister* består av objektmängder av grundläggande betydelse, används för att bilda *standardiserade populationer*
- övriga *statistiska register* innehåller statistiskt intressanta variabler
- *länkar* mellan objekt i olika basregister samt *länkar* mellan basregister och övriga statistiska register
- *registerstatistiska metoder* och rutiner för kvalitetssäkring
- *metadata*, registerdokumentation av definitioner, kvalitet, jämförbarhet
- *IT-verktyg* för registerbearbetningar
- *rutiner för integritetsskydd*
- *standardiserade variabler*, variabler av grundläggande betydelse.

Basregister, statistiska register, länkar diskuteras i kapitel 2-3

Registerstatistiskt metodarbete och kvalitet diskuteras i kapitel 5-7

Metadata diskuteras i kapitel 8

IT-verktyg diskuteras i kapitel 9

Integritetsskydd diskuteras i kapitel 10

Standardiserade variabler och populationer diskuteras i kapitel 11

1.5 Hur kan man använda SCB:s registersystem?

Ett statistiskt registersystem kan användas på tre sätt:

- registersystemet används för att framställa registerstatistik
- registersystemet används för urvals- och totalundersökningar
- de statistiska registren och urvalsundersökningarna ses som *ett* system där registerstatistiken och urvalsundersökningarna kompletterar varandra.

1.5.1 Systemet används för registerstatistik

Registersystemet kan användas för ren registerstatistik. Med ett befintligt statistiskt register kan tabeller med *skattningar* produceras och publiceras. Termen skattning används allmänt för urvalsundersökningar, men bör även användas inom registerstatistiken. Även här måste man skilja på det verkliga värdet och den skattning som registret ger.

Vårt exempel med unga personers övergång från studier till arbetsmarknaden visar hur sådan registerstatistik kan framställas. I avsnitt 3.4 ges exempel på en rad andra registerstatistiska tillämpningar inom individstatistiken.

1.5.2 Registersystemet används för urvalsundersökningar

I samband med urvalsundersökningar kan man använda registersystemet på följande sätt:

- I samband med urvalet används register som urvalsram och registervariabler används för att stratifiera populationen.
- Mätningarna underlättas av att man slipper fråga efter de uppgifter som redan finns i befintliga register.
- Under estimationsfasen kan man använda registervariabler som hjälp-information för att öka precisionen och kompensera för bortfallet.

Urval

Basregistren används som *urvalsram* och registervariabler kan användas för att göra ett stratifierat urval.

Exempel: För Arbetskraftsundersökningarna används Befolkningsregistret för att göra urval bland befolkningen i åldrarna 16-64 år, där urvalet stratifieras efter region, kön och medborgarskap enligt Befolkningsregistret samt förvärvs-arbetande/ej förvärvsarbetande enligt Sysselsättningsregistret.

Datainsamling – mätningar

Att inte fråga efter uppgifter som redan finns i SCB:s register innebär dels att uppgiftslämnarbördan minskar, dels att intervju tid sparas och dels att intervjuarna slipper ställa känsliga och svåra frågor om t.ex. inkomst och utbildning.

Exempel: Inkomstfördelningsundersökningen använder i huvudsak inkomstvariabler från register. De utvalda personerna behöver bara besvara frågor om hushållets sammansättning, yrke, om de arbetar hel- eller deltid och om vissa slag av inkomster som inte finns särredovisade i registren.

Ett annat sätt att kombinera en urvalsundersökning med registerdata är att undersöka nuläget med enkät- eller intervjufrågor och att använda registerdata

för att undersöka historiken. *Exempel:* I en hälsoundersökning ställs frågor om det aktuella hälsoläget och aktuella levnadsförhållanden. Registerdata kan användas för att belysa utbildningsbakgrund, förvärvsarbetshistorik etc.

Estimationsfasen

Registren har också en viktig roll under estimationsfasen. Registersystemet innehåller många variabler vilka kan användas som hjälpinformation för att göra bättre skattningar. Dessa hjälpvariabler minskar felmarginalerna för skattningarna vid given urvalsstorlek eller minskar urvalsstorleken och kostnaderna vid ett givet precisionskrav.

Exempel: I estimationsfasen i Arbetskraftsundersökningarna, utnyttjas registervariablerna kön och ålder enligt Befolkningsregistret, förvärvsarbete efter näringsgren enligt Sysselsättningsregistret samt arbetssökandekategori enligt AMS:s sökanderegister. En rimlig bedömning är att dessa hjälpvariabler minskar felmarginalerna med cirka 20 % vid given urvalsstorlek vilket motsvarar en minskad urvalsstorlek med cirka 36 % för samma precisionskrav.

Källa:
Hörngren (1992)

Bortfallet i urvalsundersökningar är ett stort problem och *bortfallets storlek* ökar. Registerdata kan användas för att minska bortfallets storlek då man slipper fråga efter vissa känsliga variabler som inkomst, utbildning och ålder. Dessutom blir frågeformuläret kortare om man använder registervariabler i stället för att fråga. Även detta kan minska bortfallets storlek.

Registervariabler kan också användas för att minska effekten av *bortfallsfel*. Det är då en fördel att ha tillgång till många variabler så att man med ledning av bortfallets struktur och samband med undersökningsvariablerna kan välja den korrigeringsmetod som passar bäst för den aktuella undersökningen.

Källa:
Lundström och Särndal
(2001)

1.5.3 Register och urval kompletterar varandra

En registerbaserad undersökning och en urvalsundersökning kan komplettera varandra innehållsmässigt. Registerstatistiken ger grundläggande uppgifter om skillnader mellan olika kategorier och om förändringar över tiden. Det är sedan möjligt att göra en urvalsundersökning som ger en utförligare beskrivning av orsakerna till dessa mönster. Med hjälp av registren kan urvalet avgränsas till särskilt intressanta delgrupper.

Exempel: Med Mervärdesskatteregistret skulle man kunna göra ett urval av företag som minskat sin omsättning under det senaste året. De utvalda företagen intervjuas om orsakerna till nedgången och hur man bedömer framtiden.

Exempel: Med Aktivitetsregistret (detta register beskrivs i avsnitt 2.2.2) skulle man kunna göra ett urval av personer som bytt arbetsgivare under de två senaste åren och intervjua dem om orsakerna till bytet.

1.6 Varför skall statistik baseras på administrativa data?

Det finns både fördomar och berättigad kritik mot registerstatistik. Enligt fördomarna är statistik baserad på administrativa data billig men dålig jämförd med "riktig" surveystatistik. Berättigad kritik kan avse relevansfel, jämförbarhetsproblem och att kvaliteten inte är under kontroll. Tyvärr kan liknande kritik

riktas mot urvalsundersökningar, bortsett från att urvalsfelen kontrolleras, kan mätfel och bortfallsfel vara stora och okända.

Vårt svar på synpunkter av detta slag är att samhällets statistik skall bestå av *både* registerstatistik och surveystatistik. Det är inte fråga om att den ena metoden är bättre än den andra, utan i vissa situationer är registerstatistik effektivast, medan i andra fall är urvalsundersökningar lämpligast.

1.6.1 För- och nackdelar med surveystatistik och registerstatistik

Det finns en vanlig men ofta diffus uppfattning om att statistik baserad på administrativa data har låg kvalitet: ”Vi har ju ingen kontroll över datainsamlingen.” Är denna uppfattning befogad för de administrativa källor som SCB använder? En mycket stor del av SCB:s registersystem baseras på data från folkbokföring och skatteadministration. Skulle denna statistik ha högre kvalitet om SCB gjorde en egen datainsamling parallellt med Skatteverkets insamling av administrativa data? Så är knappast fallet – SCB:s egna försök att samla in dessa uppgifter skulle vara dyra, öka uppgiftslämnarbördan och ge data med större mätfel.

Från och med den
1/1 2004 bytte
Riksskatteverket
(RSV) namn till
Skatteverket (SKV)

Bild 1.7 För- och nackdelar med surveystatistik och registerstatistik

	Fördelar	Nackdelar
Undersökningar baserade på egen datainsamling	Du kan ställa de frågor du vill Kan vara aktuell	Vissa respondenter förstår inte frågan ... har glömt hur det var ... svarar inte (bortfall) ... svarar slarvigt Uppgiftslämnarbördan kan bli stor Dyrt Låg kvalitet på skattningar för små redovisningsgrupper (gäller urval)
Registerbaserade undersökningar	Ingen ytterligare uppgiftslämnarbörda för statistiken Låga kostnader God täckning av populationens objekt Data för varje år Respondenterna svarar omsorgsfullt på administrativt viktiga frågor Goda möjligheter till redovisning för små grupper, regional statistik och longitudinella studier	Du kan inte ställa frågor själv Du är beroende av de administrativa systemens populations- och objektavgränsningar samt variabeldefinitioner I vissa fall lång tid från händelse till att data blir tillgängliga Förändringar i de administrativa systemen försvårar jämförelser Variabler som är mindre viktiga för det administrativa arbetet kan ha lägre kvalitet

I denna tabell har vi sammanfattat fördelar och nackdelar som finns med de två undersökningsmetoderna. Hur allvarliga nackdelarna är kan vara olika vid olika undersökningssituationer.

Vårt exempel med ungdomars övergång från studier till förvärvsarbete kan illustrera fördelar och nackdelar med registerstatistik. Vi har redan nämnt att registeransatsen är lämplig vid longitudinella studier. En annan fördel är att man kan redovisa resultat på många undergrupper, olika kurser, studieprogram

och regioner. Men en nackdel med vissa typer av administrativa källor är att de administrativa systemen behöver lång tid från händelse till att data blir tillgängliga, detta försenar registerstatistiken. I vårt exempel publicerades rapporten som beskrev perioden 1987-1993 i juni 1995, dvs. efter ett och ett halvt år.

Det gäller inte enbart *om* man skall utnyttja administrativa data, utan i högsta grad *hur* man skall utnyttja dessa. Vårt svar på *hur* är att administrativa data som regel inte skall användas som de är, utan bearbetas för statistiska syften. Man skall sträva efter att bygga ett system av samverkande register – detta ger både kvalitetsmässiga och ekonomiska fördelar, och kvalitetssäkring skall finnas som en viktig komponent i systemet.

Det finns många utvecklingstendenser som hotar surveystatistikens kvalitet. Ökande användning av telemarketing och nummerpresentatörer gör det svårare att genomföra telefonintervjuer. Om uppgiftslämnarnas motivation minskar så ökar bortfallet och svars kvaliteten minskar. Det blir allt svårare att motivera dubbelt uppgiftslämnande – varför svara på en enkät om företagets omsättning då man också lämnar momsdeklaration till skattemyndigheten där uppgift om omsättning ingår? Allt detta talar för att registerstatistikens betydelse kommer att öka. Att dubbelt uppgiftslämnande både till SCB och till en annan myndighet uppfattas som orimligt framgår av följande tidningsklipp:

Mariestads Tidning den 26 juni 2000:

Vägrar skicka in statistik till SCB

Bo R på B-torp tycker att myndigheterna kan leta fram uppgifterna ur sina egna register

Bo R vägrar skicka in statistik till SCB i Örebro. Eftersom han skickar in uppgifter varannan vecka till Jordbruksverket tycker han att myndigheterna i stället borde samarbeta. ...

1.6.2 Kostnadsaspekten – är registerstatistik billigt?

”It is quite clear that the Member States find themselves in the paradoxical situation of having to face a number of budget cutbacks at the same time as providing users with an increasing volume of high-quality relevant information.” Med dessa ord inledde Y. Franchet ett seminarium (Eurostat, 1997) som behandlade användningen av administrativa källor för statistiska ändamål. Citatet illustrerar kraven på mer effektiva statistiska system. Dessa krav på ökad effektivitet kan mötas genom att kombinera två strategier:

- Genom att fler administrativa källor används kan kostnaden för datainsamling reduceras både för uppgiftslämnare och statistikproducent.
- Genom mer effektivt och flexibelt utnyttjande av redan befintliga data kan ny efterfrågan mötas utan dyrbar insamling av nya data. Detta uppnås genom att de administrativa källorna används för att skapa ett system av samordnade statistiska register.

Att bygga upp ett sådant registersystem kostar, men då systemet används blir marginalkostnaden låg. I exemplet med ungdomars övergång från studier till arbetsliv var produktens marginalkostnad 0,4 miljoner kronor.

Yves Franchet,
dåvarande general-
direktör för Eurostat

Uppgiftslämnarbördan är också en samhällsekonomisk kostnad som kan minskas genom att gå över från surveystatistik till registerstatistik.

Urvalsundersökningar är en allt dyrare metod, då kostnaderna för bortfallsuppföljning ökar. Kostnaderna för urvalsundersökning per invånare är höga för små länder – för samma precision krävs lika stort urval i ett litet land som ett stort land. Det blir därför särskilt viktigt för små länder att bygga upp ett statistiskt registersystem. I Danmark har man konsekvent följt denna strategi.

1.6.3 Register, integritet och allmänna opinionen

Folkräkningarna i Sverige har hittills varit traditionella totalundersökningar med frågeformulär som skickats ut till alla hushåll. Sedan Folk- och Bostadsräkningen 1970 har det förekommit integritetsdebatter i samband med varje ny folkräkning. Dessa debatter har ökat bortfallet även i andra undersökningar, som t.ex. Arbetskraftsundersökningarna. Om Folk- och Bostadsräkningen upphör och ersätts av löpande registerbaserad statistik, så kommer troligen dessa negativa debatter att försvinna och den minskade uppgiftslämnarbördan kommer istället att uppfattas positivt. De traditionella totalundersökningarnas tid är nog ute – kostnaderna och uppgiftslämnarbördan uppfattas som orimliga och kvaliteten minskar genom ökande bortfall.

1.7 Administrativa och statistiska informationssystem

Att utnyttja administrativa data för statistiska ändamål är inget specifikt för SCB eller andra länders statistiska centralbyråer, utan är vanligt inom alla större företag och organisationer. Administrativa system används allmänt som källa för statistisk information och det är ingen principiell skillnad mellan följande företagsexempel och SCB:s registerstatistik:

- Personalstatistik och lönestatistik inom ett företag får man med hjälp av det personaladministrativa systemet.
- Befolkningsstatistik och inkomststatistik tas fram inom SCB med hjälp av data från Skatteverkets skatteuppbördssystem för folkbokföring och taxering.

Registerbaserade undersökningar har blivit vanligare inom företag och organisationer. Kunskaper om registersystem, registerstatistik och registerkvalitet behövs inte endast inom SCB utan också mer allmänt. Följande utdrag ur en platsannons illustrerar detta:

Marknadsanalytiker

Som analytiker på marknadsavdelningen kommer du att vara en viktig kugge i företagets fortsatta expansion. Du får förvalta och utveckla användandet av en av företagets värdefullaste tillgångar – kundregistret. Du får jobba med kampanjanalyser, att skapa rapporter, segmenteringar och registerkvalitet. Du kommer att ha kontakt med externa registerföretag och arbeta nära marknadsansvariga.

Vissa informationssystem är uppbyggda för enbart statistiska ändamål, som t.ex. SCB:s Arbetskraftsundersökningar. Sådana system kan då helt utformas efter statistiska principer.

Andra informationssystem används både för administrativa och statistiska syften och det kan då finnas konflikter mellan administrativa och statistiska principer med avseende på systemets uppbyggnad. Som regel är dessa system i första hand avsedda för administrativa ändamål och den statistiska informationen blir då en biprodukt.

Mellan ett rent administrativt och ett rent statistiskt system finns flera rent principiella skillnader. Administrativa och statistiska system jämförs nedan.

1.7.1 Olika syften

Informationen i ett administrativt system används som underlag när man vidtar administrativa åtgärder som berör objekten i systemet. *Exempel:* Ett personaladministrativt system används för att verkställa löneutbetalningar.

Informationen i ett statistiskt system används som underlag för analys och kunskapsbildning. Denna kunskap kan sedan vara underlag för beslut av policykaraktär.

Exempel: Ett lönestatistiskt system används för att studera lönestrukturen. Hur har den förändrats? Vilka löneskillnader finns mellan olika personalkategorier? Denna analys kan sedan medföra ändrad policy i lönefrågor.

1.7.2 Olika roller för enskilda objekt

I ett administrativt system fattar man beslut och vidtar åtgärder som avser enskilda objekt. I samband med detta tar man fram uppgifter som avser ett objekt i taget.

Exempel: Man betalar lön till var och en av de anställda. Man kontrollerar vissa administrativa uppgifter och beräknar lön och skatt för de anställda.

I ett statistiskt system har de enskilda objekten inget intresse var för sig. I den statistiska analysen används grupper av objekt för att beräkna olika mått.

Exempel: Lönesummor, medellöner, lönespridningar etc. beräknas för olika personalkategorier.

1.7.3 Synen på fel

Från administrativa synpunkter måste vissa slag av individuella uppgifter vara helt korrekta medan andra slag av uppgifter kan ges en mer ungefärlig redovisning. Från statistisk synpunkt får fel förekomma, med de skall vara under kontroll och man strävar efter att minska de fel som är betydelsefulla för de statistiska slutsatserna. Man kan acceptera fel i enstaka uppgifter om de sammantaget bara har en begränsad effekt.

Exempel: Personnumret i ett personaladministrativt system måste vara helt korrekt ur löne- och skatteadministrativ synpunkt. Personnumret används även då man beskriver personalens åldersstruktur – om t.ex. 30 % av personalen fått felaktigt månadsnummer så störs inte den statistiska analysen nämnvärt medan däremot löne- och skatterutinerna omöjliggörs.

Kapitel 2. Hur strukturera ett registersystem?

En statistikproducent som SCB behöver en modell som beskriver de register som finns inom verksamheten och som också beskriver sambanden mellan dessa register.

Mängden av alla register betraktas som ett statistiskt informationssystem och förståelsen av detta system underlättas av en *konceptuell modell* som utan att vara datateknisk beskriver systemets objekttyper och relationer på det sätt som anges i Sundgren (1992). Dels behövs modeller som beskriver hur systemet faktiskt ser ut för närvarande och dels behövs modeller som anger en tänkt förbättrad struktur som kan ligga till grund för utvecklingen av systemet.

2.1 Objekttyper och relationer

Den färdiga modellen har tidigare beskrivits av Registerprojektet (2001) och beskrivs närmare i kapitel 3. Arbetet med att ta fram denna konceptuella modell av registersystemet utgick från en inventering av de register och registerprodukter som fanns inom SCB. En statistisk undersökning börjar med att man definierar populationens objekt. På motsvarande sätt började vi inom Registerprojektet det konceptuella modellarbetet med att sortera de inventerade register efter typ av objekt. Det är uteslutande statistiskt betydelsefulla objekttyper som studeras, dvs. objekttyper som ingår i undersökningspopulationer.

Vissa register innehåller hierarkier av objekttyper och det är därför naturligt att gruppera på följande sätt:

- person och familj/hushåll (idag familj, i framtiden hushåll)
- arbetsställe och företag (även organisationer/myndigheter kallas företag)
- fastighet, värderingsenhet och taxeringsenhet
- fordon

Mellan olika typer av objekt kan det finnas *relationer* av olika slag:

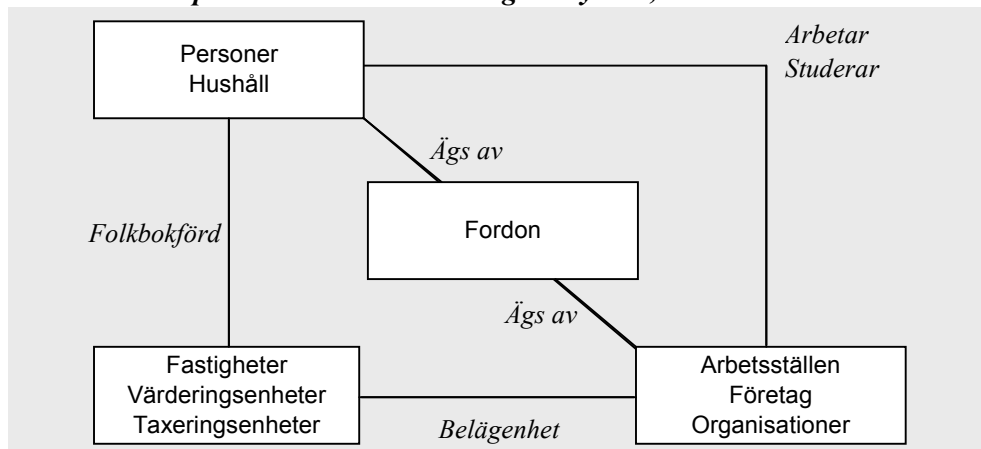
- Mellan person och arbetsställe: En person *är anställd* av ett visst företag eller organisation och *arbetar* på ett visst arbetsställe. En liknade relation gäller mellan studerande och utbildningsställe: En person *deltar i en utbildning* som anordnas av en viss organisation eller ett visst företag och undervisningen sker på ett visst utbildningsställe/arbetsställe.
- Mellan person och fastighet: En person *är folkbokförd* på en viss fastighet.
- Mellan arbetsställe och fastighet: Ett arbetsställe *är beläget* på en viss adress och adressen anger en viss fastighet.
- Mellan person och fordon respektive företag och fordon: Ett visst fordon *ägs* av en person eller ett företag.

Dessa relationer är mycket viktiga ur administrativ och juridisk synvinkel. Det är därför som de administrativa källorna innehåller information av god kvalitet om dessa relationer. Detta är också mycket viktigt för ett fungerande register-

system – när de administrativa källorna innehåller uppgifter om relationerna får man *länkar* i systemet. En länk mellan två objekt består av en eller flera gemensamma *kopplingsvariabler* som innehåller den information som behövs för att identifiera relationen mellan objekten. Eftersom de relationer som anges ovan avser relationer mellan olika typer av objekt har dessa länkar en mycket viktig roll i registersystemet – i den färdiga modellen utgör de länkarna mellan systemets basregister.

Med dessa objekttyper och relationer har vi skapat modellen nedan. Rektanglarna representerar *objektmängder* och linjerna representerar *relationer* mellan objekt. Objektmängderna ger upphov till register i systemet och relationerna ger upphov till länkar.

Bild 2.1 Konceptuell modell av SCB:s registersystem, version 1



I bilden ovan finns fyra register som innehåller olika typer av objekt av stort statistiskt intresse. De fyra rektanglarna motsvaras av fyra SCB-register: *Befolknings-, Fordons-, Fastighets- och Företagsregistret*.

När vi talar om t.ex. Befolkningsregistret (singularis) avser vi egentligen en mängd register (pluralis) då Befolkningsregistret finns i många versioner: en personversion, en familjeversion, olika versioner för befolkningen vid olika tidpunkter etc.

Relationen *Folkbokförd* anges i Befolkningsregistret genom en kopplingsvariabel som anger identiteten för den bostadsfastighet, där en viss person är folkbokförd. Relationerna *Ägs av* anges i Fordonsregistret genom att fordonsägarers person- eller företagsidentitet ingår som kopplingsvariabel i registret.

Varken Befolknings- eller Företagsregistret innehåller kopplingsvariabler som anger relationerna *Arbetar/Studerar*. Men det finns andra register som innehåller information om förvärvsarbete respektive studieaktiviteter:

- Kontrolluppgifts- och lönerregistren innehåller variabler som ger information om anställningar. Objekttypen anställning definieras av relationen mellan objekttyperna anställd person, arbetsgivare och arbetsställe. För varje kombination av personidentitet, företags/organisationsidentitet och arbetsställe-identitet finns uppgifter om förvärvsinkomst/lön etc.

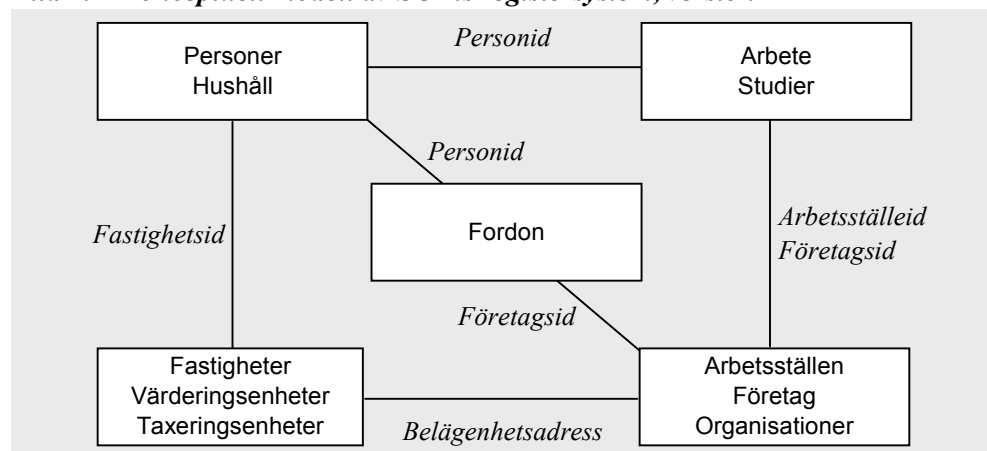
- Studeranderegistren innehåller variabler som anger relationen mellan personer som studerar, utbildningsanordnare och utbildningsställe. Utbildningsställen är arbetsplatser för skolpersonal och alltså en undertyp av objekt-typen arbetsställe.

Information om relationerna *Arbetar/Studerar* skulle kunna lagras i Befolkningsregistret men då många förvärvsarbetande har relationer till mer än en arbetsgivare skulle denna information bli ofullständig då man skulle tvingas att begränsa informationen till enbart en arbetsgivare och då tvingas att välja den huvudsakliga arbetsgivaren.

En bättre lösning är att skapa ett speciellt register som består av *relationsobjekt* dvs. objekt som beskriver relationen mellan person, företag och arbetsställe, där samma person kan ha flera relationer till olika företag och arbetsställen. I bild 2.2 har vi infört relationsobjekten *Arbete/Studier* i ett särskilt register, som vi kommer att kalla *Aktivitetsregistret*. Relationerna i bild 2.1 har i bilden nedan ersatts med de variabler som fungerar som länkar mellan objekt i olika register.

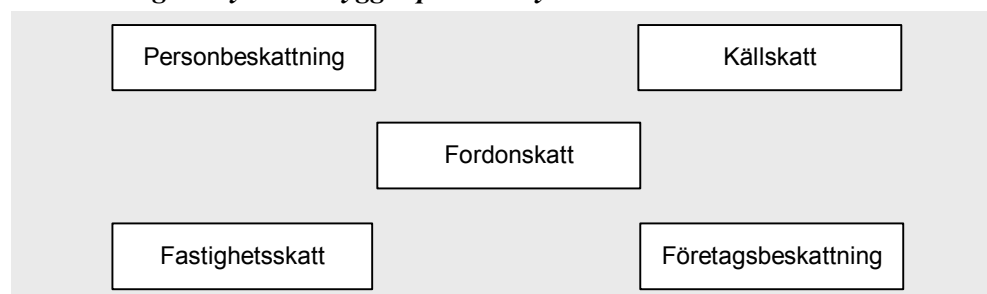
Relationsobjekt
beskrivs i
Sundgren (1992)

Bild 2.2 Konceptuell modell av SCB:s registersystem, version 2



SCB:s system av statistiska register bygger på olika administrativa källor. I bilden nedan anges de fem delar av skattesystemet som är källorna till de fem statistiska register som anges i bild 2.2 ovan.

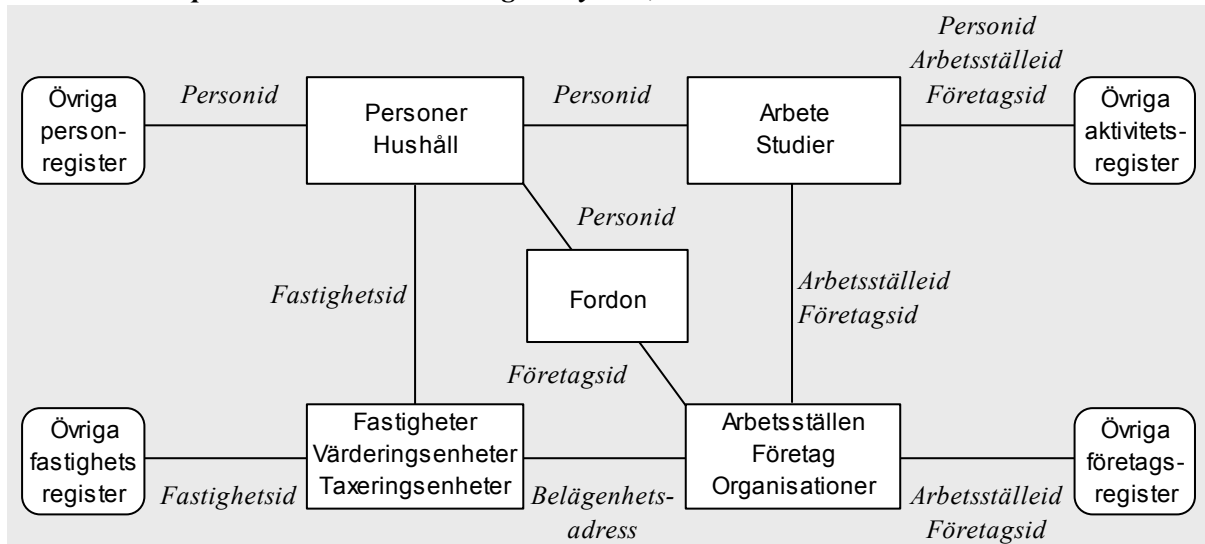
Bild 2.3 Registersystemet bygger på skattesystemets olika delar



Förutom de fem register som anges i bild 2.2 finns ytterligare cirka 50 register på SCB. Med hjälp av befintliga identitetsvariabler kan dessa kopplas till något av Befolknings-, Aktivitets-, Fastighets- eller Företagsregistret. Dessa fyra

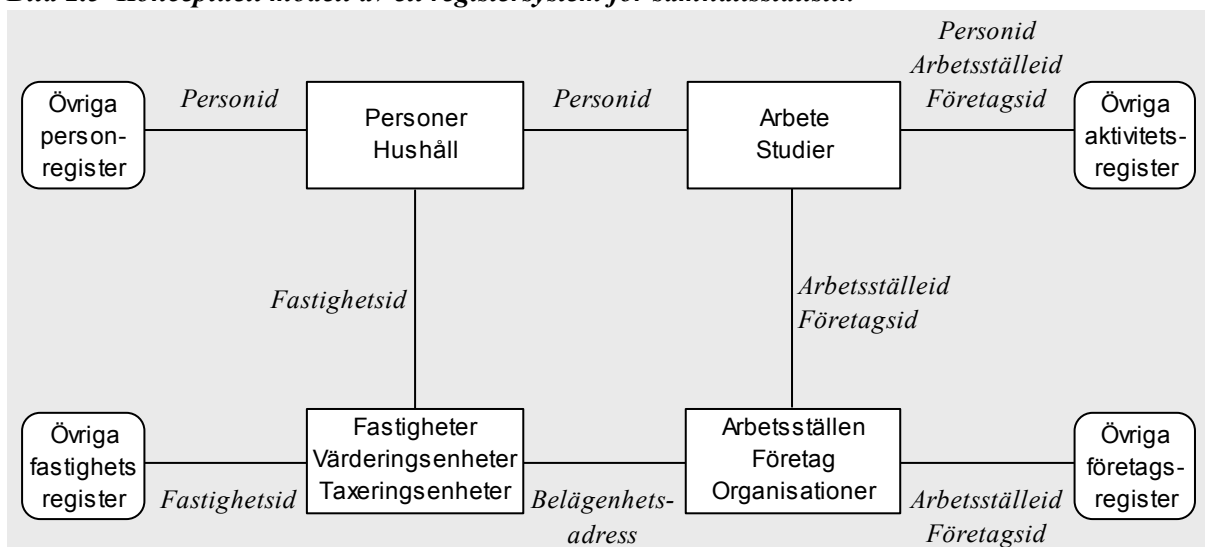
register har en särskild betydelsefull roll inom registersystemet och kallas därför för systemets *basregister*. Hela registersystemet beskrivs i den fullständiga konceptuella modellen i bild 2.4. Registersystemet innehåller ett system av objekt som förändras över tiden. Vid varje tidpunkt har varje objekt ett antal variabelvärden och står i vissa relationer till andra objekt i systemet.

Bild 2.4 Konceptuell modell av SCB:s registersystem, version 3



Fordonsregistret innehåller objekttypen fordon som har en mer begränsad roll för systemet som helhet, varför vi av pedagogiska skäl låter personägda fordon ingå bland övriga personregister och företagsägda fordon bland övriga företagsregister. Fordonsregistret skulle annars dominera bilden av registersystemet. Den slutliga modellen beskrivs nedan. Kan denna modell användas generellt? Då samma objekttyper och variabelinnehåll finns i många länder, tror vi att modellen är generellt användbar för att beskriva system för samhällsstatistik.

Bild 2.5 Konceptuell modell av ett registersystem för samhällsstatistik



2.2 Systemet av basregister

De fyra basregistren är grundvalen för hela registersystemet då de innehåller viktiga objekttyper och länkar. Om något av dessa basregister saknas eller har låg kvalitet skulle hela registersystemet vara betydligt mindre användbart för statistiska ändamål. Basregistrens roll är att definiera objekten i systemet där bra objektavgränsningar och bra täckning är viktiga egenskaper som är avgörande för kvaliteten i hela systemet.

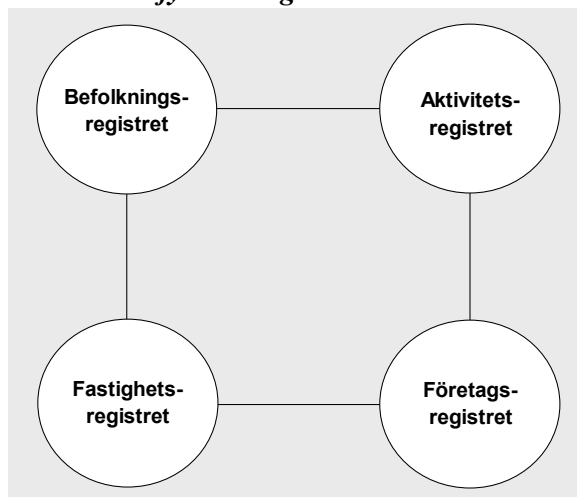
De variabler som är viktiga i ett basregister är olika variabler som identifierar objekt och kan användas för att koppla objekten till objekt i andra register. Tidsuppgifter för olika händelser som berör objekten behövs för att kunna skapa bestånd avseende viss tidpunkt/tidsperiod.

Förutom att de skall bygga på stabila och tillförlitliga administrativa källor skall basregistren ha följande egenskaper:

1. definiera viktiga objekttyper
2. definiera viktiga objektmängder och *standardiserade populationer*, som kan följas över tiden
3. ha länkar till objekt i andra basregister
4. ha länkar till övriga register som avser samma objekttyp
5. vara viktiga för systemet som helhet – därför är det viktigt att de har hög kvalitet och är väldokumenterade
6. vara viktiga som urvalsramar
7. kunna användas för demografisk statistik avseende personer, aktiviteter, fastigheter respektive företag. På samma sätt som man beskriver befolkningens åldersfördelning och befolkningens förändringar (födda och döda), kan man beskriva åldersfördelning och förändringar bland stocken av alla jobb (förvärvsarbetsaktiviteter), byggnader eller arbetsställen.

De fyra basregistren beskrivs i bild 2.6 nedan. I fortsättningen avbildas dessa som cirklar i modellerna på det sätt som infördes av Registerprojektet 1996.

Bild 2.6 De fyra basregistren



T.ex. "Befolkningsregistret" i bilden avser många registerversioner då varje basregister kan finnas i flera versioner:

- Aktuellt bestånd
- Beståndet en viss tidpunkt, t.ex. den 31/12
- Förändringar under en viss period, t.ex. kalenderår
- Alla objekt som funnits under en viss period, t.ex. årgångsregister

Vi kommer här att diskutera dessa fyra basregister och hur de befintliga register som fanns år 2003 överensstämmer med de sju krav som anges ovan.

2.2.1 Befolkningsregistret

Detta basregister har kommit längst när det gäller att utveckla registerstatistisk metodik för arbetet med ett basregister. Alla de sju krav på basregister som vi anger på sidan 28 ovan är uppfyllda. Befolkningsregistret kan därför användas som inspirationskälla för de övriga basregistren. De som arbetar med Befolkningsregistret kan å jourhållning, har en lång tradition av att producera avancerad registerstatistik, gör regelmässigt demografisk analys och är vana att skapa integrationsregister för olika ändamål. Man har utvecklat ett avancerat produktionssystem där objektens historik lagras.

Registret bygger på administrativa uppgifter från Skatteverkets folkbokförings-system. SCB hämtar dagligen dessa uppgifter från SKV:s elektroniska brevlåda samt vissa årliga uppgifter från Migrationsverket. Uppgifterna avser variabler för identifiering och lokalisering av personer samt demografiska variabler som ålder, kön civilstånd etc.

I samband med bytet från stordator till PC-nätverk i slutet av 1990-talet utökades variabelinnehållet i Befolkningsregistret. Bättre *adressuppgifter*, bättre uppgifter om *relationer* till andra personer samt *datum* för alla folkbokföringshändelser ger nya möjligheter för statistikproduktionen. Ändrade och tillfälliga personnummer registreras så att matchningar underlättas. Dessutom registreras olika slag av rättelser så att det är möjligt att skilja mellan verkliga händelser och rättelser. Dessa förbättringar är viktiga för registrets roll som basregister. Wilén och Johannesson (2002) beskriver det nya Befolkningsregistret.

Vissa svagheter finns emellertid:

- Hushållsavgrensningarna har ännu år 2003 brister som kan avhjälpas med ett bra lägenhetsregister och att personer är folkbokförda på lägenhet.
- Övertäckning finns för vissa kategorier av invandrare. Detta upptäcktes inom SCB och Skatteverket (SKV) informerades, se Nilsson (1995). SKV har därefter förbättrat kontrollen av folkbokföringen för att få ned övertäckningen.

2.2.2 Aktivitetsregistret

Detta basregister var år 2002 under utveckling. Registret skall bestå av tre olika delar: *förvärvsaktiviteter*, *studieaktiviteter* och *andra arbetsmarknadsrelaterade aktiviteter*. Som vi tidigare nämnt består registret av relationsobjekt. Varje sådant objekt identifieras med tre kopplingsvariabler: personnummer, företagets organisations- och arbetsställenummer enligt Företagsregistret.

Uppgifter om *förvärvsaktiviteter* bygger på arbetsgivarnas årliga kontrolluppgifter, som ger förvärvsinkomst för varje kombination av *anställd* och arbetsställe. Självdeklarationer ger uppgifter om de personer som är aktiva som *egna företagare*. För offentliganställda får SCB månadsvisa uppgifter från arbetsgivarnas löneregister. År 2002 var registret uppbyggt för att beskriva aktiviteten förvärvsarbete och används i huvudsak för att bilda det personregister som

beskriver befolkningens sysselsättning. I princip är förvärvsinkomst, månadslön, tjänstgöringens omfattning, yrke och pendling egenskaper hos relationsobjekten i Aktivitetsregistret. I det årliga Sysselsättningsregistret är objektet däremot person och registret beskriver sysselsättning för varje person.

Uppgifter om *studieaktiviteter* bygger på uppgifter från skolor och universitet. Det finns en rad register över olika slag av studerande. Dessa register innehåller de studerandes personnummer och uppgift om skola eller studieställe, som i sin tur är arbetsställen i Företagsregistret. Genom att ge dessa studieställen arbetsställesnummer kommer t.ex. Skolregistret att kunna kopplas till Företagsregistret och aktiviteten studier kommer i Aktivitetsregistret att vara en relation mellan ett personnummer och ett arbetsställesnummer. På samma sätt som man bildar pendling för förvärvsarbetande med personnummer och arbetsställesnummer, som ger bostadens respektive arbetsställets belägenhet, kan man då bilda pendling för studerande.

Även uppgifter om *andra arbetsmarknadsrelaterade aktiviteter*, relaterade till försörjning, skall ingå i det nya registret. Olika myndigheter har information om värnplikt, sjuk- och förtidspensionering, arbetsmarknadsåtgärder, registre-rad arbetslöshet och slutna sjukvård. Denna information ger en kompletterande bild av arbetsmarknadsstatus utöver förvärvsarbete och studier. Det finns uppgifter i de administrativa källorna som tidslokaliserar aktiviteterna även om kvaliteten på tidslokaliseringen ibland är låg.

Det finns en stor mängd administrativa data som innehåller kopplingar till både person och företag. Rent datatekniskt skulle dessa uppgifter kunna betraktas som antingen person- eller företagsdata. Vi har istället valt att betrakta dessa uppgifter som data som beskriver *aktiviteter*. Dessa aktiviteter uppfattar vi som en egen objekttyp av två skäl, de är dels en statistiskt intressant objekttyp och dels måste denna objekttyp särskiljas för att registersystemet som helhet skall få en pedagogiskt klar struktur.

Uppfyller Aktivitetsregistret de sju kraven på ett basregister (enligt sidan 28)?

1, 2, 7. *Definierar viktiga objekttyper och objektmängder, populationer. Kan användas för demografisk statistik avseende aktiviteter.*

När utbud och efterfrågan på arbetsmarknaden möts skapas relationer mellan personer och företag/organisationer. Dessa relationer är viktiga undersökningsobjekt för arbetsmarknadsstatistiken. Demografisk statistik, som beskriver hur t.ex. stocken av förvärvsaktiviteter förändras genom så kallad *job-creation* och *job-destruction*, måste vara mycket relevant för studier av arbetsmarknaden.

3. *Har länkar till objekt i andra basregister*

Kopplingsvariabeln personnummer är länken till Befolkningsregistret och kombinationen av kopplingsvariablerna organisationsnummer och arbetsställesnummer är länken till Företagsregistret.

4. *Har länkar till övriga register som avser samma objekttyp*

Kopplingsvariablerna person-, organisations- och arbetsställesnummer var år 2003 länken till Kontrolluppgiftsregistret, Privatanställdas löner och Yrkesregistret. När övriga löneregister och studeranderegistren standardiserats med

avseende på de variabler som identifierar arbetsställen och utbildningsställen kommer aktivitetsregistret att ha länkar till samtliga register som består av dessa relationsobjekt.

5. Viktigt för systemet som helhet

Bland användarna finns ett stort intresse för statistik där persondata och företagsdata kombineras. Aktivitetsregistret har en mycket viktig roll som brygga mellan individ- och företagsstatistiken.

6. Viktigt som urvalsram

Kanske inte direkt, men vissa kategorier av personer eller företag skulle kunna selekteras med hjälp av Aktivitetsregistret.

Aktivitetsregistret kan användas för att skapa både person- och företagsregister

Aktivitetsregistret bygger direkt på den administrativa källan – för varje kombination av person och arbetsställe finns uppgift om årets lönesumma. Genom att summera löneuppgifterna för varje person kan vi skapa ett personregister. Genom att summera löneuppgifterna i aktivitetsregistret för varje arbetsställe kan vi skapa ett arbetsställeregister med lönesummor.

Bild 2.7 Sambandet mellan person-, aktivitets- och företagsregister

Personregister		Aktivitetsregister – jobb			
Person	Årslön	Jobb	Person	Arbetsställe	Årslön
Pnr 1	450 000	J1	Pnr 1	AE 1	220 000
Pnr 2	210 000	J2	Pnr 3	AE 1	180 000
Pnr 3	270 000	J3	Pnr 1	AE 2	230 000
		J4	Pnr 2	AE 2	210 000
		J5	Pnr 3	AE 2	90 000

Företagsregister – arbetsställen	
Arbetsställe	Lönesumma
AE 1	400 000
AE 2	530 000

2.2.3 Företagsregistret

SCB får administrativa data avseende *juridiska enheter* från Patent- och registreringsverket (PRV) och Skatteverket (SKV). Data från PRV, SKV och Svensk Adressändring AB aviseras regelbundet.

Ett omfattande arbete sker inom SCB för att samla in uppgifter från företag med verksamhet vid mer än ett arbetsställe för att kunna skapa *arbetsställe-enheter*. Detta arbete är uppdelat mellan två avdelningar. Inom avdelningen för arbetsmarknadsstatistik arbetar man med kontrolluppgifter och tar många kontakter med företag i samband med att kontrolluppgifterna granskas. Kontrolluppgiften är ett exempel på *integrerad datainsamling*, där dels skatteadministrativa uppgifter samlas in men även vissa uppgifter för statistiska ändamål som bl.a. avser avgränsningen av arbetsställen. Detta arbete avser arbetsställepopulationen föregående kalenderår, till skillnad från registervårdsarbetet inom avdelningen för ekonomisk statistik, som är inriktat på det aktuella beståndet av arbetsställen. Även här gör man flerarbetsställeenkäter och tar många kontakter med företag. Inom Företagsregistret skapas även *företags-, verksamhets- och lokala verksamhetsenheter*.

Integrerad datainsamling

Två principer genomsyrade det utvecklingsarbete som ägde rum i samband med övergången från stordatordrift till PC-nätverk:

- Inga parallella objektmängder skall förekomma inom den ekonomiska statistiken, dvs. populationen av t.ex. industriföretag skall vara densamma både inom industristatistiken och i Företagsregistret.
- Alla inom avdelningen för ekonomisk statistik skall delta i vården av Företagsregistret, dvs. de som t.ex. samlar in data från industriföretag skall vidarebefordra den information de får om industripopulationen till Företagsregistret.

Dessa principer skall gälla alla basregister, inte enbart Företagsregistret. Dessutom skall alla delta i vården av ett basregister oberoende av avdelningsgränser. Konkret innebär detta att de som arbetar med Lantbrukets företagsregister skall lämna underlag för lantbruksdelen av Företagsregistret och att de som arbetar med Skolregistret lämnar underlag så att Företagsregistrets skoldel blir så bra som möjligt.

Uppfyller Företagsregistret de sju kraven på ett basregister (enligt sidan 28)?

1, 2, 7. Definierar viktiga objekttyper och objektmängder, populationer. Kan användas för demografisk statistik avseende arbetsställen och företag.

Ett flertal viktiga objekttyper och populationer för den ekonomiska statistiken finns inom Företagsregistret. Statistik som beskriver företagens och arbetsstäl- lenas demografi efterfrågas av många användare. Johansson (1997 och 2001) innehåller exempel på hur Företagsregistret kan användas för demografisk företagsstatistik inom ekonomisk forskning. Även exemplet nedan visar att förändringar i företagspopulationen är viktiga att beskriva.

Exempel: Statistik från företagsregistret kan hamna på första sidan i DN:

Dagen Nyheter den 12 juli 2002:

Här finns de nya jobben

Vill du ha jobb? Sök dig till ett tjänsteföretag. Det är de som anställer mest ny personal, enligt statistik som SCB tagit fram åt DN...

SCB:s statistik ska tolkas med ganska stor försiktighet. De företag som ser ut att ha växt allra mest är ofta dotterbolag i en större koncern, där verksamheten organiserats om...

Exemplet visar att statistik baserad på Företagsregistret kan ha stort nyhetsvärde dvs. vara av allmänt intresse för en stor läsekrets. Innehållet i artikeln berör företagens demografi. Bearbetningar av Företagsregistret bör anpassas för att möta statistikens krav – statistiskt intressant tillväxt får inte blandas samman med ointressanta administrativa förändringar som företagsköp och fusioner. Företagens demografi är ett viktigt område som har utvecklats starkt under de senaste åren för att beskriva statistiskt intressanta förändringar i företagspopulationen.

3. Har länkar till objekt i andra basregister

Kombinationen av kopplingsvariablerna organisationsnummer och arbetsstäl- lenummer är länken till Aktivitetsregistret. Arbetsställets belägenhetsadress är

länken till Fastighetsregistret, fortfarande finns det dock år 2003 brister med denna kopplingsvariabel.

4. Har länkar till övriga register som avser samma objekttyp

Kopplingsvariablerna organisations- och arbetsställenummer är länken till andra register som avser företag/organisationer och arbetsställen. För personliga företag fungerar företagarens personnummer som organisationsnummer. Även objekttypen *Företagsenheter* identifieras med en sifferkod som kan användas som kopplingsvariabel.

Lantbrukets företagsregister och Skolregistret med utbildningsanordnare och utbildningsställen hade år 2003 bristfälliga länkar till Företagsregistret.

5. Viktigt för systemet som helhet

Företagsregistret är betydelsefullt inte enbart för företagsstatistiken utan även för arbetsmarknads- och individstatistiken. Uppgifter om arbetsställets näringsgren och belägenhet används inom många statistikprodukter.

6. Viktigt som urvalsram

Används som ram för många urvalsundersökningar som avser företag, företagsenheter eller arbetsställen.

2.2.4 Fastighetsregistret

År 2003 fanns inte ett renodlat basregister för fastigheter. Det Fastighetstaxeringsregister som har funnits länge har inte samma struktur som de övriga tre basregistren. Ett basregister skall avgränsa populationer av objekt. Viktiga variabler är då sådana som identifierar och lokaliserar objekten och anger tidpunkter för viktiga händelser. Fastighetstaxeringsregistret i nuvarande form är inte något idealt basregister för fastigheter. Ett sådant bör på sikt byggas upp med hjälp av Lantmäteriverkets fastighets- och byggnadsregister och kompletteras med det planerade lägenhetsregistret samt det nuvarande Fastighetstaxeringsregistret. I ett nytt basregister för fastigheter skall det finnas länkar till taxeringsenheter och byggnader. Basregistret kan uppdateras med lagfartsuppgifter, bygglagsuppgifter och fastighetstaxeringsuppgifter.

Ett renodlat basregister för fastigheter och andra objekttyper som hör samman bör bildas. Ett sådant register skall uppdateras ofta så att uppdelningar och sammanslagningar av fastigheter kan följas. Registret skall uppfylla alla de krav vi ställer på basregister (enligt sidan 28). Följande objekttyper bör ingå:

fastighet

byggnad

lägenhet (både bostäder och lokaler)

taxeringsenhet

värderingsenhet

Viktiga variabler i ett sådant basregister är identiteter på de olika objekt som tillhör ovanstående objekttyper. Belägenhetsadresser och geografiska koordinater och koder skall också ingå. Adresserna skall förbättras så att fastigheter med bostäder och arbetsställen kan knytas till standardiserade adresser.

2.3 Grundläggande variabler i registersystemet

Vissa variabler används inom många register och har en viktig roll för systemet som helhet. Det register som tar in en variabel från en annan myndighet eller själva skapar en variabel ska ha ansvaret för denna på SCB. Att ha ansvar för en *standardiserad variabel* innebär följande:

- ha kontakt med registerlämnande myndighet och ansvara för att variabeln och information om variabeln hämtas från myndigheten till SCB
- ansvara för granskning, bearbetning, namngivning och dokumentation.

Andra register inom SCB som använder en standardiserad variabel skall använda den bearbetade variabeln utan ändringar, dess korrekta namn och den ursprungliga dokumentationen. På detta sätt minskas dubbelarbete och sambearbetningar underlättas samtidigt som konsistensen inom systemet förbättras. Två kategorier av standardiserade variabler har speciella roller inom systemet, länkar och klassifikationer.

2.3.1 Länkar

De länkar som systemet bygger på anges i bild 2.5 och baseras på följande identiteter: *Person-*, *Arbetsställe-*, *Företags-* och *Fastighetsidentitet*. Vidare är *Arbetsställets belägenhetsadress* länk mellan Företags- och Fastighetsregistret.

När ett register skapas i registersystemet görs många matchningar mellan olika register. Ett administrativt register matchas mot ett basregister för att avstämma objektmängden, man matchar mot olika andra statistiska register för att kunna hämta variabler till det nya registret. Alla dessa matchningar förutsätter att de olika registren innehåller bra länkar. En länk består av en eller flera variabler som identifierar enskilda objekt. För att underlätta matchning av olika register fodras att samma identifierande variabler finns i många register. När man matchar register som avser olika årgångar behöver man länkar som är stabila över tiden dvs. ett enskilt objekt som är oförändrat skall ha samma värden på de identifierande variablerna.

Exempel: Personnummer är en mycket stabil variabel vilket underlättar registerbearbetningarna. Fastighetsidentitet består av län, kommun och fastighetsbeteckning. Detta är ett exempel på en mycket instabil variabel, om läns- eller kommunkod ändras kommer många fastighetsidentiteter att ändras.

2.3.2 Klassifikationer

Näringsgren, varugrupp, utbildning, yrke m.m. är exempel på viktiga statistiska *standarder* och *klassifikationer*. Dessa baseras på internationella rekommendationer och är innehållsmässigt betydelsefulla och används inom många undersökningar, både registerbaserade och andra.

De administrativa källorna innehåller uppgifter om dessa klassifikationer och denna information används för att skapa variablerna inom registersystemet. Dessutom måste de som ansvarar för en standard inom SCB kunna *koda* nya yrkesbenämningar, utbildningar etc. så att de nya benämningarna förs till lämpliga kategorier inom klassifikationen.

En *standardiserad variabel* är så viktig att ansvaret för variabeln klargörs genom särskilt beslut.

I läroböcker används samma term för ett annat begrepp: En variabel transformeras till en *standardiserad variabel* så att medelvärdet blir 0 och standardavvikelsen blir 1

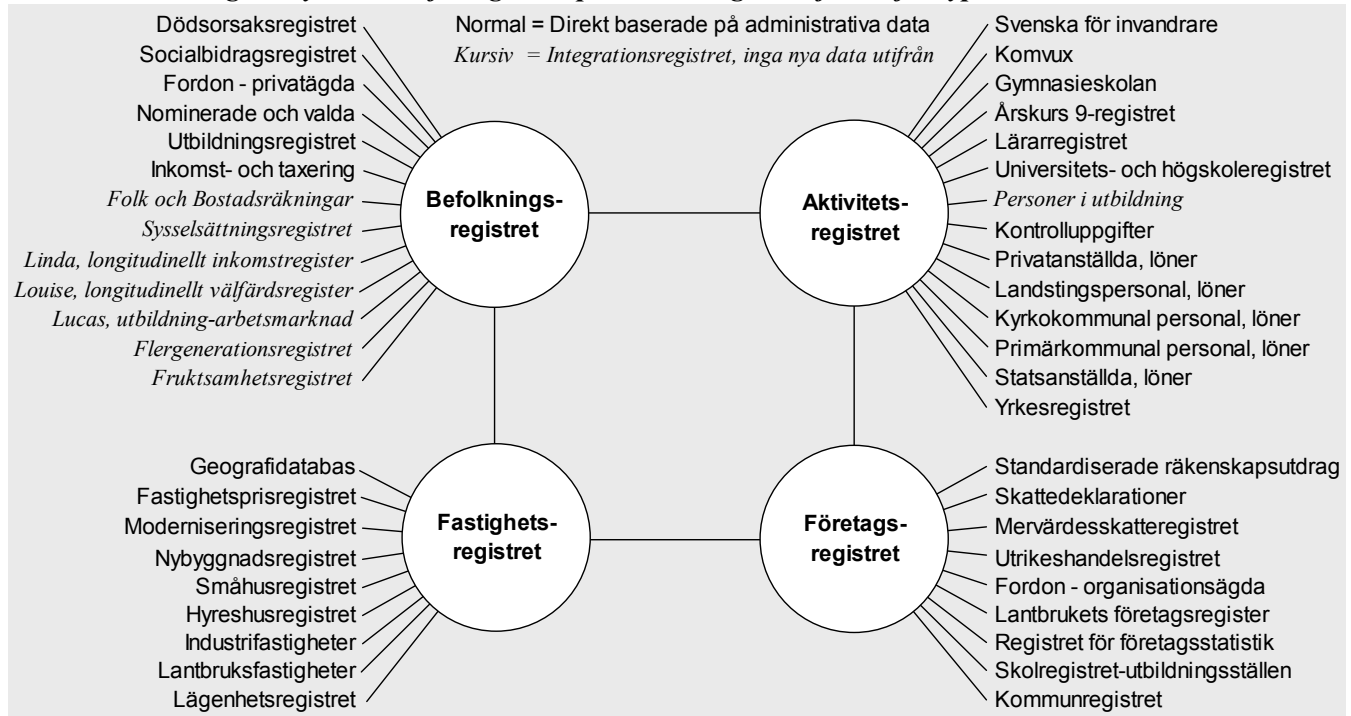
Standarder
Klassifikationer

Kodning

2.4 Hela registersystemet

Nästa steg är att koppla övriga statistiska register till respektive basregister. Cirkarna i bild 2.8 representerar basregistren och linjerna befintliga eller planerade länkar mellan objekten i olika register. Ett väl fungerande system fordrar att objektdefinitioner har samordnats, att tidsreferenser och gemensamma variabler är harmoniserade och att det finns bra identifieringsvariabler som kan användas för att länka objekt i olika register.

Bild 2.8 SCB:s registersystem – befintliga och planerade register efter objekttyp



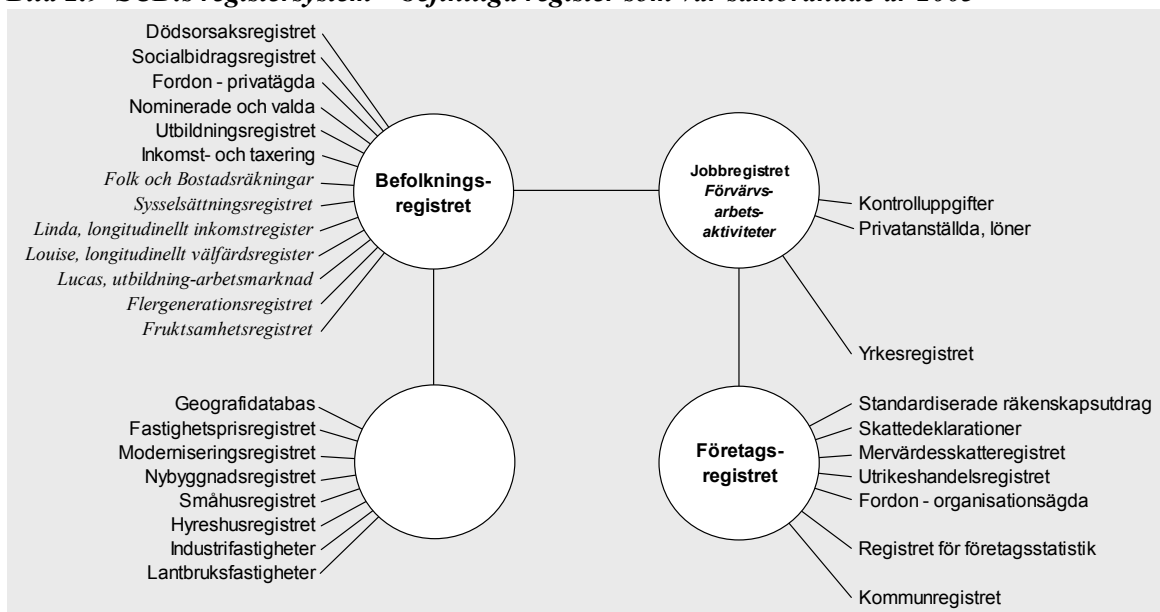
De basregister, övriga statistiska register och länkar som ingår i SCB:s registersystem visas i bild 2.8 ovan. De flesta komponenterna i systemet fanns år 2003, men ett antal register och länkar bör samordnas och utvecklas för att registersystemet skall fungera väl.

Den verklighet registersystemet avbildar förändras kontinuerligt – objekt tillkommer eller försvinner. Relationer mellan objekt förändras och egenskaper och variabelvärden hos objekten ändras. Varje register i modellen ovan finns för flera årgångar och kan finnas i flera versioner.

I bild 2.8 skiljer vi på tre typer av register: basregister (cirklar), register som baseras direkt på administrativa data (normal stil) och register som helt är baserade på befintliga register (*kursiv* stil). Även utvecklingsarbetet inom SCB medför att registersystemet förändras över tiden. Nya register baserade på nya administrativa källor kan tillkomma och nya integrationsregister kan bildas. Vidare kan befintliga register förändras och inom systemet som helhet kan samordningen förbättras.

I bilden nedan beskrivs de delar av systemet som hade fungerande länkar år 2003. Aktivitets- och Fastighetsregister finns, men har inte den struktur som är lämplig för ett basregister.

Bild 2.9 SCB:s registersystem – befintliga register som var samordnade år 2003

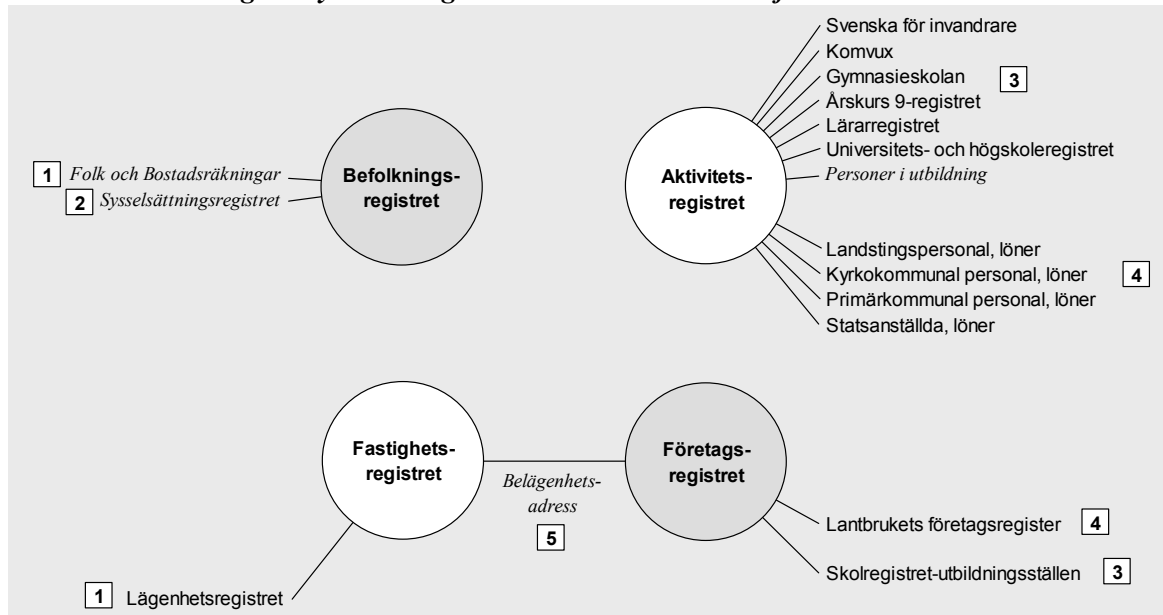


Följande delar av systemet skall förbättras enligt de förslag som presenteras i Registerprojektets slutrapport (2001).

1. För att helt ersätta de traditionella Folk- och Bostadsräkningarna med registerbaserad statistik förutsätts att ett Lägenhetsregister finns som kan användas för att avgränsa hushåll.
2. Sysselsättningsregistret innehåller en variabel som anger om en person förvärvsarbetar under november. Idag är denna variabel endast baserad på Kontrolluppgifter, men genom att använda uppgifter från fler administrativa källor, t.ex. löneregister, skulle kvaliteten i denna variabel kunna förbättras.
3. Skolor och utbildningsställen finns både i Företagsregistret och i olika register som avser skol- och universitetsutbildningar. Men idag används olika identiteter inom dessa register. Då skolregistren innehåller samma organisations- och arbetsställenummer som Företagsregistret blir matchningar möjliga. Även skolornas belägenhetsadresser måste förbättras.
4. På motsvarande sätt bör alla löneregister samt Lantbrukets företagsregister innehålla samma organisations- och arbetsställenummer som i Företagsregistret.
5. Genom att förbättra adresser så att de blir entydiga kan arbetsställes belägenhetsadresser kopplas till rätt fastighet. Länken mellan Företags- och Fastighetsregistret kommer då att få högre kvalitet.

Siffrorna i bilden nedan refererar till dessa planerade förbättringar. Aktivitets- och Fastighetsregistret skall utvecklas som basregister, Lägenhetsregistret planeras och de övriga förbättringarna avser ändrade eller förbättrade länkar.

Bild 2.10 SCB:s registersystem – register och länkar som skall förbättras



Varför fyra basregister?

I vissa andra länder har man utgått från en struktur med tre basregister, individer, företag och fastigheter. Denna struktur med tre basregister avser de *administrativa* register som används för att administrera dessa tre slag av objekt. Då individer, företag och fastigheter administreras fristående från varandra är länkar mellan de administrativa registren inte så betydelsefullt.

Varje administrativt register är en viktig källa till motsvarande *statistiska* basregister, men dessa bygger som regel på fler källor. Dessutom är länkarna mellan basregistren mycket betydelsefulla för det statistiska registersystemet och Aktivitetsregistret behövs därför som länk mellan individer och företag. En registermodell med fyra basregister har en klar struktur och framhäver *aktiviteter* som en statistiskt intressant objektmängd som kan belysas med många administrativa källor. Registermodellen skall kunna användas för pedagogiska syften för att ge en åskådlig bild av systemet och hur de olika delarna samverkar. Detta är ytterligare ett skäl för en modell med fyra basregister.

Strukturen med fyra basregister utgår *inte* från SCB:s indelning i avdelningar. När strukturen förslögs av Registerprojektet 1996 fanns tre av dessa basregister inom samma avdelning

2.5 Statistiska registersystem utanför SCB

Registerstatistik och statistiska register finns även utanför SCB. Våra olika statistikansvariga myndigheter har många register som baseras på administrativa data. Statistikmyndigheter i andra länder blir mer och mer intresserade av att basera statistik på administrativa källor. Dessutom är den interna statistiken inom större företag helt registerbaserad.

2.5.1 Registerstatistik i andra länder

I andra länders statistiska centralbyråer arbetar man med undersökningar som avser samma objekttyper som finns i SCB:s registersystem. Det innebär att man skulle kunna strukturera sina befintliga eller planerade registersystem på liknande sätt som SCB:s system enligt bild 2.5 eller 2.8.

Det förekommer att man av tradition strukturerar registerstatistiken i två delsystem, ett system för person- eller socialstatistik och ett annat för ekonomisk statistik. Då förbiser man de möjligheter som finns om man kan integrera informationen i både personregister och företagsregister. Med det registersystem som vi beskriver i bild 2.8 kan man t.ex. beskriva *personer* med egenskaper hos de *företag* där de är anställda – sysselsatta personer efter näringsgren, efter företagets storlek, efter företagets ålder, etc. På motsvarande sätt kan *företagen* beskrivas med egenskaperna hos de *personer* som tillhör företagets personal – andel kvinnor/män, andel unga, andel högutbildade, etc. För att kunna tillgodose statistikbehov av denna typ fordras ett sammanhängande registersystem.

2.5.2 Statistiska register hos andra statistikansvariga myndigheter

Eftersom produktionen av samhällsstatistik i Sverige har decentraliserats så har flera statistikansvariga myndigheter (SAM) ansvar för en rad olika register. Vissa av dessa register sköts på uppdragsbasis av SCB medan andra register finns hos respektive myndighet utanför SCB. I avsnitt 3.5 diskuterar vi dessa SAM-register och hur deras anknytning till registersystemet.

2.5.3 Statistiska register inom företag

Företag och organisationer har egna administrativa registersystem som även kan användas för statistiska syften. Företag inom tillverkningsindustrin har administrativa system som består av två huvuddelar, ekonomisystemet och MPS-systemet (material- och produktionsstyrning). Dessa system innehåller 100-tals register med sammanlagt 1000-tals variabler. På motsvarande sätt som man på SCB använder administrativa register för att skapa statistiska register använder man t.ex. register i MPS-systemet för att skapa särskilda statistiska register som sedan används som källor för företagets interna registerstatistik avseende försäljning. Dessa statistiska register innehåller mikrodatabaser för alla transaktioner som avser fakturering, orderingång och orderstock.

I ett tillverkande företags registersystem finns även vissa register som kan uppfattas som basregister. Artikelregistret och kundregistret är två exempel på sådana register som definierar viktiga objekttyper. Viktiga kopplingsvariabler i systemet är datum, kundidentitet och artikelidentitet.

På samma sätt som vi i denna bok beskriver hur samhällets administrativa system kan användas för statistiska syften, skulle man kunna utreda principerna för hur företagens administrativa data kan användas statistiskt. Att det finns ett behov av detta illustreras av den snabba framväxten av "Data Mining". Statistikvetenskapen skulle kunna bidra till att utveckla registerstatistiska metoder inom detta område. I Wallgren (1986) redogörs för vissa erfarenheter av användningen av företagens administrativa system för statistiska ändamål.

Kapitel 3. Hur skapas ett registersystem?

I kapitel 2 går vi igenom registersystemets struktur och dess olika delar. I detta kapitel beskriver vi arbetsgången från mottagandet av administrativa data fram till skapandet av de register som är underlag för registerstatistiken.

De administrativa register, som finns hos olika myndigheter och från vilka SCB får leveranser, kallar vi för administrativa *källregister*. Med hjälp av dessa administrativa källor skapas inom SCB olika *statistiska register*. Varje sådant statistiskt register kan vara underlag för en eller flera *statistiska produkter*, dvs. undersökningar. Om dessa statistiska register ingår i ett samordnat system kan de sambearbetas och användas så att nya produkter kan framställas.

Register 1	Produkt A
Register 2	Produkt B Produkt C
Register 3	Produkt D
Register 2 o 3 sambearbetas	Produkt E

3.1 Administrativa register omvandlas till statistiska register

Den ursprungliga databildningen sker hos myndigheter och organisationer. Definitioner av objekt och variabler är avpassade för olika administrativa ändamål. Varje myndighet gör kontroller, rättelser och andra bearbetningar som passar för deras administrativa syften. När en myndighet levererar data till SCB görs eventuellt ytterligare selektioner och bearbetningar för att tillfredsställa SCB:s önskemål. Leveransen skall vara fullständig, internt konsistent och avse en viss angiven tidsperiod eller tidpunkt. Inom respektive myndighet finns även metadata i form av kunskaper om definitioner, databildning och kvalitet. Denna typ av information är också viktig för dem som tar emot data inom SCB.

Exempel: SKV:s leveranser till SCB:s Inkomst- och taxeringsregister

1. *Databildningen hos Skatteverket (SKV)*
Den årliga inkomsttaxeringen bygger på inkomsttagarnas självdeklARATIONER och lokala skattemyndigheters taxeringsbeslut. Både inkomsttagare och skattemyndigheter använder kontrolluppgifter avseende lön, sjukpenning etc. och räntor som arbetsgivare, försäkringskassor och kreditinstitut ansvarar för. Denna information sammanställs slutligen hos SKV. Deklaration, kontrolluppgifter och taxeringsbeslut kan ändras och kompletteras.
2. *Mikrodataleveranser till SCB:s Inkomst- och taxeringsregister*
SKV skapar årligen ”taxeringsband” som innehåller uppgifter om Sveriges befolkning. *Taxeringsbandet* för ett år – innehåller ca 9 miljoner poster och varje post innehåller ca 300 variabler – levereras direkt till SCB:s Inkomst- och taxeringsregister. *Kontrolluppgifterna* sammanställs av SKV och levereras till SCB:s Kontrolluppgiftsregister som sedan skickar en kopia till Inkomst- och taxeringsregistret. Leveransen omfattar ca 70 miljoner poster och med över 300 variabler fördelade på 20 olika posttyper.
3. *Metadata till SCB:s Inkomst- och taxeringsregister*
Postbeskrivningar med variabelnamn och variabeldefinitioner medföljer leveranserna från SKV. Deklarationsblanketter, kontrolluppgiftsblanketter, taxeringsbestämmelser, deklarationsanvisningar och anvisningar till arbetsgivare behövs också för att kunna tolka data.

Inkomst- och taxeringsregistret (IoT) kallades tidigare för Inkomst- och förmögenhetsregistret (IoF)

Vad händer när data levererats till SCB?

Det är som regel inte lämpligt att framställa statistik direkt från de mottagna administrativa registren eftersom dessa inte är anpassade för statistikens krav. Objekt mängder, objektavgränsningar och variabler behöver kontrolleras och därefter kan det vara nödvändigt med vissa bearbetningar så att registret uppfyller statistikens krav på objekt och variabler. De registerstatistiska bearbetningar, som syftar till att omvandla ett eller flera administrativa register till ett statistiskt register, skall baseras på genomtänkta och standardiserade *registerstatistiska metoder*. Dessa metoder diskuteras mer utförligt i kapitel 5. Följande bild visar de olika moment som ingår i det metodstatistiska arbetet.

Bild 3.1 Från administrativa register till statistiskt register



Arbetet med att ta emot stora administrativa material samt granska dessa är viktiga moment som är likartade för många registerenheter vid SCB. Det är därför viktigt med erfarenhetsutbyte, gemensam metodutveckling och att väl-etablerade och väldokumenterade metoder används allmänt på det sätt som länge har varit självklart i samband med SCB:s urvalsundersökningar. På motsvarande sätt bör gemensamma metoder och verktyg för bortfallshantering, sambearbetningar och evalveringar användas.

Exempel: SCB:s Inkomst- och taxeringsregister (IoT)

1. *Leverantörskontakter, mottagningskontroll och granskning av grunddata*
IoT-registret får data från elva olika leverantörer både utanför och inom SCB. Data som kommer utifrån granskas och åtgärdas. Data som levereras från andra SCB-register är redan granskade och åtgärdade. Leverantörskontakterna är viktiga för att få kunskap om förändringar i de administrativa systemen, vilket är viktigt för att ha kontroll av kvaliteten hos registerstatistiken – administrativa förändringar får inte misstolkas som verkliga inkomstförändringar. I avsnitt 5.7 behandlas granskning av registerdata.
2. *Sambearbetningar, selektioner*
Det är ett stort antal register som skall sambearbetas för att skapa de olika delregister som ingår i IoT-registret. Här sker då matchningar med hjälp av personnummer och samtidigt aggregeras t.ex. alla kontrolluppgifter för en viss person så att personens arbetsinkomst bildas. Taxeringsmaterialet inne-

håller data som avser ett helt år. En typ av bearbetning är att selektera personer 16 år och däröver som ingick i befolkningen den 31/12.

3. Härledda objekt bildas

Med information om vissa relationer bildas familjer. Mellan vuxna medför relationerna *gifta* eller *sammanboende med gemensamma barn* att man förs till samma familj. Dessa relationer anges med de anhörigas personnummer som finns i taxeringsmaterialet och i SCB:s befolkningsregister.

Härledda objekt diskuteras i avsnitt 5.5

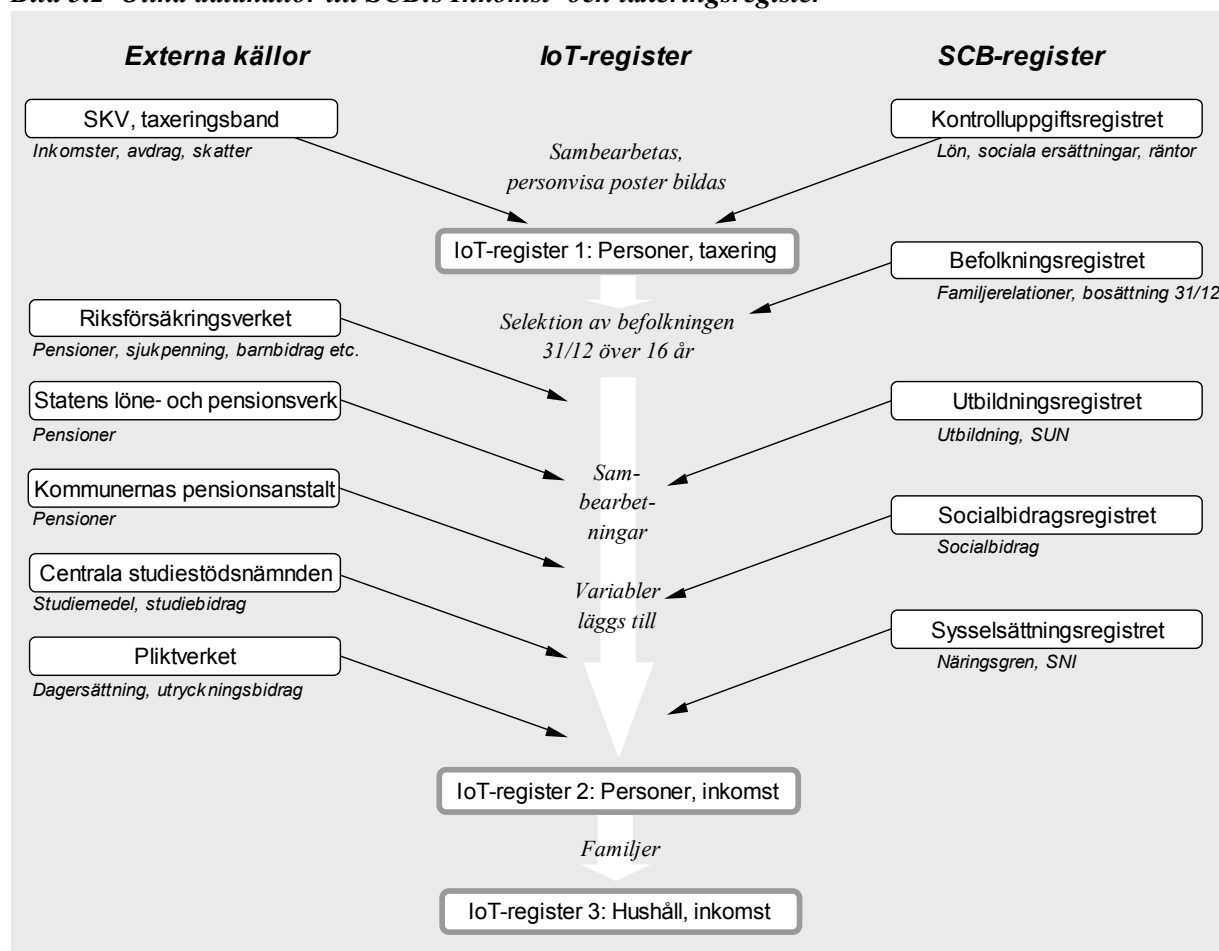
4. Härledda variabler bildas

Ett stort antal härledda inkomstvariabler bildas. T.ex. aggregeras lönebelopp på olika kontrolluppgifter till personens *arbetsinkomst*. Varje persons samtliga inkomster och sociala ersättningar minus skatter blir personens *disponibla inkomst*. För familjer bildas bl.a. variablerna *familjetyp*, *antal konsumtionsenheter* och *disponibel inkomst*.

Härledda variabler diskuteras i avsnitt 4.5.3 och avsnitt 5.8

Bilden nedan visar hur man på Inkomst- och taxeringsregistret tar emot administrativa data från en rad olika externa källor och vissa SCB-register. Med *källregister* avser vi både de administrativa källorna och de SCB-register som används för att skapa det nya registret. De olika arbetsmomenten när källregistren omvandlas till nya statistiska register anges i bildens mittfält.

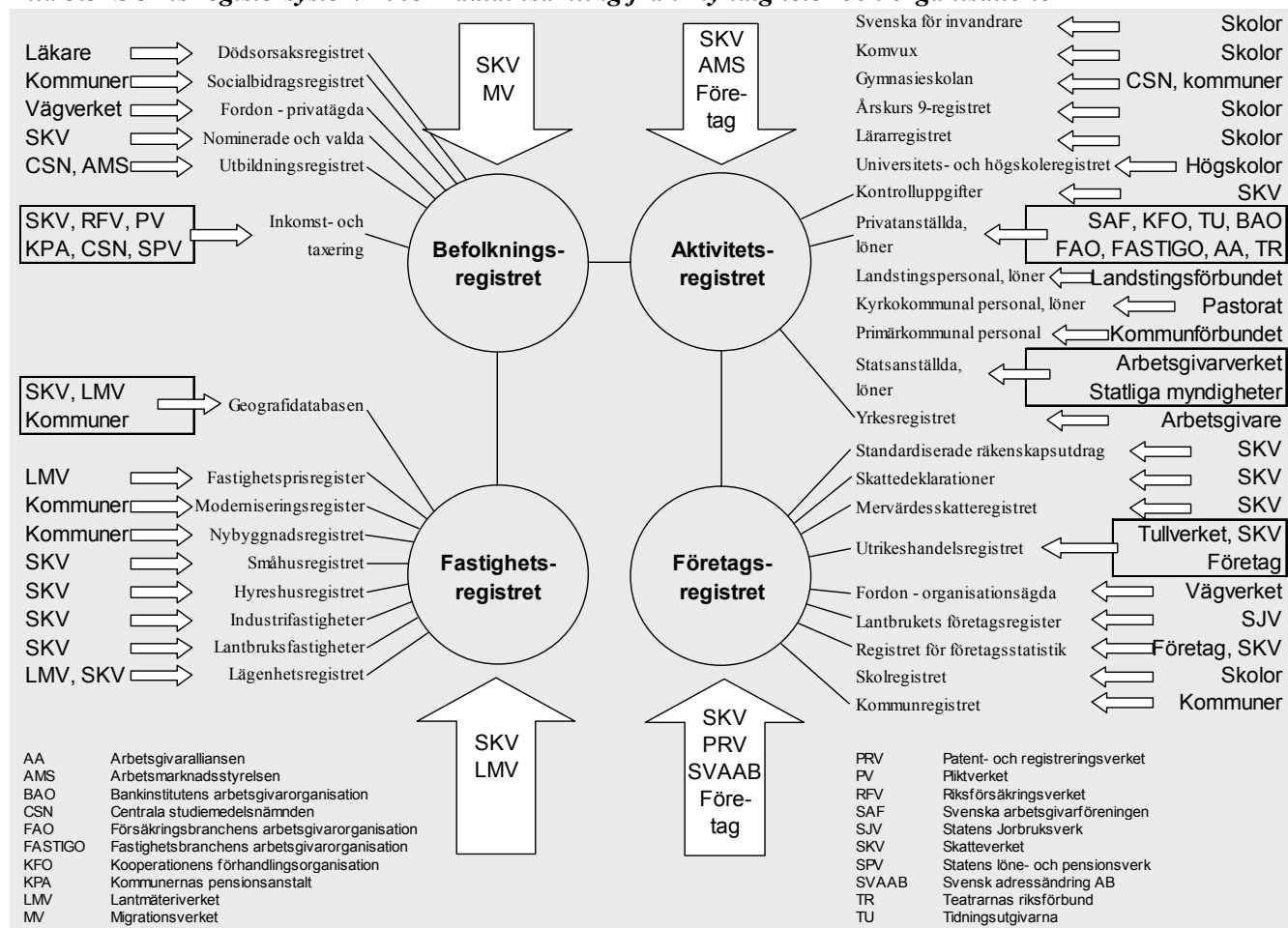
Bild 3.2 Olika datakällor till SCB:s Inkomst- och taxeringsregister



3.2 Systemet får data från många administrativa källor

Från ett stort antal registerlämnande myndigheter och organisationer får olika registerenheter inom SCB administrativa data som bearbetas så att ett antal statistiska register bildas. Bilden nedan beskriver registersystemets indatadel.

Bild 3.3 SCB:s registersystem 2003 – datainsamling från myndigheter och organisationer



Av bilden framgår att många enheter har kontakter med registerlämnare och att många olika myndigheter och organisationer är involverade. SCB:s försörjning av administrativa data organiseras och tryggas genom goda och välorganiserade kontakter med registerlämnande organisationer. Dessa kontakter sker på olika nivåer men bör vara samordnade inom SCB. Huvuddelen av SCB:s datalager kommer ursprungligen från de datainflöden som markeras med vita pilar i bilden.

Exempel på olika sätt att ta emot administrativa data

I avsnitt 3.1 beskrivs hur man inom SCB:s Inkomst- och taxeringsregister tar emot administrativa data från olika myndigheter. Dessa leveranser är *årsvisa* och kommer från ett *begränsat antal centrala myndigheter* och leveranserna sker på *magnetband*. För andra SCB-register kan förhållandena vara helt annorlunda och sättet att ta emot data kan struktureras på följande sätt:

1. Leveranserna kan vara löpande och avser då förändringar i beståndet av objekt, eller avse månadsvisa eller årsvisa uppgifter. Det är intressant att så stor del av registersystemet får data månadsvis eller oftare – registerstatistik kan alltså ha hög aktualitet.

Bild 3.4 Registersystemets indata efter den administrativa källans periodicitet

Får data löpande	Får data månadsvis	Får data årsvis
Befolkningsregistret Dödsorsaksregistret Fordonsregistret Företagsregistret Skolregistret Fastighetsprisregistret <i>Det planerade Fastighetsregistret kommer att få löpande aviseringar</i>	Jobbregistrats AMS-del Universitets- och högskoleregistret Statsanställda, löner Skattedeklarationer Mervärdesskatteregistret Utrikeshandelsregistret Geografidatabasen Moderniseringsregistret Nybyggnadsregistret	Utbildningsregistret Inkomst- och taxeringsregistret Svenska för invandrare (3 ggr/år) Komvux (2 ggr/år) Gymnasieskolan Årskurs 9-registret Lärarregistret Kontrolluppgifter Privatanställda, löner Landstingspersonal, löner Kyrkokommunal personal, löner Primärkommunal personal, löner Standardiserade räkenskapsutdrag Lantbrukets företagsregister Registret för företagsstatistik Kommunregistret Fastighetstaxeringsregistret <i>(Småhus, Hyreshus, Industri- och Lantbruksfastigheter ingår här)</i>

2. SCB kan få data på tre olika sätt:
 - a) Som *rikstäckande register* från ett begränsat antal centrala myndigheter, t.ex. SKV, RFV, PV, KPA, CSN och SPV lämnar data till Inkomst- och taxeringsregistret.
 - b) Som data från ett *stort antal lokala register*. När det inte finns ett rikstäckande register kan data hämtas från de lokala register som finns i olika regioner. T.ex. så lämnar varje högskola data från sitt lokala studie-dokumentationssystem till SCB:s Universitets- och högskoleregister.
 - c) Som uppgifter på *blanketter som är ifyllda av tjänsteman* med hjälp av data från organisationens administrativa system. T.ex. så fyller läkare i dödsorsaksintyg som skickas till SCB:s Dödsorsaksregister. Företag, kommuner och skolor fyller i enkäter från SCB för olika totalundersökningar. Arbetet med att fylla i blanketterna kan ibland vara omfattande, det går t.ex. åt ca 40 timmar för en uppgiftslämnare att räkna om uppgifterna i kyrkokommunernas bokföring till nationalräkenskapernas ändamålsgruppering.
3. Leveransernas form kan vara någon form av datamedium eller pappersblankett. I vissa fall kan telefonkontakt förekomma. De SCB-register som tar emot data från många lokala myndigheter kan få data på flera slag av datamedier och samtidigt även på pappersblankett. T.ex. får SCB-registret Svenska för invandrare data på blankett, diskett med data på textfil eller Excel-fil. Filerna kan även skickas via e-post.

Källor:
 Pettersson (1993)
 Voss (1994)
 Lindström (1999)

3.3 Helheten skapas genom samordning och samarbete

Det exempel med *Inkomst- och taxeringsregistret* (IoT) som diskuteras i avsnitt 3.1 visar att ett statistiskt register på SCB kan innehålla tre typer av variabler:

- variabler hämtade direkt från de administrativa källorna
- variabler hämtade från andra statistiska register inom SCB
- variabler som har härletts inom det egna statistiska registret.

Man gör även medvetna avgränsningar av objektmängden – inom IoT-registret avgränsas populationen till personer 16 år och däröver som ingick i befolkningen den 31/12 enligt Befolkningsregistret.

Vissa register är så kallade integrationsregister som *uteslutande* innehåller variabler från andra statistiska register. Man tar inte själv emot några data från administrativa myndigheter och organisationer. Lukas-registret som beskrivs i avsnitt 1.3.2 är ett exempel på ett sådant integrationsregister.

Bild 3.5 De olika variabeltyperna inom ett statistiskt register

Variabler från:						
Administrativ källa		Andra SCB-register			Egna härledda variabler	
Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7

Bild 3.6 De olika variabeltyperna inom ett integrationsregister

Variabler från:						
Andra SCB-register					Egna härledda variabler	
Var 11	Var 12	Var 13	Var 14	Var 15	Var 16	Var 17

Detta visar på betydelsen av samordning och samarbete mellan olika register inom SCB. Inom ett visst register behövs variabler från andra register och inom andra register kan man behöva variabler från det aktuella registret. T.ex. så finns det flera register inom SCB, där man behöver olika inkomstvariabler från IoT-registret.

För att registersystemet skall fungera effektivt fordras att:

- Kopplingsvariablerna i basregistren är standardiserade och att träffprocenten är hög när basregister matchas.
- Alla övriga register skall kunna länkas till sitt basregister med de standardiserade kopplingsvariablerna.
- Objektavgränsningar och objektmängderna är avstämda mot respektive basregister.
- Variablerna i systemets olika register är samordnade. Variabelnamn och definitioner överensstämmer för samma variabel i olika register. Metadata samutnyttjas för att undvika dubbelarbete.
- Dubbelarbete undviks genom att statistiska metoder och IT-verktyg utvecklas gemensamt.
- Gemensamma sekretessregler tillämpas av alla.

Samordning och samarbete ger oss ett bättre fungerande registersystem med större möjligheter. Befintlig statistik blir mer konsistent och får högre kvalitet, samtidigt som möjligheterna till ny registerbaserad statistik ökar.

Det som är utmärkande för statistiska register är att de skall kunna användas för många syften. Ett visst register skall dels användas för de olika produkterna på den egna enheten men även kunna användas av andra enheter för undersökningar med helt andra syften. All SCB-statistik bygger på registersystemet – även urvalsundersökningarna och nationalräkenskaperna!

Exempel 1: Populationsavgränsningar

Många produkter inom SCB avser årlig personstatistik där populationen består av befolkningen den 31/12. Denna avgränsning görs inom Befolkningsregistret som sedan används av olika produkter. Detta medför att det blir konsistens mellan olika statistikprodukter med avseende på antal personer totalt och uppdelat efter ålder, kön och region. På motsvarande sätt kan fastighets-, företags- och arbetsställepopulationer definieras för någon viss tidpunkt under året.

Nationalräkenskaperna behöver en annan typ av populationsdefinition. Deras årliga statistik avseende befolkningens inkomster avser de personer som ingick i befolkningen någon gång under året. I ett flexibelt registersystem måste det vara möjligt att bilda både årsskiftes- och årgångspopulationer.

Exempel 2: Andelen träff vid matchning

När man framställer det årliga Sysselsättningsregistret matchas kontrolluppgiften för varje persons huvudsakliga arbetsgivare mot Företagsregistret. Avsikten är att hämta uppgifter om arbetsställets näringsgren och belägenhet från företagsregistret.

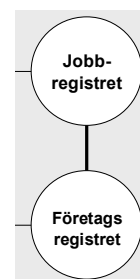
För den första årgången av Sysselsättningsregistret som avsåg 1985 var andelen träff vid denna matchning 93,6 %. För 1999 års Sysselsättningsregister var andelen träff 98,3 %. Länken mellan Jobb- och Företagsregistret har förbättrats mellan 1985-1999 tack vare av det kompletteringsarbete som har lagts ner. Arbetet består av att ta fram listor över förvärvsarbetande som inte fått träff vid matchning mellan kontrolluppgifter och Företagsregistrets arbetsställen och att göra enkäter till de berörda företagen.

Exempel 3: Objektavgränsningar

Inom företagsstatistiken är det ett centralt problem att kunna avgränsa statistiskt intressanta objekt. De administrativa materialen avser i första hand juridiska enheter (JE) medan den statistiskt intressanta verksamheten bedrivs inom en organisation där flera juridiska enheter kan ingå. Inom Företagsregistret läggs mycket arbete ner på att bilda företagsenheter (FE) och arbetsställen (AE) – två statistiskt betydelsefulla objekttyper.

Exempel 4: Samutnyttja metadata

Metadata behövs både i statistikproduktionen och när resultaten används. För att undvika dubbelarbete bör variabler namnges och dokumenteras endast en gång hos det ansvariga registret som har tagit emot eller skapat variablerna. När andra register hämtar variabler från det ansvariga registret bör metadata överföras samtidigt.



Exempel 5: Konsistens vid granskning och rättning

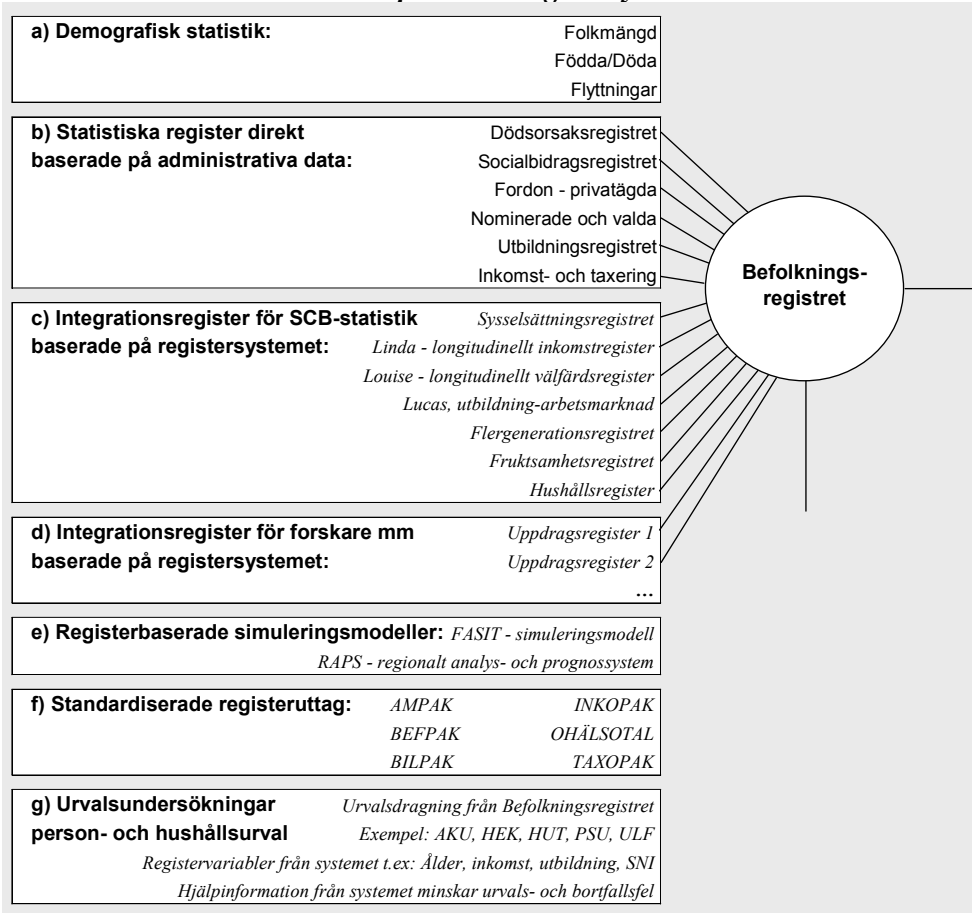
Hur skall granskning och rättning av administrativa data ske? Flera register kan bygga på samma administrativa källor och man gör var sin granskning. Om man då hittar olika fel och åtgärdar felen på olika sätt, får registersystemet som helhet sämre kvalitet och onödiga konsistensbrister. Genom samarbete bör upptäckta fel och lämpliga åtgärder förmedlas till berörda register. Av bild 3.2 framgår att kontrolluppgiftsdata levereras till IoT-registret från Kontrolluppgiftsregistret som tagit emot data från SKV. Dessa två SCB-register bör alltså samarbeta om granskning och rättning av detta administrativa material.

3.4 Skapa ytterligare register och användningar

Den del av registersystemet som avser personstatistik är den mest utvecklade och kan användas för att visa på de möjligheter som ett samverkande system kan ge.

3.4.1 Personstatistik baserad på SCB:s registersystem

De produkter som anges i ruta a) i bild 3.8 bygger helt på data från Befolkningsregistret.

Bild 3.8 Personstatistik baserad på SCB:s registersystem

AKU=Arbetskraftsundersökningen

HEK=Hushållens ekonomi

HUT=Hushållens utgifter

PSU=Partsympatiundersökningen

ULF=Undersökningen av befolkningens levnadsförhållanden

De register, som finns i ruta b), har bildats med administrativa data från olika registerlämnande myndigheter. Inkomst- och taxeringsregistret som har tagits upp i flera exempel ingår i denna grupp. Varje sådant register ensamt är grunden för olika statistikprodukter.

Genom att sambearbeta flera register inom systemet kan olika *integrationsregister* bildas. Dessa kan vara tillfälliga för specifika uppdrag eller permanenta för nya SCB-produkter. De register som anges under c) och d) i bild 3.8 är sådana integrationsregister och de har bildats utan ytterligare datainsamling. Förmågan att göra sambearbetningar beror på hur samordnat och välutvecklat SCB:s registersystem är. Datalagret skall vara bra organiserat och det skall finnas effektiva IT-verktyg för olika slag av registerbearbetningar.

Registerbaserade simuleringsmodeller (ruta e) har utvecklats för att bl.a. visa hur förändringar i skattesystemet påverkar ekonomin för personer och hushåll. De standardiserade uttagen av regionala tabeller (ruta f) har stor efterfrågan. Slutligen baseras alla urvalsundersökningar (ruta g) på registersystemet.

3.4.2 Varför är denna del av systemet så bra?

Basregistret har under en lång tid använts för att framställa förstklassig demografisk statistik. Detta har stimulerat framväxten av en god registerstatistisk tradition som i sin tur har bidragit till att basregistret har hög kvalitet. I samband med plattformbytet från stordator till den nya databastekniken 1998 fick Befolkningsregistret en effektiv IT-struktur som underlättar framställningen av olika registerversioner. Historiska uppgifter från tidigare årgångar finns lagrade så att longitudinella register kan skapas relativt enkelt. Registrets struktur beskrivs av Wilén och Johannesson (2002).

Personnumret har använts länge inom folkbokföringen och kvaliteten på denna kopplingsvariabel är hög. Personnumret är därför en bra länk mellan Befolkningsregistret och alla andra personregister och även till Aktivitetsregistret. Även länken till Fastighetsregistret (fastighetsbeteckning) har god kvalitet så att personer kan knytas till bostadsfastigheten.

De andra registren i denna del av systemet använder Befolkningsregistret som avser befolkningen den 31/12 varje år för att definiera sina populationer. Detta ger konsistens mellan olika register produkter med avseende på population.

Befolkningen den 31/12 är ett exempel på en standardiserad population

Ansaret för viktiga variabler av gemensamt intresse är decentraliserat så att man inom olika register ansvarar för "sina" variabler. Dessa variabler samutnyttjas, vilket leder till olika register är konsistenta med avseende på variabler.

Befolkningsregistret	Demografiska variabler, civilstånd, födelseland
Utbildningsregistret	Utbildning enligt SUN2000
Sysselsättningsregistret	Sysselsättningsstatus
Inkomstregistret	Olika inkomstvariabler

Dessa variabler är exempel på variabler som bör vara standardiserade

Registerstatistikens fördelar utnyttjas genom att många kvalificerade integrationsregister har bildats helt med befintliga data. Många av dessa register är longitudinella och har stor betydelse för samhällsvetenskaplig och medicinsk forskning.

Standardisering ger konsistens, detta diskuteras i kapitel 11

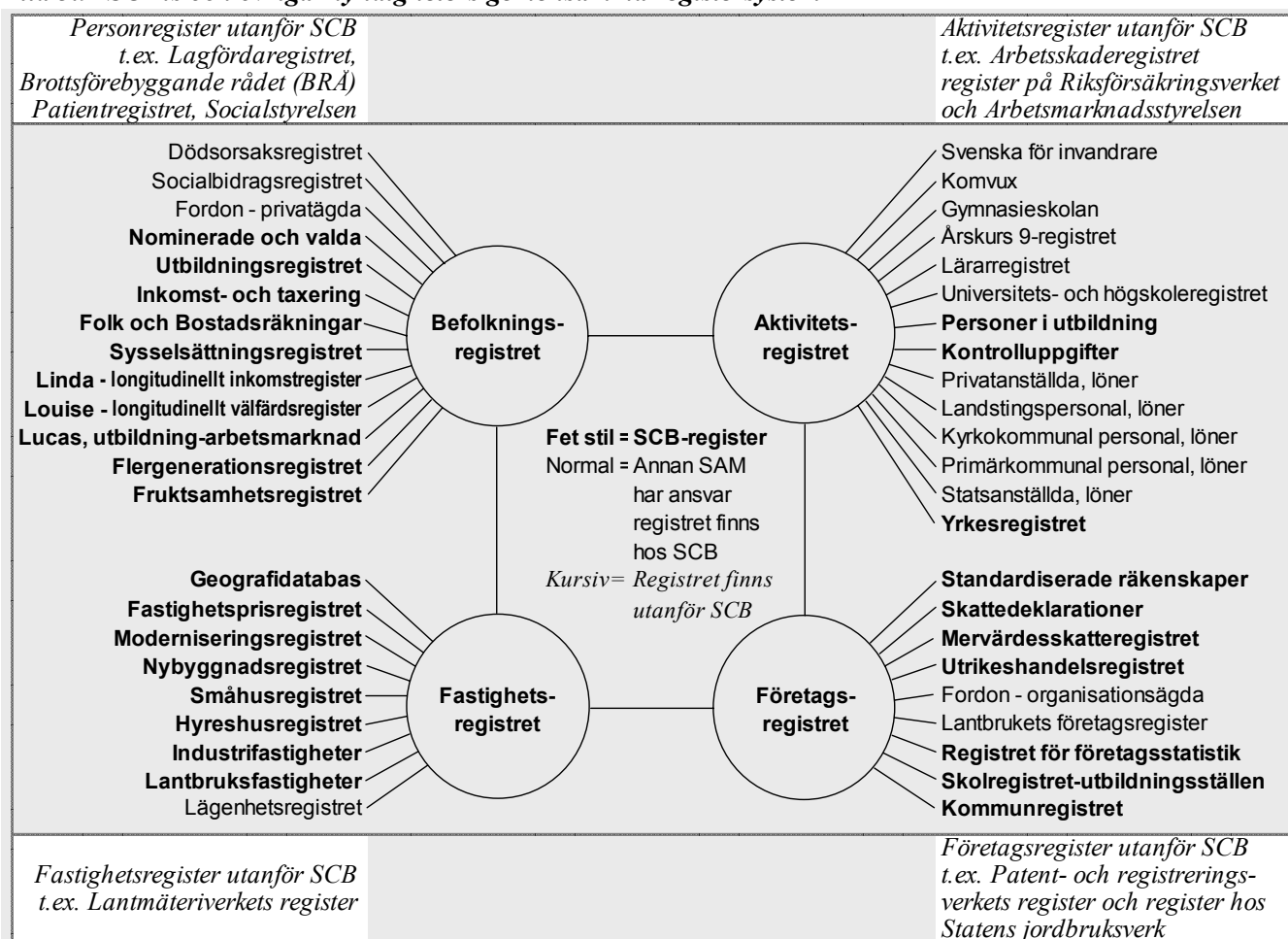
3.5 SCB:s och övriga SAM:s register samverkar

Vi har hittills enbart diskuterat SCB:s registersystem. De register som ingår i detta system kan dels vara register som SCB har ansvar för och dels register som andra *statistikansvariga myndigheter* (SAM) ansvarar för. Dessutom finns utanför SCB många register som skulle kunna kopplas till SCB:s basregister.

I kapitel 2 och 3 där vi beskriver hur SCB:s registersystem som helhet skapas, påpekar vi, att ett effektivt system förutsätter samordning och samarbete mellan avdelningar och enheter inom SCB. På motsvarande sätt bör SCB och övriga statistikansvariga myndigheter samarbeta i registerfrågor för att skapa ett bra registersystem för hela det svenska statistiska systemet.

Det är viktigt för SCB att ha tillgång till bra register från andra myndigheter. På motsvarande sätt är det viktigt för andra myndigheter att ha tillgång till SCB:s registersystem. Samarbete kring registerfrågor är därför av gemensamt intresse.

Bild 3.9 SCB:s och övriga myndigheters gemensamma registersystem



Fordonsregistret (privat- och organisationsägda fordon) används för SCB:s nationalräkenskaper och skulle kunna användas i andra sammanhang t.ex. inom välfärdsstatistiken. Genom att knyta bilägaren till utbildning, inkomst mm. från SCB:s system skulle SIKA, som är den myndighet som ansvarar för Fordonsregistret, kunna göra ny egen statistik.

SIKA
Statens institut för
kommunikationsanalys

De olika utbildnings- och löneregistren kommer att få ökad betydelse för SCB:s statistik. De kan användas för Aktivitets- och Sysselsättningsregistrets behov så att kvaliteten hos den registerbaserade sysselsättningsstatistiken förbättras. Detta är exempel på register som andra SAM än SCB har ansvaret för – ett gott samarbete med dessa myndigheter är därför viktigt för många statistikprodukter inom SCB.

För att Lantbrukets företagsregister skall kunna kopplas till Företagsregistret måste enhetsbegreppen stämma överens. Metodiken för att avgränsa jordbruksföretag blir då inte endast betydelsefull för Statens jordbruksverk utan är även en viktig fråga för SCB och hela det statistiska systemet.

Statistiken över nystartade företag är en del av den demografiska företagsstatistiken. Företagsregistret har en nyckelroll för denna typ av statistik. Även om en annan SAM-myndighet (ITPS) ansvarar för denna officiella statistik måste statistiken vara avstämd mot förändringar i SCB:s Företagsregister.

ITPS
Institutet för tillväxt-
politiska studier

Ofta behöver olika forskningsinstitut möjlighet att analysera data som finns i SCB:s registersystem. Dessa behov skulle kunna tillgodoses antingen genom att forskarna analyserar data som finns hos SCB eller att forskningsinstitutet får avidentifierade kopior av SCB-register för att bygga upp ett eget system. Detta aktualiserar tre problem:

- onödiga kostnader och risk för kvalitetsbrister på grund av ett omfattande dubbelarbete med att hålla många registerkopior aktuella och konsistenta om flera institut har egna versioner av SCB-register
- sekretess- och integritetsproblemen ökar om kopior på register finns utanför SCB
- det måste finnas ett nära samarbete mellan SCB och forskarna, så att forskarna förstår de möjligheter som SCB:s registersystem erbjuder och att SCB förstår forskarnas behov.

I flera länder lämnas inte data ut, utan analyserna får ske hos den statistiska myndigheten. I Sverige lämnas ibland data ut, ett exempel på detta är Institutet för arbetsmarknadspolitisk utvärdering (IFAU), som bland annat har delar av fyra avidentifierade SCB-register, Kontrolluppgiftsregistret, Linda, Louise och Sysselsättningsregistret.

Kapitel 4. Olika synsätt – olika språkbruk

Skillnader i synsätt och språkbruk kan vara stora hinder för registerstatistikens utveckling. Arbetet med registerstatistiken har under många år haft en undanskymd roll inom SCB. Vi tror att en viktig orsak till detta är de synsätt som varit förhärskande. För att utveckla registerstatistiken bör synsätten vidgas.

4.1 Olika synsätt inom statistikvetenskapen

SCB kan uppfattas som ett ”mångkulturellt samhälle” om man tänker på personalens yrkesmässiga inriktning och utbildningsbakgrund. Personalen består av olika kategorier med skilda synsätt och olika språkbruk. Ämnesstatistikerna är en stor kategori med mycket varierande ämnesinriktning. Gruppen metodstatistiker har olika specialiseringar: urval och estimation, mätteknik, tidsserier, registerstatistik, demografi, prisindex, analys- eller presentationsmetodik. En tredje kategori är IT-specialisterna med egna termer och yrkestraditioner.

För att personer från olika kategorier skall kunna samverka effektivt fordras att de förstår varandra. Därför måste gemensamma facktermer och förståelse för olika synsätt utvecklas. Våra yrkesmässiga och vetenskapliga synsätt, paradigm eller tankemönster, styr vårt tänkande. Dessa tankemönster kan vara mer eller mindre medvetna, men de styr vårt sätt att uppfatta den statistiska verksamheten på SCB. Våra paradigm avgör vilka problem vi ser och uppfattar som viktiga och vilka problem vi inte ser eller uppfattar som mindre viktiga.

Följande exempel illustrerar vilket paradigm som är dominerande då officiell statistik diskuteras. I *Journal of Official Statistics* (JOS, mars 2001) diskuteras kvaliteten hos den statistik som framställs av de nationella statistiska byråerna av 23 kända och kvalificerade statistiker från flera länder. Diskussionen avser olika aspekter på surveyundersökningar. Registerstatistik nämns i ett avsnitt i den inledande uppsatsen, men huvuddelen av diskussionen behandlar urvalsundersökningar. Mättekniska frågor nämns mycket kortfattat och tidsserier, demografi, prisindex, analys- eller presentationsmetodik nämns inte alls.

Att tidsserieteorin inte nämns är enligt vår mening en brist, då det flesta urvalsundersökningar, som statistikbyråerna gör, resulterar i månads- eller kvartalsstatistik. Hur kan man diskutera kvaliteten hos dessa undersökningar om man bortser från deras tidsserieanvändning? Tidsseriebruset och andra störningar av tidsseriemönstret är avgörande för statistikens kvalitet.

Registerstatistikens problem nämns endast av två av de 21 som diskuterar den inledande uppsatsen. I kapitel 7 kommer vi att utgå från Platek och Särndal (2001) när vi diskuterar registerstatistikens kvalitet.

Vår kommentar till diskussionen i JOS är att surveyteorin tycks behöva ett paradigmskifte – mätteknik och tidsserieteori bör kopplas till surveyteorin. Dessutom bör registerstatistiken erkännas som en egen specialitet. Detta är en förutsättning för att registerstatistikens problem skall uppmärksammas och att teoretiska principer och begrepp kan utvecklas inom detta område.

Platek och Särndal
(2001)

Holt (2001)
Nanopoulos (2001)

4.1.1 Urvalsundersökningar – registerbaserade undersökningar

Att planera en registerbaserad undersökning är en helt annan uppgift än att planera en urvalsundersökning. Vid en *urvalsundersökning* är första steget att bestämma populationen och vilka parametrar som skall skattas och för vilka redovisningsgrupper. Detta bestämmer karaktären på undersökningen med avseende på urvalsdessign och estimation. Definition av population och parametrar kommer först och datainsamling kommer därefter. Som regel betraktas *en* undersökning åt gången med ett begränsat antal parametrar och redovisningsgrupper. Vid arbetet med urvalsundersökningar tänker man ofta i begreppsparen *population – urval* och *parameter – estimator*.

Andelen arbetslösa i populationen är en *parameter*, kallas även *storhet* inom SCB

Vid en *registerbaserad undersökning* är ansatsen annorlunda, då data redan har samlats in och finns tillgängliga i olika register, där de inte är avpassade för en viss speciell undersökning. Med hjälp av tillgängliga register görs en selektion av objekt och variabler som är relevanta för den aktuella undersökningen. Det kan inträffa att man med befintliga register behöver bilda nya variabler – och ibland även nya objekt. I utgångsläget finns redan data och sedan avgränsas populationen. Tabellplanen bestäms därefter utan att man tänker i termer av parameterskattningar och redovisningsgrupper. Urvalsfel begränsar inte möjligheterna att välja redovisningsgrupper för den kommande analysen och resultatredovisningen. Vid arbetet med registerstatistik finns inte distinktionen *population – urval* och man brukar inte tänka i termerna *parameter – estimator*.

En *estimator* är en metod att skatta en parameter med urvalets värden

Vid utformningen av de statistiska registren och registersystemet är det önskvärt att göra dem så flexibla som möjligt och därigenom brett användbara. Därför är en mycket viktig del av registerstatistiskt metodarbete att strukturera och förbättra helheten, dvs. att finna den bästa utformningen av registersystemet. I detta ingår det långsiktiga arbetet med att övervaka och påverka tillgången till administrativa data för statistiska ändamål.

I kapitel 6 diskuterar vi skattningsmetoder för register.

Så även om mycket av den statistiska metodiken uppfattas som gemensam för urvalsundersökningar och registerstatistik – som problem med vissa felkällor (utom urvalsfelet) och arbetet med analys och presentation – så är sättet att tänka olika, då urvalsfel och designproblem är så centrala för urvalsundersökningar. Inom registerstatistiken är *systemtänkandet* fundamentalt – för att förbättra kvaliteten kan man inte se på bara ett register i taget, utan man måste betrakta systemet som en helhet och lägga ner speciell uppmärksamhet på identifieringsvariabler som används som länkar mellan olika register. Det registerstatistiska metodarbetet har olika nivåer:

- *Arbete på registernivå*: Arbete med att skapa ett statistiskt register med hjälp av olika administrativa källor och befintliga statistiska register i systemet.
- *Arbete på systemnivå*: Arbete med att koordinera flera statistiska register till ett väl fungerande system.

Att våra synsätt styr vad vi ser och uppfattar som viktigt, illustreras av följande händelse som inträffade vid ett metodseminarium. Den förste talaren berättade om mätfel vid intervjuundersökningar. Det hade visat sig att ca 40 % av intervjupersonerna i urvalet hade missuppfattat en fråga om ”huvudsaklig sysselsättning”. Näste talare berättade om registerbaserad sysselsättningsstatistik.

Flera seminariedeltagare kommenterade detta med att ”det är väl ingen vidare kvalitet på registerstatistiken”. Trots att den förste talaren hade visat att man kan få stora kvalitetsproblem med surveyansatsen, så hade seminariedeltagarna fortfarande uppfattningen att registerstatistiken är sämre.

4.1.2 Även granskning, bortfall, mätfel och presentation skiljer sig

Granskningsprocessen har olika förutsättningar för urvalsundersökningar och registerbaserade undersökningar. När man vid granskningen upptäcker orimliga värden kan man vid urvalsundersökningar återkontakta uppgiftslämnarna. Alternativt kan man redan vid en intervju upprepa frågan om man fått ett konstigt svar. Dessa möjligheter har man inte vid registerbaserade undersökningar. Man kan däremot ställa frågan till den registerlämnande myndigheten vad som är rimliga värden och hur fel kan uppkomma. Det är viktigt att man inom SCB vet vilka variabler som myndigheten har granskat och åtgärdat.

Även i andra situationer har registerbaserade undersökningar speciella granskningsproblem. När man tar emot stora register kan det finnas behov av att granska och åtgärda tekniska fel som att t.ex. att ett skannat material kan ha värden som hamnat i fel positioner. I vissa fall samlas inte data in centralt utan av olika regionala myndigheter och då kan man inom SCB undersöka om det finns strukturskillnader mellan olika myndigheter med avseende på tidsfördröjningar, svarsmönster etc.

Även *bortfallsproblemet* har olika karaktär för urvals- och registerbaserade undersökningar. Inom registerstatistiken finns inte problemen med ej anträffade eller vägrare och inte heller är det aktuellt att skicka påminnelser och bestämma när man skall avbryta uppgiftsinsamlingen. Men däremot förekommer även inom registerstatistiken partiellt bortfall och även att objekt saknas. Vid urvals- och totalundersökningar korrigeras skattningarna för att motverka bortfallets effekter. Idag förekommer inte detta inom den registerbaserade statistiken.

Mätfel finns i både urvals- och registerbaserade undersökningar. I urvalsundersökningar inriktar vi oss på att minska mätfelen genom att testa och förbättra frågeblanketterna. Inom registerstatistiken kan vi inte arbeta på samma sätt, då det är de administrativa myndigheterna som utformar blanketterna. SCB:s roll här är att påverka olika myndigheter så att blanketter och anvisningar förbättras.

Presentationen av undersökningsresultat har något olika förutsättningar. Vid urvalsundersökningar kan vi inte dela upp materialet i för små grupper och det är urvalsfelen som sätter gränserna. De tabeller som presenteras kan därför inte vara indelade efter många variabler samtidigt och får inte ha för många celler. Inom registerstatistiken förekommer däremot stora detaljerade och flervägsindelade tabeller. Sådana tabeller är ofta svårtolkade och ställer stora krav på presentationsmetoden. Även här finns det gränser för hur detaljerade tabellerna kan vara på grund av *röjanderisker* för enskilda personer och företag.

Dessa exempel visar att urvalsundersökningar och registerbaserade undersökningar ofta har olika typer av metodproblem. Den metodutveckling som gäller granskning, bortfall och mätfel har hittills endast avsett urvalsundersökningar. Detta förklaras av att gällande paradigm inte ”ser” registerstatistikens behov.

Partiellt bortfall: vissa variabelvärden saknas för objekt som finns i registret

Objektbortfall eller undertäckning: objekt saknas helt i registret

Mer om röjanderisker i kapitel 10

Mer om detta i kapitel 5-6

4.2 Olika synsätt inom en statistikbyrå

4.2.1 Mitt register – registersystemet

För många som arbetar med registerbaserade undersökningar, är det vardagliga perspektivet begränsat till arbetet med ett fåtal register. Man tar regelbundet emot administrativa data och bearbetar dessa för den egna produktens behov. Vid dessa bearbetningar använder man ett eller några få andra SCB-register för sambearbetningar. Man använder produktens register till vissa återkommande publikationer eller andra registertillämpningar. Man kommer då endast i kontakt med en liten del av registersystemet och synsättet påverkas av detta.

För dem som arbetar med att framställa SCB:s integrationsregister ser vardagen något annorlunda ut. Man tar inte emot data från någon myndighet utan arbetar helt med att sambearbeta ett stort antal SCB-register. I exemplet med Lucasregistret i avsnitt 1.3 sambearbetas 21 olika SCB-register för att skapa det longitudinella register som används för att undersöka unga personers övergång från studier till förvärvsarbete. Detta medför att man kommer i kontakt med en större del av registersystemet och man får en annan syn på systemet och möjligheterna till sambearbetningar och de svårigheter som finns.

De som arbetar med registeruppdrag arbetar under andra förutsättningar. Man har inga fasta publikationer utan olika uppdragsgivare har olika behov, vilket medför att nya registerkombinationer kan bli aktuella för de bearbetningar som måste göras. För denna kategori är systemets möjligheter och begränsningar ständigt aktuella. Man får kunskaper om de möjligheter och brister som finns.

Detta innebär att synen på registersystemet kan variera. Många får genom sitt arbete perspektivet ”mitt register är huvudsaken”, medan andra ständigt får arbeta med hela systemet. Det är viktigt att alla får uppleva att deras register ingår i ett gemensamt registersystem och att det som jag gör med ”mitt register” är betydelsefullt för helheten.

Registerstatistik och utveckling av registersystemet kräver samverkan över enhets- och avdelningsgränser. Den gamla verkskulturen med egna små registersystem avpassade för den egna verksamheten är ett hinder för ett helhetstänkande – ett samordnat registersystem för hela SCB!

4.2.2 Administrativa data skall bearbetas för statistiska syften

Kan jag som SCB:are verkligen ändra i de administrativa data som kommer från en annan myndighet? Den administrativa myndigheten samlar ju in uppgifterna och har det primära ansvaret för registret. En SCB:are som arbetar med en viss produkt eller undersökning, kanske inte kommer på tanken att dessa data behöver ändras och kompletteras för att passa till den aktuella produktens statistiska syften.

I kapitel 3 påpekas att administrativa register skall bearbetas så att objekt och variabler svarar mot statistikens behov. Detta innebär att de SCB:are som tar emot administrativa data har både frihet och skyldighet att göra sådana ändringar så att statistikens kvalitet förbättras. För att göra detta måste man vara erfaren och självständig och ha stöd av ett nätverk av registerstatistiker.

De problem vi tar upp här finns nog inom flera statistikbyråer, även om vi hämtar våra exempel från SCB



4.3 Termer – olika språkbruk

För att kunna utveckla registerstatistiken behövs ett gemensamt och rikt registerstatistiskt språk. Inom surveyteorin är ett gemensamt språk en självklarhet. Surveytermer som urvalsram, sannolikhetsurval, estimator och medelfel är välkända och har en klart definierad innebörd. Inom registerstatistiken finns samma behov av väletablerade termer. Vi tänker med ord och om våra registerstatistiska termer är diffusa blir tänkandet diffust. Har vi i stället många bra och väldefinierade begrepp stimuleras kunskapsutbyte och metodutveckling.

Registerbearbetningar har dominerats av IT-tänkande och IT-termer har använts flitigt medan statistiska termer har saknats. Ibland har dessutom termer olika innebörd, när de används som IT-term har de en betydelse, men som statistisk term har de en annan betydelse. Syftet med de registerstatistiska termer vi inför i denna bok är att de skall vara entydiga för att undvika missförstånd och ha välpreciserad innebörd så att ett gemensamt registertänkande utvecklas.

4.3.1 Begrepp, termer och tidens tand

Ett *begrepp* är det abstrakta innehållet i en språklig *term*. Det kan finnas olika termer som avser samma begrepp. Nedan ges flera exempel på att en statistisk term och en IT-term kan vara språkligt olika trots att de avser samma begrepp.

Den snabba IT-utvecklingen gör att vissa facktermer byts ut vid teknikskiften. Detta kan skapa språkförbistring och dölja att det är samma statistiska begrepp som man fortfarande avser. När SCB övergick från stordatordrift till SQL-serverdrift kom gamla termer som *flat fil* med poster och positioner att ersättas av termen *databastabell* med rader och kolumner. Det behövs registerstatistiska termer som överlever sådana teknikbyten.

Dessutom påverkas språket av vilka programpaket som används. Vi kommer att använda termen *datamatrix* i stället för olika termer som *arbetsblad*, *worksheet*, *dataset*, *databastabell* eller *flat fil*.

4.3.2 Termen statistisk undersökning

I avsnitt 1.2 anges fem olika typer av statistiska undersökningar. Tre av dessa, registerbaserade undersökningar, total- och urvalsundersökningar, är grunden för SCB:s statistikproduktion. Men termen undersökning används oftast inom SCB för att beteckna en urvalsundersökning. Ett exempel på detta är ”Att köpa statistiska undersökningar”, SCB (1998). I denna skrift som vänder sig till potentiella köpare av statistiska undersökningar nämns inte registerbaserade undersökningar trots att de utgör en stor del av SCB:s uppdragsverksamhet och att denna del dessutom är mycket konkurrenskraftig.

Varför skall även registerbaserade undersökningar kallas för statistiska undersökningar? Därför att man i båda fallen skall tänka undersökningsmetodologiskt. Även vid registerbaserade undersökningar skall man definiera undersökningsproblem, definiera population och variabler, införskaffa data (från någon myndighet eller från SCB:s registersystem), arbeta med kvalitetssäkring, analys och presentation. Genom att använda termen *undersökning* poängteras detta.

4.4 Registertermer

Register och tabell är två viktiga termer som kan ha olika innebörd, men det är önskvärt att alla använder dessa termer så att missförstånd inte uppstår. Termen *register* infördes i avsnitt 1.3 och fick där följande definition: Ett register avses vara en *fullständig* förteckning över objekten i en viss objektmängd eller population och registret skall innehålla uppgifter om objektens *identitet* så att registret kan uppdateras med nya variabelvärden för objekten. Denna definition gäller både för administrativa och statistiska register.

Alla registertermer finns samlade i bilaga 3, Registerstatistiska termer under *Register*

4.4.1 Register, datamatrix och tabell som statistiska termer

Termerna register, datamatrix och tabell diskuterar vi med hjälp av följande exempel med data från ett tänkt statistiskt register. Antag, att vi har ett register som innehåller uppgifter om alla företag i landet vid en viss tidpunkt. Antalet objekt i registret betecknas med N och registret innehåller 6 variabler.

Bild 4.1 Exempel på ett register och en datamatrix

	Variabel 1	Variabel 2	Variabel 3	Variabel 4	Variabel 5	Variabel 6
	Företagsnamn	Adress	Org.nummer	Omsättning	Anställda	Näringsgren
Objekt 1	A:sons Måleri	Adress 1	165512341234	12	9	F
Objekt 2	B:sons Bageri	Adress 2	165512341235	3	4	D
Objekt 3	C:s Salong	Adress 3	195512341236	7	10	O
•	•	•	•	•	•	•
Objekt N	Ö:s Fabrik	Adress N	165512349999	211	76	D

I detta enkla exempel överensstämmer register och datamatrix

Vanligen så gäller inte detta. Begreppen register och datamatrix jämförs på sidan 58

I en *datamatrix* har statistiska uppgifter ordnats så att matrixens kolumner är variablerna och matrixens rader är observationerna för objekten. Registret i bild 4.1 representeras av en datamatrix med N rader och 6 kolumner. Datamatrixen har markerats med en ram och vi har dessutom lagt till förklaringstext i överspalten och förspalten.

Varje statistisk undersökning syftar till att skapa en eller flera datamatrixer med mikrodata som sedan skall bearbetas för vissa statistiska syften. Termen *datamatrix* kan ses som ett statistiskt begrepp för en sådan datamängd.

Kolumnerna i matrixen innehåller *mätvärden* på *variabler*, raderna i matrixen innehåller *observationer* för objekten i registret. Observationen för objekt 2 har markerats med vitt i bilden. Observationerna är i detta fall 6-dimensionella. Om vi får nya data med t.ex. organisationsnummer och reviderad omsättning kan registret i bild 4.1 uppdateras genom att bilda en ny version där gamla värden har ersatts med nya. En datamatrix innehåller observationer för *enskilda* objekt, detta kallas *mikrodata*. Om registret *avidentifieras* får vi följande datamatrix:

Variabel, mätbar egenskap hos objekt

Mätvärde, uppmätt värde på en variabel hos ett visst objekt

Observation, alla mätvärden för ett visst objekt, observation kallas även *post*

Bild 4.2 Avidentifierad datamatrix

	Variabel 1	Variabel 2	Variabel 3
	Omsättning	Anställda	Näringsgren
Objekt 1	12	5	F
Objekt 2	3	4	D
Objekt 3	7	10	O
•	•	•	•
Objekt N	211	76	D

Denna datamatrix innehåller inga uppgifter om objektens identitet och man kan inte lägga till nya mätvärden. Den är därför *inte* ett register. Ibland används uttrycket *ett avidentifierat register*. Detta är mindre lämpligt, då det strider mot definitionen av termen register.

SCB lämnar *inte ut register!* Däremot kan avidentifierade datamatrixer lämnas ut för forskningsändamål

Ett urval från registret ger en tredje datamatrix och då urvalet inte är en fullständig förteckning över en population är denna datamatrix inte ett register i vår bemärkelse. I datamatriser för urval får varje observation i urvalet representera många observationer i populationen. I bild 4.3 representerar den första observationen 411,7 observationer i populationen. En betydelsefull del av arbetet med urvalsundersökningar är att beräkna dessa *vikter* eller *uppräkningsstal*.

Bild 4.3 En datamatrix från en urvalsundersökning

	Variabel 1	Variabel 2	Variabel 3	Variabel 4	Variabel 5	Variabel 6
	Företagsnamn	Adress	Org.nummer	Omsättning	Anställda	Vikt
Objekt 1	X AB	Adress 1	165512341111	111	57	411,7
Objekt 2	Y HB	Adress 2	165512342222	1	2	823,4
.	.	.	.	.	.	.
Objekt n	Z AB	Adress n	165512349998	56	38	411,7

Genom att bearbeta datamatriken i bild 4.1, 4.2 eller 4.3 kan man skapa tabeller av det slag som illustreras nedan. Genom att gruppera variabelvärden skapas *grupperade variabler* och *klassindelade variabler*. I den tvåvägsindelade tabellen 4A nedan finns två *indelningsvariabler*, näringsgren och antal anställda. Variabeln *näringsgren* har grupperats och variabeln *antal anställda* har klassindelats.

Tabell 4A. Exempel på en tabell skapad med data i bild 4.1, 4.2 eller 4.3

Företag efter näringsgren och antal anställda år 2000, antal företag

Näringsgren	Antal anställda					
	0	1-9	10-49	50-99	100-199	200-
A-F, varuproduktion	198 006	45 124	8 936	1 144	587	626
G-K, privata tjänster	231 910	107 296	14 830	1 424	645	568
L-Q, offentliga tjänster	65 783	24 621	3 713	434	200	548
okänd	105 823	2 515	0	0	0	0

I en *tabell* består överspalt och förspalt av variabelvärden eller grupper av variabelvärden för indelningsvariablerna. Genom att kombinera indelningsvariablerna definieras tabellens celler. Observationerna fördelas så att varje observation tillhör endast en cell i tabellen. Värdet i en tabellcell ger en sammanfattande beskrivning av cellens observationer. Sådana sammanfattande beskrivningar av grupper av observationer kallas *aggregerade data* eller *makrodata*. Med samma datamatrix kan man bilda flera tabeller med olika innehåll, t.ex. summa omsättning eller genomsnittlig omsättning per anställd.

4.4.2 Termerna databas och databastabell

Datamatriser av de typer som illustreras i bild 4.1-4.3 lagras i databaser och det finns då olika IT-termer som beskrivs i detta avsnitt. I avsnitt 2.1 beskrivs konceptuella modeller men dessa skiljer sig ofta från den fysiska implementeringen i en databas. *I denna bok diskuterar vi alltid på den konceptuella nivån, dvs. hur man tänker sig att register och datamatriser "ser ut"*. I bild 4.1 är det t.ex. fullständig överensstämmelse mellan registret, datamatriken och databastabellen, men i den konkreta IT-lösningen behöver detta inte vara fallet.

Inom SCB förekommer även termen spridningsvariabel i stället för indelningsvariabel

Bild 4.4 illustrerar olika sätt att beskriva samma databas. Den övre delen av bilden visar det traditionella sättet att beskriva en databas. Den undre delen visar ett exempel på databasens innehåll. Bilden visar strukturen på systemlösningen, dvs. hur data lagras rent fysiskt. Databasen är *normaliserad*, dvs. optimerad för konsistens och utrymmeseffektiv lagring.

Syftet med systemlösningen är att databasen skall vara en flexibel grund för många användningar

Bild 4.4 En persondatabas med tre databastabeller

A. Konceptuell databasmodell



B. Exempel på innehåll i databasen

Samtliga				Förvärvsarbetande			
PersonNr	Ålder	Kön	BoKom.	PersonNr	SNI	Lön	ArbKom.
Pnr 1	20	K	0586	Pnr 2	G	52 000	0586
Pnr 2	23	M	0586	Pnr 3	G	287 000	0580
Pnr 3	31	M	0586	Pnr 4	A	193 000	0586
Pnr 4	32	K	0586	Pnr 6	D	291 000	0586
Pnr 5	33	M	0586	Pnr 7	D	314 000	0580
Pnr 6	40	K	0586				
Pnr 7	59	K	0586				
Pnr 8	65	M	0586				
Pnr 9	71	K	0586				

Studerande			
PersonNr	UtbSlag	Hel/Deltid	StudKom.
Pnr 1	Komvux	100	0586
Pnr 2	Univ	100	0580
Pnr 5	Univ	100	0580

Med databasen ovan är det möjligt att göra statistiska bearbetningar för olika syften. I bild 4.5 nedan visar vi exempel på två olika datamatriser som kan skapas med denna databas. Dessa datamatriser kan skapas och lagras fysiskt, men de kan också bildas tillfälligt vid bearbetningen. En konceptuell datamatrix fokuserar på det specifika problem som skall lösas, dvs. den statistiska analys som skall göras.

Datamatrixens struktur kan även sparas i en så kallade vy.

I bild 4.5 visas exempel på två datamatriser för sysselsättning respektive pendling. Datamatrixerna i bilden skulle i databastermer kallas denormaliserade, vilket syns genom att det finns många celler där värde saknas.

Databaser, databastabeller och vyer diskuteras i kapitel 9

Bild 4.5 Två datamatriser för olika statistiska ändamål

A. En datamatrix för ett Sysselsättningsregister

PersonNr	Ålder	Kön	SNI	Lön
Pnr 1	20	K	-	-
Pnr 2	23	M	G	52 000
Pnr 3	31	M	G	287 000
Pnr 4	32	K	A	193 000
Pnr 5	33	M	-	-
Pnr 6	40	K	D	291 000
Pnr 7	59	K	D	314 000
Pnr 8	65	M	-	-
Pnr 9	71	K	-	-

B. En datamatrix för ett Pendlingsregister

PersonNr	BoKom.	ArbKom.	StudKom.	Pendling
Pnr 1	0586	-	0586	0
Pnr 2	0586	0586	0580	1
Pnr 3	0586	0580	-	1
Pnr 4	0586	0586	-	0
Pnr 5	0586	-	0580	1
Pnr 6	0586	0586	-	0
Pnr 7	0586	0580	-	1
Pnr 8	0586	-	-	0
Pnr 9	0586	-	-	0

Register

Består ofta av många delregister och registerversioner

Består av alla objekt som tillhör en viss definierad objekt-mängd/population

Objekten kan identifieras

Innehåller flera slag av variabler (se avsnitt 4.5.4), både statistiska variabler och andra

Lagras i en eller flera databaser

Datamatrix

En uppsättning mikrodata som skall användas för vissa statistiska bearbetningar

Innehåller huvudsakligen statistiska variabler

Kan vara avidentifierad

Kan komma från en urvals-, total- eller registerbaserad undersökning

Kan lagras i en databastabell, men kan även skapas vid beräkningstillfället

IT-termen *tabell* kan användas i stället för den fullständiga termen *databastabell* (egentligen relationstabell). Då den statistiska termen *tabell* avser ett sätt att presentera aggregerade data – *makrodata*, så kan detta språkbruk ge upphov till missförstånd i en miljö med både statistiker och IT-specialister. Databastabeller bör inte blandas samman med statistiska tabeller med makrodata.

Vad menar vi med termen *register*? Exemplet i bild 4.4 och 4.5 avser hela populationer, inte urval. Objektens identiteter finns med, detta innebär att både bild 4.4 och bild 4.5 beskriver register i vår mening. Vi kallar även databasen i bild 4.4 för ett register, som innehåller flera möjliga delregister. När registret dokumenteras bör både systemlösningen i bild 4.4 beskrivas och de statistiskt och innehållsmässigt intressanta datamatrixerna i bild 4.5.

Termer skall även förstås av allmänheten

Internt inom SCB förekommer att t.ex. termen *register* används i olika betydelser. Även termen *tabell* kan ibland avse en databastabell och ibland en statistisk tabell. Även om vi inom SCB förstår varandra, så måste de termer som används utåt, i våra publikationer och på vår webbplats, kunna förstås av allmänheten.

Allmänheten tolkar nog termen *register* så att hela populationen avses – ett medlemsregister innehåller alla medlemmar, inte ett urval. Även termen *tabell* kopplas nog till en statistisk tabell av de som besöker SCB:s webbplats. I vår kommunikation utåt bör vi därför alltid använda termerna *register* och *tabell* på ett konsekvent och för användarna lättförståeligt sätt.

4.4.3 Termer för olika slag av register

Vi har tidigare i avsnitt 3.1 förklarat att SCB tar emot *administrativa register* och bearbetar dessa för att skapa *statistiska register*. Sådana statistiska register som *direkt* baseras på åtminstone en administrativ källa kallar vi från och med detta avsnitt för *primära statistiska register*.

Basregister betecknar de fyra för registersystemet viktiga primära statistiska register, som beskrivs i avsnitt 2.2 och bör vara samlingsbenämning för enbart dessa fyra. Men det förekommer att termen används omväxlande med termen *källregister* i betydelsen register som är grunden för en viss produkt eller statistiskt register. Sådana källregister kan vara ett ursprungligt administrativt register eller ett annat SCB-register. Termen *basregister* bör endast avse någon av Befolknings-, Aktivitets-, Företags- eller Fastighetsregistret. Termen *källregister* bör användas för att beteckna register som är källan för en viss produkt.

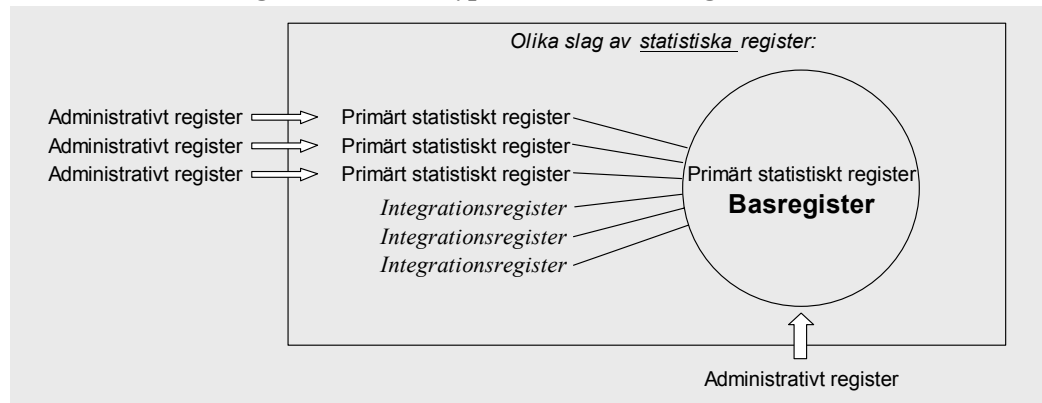
Integrationsregister är statistiska register som skapats genom att *enbart* sammanföra information som redan finns i befintliga statistiska register i systemet. Syftet med denna term är att poängtera detta viktiga sätt att utnyttja registersystemet för att göra nya undersökningar enbart med befintliga data.

Man kan sammanföra information från flera register för att skapa ett nytt register utan att det är fråga om ett integrationsregister. Så t.ex. skapas IoT-registret (Inkomst och förmögenhet) genom att integrera 6 administrativa källor och 5 SCB-register (se bild 3.2 i avsnitt 3.1). Eftersom IoT-registret är direkt baserat på åtminstone en administrativ källa är det ett primärt statistiskt register.

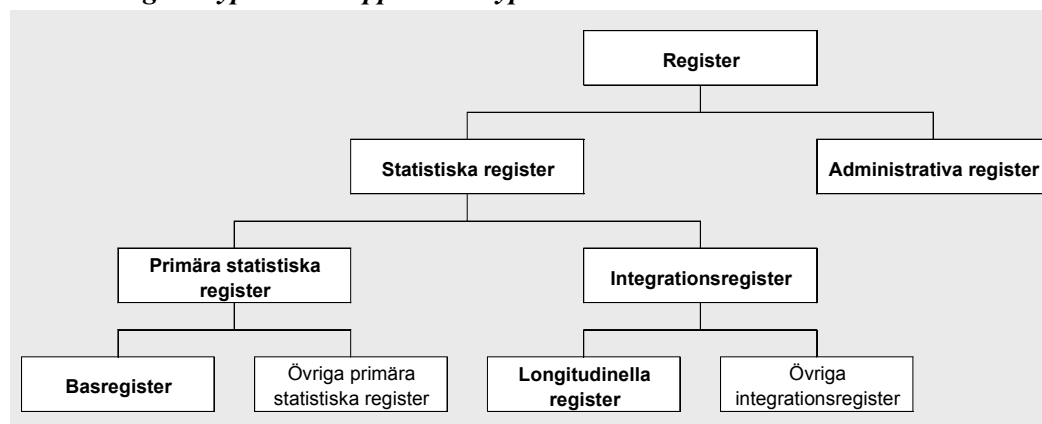
Vi skiljer på *permanent*a integrationsregister, som regelmässigt används för SCB:s återkommande produkter, och *temporära* integrationsregister som finns under en begränsad tid i samband med uppdrag. Ett *longitudinellt register* är en speciell typ av integrationsregister där man har integrerat informationen från flera årsvisa register så att man kan följa samma objekt över tiden.

Bild 4.6 Samband mellan olika registertermer

A. Administrativa register och olika typer av statistiska register



B. Olika registertyper delas upp i undertyper



I den undre bilden har vi angett de viktigaste termerna i fetstil. När vi skiljer på olika slag av statistiska register är det främst de tre termerna *basregister*, *primära statistiska register* och *integrationsregister* som används.

4.4.4 Register och tidsdimensionen

Enskilda objekt, och därmed även objektmängden, förändras över tiden. Objekt föds, flyttar, förändras eller upphör att existera. Dessa fyra olika typer av händelser kallar vi för de *demografiska händelserna*, det är dessa som förändrar objektmängden. När vi definierar register med hänsyn till tidsdimensionen måste vi skilja på följande registertyper:

- *Aktuellt bestånd* eller aktuell stock, är den registerversion som uppdaterats med all tillgänglig information för att avse nu aktiva/levande objekt.

- *Beståndet* eller stocken för *en viss tidpunkt*, t.ex. ett årsskifte, är den version av ett register som uppdaterats för att beskriva objektmängden vid tidpunkten. Denna uppdatering görs *efter* tidpunkten, när information om alla händelser fram till tidpunkten finns tillgängliga.
- *Årgångsregistret* är den registerversion som innehåller alla objekt som har funnits någon gång under ett visst år. Objekt som tillkommer eller upphör under året skall ingå med angivna händelsedatum.

Bild 4.7 Årgångsregister för år 2002

Objektidentitet	Fanns 1/1	Tillkom	Upphörde	Fanns 31/12	Övriga variabler
Idnr 1	Ja	-	20020517	Nej	...
Idnr 2	Ja	-	-	Ja	...
Idnr 3	Nej	20020315	20020925	Nej	...
Idnr 4	Nej	20020606	-	Ja	...

- *Händelseregistret* för en viss period, är det register som innehåller information om alla demografiska händelser, som inträffat under perioden. För varje slag av händelse skapas ett register.

Bild 4.8 Händelseregister för år 2002 avseende flyttningar

Objektidentitet	Adress 1/1	Flyttdatum	Ny adress
Idnr 1	Adress 11	20020517	Adress 21
Idnr 2	Adress 12	20020606	Adress 22
Idnr 3	Adress 13	20020911	Adress 23

- *Historikregistret*, är det register som innehåller information om alla demografiska händelser som någonsin inträffat för ett objekt. Ett objekt som har varit med om t.ex. tre händelser återfinns på tre rader i datamatrixen etc.

Bild 4.9 Historikregister avseende flyttningar

Objektidentitet	Från adress	Flyttdatum	Till adress
Idnr 1	född	19670517	Adress 1
Idnr 1	Adress 1	19810606	Adress 2
Idnr 1	Adress 2	20020911	Adress 3

- *Longitudinellt register* för en tidsperiod (3 år i bild 4.10), är det register som innehåller information om de demografiska händelser som inträffat under perioden och värden på de statistiska variablerna för alla delperioder (i detta fall år) inom den längre tidsperioden.

Bild 4.10 Longitudinellt register för år 2000-2002

Objektidentitet	Fanns 1/1 2000	Tillkom	Upphörde	Inkomst 2000	Inkomst 2001	Inkomst 2002
Idnr 1	Ja	-	20010517	183 450	97 600	-
Idnr 2	Ja	-	-	273 500	281 360	258 340
Idnr 3	Nej	20010315	20020925	-	193 570	204 520
Idnr 4	Nej	20020911	-	-	-	56 300

Som vi påpekat tidigare lagras register i databaser. I denna bok diskuterar vi på den konceptuella nivån, dvs. hur man tänker sig att register och datamatriser ”ser ut”. De IT-tekniska lösningarna kan ha en annan struktur än de enkla datamatriskerna ovan, som visar registrens konceptuella struktur.

4.4.5 Termen objekt

Vi har huvudsakligen använt termen objekt. Det finns en rad begrepp och termer som hör samman med denna term.

IT-termer förknippade med objekt diskuteras i Sundgren (1992)

Population, objekt, objekttyp, objektinstans, klass, enhet, element

”Populationen U består av N element/enheter/objekt.” Inom surveyteorin börjar ofta resonemangen på detta sätt. I läroböcker i statistik används oftast termerna element eller enheter, inom SCB är termen objekt vanligast. Motsvarande term vid konceptuell modellering av ett IT-system, är *objektinstans*. Med den terminologin motsvaras en population av personer av *objekttypen* person där varje enskild person i populationen är en objektinstans. Termen *klass* används ofta, framför allt vid objektorienterad programmering, som synonym till objekttyp. Inom SCB används ofta termen *objekt* som synonym till objektinstans.

För att definiera en population måste man definiera objekttypen, t.ex. vad som menas med hushåll, arbetsställe etc. *Enhetsavgränsningar* är en annan term för sådana objekttypsdefinitioner. Samtidigt måste man definiera den mängd objekt som skall ingå i populationen och i denna definition ingår alltid en rums- och en tidsangivelse, t.ex. hushållen i en viss kommun vid en viss tidpunkt.

Termen objekt inom statistik och IT

Inom statistikvetenskapen används termen enhet eller objekt endast för de objekt som en viss statistisk undersökning avser. I avsnitt 2.1 anges de olika objekttyper som förekommer i SCB:s registerbaserade undersökningar:

- person och familj/hushåll (idag familj, i framtiden hushåll)
- arbetsställe och företag (även organisationer/myndigheter kallas företag)
- fastighet, värderingsenhet och taxeringsenhet
- fordon
- aktivitet, förvärvsarbete och studier

Inom IT används termen objekt flitigt och ibland med definitioner som avviker från den statistiska termen. I en databaslösning kan rader i vissa databastabeller kallas objekt utan att vara objekt i statistisk eller konceptuell mening. Detta kan skapa missförstånd. När undersökningar dokumenteras skall endast objekt, som ingår i registerpopulationer, kallas objekt i den statistiska delen av dokumentationen. Den del av dokumentationen som beskriver IT-systemet kan dessutom innehålla andra systemrelaterade objekt.

I t.ex. ett arbetsställe-register är endast arbetsställen objekt. Däremot är inte branschkode objekt i statistisk bemärkelse, även om de lagras som rader i en av registrets databastabeller.

4.5 Termer för olika slag av variabler

Kolumnerna i en datamatrix innehåller mätvärden på olika variabler. Variabler kan ha olika karaktär och används på olika sätt i de statistiska bearbetningarna. Variabler som ingår i ett registersystem kan ha skapats på olika sätt och ha

Alla variabeltermer finns i bilaga 3 under *Variabel*

speciella funktioner i systemet. Variabler behandlas både inom statistikvetenskapen och systemvetenskapen.

4.5.1 Allmänt om variabler inom statistikvetenskapen

En variabel är en mätbar egenskap hos objekt. När vi samlar in data får vi mätvärden som är mätningar av dessa egenskaper. Det är dessa mätvärden som sammanställs i datamatriken. Vi måste tankemässigt skilja mellan egenskapen hos ett objekt och det mätvärde vi samlat in.

En statistisk variabel definieras genom vilken objekttyp som har egenskapen (t.ex. inkomst för personer och inkomst för hushåll är två olika variabler), genom den använda mätmetoden och mätskalan samt genom att ange den tidpunkt eller tidsperiod mätningen avser.

Exempel: Egenskapen ålder hos en person vid en viss tidpunkt kan mätas genom bedömning (ung, medelålders eller gammal) eller genom att fråga efter födelsedatum och sedan beräkna åldern.

Flödesvariabler anger summor för olika tidsperioder (t.ex. löneinkomst i tkr under en viss månad för en person, orderingång i milj kr under en månad för ett företag), *stockvariabler* anger tillståndet vid en viss tidpunkt (t.ex. ålder hos en person vid en viss tidpunkt, antal anställda vid årets slut för ett företag).

En variabel kan vara *kvantitativ* eller *kvalitativ*. Egenskaper som kan beskrivas med talvärden (t.ex. ålder hos en person) kan mätas med både kvantitativa (t.ex. ålder i år) och kvalitativa variabler (t.ex. ung, medelålders eller gammal).

Egenskaper som anger kategoritillhörighet (t.ex. kön) kan endast mätas med *kvalitativa* variabler. Även då man använder sifferkoder för att ange olika kategorier är variabeln kvalitativ (t.ex. 1 för man och 2 för kvinna). En kvalitativ variabel bildas genom att objekt klassificeras, dvs. indelas i ett antal grupper, sådana variabler kallas därför ibland för *klassifikationer*.

Vissa kvalitativa variabler innehåller många kategorier, som dessutom är *hierarkiskt* ordnade, t.ex. näringsgren för företag. På den högsta nivån i en hierarkisk klassifikation (en-siffernivån) gör man en grov indelning, på nästa nivå (två-siffernivån) gör man en något finare uppdelning osv.

För *kvantitativa* variabler kan man göra beräkningar med mätvärdena och i datamatriken kan man t.ex. summera mätvärden i variabelns kolumn. Därför kallas ibland kvantitativa variabler för *summerbara variabler*. För kvalitativa variabler kan man endast räkna *antal* med en viss kod, t.ex. antal 1:or respektive 2:or i datamatrikens kolumn med variabeln kön. Variabelvärdena är här alltså inte summerbara.

4.5.2 Allmänt om variabler inom systemvetenskapen

Inom SCB används inom IT-området en rad termer för att skilja på olika slag av variabler. Dessa termer brukar tyvärr inte nämnas i statistisk litteratur, men eftersom termerna är viktiga, särskilt i registerstatistiska sammanhang, så inför vi några av IT-termerna här. I detta avsnitt använder vi den fullständiga termen objektinstans i stället för objekt.

En kvalitativ variabel kan överföras till kvantitativa, (summerbara) variabler genom 0-1 kodning. T.ex. kan kön överföras till två 0-1 kodade variabler:

$x_1 = 1$ för kvinna,
0 annars

$x_2 = 1$ för man,
0 annars

Vill man beskriva andelen kvinnor används x_1 , vill man beskriva andelen män används x_2

För både x_1 och x_2 kan medelvärden och standardavvikelser beräknas

Detta avsnitt och nästa bygger på kapitel 6.7 och 9.2-9.4 i Sundgren (1992)

Mängden av de värden en variabel antar, eller kan tänkas anta, för någon objektinstans kallas för variabelns *värdemängd*. *Exempel:* Den kvantitativa variabeln *ålder på en person* har värdemängden 0 - 115 år.

Exempel: Den kvalitativa variabeln *boendelän för en person* har värdemängden Stockholms län, ..., Norrbottens län och den kvalitativa variabeln *länskod* har värdemängden 01, 03, ..., 10, 12, 13, 14, 17, ..., 25.

Envärd variabel är en variabel som antar endast *ett* värde för varje objektinstans. Som regel är variabler envärda. *Exempel:* Ålder på en person.

Flervärd variabel är en variabel som kan anta *flera* värden för vissa objektinstanser.

Exempel: Näringsgren hos ett företag, ett och samma företag kan bedriva verksamhet inom flera näringsgrenar samtidigt. Antalet näringsgrenar kan variera mellan olika företag.

Enkel variabel är en variabel som inte är sammansatt av flera variabler. Som regel är variabler enkla. *Exempel:* Ålder på en person.

Sammansatt variabel är en variabel som är sammansatt av flera variabler. *Exempel:* Telefonnummer till ett arbetsställe är sammansatt av riktnummer och abonnentnummer.

Vissa variabler är *både sammansatta och flervärda*:

Exempel: Namn på en person är sammansatt av två flervärda variabler, förnamn och efternamn. En person kan flera förnamn och även ha flera efternamn.

4.5.3 Härledda variabler

När vi bearbetar statistiska datamaterial är ett viktigt moment att skapa olika *härledda variabler*. Dessa bildas med hjälp av de variabler som finns i det aktuella registret eller genom att importera variabler från andra register. Man bör skilja på fyra olika slag av härledda variabler som beskrivs nedan.

Härledda variabler diskuteras också i avsnitt 5.8

Härledning med hjälp av variabler inom samma register, fall 1-2

1. Härledning av variabler genom *gruppering av värden och klassindelning*.

En *kvalitativ variabel grupperas*. Glesbygdverket delar t.ex. in kommunerna i fyra grupper: Skogslänens inland, övriga Norrland, Storstadsregioner, Övriga kommuner. En *kvantitativ variabel klassindelas*, t.ex. ålder och inkomst.

2. Härledning av variabler genom *beräkningar* med de variabler som finns i datamatriken. Med kvantitativa variabler gör man aritmetiska eller logiska operationer, med kvalitativa gör man logiska operationer.

Exempel: Disponibel inkomst för personer = förvärvsinkomst + kapitalinkomst + skattefria sociala ersättningar – skatt – negativa transfereringar

Exempel: Antal konsumtionsenheter i ett hushåll = $0,2 \cdot (\text{om endast en vuxen}) + 0,96 \cdot (\text{antal vuxna}) + 0,76 \cdot (\text{antal barn 11-17 år}) + 0,66 \cdot (\text{antal barn 4-10 år}) + 0,56 \cdot (\text{antal barn 0-3 år})$

logiskt villkor understruket = 1 om sant, 0 annars

Ett hushåll med en vuxen motsvarar $0,2 \cdot 1 + 0,96 \cdot 1 = 1,16$ konsumtionsenheter
Ett hushåll med två vuxna motsvarar $0,96 \cdot 2 = 1,92$ konsumtionsenheter

Härledning med hjälp av variabler i ett annat register, fall 3-4**3. Härledning av variabler genom adjungering.**

Man avser att skapa en härledd variabel i ett register med hjälp av variabler från ett annat register. Objekten i det första registret kan knytas till objekten i det andra registret genom en *ett-till-ett* relation eller en *många-till-ett* relation. Detta innebär att varje objektinstans i det första registret kan knytas till högst en objektinstans i det andra registret. Genom denna relation kan variabler i det andra registret adjungeras till objekten i det första registret.

Exempel: I ett personregister kan personerna tilldelas bostadens geografiska koordinater. Personregistret skall innehålla identiteten på den fastighet där personen är folkbokförd. Fastighetsidentiteten är länken till Fastighetsregistret. Man för över bostadsfastighetens koordinater från Fastighetsregister till det aktuella personregistret. Här länkas olika objekttyper med en *många-till-ett* relation, då varje person är folkbokförd på endast en fastighet och många personer kan vara folkbokförda på samma fastighet.

4. Härledning av variabler genom aggregering.

Man avser att skapa en härledd variabel i ett register med hjälp av variabler från ett annat register. Objekten i det första registret kan knytas till objekten i det andra registret genom en *ett-till-många* relation. Detta innebär att varje objektinstans i det första registret kan knytas till många objektinstanser i det andra registret. Man aggregerar, på ett för undersökningen relevant sätt, värdena för de *mikroobjekt* i det andra registret som är kopplade till respektive *makroobjekt* i det första registret.

Exempel: Hushållsinkomst är en aggregerad variabel som bildas genom att summera värdena på variabeln personinkomst för de personer som tillhör ett visst hushåll. Hushåll är här makroobjekt och person är mikroobjekt.

Exempel: Hur kombinerar man information från ett personregister med informationen från ett företagsregister? För arbetsställen bildas en aggregerad variabel, *andel högskoleutbildade*. Denna variabel för makroobjektet arbetsställe bildas genom att beräkna andelen högskoleutbildade bland arbetsställets samtliga anställda (mikroobjekten).

Både för adjungerade variabler och aggregerade variabler kommer mismatch att ge upphov till saknade värden i de härledda variablerna. I ett väl fungerande registersystem med låg andel mismatch finns det goda möjligheter att bilda statistiskt intressanta adjungerade variabler och aggregerade variabler.

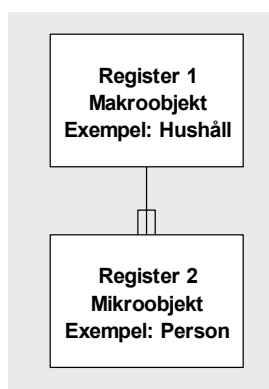
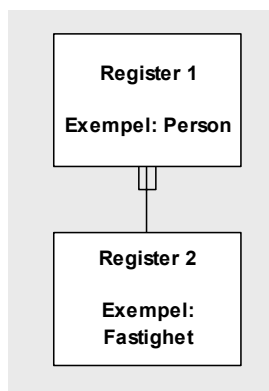
4.5.4 Variabler i registersystemet

I detta avsnitt beskriver vi registersystemets variabler dels med avseende på variablernas ursprung och dels med avseende på deras funktion i systemet.

Variabler med olika ursprung

Variablerna i statistiska register kan ha fyra olika *ursprung*. De kan ha:

1. hämtats direkt från ett administrativt register och vissa bearbetningar kan ha gjorts (formatet kan ha ändrats och variabelvärden kan ha ändrats efter granskningen), sådana variabler kallar vi för *lokala primära variabler*,



I inledningen till avsnitt 6.2 beskrivs ett femte sätt att bilda härledda variabler:

I stället för att som i bilden ovan aggregera värden för olika mikroobjekt så adjungeras alla värden från register 2 så att en flervärd variabel skapas i register 1

2. bildats som en *lokalt härledd variabel* i det aktuella registret,
3. hämtas från de primära variablerna i ett annat statistiskt register, sådana variabler kallar vi för *importerade primära variabler*,
4. hämtas från de härledda variablerna i ett annat statistiskt register, sådana variabler kallar vi för *importerade härledda variabler*.

Dessa begrepp är användbara då register dokumenteras. För ett visst register behöver endast lokala variabler dokumenteras. Dokumentationen av importerade variabler bör enkelt kunna överföras i samband med importen.

Vissa lokala variabler som är viktiga för många register kan utses till *standardiserade variabler*, som har extra höga krav på kvalitet och dokumentation.

Bild 4.11 Förenklad bild av Inkomst- och taxeringsregister för personer

Lokala primära variabler				Lokalt härledda variabler			Importerade variabler	
Pnr	Inkomst1	...	Inkomst200	Härl. Ink1	...	Disp Ink	Högsta utb nivå	BoLän
Pnr 1	10 923	.	5 223	25 766	.	197 870	5	01
Pnr 2	2 344	.	0	2 344	.	23 411	0	05
...	...	.	...	...	.	...	.	.
Pnr N	73 678	.	4 311	112 973	.	213 560	6	17

De lokala primära variablerna i datamatriisen bygger på de administrativa uppgifter som kommer från Skatteverket till Inkomst- och taxeringsregistret. Med dessa har man bildat lokalt härledda inkomstvariabler. Alla dessa lokala variabler är enkla och envärda. Vidare har man importerat två variabler – *Högsta utbildningsnivå* är en härledd variabel från Utbildningsregistret och *BoLän* är en primär variabel som har importerats från Befolkningsregistret för att få personens aktuella boställningslän enligt folkbokföringen.

Högsta utbildningsnivå i bild 4.11 är i grunden en flervärd variabel, en person kan t.ex. ha två doktorsexamina. Med en transformationsregel har denna flervärda variabel gjorts till en envärd variabel genom definitionen: senast avlagda högsta examen.

Variabler med olika funktion i systemet

Vi skiljer på sex slag av variabler med olika roller i systemet:

1. *Identifierande variabler* som identitetsnummer etc. används för att entydigt identifiera objekt. Motsvarande IT-term är *primärnyckel*. En identifierande variabel skall helst vara en helt stabil variabel, dvs. den bör ha samma värde under objektets hela livslängd. Identifierande variabler används när register som innehåller samma objekttyp matchas för att få träff mellan identiska objekt. Variabler som namn, adress och telefonnummer är besvärliga variabler vid matchning då de kan vara både sammansatta och flervärda. Därför är identitetsnummer att föredra vid registerbearbetningar.
2. *Kommunikationsvariabler* som namn, adress och telefonnummer används då SCB behöver kontakta objekt i samband med enkät eller intervju.
3. *Referensvariabler (främmande nycklar)* används för att beskriva relationer mellan olika objekt. Vid matchning av register som innehåller uppgifter om olika objekt ger referensvariabler träff mellan relaterade objekt.

4. *Tidsreferenser* är variabler som anger tidpunkt för händelser som berör objekt eller uppdateringar av register. Dessa variabler används då olika registerversioner skapas, t.ex. stocken av objekt vid en viss tidpunkt, samt för att beskriva flödet av händelser under givna tidsperioder.
5. *Tekniska variabler* eller variabler för intern *registeradministration*. Man anger källa, ger kommentarer till enstaka poster eller mätvärden. T.ex. kan källa till ett företags SNI-kod vara PRV eller SKV. Man kan även ange vilka värden som har imputerats, ange rättelsekoder och felkoder. Variabler med uppräkningsstal och vikter används i samband med beräkningar.
6. De egentliga *statistiska variablerna* används då datamatriken analyseras och beskrivs. Vissa variabler, *indelningsvariablerna*, används för att dela in observationerna i redovisningsgrupper och för varje grupp beräknas beskrivande mått för andra statistiska variabler, *aggregeringsvariablerna*. Indelningsvariabel kallas även fördelningsvariabel, spridningsvariabel eller gammavariabel.
En viss statistisk variabel kan ibland vara indelningsvariabel och i andra sammanhang vara aggregeringsvariabel. Om t.ex. medellön beräknas för olika kön är kön indelningsvariabel och lön aggregeringsvariabel. Om andel kvinnor beräknas för olika yrken är yrke indelningsvariabel och kön aggregeringsvariabel. En viktig uppgift för basregistren är att innehålla indelningsvariabler som är viktiga för många användningar.

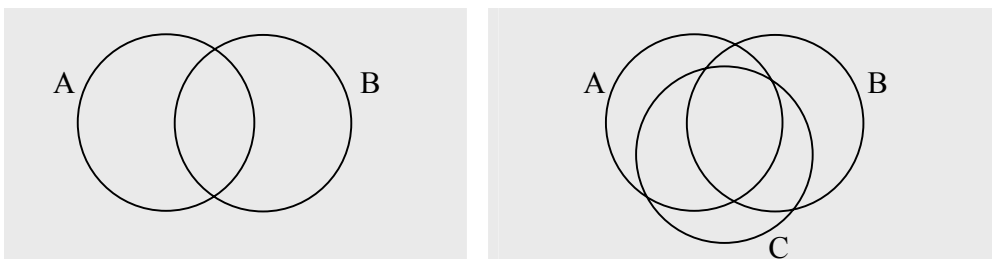
Variabler kan behöva dokumenteras på olika sätt, beroende på vilken av dessa sex olika funktioner de har i ett register.

Variabler som används vid matchningar

Man bör skilja på två olika slag av matchningar:

- Matchningens syfte är att hitta *identiska objektinstanser* i olika register eller databastabeller. Vid matchningen används som kopplingsvariabler en eller flera identifierande variabler som finns i de aktuella registren.
Exempel: Två personregister matchas, kopplingsvariabeln personnummer finns i båda.
- Matchningens syfte är att hitta de objektinstanser som har en *viss typ av relation* till varandra. Dessa objektinstanser kan finnas inom samma databastabell eller i olika register som avser olika objekttyper. Vid matchningen används som kopplingsvariabler en referensvariabel i det ena registret och motsvarande identifierande variabel i det andra registret.
Exempel: Ett personregister innehåller dels den identifierande variabeln personnummer på respektive person men även en referensvariabel som anger personnumret på personens maka/make. Två kopior av detta register matchas med personnummer som primärnyckel i den ena kopian och referensvariabeln som främmande nyckel i den andra kopian.
Exempel: Ett personregister med personnummer som primärnyckel kan matchas mot ett aktivitetsregister där förvärvsaktiviteter identifieras av tre variabler: Personnummer, arbetsställenummer och organisationsnummer. Personnummer i aktivitetsregistret är främmande nyckel vid matchningen.

När två register A och B matchas skapas tre nya objektmängder: maka objekt AB, omaka från A och omaka från B. Samtliga dessa tre objektmängder skall sparas och dokumenteras. När tre register A, B och C matchas skapas sju nya objektmängder som bör dokumenteras: maka objekt ABC, maka i endast två register (AB, AC eller BC) och helt omaka från A, B och C.



Matchningar görs med hjälp av *kopplingsvariabler* som är en sammanfattande benämning på identifierande variabler och referensvariabler. En *länk* mellan två register består av en eller flera kopplingsvariabler. Den statistiska termen *länk* motsvaras av IT-termen *nyckel*. En identifierande variabel kan vara *primär nyckel* och en referensvariabel kan vara en *främmande nyckel*.

Registertyper och variabeltyper

Som en sammanfattning av diskussionen i kapitel 3 och 4 kan de olika typerna av register i registersystemet jämföras. Det finns skillnader mellan register med avseende på vilka typer av variabler som är betydelsefulla. Dessutom skiljer sig olika typer av register när det gäller inriktning av arbetet och registrets roll.

Bild 4.12 Register- och variabeltyper och registrets huvudsakliga roll i systemet

Typ av register	Vanliga variabeltyper	Inriktning och ansvar
Basregister	<i>Lokala primära variabler:</i> Identifierande variabler Kommunikationsvariabler Referensvariabler Tidsreferenser	Ta emot administrativa data Skapa objektmängder Göra objektavgränsningar Skapa vissa grundläggande indelningsvariabler Göra demografisk statistik
Primära statistiska register	<i>Lokala primära variabler:</i> Identifierande variabler Statistiska variabler	Ta emot administrativa data Skapa de egentliga statistiska variablerna
Integrationsregister	<i>Importerade variabler:</i> Identifierande variabler Statistiska variabler Lokalt härledda variabler, adjungerade och aggregerade variabler	Skapa ny information utan datainsamling Sammanställa information från olika områden Sammanställa information från olika tidsperioder longitudinellt

Kapitel 5. Hur skapas ett statistiskt register?

Metodik betyder enligt ordboken metodlära

Behovet av metodik är lika stort för registerstatistiken som för surveystatistiken. Inom många registerenheter vid SCB görs sedan länge kvalificerade bearbetningar och man har byggt upp stor kompetens när det gäller hur registerstatistik skall produceras. Man använder visserligen metoder, men har ingen generell formulerad metodik att stödja sig på. I detta kapitel formulerar vi vissa generella begrepp och principer som kan tjäna som en början till en allmän registerstatistisk metodik.

Vi skiljer på register och registerbaserade undersökningar

Surveymetodiken ger svar på frågan *"hur gör man en urvalsundersökning respektive en totalundersökning?"* På liknande sätt skall den registerstatistiska metodiken ge svar på frågan *"hur gör man en registerbaserad undersökning?"* Men eftersom ett visst statistiskt register inte endast är avsett för en viss undersökning, utan skall ensamt eller i kombination med andra register kunna användas till flera olika undersökningar, måste man tänka mer flexibelt. Den registerstatistiska metodiken ska därför även ge svar på frågan *"hur bygger man upp ett statistiskt register, som kan användas för vissa givna undersökningar, men som även bidrar till möjligheterna att göra andra undersökningar med registersystemet?"* Detta är frågor som både skall diskuteras när man överväger ett nytt register, men även när man vidareutvecklar befintliga register.

5.1 Hur strukturera registerbaserade undersökningar?

För att inte blanda samman olika begrepp är det viktigt att tydligt skilja på de tre typerna av undersökningar som förekommer inom SCB. Vi jämför här likheter och skillnader i förutsättningar för urvals-, total- och registerbaserade undersökningar. De senare som diskuteras i detta kapitel har skuggats i bild 5.1.

Bild 5.1 Likheter och skillnader mellan olika typer av undersökningar

Urvalsundersökning	Totalundersökning	Registerbaserad undersökning
Ingår inte i registersystemet	Ingår i registersystemet – kan användas för andra registerbaserade undersökningar	
Använder registersystemet för att avgränsa populationer och som variabelkälla		
Viktigt med urvalsdesign, estimation, osäkerhetsmått	Viktigt med systemtänkande och samordning med andra registerbaserade undersökningar	
Egen datainsamling – gör egna frågeformulär	Använder andras administrativa register	
Granskning – kan kontakta uppgiftslämnare	Granskning - kan kontakta registerlämnande myndighet	
Bortfall – påminnelser, när avbryta datainsamlingen?	Bortfall/undertäckning kan synas som mismatch	
Kvalitetsbrister - urvalsfel, mätfel	Kvalitetsbrister - mätfel	Kvalitetsbrister – relevansfel, bristande jämförbarhet
Små tabeller – kan inte dela upp urvalet i för små grupper	Presentation – stora flervägsindelade tabeller	

Urvalsundersökningar görs idag av objekt inom alla basregister. Personer, företag/arbetsställen och fastigheter kan väljas direkt medan hushåll och aktiviteter väljs genom tvåstegsurval:

- först väljs person, sedan undersöks hela den valda personens hushåll
- först väljs företag, sedan undersöks alla anställningar inom det valda företaget med avseende på t.ex. lön eller yrke. Anställning är ett exempel på en typ av aktivitet.

Totalundersökningar görs idag endast av objekt inom Företagsregistret. Kommuner, skolor respektive företag totalundersöks med postenkäter eller elektroniska frågeformulär. Undersökningsmetoden medför en tung uppgiftslämnarbörd och en förutsättning är att uppgiftslämnandet är obligatoriskt. Inom företagsstatistiken förekommer *kombinationsundersökningar* där register används för små företag, mellanstora företag undersöks genom urval och de stora företagen totalundersöks. Syftet med detta är att minska kostnader och uppgiftslämnarbörd.

Totalundersökning är en undersökning med egen datainsamling

Registerbaserade undersökningar görs av objekt inom alla basregister. Det är metodiken för denna typ av undersökningar som beskrivs i detta kapitel. De skuggade delarna i bild 5.1 är viktiga förutsättningar för de registerstatistiska metoder som diskuteras.

5.1.1 Metodarbete på olika nivåer

Registerstatistiskt metodarbete bedrivs på tre nivåer:

- *Arbetet på systemnivå.* Arbetet innebär att strukturera och samordna ett stort antal register till ett fungerande system av statistiska register. Systemet som helhet skall utvecklas för att möjliggöra ny och förbättrad statistik. Detta diskuteras i de tidigare kapitlen och även i kapitel 11.
- *Arbetet med att skapa ett statistiskt register inom det befintliga systemet.* Arbetet innebär ansvar för de metoder som används när ett statistiskt register skapas. Hur använder man administrativa data för att bygga upp ett primärt statistiskt register? Hur bygger man upp ett basregister? Hur använder man registersystemet för att bygga upp ett integrationsregister? Arbetet syftar till att skapa den eller de datamatriser som registret skall bestå av. Denna typ av frågor diskuteras i detta kapitel.
- *Arbetet med en undersökning baserad på ett befintligt statistiskt register.* Arbetet syftar till att med utgångspunkt från undersökningsproblemen analysera registrets datamatriser genom olika beräkningar, aggregeringar och skattningar. Hur skall beräkningarna göras för att ta hänsyn till olika metodfrågor som täckningsproblem, saknade värden eller tidsseriebrott? Hur bearbetas register med flervärda variabler? Som resultat av registerbearbetningarna får man ett antal statistiska tabeller för en viss undersökning. Denna typ av frågor behandlas i kapitel 6.

Administrativa data skall bearbetas för att svara mot statistiska behov. Med hjälp av alla tillgängliga källor som har anknytning till det statistiska register som skall skapas, avgränsas registrets objektmängd, objekt och variabelinnehåll

på bästa sätt. Vilka registerbearbetningar skall göras? Hur skall de göras? I detta kapitel beskriver vi dessa bearbetningar från ett statistiskt perspektiv och i kapitel 9 beskrivs bearbetningarna från ett datatekniskt perspektiv.

5.1.2 Arbetsgången då ett statistiskt register skapas

Hur bygger man upp ett statistiskt register som dels skall användas för vissa givna undersökningar och dels bidrar till möjligheterna att med registersystemet göra andra undersökningar? På liknande sätt som arbetet med en urvalsundersökning brukar delas upp i olika steg bör även arbetet med att skapa ett statistiskt register struktureras i olika faser:

1. Problemformuleringsfasen: Vilka är de statistiska behov och önskemål som skall uppfyllas med registret? Detta diskuteras i avsnitt 5.2.
2. Inventeringsfasen: Vilka källor finns när man skall skapa det nya registret? Vilka administrativa källor finns och vilka befintliga statistiska register i systemet kan användas? Detta diskuteras i avsnitt 5.3.
3. Planeringsfasen kan delas in i tre delar:
 - a. Hur skall man avgränsa registrets objektmängd? Detta diskuteras i 5.4.
 - b. Hur skall man avgränsa registrets objekt? Detta diskuteras i avsnitt 5.5.
 - c. Hur bestämmer man registrets variabelinnehåll? Detta diskuteras i 5.6.
4. Leverantörskontakter och mottagande av administrativa data: Mottagningskontroll och granskning av administrativa data diskuteras i avsnitt 5.7.
5. Bearbetningsfasen kan delas in i tre delar:
 - a. Hur bearbetas befintliga källor så att det nya registret innehåller den önskade objektmängden? Administrativa data kontrolleras, granskas och åtgärdas så att objektmängden blir den önskade. Olika källor matchas och objekt selekteras. Tidsreferenserna bearbetas för att skapa objektmängden för den avsedda tidpunkten eller tidsperioden. Detta diskuteras i avsnitt 5.4.
 - b. Vilka bearbetningar skall göras för att kontrollera och ändra objektavgränsningarna? Administrativa data kontrolleras, granskas och åtgärdas så att objektavgränsningarna blir de önskade. Härledda objekt bildas i det nya registret. Detta diskuteras i avsnitt 5.5.
 - c. Vilka bearbetningar skall göras för att skapa de avsedda variablerna? Variablerna i de administrativa källorna kontrolleras, granskas och åtgärdas. Åtgärder vidtas om variabelvärden saknas. Olika källor matchas och variabler selekteras och importeras till det nya registret. Härledda variabler bildas i det nya registret. Detta diskuteras i 5.6 - 5.8.

Vilka bearbetningar skall göras? Det register som skall skapas är antingen ett primärt statistiskt register, ett basregister eller ett integrationsregister. Kraven på bearbetningarna kan vara olika för dessa tre olika slag av register.

Arbetet med en registerbaserad undersökning bör struktureras så att man är medveten om tre processer som pågår samtidigt under arbetet med undersökningen. Den första processen består av arbetet med att skapa det statistiska

registret och beskrivs i punkterna 1– 5 ovan. Parallellt med detta arbete så skall samtidigt registret och bearbetningarna *kvalitetssäkras* och *dokumenteras*.

Kvalitetssäkring

Arbetet med att skapa statistiska register skall kvalitetssäkras. Registrets kvalitet bör bedömas och även beskrivas med olika kvalitetsindikatorer. Även dokumentationsarbetet är en viktig del av kvalitetssäkringen. Felaktig och okritisk användning av administrativa data kan förhindras med metadata som ger information om jämförbarhetsproblem. Dessa frågor tas upp i kapitel 7.

Dokumentation

Eftersom ett statistiskt register skall kunna användas av flera användare som utnyttjar registersystemet, måste alla register vara dokumenterade på sådant sätt att alla kan utnyttja dokumentationen. Metadata har en mycket betydelsefull roll vid arbetet med registerstatistik. När man sambearbetar olika register måste man känna till definitioner och jämförbarhetsproblem. Det är viktigt att även bearbetningsmetoderna dokumenteras så att metodutveckling och erfarenhetsutbyte underlättas. I kapitel 8 diskuteras metadata och dokumentation.

Bild 5.2 De tre parallella arbetsprocesserna med en registerbaserad undersökning

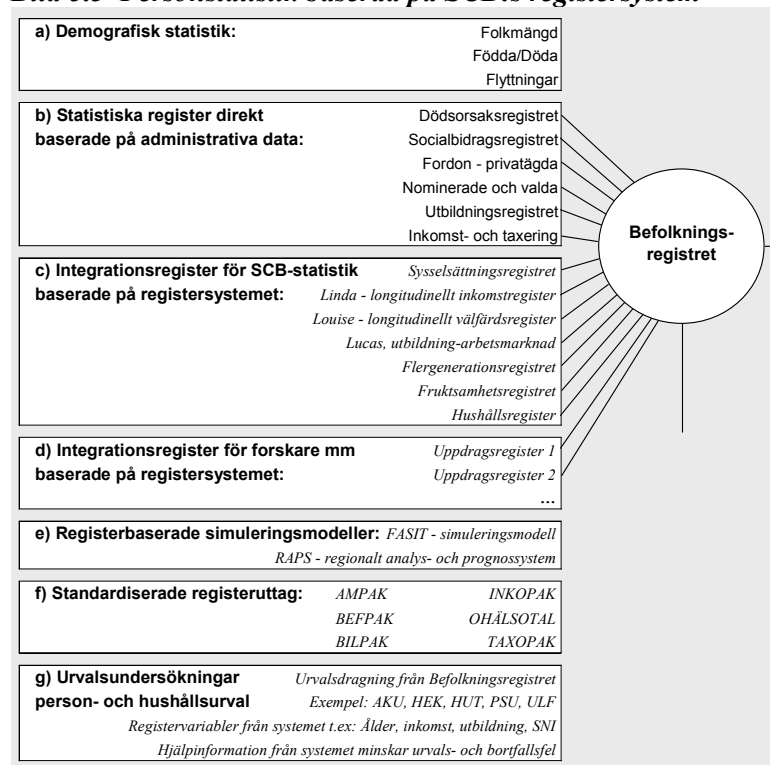
Skapa och använd ett register	Kvalitetssäkring (kap 7)	Dokumentation (kap 8)
Att skapa ett register (kap 5) 1. Problemformulering 2. Inventering 3. Planering 4. Mottagande av data 5. Bearbetningar: - Skapa population - Skapa objekt - Skapa variabler Beräkningar och tabeller (kap 6) Val av skattningsmetoder för att göra beräkningar och ta fram tabeller	Leverantörskontakter Mottagningskontroll Bortfall, omfattning och orsaker Mismatch, omfattning och orsaker Evalvera kvaliteten hos objekt och variabler Enkäter för registerhållning Inkonsistenser mellan olika källor undersöks Dokumentation	Samtidigt som mikrodata hämtas från olika källregister skall man samtidigt hämta metadata för källorna: Registerpopulation för varje källa Definitioner av importerade variabler Vilka granskningar och bearbetningar har gjorts i olika källor? Vad vet man om kvaliteten hos de olika källorna? Därefter skall det färdiga registret dokumenteras

5.2 Problemformuleringsfasen

Vilka undersökningar skall göras, vilka frågeställningar skall besvaras? Vi diskuterar detta med utgångspunkt från bilden på nästa sida (bild 3.8 från kapitel 3), som beskriver befintliga register inom systemets befolkningsdel. Inom systemets övriga delar och när man diskuterar helt nya register skall man resonera på samma sätt.

Vi har hittills betraktat bilden nedan som en del av ett registersystem. Men för varje register finns speciella frågeställningar som motiverar registret. Bilden ger då även en struktur på ett undersökningsområde (personstatistik) och dess delar.

Bild 5.3 Personstatistik baserad på SCB:s registersystem



a) Demografisk statistik

Basregistret ger underlag för demografisk statistik, vars syfte är att beskriva populationens struktur och hur strukturen förändras över tiden. Dessa demografiska frågeställningar är en egen klass av undersökningsproblem med väletablerad metodik.

b) Statistiska register direkt baserade på administrativa data

Ett antal viktiga administrativa källor har sedan länge använts som underlag för statistik. Utbildning och inkomst är områden där de statistiska registren har byggts upp successivt i samspel med olika forskare och andra användare.

c) Integrationsregister för SCB-statistik

Kreativa personer inom SCB med goda kontakter med olika användare ser nya möjligheter att använda redan befintliga registerdata. När man lyckas kombinera nya behov med systemets möjligheter skapas ett nytt integrationsregister.

d) Integrationsregister för forskare

Forskare kan få tillgång till SCB:s befintliga register för egna analyser. SCB kan även skapa ett särskilt register för forskarens behov genom att kombinera olika SCB-register. Alla samkörningar sker enligt sekretesslagen och endast avidentifierade data lämnas ut efter prövning. Information om individregister finns i SCB-data för forskare, SCB (2001b).

e) Registerbaserade simuleringsmodeller

Simuleringsmodeller som baseras på data från SCB-register kan visa hur tänkta förändringar kan påverka t.ex. inkomstfördelningen eller en regions utveckling. Förutom FASIT (simuleringsmodell för hushållens inkomster) och RAPS finns även en simuleringsmodell baserad på företagsdata (FRIDA).

f) Standardiserade registeruttag

Med olika tabellpaket kan regional statistik tas fram för valfri regional indelning. Även *Boendeprofiler* och *Marknadsprofiler* är exempel på registerbaserade produkter med regionala tabeller som kan användas av företagskunder för olika marknadsanalyser.

Genomgående för problemformuleringsfasen är att SCB:are bygger upp kunskaper om olika användares behov. Dessa SCB:are måste samtidigt vara ämnesmässigt engagerade, känna registersystemets möjligheter och veta vilka registerstatistiska metoder som skulle kunna komma till användning.

5.3 Inventeringsfasen – att inventera olika källor

Efter att ha strukturerat undersökningsproblemet bör man inventera vilka olika källor som skulle kunna användas för att skapa det nya registret. Det är tre slag av källor (källregister) som kan vara av intresse:

- Befintliga statistiska register i SCB:s registersystem. Information om vad dessa register innehåller skall framgå av befintlig registerdokumentation.
- Administrativa källor som man inom SCB känner till, men som för närvarande inte används i registersystemet. Det kan t.ex. finnas delar av leveranser som inte används av SCB idag eller delar av de administrativa registren som SCB avstår från att beställa från de administrativa myndigheterna. En tredje möjlighet är att göra en så kallad *integrerad datainsamling*, dvs. en administrativ myndighet lägger till en fråga för SCB:s behov på sina blanketter. Information om dessa möjliga källor finns hos de enheter inom SCB som har kontakt med respektive myndighet.
- Helt nya administrativa register som man inom SCB inte känner till idag. Det kan finnas administrativa källor hos företag eller myndigheter som fortfarande är outnyttjade för statistiska ändamål och som skulle kunna användas av SCB. Här är det fråga om helt nya källor varför man måste ta kontakter utanför SCB.

Observera, att alla tillgängliga källor med anknytning till problemområdet skall undersökas när det nya registret skapas. Detta är en viktig princip, då varje ny källa kanske kan användas för att förbättra det nya registrets täckning med avseende på objekt och variabler. En ny källa kan även innebära att inkonsistenser upptäcks vilket kan bidra till att höja kvaliteten.

Även en källa som ensam har kvalitetsbrister kan användas vid bearbetningarna. Källan ensam skulle kanske ge registerstatistik med låg kvalitet, men *som en del i registersystemet* kan den ändå bidra till att systemets täckning och kvalitet förbättras.

5.4 Att avgränsa registrets objektmängd

I varje statistisk undersökning skall man definiera undersökningens population. Här diskuterar vi avgränsningen eller definitionen av objektmängden och i nästa avsnitt behandlas avgränsningen av objekt.

Då ett register skall kunna användas för flera undersökningar bör man med registret kunna definiera olika populationer. För att möta dessa krav skall man kunna göra olika objektsselektioner från registret så att olika delmängder passar för olika undersökningar. En annan möjlighet är att göra olika versioner av registret för olika tillämpningar.

Vi vill reservera begreppet *population* till att avse en objektmängd som hör till en viss undersökning. När vi beskriver ett register utan att avse en viss undersökning använder vi begreppet *registrets objektmängd*. Men varje statistiskt register skapas för en eller flera huvudsakliga användningar eller undersökningar. Det är därför vanligt att registrets objektmängd överensstämmer med

Population är en objektmängd för en viss undersökning

huvudundersökningens population. När man skapar ett nytt register för en viss undersökning skall det nya registrets population avgränsas. Varje källregister har sin objektmängd som helt eller delvis kommer att ingå i det nya registret.

5.4.1 Att definiera en population

Problemformuleringsfasen i en undersökning diskuteras i Jeppesen (1973)

Varje undersökning utgår från en frågeställning som formuleras i teoretiska eller allmänna termer. De teoretiska begreppen i frågeställningen måste sedan *operationaliseras*, dvs. översättas till mätbara begrepp. När man definierar dessa mätbara begrepp avgör man *vad* som skall undersökas.

Populationsdefinitionen skall entydigt ange vilka objekt som ingår i populationen. Objekttypen skall vara klart angiven. Dessutom ingår alltid en tidsreferens och en geografisk avgränsning. I den geografiska avgränsningen ingår också att ange vilken relation som finns mellan objekten och det geografiska området.

Exempel på populationsdefinition:

”*Permanent boende* (= relation) *individer* (= objekttyp) i *Sverige* (= geografiskt område) den *31/12 2002* (= tidpunkt). Med permanent boende avses ...”

SCB:s riktlinjer för kvalitetsbegrepp och kvalitetsdeklaration (SCB, 2001a) innehåller tre populationsbegrepp:

- *Intressepopulationen* avser populationen i den teoretiska frågeställningen.
- *Målpopulationen* avser den operationaliserade populationen, den teoretiska intressepopulationen har översatts till en konkret och undersökningsbar population, dvs. den population man har som *mål* att undersöka.
- *Rampopulationen* avser den objektmängd som *ramen* och *ramförfarandet* faktiskt leder till. Dessa begrepp hör hemma inom teorin för urvals- och totalundersökningar med egen datainsamling.

Det finns skillnader mellan ram- och registerpopulationer.

En rampopulation avgränsas *före* datainsamlingen, en registerpopulation skapas *efter* referensperioden när alla administrativa uppgifter har kommit in.

En ram kan bestå av annat än de objekt man vill undersöka, t.ex. en karta för att undersöka hushåll.

För registerbaserade undersökningar använder vi endast de två första begreppen. Då urvalsram och ramförfarande inte förekommer vid registerbaserade undersökningar måste det tredje begreppet *rampopulation* ersättas:

- *Registerpopulationen* avser objektmängden i det register man har skapat för den aktuella undersökningen, dvs. den population man *faktiskt* undersöker.

Exempel: Undersökning av befolkningens struktur med Befolkningsregistret

Vi vill undersöka den svenska befolkningens sammansättning den 31/12 ett visst år. Intressepopulationen är *de permanent boende i Sverige den 31/12* men detta vaga begrepp måste definieras närmare. Denna definition kan göras på olika sätt. Allmänt gäller att en bra definition skall uppfylla följande krav:

- den skall vara *adekvat*, dvs. överensstämma med vad man vill undersöka
- den skall vara *funktionell*, dvs. den skall vara praktiskt användbar

När man gör dessa definitioner ställs man inför svåra avvägningar mellan vad man vill undersöka (adekvata definitioner) och vad man kan undersöka (funktionella definitioner). I detta exempel kan man bl.a. tänka sig följande definitioner av ”*de permanent boende i Sverige den 31/12 år t*”:

1. Folkbokförda i Sverige den 31/12 år t
2. Folkbokförda i Sverige den 31/12 år t enligt uppgifter som är tillgängliga den sista januari år $t+1$

De som arbetar med Befolkningsregistret brukar vänta till utgången av januari innan man skapar den registerversion som skall avse den 31/12. Man hoppas då att alla folkbokföringshändelser för året har hunnit rapporterats in. Det register som då har skapats får vara det som gäller trots att ett fåtal aviseringar som avser år t kan tillkomma. Definition 2 anger alltså hur man avgränsar registerpopulationen för denna undersökning.

Däremot så anges i kvalitetsdeklarationen för den årliga befolkningsstatistiken att undersökningen avser att beskriva en målpopulation enligt definition 1. Skillnaden mellan målpopulation och registerpopulation är då registrets *täckningsfel*. Dödsfall och flyttningar ur landet som inte har rapporterats före den sista januari ger *övertäckning*, medan födlsar och inflyttning som inte har hunnit rapporterats ger *undertäckning*.

Förutom dessa täckningsfel får man också relevansfel om den använda definitionen inte är adekvat. Skillnaden mellan intressepopulationen och målpopulationen är ett av undersökningens *relevansfel*. Det finns troligen mellan 25 000 - 50 000 personer som är folkbokförda i Sverige, men som inte bor permanent i Sverige. Man bedömer att mellan 4 % - 8 % av de utomnordiska invandrarna har lämnat Sverige utan att rapportera detta. Detta relevansfel påverkar statistiska mått som beskriver dödlighet, medelinkomst etc. för utomnordiska invandrare så att måtten blir missvisande.

Uppskattningen avser 1998. Se Qvist (1999)

I exemplet ovan ser vi att SCB:s befolkningsstatistik använder en administrativ definition, den folkbokförda befolkningen, när undersökningens målpopulation definieras. Att använda administrativa begrepp ger alltid definitioner som är funktionella. Det är förnuftig statistisk praxis att använda dessa administrativa begrepp för att definiera målpopulationen om relevansfelen blir små. Men grundregeln är att populationsdefinitionen skall svara mot den statistiska undersökningens krav. Är de administrativa begreppen inte tillräckligt relevanta måste man göra egna definitioner och göra de registerbearbetningar som krävs för att registrets objektmängd skall avspegla den definierade populationen så väl som möjligt.

5.4.2 Att definiera en population – primära statistiska register

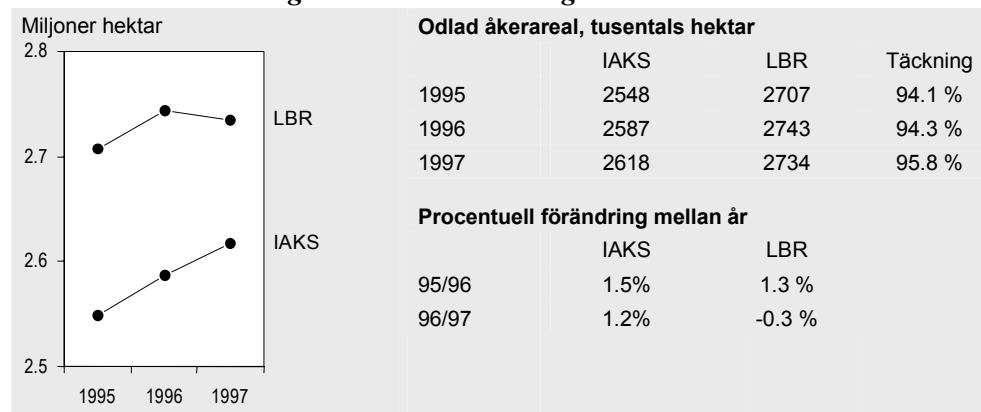
Hur skall de som arbetar med SCB:s primära statistiska register definiera målpopulationer och avgränsa registerpopulationer? De primära statistiska registren bygger helt eller delvis på administrativa register, och det finns då en risk för att de administrativa systemens objektmängder kommer att styra valet av registerpopulation på ett olämpligt sätt.

Objektmängden i de administrativa registren kanske inte helt täcker den målpopulation som är statistiskt intressant. Den administrativa objektmängden består endast av de objekt som ingår i det administrativa systemet och det kan då finnas både över- och undertäckning jämfört med den statistiskt önskvärda målpopulationen.

IACS = Integrerat
Ansöknings- och
KontrollSystem

Exempel: Den administrativa objekt mängden – duger den som målpopulation?
Lantbruksstatistiken kommer i framtiden att baseras på de ansökningar om EU-bidrag som lantbrukarna lämnar till länsstyrelserna. Dessa ansökningar registreras i IACS-systemet som används för att administrera lantbruksstödet. I bild 5.4 jämförs de arealuppgifter som fanns i ansökningarna om arealbidrag 1995-97 inom IACS-systemet med motsvarande uppgifter i Lantbrukets företagsregister, LBR, som baserats på totalundersökningar gjorda av SCB. Skillnaderna mellan IACS och LBR beror på undertäckning i IACS-registret – vissa lantbrukare ansöker inte om arealbidrag trots att de är aktiva odlare.

Bild 5.4 Undertäckning i ett administrativt register



Slutsatser: Trots att IACS-registret kanske anses ha god täckning så medför till synes små variationer i täckningen att tidsserien för odlad åkerareal blir helt missvisande – en verklig nedgång 1996-97 ser ut som en fortsatt uppgång.

Man skall inte bortse från bristande täckning i de administrativa registren utan här bör målpopulationen definieras enligt statistikens krav. Därefter skapar man ett nytt register som innehåller den avsedda registerpopulationen med hjälp av det aktuella basregistret, i detta fall genom att selektera objekt från Företagsregistret. Sedan kan detta nya register matchas mot IACS-registret och då kommer man att upptäcka både eventuell över- som undertäckning i IACS-registret. Övertäckningen är indikation på eventuella brister i Företagsregistret. Undertäckningen i IACS-registret blir bortfall i det nya registret. Detta bortfall kan åtgärdas antingen genom bortfallskorrektion eller genom att komplettera med en speciell undersökning för att samla in data för den del av målpopulationen som inte finns med i det administrativa registret.

5.4.3 Att definiera en population – integrationsregister

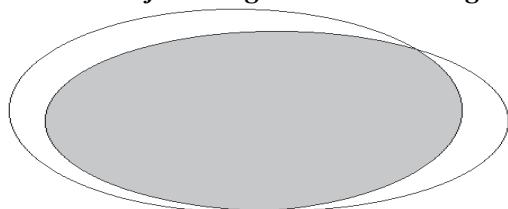
Hur skall de som arbetar med integrationsregister definiera målpopulationer och avgränsa registerpopulationer? Vi tar här upp grundprinciperna för detta.

Exempel: Registeruppdrag med sambearbetningar – val av målpopulation

I många registeruppdrag behöver flera register sambearbetas. Man gör en rad matchningar och importerar variabler från olika register. Slutresultatet blir ett integrationsregister med många variabler som är intressanta för uppdragsgivaren. Men hur har undersökningens målpopulation definierats? Det är lätt hänt

att integrationsregistrets objektmängd blir snittmängden av de matchade registrens objektmängder. Motsvarar denna mängd en adekvat målpopulation för detta uppdrag? Det är inte alls självklart att detta är fallet.

Bild 5.5 Objektmängder vid matchning av två register



Respektive ellips representerar objektmängden i två olika register.

Den skuggade ytan motsvarar de objekt som gav träff vid matchning, dvs. snittmängden.

Även vid ett sådant registeruppdrag skall man börja med att definiera målpopulationen med hänsyn till undersökningsproblemet. Sedan selekterar man en lämplig objektmängd från det aktuella basregistret. Denna objektmängd är registerpopulationen, som sedan matchas mot de register som innehåller de intressanta variablerna. För de objekt där man fått träff, importeras variabelvärdena till det nya registret. För de objekt där man inte fått träff anges partiellt bortfall, dvs. att variabelvärden saknas.

Slutsatser: Snittmängden – vilken målpopulation motsvarar den? Den kan aldrig vara målpopulation, då man inte skulle kunna ge en definition av denna. Den bör inte heller vara registerpopulation, då bortfallet på grund av mismatch skulle riskera att glömmas bort. Att vara medveten om bortfallet är viktigt och dessutom kan man göra bortfallskorrekationer då bortfallets omfattning och struktur är kända. Detta beskrivs i kapitel 6.

5.4.4 Basregistren skall användas då populationer avgränsas

Både de som arbetar med integrationsregister och de som skapar primära statistiska register med hjälp av administrativa register bör använda basregistren när objektmängderna avgränsas. Det finns två skäl till detta, dels skall basregistren vara ”bäst” på objektmängder och dels måste SCB:s registerstatistik vara inbördes konsistent. De som arbetar med primära register och integrationsregister bör utgå från respektive basregisters objektmängder och använda någon av de *standardiserade populationer* som skapats för allmän användning:

- *årsskiftesversionen* passar för årlig stockstatistik (t.ex. befolkningens förmögenhet den 31/12),
- *årgångsversionen* passar för årlig flödesstatistik (t.ex. befolkningens inkomster ett visst år),
- en *månadsversion* passar för månadsstatistik.

Generell metod:

- Definiera målpopulationen
- Selektera avsedd objektmängd från basregistret, detta ger registerpopulationen
- Matcha mot de register som innehåller intressanta variabler
- Vid träff: Importera variabelvärden till det register som skapas
- Vid mismatch: Ange partiellt bortfall

5.4.5 Kraven på ett basregister

Det ställs tre slags krav på ett basregister för att de skall kunna användas för att avgränsa registerpopulationer:

- basregistret skall innehålla tidsreferenser, dvs. alla händelser som berör registrets objekt skall vara daterade,
- basregistret skall ha god täckning,
- länkar och indelningsvariabler skall ha hög kvalitet.

Tidsreferenser

För varje objektinstans måste tidpunkten då objektet tillkom vara känd liksom tidpunkten då objektet upphör. Även tidpunkter för andra viktiga händelser, som flyttningar eller förändrad kategoritillhörighet, skall vara kända.

Med hjälp av dessa tidsreferenser kan man skapa registerpopulationer som speglar populationens tillstånd vid en viss önskad tidpunkt eller period.

Tidsreferenserna kan ha olika precision, händelser för personer är som regel kända så att en viss dag kan anges, händelser avseende företag kanske endast är kända så att månad eller endast år kan anges.

Det är viktigt att man i basregistren skiljer på statistiskt relevanta förändringar och statistiskt triviala förändringar. Byte av riktnummer eller postnummer för ett helt område skall inte uppfattas som objekt har flyttat. På samma sätt får inte ett företags byte av bolagsform uppfattas som att ett företag har upphört och ett annat företag har bildats.

Täckning, indelningsvariabler och identifierande variabler

Med hjälp av något basregister avgränsas de flesta populationer i SCB:s undersökningar. Detta ställer höga krav på att basregistrens objekt mängder har god täckning för många undersökningars behov och att länkar och variabler som används för att selektera fram delpopulationer eller dela in i redovisningsgrupper (*indelningsvariabler*) har hög aktualitet och kvalitet. Kvalitetsbrister hos dessa variabler kan ge täckningsfel i delpopulationer, t.ex. genom att uppgift om näringsgren (SNI-kod) saknas eller att SNI-koden är felaktig.

Exemplet ovan som handlar om lantbruksföretag visar vilka krav som kan ställas på ett basregister. I detta fall måste man med Företagsregistret kunna göra en selektion av lantbruksföretag så att det nya registret får en god täckning av den aktuella målpopulationen. För att lyckas med detta bör Företagsregistret dels ha en god täckning av alla aktiva företag, dels bör näringsgrenskoden ha god kvalitet så att den kan användas för att selektera den önskade registerpopulationen. Dessutom innehåller Företagsregistret identifierande variabler som person/organisationsnummer (PeOrgNr), företagsnamn, adress och telefonnummer som är viktiga när man skall matcha mot IAKS-registret.

För att få god täckning i ett basregister skall man kombinera och använda alla relevanta administrativa källor. Detta är en allmän princip för arbetet med att skapa statistiska register – *genom att kombinera många källor förbättras kvaliteten.*

Exempel: Novemberversionen av Företagsregistret anses ha god kvalitet och används som urvalsram för många undersökningar. Den goda kvaliteten behöver inte gälla småföretag och alla näringsgrenar. Matchas novemberversionen år t med IAKS-registret från sommaren år t ser man att undertäckningen är betydande. Genom att komplettera Företagsregistret med kontrolluppgifter, moms och standardiserade räkenskapsutdrag för år t blir täckningen god.

IAKS-registret innehåller lantbruksföretagens ansökningar om arealbidrag

Tabell 5A. Resultat av olika matchningar av Företags- och IAKS-registret 1995

IAKS-registret matchas mot	Kopplingsvariabler	Andel matchande i IAKS
Företagsregistret, november 1995	PeOrgNr	75,2 %
<i>Företagsregistret kompletterat med:</i>		
KU, MOMS och SRU för 1995	PeOrgNr	96,6 %

KU = Kontrolluppgifter

MOMS = Företagens momsdeklarationer

SRU = Företagens standardiserade räkenskapsutdrag

Slutsatser: En slutsats av detta exempel är att Företagsregistret hade dålig täckning av lantbruksföretag. Genom att komplettera Företagsregistret med andra källor kan täckningen förbättras avsevärt. I detta exempel finns fyra källor som skulle kunna användas för att förbättra denna del av Företagsregistret. IAKS-registret är den mest aktuella källan som blir tillgänglig cirka ett år tidigare än de övriga källorna. Dessutom vet man att IAKS-företagen är aktiva lantbruksföretag, dvs. IAKS-registret innehåller även information om SNI.

Om vissa av de variabler i ett basregister, som ofta används för att selektera eller matcha är inaktuella eller ofullständiga, bör man göra *registerhållningsenkäter*. En sådan enkät, som bör vara kortfattad för att inte medföra onödig uppgiftslämnarbyrå, kan skickas till de objekt i basregistret där uppgift saknas eller där uppgiften kan misstänkas vara inaktuell.

Exempel: För att upprätthålla kvaliteten i Företagsregistret görs två enkäter per år till alla företag med mer än ett arbetsställe. Denna enkät görs visserligen för att upprätthålla kvaliteten på populationen av arbetsställen, men man frågar samtidigt efter aktuella näringsgrenar och adressuppgifter.

Det är naturligt inom Företagsregistret att man lägger ner mer registervårdsarbete på de stora företagen som ju har stor ekonomisk betydelse i de företagsundersökningar som SCB gör, men i en dynamisk ekonomi som förändras kraftigt kan detta statiska synsätt leda fel. I Johansson (2001) finns följande exempel där tillväxten i IT-branschen studeras.

Exempel: Småttingar kan vara jätteviktiga!

Johansson beskriver tillväxten inom IT-branschen mellan 1993 och 1998. Tabellen nedan bygger på uppgifter från Företagsregistret (Johansson, s 82-83).

Tabell 5B. Sysselsättning inom IT-branschen efter företagens storleksklass

Storlek, antal anställda	0	1	2-4	5-9	10-19	20-49	50-99	100-199	200-499	500+	Totalt
Antal sysselsatta, 1993	9 157	2 219	6 724	7 548	8 082	10 678	10 123	6 544	16 597	82 588	160 230
Antal sysselsatta, 1998	17 825	3 084	8 459	9 634	11 582	16 434	13 206	12 933	21 045	85 600	199 802
Förändring 1993-98, antal	8 668	865	1 735	2 086	3 500	5 756	3 083	6 389	4 478	3 012	39 572
Förändring 1993-98, %	95 %	39 %	26 %	28 %	43 %	54 %	30 %	98 %	27 %	4 %	25 %

Anm: För företag med 0 anställda har vi räknat med 1 sysselsatt, Johansson har räknat med 0.

Slutsats: Tabell 5B visar att de allra minsta företagen hade antalsmässigt den största tillväxten. Detta visar att det kan vara riskabelt att bortse från småföretagen vid registervårdsarbetet. Viktiga tillväxtmått kan då få låg kvalitet.

Exemplet ovan innehåller flera viktiga metodproblem – hur få jämförbarhet mellan åren, hur mäta sysselsättning och tillväxt? Jämförbarheten är ett stort problem för Johanssons undersökning, då Företagsregistret har ett tidsseriebrott 1996/97 som är störande för studier av småföretag.

Exempel: Populationen i Företagsregistret

Antal företag enl Företagsregistret		
	Gammal serie	Ny serie
1995	562 765	
1996	585 571	
1997	601 385	791 385
1998		810 337
1999		797 338

Registerpopulationen i Företagsregistret består av företag som är momspliktiga, registrerade som arbetsgivare eller F-skatteskyldiga. Från 1996 är ett företag momspliktigt oavsett omsättningens storlek, tidigare var gränsen en årsomsättning på 200 000 kronor. Den nya principen innebär att antalet företag ökade med ca 190 000 till totalt 791 385 företag under 1997 då de nyblivna momspliktiga företagen infördes i registret. Man har alltså låtit administrativa momsregler styra avgränsningen av registerpopulationen. Enligt exemplet i tabell 5A på föregående sida har den gamla serien för jordbruksföretag en undertäckning på ca 25 %. Genom att kombinera flera källor skulle undertäckningen kunna minskas och samma statistiskt relevanta populationsavgränsning kanske skulle kunna ha använts både före och efter momspliktsförändringen.

Ett basregister bygger på källor som kan ha helt olika periodicitet och olika tidsfördröjningar. Registrets täckning varierar då under året.

Exempel: Varierande täckning i Företagsregistret

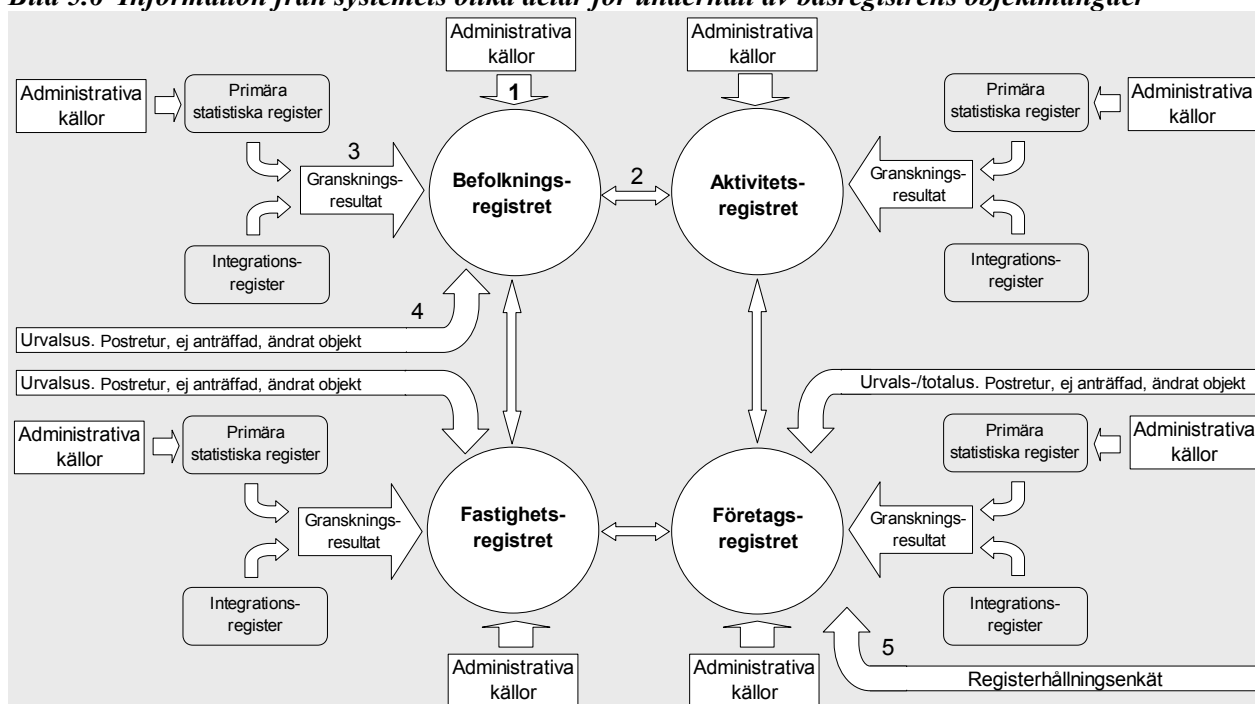
Företagsregistret bygger på aktuella månadsdata avseende skattedeklarationer för moms, arbetsgivaravgifter och avdragen skatt för företagens anställda. Men för små företag som inte lämnar månadsvisa skattedeklarationer används de årliga självdeklarationerna som kommer året efter inkomståret. Företagsregistret får kontinuerligt information om ny- och ombildade företag med varierande tidsfördröjning. Sådana ny- och ombildningar är särskilt vanliga vid årsskiften. Det krävs både tid och arbete innan dessa förändringar har införts.

Alla basregister måste uppfylla de krav som vi behandlat i detta avsnitt även om vi mest exemplifierar med Företagsregistret.

5.4.6 Alla skall stötta basregistren!

Hela systemet skall alltså samverka när det gäller att avgränsa objektmängder och objekt. Urvals- och totalundersökningar hämtar sina rampopulationer och de registerbaserade undersökningarna hämtar sina registerpopulationer från basregistren. Alla som gör undersökningar av dessa populationer får kunskaper som måste vidarebefordras till basregistren, så att basregistren kan vara SCB:s bästa källa avseende objektmängder.

Bild 5.6 visar de olika informationsflöden som bidrar till att basregistren kan hålla hög kvalitet med avseende på objekt och objektmängder. Dels skall de som arbetar med basregistren skaffa sig information men dessutom skall de som arbetar med andra register i systemet bidra med information som kan förbättra basregistrens kvalitet med avseende på objektmängder.

Bild 5.6 Information från systemets olika delar för underhåll av basregistrens objektmängder

Siffrorna i bild 5.6 refererar till punkterna nedan:

1. Inom varje basregister skall man hålla kontakt med de *myndigheter* som levererar administrativa data. Förändringar i det administrativa systemet kan påverka registrets objekt och täckning.
2. Inom varje basregister kan man få information från *angränsande basregister* där relaterade objekt kan ge upplysningar om det egna basregistrets objekt.
Exempel: Förändringar i fastigheter med bostäder måste också innebära att Befolkningsregistret skall ändras. Om en person saknar vissa aktiviteter i Aktivitetsregistret kan detta innebära att personen är övertäckning i Befolkningsregistret.
3. Granskningsresultat från olika *statistiska register* kan ge information om täckningsfel eller felaktiga objektavgränsningar.
4. *Urvals- och totalundersökningar* kan via postreturer och orsak till "ej anträffad" ge information om övertäckning och om att objekt har förändrats.
5. *Registerhållningsenkäter* görs av ett basregister till delar av objektmängden där man misstänker att uppgifterna är ofullständiga eller inaktuella.

Basregistren har alltså huvudrollen när det gäller objektmängder och även när det gäller objektavgränsningar, som diskuteras i nästa avsnitt. För att basregistren skall få så hög kvalitet som möjligt skall alla som använder SCB:s basregister vidarebefordra information om upptäckta kvalitetsbrister. Det är tre kategorier användare som skall vidarebefordra sådan information:

Information från urvals- och totalundersökningar

De som gör urvals- eller totalundersökningar har direktkontakt med uppgiftslämnare. Postreturer och spårning för telefonintervjuer kan ge information om

övertäckning, inaktuella adresser, att objekt har ombildats eller andra förändringar som är intressanta för dem som arbetar med registervård av det aktuella basregistret. Hushållsavgränsningar kan kontrolleras med urvalsundersökningar riktade mot hushåll. Företagsenheter kan kontrolleras i samband med olika företagsundersökningar.

Exempel: Övertäckning i Befolkningsregistret

Greijer (1995) analyserar bortfallet i Arbetskraftsundersökningen och ger en skattning av antal utrikes födda som troligen var övertäckning i Befolkningsregistret. Greijer (1996) analyserar postreturer och bortfall i en enkät till utrikes födda och skattar övertäckningen för olika födelseländer, åldrar och kön.

Information från primära statistiska register

De som arbetar med de primära statistiska registren tar emot administrativa register. Dessa kan innehålla objekt som inte finns i basregistren. Dessutom kan objekt ha försvunnit ur det administrativa registret, vilket kan ge information om att objektet har förändrats, ombildats eller upphört.

Exempel: Övertäckning i Befolkningsregistret

Greijer (1997a) analyserar de som saknade disponibel inkomst i Inkomstregistret och ger skattningar på övertäckningen för olika kategorier utrikes födda. Hon visar att man måste korrigera för övertäckning för att inte inkomststatistik efter födelseland skall bli missvisande.

Information från integrationsregister

De som arbetar med integrationsregister kan jämföra variabelvärden, som skall avse samma objekt från olika källor. Om man då upptäcker fel och inkonsistenser kan orsaken till detta vara att objekten har förändrats.

Exempel: Felaktiga objektavgränsningar i lantbruksregister

IAKS-registret innehåller lantbrukens ansökningar om arealbidrag. Wallgren (1999) matchade detta register mot SCB:s Lantbruksregister med länken PeOrgNr. Alla träffar kontrollerades genom att åkerareal från de två registren jämfördes. Ca 9 % falska träffar upptäcktes. Genom att föra samman enheter med samma telefonnummer kunde nya enheter härledas där åkerarealen stämde. Objektavgränsningar diskuteras i nästa avsnitt

5.5 Att avgränsa och härleda objekt

Registersystemets objekttyper kan ha skapats på fyra olika sätt:

- De kan ha skapats i ett administrativt system för administrativa ändamål, sådana objekt är ofta juridiskt betydelsefulla.
- De kan ha skapats i samarbete med en administrativ myndighet för SCB:s behov.
- De kan ha skapats inom SCB genom att samla in information som gör det möjligt att avgränsa objekten.
- De kan ha skapats inom SCB genom registerbearbetningar, det är denna kategori av objekttyper som kallas *härledda objekt*.

5.5.1 Administrativa objekttyper

Många av de administrativa objekttyperna är intressanta för statistiska syften, och de variabler som finns i de administrativa registren beskriver då statistiskt intressanta objekt.

Exempel på statistiskt intressanta administrativa objekttyper: Person, jobb (jobb är anställning eller eget företagande), fastigheter och fordon.

Exempel på administrativa objekttyper som inte alltid är statistiskt intressanta: Det skatterättsliga familjebegreppet (gifta eller sammanboende med gemensamma barn), juridisk enhet i administrativa företagsregister.

5.5.2 Objekttyper skapade i samarbete med administrativa myndigheter

De statistiska objekttyper som skapas av SCB i samarbete med en administrativ myndighet har fördelen att vissa administrativa uppgifter kan knytas till dessa statistiskt intressanta objekt, man kan alltså beskriva dessa objekttyper med statistiska variabler utan resurskrävande datainsamling.

Exempel på statistisk objekttyp som skapats i samarbete med administrativa myndigheter: Arbetsställe, information om dessa insamlas av Skatteverket för SCB:s behov genom kontrolluppgifterna. Detta är ett exempel på integrerad datainsamling, dvs. SCB och en administrativ myndighet samarbetar.

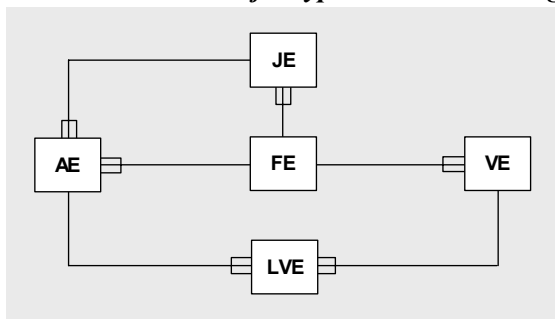
5.5.3 Statistiska objekttyper skapade av SCB via datainsamling

Genom undersökningar med egen datainsamling kan man samla in uppgifter för statistiskt intressanta objekttyper. För att kunna beskriva dessa objekttyper med statistiska variabler krävs alltså resurskrävande datainsamling.

Exempel på statistiska objekttyper som skapats efter egen datainsamling: De kosthushåll och bostadshushåll som skapas i SCB:s hushållsundersökningar (åtminstone fram till 2003). Inom den ekonomiska statistiken bildas företagsenhet, verksamhetsenhet och lokal verksamhetsenhet.

Alla administrativa företagsregister avser objekt som kallas juridiska enheter (JE) och som identifieras med person-/organisationsnummer (PeOrgNr). För statistiska ändamål bildas inom Företagsregistret objekttyperna företagsenhet (FE), arbetsställeenhet (AE), verksamhetsenhet (VE) och lokal verksamhetsenhet (LVE). Alla dessa enheter får egna identitetsnummer och enheterna måste vara observerbara så att data kan samlas in. Det är ett mycket omfattande arbete som krävs för att upprätthålla god kvalitet på dessa enheter då företag ofta förändras. I bild 5.7 nedan visas relationerna mellan dessa objekttyper.

Bild 5.7 De olika objekttyperna inom Företagsregistret



Det vanligaste är att en JE motsvarar en FE och en AE. Genom kontakt med stora företag skapas FE som är unionen av vissa JE. Dessutom delas vissa FE upp i ett antal VE, som är så branschrena som möjligt. Genom telefonkataloger och andra källor får SCB information om vilka företag som har flera AE. Via registerhållningsenkäter får dessa företag lämna uppgifter om dessa AE. Efter kontakt med företagen kan AE delas upp i förhoppningsvis branschrena LVE.

5.5.4 Statistiska objekttyper skapade av SCB via härledning

De statistiska objekttyper som skapas inom SCB genom härledning bygger på administrativa objekt och administrativa data för dessa objekt kan då användas för att beskriva de härledda statistiska objekten.

Exempel på statistiska objekttyper skapade genom härledning: Bostadshushåll baserade på den planerade folkbokföringen av personer på lägenhet. I lantbruksexemplet i föregående avsnitt skapas härledda företagsenheter genom att småföretag med samma telefonnummer definieras som en företagsenhet.

5.5.5 Objekt och identiteter

Förutom de viktiga uppgifter som basregistren har enligt avsnitten ovan, är det även inom basregistren som härledda objekt skapas. En annan viktig och svår uppgift är att följa objekten och deras identiteter över tiden och registrera statistiskt relevanta förändringar.

Varje objekttyp skall ha en entydig och klagörande definition. När man använder administrativa objekttyper för statistiska ändamål måste man sätta sig in i det administrativa systemets definitioner. De som skapar objekt inom SCB har ansvar för att dessa objekttyper definieras. Både de administrativa och statistiska definitionerna skall dokumenteras. Ett exempel på en sådan dokumentation är *Företagsenheten i den ekonomiska statistiken*, SCB (2002a).

När register matchas är det viktigt att *samma identitet* i olika register verkligen *avser samma objekt* (objektinstans med IT-terminologi). Om man vid matchning får falska träffar och importerar variabelvärden till ett objekt från ett annat objekt, men med liknande identitet, blir undersökningsresultaten felaktiga. Risken för sådana falska träffar är särskilt stor när man matchar register från olika tidpunkter eller tidsperioder, då identiteterna kan ha förändrats över tiden.

Denna typ av problem finns i alla delar av registersystemet, men har olika omfattning. Samma personnummer i register från olika år kan i sällsynta fall avse olika personer. Det förekommer även personnummerbyten så att samma person har olika personnummer i olika årgångar av register. Inom Befolkningsregistret har man uppgifter om dessa personnummer och skall man matcha personregister från olika år bör man anlita Befolkningsregistret.

Även fastighetsidentiteter (län, kommun, församling, fastighetsbeteckning) förändras när den regionala indelningen förändras. Detta innebär att man vid matchning av register från olika år måste använda nycklar mellan nya och gamla regionala koder.

Aktivitetsregistret berörs både av att personnummer och företagsidentitet kan förändras. När Aktivitets- och Företagsregistret matchas måste man med arbetsställeidentiteterna i de två registren kunna koppla samman identiska arbetsställen, AE. Detta försvåras av att företagen som fyller i kontrolluppgifter använder andra identitetsnummer än vad som används i Företagsregistret.

5.5.6 Primära statistiska register

De som tar emot administrativa register för att skapa primära statistiska register gör följande bearbetningar så att det statistiska registrets *objekt* blir de avsedda:

- Identitetsnumren i det administrativa registret ges det korrekta SCB-formatet (personnummer börjar med sekelsiffror etc.). Kontroll av dubletter i det administrativa registret skall göras.
- Det administrativa registret matchas mot det basregister som innehåller samma objekttyp. Denna matchning görs mot den version av basregistret som avser samma tidsperiod. Identiteten hos objekt som ger mismatch kontrolleras och åtgärdas med hjälp av namn, adress och telefonnummer.
- Om det statistiska registret skall avse en annan objekttyp än den som finns i den administrativa källan, kan det bli aktuellt att antingen lägga ihop eller dela upp värden för administrativa objekt. Detta görs med det basregister där de administrativa objekten kopplas till avsedd objekttyp. *Exempel:*

Bild 5.8 Juridiska enheter (JE) läggs ihop till företagsenheter (FE)

Basregistret		Administrativt register		Statistiskt register	
FE-nr	JE-nr	JE-nr	Lönesumma	FE-nr	Lönesumma
FE1	JE11	JE11	180	FE1	180 + 450 = 630
FE1	JE12	JE12	450	FE2	270 + 500 = 770
FE2	JE21	JE21	270		
FE2	JE22	JE22	500		

Anm: Olika ekonomiska variabler passar mer eller mindre bra till att adderas. Om t.ex. orderingång adderas kan order till en underleverantör komma att räknas två gånger.

Bild 5.9 Juridiska enheter (JE) delas upp i arbetsställeenheter (AE)

Basregistret				Beräkningar		Adm. register		Statistiskt register			
JE-nr	AE-nr	SNI	Antal anst	SNI	Oms/anst	JE-nr	Omsättning	AE-nr	SNI	Anst	Modellberäknad omsättning
JE1	AE11	DG	120	DA	2,1	JE1	300	AE11	DG	120	$300 \cdot 120 \cdot 1,5 / (120 \cdot 1,5 + 60 \cdot 2,1) = 176,5$
JE1	AE12	DA	60	DB	1,2	JE2	250	AE12	DA	60	$300 \cdot 60 \cdot 2,1 / (120 \cdot 1,5 + 60 \cdot 2,1) = 123,5$
JE2	AE21	DD	50	DD	1,8			AE21	DD	50	$250 \cdot 50 \cdot 1,8 / (50 \cdot 1,8 + 100 \cdot 1,2) = 107,1$
JE2	AE22	DB	100	DG	1,5			AE22	DB	100	$250 \cdot 100 \cdot 1,2 / (50 \cdot 1,8 + 100 \cdot 1,2) = 142,9$

Bild 5.9 visar hur man med en modell kan dela upp omsättning för en juridisk enhet på arbetsställen. Här används antal anställda och branschvis omsättning per anställd för branschrena företag (JE). Valet av modell är subjektivt – man kunde ha använt lönesummor i stället för antal anställda.

5.5.7 Kontroll av objektmängder och objektavgränsningar

De som hämtar data från flera statistiska register i systemet för att antingen bilda integrationsregister eller importera data till det "egna" registret, bör göra granskningar och kontroller för att upptäcka brister och inkonsistenser med avseende på objektmängder och objektavgränsningar.

- Man skall kontrollera att objektmängderna i de olika registren avser samma tidsperiod eller tidpunkt.
- Alla objekt som inte får träff vid matchning bör utredas – vad är orsaken till detta? Om alla inblandade register är avstämda mot sina respektive basregister bör det inte finnas mismatch, däremot kan det finnas bortfall.

- Granska variabelvärden från olika register för att hitta falska träffar.
- Meddela berörda register om brister och inkonsistenser som upptäcks.

5.5.8 Att skapa härledda objekt

När de administrativa objekten inte är helt lämpliga för statistiska ändamål kan man genom att använda identifierande variabler och referensvariabler bilda nya objekt som är kombinationer av två eller flera administrativa objekt.

Exempel på företagsenheter inom jordbruket skapade genom härledning:

För jordbruksföretagen kan objektavgränsningen ibland vara komplicerad. Det finns olika slag av administrativa material där i huvudsak personnummer används som identitet för juridiska enheter. Inom Statens jordbruksverks (SJV) administrativa stödutbetalningsregister (IAKS) identifieras ansökningarna med personnummer. Momsregistret och Standardiserade räkenskapsutdrag (SRU) innehåller också personnummer som identitet på jordbruksföretag. En liten del av jordbruksföretagen är handelsbolag eller aktiebolag med organisationsnummer som identitet. Inom Lantbrukets företagsregister (LBR) används ett speciellt nummer som identifierar en enhet med gemensam driftsledning. Följande identitetsvariabler finns inom de register vi nämnt:

- SRU-, Moms- och Företagsregistret (FDB) har personnummer eller organisationsnummer för *juridiska enheter*
- IAKS-registret har *ansökarens* person- eller organisationsnummer
- Lantbrukets företagsregister (LBR) har driftsenhetens LBR-nummer och *driftsledarens* person- och/eller organisationsnummer.

Lantbrukets företagsregister är uppbyggt för statistiska ändamål och innehåller statistiska objekt. SRU-, Moms- och Företagsregistret och IAKS-registret innehåller administrativa objekt. I bilden nedan anges tre exempel på hur dessa objekt kan vara relaterade. I det första fallet överensstämmer objekten i de tre registren, lantbruksföretag av denna typ ger inga problem – alla variabler i de tre registren kan användas för att beskriva ett statistiskt meningsfullt objekt. I de övriga fallen kan antingen två enheter i LBR slås samman eller två enheter i FDB respektive IAKS slås samman till en härledd företagsenhet.

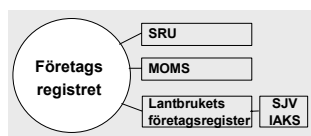


Bild 5.10 Objektavgränsning av jordbruksföretag inom tre register

FDB	LBR	IAKS	Hur bör statistiskt relevanta objekt avgränsas?
●	●	●	I detta enkla fall motsvaras en juridisk enhet (JE) i FDB av en driftsenhet i LBR och av en ansökan i IAKS och denna enhet bör avgränsas som en företagsenhet (FE), med en arbetställeenhet (AE).
●	●	●	Här motsvaras en JE i FDB av två driftsenheter i LBR och av en ansökan i IAKS. Detta fall kan inträffa om en person äger två gårdar, driver den ena själv och låter en anställd driva den andra gården. Ägaren gör en bidragsansökan och sköter momsbetalningar. Denna enhet bör avgränsas som en FE med två AE.
●	●	●	Här motsvaras två JE i FDB av en driftsenhet i LBR och av två ansökningar i IAKS. Detta fall kan inträffa om ett gift par driver ett jordbruk. Två bidragsansökningar lämnas in för två delar av verksamheten. Båda deklarerar inkomst av näringsverksamhet samt moms. Denna enhet bör avgränsas som en FE med en AE.

När flera register sambearbetas är det av stor betydelse för statistikens kvalitet att objekten med samma identitet har samma avgränsning, dvs. motsvarar samma verklighet. Kopplas t.ex. företagets omsättning samman med gårdens spannmålsareal måste samma företagsenhet (FE) avses. I ett effektivt registersystem skapas objekttyperna inom basregistren. Andra register skall vara avstämda mot respektive basregister med avseende på objektavgränsningar.

5.6 Att bestämma registrets variabelinnehåll

En variabel är en mätbar egenskap hos objekt. I kapitel 4.5 beskrivs begreppet variabel både inom statistik- och systemvetenskapen. Inom registersystemet skiljer vi på variabler med olika ursprung:

- *lokala primära variabler* har hämtats från ett administrativt register eller genom en totalundersökning till det aktuella registret,
- *lokalt härledda variabler* har bildats i det aktuella registret med hjälp av andra variabler i registret,
- *importerade primära variabler* är primära variabler som har hämtats till det aktuella registret från ett annat register i registersystemet och
- *importerade härledda variabler* är härledda variabler som hämtats till det aktuella registret från ett annat register i systemet.

5.6.1 Variabeldefinitioner

Varje undersökning utgår från en frågeställning som formuleras i teoretiska eller allmänna termer. De teoretiska begreppen i frågeställningen måste sedan *operationaliseras*, dvs. översättas till mätbara begrepp. När man definierar dessa mätbara begrepp avgör man *vad* som skall undersökas. Detta har vi tidigare påpekat när vi diskuterade definitionen av undersökningens population. Även när det gäller variabeldefinitioner så skall teoretiska begrepp som avser objektens egenskaper översättas till klart definierade statistiska variabler.

En variabeldefinition skall ange vilken objekttyp som variabeln avser samt variabelns referenstidpunkt eller referensperiod, dvs. att egenskapen avser en viss tidpunkt/period. Vidare skall definitionen ange hur egenskapen skall mätas och vilken mätskala som används. Två exempel avseende ålder får belysa detta:

Definition 1: Med ålder avses personers ålder i uppnådda hela år vid undersökningens årets slut. *Exempel:* En person född den 1 januari 1980 har för en undersökning som avser den 31/12 år 2000 åldern 20 år.

Definition 2: Med ålder avses personers ålder i år avrundat till en decimal vid referenstidpunkten. *Exempel:* En person född den 1 januari 1980 har för en undersökning som avser den 31/12 år 2000 åldern 21,0 år.

En *primär variablers definition* avgörs när man formulerar en fråga på en frågeblankett eller administrativ blankett. Även anvisningar till frågan ingår i definitionen. I arbetet med att dokumentera statistik baserad på administrativa data skall man därför alltid samla in och arkivera den registerlämnande myndighetens blanketter och broschyrer med anvisningar, helst även i elektronisk form.

En *härledd variabels definition* bestäms dels av definitionerna av de variabler som den bygger på och dels av regeln för hur den härledda variabeln bildas.

Även här är det önskvärt att definitionerna är både *adekvata*, dvs. överensstämmer med vad man vill undersöka, och *funktionella*, dvs. definitionerna skall vara lätta att använda.

Adekvata variabeldefinitioner

Att en variabeldefinition är adekvat innebär även att *definitionen överensstämmer med andra variabler* i det register som skapas för undersökningens syfte. I ett företagsregister skall definitioner av variabler som produktion, försäljning, ingående och utgående lager överensstämma så att den logiska relationen mellan variablerna gäller. Produktion minus försäljning skall vara lika med periodens lagerförändring.

Eftersom SCB:s register ingår i ett system så måste även *variabeldefinitionerna i olika register överensstämma* för att vara adekvata. T.ex. i alla personregister är grundregeln att variabler som ålder, civilstånd etc. definieras på samma sätt, annars skulle samutnyttjande av registerdata försvåras och registerstatistiken skulle få sämre kvalitet genom bristande konsistens och jämförbarhet.

Även *grupperingar och klassindelningar* av variabler bör vara samma i olika register. T.ex. om åldersklasserna är de samma i olika personregister kommer de publicerade tabellerna att vara jämförbara, vilket ökar statistikens sam användbarhet, som är en del av kvalitetsbegreppet.

Funktionella variabeldefinitioner och härledda variabler

Att använda de administrativa variabelernas definitioner är alltid funktionellt. Om de administrativa variablerna inte är tillräckligt adekvata för de statistiska behoven kan man försöka bilda härledda variabler (se avsnitt 4.5.3). För att kunna bilda statistiskt intressanta variabler skall registersystemets möjligheter utnyttjas. Genom att importera variabler både från administrativa och statistiska källregister skapas ett register med rikt innehåll och många intressanta härledda variabler kan då bildas.

Exempel: Inkomst- och taxeringsregistret (beskrivs i avsnitt 3.1.1) bygger på 100-tals skatteadministrativa inkomstvariabler. Med hjälp av dessa bildas ett antal viktiga statistiska variabler som disponibel inkomst mm.

Namnsättning av variabler och standardiserade variabler

Registervariabler som används av flera undersökningar eller produkter bör *standardiseras*. Detta innebär att ansvar för namnsättning, kvalitet och dokumentation av variabeln finns hos det register där variabeln skapas för första gången inom SCB. För detta register är alltså den standardiserade variabeln en lokal primär variabel eller en lokalt härledd variabel.

När en standardiserad variabel används av andra inom SCB skall man använda variabeln med oförändrat namn och definition. Dokumentationen skall vara lätt åtkomlig för alla. De som importerar en standardiserad variabel behöver då inte göra en egen variabeldokumentation. Den standardiserade variabelns namn får inte användas som namn på andra variabler.

En av de stora fördelarna med ett registersystem är just möjligheterna att kunna kombinera många källor och få ett rikt innehåll

Standardiserade variabler nämns i avsnitt 1.4 och diskuteras även i kapitel 11

5.6.2 Arbetsgång när registrets variabler bildas

När det nya registrets population har skapats och kontrollerats sker den egentliga variabelsektionen. Undersökningsvariablerna, dvs. de statistiskt intressanta variablerna, importeras från olika källor.

1. Problemformuleringsfasen ger svar på vilket innehåll registret skall ha. Vilka är användarnas behov, vilka möjligheter till uppdrag finns?
2. Från alla relevanta källor skall variabler hämtas till det nya registret. När man skapar ett statistiskt register skall alla registeruppgifter som kan belysa undersökningsproblemet användas. Man skall skapa bästa möjliga variabler. Det statistiska registret skall vara rensat från uppenbara fel och vara konsistent, dvs. inte innehålla motstridiga uppgifter. Även variabler som skall användas vid granskningen importeras. Dessa variabler har anknytning till undersökningsvariablerna så att granskningsvillkor kan formuleras. Tidigare versioner av det aktuella registret bör användas om det är fråga om en återkommande undersökning.
3. Adjungerade härledda variabler bildas genom matchning med andra register. Aggregerade härledda variabler bildas genom matchning och bearbetningar av andra register. Dessa typer av härledda variabler beskrivs i avsnitt 4.5.3.
4. Registrets variabler granskas. Hur upptäcker man uppenbara och misstänkta fel? Granskningsarbetet diskuteras i avsnitt 5.7.
5. Hur ersätta orimliga och saknade värden? I nästa kapitel, avsnitt 6.4 beskrivs hur det partiella bortfallet kan åtgärdas.
6. Bilda härledda variabler med registrets variabler. Detta behandlas i avsnitt 5.8 nedan.
7. Dokumentera bearbetningar, granskningsresultat och de färdiga variablerna.

Det register som skapas kommer att ingå i SCB:s registersystem vilket innebär krav på samordning och samverkan. De granskningar och bearbetningar som görs skall duga för alla användningar inom registersystemet. Genom detta så undviks dubbelarbete och värdet ökar på det register som skapas.

5.7 Granska och åtgärda grunddata

I Granquist m.fl. (2002) *Guide till granskning* beskrivs principer och metoder för granskningsarbetet. Där diskuteras främst undersökningar med egen datainsamling, dvs. urvals- och totalundersökningar. Granskning av register baserade på administrativa data är annorlunda. Det går t.ex. inte att kontakta den ursprungliga uppgiftslämnaren, i stället måste man ha goda kontakter med den registerlämnande myndigheten. Granskning av registerdata är ett internationellt sett eftersatt område och något systematiskt utvecklingsarbete har inte gjorts vid SCB.

Vid företagsundersökningar med egen datainsamling lägger man i vissa fall ner mycket arbete med att återkontakta uppgiftslämnarna för att åtgärda misstänkta fel. Enligt Granquist m.fl. är det tveksamt om detta är ett effektivt sätt att arbeta. Vid registerbaserade undersökningar förekommer inte detta arbetssätt.

Granskningens roll är att öka kvaliteten på både kort och lång sikt:

- Upptäcka och åtgärda fel i det aktuella datamaterialet. I administrativa material kan det finnas uppenbara fel som kan åtgärdas maskinellt.
- Hitta felkällor och i samverkan med den registerlämnande myndigheten minska omfattningen av dessa felkällor.
- Bidra till att den ämnesmässiga kompetensen ökar hos den personal som arbetar med registret. Granskningsarbetet bidrar till att man lär känna datamaterialets egenskaper och får förståelse för hur data kan användas.

För att undvika dubbelarbete inom registersystemet skall man för varje register granska de *lokala primära variabler* som har hämtats från ett administrativt register eller genom en totalundersökning till det aktuella registret. När sedan dessa variabler importeras av andra register skall ingen ny granskning behövas.

Värden som saknas eller som måste underkännas ger upphov till partiellt bortfall. Dessa saknade värden kan ersättas med imputerade värden. I avsnitt 6.4.4 diskuteras hur partiellt bortfall kan åtgärdas genom imputering. Det finns metoder för automatisk granskning och imputering som för närvarande inte används i Sverige.

Vi redovisar nedan fyra fallstudier som belyser granskningsarbetet vid olika register inom SCB. Därefter följer en sammanfattning av fallstudiernas resultat.

5.7.1 Granskningsarbetet inom Inkomst- och taxeringsregistret

Som första exempel på hur granskningsarbetet kan läggas upp använder vi Inkomst- och taxeringsregistret. Detta register används för att med självdeklarationer och andra administrativa register belysa Inkomst- och taxeringsfördelningen för individer och familjer.

Registrets variabler används även i mikrosimuleringsmodellen FASIT. Denna simuleringsmodell används av bl.a. Finansdepartementet för att studera effekter av planerade förändringar av skatter och transfereringar. Registret måste uppfylla mycket stora kvalitetskrav på mikronivå, dvs. Inkomst- och taxeringsvärden för enskilda personer och familjer måste vara fullständiga och konsistenta. För att inte få märkliga simuleringsresultat måste man därför lägga ner ett omfattande arbete med att granska och korrigera variabelvärden.

En beskrivning av arbetet med att skapa registret finns i avsnitt 3.1 där Inkomst- och taxeringsregistret används för att illustrera hur administrativa register omvandlas till statistiska register. De olika källregister som används anges i bild 3.2.

Steg 1 – Registervisa kontroller av administrativa källregister

Totalt tar man varje år emot ca 30 administrativa register, det första som görs är att vart och ett av dessa kontrolleras på följande sätt:

Först kontrolleras *postbeskrivningarna* som kommer tillsammans med de nya administrativa registren. Dessa postbeskrivningar ändras som regel varje år, variabelnamn kan ha ändrats och nya variabler kan ha tillkommit. Postbeskrivningarna kan missförstås, vilket kommer att generera fel. Det är därför viktigt

med en nära kontakt med personer inom den registerlämnande myndigheten som kan ge de förklaringar som behövs.

Därefter undersöks *extremvärden* på de kvantitativa variablerna. Man jämför då med *föregående års värden* både på aggregerad nivå och på individnivå. Vissa variabler som t.ex. sjukpenning har takvärden som kan användas vid dessa kontroller. Därefter görs *logiska kontroller* av att summor verkligen är summan av sina delar. I alla de fall där man upptäcker att aggregerade värden ligger på en oväntad nivå, att personposter har extremvärden eller att de har summor som inte stämmer med sina delar, så kontaktas den registerlämnande myndigheten och värdena rättas. Felorsaken kan vara missuppfattningar av postbeskrivningen, eller att extremvärden beror på någon taxeringsåtgärd som inte bör påverka statistiken. För personposterna är felfrekvensen så låg att man kan göra individuella rättelser i samråd med den administrativa myndigheten.

Vissa variabler rapporteras från underordnade myndigheter till den centrala myndigheten, som i sin tur lämnar data till SCB. I sådana situationer är det lämpligt att kontrollera att alla underordnade myndigheter har lämnat uppgifter. T.ex. så brukar uppgifter om socialbidrag saknas för några kommuner varje år. Inom Inkomstregistret måste man dokumentera vilka kommuner som inte lämnat uppgifter och göra imputeringar för dessa saknade värden. Föregående års socialbidrag brukar då användas.

Granskningsarbetet har inte endast direkta effekter på registrets kvalitet. Om arbetet organiseras så att flera personer delar ansvaret, så bidrar granskningen till samverkan och erfarenhetsutbyte inom arbetsgruppen. Därigenom ökar *ämneskompetensen* och indirekt även registrets kvalitet. När man sedan dokumenterar arbetet och de åtgärder som gjorts för att korrigera data, så befästs ämneskompetensen ytterligare.

Leverantörskontakterna har flera betydelsefulla effekter. Man skall dels förklara för personal inom den registerlämnande myndigheten hur man inom SCB använder myndighetens data och till vad. Myndighetens personal bör få förståelse för de konsekvenser som kvalitetsbrister kan ha för statistikens användare. För SCB-personalens ämneskompetens är kontakterna med leverantörerna viktiga, därför har man inom Inkomstregistret regelbundna träffar med Skatteverket två gånger per år. Dessa kontakter används också för det mycket viktiga arbetet med att söka efter nya administrativa källor.

De statistiska register inom SCB:s registersystem, som används som källor till Inkomst- och taxeringsregistret (IoT), skall i princip inte behöva granskas en gång till när de tas emot av andra användare inom SCB – granskningen skall vara gjord vid det primära statistiska registret.

Något som fungerar dåligt hittills är erfarenhetsutbytet mellan olika enheter inom SCB. Hur många vet hur mycket som samlas in för IoT-registret? De som har egna data som skulle kunna användas för IoT-registret behov, vet de vilka krav som IoT-registret ställer? Uppfyller t.ex. Fastighetstaxeringsregistret de krav som skatte- och förmögenhetsberäkningarna inom IoT-registret ställer? Det skall ju inte vara så att dålig samverkan medför att flera enheter inom SCB

tar emot och granskar samma administrativa data. Dubbelarbete för SCB och de myndigheter som lämnar register skall undvikas.

Steg 2 – Slutlig kontroll av det totala registret

I det första steget ovan har alltså data från varje myndighet kontrollerats var för sig. I nästa steg sammanförs alla variabler från samtliga källor i ett totalregister så att de olika källorna kan jämföras. Därefter bildar man alla härledda variabler. På detta sätt kan man göra nya konsistenskontroller, t.ex. av att en summa av variabelvärden från olika källor överensstämmer med summan i en annan källa. På detta sätt kommer ytterligare fel att upptäckas.

Denna del av granskningsarbetet görs på *ett urval* av poster i totalregistret. Detta beror dels på att arbetet annars skulle vara för tungt och omfattande och dels på att man har särskilt höga kvalitetskrav på den del av totalregistret som skall användas i FASIT-modellen. Totalregistret består av ca 9 miljoner poster med ca 500 variabler, därför är urvalsmetoden en förutsättning för att klara av den noggranna granskningen. Det är viktigt att påpeka att granskningsarbetet ökar personalens ämneskompetens – att konkret arbeta med data på mikronivå ger viktiga kunskaper – att arbeta med ett urval kan vara en bra metod för att lära känna datamaterialet.

De fel man hittar på detta sätt är inte fler än att de kan rättas manuellt i samråd med de administrativa myndigheterna. Om rättelserna i urvalet inte ger för stora effekter på skattningarna i urvalet så godkänner man totalregistret och denna kontrollfas är avslutad. Om rättelserna i urvalet är för omfattande skulle man ha behov av metoder för automatisk kontroll och rättning av alla poster i totalregistret. Sådana metoder saknar man för närvarande.

Steg 3 – Kontroll av färdiga skattningar

I detta steg bildar man alla viktiga tabeller med hela registret som bas. Skattningarna granskas och jämförs med föregående års värden. Dessutom gör man med FASIT-modellen ett antal simuleringar vars enda syfte är att pröva datakvaliteten. Om t.ex. bostadsbidragen hålls oförändrade i modellen så skall modellen generera modellvärden som överensstämmer med de tabeller man tidigare tagit fram.

5.7.2 Granskningsarbetet inom Kontrolluppgiftsregistret

Kontrolluppgiftsregistret används dels för att beräkna branschvisa lönesummor och dels när Aktivitets- och Sysselsättningsregistret skapas. I detta avsnitt redogörs för granskningen av de definitiva kontrolluppgifterna för inkomståret 2001 som SCB tog emot fram till oktober 2002. Kontrolluppgifterna granskas av de som ansvarar för Kontrolluppgiftsregistret och det granskade registret används sedan som källa för andra register inom registersystemet.

Kontroll av populationsavgränsning

Första steget i granskningen är en avstämning av att antalet mottagna kontrolluppgifter stämmer med de som sända av Skatteverket.

Andra steget i arbetet är att skapa en datamatrix med de slutliga kontrolluppgifterna efter alla ändringar som finns i sändningen. SKV ändrar inte i inlagda

uppgifter utan när uppgiftslämnarna (i detta fall arbetsgivarna) lämnar ändringar till SKV, skapas nya poster som motsvarar borttagning, ändring eller ersättning av tidigare poster. På grund av detta krävs registerbearbetningar för att ta bort ogiltiga poster och att göra flera dubblettkontroller. För ca 10 300 poster förs variabelvärden över från den ursprungliga kontrolluppgiften till ändringsposten, då ändringsposterna kan vara ofullständigt ifyllda.

Det tredje steget i granskningsarbetet är att kontrollera alla identiteter. Då kontrolluppgifterna innehåller både personidentitet och företagsidentitet skall både personnummer och organisationsnummer kontrolleras. Ca 7 600 personnummer var felaktiga varav ca 5 000 personnummer kunde rättas maskinellt.

Det fjärde steget innebär att personnumren på kontrolluppgifterna matchas mot Befolkningsregistret för den 31/12 2001 och att företagsidentiteten *PeOrgNr* matchas mot Företagsregistret för mars 2002. I båda fallen upptäcks flera omaka poster – KU-registret innehåller fler personnummer än Befolkningsregistret och ca 12 200 fler *PeOrgNr* än Företagsregistret. Eftersom kontrolluppgifterna avser hela året så borde man ha matchat mot årgångsversioner av Befolknings- och Företagsregistret. De 12 200 posterna som inte fanns i Företagsregistret består huvudsakligen av *PeOrgNr* som används tillfälligt för konkursbolag etc. samt i viss utsträckning av felaktiga personnummer.

Kontroll av variabelvärden

I *det femte steget* granskas avvikelsefel genom 16 olika rimlighetskontroller. Relationen mellan löneinkomst och skatt används på flera sätt och dessutom söker man poster med extremt hög löneinkomst eller extremt hög skatt. Ca 5 000 poster med extrema värden upptäcks i dessa kontroller. Dessa granskas översiktligt och endast ett fåtal kontrolleras med SKV. Efter detta granskningsmoment kan kontrolluppgiften accepteras, ersättas med en ny uppgift eller tas bort ur registret.

Kompletteringsarbetet – kontroll av den viktigaste variabeln

Den absolut viktigaste fasen i granskningsarbetet innebär att kontrollera att de anställda personerna kopplas till rätt arbetsställe. Denna koppling är viktig för hela registersystemet som gör det möjligt att redovisa förvärvsarbetande efter näringsgren och region. Svårigheterna att koppla de anställda till rätt arbetsställe gäller företag med fler än ett arbetsställe. Trots att arbetsgivarna är skyldiga att ange arbetsstället på varje kontrolluppgift så saknas denna uppgift ofta och dessutom är uppgiften om arbetsställe ibland orimlig.

Orimliga arbetsställenummer upptäcks genom att antal anställda jämförs med motsvarande uppgift i Företagsregistret och med föregående års version av Kontrolluppgiftsregistret. Rimligheten i pendlingsavstånd bedöms också.

När arbetsställeuppgift saknas på kontrolluppgifter för flerarbetsställeföretag tas kontakt med arbetsgivare. Samma gäller då arbetsställeuppgiften verkar orimlig. Detta partiella bortfall åtgärdas alltså genom särskild uppgiftsinsamling genom en registerhållningsenkät. De som ansvarar för Kontrolluppgiftsregistret samarbetar med de ansvariga för Företagsregistret angående arbetsställenas kommun och näringsgren för att fånga upp förändringar.

Outputgranskning

Kontrolluppgiftsregistret används som källa till Sysselsättningsregistret och genom att outputgranska Sysselsättningsregistret kontrolleras även kvaliteten på Kontrolluppgiftsregistret. Detaljerade tabeller med sysselsatta efter näringsgren och kommun tas fram och jämförs med föregående års tabeller. Avvikelserna granskas och granskningsresultaten dokumenteras. Denna dokumentation är mycket värdefull då många användare hör av sig och ifrågasätter resultaten. Då finns dokumentationen, ”vi har kollat och det stämmer så långt vi kan bedöma!”

5.7.3 Granskningsarbetet inom Företagsstatistiken - SRU

Företagsstatistiken är en årlig undersökning där bokslutsdata används. Populationen består av alla företag utom de finansiella företagen och omfattar c:a 700 000 företag. Företagsenheter (FE), verksamhetsenheter (VE) och industrins lokala verksamhetsenheter (LVE) undersöks. I mars sänds enkäter till c:a 5 700 större företag, för övriga företag i populationen används SRU-data (Standardiserade RäkenskapsUtdrag) från Skatteverket.

I arbetets *första steg* skapas registerpopulationen genom att använda Företagsregistrets novemberversion (SAMU-versionen) över aktiva företag. Om företagens SRU-data avser år 2001, så används Företagsregistret för november 2001 och de preliminära SRU-blanketterna kommer till SCB under augusti-september 2002. Efter ändringar och taxeringsbeslut finns definitiva SRU-data, men dessa används inte av Företagsstatistiken för närvarande.

Det *andra steget* i arbetet är mottagandet av SRU-sändningen från den registerlämnande myndigheten, i detta fall SKV. Sändningen är omfattande och data har inte den struktur som är vanlig inom SCB utan består av en mängd delposter: *Identitet, posttyp, värde*. Uppgifterna skickas variabelvis och inte radvis eller objektvis. Överföringen av data kan ta över en vecka, materialet består av ca 1 miljon blanketter med mellan 300 och 400 variabler.

Det *tredje steget* i arbetet är att omstrukturera hela datamaterialet till datamatriser med alla variabler, som avser blanketten från ett visst företag, på en rad i matrisen. En datamatrix skapas för varje blankettyper. Noggranna granskningar behövs för att kontrollera att alla uppgifter kommer med och att data verkligen blir rätt tolkat vid omstruktureringen. Ändringar mellan åren sker i datamaterialet vilket kräver kontakter med SKV och tillgång till årets samtliga blankettyper. Dessa blanketter är komplicerade med många företagsekonomiska och skattetekniska begrepp, detta kräver hög ämnesmässig kompetens inom den enhet som tar emot SRU-data så att datamaterialet tolkas och utnyttjas på rätt sätt. Minst fem andra enheter inom SCB använder därefter dessa datamatriser (uppgifterna i detta avsnitt avser företagets verksamhet år 2001).

I arbetets *fjärde steg* kontrolleras och kompletteras SRU-uppgifterna innan de importeras till det nya registret. Enkätuppgifterna granskas för sig.

- Fyra summakontroller av resultat- och balansräkningarnas totaler görs.
- Mellan 6 000 och 8 000 maskinella rättningar görs, som regel ändras tecken, men även vissa fält som saknar värden fylls i automatiskt.

- Investeringar och förändring i eget kapital beräknas. Dessa jämförs med föregående års värden.
- Antal helårsanställda matchas på från en extern källa (MMP) som säljer årsredovisningsdata för aktiebolag. För de företag som inte finns med i denna källa imputeras värden för antal årsanställda. Företagets lönesumma divideras med branschens medellön per helårsanställd som beräknas med MMP-källan.

I det *femte steget* importeras variablerna från SRU-materialet till det nya registret och därefter görs ett antal kontroller som sedan bearbetas manuellt. Det är främst stora förändringar och härledda variabler som kontrolleras. Samtliga SRU-blanketter används, men endast de som ingår i Företagsstatistikens SAMU-population ingår i den ordinarie redovisningen. För ca 90 000 företag saknas uppgift om SNI, vissa av dessa är nystartade företag som ännu inte blivit SNI-kodade. Dessa 90 000 ingår inte i SAMU-populationen men kan användas i specialuttag t.ex. för att kvantifiera undertäckningen i SAMU-populationen. De poster som inte går att matcha mot SAMU-populationen motsvarar en årsomsättning på ca 35 miljarder kronor.

Efter dessa bearbetningar finns det partiellt bortfall i det nya registret, vilket innebär att ekonomiska variabelvärden saknas för ca 100 000 företag i SAMU-populationen. Men uppgift om SNI och antal årsanställda finns för dessa företag. Arbetets *sjätte steg* innebär att värden imputeras för dessa ca 100 000 företag. För de företag där ekonomiska variabelvärden finns, beräknas medelvärden för olika SNI och antal anställda. Dessa medelvärden används sedan för att imputera värden för de 100 000 företag som saknar ekonomiska variabelvärden. Därefter sker vissa modellberäkningar av härledda variabler där intäkter och kostnader fördelas på olika poster med hjälp av kvoter som hämtas från intermittenta branschundersökningar och industrins varuproduktion. Detta sker för nationalräkenskapernas behov och publiceras inte.

Slutprodukten består av tre registerversioner som avser företagsenheter, verksamhetsenheter och för industrin även lokala verksamhetsenheter. Här ingår även enkätdata. SRU är knutet till den juridiska enheten (JE). FE som är sammansatta av flera JE och företag med fler än en VE undersöks via enkät. Även industriföretag med flera LVE undersöks via enkät.

5.7.4 Granskning av SRU-data för FRIDA-registret

Företagsstatistiken i exemplet ovan använder preliminära SRU-data som levereras under augusti/september. FRIDA-registret bygger på definitiva SRU med taxeringsvariabler som levereras fyra månader senare. Detta register används på ett likartat sätt som FASIT-modellen i avsnitt 5.7.1, men syftet är att studera skatteeffekter på företag och företagare. FRIDA-registret har utvecklats under flera år i nära samarbete med Finansdepartementet. Registret finns dokumenterat, senast i SCB (2002b), *Dokumentation av FRIDA – inkomståret 2000*. Dokumentationen är även ett intressant exempel på hur ett register baserat på administrativa data kan dokumenteras, granskningsmetodiken beskrivs och de administrativa blankettyperna ingår i rapporten.

FRIDA-registret är uppdelat på fem delregister, ett delregister för varje företagstyp som har sin speciella blankett

Ursprungstanken var att skapa ett register med alla företag. Men man fann att datakvaliteten var alldeles för dålig för att detta skulle vara möjligt. Granskningsarbetet skulle ha varit alldeles för betungande om man ville skapa ett totalregister med hög kvalitet. Datakvaliteten har under årens lopp blivit bättre, men framförallt har det granskningsprogram som man utvecklat blivit så avancerat, så att det nu skulle kunna övervägas att skapa ett totalregister. En förutsättning för att FRIDA-registret skall kunna användas för simuleringsmodeller är att granskningen är så effektiv att datakvaliteten blir hög. De höga kvalitetskraven är orsak till att man har valt att arbeta med definitiva SRU-data och att jämföra dessa med de taxeringsuppgifter som blir tillgängliga samtidigt med de definitiva SRU-uppgifterna. Det komplicerade datamaterialet har krävt att man arbetat långsiktigt med att bygga upp ämneskompetens och att ha täta kontakter med både registerlämnande myndighet och uppdragsgivare.

Efter att ha tagit emot och strukturerat data på samma sätt som i avsnitt 5.7.3 är nästa steg i arbetet att skapa *härledda företagsenheter* av enskilda näringsidkare. Detta är en svärbearbetad företagsgrupp med flera typer av blanketter och många företagare bedriver flera olika verksamheter. Alla företag som kan kopplas till personerna i samma taxeringshushåll bildar en företagsenhet.

En första granskning görs på totalmaterialet för att hitta vissa tekniska fel. En sådan typ av fel är att rutans nummer på blanketten flyter ihop med beloppet. Om t.ex. ett redovisat resultat på 1 000 000 kr skall anges i ruta 605 så kan beloppet felaktigt ha blivit 6051000000 kr. Ett resultat på 0 kr kan av samma anledning felaktigt bli 605 kr. Granskningsprogrammet skall då kunna känna igen mönstret "605". Dessa fel rättas sedan manuellt.

FRIDA-registret är helt baserat på administrativa data, dvs. SCB gör ingen egen datainsamling.

Trots att man idag är tvungen att göra urval hänför vi FRIDA till registersystemet

Populationen av företag stratifieras efter företagstyp och storlek och därefter görs urvalen till FRIDA-registret. Vilande företag rensas sedan bort och de poster som inte har definierats som vilande går därefter in i det automatiska gransknings- och rättningsprogrammet.

Felsummeringar, skrivfel och taxeringsändringar, som inte korrigerats i SRU-materialet från SKV, är vanliga fel i materialet. Gransknings- och rättningsprogrammet kontrollerar alla summeringar och att värden som flyttas från en ruta på blanketten till en annan ruta stämmer. De olika feltyper som personalen som arbetar med SRU-data har upptäckt under årens lopp har lagts in i programmet och rättning sker automatiskt vid körningen. Om en summa stämmer efter ändring enligt en viss feltyp så görs rättning enligt feltypen. På detta sätt kan teckenfel, 1000-fel och andra fel åtgärdas.

Ett delurval består av ca 40 000 aktiebolag. Av dessa passerar ca 30 000 företag granskningsprogrammet helt utan åtgärd, alla summor stämmer. Men för ca 10 000 företag upptäcks fel. Av dessa kan ca 8 000 rättas automatiskt genom alla de feltyper som finns i programmet. De övriga 2 000 granskas ytterligare. Av dessa har ca 1 700 procentuellt sett små fel, dessa poster godkänns därför utan ytterligare åtgärd. Då återstår ca 300 poster med stora fel som granskas manuellt. Dessa 300 jämförs med företagets årsredovisning som lämnats till Patent- och registreringsverket (PRV). Ca 60 poster kan då rättas, men de övriga 240 går inte att åtgärda. För varje felaktig summa skapas en särskild

variabel där differensen sparas. Varje rättning som görs tilldelas en rättningskod. Poster som innehåller fel som inte kan rättas tilldelas en felkod.

Skatteregler ändras så att datamaterialet ändras varje år, variabler försvinner och nya tillkommer. Detta innebär att granskningsprogrammet måste modifieras regelbundet. Dessutom ändras även användarens behov, förändringarna av granskningen måste därför diskuteras med användaren, så att arbetet vid SCB fokuseras på de viktiga problemen. Genom att personalen vid SCB "lever med grundmaterialet på mikronivå" så pågår en lärprocess som leder till högre ämneskompetens. Det är viktigt att även SCB-personalen använder data och regelbundet diskuterar användarnas problem. Kraven på data och analyserna av data ändras över tiden och inom SCB måste man följa denna utveckling!

5.7.5 Sammanfattning – granskningsprocessen i registersystemet

En analys av dessa fyra fall visar att granskningen av data från administrativa källor kan struktureras i följande sex moment. Under varje rubrik citerar vi ur fallbeskrivningarna ovan:

1. Skapa en datamatrix och *föra samman alla poster* som hör till samma objekt.

Kontrolluppgiftsregistret:

En datamatrix skapas med de slutliga kontrolluppgifterna efter alla ändringar som finns i sändningen. På grund av detta krävs registerbearbetningar för att ta bort ogiltiga poster och att göra flera dubblettkontroller.

Företagsstatistiken:

Datamaterialet omstruktureras till datamatriser med alla variabler, som avser blanketten från ett visst företag, på en rad i matrisen.

2. Granskningen skall göras för att *kontrollera registerpopulationen*, dvs. att inga objekt saknas eller att främmande objekt eller dubletter förekommer. Detta görs genom avstämning mot respektive basregister. I detta ingår att alla identitetsvariabler kontrolleras och ges korrekt SCB-format.

Kontrolluppgiftsregistret:

Ca 7 600 personnummer var felaktiga varav ca 5 000 personnummer kunde rättas maskinellt. Personnumren på kontrolluppgifterna matchas mot Befolkningsregistret för den 31/12 2001 och att företagsidentiteten PeOrgNr matchas mot Företagsregistret för mars 2002. I båda fallen upptäckts flera omaka poster.

Företagsstatistiken:

Registerpopulationen skapas genom att använda Företagsregistrets novemberversion. Om företagens SRU-data avser år 2001, så används Företagsregistret för november 2001 och de preliminära SRU-blanketterna kommer till SCB under augusti-september 2002.

3. Granskningen kan även behöva göras för att *kontrollera objekten*, dvs. att data avseende en viss identitet från olika källor avser samma objekt. Detta kan göras genom att jämföra värden på likartade variabler från olika källor.

Företagsstatistiken:

Slutprodukten består av tre registerversioner som avser företagsenheter, verksamhetsenheter och för industrin även lokala verksamhetsenheter.

FRIDA:

Härledda företagsenheter avseende enskilda näringsidkare skapas. Alla företag som kan kopplas till personerna i samma taxeringshushåll bildar en företagsenhet.

Avsnitt 5.4.6, sista exemplet:

IAKS-registret innehåller lantbrukens ansökningar om arealbidrag. Wallgren (1999) matchade detta register mot SCB:s Lantbruksregister med länken PeOrgNr. Alla träffar kontrollerades genom att åkerareal från de två registren jämfördes. Ca 9 % falska träffar upptäcktes.

4. Kontroller skall göras av att *leveransen är fullständig* både med avseende på objekt och variabler. Många administrativa variabler anges endast för de objekt som berörs, om värdet är noll, så finns ingen administrativ uppgift. Ofullständig variabelleverans undersöks genom att granska tabeller där summor eller frekvenser jämförs med föregående års värden. Summor/frekvenser beräknas för lämpliga grupper, t.ex. kommuner, så att ofullständiga leveranser upptäcks.

Kontrolluppgiftsregistret:

Första steget i granskningen är en avstämning av att antalet mottagna kontrolluppgifter stämmer med de som sända av Skatteverket.

Inkomstregistret:

T.ex. så brukar uppgifter om socialbidrag saknas för några kommuner varje år. Inom Inkomstregistret måste man dokumentera vilka kommuner som inte lämnat uppgifter och göra imputeringar för dessa saknade värden. Föregående års socialbidrag brukar då användas.

5. Granskning skall göras för att *kontrollera variabelvärden* så att uppenbara och misstänkta fel upptäcks och åtgärdas. Missuppfattade postbeskrivningar och tekniska fel ger upphov till *uppenbara fel* som om de upptäcks enkelt kan rättas.

Inkomstregistret:

Dessa postbeskrivningar ändras som regel varje år, variabelnamn kan ha ändrats och nya variabler kan ha tillkommit. Postbeskrivningarna kan missförstås, vilket kommer att generera fel. Det är därför viktigt med en nära kontakt med personer inom den registerlämnande myndigheten som kan ge de förklaringar som behövs.

Därefter undersöks extremvärden på de kvantitativa variablerna. Man jämför då med föregående års värden både på aggregerad nivå och på individnivå.

I nästa steg görs logiska kontroller av att summor verkligen är summan av sina delar. Felorsaken kan vara missuppfattningar av postbeskrivningen, eller att extremvärden beror på någon taxeringsåtgärd som inte bör påverka statistiken. Felfrekvensen är så låg att man kan göra individuella rättelser i samråd med den administrativa myndigheten.

Kontroll av färdiga skattningar: I detta steg bildar man alla viktiga tabeller med hela registret som bas. Skattningarna granskas och jämförs med föregående års värden.

Kontrolluppgiftsregistret:

I det femte steget granskas avvikelser genom 16 olika rimlighetskontroller. Relationen mellan löneinkomst och skatt används på flera sätt och dessutom söker man poster med extremt hög löneinkomst eller extremt hög skatt. Ca 5 000 poster med extrema värden upptäcks i dessa kontroller. Dessa granskas översiktligt och endast ett fåtal kontrolleras med SKV. Efter denna kontroll kan kontrolluppgiften accepteras, ersättas med en ny uppgift eller tas bort ur registret.

Detaljerade tabeller med sysselsatta efter näringsgren och kommun tas fram och jämförs med föregående års tabeller. Avvikelsena granskas och granskningsresultaten dokumenteras.

Företagsstatistiken:

I fjärde steget kontrolleras och kompletteras SRU-uppgifterna innan de importeras till det nya registret. Fyra summakontroller av resultat- och balansräkningarnas totaler görs. Mellan 6 000 och 8 000 maskinella rättningar görs, som regel ändras tecken, men även vissa fält som saknar värden fylls i automatiskt.

Investeringar och förändring i eget kapital beräknas. Dessa jämförs med föregående års värden. Antal helårsanställda matchas på från en extern källa som säljer årsredovisningsdata för aktiebolag. För de företag som inte finns med i denna källa imputeras värden för antal årsanställda.

6. Granskning skall *dokumenteras*, dels vilken metodik som används, dels omfattningen av olika feltyper och hur de har åtgärdats. Den registerlämnande myndigheten bör få feedback när fel upptäcks. Granskningsresultat dokumenteras både genom fel- och rättelsekoder i det färdiga registret och med en beskrivning av granskningens effekter och omfattning.

Kontrolluppgiftsregistret:

Detaljerade tabeller med sysselsatta efter näringsgren och kommun tas fram och jämförs med föregående års tabeller. Avvikelserna granskas och granskningsresultaten dokumenteras. Denna dokumentation är mycket värdefull då många användare hör av sig och ifrågasätter resultaten. Då finns dokumentationen, "vi har kollat och det stämmer så långt vi kan bedöma!"

FRIDA:

För varje felaktig summa skapas en särskild variabel där differensen sparas. Varje rättning som görs tilldelas en rättningskod. Poster som innehåller fel som inte kan rättas tilldelas en felkod. Registret finns dokumenterat, senast i SCB (2002b), Dokumentation av FRIDA – inkomståret 2000. Dokumentationen är även ett intressant exempel på hur ett register baserat på administrativa data kan dokumenteras, granskningsmetodiken beskrivs och de administrativa blankettyperna ingår i rapporten.

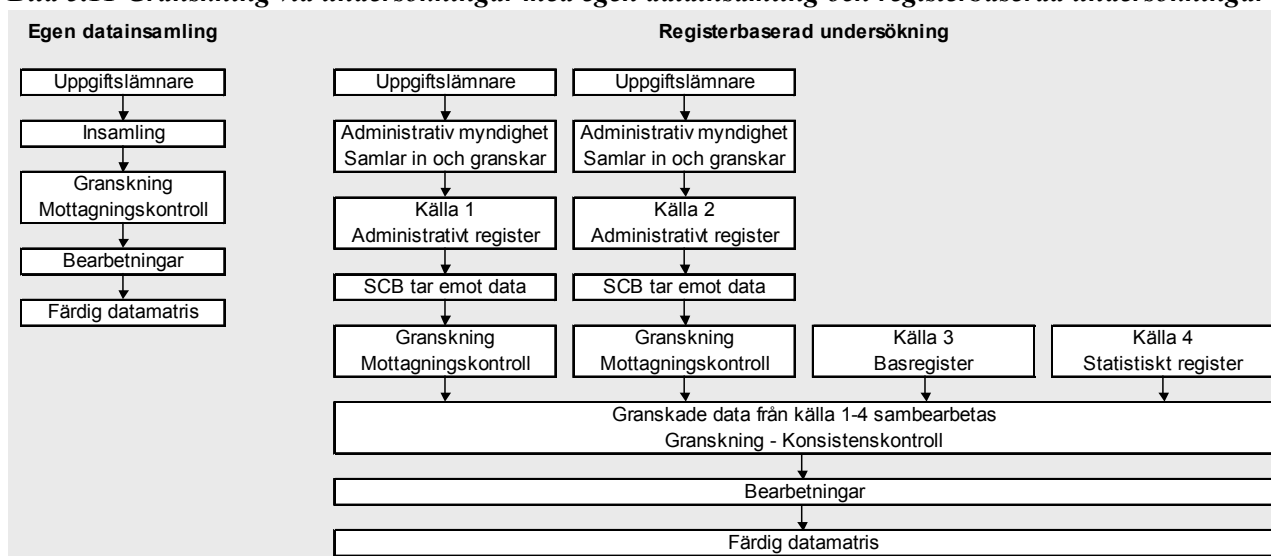
5.7.6 Vad lär vi oss mer av dessa exempel?

Exemplen ovan visar att de administrativa data som tas emot inom SCB innehåller fel som kan kräva granskning på mikronivå. Har man väl upptäckt dessa fel så kan de vara enkla att rätta. Vilka krav granskningen skall uppfylla beror på hur registret skall användas. För kvalificerade analysbehov så måste kvaliteten på mikronivå vara högre än då endast enkla tabeller tas fram, detta ställer höga krav på granskningen. Det är framförallt för longitudinella studier som det ställs höga krav då sambanden på mikronivå måste vara rimliga.

Egen datainsamling och registerbaserade undersökningar

Undersökningar med egen datainsamling har ett granskningsmoment som innebär mottagningskontroll av insamlade data. Det går att kontakta uppgiftslämnarna för att korrigera orimliga uppgifter. Om granskningen kräver stora resurser kan detta vara ett tecken på att frågeformuläret behöver omarbetas.

Vid registerbaserade undersökningar granskas varje administrativ källa genom en mottagningskontroll. Men dessutom finns ett mer kvalificerat moment där alla källor samgranskas för att söka efter ytterligare fel och inkonsistenser. Denna konsistenskontroll är ett viktigt moment som saknas vid undersökningar med egen datainsamling.

Bild 5.11 Granskning vid undersökningar med egen datainsamling och registerbaserad undersökningar

Ämneskompetens och leverantörskontakter

En genomgående slutsats är att ämneskompetensen har en stor betydelse för granskningens effektivitet. Vid undersökningar med egen datainsamling räcker det att kunna den egna undersökningen som ändras sällan. Vid registerbaserade undersökningar måste man kunna det administrativa system som genererar data. Ett sådant system kan innehålla många komplicerade variabler som ofta ändras.

Granskningsarbetet har inte endast direkta effekter på registrets kvalitet. Om arbetet organiseras så att flera personer delar ansvaret, så bidrar granskningen till samverkan och erfarenhetsutbyte inom arbetsgruppen. Därigenom ökar ämneskompetensen och indirekt även registrets kvalitet. När man sedan dokumenterar arbetet och de åtgärder som gjorts för att korrigera data, så befästs ämneskompetensen ytterligare.

Metoden med att arbeta ingående med ett urval från ett stort register bör kunna användas generellt. För IoT-registret är det naturligt att arbeta med FASIT-urvalet, men även andra register bör kunna använda denna metod för att testa kvaliteten och lära känna datamaterialet genom att arbeta med dataanalys på ett sätt som är omöjligt med det totala registret.

Exemplen visar också betydelsen av samarbete och kompetensutveckling inom den arbetsgrupp som tar emot administrativa register och att man har goda kontakter med de myndigheter som lämnar data. Dessutom måste man inom SCB lära sig att samverka mellan produkter och enheter så att de administrativa datamaterialen utnyttjas effektivt.

Genom att personalen vid SCB ”lever med grundmaterialet på mikronivå” så pågår en lärprocess som leder till högre ämneskompetens. Denna lärprocess förstärks om man har nära kontakt med användare.

Dubbelarbete och erfarenhetsutbyte

Exemplen visar att dubbelarbete förekommer, granskningen av SRU-data skulle kunna ske i samverkan mellan flera enheter inom SCB. Även kring fastighets-taxeringsdata skulle samverkan kunna ske. Eftersom granskning av registerdata ännu inte uppmärksammas sker inte det viktiga erfarenhetsutbyte som är nödvändigt för att denna del av det registerstatistiska metodarbetet skall förbättras.

Gransknings- och rättningsprogram

Datoriserade granskningsrutiner bör utvecklas som upptäcker och åtgärdar fel i våra stora och ofta överskådliga register. Det som behövs är utveckling av metoder för att granska och åtgärda mikrodata i stora register – de metoder som utvecklats för att makrogranska urvalsundersökningar kanske inte räcker för registerstatistikens behov. FRIDA-exemplet kan här användas som inspirationskälla. Det gransknings- och rättningsprogram som används för FRIDA-registret bygger på ett mycket stort ämneskunnande som byggts upp gradvis.

Inom EU-projektet Euredit har metoder för helautomatisk granskning och imputering utvecklats, flera av dessa metoder utnyttjar neurala nätverk. Om denna typ av programvara kan användas för att granska administrativa data återstår att undersöka. Automatiken får aldrig hämma uppbyggnaden av ämneskompetens.

Basregister – lämplig version saknas

Alla de fyra fall som beskrivs ovan behöver årgångsversioner av Befolknings- respektive Företagsregistret för att avgränsa sina registerpopulationer. Företagsstatistikens registerpopulation har både över- och undertäckning då den bygger på novemberversionen av Företagsregistret trots att SRU-data levereras ca 10 månader senare. Enligt de principer, som beskrivs i avsnitt 5.4, så bör man försöka skapa en så bra årgångsversion som möjligt med alla källor som är tillgängliga när SRU-data har levererats. Årgångsversionen skall innehålla alla företag som bedrivit verksamhet någon gång under referensåret.

Kompletterande datainsamling kan behövas

När en variabel i det administrativa materialet visar sig ha för låg kvalitet för att kunna användas för statistiska ändamål kan man vara tvungen att göra en kompletterande datainsamling för att få tillräckligt hög kvalitet. Kompletteringsarbetet inom Kontrolluppgiftsregistret är ett exempel på detta.

Granskningen är underlag till kvalitetsdeklarationen

Om registrets personal ”lever med grundmaterialet på mikronivå” får de en god bild av registrets kvalitet och denna kunskap är en god grund för arbetet med att skriva kvalitetsdeklarationer.

5.8 Att bilda härledda variabler med modeller

Härledda variabler diskuteras i avsnitt 4.5.3 där fyra typer av härledda variabler beskrivs:

1. Härledning genom gruppering eller klassindelning av värden.
2. Härledning av en (statistisk) variabel genom beräkningar och logiska operationer med flera (administrativa) variabler som finns i datamatrixen.
Det är denna typ av härledda variabler som diskuteras i detta avsnitt.
3. Härledning genom adjungering av en variabel i ett annat register som avser en annan objekttyp. T.ex. näringsgren för en förvärvsarbetande person bildas genom att ange näringsgren för de arbetsställen där personen är anställd. En adjungerad variabel kan vara flervärd.
4. Härledning genom aggregering av en variabel i ett annat register. T.ex. hushållsinkomst bildas genom att summera värden på variabeln personinkomst för alla de som tillhör ett visst hushåll.

Med hjälp av ett antal administrativa variabler vill man bilda en statistiskt meningsfull variabel. I detta avsnitt betecknar vi den härledda variabeln med y och de administrativa variablerna med $x_1, x_2, \dots$. Vi skiljer på situationer då y är en kvalitativ variabel och då y är en kvantitativ variabel och på exakt beräknade variabelvärden och variabelvärden skattade med en modell.

Att härleda variabelvärden genom beräkningar är besläktat med att *imputera* variabelvärden. Skillnaden är att en härledd variabel skapas för alla objekt i ett register, medan imputerade variabelvärden endast bildas för de objekt i registret där värden saknas. I avsnitt 5.8.4 jämförs härledning och imputering. Diskussionen nedan är även av intresse när man skall beräkna imputerade värden.

5.8.1 Exakt beräkning av värden för en härledd variabel med en regel

I många situationer är det lämpligt att bilda en ny variabel med en regel, som skall vara välmotiverad och dokumenterad.

Exempel, kvalitativ variabel:

I Befolkningsregistret skapas variabeln utrikes eller inrikes född y , med hjälp av fyra variabler:

Bild 5.12 Klassificering av utrikes och inrikes födda

y , utrikes eller inrikes född, kod:	x_1 , personens födelseland	x_2 , personens bosättningsstid	x_3 , faderns födelseland	x_4 , moderns födelseland
1.1 utrikes född med bosättningsstid 0-4 år	Utrikes född	0 - 4 år	-	-
1.2 utrikes född med bosättningsstid 5- år	Utrikes född	5 år och mer	-	-
2.1 inrikes född med två utrikesfödda föräldrar	Född i Sverige	-	Utrikes född	Utrikes född
2.2 inrikes född med en inrikes och en utrikes född förälder	Född i Sverige	-	Utrikes född Född i Sverige	Född i Sverige Utrikes född
2.3 inrikes född med två inrikesfödda föräldrar	Född i Sverige	-	Född i Sverige	Född i Sverige

Exempel, kvantitativ variabel:

Disponibel inkomst y , beräknas med en regel som anger vilka inkomstvariabler x_{i1} , x_{i2} , ... som skall summeras och vilka skatter x_{s1} , x_{s2} , ... som skall dras bort.

I båda exemplen ovan bildar man regler för hur den härledda variabeln skall definieras. Dessa regler baseras på ämneskunnande och bedömningar. Beräkningarna är exakta, dvs. om x -variablerna inte har några mätfel så har inte y -variabeln något mätfel.

5.8.2 Skatta värden för en härledd variabel med en regel

Även i de två exemplen nedan används regler för hur den härledda variabeln skall definieras. De regler som används är modeller som är helt baserade på ämneskunnande. Men beräkningarna är här inte exakta, dvs. även om x -variablerna inte har mätfel så kan y -variabeln innehålla fel. Dessa fel i y -variabeln kallas *modellfel*. I avsnitt 5.8.3 diskuteras modeller som baseras på sambandsanalys.

Exempel: Yrke inom statlig sektor, prioritering av källor

De statliga arbetsgivarnas personalregister innehåller administrativa variabler som används av SCB för att klassificera de anställda efter yrke enligt standard för svensk yrkesklassificering, SSYK. Variablerna *Tjänstebenämning* och *TNS-kod* har använts enligt vissa regler. Dessa regler ändrades från och med år 2000. De gamla och nya reglerna har följande utseende där man *prioriterar* mellan de administrativa variablerna på olika sätt:

Bild 5.13 Klassificering av yrke inom statlig sektor

T.o.m. år 2000	Fr.o.m. år 2001
<u>1:a hand:</u> (ca 90 % av fallen) Användbar Tjänstebenenämning finns och används. <i>Exempel:</i> Tjänstebenenämning: 93460 "webbredaktör" blir SSYK: 2451 "journalister, författare, informatörer m fl" <u>2:a hand:</u> (ca 8 % av fallen) Användbar Tjänstebenenämning saknas men TNS-kod finns och används. <i>Exempel:</i> TNS: 1316 "skriver, testar och dokumenterar program" blir SSYK: 3121 "datatekniker" <u>3:e hand:</u> (ca 2 % av fallen) Varken användbar Tjänstebenenämning eller TNS-kod finns. Blir bortfall, SSYK saknas	<u>1:a hand:</u> (ca 48 % av fallen) TNS-kod finns och används. <i>Exempel:</i> TNS: 1316 "skriver, testar och dokumenterar program" blir SSYK: 3121 "datatekniker" <u>2:a hand:</u> (ca 51 % av fallen) TNS-kod saknas, men användbar Tjänstebenenämning finns och används. <i>Exempel:</i> Tjänstebenenämning: 93460 "webbredaktör" blir SSYK: 2451 "journalister, författare, informatörer m fl" <u>3:e hand:</u> (ca 1 % av fallen) Varken TNS-kod eller användbar Tjänstebenenämning finns. Blir bortfall, SSYK saknas

Även om *TNS-kod* och *Tjänstebenenämning* är korrekta så kan yrket enligt SSYK ibland bli felaktigt. Det finns alltså inget exakt samband mellan de två administrativa variablerna och verkligt yrke hos en person.

Vi ser dels att regler inte är självklara, de administrativa variablerna kan användas enligt olika principer, och dels att tidsseriebrott kan uppstå. Trots att SSYK är densamma, uppstår tidsseriebrott då man förbättrar reglerna. Om kvaliteten på TNS-koden ökar så är det lämpligt att prioritera denna variabel.

På motsvarande sätt finns regler som översätter kommunernas, landstingens och de privata arbetsgivarnas koder till SSYK.

Exempel: Sysselsättningsstatus i Sysselsättningsregistret 1985 - 1992.

I den första versionen av Sysselsättningsregistret (1985) var avsikten att med kontrolluppgiftsdata mäta sysselsättning som i en traditionell folkräkning – minst en timmes förvärvsarbete under mätveckan i november. Kontrolluppgifterna med uppgifter om alla transaktioner mellan arbetsgivare och anställda tolkades med följande regler:

- Kontrolluppgift för en del av ett år, där november ingick: Om total inkomst per månad (inklusive sjukersättning) var större än 200 SEK klassificerades personen som förvärvsarbetande i november.
- Kontrolluppgift för helt år: Om total årsinkomst var större än 21 800 SEK klassificerades personen som förvärvsarbetande i november.
- Om inte något av villkoren ovan var uppfyllda klassificerades personen som ej förvärvsarbetande i november.
- Dessutom fanns särskilda regler för sjömän och egna företagare.

Även om alla variabler på kontrolluppgifterna är korrekta så kan regeln ovan ge felklassificeringar. Dels kan personer, som är förvärvsarbetande enligt FoB-definitionen, bli felaktigt klassificerade som ej förvärvsarbetande och dels kan ej förvärvsarbetande bli klassificerade som förvärvsarbetande. En evalverings-

studie genomfördes och man bedömde att kvaliteten hos den härledda sysselsättningsvariabeln var godtagbar. Mellan 1986-1992 användes dessa regler, men med inkomstgränser som justerades varje år med ett löneindex.

Den statistiska variabeln *sysselsättningsstatus* bygger här på tre administrativa variabler: Tidsmarkeringen på kontrolluppgiften (första och sista månad i den period som anställningen avser), lönesumma och sjukersättning. Regeln säger hur dessa variabler skall tolkas för att en person skall klassificeras som förvärvsarbetande respektive ej förvärvsarbetande i november.

5.8.3 Skatta värden för en härledd variabel med en sambandsmodell

I båda exemplen i avsnitt 5.8.2 bildar man regler för hur den härledda variabeln skall definieras. Dessa regler baseras på ämneskunnande och bedömningar. Ett alternativ till att använda regler är att studera samband mellan den sökta variabeln y och de administrativa variablerna (som vi här kallar för $x_1, x_2, \dots$) som kan tänkas ha samband med denna y -variabel. Med hjälp av dessa samband skapas därefter den härledda variabeln.

När vi skapar härledda variabler med *sambandsmodeller* arbetar vi i två steg med olika datamatriser:

1. Ett *testdatamaterial* från t.ex. en urvalsundersökning innehåller både y -variabeln och x -variablerna. Med denna datamatrix bygger man först en sambandsmodell som anger hur man bäst skall skatta y för givna värden på x -variablerna.
2. Modellen används sedan på *registrets datamatrix*, där endast x -variablerna finns. Med den skattade sambandsmodellen beräknas för varje objekt i registret ett y -värde med hjälp av objektets x -värden.

Fördelen med en sambandsmodell, jämfört med en regel baserad på ämneskunnande och bedömningar, är att en bra sambandsmodell visar hur man kan använda många administrativa variabler på bästa möjliga sätt. Modellen kan innehålla många variabler till skillnad från en regel baserad på ämneskunskap. Nackdelen är att man måste generalisera analysresultaten från testdatamaterialet till registrets datamatrix – att modellen är bra för testdatamaterialet behöver inte innebära att den också är bra för registrets datamatrix.

Kvantitativa härledda variabler

När y är en kvantitativ variabel finns det tre olika typer av modeller som kan användas:

- *Gruppvisa medelvärden*: Man studerar sambandet mellan y -variabeln och x -variablerna i en tabell med y -medelvärden baserad på testdatamaterialet.
- *Kvot*: Om y kan antas vara proportionell mot en x -variabel kan den skattade kvoten $\Sigma y / \Sigma x$ från testdatamaterialet användas. Olika kvoter kan beräknas för olika grupper av objekt.
- *Regressionsmodell*: Om man kan bygga en regressionsmodell $\hat{y} = f(x_1, x_2, \dots)$ med testdatamaterialet kan funktionen $f(x_1, x_2, \dots)$ användas.

Exempel: Industrins energianvändning – gruppvisa kvoter

Bland arbetsställen med 10-49 anställda görs en urvalsundersökning. För urvalet som består av ca 1 800 arbetsställen känner man därefter antal anställda, SNI och energiförbrukning. Bland arbetsställen med *färre än 10 anställda* (ca 50 000 st.) görs inget urval utan man antar att inom varje bransch förbrukas lika mycket energi per anställd, som bland företag med 10 – 49 anställda.

Urvalet av 1 800 företag är här testdatamaterialet som används för att beräkna kvoter $\Sigma y / \Sigma x$ = energiförbrukning/antal anställda för olika SNI. Dessa kvoter används genom att för alla arbetsställen med färre än 10 anställda beräknas skattad energiförbrukning = skattad kvot • arbetsställets antal anställda.

En nackdel med detta är en modell som ser bra ut för företag med 10-49 anställda inte behöver vara lämplig för de mycket små företagen.

Kvalitativa härledda variabler

Exemplet från Sysselsättningsregistret ovan väcker frågan om regeln verkligen använder de administrativa variablerna på bästa sätt. Om t.ex. tidsmarkeringen på kontrolluppgiften har låg kvalitet, bör man då ha en regel som är starkt beroende av denna variabel? Genom att analysera ett testdatamaterial kan man studera sambandet mellan förvärvsarbete och kontrolluppgifternas variabler.

Om y är en kvalitativ variabel kan följande typer av modeller användas:

- *Regressionsmodell*: Om y endast har två klasser så kan sambandet undersökas med en regressionsmodell $\hat{y} = f(x_1, x_2, \dots)$.
- *Diskriminantanalys*: Om y har fler än två klasser kan en diskriminantanalysmodell användas för att undersöka hur x -variablerna bäst kan användas för att klassificera objekten i olika y -klasser.
- *Data mining-modeller*: Programvara för data mining kan även användas.

Exempel: Sysselsättningsstatus i Sysselsättningsregistret från och med 1993.

I den första versionen av Sysselsättningsregistret användes följande regler:

- Kontrolluppgift för en del av ett år, där november ingick: Om total inkomst per månad (inklusive sjukersättning) var större än 200 SEK klassificerades personen som förvärvsarbetande i november.
- Kontrolluppgift för helt år: Om total årsinkomst var större än 21 800 SEK klassificerades personen som förvärvsarbetande i november.

Men denna regelbaserade metodik har allvarliga nackdelar. Unga och gamla klassificeras med samma inkomstgränser trots att lönenivåerna skiljer sig åt. Inkomstmönster, fördelningen mellan fasta och tillfälliga anställningar samt strukturen på kontrolluppgiftsdata förändras över tiden. Även sjukersättnings-systemet förändras och allt detta innebär att statistik från olika år inte är jämförbar, trots att samma regler använts.

För att lösa jämförbarhetsproblemen ovan infördes en härledd variabel i 1993 års version av Sysselsättningsregistret. För de personer som deltog i Arbetskraftsundersökningen (AKU) i november 1993 kombinerades kontrolluppgiftsdata med sysselsättningsstatus enligt AKU. Med regressionsanalys byggdes

Metodiken beskrivs i
Sandgren m.fl. (1994)

sambandsmodeller för olika kombinationer av kön och ålder för dessa testdata. På detta sätt får olika grupper olika inkomstgränser, men alla gränser motsvarar sysselsättningsdefinitionen i AKU. En annan fördel är att administrativa variabler med hög kvalitet får stor inverkan på klassifikationen medan variabler med låg kvalitet får liten inverkan. I stora drag görs analysen på följande sätt:

1. Med testdatamaterialet med känd sysselsättningsstatus enligt AKU skattas regressionsmodeller, där sysselsättningsstatus enligt AKU är y -variabel med två kategorier (förvärvsarbetande/ej förvärvsarbetande) och kontrolluppgiftsvariablerna är regressionsmodellens x -variabler. Materialet delas upp i delgrupper som har samma typ av kontrolluppgift, åldersklass och kön. För varje delgrupp görs separata analyser.
2. Med hjälp av den skattade modellen för en delgrupp beräknas anpassade y -värden med x -variablerna från kontrolluppgifterna. De som enligt AKU är förvärvsarbetande kommer ha anpassade värden som är markant skilda från de ej förvärvsarbetande om analysen har lyckats.
3. Man bestämmer ett gränsvärde så att de med anpassade y -värden på ena sidan gränsvärdet klassificeras som förvärvsarbetande och de övriga som ej förvärvsarbetande. Gränsen sätts så att antalet klassificerade förvärvsarbetande blir lika stort som motsvarande antal enligt AKU i testmaterialet.
4. Därefter används gränsvärdena för de olika delgrupperna på hela registret så att alla personer i registerpopulationen blir klassificerade med hjälp av de administrativa variablerna på kontrolluppgifterna.

För t.ex. 2001 års version av registret användes 2001 års november-AKU för att ta fram nya inkomstgränser genom nya regressionsanalyser – dessa nya gränser motsvarar samma definition av förvärvsarbetande i AKU som tidigare. På detta sätt är det möjligt att göra relevanta jämförelser mellan olika år.

Kvalitativa härledda variabler som bildats med hjälp av en sambandsmodell skall evalveras så att *klassifikationsfelen* kan bedömas. Med en bra sambandsmodell blir både *nettofelen* och *bruttofelen* små. Tabellerna nedan innehåller en jämförelse mellan den gamla och nya metoden att avgränsa sysselsatta i Sysselsättningsregistret.

Tabell 5C. Klassifikationsfel i Sysselsättningsregistret (SyssReg) 1993

<i>Antal personer i testdatamaterialet</i>	Skattning enl. nya SyssReg			Skattning enl. gamla SyssReg		
	Sysselsatt	Ej sysselsatt	Totalt	Sysselsatt	Ej sysselsatt	Totalt
Sysselsatt AKU	22 360	1 158	23 518	22 472	1 046	23 518
Ej sysselsatt AKU	1 068	6 872	7 940	1 329	6 611	7 940
Totalt	23 428	8 030	31 458	23 801	7 657	31 458

<i>Procent av totala antalet personer</i>	Skattning enl. nya SyssReg			Skattning enl. gamla SyssReg		
	Sysselsatt	Ej sysselsatt	Totalt	Sysselsatt	Ej sysselsatt	Totalt
Sysselsatt AKU	71,1	3,7	74,8	71,4	3,3	74,8
Ej sysselsatt AKU	3,4	21,8	25,2	4,2	21,0	25,2
Totalt	74,5	25,5	100,0	75,7	24,3	100,0

<i>Klassifikationsfel</i>	Nettofel:	74,5 – 74,8 = – 0,3 %	Nettofel:	75,7 – 74,8 = 0,9 %
	Bruttofel:	3,7 + 3,4 = 7,1 %	Bruttofel:	3,3 + 4,2 = 7,5 %

Det skattade *bruttofelet* är en skattning av andelen felklassificerade i hela registret, skattningen av *nettofelet* är en skattning av det systematiska felet hos metoden att avgränsa förvärvsarbetande i Sysselsättningsregistret (om man antar att AKU ger korrekta skattningar). För att få bra skattningar av brutto- och nettofel bör man ha två testdatamaterial, ett som används för att bygga sambandsmodellen och ett annat för att skatta klassifikationsfelen.

5.8.4 Härledda variabler och imputerade variabelvärden

Som vi tidigare påpekat, är arbetet med att härleda variabelvärden genom beräkningar, besläktat med att *imputera* variabelvärden. Skillnaden är att en härledd variabel skapas genom beräkningar för *alla* objekt i ett register, medan imputerade variabelvärden endast beräknas för de objekt i registret där värden saknas. I bilden nedan jämförs dessa två slag av bearbetningar.

Imputering behandlas i avsnitt 6.4.4

Bild 5.14 beskriver ett longitudinellt företagsregister med lönesumma för år 1 och år 2 (*Lsum1* och *Lsum2*), samt antal anställda år 1 och år 2 (*Anst1* och *Anst2*). Vissa värden saknas för antal anställda. Imputerade värden beräknas genom att dividera lönesumman med genomsnittlig lönesumma per anställd. Därefter beräknas två tillväxtmått som härledda variabler. *Ldiff* = 1 om lönesumma ökat och *Adiff* = 1 om antal anställda ökat mellan år 1 och år 2.

Bild 5.14 Imputerade värden och härledda variabler i ett företagsregister

Före bearbetning

Ftgid	Lsum1	Lsum2	Anst1	Anst2
1	12 132	12 344	34	32
2	1 775	1 438	5	4
3	893	914	2	saknas
4	18 923	17 835	53	47
5	239	346	saknas	saknas
6	6 221	7 583	17	20
7	549	514	3	2

Med imputerade värden och härledda variabler

Ftgid	Lsum1	Lsum2	Anst1	Anst2	Ldiff	Adiff
1	12 132	12 344	34	32	1	0
2	1 775	1 438	5	4	0	0
3	893	914	2	2	1	0
4	18 923	17 835	53	47	0	0
5	239	346	1	1	1	0
6	6 221	7 583	17	20	1	1
7	549	514	3	2	0	0

Hur bildas det imputerade värdet för Ftgid 5 för antal anställda år 1?

För de sex företag där värdena är kända är:
Lönesumman = 40 493
Antal anställda = 114

Imputerat värde avrundas
= $239 / (40\,493 / 114) = 1$

5.8.5 Att bilda variabler genom kodning

I vissa fall används uppgifter i textform för att skapa statistiskt användbara variabler. Genom en kodningsprocess överförs information i delvis ostrukturerad text till helt strukturerade variabelvärden.

Arbetsställets belägenhetsadress är länken mellan Fastighets- och Företagsregistret. Det finns flera problem med dessa adresser som för närvarande inte anges på ett standardiserat sätt. T.ex. kan samma gatuadress anges på olika sätt:

Storgatan 17

Storg 17

Storg. 17

dessutom kan felstavningar förekomma: Storgtan 17

Genom att bearbeta adressuppgifter i ett översättningsprogram där faktiska adresser jämförs med alias-listor kan många adressuppgifter överföras till strukturerad form (t.ex. alla varianterna ovan blir Storgatan 17). De adresser som inte kan klaras av med en sådan databearbetning kan sedan kodas manuellt. När både fastighetsadresserna i Fastighetsregistret och belägenhetsadres-

serna i Företagsregistret har samma strukturerade form så kan registren matchas med hjälp av adressvariabeln.

Det finns flera exempel på att viktiga variabler i registersystemet skapas genom kodning. För vissa företag bestäms näringsgren genom att man använder telefonkatalogens Gula sidor. Dödsorsaker och typ av arbetsskada kodas med hjälp av textinformation från administrativa blanketter.

Dödsorsaker

Efter utredning av dödsorsaken utfärdar en läkare ett intyg om dödsorsaken. Dessa intyg skannas och dataregistreras. Diagnosuttrycken på dessa dataregistrerade blanketter kodas med hjälp av en särskild programvara som har utvecklats vid SCB. Programvaran omfattar automatisk kodning av diagnoser från klartext till en kod enligt en internationell statistisk dödsorsaksklassifikation. Osäkra fall sorteras fram och kodas manuellt efter kontakt med läkare.

Arbetsskador

Anmälan av arbetsskador görs enligt en särskild blankett som arbetsgivaren är skyldig att skicka till försäkringskassan. Den fria beskrivningen av arbetsskadan kodas av den statistikansvariga myndigheten Arbetsmiljöverket till ett antal statistiska variabler som *händelse*, *huvudsaklig yttre faktor*, *misstänkt orsak* och *diagnos*. Kodningsinstruktioner, utbildning av kodare samt kontrollkodning är viktiga moment för att få kvalitet på kodningsresultatet.

5.9 Att skapa longitudinella register

Vad skall ett statistiskt register användas till? Registrets användning avgör vilka kvalitetskrav som måste uppfyllas och detta blir styrande för vilken typ av granskning som behövs och vilka registerbearbetningar som måste göras.

Ett register som endast används för att producera årlig officiell statistik skall ge skattningar av god kvalitet på aggregerad nivå. Men om registret även skall användas för longitudinella analyser måste *objekten* avgränsas så att de kan följas över tiden utan att statistiskt ointressanta förändringar stör.

Exempel: Om ett företag byter juridisk form från handelsbolag till aktiebolag skall detta inte uppfattas som att ett företag upphör och ett helt nytt företag bildas. En ny statistikprodukt inom SCB, *Företagens och arbetsställets dynamik* (FAD) syftar till att ge en mer nyanserad bild av näringslivets strukturella förändringar och förbättra klassificeringen av händelser inom företag som ger nya möjligheter att t.ex. studera arbetskraftens rörlighet.

Primärt skapas stabila identiteter för företagen vilket ger möjlighet att följa dem över tiden på ett mer analytiskt sätt än vad man kan göra med Företagsregistret. Med FAD-registret kartläggs samtidigt nybildningar, nedläggningar, uppdelningar och sammanslagning av företag. Som indikator på förändringar tas främst hänsyn till företagets bestånd av personal vid olika tidpunkter medan mindre vikt läggs vid förändringar av ägare, näringsgren och belägenhet.

Grundregeln i FAD-registret är att om en majoritet av de sysselsatta år 1, också utgör en majoritet av de sysselsatta år 2, betraktas enheterna för de två åren som

Bolin m.fl. (1996)
redovisar metoder för
att utvärdera och
förbättra kodbarheten i
Dödsorsaksregistret

samma företag oavsett organisationsnummer. Genom att på motsvarande sätt ställa upp andra regler för personalflöden mellan två år, kan man klassificera hopslagningar och uppdelningar av företag.

På samma sätt måste även *variablerna* definieras så att förändringar över tiden blir statistiskt meningsfulla. När administrativa regler förändras på t.ex. skatteområdet blir inte de variabler som berörs jämförbara över tiden.

Exempel: Reglerna för beskattning av fåmansbolagens redovisade vinst har förändrats flera gånger. Reglerna påverkar om bolagens ägare tar ut överskottet som vinst eller som lön. Statistik som innehåller lönesummor och driftsöverskott påverkas av detta. De administrativa variablerna lön till företagsägare och redovisad vinst bör då läggas ihop till en meningsfull statistisk variabel för dessa fåmansbolag.

Även de variabler som bildas inom SCB kan vara mer eller mindre lämpliga för longitudinella analyser. Den metod som används för att bilda variabeln *förvärvsarbetande i november* i Sysselsättningsregistret, vilket beskrivs i avsnitt 5.8.3 ovan, kan ge skenbara förändringar för personer som ligger nära de inkomstgränser vilka används i den sambandsmodell den härledda variabeln bygger på.

Tabell 5D. Skenbara förändringar i sysselsättningsstatus för en person

	År 1	År 2	År 3	År 4	År 5
Personens inkomst respektive år	47 400:–	48 500:–	49 600:–	52 800:–	53 900:–
Gräns i modellen respektive år	47 000:–	48 600:–	49 400:–	52 100:–	54 200:–
Sysselsättningsstatus, skattat värde	Syss	Ej syss	Syss	Syss	Ej syss

Med denna typ av härledd variabel i Sysselsättningsregistret kan man få skattningar på aggregerad nivå som medger jämförelser över tiden som har god kvalitet. Men om samma variabel används för longitudinella analyser, där förändring i sysselsättningsstatus undersöks för grupper av personer, kan skenbara förändringar medföra att kvaliteten inte är tillräckligt hög. Andelen sysselsatta kan skattas med god kvalitet, men andelen personer som ändrar sysselsättningsstatus kanske inte kan skattas med tillräckligt god kvalitet.

De kvalificerade användningarna av statistiska register, som longitudinella analyser, simulering- och prognosmodeller, ställer högre krav på granskning och bearbetning. Reglerna för att avgränsa och kontrollera objekt måste anpassas till användningens krav. Dessutom behöver registret kompletteras med en rad nya variabler som ofta kräver omfattande bearbetningar. Dessa bearbetningar avser olika statistiska mått, nya klassificeringar och nya härledda variabler. Dessutom kompletteras registret ofta med variabler som importeras från andra register.

De bearbetningar som görs i samband med kvalificerade användningar måste dokumenteras utförligt. Detta gäller även bearbetningar som görs för ett enstaka uppdrag åt forskare. Denna dokumentation behövs för metoddiskussioner och för att ge idéer till framtida uppdrag och andra SCB-produkter.

Bearbetningarna i kapitel 5 syftar till att skapa den eller de datamatriser som ingår i registret

Kapitel 6. Beräkningsmetoder

Efter att ha gjort de bearbetningar som beskrivs i kapitel 5 är registrets datamatrix eller datamatriser färdiga. Nästa steg är att med hjälp av den färdiga datamatrixen bilda de statistiska tabeller som är relevanta för problemställningen. I detta kapitel beskriver vi beräkningsmetoder som används och skulle kunna användas för att bilda skattningar och tabeller. Vi diskuterar några kvalitetsproblem och ger förslag på lösningar av dessa problem, som bygger på speciella beräknings- eller skattningsmetoder. Vissa av dessa skattningsmetoder bygger på att vikter används även för registerbaserad statistik, på ett liknande sätt som vid urvalsundersökningar. Sådana vikter är en speciell typ av härledda variabler.

När man använder datamatrixen för att skapa statistiska tabeller skall tabellcellerna innehålla antal, summor eller andra statistiska mått. När vikter används vid skattningarna beräknas vägda antal eller summor.

Bild 6.1 Samband mellan undersökningens olika faser och bokens kapitel

Kapitel 5	Problemformulering ↓ Datamatrix/-er	Data från olika källor skall sambearbetas genom objekt- och variabelselektioner och matchningar. Data skall granskas och härledda variabler bildas.
Kapitel 6	Datamatrix/-er Råtabeller med skattningar i tabellcellerna Färdiga tabeller	Vid registerbaserade undersökningar och totalundersökningar kan man bilda detaljerade råtabeller Här brukar man endast göra enkla summeringar, men uppgifterna i tabellens celler bör betraktas som skattningar som skall göras på bästa möjliga sätt. De stora råtabellerna behöver ofta bearbetas med t.ex. standardvägningar till mer överskådliga tabeller
Ej i denna bok	Färdiga tabeller ↓ Diagram och text	Den färdiga presentationen skall belysa frågorna i problemformuleringen. Presentationen består av lättlästa tabeller, diagram, kartor och text. Varje diagram eller karta motsvaras av en viss tabell.

I avsnitt 6.2 till 6.5 beskriver vi skattningsmetoder som kan användas för *flervärda variabler*, vid problem med *övertäckning*, *partiellt bortfall* och *tidsseriebrott*. Idag brukar inte dessa kvalitetsproblem åtgärdas i samband med att registerstatistik framställs, men med metoderna i detta kapitel skulle man kunna korrigera för dessa felkällor så att felen minskar.

Eftersom registren i registersystemet samverkar så påverkar t.ex. bortfallet i ett visst register även andra register som importerar data från detta register. Även sättet att korrigera för det partiella bortfallet påverkar andra register i systemet. De metoder vi föreslår måste även kunna fungera inom systemet så att statistik från olika register blir konsistent.

Metoderna som beskrivs illustreras med konkreta exempel. Dessa exempel måste vi av pedagogiska skäl renodla så att de allmänna principerna framgår. I verkliga situationer skulle kanske tillämpningen av metodiken vara annorlunda än i exemplen. Vissa exempel bygger på verkliga SCB-data, medan andra exempel bygger på små fiktiva datamaterial.

6.1 Skattningsmetoder i survey- och registerstatistik

Termen *skattning* används allmänt för urvalsundersökningar, men bör även användas inom registerstatistiken. Även här måste man skilja på det verkliga värdet i målpopulationen och den skattning som registret ger.

6.1.1 Skattningsmetoder som använder vikter

Vid urvalsundersökningar bildas skattningar med formel (1) nedan. Design-vikterna d_i beror på hur urvalet har fördelats på olika strata och vikterna g_i i formel (1) är baserade på hjälpvariabler från statistiska register och används för att minska urvalsfelen och de fel som orsakas av bortfallet. Lundström och Särndal (2001) behandlar denna typ av skattnings- eller estimationsmetoder. För varje cell i en tabell beräknas summor för kvantitativa variabler t.ex. lön eller omsättning med formel (1). Om antalen i tabellens celler skall beräknas, sätter man variabeln $y = 1$ för alla observationer och får då det vägda antalet i cellen. Vid beräkning av medelvärden divideras den vägda summan med det vägda antalet.

$$\hat{Y} = \sum_{i=1}^r d_i g_i y_i \quad \begin{array}{l} r \text{ är antal objekt i urvalet som svarat i en viss cell} \\ \text{(alla formler nedan avser en viss cell i en tabell)} \end{array} \quad (1)$$

För totalundersökningar, där SCB samlar in data, uppstår bortfall på motsvarande sätt som vid urvalsundersökningar. Med de metoder som beskrivs av Lundström och Särndal (2001) kan man även här beräkna vikter g_i för att minska de fel som orsakas av bortfallet.

$$\hat{Y} = \sum_{i=1}^R g_i y_i \quad \begin{array}{l} R \text{ är antal objekt i totalundersökningen som svarat i} \\ \text{en viss tabellcell med värdena } y_1, y_2, \dots, y_R \end{array} \quad (2)$$

Även för registerbaserade undersökningar utan egen datainsamling kan motsvarigheten till partiellt bortfall uppstå dvs. vissa variabelvärden saknas. Även här kan man kompensera för sådant bortfall genom att använda vikter, vilket diskuteras i avsnitt 6.4.

6.1.2 Skattningsmetoder för registerstatistik

Vi kommer här att diskutera två frågor:

- Finns det skattningsmetoder inom registerstatistiken?
- Kan god kvalitet uppnås med skattningsmetoder där vikter används?

Finns det skattningsmetoder inom registerstatistiken?

Idag används inte några speciella metoder när registerstatistik produceras utan beräkningar och summeringar görs på enklast möjliga sätt.

$$\hat{Y} = \sum_{i=1}^R y_i \quad R \text{ är antalet objekt i registret i en viss tabellcell} \quad (3)$$

Vi uppfattar dessa skenbart enkla summeringar som skattningar och värdena på dessa skattningar beror på de metoder som användes när registret bildades. Både i kapitel 3 och 5 beskrivs det arbete som görs för att bilda ett statistiskt register. Om detta arbete görs på olika sätt får man olika numeriska värden på

den registerstatistik som framställs med det färdiga registret. Att välja metodik för arbetet med att skapa ett register innebär att man väljer skattningsmetodik.

Vid urvalsundersökningar ägnas stora delar av metodarbetet på *hur* man skall summera, dvs. hur vikterna d_i och g_i skall bestämmas. Metodarbetet inom registerstatistiken läggs i stället ner på *vad* som skall summeras, dvs. hur man definierar registerpopulationen, hur objekten definieras i denna population och hur registrets variabler bildas med tillgängliga data. Det sätt på vilket vi bildar ett statistiskt register avgör vilka skattningar vi kommer att kunna bilda med registret. Det finns alltså skattningsmetoder även inom registerstatistiken.

Kan god kvalitet uppnås med skattningsmetoder där vikter används?

Förutom de skattningsmetoder som bestäms av sättet att skapa registret, inför vi i detta kapitel vikter w_i för att lösa några olika kvalitetsproblem. Vikterna beräknas på olika sätt för olika problem och med vikterna kan man korrigera för olika typer av fel, dvs. att registerskattningarna ligger på en felaktig nivå.

$$\hat{Y} = \sum_{i=1}^R w_i y_i \quad \begin{array}{l} R \text{ är antalet objekt i registret i en viss tabellcell, } w_i \text{ är} \\ \text{vikter. Vid de traditionella metoderna är alla } w_i = 1 \end{array} \quad (4)$$

De typer av fel vi tar upp är: fel som beror på att man bortser från information i flervärda variabler, fel som beror på undertäckning, fel som beror på partiellt bortfall och fel som beror på tidsseriebrott. Metodiken skulle kunna användas för fler typer av fel.

6.1.3 Från datamatrix till statistiska tabeller

Hur används vikter i ett register för att göra skattningar? Detta illustreras med ett fiktivt löneregister i bilden nedan. Registret innehåller uppgifter om personers faktiska månadslön under en viss månad samt uppgift om tjänstgöringens omfattning (Tjomf), där 1 innebär heltid. Den faktiska lönen räknas om till heltidslön med variabeln Tjomf. Personernas yrke anges med yrkeskod enligt SSYK. Yrkesvariabeln kan grupperas till kompetensnivå, där 1 är lägsta nivån och 5 avser chefer. Även variablerna ålder och heltidslön har grupperats. För att korrigera för någon form av fel, t.ex. bortfall, har vikter w_i beräknats. Hur man bildar sådana vikter diskuteras från och med avsnitt 6.2.

Bild 6.2 Ett löneregister där observationerna har vikter

Person (1)	Kön (2)	Ålder (3)	SSYK (4)	Nivå (5)	Lön (6)	Tjomf (7)	LönHeltid (8)	Löneklass (9)	w_i (10)	$w_i \cdot \text{lön}_i$ (11)	$w_i \cdot \text{tjomf}_i$ (12)	$w_i \cdot \text{lönheltid}_i$ (13)
Pnr 1	kvinn	50-54	4190	2	14850	1,00	14850	14-14,9	1,028	15271,4	1,028	15271,4
Pnr 2	kvinn	40-44	2330	4	16630	0,95	17505	17-17,9	1,031	17147,5	0,980	18049,8
Pnr 3	man	50-54	2492	4	17807	1,00	17807	17-17,9	1,083	19285,5	1,083	19285,5
Pnr 4	kvinn	40-44	2330	4	1485	0,09	16500	16-16,9	1,031	1531,2	0,093	17013,5
Pnr 5	kvinn	40-44	5133	2	6497	0,50	12994	12-12,9	1,031	6699,2	0,516	13398,4
Pnr 6	kvinn	40-44	5131	2	14102	1,00	14102	14-14,9	1,031	14540,9	1,031	14540,9
Pnr 7	man	50-54	5131	2	858	0,06	14300	14-14,9	1,083	929,2	0,065	15487,3
***	***	***	***	***	***	***	***	***	***	***	***	***

Registret innehåller kolumnerna (1) – (10), men när skattningarna skall göras bildas kolumnerna (11) – (13) tillfälligt. Råtabellen nedan bildas genom att summera aggregeringsvariablerna i registrets kolumn (10) – (13) för alla kom-

binationer av indelningsvariablerna kön, åldersklass, SSYK (därigenom även nivå) och löneklass. Den fullständiga tabellen består av ca 2 200 rader.

Bild 6.3 En råtabell som kan användas för att bilda många tabeller

Kön	Ålder	SSYK	Nivå	Löneklass	$\sum w_i$	$\sum w_i \cdot \text{lön}_i$	$\sum w_i \cdot \text{tjomf}_i$	$\sum w_i \cdot \text{lönheltid}_i$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
kvinna	17-24	2330	4	12-12,9	42,52	429170	34,55	526165
kvinna	17-24	2330	4	13-13,9	95,67	1293410	95,35	1297704
kvinna	17-24	2330	4	14-14,9	42,52	201399	14,14	622852
kvinna	17-24	2330	4	15-15,9	10,63	159444	10,63	159444
kvinna	17-24	2330	4	16-16,9	53,15	163111	9,89	876942
...	...	...	...	...	...	...	...	...
man	60-64	8320	2	13-13,9	21,24	290107	21,23	290107
man	60-64	8320	2	14-14,9	10,62	149300	10,62	149300
man	60-64	9140	1	12-12,9	10,62	136422	10,62	136422
man	60-64	9140	1	13-13,9	10,62	71348	5,31	142697
man	60-64	9140	1	14-14,9	21,24	308040	21,23	308040

Registerbaserade undersökningar lämpar sig för att bilda stora detaljerade råtabeller som kan användas för att skapa många olika läsbara tabeller för olika syften

Genom att ytterligare aggregera denna råtabell på olika sätt kan flera tabeller för olika syften bildas. Variabeln lön används både som indelningsvariabel (kol 5) och som aggregeringsvariabel (kol 7 och 9) i råtabellen ovan. I en verklig statistisk tabell används lön endast i den ena av dessa roller.

Uttag av frekvenstabeller – lönefördelningen för olika redovisningsgrupper

Genom att summera kolumn (6), som innehåller vägda antal personer, för olika kombinationer av kompetensnivå, kön och löneklass får man tabell 6A nedan. De absoluta frekvenserna i 6A räknas om till relativa frekvenser i tabell 6B som illustrerar sambandet mellan i första hand kön och lön och i andra hand sambandet mellan kompetensnivå och lön.

Tabell 6A. Lönefördelningen efter nivå och kön, antal personer

Nivå:	1		2		3		4		5	
Lön	Kvinnor	Män	Kvinnor	Män	Kvinnor	Män	Kvinnor	Män	Kvinnor	Män
9-11,9	615	107	1823	484	83	32	31			
12-12,9	1138	108	2806	434	199	32	73			
13-13,9	2220	381	10382	1686	239	130	397	174		
14-14,9	560	162	9675	968	900	347	831	194		
15-15,9	114	54	4246	565	1719	533	911	228	21	
16-16,9	21	44	1709	651	1758	467	1293	454	10	
17-17,9			1389	520	1054	468	1675	576	124	
18-18,9	10	11	765	251	786	271	1729	721	114	11
19-19,9			196	122	487	229	1076	644	21	11
20-20,9			73	43	289	110	1492	882	31	21
21-22,9			21	22	237	66	550	567	62	44
23-25,9				11		22	238	412	238	250
26-29,9					10	11	114	205	186	163
30-34,9							52	151	10	99
35-39,9							155	230		44
40-125							145	492		33
Totalt	4677	869	33084	5758	7762	2717	10763	5930	817	675

6B. Som 6A, men procent

1	...	5		
Kvinnor	Män	Kvinnor	Män	
13,2	12,3	...	0,0	0,0
24,3	12,5	...	0,0	0,0
47,5	43,8	...	0,0	0,0
12,0	18,7	...	0,0	0,0
2,4	6,3	...	2,5	0,0
0,4	5,1	...	1,3	0,0
0,0	0,0	...	15,2	0,0
0,2	1,3	...	13,9	1,6
0,0	0,0	...	2,5	1,6
0,0	0,0	...	3,8	3,2
0,0	0,0	...	7,6	6,4
0,0	0,0	...	29,1	37,1
0,0	0,0	...	22,7	24,2
0,0	0,0	...	1,3	14,6
0,0	0,0	...	0,0	6,5
0,0	0,0	...	0,0	4,9
100,0	100,0	...	100,0	100,0

Uttag av kvottabeller – medellön etc. för olika redovisningsgrupper

Genom att summera aggregeringsvariablerna i kolumn (6) – (9) i råtabellen i bild 6.3 för olika kombinationer av indelningsvariablerna i kolumn (1) – (4) får man tabeller med summor som kan användas för att bilda medelvärden och kvoter för olika redovisningsgrupper.

Tabell 6C. Tabell som är underlag för flera olika kvottabeller

Kön (1)	Nivå (2)	$\sum w_i$ (3)	$\sum w_i \cdot l_{0i}$ (4)	$\sum w_i \cdot t_{j0mfi}$ (5)	$\sum w_i \cdot l_{0nheltidi}$ (6)	(6)/(3) (7)	(4)/(5) (8)	(5)/(3) (9)
Kvinnor	1	4677	40635041	3061	61701517	13 191	13 274	0,65
Kvinnor	2	33084	360320838	25085	474211547	14 334	14 364	0,76
Kvinnor	3	7762	106558725	6416	129063365	16 627	16 607	0,83
Kvinnor	4	10763	175129111	9313	202168595	18 784	18 805	0,87
Kvinnor	5	817	17415674	771	18333690	22 431	22 583	0,94
Män	1	869	10905658	797	11795151	13 579	13 690	0,92
Män	2	5758	73156416	4898	85083524	14 777	14 936	0,85
Män	3	2717	42580548	2532	45632570	16 796	16 815	0,93
Män	4	5930	125232247	5430	136222043	22 971	23 063	0,92
Män	5	675	18050648	655	18594809	27 531	27 564	0,97

Tabell 6D

Medellön efter kön och kompetensnivå		
Nivå	Kvinnor	Män
1	13 274	13 690
2	14 364	14 936
3	16 607	16 815
4	18 805	23 063
5	22 583	27 564

Med kolumn (8) i tabell 6C bildas tabell 6D ovan

Kolumn (3) innehåller vägda antal personer inom varje redovisningsgrupp. Kolumn (4) innehåller vägda faktiska lönesummor, kolumn (5) vägda antal heltidstjänster och kolumn (6) vägda heltidslönesummor. Kolumn (7) innehåller genomsnittlig heltidslön per person, kolumn (8) genomsnittlig lön per heltidstjänst och kolumn (9) genomsnittlig tjänstgöringsomfattning per person.

I bild 6.4 visas delar av en tabell där medellöner för kvinnor och män jämförs inom samma åldersklass och yrke. Hela tabellen innehåller ca 800 rader och det är lämpligt att sammanfatta tabellens innehåll genom att beräkna standardvägda medellöner. Kvinnor och män har olika fördelning på ålder och yrke, vilket förklarar största delen av löneskillnaden 15 680 – 18 860. Om både kvinnor och män har samma ålders- och yrkesfördelning enligt kolumn (8) nedan skulle löneskillnaden endast ha varit 16 256 – 16 505 enligt de två sista kolumnerna.

Bild 6.4 Beräkning av standardvägda medellöner

Ålder (1)	SSYK (2)	Medellön kv. (3)	Medellön män (4)	Antal kv. (5)	Antal män (6)	Antal totalt (7)	Standardvikt (8) = (7) / 5688	Kvinnor: (3) · (8)	Män: (4) · (8)
17-24	2330	13 660	14 100	276	75	351	0,0062	84,41	87,13
...	...	...	...	...	...	...	...	...	...
60-64	7130	13 826	13 900	10	63	74	0,0013	18,09	18,19
Totalt		15 680	18 860	4 523	1 165	5 688	1,0000	16 256	16 505

6.2 Skattningsmetoder vid flervärda variabler

I kapitel 4.6 behandlas olika slag av variabler som envärda respektive flervärda variabler. En *envärd variabel* kan endast anta *ett* värde för varje objekt, som t.ex. en persons *ålder* vid en viss tidpunkt. En *flervärd variabel* kan anta *flera* värden för vissa objekt, som t.ex. näringsgren hos ett företag, då ett och samma företag kan bedriva verksamhet inom flera näringsgrenar samtidigt.

Inom SCB:s registersystem finns det många exempel på flervärda variabler:

- *Utbildning* är en flervärd variabel då en person kan ha flera utbildningar, t.ex. kan samma person vara både civilekonom och civilingenjör.
- I årgångsregister är *kommun* en flervärd variabel som antar flera värden för de objekt som flyttar från en kommun till en annan kommun under året.

När vi skapar statistiska register är det vanligt att importera variabler från andra register i systemet. Denna möjlighet är mycket viktig för möjligheterna att producera registerstatistik. Men när vi t.ex. importerar näringsgren från Företagsregistret till ett löneregister eller till Sysselsättningsregistret så importerar vi också problemet med flervärda variabler.

Registerbaserade undersökningar lämpar sig för detaljerade tabellanalyser som ofta bör kompletteras med standardvägningar
Tabellanalys och standardvägning diskuteras i Wallgren (1983)

När register länkas genom en *ett-till-många* relation, kommer också de importerade variablerna att ge upphov till flervärda variabler. I Aktivitetsregistret kan en person ha två anställningar med olika yrken. När yrke, som är en envärd variabel i Aktivitetsregistret (en anställning avser ett yrke), importeras till Sys-sättningsregistret blir yrke där en flervärd variabel som kan anta flera värden för vissa personer.

Flervärda variabler är svåra att hantera men de är både vanliga och viktiga inom registersystemet. Flervärda variabler i de statistiska registren används också i total- och urvalsundersökningar vilket innebär att problemen också påverkar dessa undersökningar.

Dessa problem brukar ”lösas” på ett drastiskt sätt – den flervärda variabeln transformeras till en envärd variabel genom att endast använda ”det viktigaste värdet” för varje objekt. Om t.ex. de yrkesverksammas fördelning på olika yrken skall undersökas kommer yrken som är vanliga som bisysslor att underskattas. Man kastar bort en del av yrkesinformationen och skattningarna kommer att få kvalitetsbrister av okänd omfattning.

Vi inleder med ett överskådligt exempel för att visa de grundläggande principerna för hur vi anser att flervärda variabler bör behandlas. Därefter behandlar vi mer komplicerade situationer som uppstår när principerna används i praktiken.

6.2.1 Yrke i Aktivitets- och Yrkesregistret

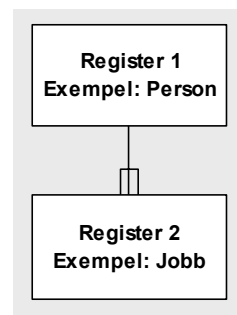
Beräkningsproblemet i detta avsnitt kan definieras på följande sätt: Hur skall man beräkna eller skatta de yrkesaktivas fördelning på olika yrken?

Princip nr 1: Vilka beräkningar skall göras? Man måste alltid precisera beräkningsproblemet innan man startar beräkningsarbetet! Detta exemplifieras nedan där tre sätt att definiera och lösa beräkningsproblemet jämförs.

I följande datamatrix anges yrken och yrkeskod (SSYK) för sex personer, två av dessa har mer än ett yrke. Objekten i matrisen är *jobb*, som är *relationsobjekt* som identifieras med personnummer och företagsidentitet. Från löneregistren hämtas variabeln *Tjombf*, dvs. tjänstgöringens omfattning, som anges i procent av heltidsarbete. Antag att datamatrixen innehåller alla yrkesverksamma i en liten region, hur skall vi skatta *personernas* fördelning på olika yrken? Detta är vårt inledande beräkningsproblem.

Bild 6.5 Aktivitetsregister över jobb med yrkesuppgifter

Person	Företag	Yrke	SSYK	Tjombf	Vikt alt 1	Vikt alt 2	Vikt alt 3
Pnr 1	JE 1	Statistiker	2211	100	1	0,80	1,00
Pnr 1	JE 2	Jordbrukare	6111	15	0	0,12	0,15
Pnr 1	JE 3	Politiker	1110	10	0	0,08	0,10
Pnr 2	JE 4	Sjukvårdsbiträde	5132	30	1	0,60	0,30
Pnr 2	JE 5	Städare	9122	20	0	0,40	0,20
Pnr 3	JE 6	Butiksbiträde	5221	10	1	1,00	0,10
Pnr 4	JE 6	Butiksbiträde	5221	50	1	1,00	0,50
Pnr 5	JE 6	Butiksbiträde	5221	20	1	1,00	0,20
Pnr 6	JE 6	Butiksbiträde	5221	100	1	1,00	1,00
Summa					6	6,00	3,55



I kapitel 7 kommer vi att kalla en flervärd variabel, där endast det viktigaste värdet används, för en *stympad flervärd variabel*.

Objekttypen *jobb* består av anställningar och eget företagande som är relationer mellan personer och arbetsgivare.

Beräkningsproblemet i alternativ 1 och 2 är att skatta personernas fördelning på yrken.

Den traditionella ansatsen är att varje person skall ha endast *ett* yrke – *huvudsakligt yrke*. Det innebär att informationen som finns för personer med flera yrken reduceras, endast yrket med högst tjänstgöringsomfattning tas med.

Vi får då en ny datamatrix, där objekten är *personer* och där fördelningen på yrke erhålls genom att summera antal personer efter yrke.

Bild 6.6 Personregister med yrkesuppgifter

Person	Företag	Huvudsakligt yrke	SSYK	Tjmf	Vikt alt 1
Pnr 1	JE 1	Statistiker	2211	100	1
Pnr 2	JE 4	Sjukvårdsbiträde	5132	30	1
Pnr 3	JE 6	Butiksbiträde	5221	10	1
Pnr 4	JE 6	Butiksbiträde	5221	50	1
Pnr 5	JE 6	Butiksbiträde	5221	20	1
Pnr 6	JE 6	Butiksbiträde	5221	100	1
Summa					6

Tabell 6E nedan innehåller den skattade yrkesfördelningen. På det sätt som är tradition inom individstatistiken får varje person samma vikt oavsett om personen arbetar 100 % eller 10 % av en heltidstjänst.

Tabell 6E. Yrkesverksamma efter yrke, alternativ 1

Huvudsakligt yrke	SSYK	Antal	Procent
Statistiker	2211	1	16,7
Sjukvårdsbiträde	5132	1	16,7
Butiksbiträde	5221	4	66,7
Summa		6	100,0

Exemplet visar att yrken som är vanliga som bisysslor underskattas, så t.ex. är både politiker och jordbrukare yrken som ofta bedrivs vid sidan om ett annat huvudsakligt yrke. Skattningar vid flervärda variabler kan i stället göras på ett sätt så att man slipper kasta bort information. Detta är möjligt om man baserar skattningarna på en datamatrix med ”kombinationsobjekt”.

Princip 2: Grundprincipen är att skapa en datamatrix så att *varje kombination av objekt och värde på den flervärda variabeln motsvarar ett objekt* i den nya datamatrixen. Objekten, dvs. raderna, i sådana datamatrixer kallar vi för *kombinationsobjekt*. Datamatrixen i bild 6.7 nedan har bildats på detta sätt, de sex personerna i personregistret i bild 6.6 ger upphov till nio kombinationsobjekt i bild 6.7.

Vi skiljer mellan den nya termen kombinationsobjekt och termen relationsobjekt som infördes i avsnitt 2.1: Relationen mellan två objekt kan göras till ett relationsobjekt

Bild 6.7 Aktivitetsregister över jobb med yrkesuppgifter

Komb.obj.	Person	Yrke	Tjmf	Vikt alt 1	Vikt alt 2
1	Pnr 1	Statistiker	100	1	0,80
2	Pnr 1	Jordbrukare	15	0	0,12
3	Pnr 1	Politiker	10	0	0,08
4	Pnr 2	Sjukvårdsbitr	30	1	0,60
5	Pnr 2	Städare	20	0	0,40
6	Pnr 3	Butiksbiträde	10	1	1,00
7	Pnr 4	Butiksbiträde	50	1	1,00
8	Pnr 5	Butiksbiträde	20	1	1,00
9	Pnr 6	Butiksbiträde	100	1	1,00
Summa				6	6,00

Raderna i matrisen består av alla kombinationer *person • yrke*. T.ex. så förekommer person *Pnr 1*, som har tre yrken, på tre rader i matrisen. Vikterna enligt alternativ 2 har beräknats med variabeln *Tjmf* så att $0,80 = 100/(100 + 15 + 10)$ etc. Vikterna för varje person summerar till 1 både i alternativ 1 och 2. Vikterna i båda alternativen summerar till 6, dvs. totalantalet personer.

Bild 6.7 ovan illustrerar flera generella principer:

Princip 3: Summan av vikterna för *en person* (den objekttyp som beräkningsproblemet avser och som vi utgick ifrån när vi bildade kombinationsobjekt) skall alltid vara 1.

Princip 4: Av princip 3 följer att summan av alla vikter är lika med totalantalet objekt (den objekttyp som beräkningsproblemet avser).

I tabellen nedan jämförs yrkesfördelningen beräknad med vikter enligt alternativ 1 och 2.

Tabell 6F. Yrkesverksamma efter yrke enligt:

Yrke	SSYK	Alternativ 1		Alternativ 2		
		Antal	Procent	Antal	Procent	
Politiker	1110	0,00	0,0	0,08	1,3	Antalen beräknas genom att summera vikterna för varje yrke. I alternativ 1 summeras Vikt alt 1 och i alternativ 2 Vikt alt 2 i bild 6.7
Statistiker	2211	1,00	16,7	0,80	13,3	
Sjukvårdsbiträde	5132	1,00	16,7	0,60	10,0	
Butiksbiträde	5221	4,00	66,7	4,00	66,7	
Jordbrukare	6111	0,00	0,0	0,12	2,0	
Städare	9122	0,00	0,0	0,40	6,7	
Summa		6,00	100,0	6,00	100,0	

Eftersom delar av yrkesinformationen har vikten 0 i alternativ 1 (som motsvarar det traditionella sättet att räkna) så uppkommer fel, dvs. skattningarna enligt alternativ 1 är skeva på så sätt att frekvensen för vissa yrken överskattas medan frekvensen för andra yrken underskattas. Skattningarna enligt alternativ 2 utnyttjar däremot all information om den flervärda variabeln yrke.

Vikterna i alternativ 2 utnyttjar variabeln *Tjomf*. Denna variabel finns endast i löneregistren, för de anställningar som inte ingår där måste man bilda vikter med annan information. Kontrolluppgiftsregistret innehåller *årslönesummor* för alla jobb och kan därför alltid användas. Vikter beräknade med årslöner blir något annorlunda än vikter beräknade med *Tjomf*. I arbetet med att skapa bra skattningar ingår att välja mellan olika *viktgenererande variabler*, det gäller att välja en variabel som är både relevant och funktionell.

Vikterna enligt alternativ 1 och 2 bygger båda på variabeln *Tjomf*, men skulle kunna ha baserats på andra variabler. De vikter man faktiskt använder kan avvika mer eller mindre från de ideala vikterna, för vissa personer kan vikterna för ett yrke vara för stor, för andra personer kan vikterna för samma yrke vara för små. Felen kan då delvis ta ut varandra när man bildar yrkesfördelningen totalt. Det relevanta kvalitetsmåttet skulle vara ett mått som anger hur nära den skattade fördelningen är den fördelning som är beräknad med ideala vikter.

Princip 5: Det är bättre med bra vikter än dåliga, även om de bra vikterna inte är helt perfekta.

I alternativ 1 och 2 är beräkningsproblemet att beskriva *personernas* fördelning på yrken. Ett tredje alternativ är att fördela *arbetsvolymen* på yrken. Arbetsvolymen kan t.ex. beskrivas med antal yrkesverksamma omräknat till heltids-sysselsatta. Detta sätt att räkna är vanligt inom den ekonomiska statistiken där man är van att räkna med volymer istället för individer.

Person *Pnr 1* har tre yrken, ett på heltid och de två övriga motsvarar 15 % respektive 10 % av en heltidssysselsättning. Matrisen med de sex personerna representerar 3,55 *heltidssysselsättningar*. Objekten i matrisen i bild 6.8 är *jobb* och variabeln yrke är en envärd variabel – varje jobb motsvaras av endast ett yrke. Heltidssysselsättningarnas fördelning på yrke får vi genom att summera variabeln *Vikt alt 3* (= *Tjomf/100* i bild 6.7) för olika yrken.

Bild 6.8 Aktivitetsregister över jobb med yrkesuppgifter

Person	Yrke	Vikt alt 1	Vikt alt 2	Vikt alt 3
Pnr 1	Statistiker	1	0,80	1,00
Pnr 1	Jordbrukare	0	0,12	0,15
Pnr 1	Politiker	0	0,08	0,10
Pnr 2	Sjukvårdsbitr	1	0,60	0,30
Pnr 2	Städare	0	0,40	0,20
Pnr 3	Butiksbiträde	1	1,00	0,10
Pnr 4	Butiksbiträde	1	1,00	0,50
Pnr 5	Butiksbiträde	1	1,00	0,20
Pnr 6	Butiksbiträde	1	1,00	1,00
Summa		6	6,00	3,55

Tabell 6G. Yrkesverksamma efter yrke enligt:

Yrke	SSYK	Alternativ 1		Alternativ 2		Alternativ 3	
		Antal	Procent	Antal	Procent	Antal	Procent
Politiker	1110	0,00	0,0	0,08	1,3	0,10	2,8
Statistiker	2211	1,00	16,7	0,80	13,3	1,00	28,2
Sjukvårdsbiträde	5132	1,00	16,7	0,60	10,0	0,30	8,5
Butiksbiträde	5221	4,00	66,7	4,00	66,7	1,80	50,7
Jordbrukare	6111	0,00	0,0	0,12	2,0	0,15	4,2
Städare	9122	0,00	0,0	0,40	6,7	0,20	5,6
Summa		6,00	100,0	6,00	100,0	3,55	100,0

Alternativ 1 och 2 avser samma beräkningsproblem, *personer* fördelade på yrken, men bygger på olika skattningsmetoder som använder olika vikter. Alternativ 3 avser ett annat beräkningsproblem, *arbetsvolymen* fördelad på yrken. Alternativ 3 bygger inte på någon speciell skattningsmetod som använder vikter, utan är endast en enkel summering.

6.2.2 Näringsgren i Företagsregistret

Näringsgren är en annan betydelsefull flervärd variabel. Den skapas i Företagsregistret och används av många register inom registersystemet. Även här brukar man välja ”viktigaste näringsgren” och bortse från den information som finns om övriga näringsgrenar för de arbetsställen eller företagsenheter som statistiken avser. Det finns idag information i Företagsregistret om alla näringsgrenar som ett företag bedriver verksamhet inom. Det finns även uppgifter om hur stor andel av verksamheten som sker inom varje näringsgren. Näringsgrenskod och andel inom respektive näringsgren har bäst kvalitet för industriföretag. Metoden att välja viktigaste näringsgren kan ge problem vid redovisning av branschstatistik och kan dessutom ge tidsserieproblem. T.ex. om ett stort företag år 1 har 51 % av verksamheten inom en viss näringsgren och år 2 endast

Branschstatistik brukar redovisas efter *primär SNI*, som inte alltid behöver vara största SNI.

49 %, kan denna lilla förändring ge stora nivåbrott i många tidsserier då alla anställda skenbart byter näringsgren mellan år 1 och 2. Inom den regionala statistiken kan dessa problem vara ännu allvarligare då ett arbetsställe kan vara dominerande så att byte av näringsgren ger stora tidsseriebrott. Små nivåförändringar är egentligen ännu lömskare eftersom de är svårare att upptäcka. Den metodik som presenteras i föregående avsnitt, som handlar om yrke, gör det möjligt att undvika dessa kvalitetsbrister.

I följande tabell anges näringsgren, SNI och antal anställda för tre arbetsställen. *Beräkningsproblemet gäller att skatta antal anställda uppdelade efter SNI.* Den information som den traditionella metoden använder finns i de skuggade tabell-cellerna, informationen i de ofärgade cellerna finns, men används inte.

Bild 6.9 Företagsregistret år 1: Datamatrix för arbetsställen

Arbetsställe	Näringsgren 1	%	Näringsgren 2	%	Näringsgren 3	%	Anställda
AE 1	DJ	100					218
AE 2	DH	51	DJ	49			293
AE 3	DJ	40	DH	30	DK	30	156

Bild 6.10 Företagsregistret år 2: Datamatrix för arbetsställen

Arbetsställe	Näringsgren 1	%	Näringsgren 2	%	Näringsgren 3	%	Anställda
AE 1	DJ	100					221
AE 2	DJ	52	DH	48			314
AE 3	DJ	36	DH	34	DK	30	143

Tabell 6H. Antal anställda efter SNI år 1-2

Näringsgren	År 1	År 2
DH	293	0
DJ	374	678
DK	0	0
Totalt	667	678

I avsnitt 6.2.8 finns ett exempel som visar de tidsserieproblem som dagens skattningsmetodik medför. Exemplet bygger på verkliga data.

I tabell 6H är antal anställda tabellerade efter första SNI, som är det vanliga sättet att presentera tidsserier baserade på näringsgren från Företagsregistret.

I bild 6.9 och 6.10 anger procenttalen proportioner, som är ett mått på storleken för varje SNI inom varje arbetsställe. Storleksmättet kan vara baserat på omsättning, antal anställda eller något annat. Vi antar först att procenttalet är baserat på antalet anställda och därefter visar vi hur man transformerar vikter, baserade på ett visst storleksmått, till vikter baserade på ett annat storleksmått.

Enligt princip 2 i föregående avsnitt skapas en ny datamatrix (se bild 6.11 nedan) som innehåller kombinationsobjekt så att varje kombination av objekt och värde på den flervärda variabeln motsvarar en rad i den nya datamatrixen.

I denna datamatrix är varje rad en kombination av arbetsställe och näringsgren. I stället för matriser med tre rader avseende tre arbetsställen får vi nya matriser med sex rader avseende alla kombinationer Arbetsställe • SNI för respektive år. För varje SNI kan vi sedan skatta antalet anställda med formeln (4):

$$\hat{Y} = \sum_{i=1}^R w_i y_i \quad \text{I bild 6.11 nedan har } w_i y_i \text{ beräknats för varje rad} \quad (4)$$

I föregående avsnitt med yrkesverksamma är variabeln $y_i = 1$ för alla kombinationsobjekten. Formel (4) betyder då att vikterna w_i summeras för varje tabellcell.

Bild 6.11 Datamatrix med kombinationsobjekt: Arbetsställe • SNI

År 1					År 2				
Arbetsställe	SNI	Vikt, w_i	Anställda, y_i	$w_i y_i$	Arbetsställe	SNI	Vikt, w_i	Anställda, y_i	$w_i y_i$
AE 1	DJ	1	218	218	AE 1	DJ	1	221	221
AE 2	DH	0,51	293	149,43	AE 2	DH	0,48	314	150,72
AE 2	DJ	0,49	293	143,57	AE 2	DJ	0,52	314	163,28
AE 3	DJ	0,4	156	62,4	AE 3	DJ	0,36	143	51,48
AE 3	DH	0,3	156	46,8	AE 3	DH	0,34	143	48,62
AE 3	DK	0,3	156	46,8	AE 3	DK	0,3	143	42,9
Summa		3		667	Summa		3		678

Vikterna w_i summerar till 3 eftersom det fortfarande är fråga om 3 arbetsställen. Summorna av produkterna $w_i y_i$ blir samma totaler som förut – *totalantalet anställda är givna totaler som inte skall ändras när vi inför vikter. Vikterna kommer endast att påverka hur de anställda fördelas på de olika näringsgrenarna*. Detta är ett exempel på en princip som gäller allmänt.

Tabell 6J. Antal anställda efter SNI år 1-2

SNI	År 1	År 2
DH	196,23	199,34
DJ	423,97	435,76
DK	46,80	42,90
Totalt	667	678

Man bör runda av till heltal när statistiken presenteras

Tidsserierna i tabell 6J har beräknats med vikterna w_i . Jämfört med tabell 6H har serierna i tabell 6J högre kvalitet med relevanta förändringar och inga nivåbrott.

6.2.3 Transformation av vikter

Vikterna för olika näringsgrenar i Företagsregistret är huvudsakligen baserade på omsättning. Om vi skall göra skattningar för andra variabler bör dessa vikter transformeras.

Exempel: Vid skattning av antalet anställda bör vikter baserade på omsättning transformeras till vikter lämpliga för antalet anställda. Denna transformation baseras på en modell som belyser sambandet mellan omsättning och antal anställda och modellen baseras på lämplig statistik som beskriver anställning och omsättning inom olika SNI för branschrena arbetsställen. I bild 6.12 visas denna transformation för arbetsställe AE 3. Vikten baserad på omsättning multipliceras med antal anställda per omsättning. Dessa värden måste sedan divideras med en konstant så att de summerar till 1 för varje arbetsställe.

Bild 6.12 Transformation av vikter

Register över arbetsställen			Aggregerade branschdata		Register över arbetsställen		
År 1		Vikter baserade	Modeller för olika SNI		Transformerade vikter baserade på modell		
Arbetsställe	SNI	på omsättning	Antal anställda/Mkr omsättning		lämpade för att skatta antal anställda		
AE 3	DJ	0,4	DJ	0,5	AE 3 DJ	$0,4 \cdot 0,5 / (0,4 \cdot 0,5 + 0,3 \cdot 0,6 + 0,3 \cdot 0,7) = \mathbf{0,34}$	
AE 3	DH	0,3	DH	0,6	AE 3 DH	$0,3 \cdot 0,6 / (0,4 \cdot 0,5 + 0,3 \cdot 0,6 + 0,3 \cdot 0,7) = \mathbf{0,30}$	
AE 3	DK	0,3	DK	0,7	AE 3 DK	$0,3 \cdot 0,7 / (0,4 \cdot 0,5 + 0,3 \cdot 0,6 + 0,3 \cdot 0,7) = \mathbf{0,36}$	

Enligt princip 3 skall summan av vikterna för *ett arbetsställe* (den objekttyp som beräkningsproblemet avser) alltid vara 1.

Den transformerade vikten för den del av arbetsstället som tillhör stålindustrin DJ, vilken är kapitalintensiv, är lägre än originalvikten, som är baserad på

omsättningen (0,34 jämfört med 0,4). Exemplet illustrerar också att den viktigaste näringsgrenen är DJ om vi använder omsättning, men troligen DK om vi använder antalet anställda.

Detta visar återigen att principen med att endast använda den flervärda variabelns ”viktigaste” värde kan medföra problem. Vidare gäller att när register har flervärda variabler bör vikter beräknas, som är lämpliga för olika beräkningsproblem.

6.2.4 Importera flervärda variabler – många flervärda variabler samtidigt

Till Sysselsättningsregistret importeras många variabler från flera register. Populationen skapas med Befolkningsregistret den 31/12 och därifrån hämtas envärda variabler som ålder och kön. Från Utbildningsregistret hämtas utbildning, från Yrkesregistret hämtas yrke, båda är flervärda variabler. Aktivitetsregistret innehåller objekten jobb som är en kombination av person och arbetsställe. Till Aktivitetsregistret importeras arbetsställets näringsgrenar och slutligen exporteras arbetsställe med näringsgrenar till Sysselsättningsregistret. Både arbetsställe och näringsgren är flervärda variabler i Sysselsättningsregistret. En härledd variabel skapas i Sysselsättningsregistret som anger om en person är förvärvsarbetande under november (kallas nedan *FarbNov*).

Den traditionella skattningsmetoden innebär att man för alla dessa flervärda variabler endast använder uppgifter som avser det ”viktigaste” variabelvärdet (skuggas i matriserna nedan). Vi visar först alla inblandade register med de uppgifter som skall knytas till en viss person *Pnr 10* och hur den traditionella skattningen görs. Därefter visas hur man med vikter för de flervärda variablerna utnyttjar all information.

1. Traditionell metodik – endast viktigaste värde används

Bild 6.13 Befolkningsregistret

Person	Kön	Ålder
Pnr 10	Kv	32

Bild 6.14 Utbildningsregistret

Person	Utb 1	Poäng 1	Utb 2	Poäng 2
Pnr 10	U1	140	U2	120

Pnr 10 har två examina på samma nivå men med olika inriktning, U2 är den senaste.

Bild 6.15 Aktivitetsregistret, med tjänstgöringens omfattning i november

Person	Arbetsställe	Tjomf
Pnr 10	AE 11	80 %
Pnr 10	AE 12	20 %

Endast arbetsstället för huvudsysslan används traditionellt

Bild 6.16 Yrkesregistret

Person	Arbetsställe	Yrke
Pnr 10	AE 11	Y1
Pnr 10	AE 12	Y2

Endast yrket för huvudsysslan används traditionellt

Bild 6.17 Företagsregistret

Arbetsställe	SNI 1	Vikt 1	SNI 2	Vikt 2
AE 11	DH	70 %	DJ	30 %
AE 12	DK	100 %		

Aktivitets- och Företagsregistret matchas med arbetsställeidentiteten som kopplingsvariabel. Största SNI importerar till Aktivitetsregistret. Därefter matchas Aktivitetsregistret med Yrkesregistret med personnummer och arbetsställeidentitet som kopplingsvariabler. Yrke importerar till Aktivitetsregistret.

Bild 6.18 Aktivitetsregistret, SNI och yrke har importerats

Person	Arbetsställe	SNI	Yrke
Pnr 10	AE 11	DH	Y1
Pnr 10	AE 12	DK	Y2

Endast huvudsysslans arbetsställe och yrke används traditionellt

Sysselsättningsregistret skapas i tre steg:

- Genom selektion av personer 16 år och äldre och variablerna kön och ålder från Befolkningsregistret skapas ett nytt register.
- Från olika register importerar viktigaste variabelvärde för de flervärda variablerna utbildning, yrke, arbetsställe och näringsgren.
- Variabeln *FarbNov* som avser förvärvsarbete i november skapas.

Den del av det färdiga Sysselsättningsregistret som avser person *Pnr 10* har följande utseende:

Bild 6.19 Sysselsättningsregistret skapas

Person	Kön	Ålder	Utb	Yrke	Arbetsställe	SNI	FarbNov
Pnr 10	Kv	32	U2	Y1	AE 11	DH	Ja

2. Metodik med vikter där all information används

Bild 6.20 Befolkningsregistret

Person	Kön	Ålder
Pnr 10	Kv	32

Bild 6.21 Utbildningsregistret

Person	Utbildning	w_{UTB}
Pnr 10	U1	0,6
Pnr 10	U2	0,4

Vikter för utbildningar skapas med hjälp av utbildningens längd uttryckt i högskolepoäng

Bild 6.22 Aktivitetsregistret, med tjänstgöringens omfattning i november

Person	Arbetsställe	w_{AE}
Pnr 10	AE 11	0,8
Pnr 10	AE 12	0,2

För objektet *Person* bildas vikter för den flervärda variabeln *Arbetsställe* med hjälp av variabeln *Tjomf*.

Bild 6.23 Yrkesregistret

Person	Arbetsställe	Yrke
Pnr 10	AE 11	Y1
Pnr 10	AE 12	Y2

Eftersom yrke är knutet till kombinationen *Person* och *Arbetsställe* är vikten för *Yrke* densamma som för *Arbetsställe*

Bild 6.24 Företagsregistret

Arbetsställe	SNI	w_{SNI}
AE 11	DH	0,7
AE 11	DJ	0,3
AE 12	DK	1,0

Med informationen i Företagsregistret skapas ett register med kombinationsobjekt *Arbetsställe* • *SNI* och vikter för olika SNI.

Ett Sysselsättningsregister bildas med kombinationsobjekt som är kombinationer av den avsedda objekttypen *Person* och alla värden på samtliga flervärda variabler, enligt den princip vi nämnt ovan. (*Princip 2: Skapa en datamatrix som innehåller kombinationsobjekt så att varje kombination av objekt och värde på den flervärda variabeln motsvarar en rad i den nya datamatrixen.*)

Som resultat av matchningarna skapas en datamatrix och den del av datamatrixen som avser person *Pnr 10* har följande utseende:

Bild 6.25 Datamatrix för kombinationsobjekten: Person • Utbildning • Arbetsställe • SNI

Person	Kön	Ålder	Utb.	Arbställe	Yrke	SNI	FarbNov	W _{UTB}	W _{AE}	W _{SNI}	W _{KombObj}
Pnr 10	Kv	32	U1	AE 11	Y1	DH	Ja	0,6	0,8	0,7	0,336
Pnr 10	Kv	32	U1	AE 11	Y1	DJ	Ja	0,6	0,8	0,3	0,144
Pnr 10	Kv	32	U1	AE 12	Y2	DK	Ja	0,6	0,2	1,0	0,120
Pnr 10	Kv	32	U2	AE 11	Y1	DH	Ja	0,4	0,8	0,7	0,224
Pnr 10	Kv	32	U2	AE 11	Y1	DJ	Ja	0,4	0,8	0,3	0,096
Pnr 10	Kv	32	U2	AE 12	Y2	DK	Ja	0,4	0,2	1,0	0,080
Totalt											1,000

Beräkningarna skall avse objekttypen *Person*. Variablerna *Utbildning*, *Arbetsställe* och *Yrke* är flervärda variabler för objekttypen *Person*. *SNI* är en flervärd variabel för objekttypen *Arbetsställe*. Varje kombination av *Person* och *Arbetsställe* motsvaras av endast ett *Yrke*, därför behövs ingen ytterligare vikt för variabeln *Yrke*. *Yrke* är här en envärd variabel för varje kombination av *Person* och *Arbetsställe*.

I detta exempel motsvarar varje kombination av person och arbetsställe ett yrke. I SCB:s löneregister förekommer ibland att en person har flera yrken på samma arbetsställe.

Princip 6: När många flervärda variabler finns i samma datamatrix skall alla vikter för de flervärda variablerna multipliceras för att få de vikter som skall användas i skattningarna.

I Bild 6.25 har person *Pnr 10* delats upp i sex kombinationsobjekt eller rader. Vikterna $w_{KombObj}$ för dessa sex kombinationsobjekt skall summera till 1 och bildas genom multiplikation, t.ex. $0,6 \cdot 0,8 \cdot 0,7 = 0,336$

3. Skattningar med traditionell metod och med vikter vid flervärda variabler

Med bakgrundsvariablerna ålder, kön, utbildning, yrke och SNI kan en rad frekvenstabeller bildas för variabeln *FarbNov*, dvs. förvärvsarbetande i november. Vi utgår från datamatrixen ovan och visar nedan hur person *Pnr 10* kommer att bidra till skattningar av antalen i olika tabellceller. Till vänster visas tabeller där antalen skattats på traditionellt sätt. Till höger visas motsvarande tabeller där antalen skattas genom att summera vikterna $w_{KombObj}$ i respektive cell. Skillnaderna mellan skattningsmetoderna är stora, med den traditionella metoden kastar man bort mycket av informationen hos de flervärda variablerna.

Tabell 6K1. Antal förvärvsarbetande i november efter ålder och kön

Traditionella skattningar

Ålder	Kv	Män	Summa
20-49	1	0	1
50-64	0	0	0
65-	0	0	0
Summa	1	0	1

Skattningar för envärda variabler som ålder och kön påverkas inte av de vikter man bildar för de flervärda variablerna.

Skattningar med vikter

Ålder	Kv	Män	Summa
20-49	1	0	1
50-64	0	0	0
65-	0	0	0
Summa	1	0	1

Tabell 6K2. Antal förvärvsarbetande i november efter yrke*Traditionella skattningar*

Yrke	Antal
Y1	1
Y2	0
Summa	1

Skattningar med vikter:
 $0,8 = 0,336 + 0,144 + 0,224 + 0,096$

Skattningar med vikter

Yrke	Antal
Y1	0,8
Y2	0,2
Summa	1

Tabell 6K3. Antal förvärvsarbetande i november efter utbildning*Traditionella skattningar*

Utbildn.	Antal
U1	0
U2	1
Summa	1

Skattningar med vikter:
 $0,6 = 0,336 + 0,144 + 0,224 + 0,120$

Skattningar med vikter

Utbildn.	Antal
U1	0,6
U2	0,4
Summa	1

Tabell 6K4. Antal förvärvsarbetande i november efter näringsgren*Traditionella skattningar*

SNI	Antal
DH	1
DJ	0
DK	0
Summa	1

Skattningar med vikter:
 $0,56 = 0,336 + 0,224$

Skattningar med vikter

SNI	Antal
DH	0,56
DJ	0,24
DK	0,20
Summa	1

Tabell 6K5. Antal förvärvsarbetande i november efter yrke och utbildning*Traditionella skattningar*

Utbildn.	Y1	Y2	Summa
U1	0	0	0
U2	1	0	1
Summa	1	0	1

Skattningar med vikter:
 $0,48 = 0,336 + 0,144$

Skattningar med vikter

Utbildn.	Y1	Y2	Summa
U1	0,48	0,12	0,60
U2	0,32	0,08	0,40
Summa	0,80	0,20	1

Tabell 6K6. Antal förvärvsarbetande i november efter utbildning och näringsgren*Traditionella skattningar*

SNI	U1	U2	Summa
DH	0	1	1
DJ	0	0	0
DK	0	0	0
Summa	0	1	1

Skattningar med vikter:
0,336 hämtas direkt från datamatrixen

Skattningar med vikter

SNI	U1	U2	Summa
DH	0,336	0,224	0,560
DJ	0,144	0,096	0,240
DK	0,120	0,080	0,200
Summa	0,600	0,400	1

Tabell 6K7. Antal förvärvsarbetande i november efter yrke och näringsgren*Traditionella skattningar*

SNI	Y1	Y2	Summa
DH	1	0	1
DJ	0	0	0
DK	0	0	0
Summa	1	0	1

Skattningar med vikter:
 $0,56 = 0,336 + 0,224$

Skattningar med vikter

SNI	Y1	Y2	Summa
DH	0,56	0,00	0,56
DJ	0,24	0,00	0,24
DK	0,00	0,20	0,20
Summa	0,80	0,20	1

6.2.5 Konsistens mellan olika variabler

För att det skall bli konsistens vid användningen av de olika flervärda variabler som finns i registersystemet bör vikterna för dessa finnas i det register som har ansvar för respektive flervärd variabel. Dessa vikter bör användas av alla.

Vissa register innehåller många variabler som skall uppfylla vissa konsistensvillkor som t.ex. posterna i en resultaträkning. Konsistensen bibehålls om samma vikt används för alla variabler eller om delposterna räknas om först med eventuellt olika vikter och summor och skillnader beräknas sist.

6.2.6 Urvalsundersökningar med flervärda registervariabler

Om Företagsregistret används som urvalsram, bör vikterna för kombinationsobjekten Arbetsställe • SNI kombineras med vikterna som används för skattningar vid urval. Summeringar skall göras med alla kombinationsobjekt som tillhör motsvarande urvalsenhet. Vikterna d_i och g_i är de vanliga vikterna som används för skattningar vid urval och w_i är vikten för den flervärda variabeln som beskriver näringsgren i Företagsregistret.

$$\hat{Y} = \sum_{i=1}^{r_r} d_i g_i w_i y_i \quad r_r \text{ är antalet svarande kombinationsobjekt} \quad (5)$$

Exempel: I Arbetskraftsundersökningarna frågar intervjuaren efter arbetsplats. Därefter kodas arbetsplatsens näringsgren med hjälp av information från Företagsregistret. Det traditionella är att man använder första SNI.

Antag att en person med Pnr 3 har blivit utvald och att personen arbetar på arbetsställe AE 4 som har verksamhet inom två näringsgrenar DJ och DH med 60 % inom DJ och 40 % inom DH. De vanliga vikterna $d_i g_i$ är lika med 353,8 för denna person.

Bild 6.26 Datamatrix för AKU

Person	Förvärvs- arbetande	Arbetade timmar	Arbets- ställe	SNI	$d_i \cdot g_i$	w_{SNI}
--------	------------------------	--------------------	-------------------	-----	-----------------	-----------

Traditionella skattningar

Pnr 3	Ja	36	AE 4	DJ	353,8	-
-------	----	----	------	----	-------	---

Skattningar med kombinationsobjekt

Pnr 3	Ja	36	AE 4	DJ	353,8	0,6
Pnr 3	Ja	36	AE 4	DH	353,8	0,4

Hur kommer person Pnr 3:s värden att bidra till skattningarna?

Traditionella skattningar innebär att:

Antal sysselsatta inom näringsgren DJ ökas med 353,8 stycken,
 antal arbetade timmar inom DJ ökas med $353,8 \cdot 36 = 12\,736,8$ timmar,
 näringsgren DH får inget tillskott vare sig till antal sysselsatta eller antal arbetade timmar.

Med den *skattningsmetod som anges i formel (5)* ovan skall:

Antal sysselsatta inom näringsgren DJ ökas med $353,8 \cdot 0,6 = 212,3$ stycken,
 antal sysselsatta inom näringsgren DH ökas med $353,8 \cdot 0,4 = 141,5$ stycken.

Antal arbetade timmar inom DJ ökas med $353,8 \cdot 0,6 \cdot 36 = 7\,642,1$ timmar,
 antal arbetade timmar inom DH ökas med $353,8 \cdot 0,4 \cdot 36 = 5\,094,7$ timmar.

Urvalsundersökningar som använder branschvisa skattningar enligt formel (5) har mindre fel och urvalsfel än traditionella skattningar.

6.2.7 Konsistens mellan skattningar från olika register

När flera register innehåller samma flervärda variabel bör skattningarna från olika register överensstämma. För det första måste alla registerpopulationer vara konsistenta. I exemplet i bild 6.27 är personerna desamma både i Person- och Aktivitetsregistret. Vidare är arbetsställe- och företagsenheter desamma i Aktivitetsregistret och i Företagsregistret.

För det andra måste alla variabler vara konsistenta. I exemplet nedan är variablerna *SNI* och *Tjomf* (tjänstgöringens omfattning) konsistenta mellan de olika registren, samma variabeldefinitioner och mätfel finns i alla register då variablerna har exporterats från ursprungsregistret till övriga register.

Det tredje villkoret för konsistens är att skattningarna görs med vikter så att all information om den flervärda variabeln *SNI* ingår i skattningarna. Om man inte använder vikter utan i stället endast använder viktigaste värde på *SNI* så kommer skattningarna ha fel och felen blir olika i olika register så att skattningarna blir inkonsistenta. Detta visas i exemplet nedan.

I alla register bildas kombinationsobjekt med *objekt* • *SNI* som får vikter. Till Jobbregistret i bild 6.27 har variablerna *SNI* och w_{SNI} importerats från Arbetsställeregistret. Genom att summera $Tjomf_J \cdot w_{SNI}$ i Aktivitetsregistret får man *Tjomf* för personer, arbetsställen respektive företagsenheter.

Bild 6.27 Konsistenta skattningar med vikter i ett system av register

Personregister – sysselsättning						
Person	SNI	Tjomf _P	w _{SNI}	w _{Jobb}	$w_P = w_{SNI} \cdot w_{Jobb}$	Tjomf _P · w _P
Pnr 1	A	0,50	0,6	0,6	0,36	0,18
Pnr 1	B	0,50	0,4	0,6	0,24	0,12
Pnr 1	D	0,50	1	0,4	0,4	0,20
Pnr 2	A	1,00	0,6	1	0,6	0,60
Pnr 2	B	1,00	0,4	1	0,4	0,40
Pnr 3	C	1,00	1	1	1	1,00
Pnr 4	C	1,00	1	1	1	1,00
Pnr 5	D	1,00	1	1	1	1,00
Summa					5	4,5

Aktivitetsregister – jobb								
Jobb	Ftgenh	Arbställe	Person	SNI	Tjomf _J	w _{SNI}	Tjomf _J · w _{SNI}	
J1	FE 1	AE 11	Pnr 1	A	0,30	0,6	0,18	
J1	FE 1	AE 11	Pnr 1	B	0,30	0,4	0,12	
J2	FE 1	AE 11	Pnr 2	A	1,00	0,6	0,60	
J2	FE 1	AE 11	Pnr 2	B	1,00	0,4	0,40	
J3	FE 2	AE 21	Pnr 3	C	1,00	1	1,00	
J4	FE 2	AE 21	Pnr 4	C	1,00	1	1,00	
J5	FE 2	AE 22	Pnr 5	D	1,00	1	1,00	
J6	FE 2	AE 22	Pnr 1	D	0,20	1	0,20	
Summa						6	4,5	

Kommentarer till Aktivitetsregistret: Registret innehåller uppgifter om 6 jobb som motsvarar 4,5 heltidstjänster. Jobb J1 och J2 har delats upp i två kombinationsobjekt vardera då AE 11 har verksamhet inom både SNI A och SNI B. Från Arbetsställeregistret hämtas vikterna 0,6 och 0,4 för dessa två SNI. Tjomf_J anger tjänstgöringens omfattning för varje jobb.

Kommentarer till Personregistret:

Summeras Tjomf_J • w_{SNI} för person Pnr 1 i Aktivitetsregistret får man 0,18 + 0,12 + 0,20 = 0,50. Detta värde blir Tjomf_P för Pnr 1 i Personregistret. För Pnr 1 bildas tre kombinationsobjekt för tre SNI. Både SNI och jobb/arbetsställe är flervärda variabler för personer. Vikterna för kombinationsobjekten bildas genom att multiplicera w_{SNI} med w_{Jobb} där w_{SNI} hämtas från Aktivitetsregistret och w_{Jobb} beräknas som varje jobs andel av personens samtliga jobb.

För Pnr 1 har jobb J1 vikten 0,30/(0,30+0,20)=0,6

Kommentarer till tabellen Heltidsanställda efter SNI:

Alla dessa fyra register ger exakt samma skattningar av hur antal heltidsanställda fördelar sig på olika SNI.

Heltidsanst efter SNI	
A	0,78
B	0,52
C	2,00
D	1,20
S:a	4,5

Företagsregister – arbetsställen				
Arbställe	SNI	Tjomf _{AE}	w _{SNI}	Tjomf _{AE} • w _{SNI}
AE 11	A	1,3	0,6	0,78
AE 11	B	1,3	0,4	0,52
AE 21	C	2,0	1	2,00
AE 22	D	1,2	1	1,20
Summa			3	4,5

Företagsregister – företagsenheter					
Ftgenh	Arbställe	SNI	Tjomf _{FE}	w _{SNI}	Tjomf _{FE} • w _{SNI}
FE 1	AE 11	A	1,3	0,6	0,78
FE 1	AE 11	B	1,3	0,4	0,52
FE 2	AE 21	C	3,2	0,625	2,00
FE 2	AE 22	D	3,2	0,375	1,20
Summa				2	4,5

Bild 6.28 nedan visar det traditionella sättet att beräkna antal heltidsanställda för olika SNI. I alla fyra register används endast största SNI för respektive arbetsställe, företagsenhet, jobb och person. På detta sätt tvingas SNI att bli en envärd variabel – man behöver inte bilda kombinationsobjekt, men skattningarna av antal heltidsanställda efter SNI blir olika för varje register. Detta beror på att när man bortser från den information som finns i kombinationsobjekten så kommer SNI-information att tas bort på något olika sätt i olika register.

Bild 6.28 Traditionella skattningar i ett system av register

Personregister – sysselsättning				Aktivitetsregister – jobb						
Person	SNI	Tjmf _P	w _P	Jobb	Ftgenh	Arbställe	Person	SNI	Tjmf _J	w _{SNI}
Pnr 1	A	0,5	1	J1	FE 1	AE 11	Pnr 1	A	0,3	1
Pnr 2	A	1,0	1	J2	FE 1	AE 11	Pnr 2	A	1,0	1
Pnr 3	C	1,0	1	J3	FE 2	AE 21	Pnr 3	C	1,0	1
Pnr 4	C	1,0	1	J4	FE 2	AE 21	Pnr 4	C	1,0	1
Pnr 5	D	1,0	1	J5	FE 2	AE 22	Pnr 5	D	1,0	1
Summa		4,5	5	J6	FE 2	AE 22	Pnr 1	D	0,2	1
				Summa					6	

I alla register i bild 6.28 betyder SNI *huvudsaklig* SNI. Varje register har dessutom endast en rad per objekt (person, jobb, arbetsställe eller företagsenhet). Det som vi kallar skillnad i beräkningsmetod innebär alltså att man skapar registren på olika sätt – med eller utan kombinationsobjekt.

I tabell 1 i bild 6.29 nedan jämförs olika skattningar av heltidsanställda efter SNI. Kolumnerna "PersonReg", ... "FtgReg-FE" är skattningar som har beräknats med registren på denna sida. Skillnaderna mellan dessa fyra kolumner beror enbart på att skattningsmetoden är olämplig. Sista kolumnen i tabell 1 innehåller de skattningar med vikter som görs med registren på föregående sida.

I tabell 2 finns dels traditionella skattningar som bildats med personregistret i bild 6.28 ovan. I den högra kolumnen finns skattningar där vikterna w_P i personregistret i bild 6.27 på föregående sida används. För t.ex. SNI A är 0,36 + 0,6 = 0,96

Företagsregister – arbetsställen			
Arbställe	SNI	Tjmf _{FE}	w _{SNI}
AE 11	A	1,3	1
AE 21	C	2,0	1
AE 22	D	1,2	1
Summa		4,5	3

Företagsregister – företagsenheter				
Ftgenh	Arbställe	SNI	Tjmf _{FE}	w _{SNI}
FE 1	AE 11	A	1,3	1
FE 2	AE 21	C	3,2	1
Summa			4,5	2

Bild 6.29 Jämförelse av olika slag av skattningar

Tabell 1 Antal heltidsanställda efter SNI					Skattningar med vikter
SNI	PersonReg	AktivitetsReg	FtgReg-AE	FtgReg-FE	
A	1,5	1,3	1,3	1,3	0,78
B	0,0	0,0	0,0	0,0	0,52
C	2,0	2,0	2,0	3,2	2,00
D	1,0	1,2	1,2	0,0	1,20
S:a	4,5	4,5	4,5	4,5	4,5

Tabell 2 Antal personer efter SNI	
Traditionellt	Med vikter
2	0,96
0	0,64
2	2,00
1	1,40
5	5,00

De traditionella skattningsmetoderna ger alltså upphov till inkonsistenser i den registerstatistik som framställs. Dessutom kan skillnader i populationsavgränsningar och variabeldefinitioner ge upphov till ytterligare inkonsistenser.

Konsistens diskuteras i avsnitt 11.5

Det finns en fjärde orsak till att registerstatistik från olika produkter kan vara inkonsistenta. Tabellerna ovan visar effekten av innehållsmässiga skillnader. Tabell 1 beskriver *heltidsanställda* medan tabell 2 beskriver *personer*. Inom person- och arbetsmarknadsstatistiken är det vanligt att beskriva personer, men inom den ekonomiska statistiken är volymer och heltidsanställda vanliga mått.

6.2.8 Flervärda variabler – hur gör vi i praktiken?

Exempel: Byte av SNI

I avsnitt 6.2.2 behandlas den flervärda variabeln näringsgren. Den nuvarande metodiken innebär att all verksamhet i ett företag med flera näringsgrenar hänförs till den största näringsgrenen. Detta ger upphov till tidsseriebrott när ett företag byter största näringsgren. Storleken på tidsseriebrottet beror på hur stort företaget är inom sin bransch eller region. Exemplet nedan har verklighetsbakgrund, men vi har transformerat uppgifterna något. Företaget X AB har under åren 1-3 verksamhet inom flera branscher, men ca 60 % av omsättningen sker inom SNI R. Under år 4 köpte X AB ett annat företag med verksamhet inom en annan bransch. Ägarbytet sker från och med kvartal 4 år 4.

Kolumn (2) i tabellen nedan innehåller omsättningen i miljarder kronor för alla företag inom SNI R utom X AB. Kolumn (3) och (4) innehåller hela omsättningen för X AB. Kolumn (5) har hämtats från Företagsregistret och anger hur stor andel av verksamheten som sker inom SNI R. Med dagens skattningsmetodik hänförs hela företagets omsättning till SNI R under år 1-4. Observera att man behåller SNI R som huvudsaklig bransch under hela år 4, då man med dagens metodik gör branschbyten vid årsskiften. Från år 5 hänförs ingen del av företagets omsättning till SNI R. Genom att addera kolumn (2) och kolumn (7) får vi tidsserien i kolumn (8) som innehåller ett tidsseriebrott.

Bild 6.30 Skattning av omsättning inom en bransch med två metoder

År kv	SNI R utom X AB	X AB före köp	X AB efter köp	X AB:s andel inom SNI R	Vikt vid traditionell skattning	X AB:s bidrag till SNI R traditionell skattning	SNI R traditionell skattning	X AB:s bidrag till SNI R enligt avsnitt 6.2	SNI R enligt avsnitt 6.2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1 1	7,684	7,354		0,60	1	7,354	15,038	4,412	12,096
1 2	7,086	7,086		0,60	1	7,086	14,172	4,252	11,338
1 3	8,142	6,788		0,60	1	6,788	14,930	4,073	12,215
1 4	9,853	8,387		0,60	1	8,387	18,240	5,032	14,885
2 1	9,192	7,684		0,60	1	7,684	16,876	4,610	13,803
2 2	9,620	8,102		0,60	1	8,102	17,722	4,861	14,481
2 3	8,065	7,165		0,60	1	7,165	15,230	4,299	12,364
2 4	11,471	7,362		0,60	1	7,362	18,832	4,417	15,887
3 1	10,418	8,566		0,58	1	8,566	18,984	4,968	15,386
3 2	10,406	8,535		0,58	1	8,535	18,940	4,950	15,356
3 3	9,784	7,792		0,58	1	7,792	17,576	4,520	14,303
3 4	13,106	9,612		0,58	1	9,612	22,718	5,575	18,681
4 1	13,071	9,259		0,57	1	9,259	22,330	5,278	18,349
4 2	13,127	9,509		0,57	1	9,509	22,636	5,420	18,547
4 3	11,253	9,499		0,57	1	9,499	20,752	5,414	16,668
4 4	12,921		15,881	0,21	1	15,881	28,802	3,335	16,256
5 1	12,782		12,397	0,21	0	0,000	12,782	2,603	15,385
5 2	13,360		12,634	0,21	0	0,000	13,360	2,653	16,013
5 3	11,098		11,621	0,21	0	0,000	11,098	2,440	13,538
5 4	12,888		13,209	0,21	0	0,000	12,888	2,774	15,662

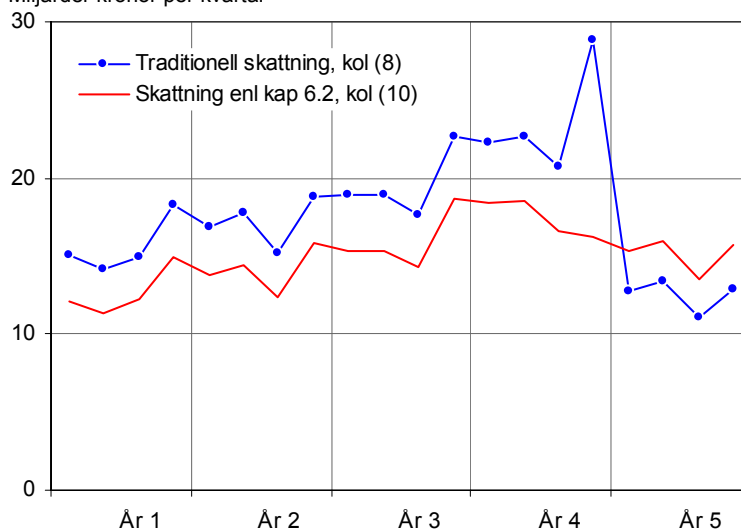
I enlighet med den skattningsmetodik som beskrivs i avsnitt 6.2 skall andelarna i kolumn (5) användas som vikter när omsättningen för bransch X skattas. Genom att multiplicera företagets omsättning i kolumn (3) och (4) med vikterna i kolumn (5) får vi den del av omsättningen som skall hänföras till SNI R. Genom att addera denna del av omsättningen i kolumn (9) med omsättningen för de övriga företagen inom SNI R får vi tidsserien i kolumn (10) som beskriver branschens omsättning utan tidsseriebrott.

Den traditionellt skattade omsättningsserien påverkas av ett tidsseriebrott som beror på de fel som metoden orsakar. Under perioden år 1 till år 4 ger den traditionella metoden överskattningar av branschens omsättning. Det felet blir extra stort under det fjärde kvartalet år 4. Från och med år 5 blir det däremot underskattningar av omsättningen inom SNI R.

Serien i kolumn (10) har skattats med hjälp av vikterna i kolumn (5). Även om dessa vikter inte är helt perfekta så är de avsevärt bättre än de traditionellt använda vikterna i kolumn (6). Genom att använda vikter minskas felen avsevärt och tidsseriekvaliteten ökar betydligt. Genom att på motsvarande sätt använda vikter för de övriga företagen med verksamhet inom SNI R kan skattningarna förbättras ytterligare.

Bild 6.31 Omsättning inom en bransch skattad med två metoder

Miljarder kronor per kvartal



Flervärda variabler – sammanfattande rekommendationer

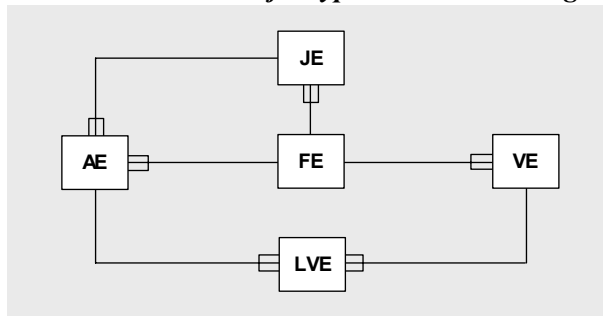
En rad viktiga variabler i registersystemet är flervärda. Det nuvarande sättet att behandla dessa variabler ger i vissa tillämpningar skattningar med fel. Genom att använda vikter vid skattningarna kan dessa fel minskas. I detta avsnitt (hela avsnitt 6.2) beskrivs en rad olika skattningsproblem vid flervärda variabler och hur dessa problem kan lösas. Exemplet ovan med byte av SNI visar att man med relativt enkla metoder, som använder vikter, kan åstadkomma kvalitetsförbättringar utan att vikterna som används är helt perfekta.

Källa: Företagsenheten i den ekonomiska statistiken, SCB (2002a)

6.2.9 Objekttyperna i Företagsregistret

I avsnitt 5.5.3 beskrivs de olika objekttyper som ingår i Företagsregistret. Administrativa data som avser juridiska enheter, JE, och arbetsställeenheter, AE, levereras till SCB:s registersystem. För att få statistiskt meningsfulla företagsenheter måste ibland vissa JE sammanföras till en FE. Vissa FE delas upp i verksamhetsenheter, VE, som är så branschrena som möjligt. Dessutom delas vissa AE upp i branschrena lokala verksamhetsenheter, LVE. Idag görs sådana uppdelningar i branschrena enheter endast i de fall där det är möjligt att samla in data som avser dessa VE och LVE.

Bild 6.32 De olika objekttyperna inom Företagsregistret



FE är den *institutionella enheten*, dvs. en organisation med självständig ledning. Även en offentlig myndighet är en FE.

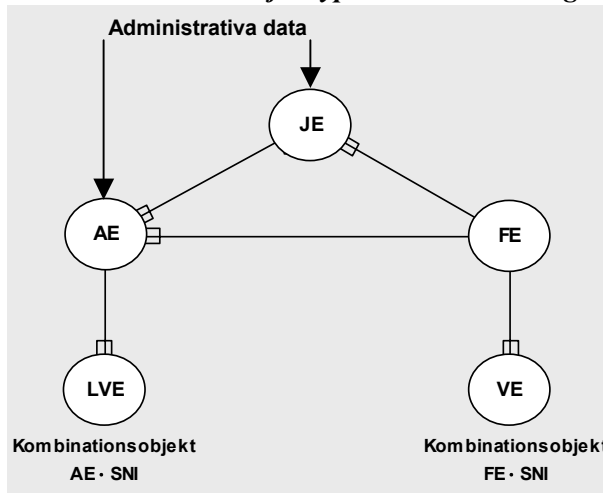
Funktionella enheter är de branschrena enheterna VE och LVE.

För den flervärda variabeln näringsgren har vi i avsnitt 6.2 bildat kombinationsobjekt för varje kombination av näringsgren och FE respektive AE. Dessa kombinationsobjekt kan tolkas så att de motsvarar objekttyperna VE och LVE. Kombinationsobjekten är branschrena och avsikten är att både VE och LVE skall vara branschrena.

I bilden nedan skiljer vi på tre slag av objekttyper:

- JE och AE för vilka det finns administrativa data,
- AE och FE som är de statistiskt intressanta objekttyperna,
- LVE och VE som är de branschvisa uppdelningarna av AE respektive FE.

Bild 6.33 De olika objekttyperna inom Företagsregistret



Den stora mängden av administrativa data avser JE, t.ex. resultat- och balansräkning.

Data avseende AE består i första hand av kontrolluppgifter, de anställda och företagens verksamhet knyts till geografiskt lokaliserade AE.

AE är viktig för arbetsmarknadsstatistiken, FE för den institutionella statistiken, där företag kan följas över tiden, VE för branschstatistiken.

Den konceptuella modellen i bild 6.33 exemplifieras av de fem registren nedan. JE 1 och JE 2 förs samman till FE 1. AE 1 delas upp i LVE 1 och LVE 2 och data samlas in för båda dessa. På liknande sätt delas FE 1 upp i VE 1 och VE 2 och data samlas in även för dessa två enheter.

Bild 6.34 Dagens registerstruktur i Företagsregistret

JE-register		Omsätt- ning	Prel skatt	Ng 1	%	Ng 2	%
JE 1	FE 1	750	0	DH	60	DJ	40
JE 2	FE 1	0	40	DH	60	DJ	40
JE 3	FE 2	550	28	DL	65	DK	35
JE 4	FE 3	300	14	DL	55	DM	45

AE-register		Antal anst	Ng 1	%	Ng 2	%
AE 1	FE 1	500	DH	60	DJ	40
AE 2	FE 2	100	DK	70	DL	30
AE 3	FE 2	250	DL	80	DK	20
AE 4	FE 3	180	DL	55	DM	45

FE-register		Omsätt- ning	Prel skatt	Ng 1	%	Ng 2	%
FE 1		750	40	DH	60	DJ	40
FE 2		550	28	DL	65	DK	35
FE 3		300	14	DL	55	DM	45

LVE-register		Antal anst	Ng
LVE 1	AE 1	300	DH
LVE 2	AE 1	200	DJ
LVE 3	AE 2	100	DK
LVE 4	AE 3	250	DL
LVE 5	AE 4	180	DL

VE-register		Omsätt- ning	Prel skatt	Ng
VE 1	FE 1	450	24	DH
VE 2	FE 1	300	16	DJ
VE 3	FE 2	550	28	DL
VE 4	FE 3	300	14	DL

I bilden nedan har LVE och VE införts konsekvent. För LVE 1, LVE 2, VE 1 och VE 2 samlas data in av SCB. Dessa enheter får då vikten 1 då variabelvärdena avser hela objektet. LVE 3 – LVE 8 och VE 3 – VE 5 är kombinationsobjekt och varje kombination av AE respektive FE och SNI får vikter i proportion till näringsgrenens storlek enligt uppgifterna i bild 6.34. Dessa vikter används vid skattningar på det sätt som beskrivits tidigare.

Fördelen med strukturen i bild 6.35 är dels att strukturen är logisk, LVE och VE är konsekvent definierade, dels får branschstatistiken högre kvalitet. När vi använder vikter fördelas t.ex. antal anställda med vikter på olika SNI. När vi samlar in data till LVE gör företagen motsvarande fördelning så gott de kan.

En ytterligare fördel med strukturen i bild 6.35 är att där man har samlat in uppgifter från VE eller LVE så kan dessa användas på vanligt sätt (med vikten 1), men dessutom kan dessa uppgifter kombineras med data som bygger på administrativa källor

Bild 6.35 Registerstruktur med LVE och VE som motsvarar kombinationsobjekt

JE-register		Samma som ovan					
-------------	--	----------------	--	--	--	--	--

AE-register		Samma som ovan				
-------------	--	----------------	--	--	--	--

LVE-register		Antal anst	Ng	Vikt
LVE 1	AE 1	300	DH	1,00
LVE 2	AE 1	200	DJ	1,00
LVE 3	AE 2	100	DK	0,70
LVE 4	AE 2	100	DL	0,30
LVE 5	AE 3	250	DL	0,80
LVE 6	AE 3	250	DK	0,20
LVE 7	AE 4	180	DL	0,55
LVE 8	AE 4	180	DM	0,45

FE-register		Samma som ovan				
-------------	--	----------------	--	--	--	--

VE-register		Omsätt- ning	Prel skatt	Ng	Vikt
VE 1	FE 1	450	24	DH	1,00
VE 2	FE 1	300	16	DJ	1,00
VE 3	FE 2	550	28	DL	0,65
VE 4	FE 2	550	28	DK	0,35
VE 5	FE 3	300	14	DL	0,55
VE 6	FE 3	300	14	DL	0,45

6.2.10 Skattningsmetoder för årgångsregister

I inledningen till avsnitt 6.2 nämns att i årgångsregister är *kommun* en flervärd variabel som antar flera värden för de objekt som flyttar från en kommun till en annan kommun under året. Som avslutning på detta avsnitt om skattningsmetoder för flervärda variabler visar vi hur skattningar kan göras för årgångsregister.

I de exempel vi diskuterat tidigare i avsnitt 6.2 har beräkningarna gällt flervärda variabler, där variabeln *vid samma tidpunkt* kan anta ett eller flera värden. I detta avsnitt behandlas variabler där *värdet ändras över tiden* under det år årgångsregistret avser. Genom att använda tiden som viktsgenererande variabel kan korrekta skattningar göras.

Medelfolkmängden i t.ex. en kommun kan beräknas på följande sätt, där vi räknar tillkomstdagen som hel dag i kommunen och dagen då personen flyttade/avled som ingen dag i kommunen.

Bild 6.36 Årgångsregister för år 2002 för befolkningen i en viss (liten) kommun

Personnummer	Fanns 1/1	Tillkom	Upphörde	Fanns 31/12	Vikt = Tid i kommunen, år
Pnr 1	Ja	-	20020517	Nej	136/365 = 0,37
Pnr 2	Ja	-	-	Ja	365/365 = 1,00
Pnr 3	Nej	20020315	20020925	Nej	194/365 = 0,53
Pnr 4	Nej	20020606	-	Ja	209/365 = 0,57
Summa	2			2	2,47

Det traditionella sättet att beräkna medelfolkmängd för 2002 är att bilda medelvärdet av folkmängden den 31/12 år 2001 och folkmängden den 31/12 år 2002. Mer noggranna beräkningar, där tiden i kommunen används som vikter, ger att medelfolkmängden under 2002 var 2,47 personer, istället för det traditionella måttet 2 personer.

I exemplet nedan avser uppgifterna företagen i en viss kommun under år 2003. Vissa företag startas eller upphör vid olika tidpunkter under året och även här kan tiden användas som viktsgenererande variabel. Vi bortser i detta exempel från att företag kan flytta geografiskt mellan kommuner.

Flödes- och stockvariabler skall behandlas olika. Flödesvariabler som t.ex. förädlingsvärde för ett företag avser endast värdet under den period under året som företaget var aktivt och behöver inte vägas. En stockvariabel som anger nivå vid en tidpunkt, som t.ex. antal anställda, måste vägas.

Totalt förädlingsvärde i kommunen år 2003 var 83 milj kr, medan genomsnittligt antal anställda under 2003 var 112,5 personer. Relationstal, som t.ex. produktivitet beräknas som $83/112,5 = 0,738$ milj kr per anställd och år.

Bild 6.37 Årgångsregister för år 2003 för företagen i en viss kommun

Företags-identitet	Fanns 1/1	Tillkom	Upphörde	Fanns 31/12	Vikt	Förädlingsvärde	Antal anställda	Vikt • Ant. anst
Idnr 1	Ja	-	20030630	Nej	0,50	10	30	$0,50 \cdot 30 = 15,0$
Idnr 2	Ja	-	-	Ja	1,00	42	45	$1,00 \cdot 45 = 45,0$
Idnr 3	Nej	20030401		Ja	0,75	31	70	$0,75 \cdot 70 = 52,5$
Summa					2,25	83		112,5

6.3 Skattningsmetoder för att korrigera för övertäckning

I detta avsnitt jämförs över- och undertäckningsproblemen vid urvals- och totalundersökningar, där datainsamlingen görs inom SCB, med de täckningsproblem som kan finnas vid registerbaserade undersökningar baserade på administrativa register. Vi visar också hur kalibreringsmetodikerna kan användas för att korrigera för övertäckning i ett register. Idag brukar inte detta kvalitetsproblem åtgärdas, men med metoderna i detta avsnitt skulle man kunna utpröva metoder som korregerar för denna felkälla.

6.3.1 Täckningsproblem vid undersökningar med egen datainsamling

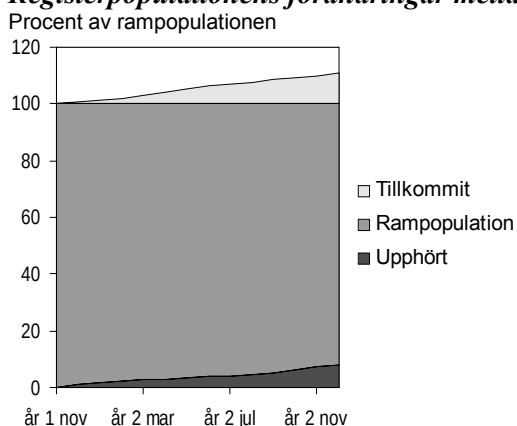
Vid undersökningar med egen datainsamling avgränsas undersökningens rampopulation *innan* data kan samlas in. Det har varit vanligt att kvartals- och månadsundersökningar av företag har varit baserade på den så kallade SAMU-versionen av Företagsregistret.

Undersökningarna skall avse en referensmånad eller ett referenskvartal under ett visst år, som vi här kallar år 2. Målpopulationen är alltså företag som är aktiva under referensperioden under år 2. Rampopulationen för dessa undersökningar definieras som den aktuella stocken av aktiva företag enligt Företagsregistrets SAMU-version i november år 1.

Företag som inte finns med i rampopulationen utan tillkommer mellan december år 1 och december år 2 ger upphov till *undertäckning* i dessa undersökningar. Rosén (1990) redogör för hur skattningar kan korrigeras för denna undertäckning.

Företag som upphört mellan december år 1 och december år 2 ger upphov till *övertäckning*. Om uteblivna svar från dessa företag missuppfattas som svarsbortfall kommer övertäckningen att ge upphov till fel. Rosén (1990) diskuterar hur man under vissa antaganden kan göra skattningar som tar hänsyn till både under- och övertäckning.

Bild 6.38 Över- och undertäckning vid undersökningar med egen datainsamling
Registerpopulationens förändringar mellan november år 1 och december år 2



De företag som tillkommer efter november år 1 ger upphov till *undertäckning* (markeras med ljusgrått i diagrammet).

Den *kvarvarande* delen av rampopulationen markeras med mellangrått i diagrammet.

De företag i ramen som har upphört efter november år 1 ger upphov till *övertäckning* (markeras med mörkgrått i diagrammet).

Vi konkretiserar, med exemplet i bild 6.39 nedan, de problem med över- och undertäckning som finns vid undersökningar med egen datainsamling. Den

rampopulation som man måste använda sig av bestäms *före* datainsamlingen och denna rampopulation jämförs med den registerpopulation som kan skapas *efter* undersökningsperioden när all administrativ information kommit in. Den rampopulation som avgränsas under november år 1 används för årsundersökningarna med egen datainsamling som avser år 1 och för de undersökningar med egen datainsamling som avser t.ex. kvartal 1 under år 2.

Information om företagsombildningar, nystartade och nedlagda företag kommer till Företagsregistret med ganska stor eftersläpning. Dessutom kan t.ex. uppgifter om företagets näringsgren också vara felaktiga i novemberramen, vilket kan komma att upptäckas under år 2 eller år 3. Någon gång under hösten efter det aktuella året har nästan all information avseende föregående år kommit in och ett årgågsregister av god kvalitet kan skapas. *Årgågsregister* kallas ibland för årgågsramar, men då de aldrig används som ram för undersökningar med egen datainsamling bör termen årgågsram undvikas.

Om ramen i bild A används för år 1 är företag *Idnr 1* övertäckning och *Idnr 5* undertäckning och SNI är fel för *Idnr 3*. Används ramen i bild A för år 2 är *Idnr 1* och *2* övertäckning, *Idnr 5* och *6* undertäckning och SNI är fel för *Idnr 3*.

Bild 6.39 Rampopulation och årgågsregister

A. Rampopulation

*bildad november år 1
för år 1 och år 2*

Företagsidentitet	SNI
Idnr 1	DE
Idnr 2	DB
Idnr 3	DA
Idnr 4	DC
-	-
-	-

B. Årgågsregister

*bildat hösten år 2
avser år 1*

Företagsidentitet	SNI
-	-
Idnr 2	DB
Idnr 3	DB
Idnr 4	DC
Idnr 5	DG
-	-

C. Årgågsregister

*bildat hösten år 3
avser år 2*

Företagsidentitet	SNI
-	-
-	-
Idnr 3	DB
Idnr 4	DC
Idnr 5	DG
Idnr 6	DC

6.3.2 Täckningsproblem vid registerbaserade undersökningar

Den stora skillnaden jämfört med undersökningar med egen datainsamling är att vid registerbaserade undersökningar skapas registerpopulationen *efter* den avsedda referensperioden. Beroende på hur snabbt det administrativa systemet får information om att nya objekt tillkommit och att gamla objekt upphört, kan man efter en tid skapa en registerpopulation som avser en viss tidpunkt eller period. De täckningsproblem som beror på att rampopulationen avgränsas *före* referensperioden, som alltid sker vid undersökningar med egen datainsamling, uppkommer inte vid registerbaserade undersökningar. Däremot kan brister i de administrativa systemen orsaka att vissa kategorier av objekt saknas eller att viktiga förändringar hos vissa objekt inte inrapporteras.

Bild 6.40 Populationsavgränsningar vid survey- respektive registerstatistik

	Fördelar	Nackdelar
Surveystatistik, egen datainsamling	Kan vara aktuell	Stora problem med över- och undertäckning och fel i indelningsvariabler om förändringar rapporteras sent
Registerstatistik	Bra täckning, mer korrekta indelningsvariabler	I vissa fall lång tid från händelse till att statistiken blir tillgänglig

6.3.3 Övertäckning i Befolkningsregistret

Den första indikationen på övertäckning i SCB:s befolkningsregister kom från demografiska studier över mortalitet. Bland några kategorier av utrikes födda var mortaliteten ovanligt låg. Dessutom fann man att andelen familjer som saknade disponibel inkomst var hög bland vissa invandrarkategorier. Nilsson (1995) och Ekberg (1995) var de första i Sverige som visade detta.

Övertäckningen i Befolkningsregistret har beräknats av Greijer (1995, 1996, 1997a, 1997b) genom analyser av bortfall i Arbetskraftsundersökningarna och i en totalundersökning av utrikes födda personer baserad på postenkäter. Med denna information är det möjligt att skatta övertäckningen bland olika kategorier av utrikes födda personer.

Data i statistiska register kan också användas för att få indikationer på övertäckning. En utrikes född person som inte har någon inkomst i registren kan ha lämnat Sverige utan att rapportera detta till skattemyndigheten.

Övertäckning kan orsaka allvarliga fel i registerbaserad statistik. T.ex. så kan genomsnittsinkomsten för personer födda i olika länder bli missvisande. För personer födda i vissa länder kan underskattningen bli omkring 20 %.

Hur skall vi kontrollera övertäckningen och förbättra kvaliteten? En strategi för att åtgärda fel orsakade av övertäckning kan bestå av följande:

- Genom att vara vaksam kommer man i samband med granskningsarbetet att lägga märke till orimliga värden i den registerbaserade statistiken. Man bör då ställa frågan om övertäckning vara orsaken till dessa extrema värden.
- Om man misstänker övertäckning så kan tillgängliga urvalsundersökningar användas för att skatta övertäckningen.
- När man har samlat tillräcklig information om övertäckningens omfattning och karaktär, så skattas övertäckningen för olika kategorier i registret.
- Därefter justeras vikterna för att korrigera för den skattade övertäckningen. Före justeringen är alla vikter lika med 1, efter justeringen kommer vikterna för de kategorier, där övertäckning förekommer, att vara mindre än 1. Använd *kalibreringsmetodiken* (se t.ex. Lundström och Särndal, 2001) för att justera vikterna när övertäckningen beskrivs av många variabler samtidigt.
- De justerade vikterna lagras i basregistret (i detta fall Befolkningsregistret).
- Alla andra statistiska produkter som använder basregistret skall använda vikterna – på detta sätt kommer all statistik som produceras att bli konsistent korrigerad för den skattade effekten av övertäckningen.

Exempel:

I ett (fiktivt) register med 1000 utrikes födda är övertäckningen totalt 10 %, dvs. registrets 1000 objekt motsvarar 900 personer. Vi antar vidare att övertäckningen har skattats för olika kategorier enligt tabell 6L nedan och att vi vill använda denna information för att beräkna justerade vikter.

Tabell 6L. Skattad övertäckning för olika kategorier av utrikes födda personer

		Antal personer före korrektion	Skattad övertäckning	Antal personer efter korrektion för skattad övertäckning
Födelseland	Europa, E	584	6,68 %	545
	Ej Europa, R	416	14,66 %	355
	Totalt	1000	10,00 %	900
År i Sverige	Få	819	7,20 %	760
	Många	181	22,65 %	140
	Totalt	1000	10,00 %	900
Inkomst	Låg	101	40,59 %	60
	Hög	899	6,56 %	840
	Totalt	1000	10,00 %	900

I tabell 6L ser det ut som att vi har 6 villkor på antalen, men det är endast 4 villkor eftersom man med 4 antal kan räkna ut de övriga antalen i tabellen. T.ex. kan man med de 4 antal som markerats med fet stil räkna ut de övriga.

6.3.4 Kalibrering av vikter

I bild 6.41 visas hur de justerade vikterna beräknas med kalibreringsmetoden. De först fem kolumnerna visar originalregistret, kolumnerna $x_{1i} - x_{4i}$ innehåller den information som skall användas vid kalibreringen. I beräkningarna används x_i' -vektorer, en vektor per rad. För $i = 1$, dvs. för Pnr 1, är $x_i' = (1 \ 0 \ 1 \ 1)$

Den sista kolumnen visar de justerade vikterna w_i , som har beräknats i tre steg:

1. $T = \sum d_i x_i x_i'$ och T^{-1} beräknas, där alla $d_i = 1$ och $i = 1, 2, \dots, 1000$
 T är en matris med kvadrat- och produktsummor, här en 4×4 matris:

$$T = \begin{bmatrix} \sum x_{1i}^2 & \sum x_{1i} x_{2i} & \sum x_{1i} x_{3i} & \sum x_{1i} x_{4i} \\ \sum x_{2i} x_{1i} & \sum x_{2i}^2 & \sum x_{2i} x_{3i} & \sum x_{2i} x_{4i} \\ \sum x_{3i} x_{1i} & \sum x_{3i} x_{2i} & \sum x_{3i}^2 & \sum x_{3i} x_{4i} \\ \sum x_{4i} x_{1i} & \sum x_{4i} x_{2i} & \sum x_{4i} x_{3i} & \sum x_{4i}^2 \end{bmatrix}$$
2. Vektorn λ beräknas: $\lambda = T^{-1}(t_x - \sum d_i x_i)$
 Vektorn t_x är de 4 villkoren på antalen som korrigerats för under-
 täckning, dvs. de fetmarkerade antalen i tabell 6L ovan, och vektorn
 $\sum d_i x_i$ är motsvarande okorrigerade antal.
3. De justerade vikterna blir : $w_i = d_i (1 + x_i' \lambda)$
 De justerade vikterna används för att beräkna vägda antal och summor.

Bild 6.41 Register med kalibrerade vikter för de första observationerna av 1000

Pnr	Födelse- land	År i Sverige	Inkomst	Vikter d_k	x_{1i} Land=E	x_{2i} Land=R	x_{3i} Få år	x_{4i} Hög inkomst	Justerade vikter, w_i
1	E	Få	Hög	1	1	0	1	1	0,992
2	R	Få	Hög	1	0	1	1	1	0,916
3	E	Få	Låg	1	1	0	1	0	0,657
4	E	Få	Hög	1	1	0	1	1	0,992
5	R	Många	Hög	1	0	1	0	1	0,770
6	R	Få	Låg	1	0	1	1	0	0,581
...	...	...	...	...	...	...	...	...	...
1 000	R	Få	Hög	1	0	1	1	1	0,916
Summa				1 000	584	416	819	899	900

Villkor

t_x	$\sum d_i x_i$
545	584
355	416
760	819
840	899

6.4 Partiellt bortfall – värden saknas

Det finns två bortfallstermer vid undersökningar med *egen* datainsamling: *objektbortfall*, dvs. inga svar inkommer från vissa objekt och *partiellt bortfall*, enstaka frågor har inte besvarats, men svar finns för de övriga frågorna.

För *registerbaserade* undersökningar kan variabelvärden saknas för vissa objekt på grund av brister i det administrativa systemet. Vid registerbaserade undersökningar skapas register oftast med hjälp av *flera* källregister. När olika källregister sambearbetas kan mismatch upptäckas som beror på att objekt saknas i vissa källregister. Detta medför värden saknas för alla variabler som importeras från dessa register. I samband med granskningsarbetet (diskuteras i avsnitt 5.7) upptäcks att värden saknas eller att vissa variabelvärden måste underkännas. Den bortfallsterm som passar bäst för registerbaserade undersökningar är termen *partiellt bortfall* som beteckning på att värden saknas.

För variablerna i ett statistiskt register skall omfattningen av det partiella bortfallet redovisas. Metoderna för bortfallskompensation skall beskrivas. Om man imputerar variabelvärden skall dessa anges i speciella variabler så att man alltid vet vad som är uppmätta värden och vad som är imputerade värden.

6.4.1 Det partiella bortfallet varierar över tiden

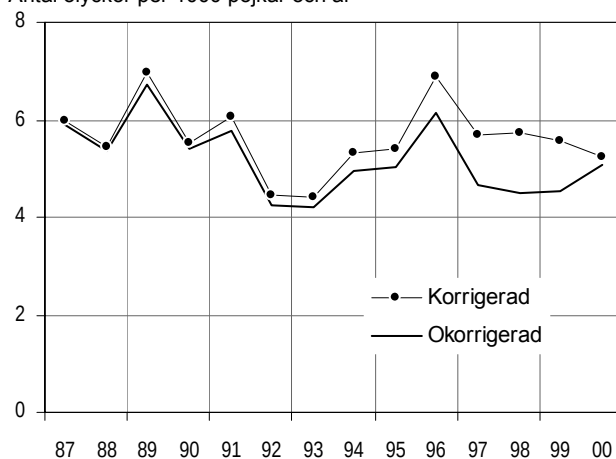
Bortfall inom registerstatistiken borde uppmärksammas och åtgärdas på motsvarande sätt som är praxis för urvalsundersökningar. För registerbaserade undersökningar är det vanligt att man *inte* korregerar för bortfall utan endast redovisar att uppgift saknas. Om bortfallets omfattning varierar över tiden och man inte gör korrekationer kommer jämförbarheten över tiden att vara låg.

Exempel: Bortfall i Patientregistret

Socialstyrelsens Patientregister innehåller uppgifter om dem som vårdats på sjukhus. Diagnos skall registreras, men partiellt bortfall för denna variabel medför att antal patienter med en viss diagnos underskattas. Bortfallet kan variera kraftigt mellan olika år och regioner beroende på hur väl de administrativa systemen på olika sjukhus fungerar.

Bild 6.42 Fallolyckor bland pojkar 0-12 år i Norrbottens län

Antal olyckor per 1000 pojkar och år



Bilden visar skillnaden mellan en tidsserie som *inte* har korrigerats för det partiella bortfallet i diagnosvariabeln och en serie som har korrigerats. Tidsseriemönstret blir felaktigt när bortfallsfrekvensen varierar över tiden. Jämförelser av okorrigerade värden för olika regioner blir missvisande om regionerna har olika bortfallsfrekvens.

Begreppet *bortfall* är avsett för undersökningar med egen datainsamling.

Vid registerbaserade undersökningar är det svårt att skilja mellan begreppen objektbortfall och undertäckning

Källa: Bearbetningar av data från *Barns skador i Sverige*, SOU 2002:99

6.4.2 Två traditioner för att åtgärda bortfall

För registerbaserade undersökningar är det, som vi tidigare nämnt, vanligt att man *inte* korregerar för bortfall utan endast redovisar att uppgift saknas.

Exempel: Bortfallet i Arbetskraftsundersökningen och Utbildningsregistret

Dessa två undersökningar redovisar sina resultat på de sätt som anges i de två tabellerna 6M1 och 6M3 till vänster nedan. Tabellerna visar hur respektive undersöknings huvudvariabel rapporteras. Arbetskraftsundersökningen (AKU) är en urvalsundersökning med ca 15 % bortfall under 2001. Bortfallet i AKU har varierat över tiden, men man har alltid korregerat för bortfallets effekter. Bortfallet i Utbildningsregistret (UREG) anges som "Uppgift saknas" och man har aldrig korregerat för detta bortfall.

Tabell 6M1. AKU enl AKU- praxis

Arbetskraftsundersökningen 2001		
Arbetskraftskategori	Tusental	% av bef.
Sysselsatta	4 239	75,3
Arbetslösa	175	3,1
Ej i arbetskraften	1 218	21,6
Befolkningen 16-64 år	5 632	100,0

Tabell 6M2. AKU enl UREG- praxis

Arbetskraftsundersökningen 2001		
Arbetskraftskategori	Tusental	% av bef.
Sysselsatta	3 603	64,0
Arbetslösa	149	2,6
Ej i arbetskraften	1 035	18,4
Uppgift saknas	845	15,0
Befolkningen 16-64 år	5 632	100,0

Tabell 6M3. UREG enl UREG- praxis

Utbildningsregistret 2001		
Utbildningsnivå	Tusental	% av bef.
Under 9 år	755	11,8
Grundskola 9 år	939	14,7
Gymnasium 2 år	1 747	27,4
Gymnasium 3 år	1 142	17,9
Högskola < 3 år	802	12,6
Högskola ≥ 3 år	848	13,3
Forskarutbildning	48	0,7
Uppgift saknas	106	1,7
Befolkningen 16-74 år	6 386	100,0

Tabell 6M4. UREG enl AKU- praxis

Utbildningsregistret 2001		
Utbildningsnivå	Tusental	% av bef.
Under 9 år	767	12,0
Grundskola 9 år	954	14,9
Gymnasium 2 år	1 776	27,8
Gymnasium 3 år	1 162	18,2
Högskola < 3 år	816	12,8
Högskola ≥ 3 år	862	13,5
Forskarutbildning	48	0,8
Befolkningen 16-74 år	6 386	100,0

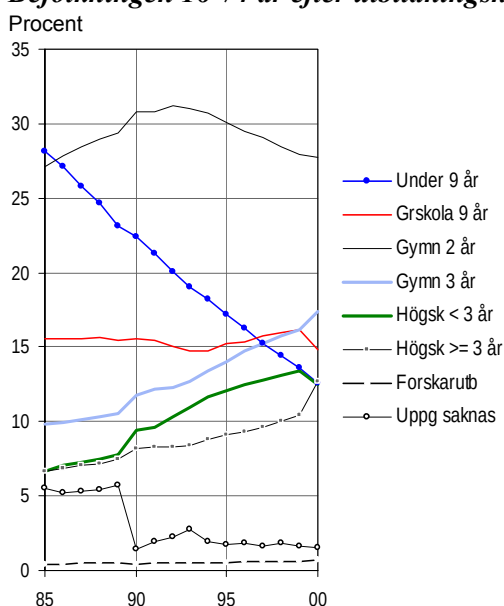
I tabell 6M2 ovan redovisar vi samma Arbetskraftsundersökning som i tabell 6M1, men bortfallet redovisas på det sätt som är praxis i Utbildningsregistret idag. Andelen sysselsatta, arbetslösa och personer ej i arbetskraften minskar jämfört med tabell 6M1. Om bortfallet i AKU skulle öka, allt annat oförändrat, så skulle andelen sysselsatta etc. minska ytterligare. Detta visar att det skulle vara orimligt att inte bortfallskorrigera.

I tabell 6M4 ovan har vi bortfallskorrigerat UREG, något som är praxis i AKU. Genom bortfallskorrekturen slipper användarna kategorin "Uppgift saknas", dessutom får man tidsserier av högre kvalitet. Om kategorin "uppgift saknas" har ett visst innehållsmässigt intresse, genom att t.ex. vara en indikator på låg utbildning, kan denna information användas för att göra bortfallskorrekturen på ett bra sätt.

Bortfallet i Utbildningsregistret har varierat kraftigt över tiden, dessa variationer ger skenbara förändringar i övriga tidsserier vilket syns i bilden nedan.

Bortfallskorrekturen i tabell 6M4 har gjorts genom *rak uppräkning* 12 % av de med känd nivå har studerat mindre än 9 år: $755/(6386-106)=0,12$ Om även 12 % av de 106 tusen där uppgift saknas har studerat mindre än 9 år kommer även 12 % av samtliga 6 386 tusen att ha studerat mindre än 9 år

Bild 6.43 Bortfallets effekter på tidsserier från Utbildningsregistret
Befolkningen 16-74 år efter utbildningsnivå 1985-2000



Mellan 1989 och 1990 minskade bortfallet från 5,7 % till 1,4 %.

Orsaken till detta var uppgifter som samlades in i samband med FoB90.

Alla serier utom Grundskola och Forskarutbildning ökar mellan 1989-1990 och ökningarna är till stor del skenbara.

Efter 1990 har bortfallet varierat mellan 2,7 % och 1,6 % vilket stör de övriga serierna något.

Det finns även ett tidsseriebrott mellan 1999 och 2000 som beror på ändrad utbildningsklassifikation och att nya källor tillkommit. Detta diskuteras vi i avsnitt 6.5.

6.4.3 Bortfallskorrektion med vikter

Vilka metoder kan användas för att korrigera för bortfall? Vi visar den enklaste metoden "rak uppräknings" för fiktiva material dels för AKU och dels för data från UREG. Vi väljer den enklaste metoden för att visa principen, för AKU används idag en bättre metodik och även för UREG skulle man kunna göra bättre korrektioner om man tar hänsyn till andra variabler som ålder och kön etc. Men även rak uppräknings kan vara bättre än ingen korrektion alls. Eftersom bortfallet kan variera kraftigt mellan undersökningar från olika tidpunkter, skulle tidsserier som inte korrigerats för bortfall bli svårtolkade.

Om AKU bygger på ett urval (OSU) av 20 000 från populationens 5 632 000 personer i åldern 16-64 år, blir designvikterna $d_i = 5\,632\,000/20\,000 = 281,6$. Om bortfallet är 3 000 personer, dvs. 15 %, så svarar 17 000 personer och korrektionsfaktorn $g_i = 20\,000/17\,000 = 1,176$. De korrigerade vikterna $d_i g_i$ blir då 331,3. De korrigerade vikterna för bortfallet blir 0. Med formel (1) i avsnitt 6.1.1 beräknas alla skattningar. I bilden nedan visas beräkningarna av de korrigerade vikterna.

Urvalet i AKU är inte ett OSU utan ett stratifierat urval, vilket innebär att vikterna d_i varierar mellan olika strata

Bild 6.44 Bortfallskorrektion av Arbetskraftsundersökningen med vikter

Person	Kön	Ålder	AK-status	d_i	$d_i g_i$
Pnr 1	Kv	34	Sysselsatt	281,6	331,3
Pnr 2	M	28	Arbetslös	281,6	331,3
Pnr 3	M	59	Uppgift saknas	281,6	0
Pnr 4	Kv	24	Ej i AK	281,6	331,3
Pnr 5	M	62	Uppgift saknas	281,6	0
Pnr 6	M	33	Sysselsatt	281,6	331,3
...	...	...	...	...	...
Pnr 20 000	Kv	54	Sysselsatt	281,6	331,3
Summa:				5 632 000	5 632 000

Motsvarande görs för UREG i bilden nedan, där ursprungsvikterna $d_i = 1$. Om bortfallet är 106 051 så finns uppgifter för 6 279 964 personer och korrektionsfaktorn $g_i = 6\,386\,015 / 6\,279\,964 = 1,01689$. De korrigerade vikterna $d_i g_i$ för bortfallet blir 0. Med formel (2) i avsnitt 6.1.1 beräknas alla skattningar.

Bild 6.45 Bortfallskorrektion av Utbildningsregistret 2001 med vikter

Person	Kön	Ålder	Utbildningsnivå	d_i	$d_i g_i$
Pnr 1	M	18	Grundskola 9 år	1	1,01689
Pnr 2	Kv	72	Under 9 år	1	1,01689
Pnr 3	M	33	Gymnasium 2 år	1	1,01689
Pnr 4	M	62	Gymnasium 3 år	1	1,01689
Pnr 5	Kv	71	Uppgift saknas	1	0
Pnr 6	Kv	26	Högskola ≥ 3 år	1	1,01689
Pnr 7	M	54	Forskarutbildning	1	1,01689
Pnr 8	M	67	Uppgift saknas	1	0
Pnr 9	Kv	39	Under 9 år	1	1,01689
***	***	***	***	***	***
Pnr 6386000	M	53	Högskola < 3 år	1	1,01689
Summa:				6 386 015	6 386 015

Korrigerade vikter i en datamatrix medför att alla tidsserier blir konsistent korrigerade och jämförbarheten över tiden förbättras. Då bortfallet varierar mellan åren kommer de korrigerade vikterna att bli olika för olika år.

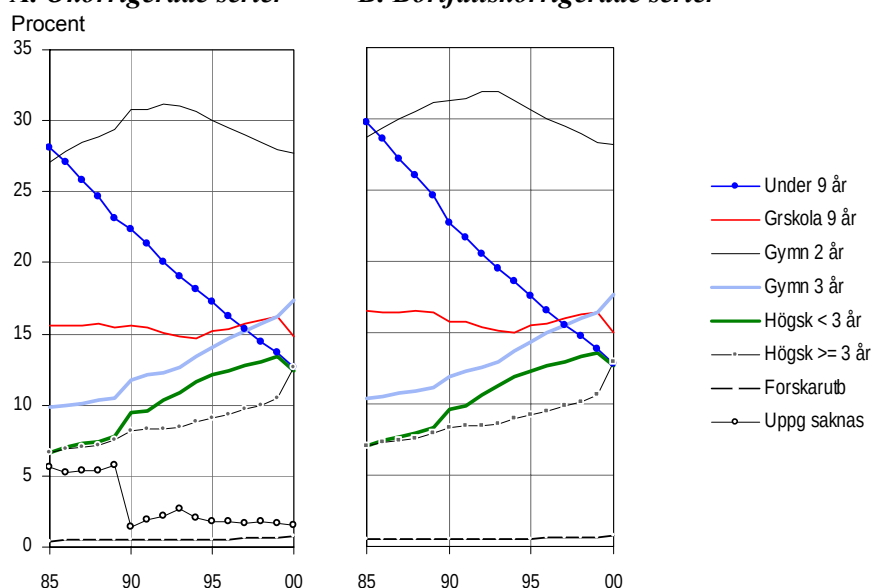
Efter bortfallskorrektion på detta enkla sätt har tidsseriekvaliteten ökat, dvs. jämförbarheten mellan år har förbättrats. De korrigerade serierna beskrivs i bild 6.46 B nedan. Här visar vi principen för bortfallskorrektion med den enklast tänkbara metoden, rak uppräknig. I en verklig situation bör man använda bättre metoder som använder mer hjälpinformation, t.ex. kön och ålder. Vill man använda data från perioden 1985-1989 tillsammans med senare uppgifter så måste man göra bortfallskorrektion för att få jämförbarhet mellan olika år.

Bortfallskorrektion med vikter som använder mer hjälpinformation beskrivs i Lundström och Särndal (2002)

Bild 6.46 Befolkningen 16-74 år efter utbildningsnivå 1985-2000

A. Okorrigerade serier

B. Bortfallskorrigerade serier



6.4.4 Bortfallskorrektion med imputering

Ett annat sätt att korrigera för det partiella bortfallets effekter är att bilda *imputerade värden* när variabelvärden saknas. De värden som saknas ersätts då med syntetiska värden. Det finns två olika sätt att bilda sådana värden:

- Värden bildas slumpmässigt med en eller flera sannolikhetsfördelningar. Denna metod avser endast kvalitativa variabler.
- Värden bildas med en deterministisk sambandsmodell på liknande sätt som för härledda variabler, vilket diskuteras i avsnitt 5.8.3.

Fördelen med imputering av variabelvärden är att man slipper räkna med vikter och att fördelningen för alla andra variabler i registret är oförändrade.

Imputerade värden för kvalitativa variabler bildas slumpmässigt

I bilden nedan visas hur man kan bilda imputerade värden för variabeln utbildningsnivå. Eftersom det är fråga om ett personregister bör värden för person Pnr 5 och 8 inte imputeras, utan man bör i stället bilda *syntetiska observationer* som inte har personnummer.

Den imputering som motsvarar rak uppräknings görs på följande sätt:

De observationer som saknar uppgift om *Utbildningsnivå* används för att bilda lika många syntetiska observationer. Dessa tilldelas värden på variabeln *Utbildningsnivå* helt slumpmässigt. Dessa syntetiska värden har samma fördelning som bland de, där uppgift om *Utbildningsnivå* är känd. Registret utökas med slumpstal och därefter skapas en datamatrix utan personnummer. Dessa slumpstal är värden på en teknisk variabel som används internt (avsnitt 4.5.4).

Bild 6.47 Bortfallskorrektion av Utbildningsregistret 2001 med imputering, rak uppräknings

A. Egentligt register

	Kön	Ålder	Utbildningsnivå	Slumptal
Pnr 1	M	18	Grundskola 9 år	0,7771
Pnr 2	Kv	72	Under 9 år	0,3168
Pnr 3	M	33	Gymnasium 2 år	0,3096
Pnr 4	M	62	Gymnasium 3 år	0,8667
Pnr 5	Kv	71	Uppgift saknas	0,1749
Pnr 6	Kv	26	Högskola ≥ 3 år	0,4114
Pnr 7	M	54	Forskarutbildning	0,1605
Pnr 8	M	67	Uppgift saknas	0,5536
Pnr 9	Kv	39	Under 9 år	0,5513
...	...	...	...	...
Pnr 6386000	M	53	Högskola < 3 år	0,7828

B. Datamatrix för tabelluttag

Kön	Ålder	Utbildningsnivå	Utb.nivå imputerad
M	18	Grundskola 9 år	Nej
Kv	72	Under 9 år	Nej
M	33	Gymnasium 2 år	Nej
M	62	Gymnasium 3 år	Nej
Kv	71	Grundskola 9 år	Ja
Kv	26	Högskola ≥ 3 år	Nej
M	54	Forskarutbildning	Nej
M	67	Gymnasium 3 år	Ja
Kv	39	Under 9 år	Nej
...	...	...	...
M	53	Högskola < 3 år	Nej

C. Sannolikhetsfördelning baserad på frekvenstabell 6M4 i avsnitt 6.4.2

Utbildningsnivå	Andel av bef.	Akkumulerad andel
Under 9 år	0,120	0,120
Grundskola 9 år	0,149	0,269
Gymnasium 2 år	0,278	0,547
Gymnasium 3 år	0,182	0,729
Högskola < 3 år	0,128	0,857
Högskola ≥ 3 år	0,135	0,992
Forskarutbildning	0,008	1,000
Befolkningen 16-74 år	1,000	

Slumptalen i registret är likformigt fördelade mellan 0 och 1.

Personer med slumpstal mellan 0 och 0,120 får nivån "Under 9 år", de med slumpstal mellan 0,120 och 0,269 får nivån "Grundskola 9 år" etc.

Genom att utnyttja sambandet mellan ålder, kön och utbildningsnivå kan imputeringen förbättras. För olika kombinationer av åldersklass och kön använder man olika frekvensfördelningar på utbildningsnivå. I tabell 6N nedan kan vi jämföra tre sådana fördelningar. Det är stora skillnader mellan dessa vilket innebär att det går att förbättra bortfallskorrekturen genom att använda olika fördelningar när imputerade värden fördelas slumpmässigt för olika kombinationer av kön och ålder.

Tabell 6N. Frekvenstabell efter ålder och kön, UREG 2001

Utbildningsnivå	Ackumulerad andel Män 65-74 år	Ackumulerad andel Kvinnor 65-74 år	Ackumulerad andel Båda könen 16-74 år
Under 9 år	0,466	0,455	0,120
Grundskola 9 år	0,507	0,532	0,269
Gymnasium 2 år	0,700	0,821	0,547
Gymnasium 3 år	0,837	0,858	0,729
Högskola < 3 år	0,901	0,918	0,857
Högskola ≥ 3 år	0,988	0,998	0,992
Forskarutbildning	1,000	1,000	1,000

I bild 6.48 nedan används samma register med samma slumptal som tidigare. Men slumptalen översätts här till utbildningsnivå genom att använda andra frekvenstabeller. Kvinnor i åldersklassen 65-74 år med slumptal mellan 0 och 0,455 får nivån "Under 9 år". Män i åldersklassen 65-74 år med slumptal mellan 0,507 och 0,700 får nivån "Gymnasium 2 år". På motsvarande sätt skulle yngre personer med "Uppgift saknas" få imputerade värden med hjälp av frekvenstabeller för sin åldersklass och sitt kön.

Bild 6.48 Bortfallskorrekturen av UREG 2001 med imputering, sambandsmodell

A. Egentligt register

	Kön	Ålder	Utbildningsnivå	Slumptal
Pnr 1	M	18	Grundskola 9 år	0,7771
Pnr 2	Kv	72	Under 9 år	0,3168
Pnr 3	M	33	Gymnasium 2 år	0,3096
Pnr 4	M	62	Gymnasium 3 år	0,8667
Pnr 5	Kv	71	Uppgift saknas	0,1749
Pnr 6	Kv	26	Högskola ≥ 3 år	0,4114
Pnr 7	M	54	Forskarutbildning	0,1605
Pnr 8	M	67	Uppgift saknas	0,5536
Pnr 9	Kv	39	Under 9 år	0,5513
...	...	...	...	...
Pnr 6386000	M	53	Högskola < 3 år	0,7828

B. Datamatrix för tabelluttag

Utb.nivå imputerad	Kön	Ålder	Utbildningsnivå
Nej	M	18	Grundskola 9 år
Nej	Kv	72	Under 9 år
Nej	M	33	Gymnasium 2 år
Nej	M	62	Gymnasium 3 år
Ja	Kv	71	Under 9 år
Nej	Kv	26	Högskola ≥ 3 år
Nej	M	54	Forskarutbildning
Ja	M	67	Gymnasium 2 år
Nej	Kv	39	Under 9 år
...	...	...	...
Nej	M	53	Högskola < 3 år

Pnr 5 som är en 71-årig kvinna får här utbildningsnivån "Under 9 år" till skillnad från imputeringen i bild 6.47, där hon fick nivån "Grundskola 9 år". Även den imputerade nivån för Pnr 8 ändras till en kortare utbildning.

När är det lämpligt att använda slumpmässigt fördelade värden?

När man vill beskriva en kvalitativ variabel med partiellt bortfall eventuellt uppdelad på olika kategorier som åldersklass, kön och region är metoden ovan lämplig. Efter en välgjord imputering är nivåerna mer jämförbara dels mellan kategorier och dels över tiden än om man inte korregerar för bortfall.

Om man istället vill studera sambandet mellan en variabel y och en kvalitativ variabel x , där x -variabeln har partiellt bortfall, bör man inte använda slumpmässigt imputerade värden på x -variabeln. T.ex. skulle det inte vara lämpligt att använda slumpmässigt imputerade utbildningsnivåer om man vill beskriva genomsnittlig månadslön för olika utbildningsnivåer. Det är i detta fall bättre att beräkna medellön endast för personer med känd utbildningsnivå.

Imputerade värden bildas med en deterministisk sambandsmodell

Imputeringsmetoden som används i bild 6.48 ovan utnyttjar sambandet mellan x -variablerna ålder och kön och y -variabeln utbildningsnivå. Men dessutom bildas de imputerade värdena *slumpmässigt*. För en viss kombination av ålder och kön bestäms inte utbildningsnivån exakt utan slumpmässigt. Nedan ger vi några exempel på imputeringsmetoder där värdena på x -variablerna bestämmer det imputerade värdet *exakt*. De sambandsmodeller som används vid denna typ av imputering kallar vi för *deterministiska sambandsmodeller*.

I avsnitt 5.8.3 diskuteras hur man kan bilda härledda variabler med deterministiska sambandsmodeller. Imputerade variabelvärden kan bildas på motsvarande sätt, skillnaden är att härledda variabelvärden beräknas för *alla* objekt i datamatriken, medan imputerade variabelvärden bildas endast för de objekt som saknar värde på grund av partiellt bortfall.

Under granskningsarbetet upptäcks att värden saknas eller att värden är orimliga och måste underkännas. Detta leder till att imputerade värden beräknas i nära anslutning till granskningen. De fallstudier av granskningsarbetet som redovisas i avsnitt 5.7 innehåller flera exempel på imputeringsmetoder.

Vid granskningen av Inkomst- och taxeringsregistret upptäcker man att socialbidrag inte har rapporterats för vissa kommuner (se avsnitt 5.7.1). För hushåll bosatta i sådana kommuner, imputeras då föregående års värden. En enkel sambandsmodell, *årets bidrag = föregående års bidrag*, används vid imputeringen. På hushållsnivå kan man då göra *modellfel* eller *imputeringsfel*, om årets bidrag avviker från föregående års. Man skall sträva efter att använda sambandsmodeller som ger så små imputeringsfel som möjligt.

När kvalitetskraven är så höga att imputeringsfelen inte kan tolereras är det lämpligt att göra en *särskild datainsamling*. De objekt som saknar värden på en viktig variabel får då via enkät eller intervju lämna de uppgifter som saknas. I avsnitt 5.7.2 beskrivs granskningen av de årliga kontrolluppgifterna. På varje kontrolluppgift skall arbetsställeidentitet anges. När denna uppgift saknas eller bedöms som orimlig kontaktas arbetsgivaren.

Granskningen av Företagsstatistikens SRU-data som beskrivs i avsnitt 5.7.3 ger exempel på olika typer av imputeringsmetoder:

1. Från årsredovisningar hämtas uppgift om antal helårsanställda. Om denna uppgift saknas bildas imputerade värden genom att beräkna skattat antal helårsanställda genom att företagets lönesumma divideras med branschens medellön per helårsanställd. Branschens medellön har först beräknats med hjälp av de företag där både antal årsanställda och lönesumma är kända.

2. I Företagsstatistikens registerpopulation saknas ekonomiska variabelvärden för ca 100 000 företag. Från Företagsregistret har uppgift om SNI och antal årsanställda importerats. För företag där SNI, antal årsanställda och ekonomiska variabler är kända bildas tabeller med medelvärden för olika ekonomiska variabler efter SNI och antal årsanställda. Dessa tabeller är en form av sambandsmodeller som för givna värden på SNI och antal årsanställda anger hur imputerade värden skall bildas med de beräknade medelvärdena.

6.4.5 Bortfall i ett system av register

När ett register skapas så importerar variabler från olika källregister. Om flera av dessa importerade variabler innehåller partiellt bortfall kommer det nya registret att innehålla partiellt bortfall i större utsträckning än källorna. Det är vissa kvalitativa variabler som används inom många register där kraven på konsistens innebär att bortfallet måste behandlas på ett konsekvent sätt.

Exempel: Sambearbetning av Företagsregistret och Utbildningsregistret

Sysselsättningsregistret innehåller uppgifter avseende förvärsarbete, med näringsgren för det arbetsställe där en person är anställd, och personens högsta utbildning. Dessa variabler importerar från Företagsregistret respektive Utbildningsregistret. I båda dessa källor finns partiellt bortfall. Enligt uppgifterna i tabell 6P är bortfallet avseende utbildningsnivå 1,7 % och bortfallet avseende SNI är $79/5647 = 1,4$ %. Av Sysselsättningsregistrets population avseende hela befolkningen 16-64 år saknar $(79 + 93 - 2)/5647 = 3$ % värde på näringsgren och/eller utbildningsnivå.

Tabell 6P. Befolkningen 16-64 år efter utbildningsnivå och näringsgren, år 2001

Tusentals personer	Sysselsatta inom SNI ...				Ej syss	Bef totalt	Sysselsatta inom SNI ...				Ej syss	Bef totalt
	A-F varu- prod	G-K privata tjänster	L-Q offentliga tjänster	SNI okänd			A-F %	G-K %	L-Q %	SNI okänd %		
Under 9 år, Grundskola 9 år	273	273	141	17	611	1315	24,7	18,9	9,7	21,8	39,3	23,3
Gymnasium 2 år	403	404	416	22	324	1570	36,6	27,9	28,5	28,5	20,8	27,8
Gymnasium 3 år	229	369	217	15	252	1081	20,7	25,5	14,9	18,7	16,2	19,1
Högskola < 3 år	107	188	273	11	177	757	9,7	13,0	18,7	13,6	11,4	13,4
Högskola ≥ 3 år	86	204	410	12	119	830	7,8	14,1	28,0	15,4	7,7	14,7
Utbildning okänd	5	10	4	2	73	93	0,4	0,7	0,3	2,0	4,7	1,7
Hela befolkningen 16-64 år	1103	1448	1462	79	1556	5647	100,0	100,0	100,0	100,0	100,0	100,0

I den tvåvägsindelade tabellen ovan finns bortfall i båda indelningsvariablerna. Även om det totala bortfallet endast är 3 % medför bortfallet att tabellen blir svårtolkad. Sambandsmönstret störs av bortfallet i tabellen. Genom att bortfallskorrigerar kommer flervägsindelade tabeller att bli enklare att tolka.

Men storleken på detta bortfall varierar när olika *kommuner* jämförs. I tabellen nedan anges de högsta och lägsta bortfallsandelarna för landets kommuner.

Tabell 6Q. Sysselsättningsregistret 2000 och 2001 – lägsta och högsta värde på partiellt bortfall bland befolkningen 16-64 år i Sveriges kommuner

Partiellt bortfall:	Lägst år 2000	Högst år 2000	Lägst år 2001	Högst år 2001
Utbildning okänd	0,3 %	3,7 %	0,4 %	4,6 %
SNI okänd	0,5 %	6,0 %	0,4 %	3,7 %
Både utbildning och SNI okända	ca 1 %	ca 7 %	ca 1 %	ca 7,5 %

Även om bortfallet på riksnivå inte är så stort och även är stabilt mellan olika år, kan bortfallet vara betydande på kommunnivå. Tabell 6Q visar att bortfallet varierar kraftigt mellan olika kommuner och även mellan olika år.

Hur jämför man en kommun med 1 % bortfall med en annan kommun där bortfallet är 7 %? För att göra sådana jämförelser måste man bortfallskorrigera, annars blir nivåskattningar avseende olika kommuner inte jämförbara. I tabell 6R nedan har vi korrigerat för bortfallet i tabell 6P. Först fördelades de med ”utbildning okänd” proportionellt över utbildningsnivåerna i samma kolumn, därefter fördelades de med SNI okänd proportionellt över de tre SNI-grupperna på samma rad. Den bortfallskorrigerade tabellen är enklare att tolka, dessutom kan motsvarande tabeller avseende t.ex. olika kommuner jämföras.

Tabell 6R. Befolkningen 16-64 år efter utbildningsnivå och näringsgren, år 2001
Bortfallskorrigerade skattningar

Tusentals personer	Sysselsatta inom SNI ...			Ej syss	Bef totalt	Sysselsatta inom SNI			Ej syss	Bef totalt
	A-F	G-K	L-Q			A-F	G-K	L-Q		
	varu- prod	privata tjänster	offentliga tjänster			%	%	%		
Under 9 år, Grundskola 9 år	281	282	145	641	1337	25.0	19.1	9.8	41.2	23.7
Gymnasium 2 år	413	415	425	340	1597	36.7	28.1	28.6	21.8	28.3
Gymnasium 3 år	234	378	222	264	1100	20.8	25.6	14.9	17.0	19.5
Högskola < 3 år	110	193	279	186	770	9.8	13.1	18.8	12.0	13.6
Högskola ≥ 3 år	88	209	418	125	844	7.8	14.1	28.1	8.0	14.9
Hela befolkningen 16-64 år	1125	1476	1490	1556	5647	100,0	100,0	100,0	100,0	100,0

Bortfallskorrektion med vikter i ett system av register

Befolknings-, Utbildnings- och Sysselsättningsregistret avser befolkningen den 31 december ett visst år. I Befolkningsregistret finns inget partiellt bortfall, i Utbildningsregistret finns partiellt bortfall i variabeln utbildningsnivå och i Sysselsättningsregistret finns partiellt bortfall i utbildningsnivå och näringsgren. Om registren bortfallskorrigeras med vikter var för sig kommer vikterna för samma person att vara olika i dessa tre register. Statistik från de tre registren kommer att bli inkonsistent, t.ex. kan antalet 18-åriga män bli olika. Om statistik från olika register, som avser samma population, skall vara konsistent måste man beräkna vikterna gemensamt och använda samma vikter i alla register. Detta kan vara svårt att uppnå, vår slutsats är att bortfallskorrektion med vikter ger problem med samordning och konsistens inom registersystemet.

Bild 6.49 Bortfallskorrektion med vikter i ett system av register

A. Befolkningsregistret

Person	Kön	Ålder	d _i
Pnr 1	M	18	1
Pnr 2	Kv	72	1
Pnr 3	M	33	1
Pnr 4	M	62	1
Pnr 5	Kv	71	1
Pnr 6	Kv	26	1
Pnr 7	M	54	1
Pnr 8	M	67	1
Pnr 9	Kv	39	1
...	...	...	...

B. Utbildningsregistret

Pnr	Utbildningsnivå	d _i g _i
Pnr 1	Grundskola 9 år	1,01689
Pnr 2	Under 9 år	1,01689
Pnr 3	Gymnasium 2 år	1,01689
Pnr 4	Gymnasium 3 år	1,01689
Pnr 5	Uppgift saknas	0
Pnr 6	Högskola ≥ 3 år	1,01689
Pnr 7	Forskarutbildning	1,01689
Pnr 8	Uppgift saknas	0
Pnr 9	Under 9 år	1,01689
...	...	...

C. Sysselsättningsregistret

Pnr	SNI	Utbildningsnivå	d _i g _i
Pnr 1	DM	Grundskola 9 år	1,02930
Pnr 2	DB	Under 9 år	1,02930
Pnr 3	-	Gymnasium 2 år	0
Pnr 4	DK	Gymnasium 3 år	1,02183
Pnr 5	-	Uppgift saknas	0
Pnr 6	DB	Högskola ≥ 3 år	1,02326
Pnr 7	DK	Forskarutbildning	1,02326
Pnr 8	DA	Uppgift saknas	0
Pnr 9	DM	Under 9 år	1,02930
...	...	...	...

Bortfallskorrektion med imputering i ett system av register

Om olika register i systemet korrigeras för bortfall genom att imputera på det sätt som beskrivs i avsnitt 6.4.4 kommer statistik från olika register att kunna vara helt konsistent. Samtidigt som man importerar en variabel från ett register så importeras även det slumptal som används i ursprungsregistret. Därefter kan man göra imputeringar som är konsistenta mellan olika register. I exemplet nedan visas hur man kan importera utbildningsnivå från Utbildningsregistret och näringsgren från Företagsregistret till Sysselsättningsregistret. Saknade värden i alla dessa tre register kan sedan ersättas med imputerade värden på ett konsistent sätt.

Bild 6.50 Bortfallskorrektion av Utbildningsregistret med imputering**A. Egentligt register**

Person	Kön	Ålder	Utbildningsnivå	Slumptal
Pnr 1	M	18	Grundskola 9 år	0,7771
Pnr 2	Kv	72	Under 9 år	0,3168
Pnr 3	M	33	Gymnasium 2 år	0,3096
Pnr 4	M	62	Gymnasium 3 år	0,8667
Pnr 5	Kv	71	Uppgift saknas	0,1749
Pnr 6	Kv	26	Högskola ≥ 3 år	0,4114
Pnr 7	M	54	Forskarutbildning	0,1605
Pnr 8	M	67	Uppgift saknas	0,5536
...	...	...	...	...

B. Datamatrix för tabelluttag

Kön	Ålder	Utbildningsnivå	Utb.nivå imputerad
M	18	Grundskola 9 år	Nej
Kv	72	Under 9 år	Nej
M	33	Gymnasium 2 år	Nej
M	62	Gymnasium 3 år	Nej
Kv	71	Grundskola 9 år	Ja
Kv	26	Högskola ≥ 3 år	Nej
M	54	Forskarutbildning	Nej
M	67	Gymnasium 3 år	Ja
...	...	...	...

Bild 6.51 Bortfallskorrektion av Företagsregistret med imputering**A. Egentligt register**

Företag	SNI	Slumptal
JE 1	DB	0,0316
JE 2	DK	0,6444
JE 3	Uppgift saknas	0,3978
JE 4	DA	0,2846
JE 5	DK	0,2044
...	...	...

B. Datamatrix för tabelluttag

SNI	SNI imputerad
DB	Nej
DK	Nej
DM	Ja
DA	Nej
DK	Nej
...	...

Bild 6.52 Bortfallskorrektion av Sysselsättningsregistret med imputering**A. Egentligt register**

Person	Företag	SNI	Slumptal SNI	Utbildningsnivå	Slumptal SUN
Pnr 1	JE 5	DK	0,2044	Grundskola 9 år	0,7771
Pnr 2	JE 1	DB	0,0316	Under 9 år	0,3168
Pnr 3	JE 3	-	0,3978	Gymnasium 2 år	0,3096
Pnr 4	JE 2	DK	0,6444	Gymnasium 3 år	0,8667
Pnr 5	JE 3	-	0,3978	Uppgift saknas	0,1749
Pnr 6	JE 1	DB	0,0316	Högskola ≥ 3 år	0,4114
Pnr 7	JE 5	DK	0,2044	Forskarutbildning	0,1605
Pnr 8	JE 4	DA	0,2846	Uppgift saknas	0,5536
...	...	...	...	...	...

B. Datamatrix för tabelluttag

SNI	SNI imputerad	Utbildningsnivå	Utb.nivå imputerad
DK	Nej	Grundskola 9 år	Nej
DB	Nej	Under 9 år	Nej
DM	Ja	Gymnasium 2 år	Nej
DK	Nej	Gymnasium 3 år	Nej
DM	Ja	Grundskola 9 år	Ja
DB	Nej	Högskola ≥ 3 år	Nej
DK	Nej	Forskarutbildning	Nej
DA	Nej	Gymnasium 3 år	Ja
...	...	...	...

Slutsatsen av diskussionen i detta avsnitt är att imputering är den lämpligaste metoden för bortfallskorrektion i ett system av register. Inom systemet ansvarar Utbildnings- respektive Företagsregistret för bortfallskorrektionen av SUN respektive SNI. Andra register skall sedan använda dessa korrekationer.

6.5 Skattningsmetoder för att korrigera för tidsseriebrott

Statistiska register från olika år används för att producera tidsserier. Tusentals tidsserier kan produceras av ett statistiskt register och när *det administrativa systemet ändras*, som är källan till det statistiska registret, kan detta ge upphov till nivåförändringar i många tidsserier. Förändringar hos statistikproducenten kan också orsaka nivåförändringar. *Klassifikationssystem*, som näringsgren, yrke eller utbildning revideras ibland och dessa förändringar orsakar nivåförändringar. *Kvalitetsförändringar* beroende på ändrade metoder eller nya källor medför också nivåförändringar i ett stort antal tidsserier. När *bortfallet* i en återkommande undersökning *varierar över tiden* orsakar även detta jämförbarhetsproblem, vilket diskuteras i avsnitt 6.4.

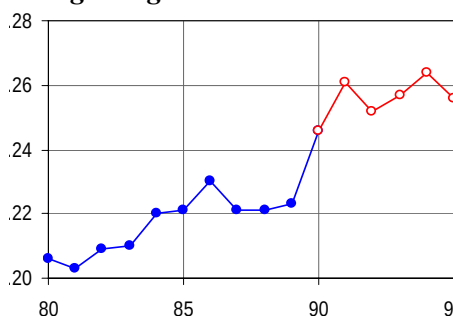
Idag brukar inte detta kvalitetsproblem åtgärdas, men med metoderna i detta avsnitt skulle man kunna utpröva metoder som korregerar för denna felkälla

6.5.1 Olika sätt att redovisa tidsseriebrott

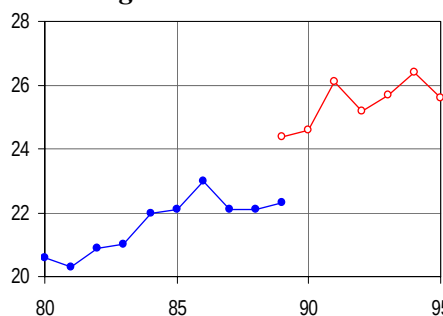
Bild 6.53 A visar situationen där producenten inte har gjort någon ansträngning för att mäta effekten av ett tidsseriebrott, i bild B har producenten gjort dubbla beräkningar för 1989 så att nivåförändringen visas. I bild C har producenten dessutom räknat om den äldre serien för tidigare år.

Bild 6.53 Tre olika sätt att redovisa tidsseriebrott – tidsserie med fiktiva data

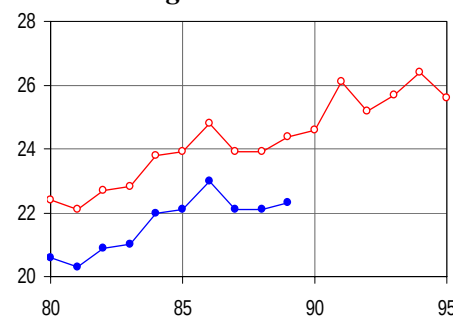
A. Inget år gemensamt



B. Ett år gemensamt



C. Revidering av äldre serie



Att bortse från tidsseriebrottet som i den vänstra bilden får inte förekomma – användarna lämnas helt åt sitt öde. Det räcker inte att enbart påpeka att det är ett tidsseriebrott utan att samtidigt förklara hur data skall tolkas. I SCB-publicationer förekommer påpekanden av typen ”Resultaten för tidigare år bör tolkas med försiktighet ...”. Ingen användare blir glad av detta, ingen vet hur man gör när man tolkar ”med försiktighet”.

Ett minimikrav är att den ansvarige gör som i bilden i mitten, dvs. anger effekten för det år då förändringen inträffade. Det bästa är om den ansvarige gör som i den högra bilden, dvs. länkar tidsserierna för att erhålla långa serier där effekten av brottet har minimerats.

Utredningen om översyn av den ekonomiska statistiken (SOU 2001:34) lyfter fram behovet av långa tidsserier och att SCB bör ta huvudansvaret för att länka ihop tidsserier. Användare tvingas idag att göra egna sammanlänkningar, vilket är ineffektivt och ger osäkra serier. Man påpekar att förbättringar och förändringar av statistiken är nödvändiga, men att det är viktigt att jämförbarheten mellan år bibehålls.

Att korrigera för tidsseriebrott kallas att *länka tidsserier*. Vi skiljer på *länkning på makronivå* och *länkning på mikronivå*. Det vanliga är att man länkar på makronivå, dvs. man arbetar endast med aggregerade tidsserievärden och försöker korrigera dessa för effekterna av tidsseriebrottet. Denna metod har nackdelen att det kan vara svårt att få flera länkade serier att bli konsistenta. Så t.ex. måste de reviderade serierna för män och kvinnor summera till en reviderad serie för befolkningen totalt. Har man många serier får man många konsistensvillkor som skall vara uppfyllda och länkningen blir besvärlig.

Det vanliga är att de som ansvarar för ett statistiskt register kan länka ett *begränsat antal* av de viktigaste tidsserierna i samband med att det har skett en förändring som påverkar registret. Men hur gör man med alla andra, kanske 1 000-tals serier som också har påverkats? Om vi antar att nivåförändringar har blivit skattade för 100 serier, men att det statistiska registret genererar totalt 10 000 serier. Vad kan då göras för de övriga 9 900 serierna?

Metoden att länka på mikronivå har inte dessa två nackdelar, dels länkas alla möjliga tidsserier samtidigt och dessutom blir alla dessa serier konsistenta. Länkning på mikronivå görs genom att beräkna reviderade vikter i datamatrixerna för de år eller perioder som skall revideras. Med de ursprungliga vikterna får man de gamla tidsserievärdena, med de reviderade vikterna får man de länkade serierna.

Vi diskuterar nedan registerbearbetningar som gör det möjligt att beräkna reviderade tidsserievärden för tidigare tidsperioder. En förutsättning för att kunna använda dessa metoder är att man har skattat effekterna av förändringen för åtminstone en tidsperiod genom att länka på makronivå.

6.5.2 Skatta effekter av tidsseriebrott – länkning på makronivå

Om det görs förändringar inom det administrativa systemet skall statistikproducenten vara förberedd och följa vilka förändringar som sker och mäta nivåförändringarna. Detta kan göras med information som finns i de administrativa registren eller med hjälp av urvalsundersökningar. Nivåförändringar för de viktigaste tidsserierna kan därefter skattas.

I en rapport av Wass (2002) analyseras effekterna av tidsseriebrottet 1999-2000 i Utbildningsregistret (UREG). Rapporten är ett bra exempel på hur man skall analysera och dokumentera tidsseriebrott.

År 2000 infördes dels en ny utbildningsnomenklatur (SUN 2000) och samtidigt började man använda nya källor. Wass redovisar hur befolkningen 16-74 år den 31/12 1999¹ fördelar sig på olika utbildningar, dels med gammal nomenklatur och endast de gamla källorna och dels med den nya nomenklaturen och de nya källorna tillagda². Vi har lagt till tre kolumner i tabellen nedan där vi även har korrigerat för bortfall.

¹ 31/12 1999 kallas i rapporten för 2000, vilket skiljer sig från SCB-praxis

² I rapporten analyseras tidsseriebrottet i två steg, för 1999 ändras SUN, för 2000 tillkom nya källor. Vi använder båda dessa förändringar för att beskriva tidsseriebrottet för 1999.

Tabell 6S. Skattning av tidsseriebrottet 1999 för olika utbildningsnivåer

Utbildningsnivå	Utan bortfallskorrektion:		Efter bortfallskorrektion, rak uppräknig:		
	1999 gammal	1999 ny	1999 gammal	1999 ny	Skillnad/brott
Under 9 år	863 332	848 289	877 948	862 428	-15 520
Grundskola 9 år	1 025 322	926 996	1 042 680	942 446	-100 234
Gymnasium 2 år	1 768 378	1 798 085	1 798 316	1 828 054	29 738
Gymnasium 3 år	1 024 001	1 091 800	1 041 337	1 109 997	68 660
Högskola < 3 år	846 876	748 071	861 213	760 539	-100 674
Högskola ≥ 3 år	660 142	776 418	671 318	789 359	118 041
Forskarutbildning	41 619	41 619	42 324	42 313	-11
Uppgift saknas	105 466	103 858	0	0	0
Befolkningen 16-74 år	6 335 136	6 335 136	6 335 136	6 335 136	0

Hur kan man nu använda denna information för att korrigera gamla tidsserievärden? En vanlig metod är att använda *kvotmetoden*, som innebär att t.ex. alla gamla tidsserievärden för serien *Personer med utbildning under 9 år*, skulle multipliceras med kvoten $862\,428/877\,948 = 0,982$. På motsvarande sätt skulle alla gamla serievärden för de sex övriga serierna multipliceras med sina motsvarande kvoter. Detta sätt att länka serier ger emellertid bristande konsistens. De sju tidsserierna kommer inte att summera till befolkningen totalt för åren före 1999.

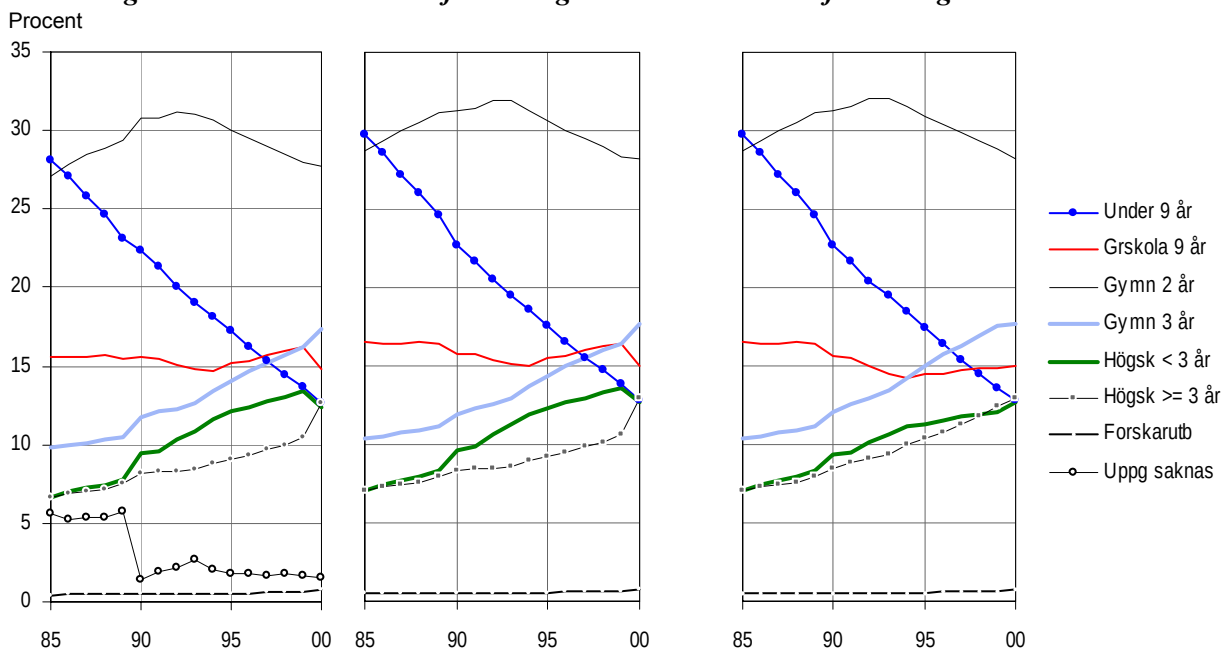
Då serierna skall adderas måste man använda en additiv länkingsmodell, inte en multiplikativ modell som kvotmetoden. I tabell 6T nedan anger vi ett möjligt sätt att länka serierna med en additiv modell. För 1999 korrigeras serierna enligt de skillnader som vi vet gäller för detta år (tabell 6S ovan). Vi har sedan antagit att dessa skillnader har uppstått gradvis under perioden 1990-1999, varför vi minskar nivåskillnaderna med en tiondel varje år. Dessa korrekationer är konsistenta, totalantalet personer 16-74 år förändras inte. Åren före 1990 och efter 1999 korrigeras inte.

Tabell 6T. Additiva tidsseriekorrekationer, antal personer

	Under 9 år	Grskola 9 år	Gymn 2 år	Gymn 3 år	Högs < 3 år	Högs ≥ 3 år	Forskarutb	Totalt
1990	-1 552	-10 023	2 974	6 866	-10 067	11 804	-1	0
1991	-3 104	-20 047	5 948	13 732	-20 135	23 608	-2	0
1992	-4 656	-30 070	8 921	20 598	-30 202	35 412	-3	0
1993	-6 208	-40 094	11 895	27 464	-40 270	47 216	-4	0
1994	-7 760	-50 117	14 869	34 330	-50 337	59 020	-5	0
1995	-9 312	-60 140	17 843	41 196	-60 404	70 824	-7	0
1996	-10 864	-70 164	20 817	48 062	-70 472	82 629	-8	0
1997	-12 416	-80 187	23 790	54 928	-80 539	94 433	-9	0
1998	-13 968	-90 210	26 764	61 794	-90 607	106 237	-10	0
1999	-15 520	-100 234	29 738	68 660	-100 674	118 041	-11	0

I bild 6.54C nedan visas effekten av denna tidsserielänkning. De stora tidsseriebrotten för Grundskola, Gymnasium 3 år och de båda Högskoleserierna har försvunnit. Genom att jämföra bild 6.54A, B och C ser man effekten av de olika stegen i arbetet med att förbättra jämförbarheten mellan åren.

Vi visar principer med en enkel länkning. Inom UREG bör man göra en ännu bättre länkning där man tar hänsyn även till utbildningsinriktning, ålder, kön etc.

Bild 6.54 Befolkningen 16-74 år efter utbildningsnivå 1985-2000**A. Okorrigerade serier****B. Bortfallskorrigerade serier****C. Bortfallskorrigerade och länkade serier**

6.5.3 Länkning av tidsserier på mikronivå med vikter

Det är möjligt att använda metoden med kalibrering av vikter även i denna situation. Kalibreringsmetodiken används i avsnitt 6.3 där vi visar hur man kan korrigera för övertäckning. Arbetet med att korrigera för tidsseriebrott kan göras på liknande sätt. Om vi fortsätter exemplet med Utbildningsregistret skulle man kunna göra arbetet i följande steg:

- Vi antar att 1999 är året med det första registret efter förändringen och att 100 viktiga serier finns både enligt de gamla och nya nivåerna för 1999 och tidigare år. Man har alltså länkat 100 serier på makronivå för perioden 1985-1999 på det sätt som vi ovan länkade sju serier. De 100 länkade värden för varje år är inbördes helt konsistenta, vilket innebär att alla summor stämmer, t.ex. antalet män + antalet kvinnor = totala antalet personer med någon specifik egenskap.
- Det är sedan möjligt att beräkna justerade vikter w_i för enheterna i registret för varje år så att nya tidsserievärden kan beräknas som blir jämförbara med de nya tidsserievärdena för 1999 och senare. De 100 omräknade värdena för ett visst år används som vektorn t_x i kalibreringsprocessen. För varje år kan justerade vikter w_i beräknas på det sätt som beskrivs i avsnitt 6.3.4.
- Med vikterna w_i kan alla tänkbara tidsserievärden för respektive år skattas med inbördes konsistens och skattningarna kommer att vara jämförbara med värdena som avser 1999 och senare.

Denna metod har använts inom SCB. Alla månadsvisa Arbetskraftsundersökningar för perioden 1987-1992 kalibrerades för att justera för de förändringar som infördes 1993. De värden som finns i våra databaser är dessa länkade

värden. Wallgren (1998) ger en närmare beskrivning av tillvägagångssättet. Där användes drygt 100 länkade serier som styrvillkor och efter en omfattande kontroll av hur metoden genererade övriga länkade serier fann vi att med lämpligt valda styrvillkor blev övriga serier länkade på ett acceptabelt sätt.

Vi visar nedan hur en sådan kalibrering går till genom att fortsätta exemplet med Utbildningsregistret, men vi använder endast de sju länkade serierna från föregående avsnitt. Detta ger sju styrvillkor för kalibreringen. Tabellen nedan har hämtats från tabell 6S i avsnitt 6.5.2 där länkning på makronivå diskuteras.

Tabell 6U. Skattning av tidsseriebrottet 1999 för olika utbildningsnivåer

Efter bortfallskorrektion, rak uppräknig:			
Utbildningsnivå	1999 gammal	1999 ny	Skillnad/brott
Under 9 år	877 948	862 428	-15 520
Grundskola 9 år	1 042 680	942 446	-100 234
Gymnasium 2 år	1 798 316	1 828 054	29 738
Gymnasium 3 år	1 041 337	1 109 997	68 660
Högskola < 3 år	861 213	760 539	-100 674
Högskola ≥ 3 år	671 318	789 359	118 041
Forskarutbildning	42 324	42 313	-11
Befolkningen 16-74 år	6 335 136	6 335 136	0

I bild 6.55 nedan visas hur de justerade vikterna beräknas med kalibreringsmetoden. Kolumnerna $x_{li} - x_{7i}$ visar $\mathbf{x}_i'$ -vektorer, en vektor på varje rad. För $i = 1$, dvs. för *Pnr 1*, är $\mathbf{x}_1' = (1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0)$

Den sista kolumnen i registret visar de justerade vikterna w_k , som beräknas på följande sätt:

1. $\mathbf{T} = \sum d_i g_i \mathbf{x}_i \mathbf{x}_i'$ och $\mathbf{T}^{-1}$ beräknas, det är här vikterna $d_i g_i$ som skall korrigeras genom kalibrering, där $i = 1, 2, 3, \dots, 6\ 335\ 136$
 $\mathbf{T}$ är matrisen med kvadrat- och produktsummor, här en 7×7 matris. Eftersom alla $x_{li} - x_{7i}$ är antingen 1 eller 0 så får matrisen ett enkelt utseende med okorrigerade antal på respektive utbildningsnivå i diagonalen och nollor för övrigt:

$$\mathbf{T} = \begin{bmatrix} 877\ 948 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1\ 042\ 680 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1\ 798\ 316 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1\ 041\ 337 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 861\ 213 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 671\ 318 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 42\ 324 \end{bmatrix}$$

2. Vektorn $\boldsymbol{\lambda}$ beräknas: $\boldsymbol{\lambda} = \mathbf{T}^{-1}(\mathbf{t}_x - \sum d_i g_i \mathbf{x}_i)$
Vektorn $\mathbf{t}_x$ är de 7 villkoren på antalen som korrigerats för tidsseriebrottet, och vektorn $\sum d_i g_i \mathbf{x}_i$ är motsvarande okorrigerade antal. Dessa antal hämtas från tabell 6U ovan, "1999 ny" och "1999 gammal".
3. De justerade vikterna blir : $w_i = d_i g_i (1 + \mathbf{x}_i' \boldsymbol{\lambda})$
De justerade vikterna ger de 7 serier vi haft som styrvillkor och används för att beräkna övriga länkade tidsserievärden.

Bild 6.55 Korrektion för tidsseriebrott av UREG 1999 med kalibrering av vikter

Person	Utbildningsnivå	x_{1i}	x_{2i}	x_{3i}	x_{4i}	x_{5i}	x_{6i}	x_{7i}	$d_i g_i$	w_i	t_x	$\Sigma d_i g_i x_i$	Skilnad
Pnr 1	Under 9 år	1	0	0	0	0	0	0	1,01693	0,99895	862 428	877 948	-15 520
Pnr 2	Grundskola 9 år	0	1	0	0	0	0	0	1,01693	0,91917	942 446	1 042 680	-100 234
Pnr 3	Gymnasium 2 år	0	0	1	0	0	0	0	1,01693	1,03375	1 828 054	1 798 316	29 738
Pnr 4	Gymnasium 3 år	0	0	0	1	0	0	0	1,01693	1,08398	1 109 997	1 041 337	68 660
Pnr 5	Högskola <3 år	0	0	0	0	1	0	0	1,01693	0,89805	760 539	861 213	-100 674
Pnr 6	Högskola $\geq$ 3 år	0	0	0	0	0	1	0	1,01693	1,19574	789 359	671 318	118 041
Pnr 7	Forskarutbildning	0	0	0	0	0	0	1	1,01693	1,01667	42 313	42 324	-11
Pnr 8	Uppgift saknas	0	0	0	0	0	0	0	0	0			
...	...	...	...	...	...	...	...	...	...	...			
Pnr 6386000	Högskola < 3 år	0	0	0	0	1	0	0	1,01693	0,89805			
Summa:									6 335 136	6 335 136			

Skilnaderna
ovan motsvarar
korrektionerna
för 1999 i
tabell 6T

På motsvarande sätt beräknas kalibrerade vikter för de övriga årens register. Vi visar nedan beräkningarna för 1998. Bortfallskorrektionen görs med värdet 1,01830 på vikten $d_i g_i$. De sju villkoren t_x och $\Sigma d_i g_i x_i$ är de länkade värdena respektive de tidsserievärden för 1998 som endast bortfallskorrigerats.

Bild 6.56 Korrektion för tidsseriebrott av UREG 1998 med kalibrering av vikter

Person	Utbildningsnivå	x_{1i}	x_{2i}	x_{3i}	x_{4i}	x_{5i}	x_{6i}	x_{7i}	$d_i g_i$	w_i	t_x	$\Sigma d_i g_i x_i$	Skilnad
Pnr 1	Under 9 år	1	0	0	0	0	0	0	1,01830	1,00299	915 075	929 043	-13 968
Pnr 2	Grundskola 9 år	0	1	0	0	0	0	0	1,01830	0,92909	939 556	1029 766	-90 210
Pnr 3	Gymnasium 2 år	0	0	1	0	0	0	0	1,01830	1,03318	1 858 639	1 831 874	26 764
Pnr 4	Gymnasium 3 år	0	0	0	1	0	0	0	1,01830	1,08046	1 074 101	1 012 306	61 794
Pnr 5	Högskola <3 år	0	0	0	0	1	0	0	1,01830	0,90862	750 649	841 256	-90 607
Pnr 6	Högskola $\geq$ 3 år	0	0	0	0	0	1	0	1,01830	1,18645	749 600	643 364	106 237
Pnr 7	Forskarutbildning	0	0	0	0	0	0	1	1,01830	1,01804	39 440	39 450	-10
Pnr 8	Uppgift saknas	0	0	0	0	0	0	0	0	0			
...	...	...	...	...	...	...	...	...	...	...			
Pnr 6386000	Högskola < 3 år	0	0	0	0	1	0	0	1,01830	0,90862			
Summa:									6 327 060	6 327 060			

Skilnaderna
ovan motsvarar
korrektionerna
för 1998 i
tabell 6T

6.5.4 Länkning av tidsserier på mikronivå med kombinationsobjekt

Det finns två begrepp som låter närbesläktade, men som har olika innebörd. I avsnitt 2.1 införs begreppet *relationsobjekt*. Relationen mellan två objekt kan göras till ett relationsobjekt, t.ex. jobb är en relation mellan en anställd och en arbetsgivare som görs till ett statistiskt intressant relationsobjekt. I avsnitt 6.2 införs begreppet *kombinationsobjekt*. Varje kombination av ett objekt och ett värde på en flervärd variabel kan göras till ett kombinationsobjekt.

I avsnitt 6.2 diskuteras skattningsmetoder vid flervärda variabler. När ett objekt samtidigt har flera variabelvärden på en flervärd variabel skapas kombinationsobjekt för varje kombination av objekt och möjligt värde. Varje sådant kombinationsobjekt får en vikt som används vid skattning av värden i tabellceller.

Samma metod kan även användas vid tidsseriebrott som beror på *ändrad klassifikation*, där nya och gamla kategorier inte är helt jämförbara. Nya och gamla koder kan ha följande relationer:

Bild 6.57 Översättning av gamla koder till nya

Relation mellan gamla och nya koder	Gammal kod	Kodnyckel	Ny kod	Kommentar
Ett-till-ett	1	● ————— ●	A	Inga problem, gamla koden 1 omkodas till nya koden A
Flera-till-ett	2	● ————— ●	B	Inga problem, gamla koderna 2 och 3 slås samman till nya koden B
	3	● ————— ●		
Ett-till-flera	4	● ————— ●	C	Ger problem, hur skall gamla koden 4 delas upp på nya koderna C och D?
		● ————— ●	D	
Flera-till-flera	5	● ————— ●	E	Samma problem som ovan, hur skall: - Gamla koden 5 delas upp i E och F? - Gamla koden 6 delas upp i E och F?
	6	● ————— ●	F	

Förutsättningen för den metod vi beskriver i detta avsnitt är att man har ett register för en viss tidpunkt eller period där varje objekt både är klassificerat enligt den gamla och den nya klassifikationen.

Om klassifikationen är en stabil egenskap så kan man koda om till ny klassifikation för identiska objekt i äldre registerversioner. Men en sådan omklassificering kan endast göras för den del av registerpopulationen som kan följas bakåt i tiden. När man inte kan omklassificera objektvis på detta sätt kan man använda metodiken med kombinationsobjekt som beskrivs nedan.

Problem uppstår när objekt med en gammal kod skall omklassificeras till en av flera nya koder. Detta löses genom att varje sådant objekt kombineras med varje möjlig ny kod och dessa kombinationer ges vikter som används vid skattning av värden i tabellceller.

Exempel: Ändrad näringsgrenskod i Företagsregistret

I detta exempel illustrerar vi metodiken med fiktiva data. År 3 ändras näringsgrensindelningen, de gamla koderna 1 och 2 ersätts med nya koder A och B. För år 3 har alla företagsenheter (FE) kodats både med den gamla och den nya koden. För år 1 och 2 finns enbart den gamla SNI-koden. I bild 6.58 beskrivs hur de gamla och nya koderna motsvarar varandra.

Bild 6.58 Nyckel mellan gammal och ny näringsgrenskod

Relation mellan koder	Gammal kod	Kodnyckel	Ny kod	Kommentar
Flera-till-flera	1	● ————— ●	A	Gamla 1:or blir nya A, men gamla 2:or skall delas upp på nya A och B
	2	● ————— ●	B	

Den gamla koden 1 ger inga problem då den alltid skall koda om till ny kod A. Men den gamla koden 2 ger problem, skall företag med gamla koden 2 få ny kod A eller ny kod B? Metoden med kombinationsobjekt innebär att företag med koden 2 ersätts med två kombinationsobjekt med koden A respektive B.

Bilden nedan visar Företagsregister för tre år med uppgifter om antal anställda (Anst) och förädlingsvärde (F-värde) för varje företagsenhet. FE 8 och 13 har två näringsgrenar och har därför två kombinationsobjekt vars vikter är kända.

Bild 6.59**Företagsregister år 1**

FE	SNI	Anst	F-värde	w_i
1	2	10	15	1
2	1	20	50	1
3	2	50	100	1
4	1	30	75	1
5	2	10	15	1
6	1	20	40	1
7	2	20	40	1
8	1	60	105	0.6
8	2	60	105	0.4
9	1	40	80	1
10	2	20	40	1
11	1	10	25	1
12	2	10	15	1

Företagsregister år 2

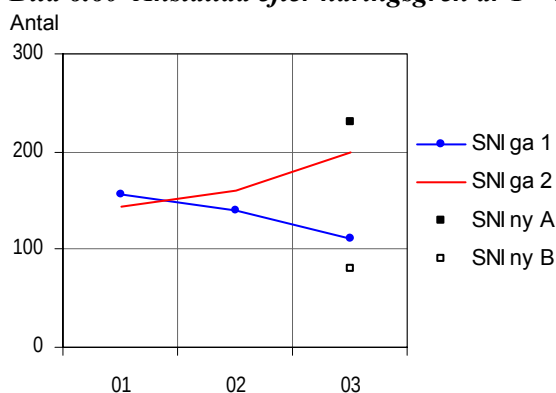
FE	SNI	Anst	F-värde	w_i
1	2	10	15	1
2	1	20	50	1
3	2	50	100	1
13	1	40	80	0.5
13	2	40	80	0.5
14	1	10	25	1
15	2	20	40	1
16	1	40	80	1
17	2	30	60	1
18	1	40	80	1
19	2	20	40	1
20	1	10	25	1
21	2	10	20	1

Företagsregister år 3

FE	SNI _{ga}	SNI _{ny}	Anst	F-värde	w_i
1	2	B	10	15	1
2	1	A	20	50	1
3	2	A	50	100	1
13	1	A	40	80	0.5
13	2	B	40	80	0.5
14	1	A	10	25	1
15	2	A	20	40	1
22	1	A	10	25	1
23	2	B	40	60	1
24	1	A	30	75	1
25	2	A	50	100	1
26	1	A	20	50	1
27	2	B	10	15	1

Med informationen i dessa tre register kan vi bilda tidsserier enligt gammal SNI. Dessutom kan för år 3 värden enligt ny SNI beräknas. Eftersom FE8 och FE13 bedriver verksamhet inom två näringsgrenar bildas kombinationsobjekt FE • SNI som ges vikter w_i som används när skattningarna för dessa tidsserier bildas på det sätt som beskrivs i avsnitt 6.2.

I bilden nedan kan vi se tidsseriebrottet år 3, men hur kan reviderade värden för tidigare år beräknas?

Bild 6.60 Anställda efter näringsgren år 1 – 3

I bild 6.61 nedan visas hur länken gör i detta fall. För år 3 känner vi både ny och gammal SNI-kod för alla företagsenheter i registret. Eftersom näringsgren är en relativt stabil variabel, använder vi den nya SNI-koden för år 3, även för år 1 och 2. Därigenom kan FE 1-3 få den nya SNI-koden för år 1 och 2 och FE13-15 kan få den nya koden för år 2.

Hur gör man med de FE som inte finns med i registret för år 3? De FE som har gammal SNI-kod 1 får den nya koden A. Eftersom gammal kod 2 både kan bli nya koden A och B skapas kombinationsobjekt för FE 5, 7, 8, 10, 12, 17, 19 och 21. Dessa har skuggats i bild 6.61 nedan. Vikterna w_i ga ger skattningar för gammal SNI, medan w_i anst och w_i fv ger skattningar av antal anställda respektive förädlingsvärde enligt ny SNI. Vikterna för ny SNI bildas med data för år 3

för FE med gammal SNI 2. Dessa uppgifter finns i tabell 6W som finns längst ner till höger under Företagsregistret för år 3 i bild 6.61.

- Antal anställda är 200 varav FE med ny SNI A har 120, dvs. $120/200 = 0,6$
- Förädlingsvärdet är 370 varav FE med SNI A har 240, dvs. $240/370 = 0,65$

Bild 6.61 Länkning på mikronivå med hjälp av kombinationsobjekt

Företagsregister år 1

FE	SNI ga	SNI ny	Anst	F-värde	w_i ga	w_i anst	w_i fv
1	2	B	10	15	1	1	1
2	1	A	20	50	1	1	1
3	2	A	50	100	1	1	1
4	1	A	30	75	1	1	1
5	2	A	10	15	1	0.6	0.65
5	2	B	10	15	0	0.4	0.35
6	1	A	20	40	1	1	1
7	2	A	20	40	1	0.6	0.65
7	2	B	20	40	0	0.4	0.35
8	1	A	60	105	0.6	0.6	0.6
8	2	A	60	105	0.4	0.24	0.26
8	2	B	60	105	0	0.16	0.14
9	1	A	40	80	1	1	1
10	2	A	20	40	1	0.6	0.65
10	2	B	20	40	0	0.4	0.35
11	1	A	10	25	1	1	1
12	2	A	10	15	1	0.6	0.65
12	2	B	10	15	0	0.4	0.35

Företagsregister år 2

FE	SNI ga	SNI ny	Anst	F-värde	w_i ga	w_i anst	w_i fv
1	2	B	10	15	1	1	1
2	1	A	20	50	1	1	1
3	2	A	50	100	1	1	1
13	1	A	40	80	0.5	0.5	0.5
13	2	B	40	80	0.5	0.5	0.5
14	1	A	10	25	1	1	1
15	2	A	20	40	1	1	1
16	1	A	40	80	1	1	1
17	2	A	30	60	1	0.6	0.65
17	2	B	30	60	0	0.4	0.35
18	1	A	40	80	1	1	1
19	2	A	20	40	1	0.6	0.65
19	2	B	20	40	0	0.4	0.35
20	1	A	10	25	1	1	1
21	2	A	10	20	1	0.6	0.65
21	2	B	10	20	0	0.4	0.35

Företagsregister år 3

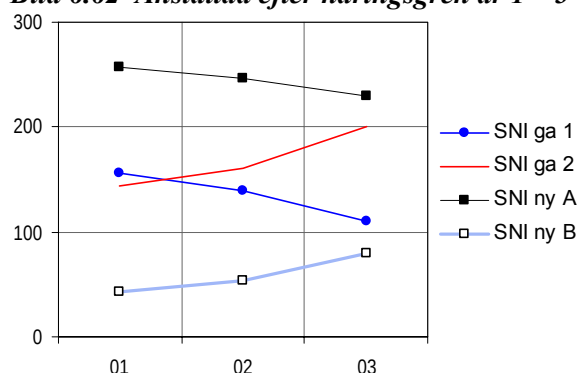
FE	SNI ga	SNI ny	Anst	F-värde	w_i ga/ny
1	2	B	10	15	1
2	1	A	20	50	1
3	2	A	50	100	1
13	1	A	40	80	0.5
13	2	B	40	80	0.5
14	1	A	10	25	1
15	2	A	20	40	1
22	1	A	10	25	1
23	2	B	40	60	1
24	1	A	30	75	1
25	2	A	50	100	1
26	1	A	20	50	1
27	2	B	10	15	1

Tabell 6W. SNI 2 under år 3

2	A	120	240
2	B	80	130
Summa		200	370

Med vikterna w_i anst och w_i fv kan länkade tidsserievärden för antal anställda respektive förädlingsvärde för år 1 och 2 beräknas. I bilden nedan jämförs tids-serierna avseende både gammal och ny näringsgrensindelning.

Bild 6.62 Anställda efter näringsgren år 1 – 3



I detta exempel har vikterna som används för att göra beräkningarna varit konstanta under alla år. Om länkningen avser en ny bransch som uppstått under de senaste åren, kan man låta vikterna avta bakåt i tiden. En fördel med denna metodik, som bygger på att kombinationsobjekt skapas, är att man inte först behöver länka serierna på makronivå.

Kvalitetsbegreppet för officiell statistik presenteras i SCB (2001a)

Kapitel 7. Registerstatistikens kvalitet

I Danmarks Statistik (1994) kapitel 4, sammanfattas de krav som ställs på ett registersystem för att man med framgång skall kunna använda administrativa data för statistiska syften. Registerstatistikens kvalitet blir bra om:

Systemet har bra täckning, innehåller många viktiga objekttyper, innehåller många viktiga relationer mellan objekt och innehåller många variabler.

Ett rikt innehåll, bra enhetsavgränsningar och nästan fullständig täckning i registren medför att registersystemet blir en bra grund för statistikproduktion.

I avsnitt 1.6 nämner vi att det finns både fördomar och berättigad kritik mot statistik baserad på administrativa data. Enligt fördomarna är registerstatistik billig men dålig jämfört med ”riktig” surveystatistik. Hur kan sådana attityder uppstå? Tidningsartiklar av den karaktär som ingår i exemplen nedan kan ge upphov till sådana attityder.

Dagens Nyheter den 2 mars 2003

Statistiken ljugar om mord

Nytt datasystem kan ligga bakom siffror som visar tre gånger fler fall än vad som skett.

Mikael Rying har skrivit en avhandling om skillnaden mellan antal anmälda och statistikförda mord i Sverige och hur många som faktiskt skett.

– Under åren 1990 -1998 var det i genomsnitt endast 57 procent av de anmälda fallen som verkligen var mord och dråp, säger Mikael Rying. De övriga fallen var allt från självmord och alkoholförgiftning till rena felkodningar i polisens rapportering.

År	Döda efter dödsorsak: Mord, dråp	Polisanmälda brott: Mord, dråp
96	110	199
97	94	157
98	98	185
99	108	188
00	90	175
Genomsnitt för 1996-2000:		
	100	181
	100/181 = 55 %	
Källa: Statistisk Årsbok för Sverige 2002 - 2003		

Detta exempel visar att man helst inte skall använda endast en administrativ källa. Grundidén med registersystemet är att många källor skall användas så att kvalitet och konsistens kan uppnås. Dessutom skall inte administrativa uppgifter användas direkt, utan administrativa data skall granskas och bearbetas så att de passar för statistikens behov. Problemet med de anmälda brotten är att rapporteringen inte klart skiljer mellan olika objekttyper: mordoffer, misstänkta gärningsmän och mordhändelser. Detta kan ge felkodningar.

En annan administrativ källa är Dödsorsaksregistret som ingår i SCB:s registersystem och är avstämt med Befolkningsregistret. Denna källa påverkas inte av de kvalitetsbrister som finns i uppgifterna avseende anmälda brott. Slutligen skulle knappast en urvalsundersökning ge mordstatistik av god kvalitet – ett frågeformulär till alla eller ett urval av landets polisstationer skulle ge upphov till dubbelt uppgiftslämnande av till stor del samma felaktiga uppgifter. Båda dessa registerbaserade undersökningar måste ha lättillgängliga kvalitetsdeklARATIONER så att användarna inte tolkar uppgifterna felaktigt.

En annan vanlig åsikt om administrativa data är att uppgiftslämnarna ger uppgifter som gynnar deras syften. I exemplet nedan vill man helst betala så lite skatt som möjligt, därför kan avdragen vara större än vad som är befogat.

Dagens Nyheter den 16 juli 2003

80 procent av svenska folkets avdrag är rent skattefusk

Skattebetalarna gör fel för miljarder i sina deklarationer. Krångliga skatteregler och en oklar lagstiftning gör det svårt för landets skattemyndigheter att kontrollera alla avdrag. Främst är det avdrag för aktieaffärer, förvaltningsavgifter och övriga avgifter som blir fel.

1. Avdrag för aktieförsäljningar

- 1/3 av aktieförsäljningarna innehåller fel
- 700 000 skattebetalare redovisar ca 50 miljarder kronor i vinster och ca 10 miljarder kronor i förluster.
- Skattefelet är svårbedömt, med det uppgår till miljardbelopp.
- Många oavsiktliga fel på grund av svåra regler.

2. Avdrag för förvaltningsavgifter

- 125 000 skattebetalare yrkade avdrag om totalt 515 miljoner kronor.
- 66 % av avdragen innehåller felaktigheter.
- Skattefelet kan totalt beräknas till ca 90 miljoner kronor.
- Avdrag för avgift till fondförvaltare är det vanligaste felet, avgiften dras automatiskt från fondresultatet.

3. Avdrag för övriga utgifter

- 700 000 skattebetalare yrkade avdrag om totalt 2,8 miljarder kronor.
- 82 % av avdragen innehåller felaktigheter.
- Skattefelet kan totalt beräknas till ca 700 miljoner kronor.

Rubriksättningen överdriver på flera sätt, ”80 procent” är överdrivet och ”skattefusk” är ofta missförstånd på grund av svåra regler:

1. Avdrag för aktieförsäljningar, visar att felen är till stor del oavsiktliga då reglerna är mycket svårtolkade.
2. Avdrag för förvaltningsavgifter, visar att felen är i genomsnitt 17 % (90/515) och det vanligaste felet kan vara oavsiktliga på grund av missförstånd.
3. Avdrag för övriga utgifter, 82 % av avdragen innehåller fel, innebär att värdet på avdragen var i genomsnitt 25 % (700/2800) felaktigt.

Ett annat perspektiv på dessa fel får man om skattefelet ovan jämförs med inkomsterna totalt för deklaranterna, felet är då i storleksordningen 0,3 %.

Att avdragen i deklarationerna är för höga, och skatten därigenom blir för låg, innebär inte att Inkomstregistrets statistik behöver få låg kvalitet, trots att den baseras på dessa deklarationer. Antag att vi har uppgifter för en deklarerant som anger för höga avdrag, men för övrigt deklarerar rätt:

Inkomst av tjänst	257 600:-	Inkomsten är korrekt
Avdrag för övriga utgifter	25 500:-	Avdraget är för högt men accepteras
Beskattningsbar inkomst	232 100:-	Beskattningsbar inkomst är felaktig enligt skattereglerna, men används inte för statistik
Skatt	100 000:-	Skatten är felaktig och för låg enligt skattereglerna, men är statistiskt sett korrekt, personen har faktiskt betalt denna skatt
Disponibel inkomst	157 600:-	Statistiskt korrekt

SCB:s statistik avseende förvärvsinkomster blir inte felaktig genom denna deklarerants uppgifter, inte heller statistiken avseende disponibel inkomst blir felaktig, då den bildas genom att beräkna skillnaden mellan inkomst och faktiskt inbetald skatt.

7.1 Vad är specifikt för registerstatistik?

Vad är statistisk kvalitet inom registerstatistiken? Holt (2001) påpekar att det är viktiga skillnader mellan surveyundersökningar med egen datainsamling och statistik som baseras på administrativa källor. Den viktigaste kvalitetsaspekten på registerbaserad statistik är enligt Holt relevans och inte noggrannhet.

Nanopoulos (2001) framhåller att länder som Danmark med väl integrerade registersystem behöver en annan begreppsapparat kring felen i statistiken än länder som i huvudsak gör urvals- och totalundersökningar. Även SCB:s vetenskapliga råd (2000) har rekommenderat att SCB utreder vad som skall menas med statistisk kvalitet i register och hur denna skall dokumenteras.

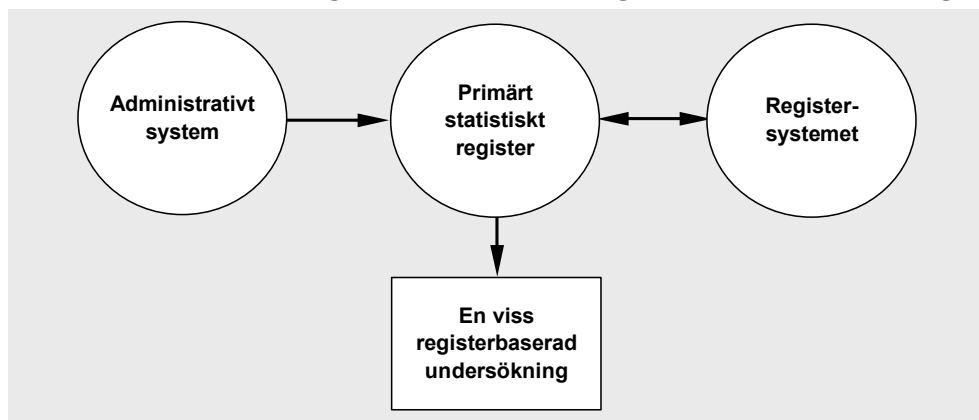
Det är viktigt att göra följande gränsdragningar när man diskuterar registerstatistikens kvalitet:

- Man måste skilja på kvaliteten hos ett statistiskt *register* och en registerbaserad *undersöknings* kvalitet.
- Man måste skilja på undersökningar med egen datainsamling och registerbaserade undersökningar. Risken finns annars att man okritiskt använder de traditionella felmodeller som utvecklats för urvals- och totalundersökningar.

Urvals- och totalundersökningar görs för *en* viss användning. När kvaliteten diskuteras avses kvaliteten hos de skattningar som görs i undersökningen. Ett statistiskt register däremot, kan ha många olika användningar, inte endast de undersökningar som görs idag utan även nya framtida undersökningar.

Bilden nedan illustrerar skillnaden mellan registret och den registerbaserade undersökningen. Dessutom visar bilden de två faktorer som är avgörande för registrets kvalitet, det administrativa systemet och det övriga registersystemet som är de två källor som det statistiska registret bygger på.

Bild 7.1 Kvalitet hos ett register, kvalitet hos en registerbaserad undersökning



Även en registerbaserad undersöknings kvalitet avser *en* speciell användning av registret. Med undersökningens kvalitet avses skattningarnas kvalitet, dvs. deras relevans och noggrannhet i förhållande till undersökningens syfte. Att beskriva kvaliteten innebär här att ange om undersökningen har *bra eller dålig* kvalitet. Med ett statistiskt registers kvalitet avses ingen speciell användning,

utan registrets egenskaper och användbarhet skall beskrivas allmänt. Att beskriva kvaliteten innebär här att ange *vilka egenskaper* som registret har. Av denna beskrivning skall framgå vilka användningar som registrets kvalitet medger. Registrets kvalitet eller egenskaper påverkar i sin tur kvaliteten hos de undersökningar som baseras på registret.

7.1.1 Undersökningar med egen datainsamling

Undersökningens kvalitet bestäms huvudsakligen av hur väl datainsamlingen fungerar. Denna process är ganska likartad i SCB:s alla undersökningar med egen datainsamling. Detta medför att om man anger de viktigaste kvalitetsfaktorerna får man ungefär samma lista för alla undersökningar:

- bortfallet,
- mätfelen och
- urvalsfelen för urvalsundersökningar.

Samma metodproblem finns i alla undersökningar vilket underlättar gemensamma diskussioner, riktlinjer och metodutveckling. Däremot påverkar inte olika urvalsundersökningar varandra, t.ex. så stör inte bortfallet i *Arbetskraftsundersökningen* andra undersökningar som *Undersökningen av befolkningens levnadsförhållanden* eller *Industrins leveranser och orderingång*.

7.1.2 Statistiska register

Registrets kvalitet bestäms av följande tre faktorer:

- det administrativa system som registret bygger på,
- de möjligheter som registersystemet ger att förbättra täckning, variabelinnehåll och konsistens, samt
- de bearbetningar som gjorts för att skapa registret.

Registrets kvalitet bestäms huvudsakligen av de administrativa system som registret bygger på, vilket illustreras av bild 7.1. Alla administrativa system är som regel unika, den administrativa databildningen sker t.ex. på helt olika sätt inom Folkbokföringen och Företagens inkomstdeklarationer. Databildningen inom de administrativa systemen skiljer sig som regel från SCB:s datainsamlingar. Även om man inom de administrativa systemen har uppgiftslämnare som fyller i blanketter, så är villkoren för uppgiftslämnandet helt annorlunda och styrs av administrativa regler och lagar.

Dessutom påverkas registrets kvalitet av vilka möjligheter registersystemet ger. Dels skall registret vara väl avstämt eller samordnat med det övriga registersystemet och dels skall systemet som helhet fungera bra.

Den tredje faktorn som avgör registrets kvalitet är de registerbearbetningar man har gjort när registrets skapades. Hur har registerpopulationen avgränsats, hur har variabelinnehållet bestämts och hur har man granskat data och åtgärdat eventuella brister enligt kapitel 5 och 6? Det administrativa systemet och de egna bearbetningarna har hittills uppfattats som specifika för varje register, vilket har försvårat erfarenhetsutbyte och metodutveckling, men registrets kvalitet påverkar andra registerbaserade undersökningar som vill utnyttja

Undersökningar med egen datainsamling:
Samma slags fel i alla undersökningar, de påverkar ej varandra

Registerbaserade undersökningar:
Olika fel i olika undersökningar, de flesta påverkar andra

registrets innehåll. Så är t.ex. det partiella bortfallet inom variabeln näringsgren i Företagsregistret ett gemensamt problem för alla register och undersökningar som innehåller näringsgren. Dessutom är alla beroende av att de ansvariga inom Företagsregistret löser detta kvalitetsproblem, annars kan inte statistik från olika register vara konsistent.

Om man anger de viktigaste kvalitetsfaktorerna får man troligen *inte* samma lista för olika registerbaserade undersökningar, däremot kommer kvaliteten i ett register som regel att påverka kvaliteten i andra register.

Även om man inte hittills haft ett gemensamt erfarenhetsutbyte i registerstatistiska metodfrågor, hoppas vi att en gemensam terminologi och ett gemensamt synsätt kommer att medföra att metodutvecklingen stimuleras.

7.1.3 Användarnas och producenternas syn på kvalitet

Hur används registerstatistik? Vilka kvalitetskrav har användarna? I Biemer och Lyberg (2003) diskuteras kvaliteten i urvalsundersökningar. Man utgår från *en* skattning och delar upp det totala felet för denna skattning i 12 olika felkomponenter. I urvalsundersökningens olika moment kan dels slumpmässiga och dels systematiska fel uppstå. *Slumpmässiga fel* gör skattningarna osäkra men leder inte till skevhet. Däremot ger *systematiska fel* upphov till skevhet, dvs. att man överskattar eller underskattar det sökta värdet.

Bild 7.2 Risken för slumpmässiga och systematiska fel för olika felkällor

Felkomponent	Risk för slumpmässigt fel	Risk för systematiskt fel
Relevansfel	Låg	Hög
Ramfel	Låg	Hög
Bortfallsfel	Låg	Hög
Mätfel	Hög	Hög
Databehandlingsfel	Hög	Hög
Urvalsfel	Hög	Låg

Källa: Biemer och Lyberg (2003), sidan 59

Detta är producentens syn på kvalitet. Som statistikproducent är det viktigt att veta vilka moment i undersökningen som fungerar bra respektive dåligt och denna kunskap används för att förbättra de processer där de största felen uppstår. Vi kommer i detta kapitel att diskutera motsvarande felkomponenter som förekommer vid registerbaserade undersökningar.

Platek och Särndal (2001) diskuterar kvaliteten hos den officiella statistiken, där de utgår från användarens synsätt. Användaren frågar sig: Kan jag lita på statistiken? Passar den för mina syften? Användarna vill ha en kvalitetsgaranti. Kan statistikerna ge denna garanti? Hur skall den utformas? Statistikbyråerna ger inte tillräckligt tydliga svar på dessa frågor. Datakvalitet betyder enligt Platek och Särndal olika för olika slag av personalkategorier på en statistikbyrå:

- metodstatistikerna engagerar sig för noggrannhet,
- ämnesstatistikerna engagerar sig för innehåll och presentation,
- IT-specialisterna vill att datasystem och bearbetningar fungerar bra,
- cheferna vill att budget och tidsplaner fungerar bra.

Registerstatistiken kan ibland användas genom att ett sammanfattande värde beräknas: *"Industriysselsättningen i vår kommun har ökat med 1,6 %."* Men ofta studeras flera stora tabeller med 100-tals skattningar där man söker efter sambandsmönster och tidsseriemönster. De mönster man finner resulterar i slutsatser: *"Arbetskraftens struktur i vår kommun har försämrats med avseende på ålder och utbildningsnivå så att högteknologiska branscher hämmas."*

En sådan slutsats bygger på många jämförelser med andra kommuner och över tiden. Här kan man inte diskutera i termer av siffermässiga fel i alla de 100-tals skattningar som slutsatsen bygger på. Att dessutom tänka sig att varje fel kan delas upp i 12 felkomponenter som skall bedömas var för sig är orimligt. Vad användaren vill veta är om man kan dra denna slutsats eller inte. Finns det några statistiska fallgropar som användaren inte tänkt på?

Från detaljkunskaper om kvalitetskomponenter till en helhetsbild av kvaliteten

Diskussionen i detta avsnitt kan sammanfattas så att olika personalkategorier har ingående kunskaper om många olika faktorer som påverkar kvaliteten. Metodstatistikern tänker i termer av undersökningens olika faser och alla de felkällor som kan finnas i de olika faserna. IT-specialisterna tänker på produktionssystemet och alla de bearbetningsfel som kan uppstå. Båda dessa kategorier kan sakna den helhetsbild som användaren behöver.

Ämnesstatistikerna är de som har nära kontakt med statistikens användning och borde därför skapa sig denna helhetsbild av kvaliteten. För att kunna skaffa sig denna helhetsbild måste ämnesstatistikerna ha goda kontakter med metodstatistikerna och IT-specialisterna och dessutom förstå användarnas behov.

Detaljkunskaperna om de olika kvalitetskomponenterna skapas genom kvalitetssäkring av de arbetsmoment som ingår i arbetet med att skapa det statistiska registret. Dessa moment beskrivs i kapitel 5 och 6. Med dessa detaljkunskaper om registret gör man därefter en sammanfattande helhetsbedömning av dels registrets kvalitet och dels den registerbaserade undersökningens kvalitet.

7.2 Kvalitetssäkring – detaljkunskaper om registrets egenskaper

Arbetet med kvalitetssäkring syftar till att undersöka och åtgärda kvalitetsbrister som kan uppstå i olika delmoment i arbetet med ett register. Den traditionella felmodellen följer en urvalsundersökningens olika steg. På motsvarande sätt beskriver vi kvaliteten i registerstatistik med grupper av indikatorer för de olika momenten i arbetet med att skapa ett statistiskt register:

Bild 7.3 Stegen då ett register skapas och motsvarande kvalitetsindikatorer

Arbetsgången då ett statistiskt register skapas	Kvalitetsindikatorer – registrets egenskaper
1. Problemformuleringsfasen	Grupp 1: Utvecklingsförmåga och ämneskompetens
2. Inventeringsfasen	
3. Planeringsfasen	
4. Leverantörskontakter, mottagande av administrativa data	Grupp 2: Leverantörskontakter och granskning
5a. Bearbetningsfasen - objektmängden	Grupp 3: Objektmängdens egenskaper
5b. Bearbetningsfasen - objekttypen	Grupp 4: Objektavgränsningens egenskaper
5c. Bearbetningsfasen - variablerna	Grupp 5: Variablernas egenskaper
6. Dokumentation	Grupp 6: Dokumentationens egenskaper

Pettersson m.fl. (1993) diskuterar kvaliteten i administrativa data

Sandgren (1993) innehåller en översikt av metodstudier under 15 år avseende statistik baserad på administrativa data

7.2.1 Problemformuleringsfasen – hur har den påverkat registret?

Vilka är de statistiska behov och önskemål som skall uppfyllas med registret? Har man kontakter med kvalificerade användare eller forskare? Gör man egna kvalificerade analyser och rapporter?

Om man vet vilka krav som dessa kvalificerade användare och de egna analyserna ställer på registret, får man en uppfattning om vilka användningar av registret som är möjliga. Dessutom finns värdefulla erfarenheter hos dessa kvalificerade användare som bör ha dokumenterats. Kvalificerade användningar är en viktig faktor för att utveckla ett register och registrets användbarhet.

7.2.2 Inventeringsfasen – hur har den påverkat registret?

Vilka källor har använts för att skapa det nya registret? Vilka administrativa källor och vilka befintliga statistiska register i systemet har använts? Har källorna tillsammans ett rikt innehåll? Om flera källor har integrerats i registret, så innebär detta att ett viktigt integrationsarbete redan har gjorts, detta kan nya användare ha glädje av.

Finns andra källor med anknytning till undersökningsområdet som inte har använts? Söker man aktivt efter nya källor?

7.2.3 Planeras förändringar?

Planeras förändringar eller förbättringar av registerpopulation, objektavgränsningar eller variabelinnehåll? Om förändringar och förbättringar planeras, så är detta dels ett tecken på brister i det nuvarande registret, men även ett tecken på att det pågår utvecklingsarbete för att öka registrets användbarhet.

De tre avsnitten ovan kan sammanfattas i följande bild:

Grupp 1. Utvecklingsförmåga och ämneskompetens hos de som skapat registret

Kvalitetsindikator	Är tecken på	Hur bedöms/mäts detta?	Se avsnitt
1a. God kontakt med användare	Utvecklingspotential	Kvalitativ bedömning	7.2.1
1b. Gör egna analyser	Ämneskompetens		
1c. Olika källor har integrerats	Mångsidig användning avses		7.2.2
1d. Söker efter nya källor	Utvecklingsvilja		
1e. Förändringar planeras			7.2.3

7.2.4 Leverantörskontakter och granskning

Hur sker leverantörskontakterna? Är kontakterna djupa så att leverantören förstår de behov som statistiken ställer och mottagaren förstår det administrativa systemets villkor? Återkommer kontakterna regelbundet så att mottagaren får information om kommande förändringar i god tid?

Hur tar man emot data och metadata från leverantören? Hur granskas mottagna data? Görs endast mottagningskontroll eller gör man även konsistenskontroll där flera källor jämförs? Vilka erfarenheter ger granskningen? Dokumenteras frekvenser av olika feltyper? Hur återrapporterar man detta till leverantören?

Dessa frågor sammanfattas i bilden nedan:

Grupp 2. Indikatorer på hur leverantörskontakter och granskning fungerar

Kvalitetsindikator	Är tecken på	Hur bedöms/mäts detta?	Se avsnitt
2a. En grupp från SCB har regelbundna kontakter med leverantören, eller en enskild SCB:are har sporadiska kontakter	Stor ämneskompetens respektive begränsad ämneskompetens	Kvalitativ bedömning	5.7.1 5.7.6
2b. Ogiltiga poster och dubbletter upptäcks vid granskningen	Administrativa källans kvalitetsbrister	Antal/andel ogiltiga poster och dubbletter	5.7.5
2c. Skillnader mellan registerpopulationen och basregistret upptäcks	Undertäckning Övertäckning	Antal/andel saknade och främmande objekt	
2d. Falsa träffar upptäcks vid granskningen	Länkar anger fel relation eller fel identitet	Antal/andel falska träffar	
2e. Objekt och/eller variabelvärden saknas, upptäcks vid granskningen	Ofullständig leverans	Antal/andel saknade objekt/variabelvärden	
2f. Uppenbart felaktiga variabelvärden upptäcks vid granskningen	Administrativa källans kvalitetsbrister	Antal/andel fel av olika slag. Felens storlek	5.7.6
2g. Utom mottagningskontroll görs konsistenskontroll av flera källor	Mångsidig användning avses	Kvalitativ bedömning	

7.2.5 Bearbetningsfasen – hur har objektmängden skapats?

Hur har befintliga källor bearbetats för att det nya registret ska innehålla den önskade objektmängden? Har administrativa data kontrollerats, granskats och åtgärdats så att objektmängden blir den önskade? Hur har olika källor matchats och objekt selekterats? Har tidsreferenser bearbetats för att skapa objektmängden för den avsedda tidpunkten eller tidsperioden?

I de två bilderna nedan anges dels indikatorer som avser basregister och dels indikatorer som gäller allmänt.

Grupp 3.1. Indikatorer på objektmängdens egenskaper i basregister

Kvalitetsindikator	Detta orsakar	Hur bedöms/mäts detta?	Se avsnitt
3.1a. Eftersläpning i rapportering	Över- och undertäckning i aktuell registerversion, ger fel i tvärsnittsdata och fel som ändras över tiden	Jämför aktuell registerversion med det register som baseras på alla tillgängliga källor när all information finns. Ange revisionsfelens storlek och hur lång tid eftersläpningen är. Ange även förändringsfrekvenser	Se nedan
3.1b. Brister i rapportering av att objekt upphört	Övertäckning, ger fel i tvärsnittsdata och fel som ändras över tiden	Orimliga värden i tabeller kan bero på detta. Använd befintliga undersökningar med egen datainsamling: Postreturer, ej anträffad vid telefonintervju. Jämför med annan registerinformation Ange övertäckningens omfattning	6.3
3.1c. Objekt med felaktiga uppgifter	Över- och undertäckning	Undersök antal/andel felaktiga aviseringsuppgifter som rättas eller annulleras	Se nedan

Eftersläpning i rapportering: Inom Befolkningsregistret planerar man att kontinuerligt följa olika månadsvisa kvalitetsindikatorer avseende fördröjningar vid registrering av födselar, dödsfall och flyttningar. Olika lokala skattemyndigheter kan ha olika fördröjningsmönster.

Inom Företagsregistret avser man dels att följa förändringsfrekvenser per månad avseende införda och avregistrerade företag, adress- och branschändringar och dels att följa aktualiteten på dessa uppgifter.

Felaktiga uppgifter avseende objekt: Inom Befolkningsregistret har man en särskild kvalitetsdatabas där man sparar alla felaktiga aviseringsuppgifter som senare rättats eller annullerats av Skatteverket.

Grupp 3.2. Indikatorer på objektmängdens egenskaper allmänt

Kvalitetsindikator	Är tecken på	Hur bedöms/mäts detta?	Se avsnitt
3.2a. Registerpopulationen stämmer inte med basregistret	Konsistensbrist, bristande sammanvändbarhet	Matcha registret mot det aktuella basregistret, analysera mismatch	5.4.3 - 5.4.4 kap 11 Se nedan
3.2b. Den administrativa källan innehåller objekt som saknas i basregistret	Undertäckning i basregistret	Analysera omfattning och orsak till denna mismatch, rapportera till basregistret	5.4.6
3.2c. Basregistret innehåller objekt som saknas i källan	Undertäckning eller bortfall i källregistret	Analysera omfattning och orsak till denna mismatch	5.5.7
3.2d. Matchningsfel – falska träffar	Fel relation, objekten hör ej samman trots samma värde på kopplingsvariablerna	Jämför variabler från de register som matchats	5.4.6 5.5.5 5.7.2

Registerpopulationen stämmer inte med basregistret: Jordbruksstatistikens jordbruksföretag stämmer inte med Företagsregistrets jordbruksföretag, utbildningsstatistikens utbildningsföretag stämmer inte med Företagsregistrets utbildningsföretag och energistatistikens energiföretag stämmer inte med Företagsregistrets energiföretag. Ett ytterligare exempel på inkonsistenta populationer är inkvarteringsstatistikens företagspopulation som har skapats oberoende av Företagsregistret.

7.2.6 Bearbetningsfasen – hur har objekten avgränsats?

Vilka bearbetningar har gjorts för att kontrollera och ändra objektavgränsningarna? Har administrativa data kontrollerats, granskats och åtgärdats så att objektavgränsningarna blir de önskade? Har härledda objekt bildats i det nya registret? Kontrolleras kvaliteten hos objektavgränsningarna? Gör man registerhållningsenkäter eller evalveringar? Hur stora är felen?

Grupp 4. Indikatorer på objektavgränsningens egenskaper

Kvalitetsindikator	Medför/är tecken på	Hur bedöms/mäts detta?	Se avsnitt
4a. Objektavgränsning avviker från norm	Konsistensbrist, bristande sammanvändbarhet	Kvalitativ bedömning	Se nedan kap 11
4b. Objektavgränsning ej jämförbar över tiden	Konsistensbrist, bristande sammanvändbarhet	Kvalitativ bedömning	5.9
4c. Görs registerhållningsenkäter?	Kvalitetsmedvetande	Hur ofta görs sådana enkäter och till vilka kategorier av objekt?	5.4.5, 5.4.6 5.5.3, 5.7.2

Objektavgränsning avviker från norm: I Fastighetstaxeringsregistret är objekt (och även variabler) administrativt definierade och överensstämmer inte med andra SCB-undersökningar, t.ex. så definieras *hyreshus* på olika sätt. På liknande sätt avviker den skattemässiga familjeavgränsningen i Befolkningsregistret från hushållsavgränsningarna i andra SCB-undersökningar.

7.2.7 Bearbetningsfasen – hur har variablerna skapats?

Vilka bearbetningar har gjorts för att skapa de avsedda variablerna? Har variablerna i de administrativa källorna kontrollerats, granskats och åtgärdats? Hur stort är det partiella bortfallet och hur har bortfallet åtgärdats? Från vilka källor har variablerna hämtats?

Finns det mätfel i variablerna? Finns det klassifikationsfel i indelningsvariablerna? Använder man urvalsundersökningar för att undersöka detta? Gör man särskilda evalveringsundersökningar? Har man gjort mättekniska studier av de administrativa blanketterna för att upptäcka källor till mätfel?

Grupp 5. Indikatorer på variablernas egenskaper

Kvalitetsindikator	Medför/är tecken på	Hur bedöms/mäts detta?	Se avsnitt
5a. Matchningsfel – uteblivna träffar	Fel i kopplingsvariabel	Identiska objekt har olika värde på kopplingsvariablerna.	Se nedan
5b. Matchningsfel – falska träffar	Fel i kopplingsvariabel	Fel relation, objekt hör ej samman trots samma värde på kopplingsvariablerna. Jämför variabler från de register som matchats, orimliga samband indikerar fel	5.4.6 5.5.5 5.7.2
5c. Variabeldefinition avviker från normen	Konsistensbrist, bristande sammanvändbarhet	Kvalitativ bedömning	Kap 11 Se nedan
5d. Partiellt bortfall	Administrativa källans kvalitetsbrister	Omfattning av och orsaker till partiellt bortfall anges för varje variabel	5.6.2, 5.7 6.4
5e. Eftersläpning i rapportering	Revisionsfel	Jämför preliminära skattningar med de definitiva skattningarna	Se nedan
5f. Felaktiga preliminära värden	Revisionsfel	Undersök antal/andel felaktiga uppgifter som rättas eller annulleras	Se nedan
5g. Variabel används ej av leverantören	Mätfel	Undersök om variabeln används och granskas av leverantören	Se nedan
5h. Misstänkt låga / höga skattningar	Mätfel	Genom urvalsundersökningar och mättekniska studier kan mätfelet bedömas	Se nedan
5i. Kodningsfel	Klassifikationsfel	Samma datamaterial kodas oberoende en gång till. Med denna kontrollkodning kan kodningsfelen skattas	5.8.4
5j. Stympad flervärd variabel	Fel i tvärsnittsdata och över tiden	Registret innehåller endast viktigaste värdet på den flervärda variabeln	6.2
5k. Klassifikationsfel i indelningsvariabel	Strukturfel Samband försvagas	Kontrollundersökningar	Se nedan
5l. Modellfel i härledd variabel	Mätfel eller klassifikationsfel	Genom urvalsundersökningar kan felen bedömas och skattas	5.8.3

Matchningsfel – uteblivna träffar: Länken mellan Företagsregistret och Fastighetsregistret är arbetsställes besöksadress. Om man med denna länk försöker koordinatsätta arbetsställen har det visat sig att det inte gick att koordinatsätta arbetsställen för ca 30 % av de sysselsatta i den kommun där matchningen gav sämst utfall. Efter granskning av alla arbetsställen med minst 25 anställda kunde andelen mismatch i sämsta kommunen minskas till ca 10 %.

Variabeldefinition avviker från normen: Utrikeshandelns definitioner avviker mot Nationalräkenskapernas och Industristatistikens definitioner. Skillnaderna beror ibland på att olika kontaktpersoner på företagen fyller i olika enkäter.

Eftersläpning i rapportering: Tidsfördröjningar vid månadsrapporteringen av bostadsbyggandet ger underskattningar av antal påbörjade lägenheter. De preliminära skattningarna kommer senare att revideras upp. Det är viktigt att följa hur lång fördröjningen är och att göra aktuella skattningar av revisionsfelets storlek. Man bör skatta både revisionsfelets systematiska och slumpmässiga komponenter.

Eftersläpning i rapportering: Förseningar i månadsvisa rapporteringen av lönesummor, arbetsgivaravgifter etc. ger upphov till bortfall. De preliminära skattningarna bygger på imputerade värden och kommer att revideras när uppgifterna kommit in.

Felaktiga preliminära värden: Uppgifter från skattskyldiga kan senare komma att ändras, både genom att den skattskyldige rättar tidigare uppgifter och genom att skattemyndigheten beslutar om ändringar.

Används ej av leverantören: På kontrolluppgifterna skall arbetsgivarna ange första respektive sista månad som anställningen avser. Denna uppgift används inte av skattemyndigheterna varför dessa tidsmarkeringar är av låg kvalitet.

Misstänkt låga/höga skattningar: Småhusägare tenderar att underskatta bostadsytan i samband med fastighetstaxeringar, då detta kan leda till lägre fastighetsskatt.

Klassifikationsfel i indelningsvariabel: Populationens struktur, dvs. fördelning enligt indelningsvariabeln (t.ex. näringsgren) blir felaktig. Samband försvagas då jämförelser mellan kategorier störs. Antag att man vill jämföra industri- och tjänsteföretag med avseende på lönsamhet. Utan klassifikationsfel i näringsgren gäller det sanna sambandet enligt tabell 7A och med 10 % klassifikationsfel ser sambandet ut som i tabell 7B nedan.

Tabell 7A. Utan klassifikationsfel

	Omsättning	Vinst	Lönsamhet
Industriftg	300	30	10 %
Tjänsteftg	250	10	4 %

Tabell 7B. Med 10% klassifikationsfel

	Omsättning	Vinst	Lönsamhet
	300-30+25=295	30-3+1=28	9,5 %
	250-25+30=255	10-1+3=12	4,7 %

7.2.8 Dokumentation som del av kvalitetssäkringen

Även dokumentationsarbetet är en viktig del av kvalitetssäkringen. Felaktig och okritisk användning av administrativa data kan förhindras med metadata som ger information om jämförbarhetsproblem. Ändringar i det administrativa systemet ger jämförbarhetsproblem och måste därför dokumenteras annars finns risk för fallgropar som leder till felaktiga slutsatser.

Eftersom ett statistiskt register skall kunna användas av flera användare som utnyttjar registersystemet, måste alla register vara dokumenterade på sådant sätt att alla kan utnyttja dokumentationen. Metadata har en mycket betydelsefull roll vid arbetet med registerstatistik. När man sambearbetar olika register måste man känna till definitioner och jämförbarhetsproblem. Det är viktigt att även bearbetningsmetoderna dokumenteras så att metodutveckling och erfarenhetsutbyte underlättas.

Grupp 6. Indikatorer på dokumentationens egenskaper

Kvalitetsindikator	Medför/är tecken på	Hur bedöms/mäts detta?	Se avsnitt
6a. Definition av registerpopulationen	Statistikens innehåll	Är definitionen fullständig: Objekttyp, vilka ingår, tidsreferens, geografisk avgränsning, objektens relation till området	5.4.1
6b. Variabeldefinitioner	Statistikens innehåll	Är definitionen fullständig: Objekttyp, variabelns tidsreferens, mätmetod, skala	5.6.1
6c. Populationen jämförbar över tiden	Tidsseriekvalitet	Tidsseriebrott och dess effekter Vilka bearbetningar har gjorts för att skapa jämförbarhet över tiden?	6.5
6d. Variabler jämförbara över tiden	Tidsseriekvalitet	Tidsseriebrott och dess effekter Vilka bearbetningar har gjorts för att skapa jämförbarhet över tiden?	6.5
6e. Populationens objekt jämförbara över tiden	Longitudinell kvalitet	Har objekten stabila identiteter? Kan de följas över tiden? Vilka bearbetningar har gjorts för att skapa jämförbarhet över tiden på objektnivå?	5.9
6f. Objektens variabelvärden jämförbara över tiden	Longitudinell kvalitet	Lämpar sig variabeln för att beskriva förändringar i variabeln på objektnivå?	5.9

7.2.9 Nivåer för kunskap om registrets kvalitet

Ofta finns en kvalitativ kunskap då den registeransvarige genom sitt ämneskunnande, genom sina iakttagelser och erfarenheter väl känner till förekomsten av ett antal feltyper men kan inte mäta hur vanliga eller betydelsefulla de är. Även sådan vag kunskap bör dokumenteras, för att dels varna användare för okritisk användning och dels kunna användas för att peka ut behov av förbättringar. Kunskapen om olika aspekter på registrets kvalitet kan klassificeras efter sin skärpa på någon av följande nivåer enligt Lindström (1999):

1. Ingen kunskap alls.
2. Vag kvalitativ information baserad på bedömningar.
3. Dokumentation av produktionsprocesserna och deras egenskaper.
4. Systematisk kvalitativ information baserad på undersökningar.
5. Kvantitativa indikatorer, t.ex. omfattning av mismatch.
6. Kvantitativa mått på skattningskvalitet.

De grupper av indikatorer som beskrivs i bilderna ovan, kan användas för att beskriva registrets egenskaper på någon av nivåerna 2 – 6.

7.2.10 Vad säger dessa indikatorer om registrets kvalitet?

Listan med indikatorerna ovan kan användas när man vill ha uppslag till hur ett visst registers egenskaper skall undersökas och beskrivas. Eftersom förutsättningarna, dvs. de administrativa källorna och möjligheterna att använda registersystemet, varierar mellan olika register kan det vara olika indikatorer som är viktiga för olika register. Man väljer då ut de viktigaste indikatorerna för det register man skall kvalitetssäkra och beskriva.

Många av de kvalitetsindikatorer vi angett ovan är unika för register och registerbaserade undersökningar och förekommer inte vid undersökningar med egen datainsamling.

7.3 Helhetsbedömning av kvaliteten

Med dessa detaljkunskaper om registret som grund gör man därefter en sammanfattande helhetsbedömning av dels registrets kvalitet och dels den registerbaserade undersökningens kvalitet.

Kvalitet hos register respektive registerbaserade undersökningar

Registrets kvalitet skall beskrivas allmänt, så att en potentiell användare kan bilda sig en uppfattning om registrets användbarhet för denna användares syften. Kvaliteten skall då beskrivas för de olika användningsområden som kan vara aktuella. Vi skiljer på tre olika sätt att använda register:

- *Tvärsnittskvalitet*, vilka jämförelser kan göras inom samma register?
- *Tidsseriekvalitet*, vilka jämförelser kan göras över tiden på aggregerad nivå?
- *Longitudinell kvalitet*, vilka jämförelser kan göras på mikronivå över tiden?

En registerbaserad undersökningens kvalitet skall beskrivas för *en* viss användning av registret. I bilden nedan jämförs sätten att beskriva kvaliteten.

Bild 7.3 Kvaliteten för register och registerbaserade undersökningar

	Register	En viss undersökning
Relevans, användbarhet	Endast definitioner anges	Är definitionerna adekvata och funktionella? Detta bör diskuteras utförligt
Tvärsnittskvalitet	Vilka jämförelser kan göras?	Kvaliteten anges endast för den aktuella användningen
Tidsseriekvalitet	Vilka jämförelser kan göras?	
Longitudinell kvalitet	Vilka jämförelser kan göras?	

7.3.1 En helhetsbedömning av ett registers kvalitet

Helhetsbedömningen av ett statistiskt registers kvalitet innehåller följande:

- Här anger man de definitioner av registerpopulation och variabler som används, men man behöver inte värdera definitionerna.
- Vilka jämförelser kan göras inom samma register?
Exempel: Om man inom regional statistik vill jämföra kommuner, är då det partiella bortfallet ungefär lika stort i alla kommuner, eller kan man med hjälp av imputerade värden göra jämförelserna?
- Vilka jämförelser kan göras över tiden på aggregerad nivå?
Exempel: Är det partiella bortfallet ungefär lika stort alla år, eller kan man med hjälp av imputerade värden göra jämförelser över tiden? Har det skett förändringar i administrativa data som försvårar jämförelser över tiden? Har man gjort bearbetningar av registret för att öka jämförbarheten mellan år?
- Vilka jämförelser kan göras på mikronivå över tiden? Rättvisande longitudinella jämförelser ställer de högsta kraven på ett register. Vilka kontroller och bearbetningar har man gjort för att möjliggöra detta?

För att kunna göra denna helhetsbedömning krävs ingående kunskap om registret, dvs. en utförlig dokumentation är nödvändig.

7.3.2 En helhetsbedömning av en registerbaserad undersöknings kvalitet

Helhetsbedömningen av en undersöknings kvalitet innehåller följande:

- Hur definieras undersökningens *målpopulation*? Är denna definition adekvat med hänsyn till undersökningens syfte? Finns det några viktiga skillnader mellan mål- och registerpopulation, den faktiska objektmängden?
- Hur definieras undersökningens *variabler*? Är definitionerna adekvata med hänsyn till undersökningens syfte? Denna diskussion inriktas mot de för undersökningen mest betydelsefulla variablerna, de viktigaste aggregeringsvariablerna och indelningsvariablerna som analysen bygger på.
- Vilka *jämförelser* vill man göra i undersökningen? Medger registrets kvalitet att man gör dessa jämförelser?
- Görs *skattningarna* på bästa sätt? Både för tvärsnittskvaliteten och tids-seriekvaliteten är det viktigt att flervärda variabler och partiellt bortfall åtgärdas genom lämpliga beräkningsmetoder. För tidsseriekvaliteten är det viktigt att tidsseriebrott länkas.
- *Tolkas resultaten* på ett rimligt sätt med hänsyn till skattningarnas osäkerhet? Tabeller baserade på statistiska register skall inte uppfattas som exakt information utan även här kan det finnas slumpvariation och andra felkällor och det är därför viktigt att inte övertolka resultaten.

För att kunna göra denna helhetsbedömning krävs kunskap om registrets kvalitet och en ingående ämneskunskap, dvs. man måste vara väl insatt i de undersökningsproblem som skall belysas.

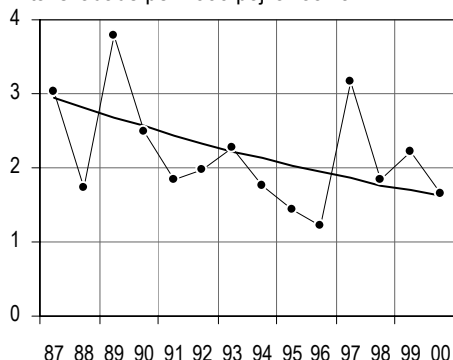
Slumpvariation i registerdata

Det kan finnas flera källor till slumpvariation i registerdata: värden imputeras med hjälp av slumpstal samt slumpmässiga mätfel, klassifikationsfel och revisionsfel. *Naturlig slumpvariation* är ett annat viktigt skäl till att skillnader och förändringar inte skall övertolkas. I bilderna nedan visas faktiska olycksfrekvenser, men tillfälliga variationer i faktiska frekvenser får inte tolkas som att den underliggande olycksrisken har ändrats. I små regioner med få barn blir den naturliga slumpvariationen kraftig (bild A), medan i stora regioner med många barn blir slumpvariationen mindre (bild B).

Bild 7.4 Vägtrafikolyckor, pojkar 0-12 år, faktiska värden och skattade trender

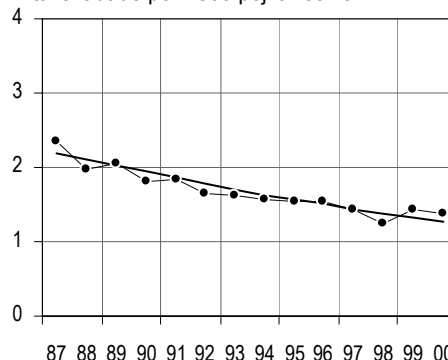
A. Gotlands län

Antal skadade per 1000 pojkar och år



B. Västra Götalands län

Antal skadade per 1000 pojkar och år



I Gotlands län finns ca 5 000 pojkar 0-12 år, i Västra Götalands finns ca 125 000 pojkar 0-12 år, dvs ca 25 gånger fler. Den naturliga slumpvariationen bör då vara 5 gånger kraftigare i bild A än i bild B

Källa: Barns skador i Sverige, SOU 2002:99

Kapitel 8. Dokumentation och metadata

Detta kapitel baseras på det arbete kring metadata som bedrivs inom Registerprojektet.

Sedan dess har arbete med ett variabelhanteringssystem påbörjats

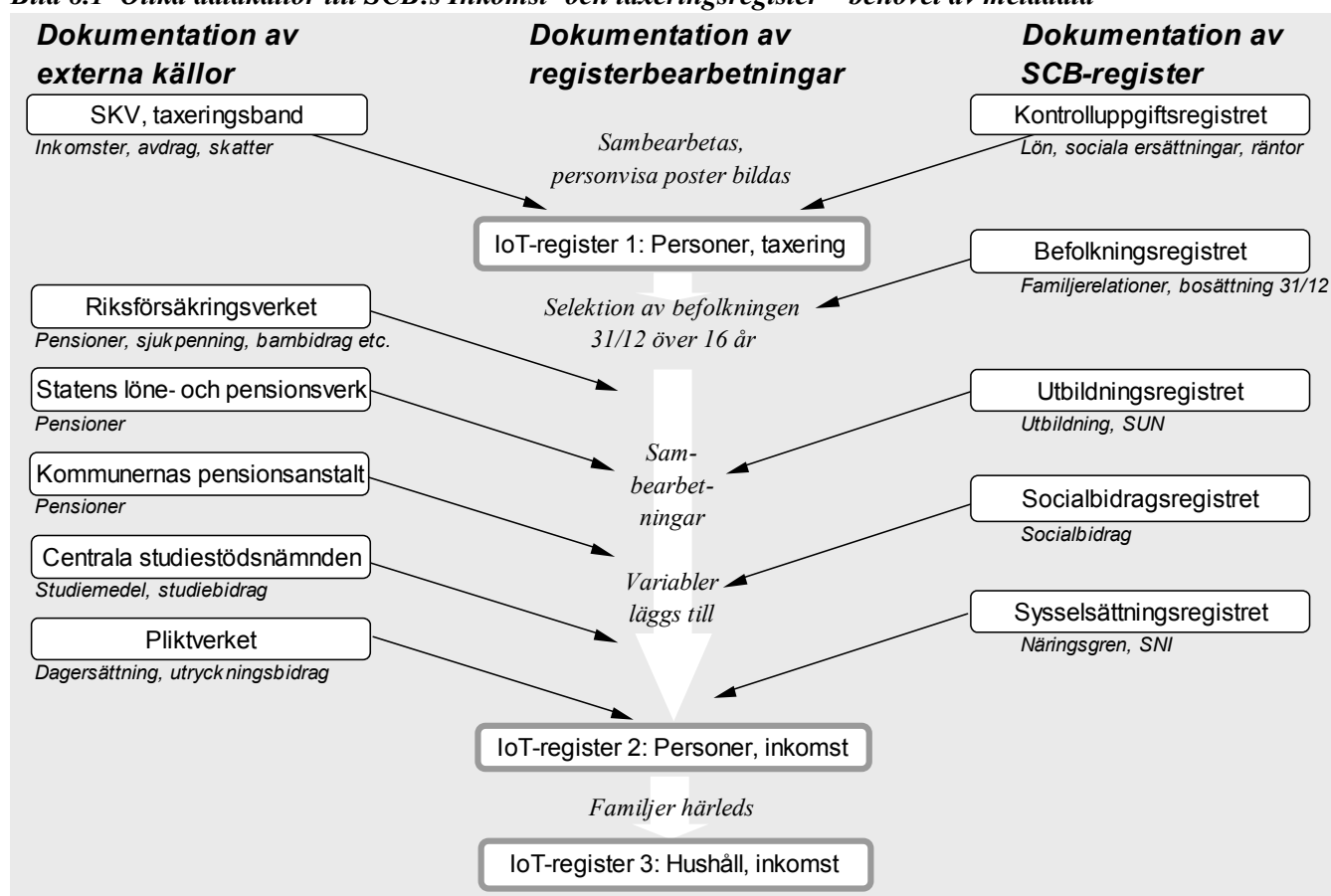
Alla undersökningar, datamatriser och databaser behöver dokumenteras. Genom detta dokumentationsarbete skapas *statistiska metadata*, dvs. information som beskriver de statistiska uppgifterna och undersökningsprocessen. Man skiljer mellan *mikrometadata* som beskriver innehållet i datamatriser med mikrodatabaser, dvs. uppgifter som avser enskilda objekt, och *makrometadata*, som beskriver innehållet i statistiska tabeller, dvs. uppgifter som avser makrodatabaser som bildats genom att aggregera uppgifter för grupper av objekt.

Mikrometadata behövs både av de som arbetar med en undersökning och av de som vill använda den färdiga undersökningen. Vi behandlar här endast de metadatabehov som man har när man arbetar med att skapa statistiska register. Vi diskuterar registersystemets behov av metadata, inte de tekniska lösningarna.

8.1 Primära statistiska register – behovet av metadata

Statistiska register skapas genom att sambearbeta olika källregister. Detta ställer speciella krav på metadata för registerbaserade undersökningar som skiljer sig från de metadatabehov som finns för undersökningar med egen datainsamling. Exemplet nedan visar de källor som behövs för att skapa ett register.

Bild 8.1 Olika datakällor till SCB:s Inkomst- och taxeringsregister – behovet av metadata



För att skapa Inkomst- och taxeringsregistret (IoT) används administrativa data från sex olika myndigheter tillsammans med uppgifter som importerats från fem olika SCB-register. Mikrodata består av ca 500 variabler, och för att förstå dessa måste man känna till de skattemässiga regler som bestämmer variablernas innehåll. Nya variabler kan tillkomma eftersom skattesystemet hela tiden ändras och dessutom kan variabelnamnen i de administrativa källorna ändras.

Detta visar att det är stora skillnader mellan registerbaserade undersökningar och undersökningar med egen datainsamling när det gäller karaktären hos metadata. Man bör även skilja på dokumentation av *register* och dokumentation av *registerbaserade undersökningar*. När man använder befintliga register för att skapa nya register är det registerdokumentationer som behövs. Denna typ av dokumentation karaktäriseras av att:

- volymen av metadata kan vara mycket stor,
- varje administrativ källa måste finnas dokumenterad,
- förändringarna i de administrativa systemen också måste dokumenteras,
- variablerna kan vara komplicerade, noggrann dokumentation behövs,
- man gör ett stort antal registerbearbetningar för att skapa objektmängder, objekt och variabler, dessa bearbetningar måste också dokumenteras.

Detta innebär att metadatasystemet måste vara anpassat till SCB:s register och de registerbaserade undersökningarnas behov.

8.1.1 Dokumentation av administrativa källor

Från leverantören får man *postbeskrivningar* m.m. som anger struktur och innehåll på de uppgifter som levererats. Dessutom skall man som mottagare skaffa sig de *blanketter med anvisningar* som används vid den administrativa datainsamlingen. Dessa blanketter bör överföras i elektronisk form. De kan då lagras i metadatasystemet så att alla som arbetar med registret lätt kan nå dem.

De som ansvarar för kontakten med en leverantör gör också *intervjuer* för att få bakgrundsinformation. Även dessa intervjuer skall dokumenteras och lagras i metadatasystemet.

Det är betydelsefullt att alla förändringar noteras noga och att alla förändringar över tiden lagras, så att det blir lätt att få översiktlig information för att bedöma jämförbarheten över tiden. Ett metadatasystem bör därför även innehålla ett *kalendarium*, som är ett IT-system med formaliserade metadata där man kan söka information med *tid*, *register* och *variabel* som sökvillkor.

De administrativa data som tas emot, omstruktureras inom SCB och genomgår mottagningskontroll. Efter detta har en datamatrix med administrativa data från leverantören skapats inom SCB. De som tagit emot leveransen gör en egen dokumentation av denna, inklusive de bearbetningar man gjort.

8.1.2 Dokumentation av källor inom SCB:s registersystem

När man behöver importera variabler från andra statistiska register som ingår i SCB:s registersystem, behöver man enkel tillgång till redan befintliga metadata för att kunna söka och välja lämpliga variabler. I samband med importen av

mikrodata skall metadata för de importerade variablerna enkelt kunna överföras till dokumentationen av det nya registret. Därigenom undviks dubbelarbete då befintliga dokumentationer kan återanvändas. Denna dokumentation måste, för att kunna återanvändas bekvämt, vara strikt formaliserad enligt gemensamma regler och lagras i en allmänt åtkomlig databas.

8.1.3 Dokumentation av det nya registret

Det nya register som har skapats dokumenteras genom att använda den dokumentation av de administrativa källorna som nämns i avsnitt 8.1.1 samt genom att importera de delar av dokumentationen från olika SCB-register som behövs och slutligen genom att komplettera med dokumentation som beskriver de bearbetningar som gjorts och de härledda variabler som skapats.

I avsnitt 4.5.4 beskrivs olika slag av variabler som bör dokumenteras på olika sätt

Bild 8.2 Vad skall dokumenteras när ett nytt register har skapats?



8.2 Förändringar över tiden – behovet av metadata

Fyra slag av händelser påverkar registerstatistiken, och för att inte dra felaktiga slutsatser av tidsserier från registerbaserade undersökningar måste man känna till följande:

- Har det skett förändringar i de administrativa system som ingår som källor, så att administrativa begrepp har fått nya definitioner?
- Har det gjorts förändringar i SCB:s sätt att bilda registret. Har man t.ex. fått tillgång till nya källor, eller infört nya skattningsmetoder? Flera kvalitetsindikatorer bör följas över tiden för att inte data skall misstolkas.
- Har det skett förändringar i de klassifikationer som finns i registret? T.ex. kan SNI eller utbildningskod ha ändrats,
- Har det inträffat yttre händelser som påverkar statistiken indirekt? Om t.ex. skattesatsen i mervärdesbeskattningen ändras, påverkas inte endast moms-data, utan även data avseende privat konsumtion. För att tolka statistiken korrekt måste man känna till sådana förändringar. Många händelser av denna typ finns i det ekonomiska kalendarium som ingår i Konjunkturinstitutets publikation *Konjunkturläget*.

Ett händelsekalendarium bör därför finnas, för att händelser som påverkar olika register skall dokumenteras och finnas samlade på ett och samma ställe. Detta skulle underlätta för både de som arbetar med att skapa register och användarna av de färdiga registren. Risken för feltolkningar av data skulle då minska.

Ett sådant kalendarium kan enligt Ryen (1999) innehålla kortfattade uppgifter med grundläggande information om:

- vad som har inträffat,
- vid vilken tidpunkt det har inträffat,
- vilket eller vilka register och vilka variabler som har påverkats,
- vilken effekt det inträffade har fått,
- referens eller källhänvisning till ytterligare information.

Bild 8.3 Exempel på innehåll i ett händelsekalendarium

Händelse:	Kunskapslyftet införs
Tid:	1997
Register:	Kommunal vuxenutbildning, årgång 1997
Variabel:	Nya variabler tillkommer (anges, en post per variabel)
Effekt:	Ökning av elevantalet med ca 40 %
Händelse:	Nya rutiner ger snabbare information om att utvandrade personer upphört med företagsverksamhet
Tid:	1997
Register:	Företagsregistret
Variabel:	-
Effekt:	Ökad tillförlitlighet genom minskad övertäckning
Händelse:	Införande av SNI 92
Tid:	1995
Register:	Sysselsättningsregistret 1993 (motsvarande post för alla berörda register)
Variabel:	Näringsgren
Effekt:	SNI 92 ersätter SNI 69
Händelse:	Tjeckoslovakien upplöses i två nya stater: Tjeckien och Slovakien
Tid:	1993-01-01
Register:	Befolkningsregistret
Variabel:	Medborgarskap
Effekt:	-

Uppgifterna *Register* och *Variabel* i händelsekalendariet kan fungera som länkar till information i den övriga databasen med formaliserade metadata.

8.3 Integrationsregister – behovet av metadata

Om man t.ex. vill skapa ett longitudinellt integrationsregister med *tio* årgångar av tre olika statistiska register skulle man behöva utnyttja befintliga metadata på ett effektivt sätt. Om de *tre* källregistren varje år innehåller sammanlagt 200 variabler så innehåller metadatasystemet 30 definitioner av de olika registerpopulationerna och 2 000 variabeldefinitioner. Detta illustrerar att registerbearbetningar kan beröra stora mängder av metadata.

Om t.ex. register 1 och 2 har samma registerpopulation som inte har ändrats under de tio åren, och register 3 har en registerpopulation som ändrats en gång, så behövs endast 4 (1+1+2) populationsdefinitioner av dessa 30.

Om t.ex. totalt 50 variabler importeras från varje årgång av de tre registren, men endast 4 variabler har ändrat definition en gång vardera under tioårsperioden, så behövs endast 54 (50+4) av de 2 000 variabeldefinitionerna.

Exemplet ovan visar att man har behov av ett verktyg som hämtar dessa fyra populationsdefinitioner och 54 variabeldefinitioner ur den stora mängd av formaliserade metadata som hör till de 30 källregistren.

8.4 Klassifikations- och definitionsdatabas

För registerstatistiken, som baseras på administrativa data, är det särskilt betydelsefullt att kunna studera variabeldefinitioner och jämföra dessa över tiden. Inkomst- och taxeringsregistret illustrerar behovet av detta, där man måste hantera ca 500 variabler som baseras på administrativa regler. Många av dessa regler och variabler ändras varje år.

8.4.1 Klassifikationsdatabas

Näringsgren, varugrupp, utbildning, yrke, regional indelning m.m. är exempel på viktiga statistiska *standarder* och *klassifikationer*. De administrativa källorna innehåller uppgifter om dessa hierarkiskt ordnade klassifikationer, och denna information används för att skapa variabler inom registersystemet. Dessa klassifikationer ändras med jämna mellanrum. Då värdemängden (mängden av alla koder/kategorier) dessutom är stor, behövs en klassifikationsdatabas för att hantera alla koder och nycklar mellan olika versioner. Denna klassifikationsdatabas är ett viktigt hjälpmedel när variablerna i ett register dokumenteras.

8.4.2 Definitionsdatabas och härledda variabler

På samma sätt som man behöver ett IT-verktyg för att hantera definitionerna av de statistiska klassifikationerna, behöver man även ett verktyg med formaliserade metadata för att hantera den stora mängden av komplicerade variabeldefinitioner som ändras över tiden. Vi visar detta med ett exempel där tre årgångar av ett tänkt inkomstregister har dokumenterats, liksom även ett longitudinellt inkomstregister (LongIoT), där de tre årgångarna har integrerats.

Vi använder begreppet *definitionsdatabas*, som underlättar arbetet med att dokumentera variabler över tiden

Bild 8.4 Dokumentation av registervariabler med hjälp av en definitionsdatabas

Register	Variabelnamn	Definitions-kod	Definitions-databas			
IoT2001	Sjukpenning	SP1	Kod	Definition	Definitionen användes	
	Havandeskapspenning	HP1			Första tid	Sista tid
	Sjukersättning	SJUK1				
IoT2002	Sjukpenning	SP2	SP1	SP1 = "....."	2001	2001
	Havandeskapspenning	HP1	SP2	SP2 = "....."	2002	-
	Sjukersättning	SJUK2	HP1	HP1 = "....."	2001	2002
IoT2003	Sjukpenning	SP2	HP2	HP2 = "....."	2003	-
	Havandeskapspenning	HP2	SJUK1	SJUK1 = SP1 + HP1	2001	2001
	Sjukersättning	SJUK3	SJUK2	SJUK2 = SP2 + HP1	2002	2002
LongIoT	Sjukersättning2001	SJUK1	SJUK3	SJUK3 = SP2 + HP2	2003	-
	Sjukersättning2002	SJUK2				
	Sjukersättning2003	SJUK3				

Det tänkta inkomstregistret innehåller tre inkomstvariabler:

- sjukpenning, där nya regler infördes 2002,
- havandeskapspenning, där nya regler infördes 2003,
- sjukersättning, som är en härledd variabel, summan av sjuk- och havandeskapspenning

Det longitudinella registret innehåller endast sjukersättning för varje år. Av bilden ovan framgår att variabler med samma namn, t.ex. *Sjukpenning*, kan ha olika definitioner. Vidare kan variabler med olika namn, t.ex. *Sjukersättning* i

IoT2001 och *Sjukersättning2001* i LongIoT, ha samma definition. Genom att definitionskoderna är unika, dvs. en viss kod motsvaras inom hela register-systemet av en och endast en variabeldefinition, bör inga missförstånd uppstå. Det är också enkelt att med en definitionsdatabas följa definitionsförändringar.

8.5 Sammanfattning – behovet av metadata för register

De som skapar statistiska register inom SCB:s registersystem har behov av olika slag av metadata och praktiska IT-verktyg för att registrera och använda dessa metadata i sitt arbete. Bilden nedan visar nio slag av metadata och de verktyg som skulle kunna användas.

Bild 8.5 Registerdokumentationens olika slag av metadata och verktyg



Idag (2004) finns Metadok-systemet för att lägga in formaliserade metadata samt SCBDOK för att skapa textdokument med metadata. I framtiden bör det finnas ett system som integrerar alla formaliserade metadata, dels vad som finns i nuvarande Metadok men även det som vi ovan kallar Kalendariet, Klassifikations- och Definitionsdatabasen. Dessutom behövs ett system för att hantera dokument med övriga metadata, t.ex. dokumentationsdelen i det nya SiP-systemet.

SiP, Systemutveckling i Praktiken, SCB:s rekommenderade modell för systemutveckling

System med formaliserade metadata kan användas för följande behov:

- dokumentation av de administrativa källornas datamatriser,
- dokumentation av det färdiga statistiska registrets datamatriser,
- Händelsekalendarium,
- Klassifikations- och Definitionsdatabas.

Den övriga dokumentationen kan bestå av olika slag av Office-dokument:

- textinformation om administrativa system,
- administrativa blanketter med anvisningar,
- minnesanteckningar från möten med registerleverantörer m.m.,
- SQL-skript med kommentarer och
- anslagstavla för alla som använder registret och hittar inkonsistenser och fel. Alla som skall stötta basregistren enligt avsnitt 5.4.6 lägger in sina bidrag på respektive basregisters anslagstavla.

Detta kapitel har skrivits av
Lars-Göran Lundell

Kapitel 9. IT-system för registerstatistik

Hittills har vi diskuterat statistiska register på en konceptuell nivå dvs. hur man tänker sig att register och datamatriser ser ut och fungerar logiskt. Detta kapitel inleds med ett fortsatt konceptuellt resonemang, men kommer även in på den fysiska realiseringen med skapande av faktiska databaser och applikationer.

Vid utveckling av ett IT-system som skall användas då statistiska register skapas bör man, liksom vid varje annat systemutvecklingsuppdrag, utgå från en standardiserad utvecklingsmodell. SCB:s rekommenderade modell kallas *SiP* (Systemutveckling i Praktiken). Den innehåller även vissa verktyg för att underlätta utvecklingsarbetet. Modellen är intuitiv och verktygen till stor del självinstruerande.

SiP är utvecklat för att i första hand vara ett stöd vid utveckling och förvaltning av undersökningar med egen datainsamling. För ett system baserat på administrativa register kommer vissa underrubriker att vara irrelevanta, medan andra har stor vikt. I nästa avsnitt utgår vi från SiP:s avsnittsindelning och försöker lyfta fram punkter som bedöms särskilt viktiga för ett IT-system för statistiska register och registerbaserade undersökningar.

Den registerdokumentation, som diskuteras i kapitel 8, behövs i arbetet med att utveckla de IT-system som skall stödja arbetet med att skapa statistiska register. Det är därför viktigt att dokumentationsarbetet påbörjas innan systemeringsarbetet startas. De olika slag av formaliserade metadata som nämns i kapitel 8 finns då tillgängliga och kan användas med hjälp av de IT-verktyg som finns för att hantera metadata.

9.1 Projekt

Syftet med ett *SiP-projekt* är att ge redskap för planering och bemanning samt att hantera risker. Tyngdpunkterna är dels projektarbetsformen i sig, dels ett antal formaliserade dokument som under projektets gång samlas i en projektdokumentation som sedan blir en del av systemdokumentationen.

En allsidig bemanning av projektgruppen är alltid viktig. Vid planering av en ny registerbaserad undersökning, eller större förändring av en befintlig, krävs alltid ett externt samarbete med den administrativa myndighet som levererar data. Samarbetet med leverantören kan gälla såväl innehåll som teknik. Teknikfrågor kan omfatta allt från variabelinnehåll till kommunikationslösningar.

De externa kontakterna gör också att det är mycket viktigt att alla underlag, minnesanteckningar och annan dokumentation samlas enligt SiP:s modell för att finnas enkelt tillgängliga vid eventuella framtida diskussioner.

Det är också viktigt att bemanna projektet inte bara för registret, utan med tanke på registersystemets behov, så att samordnings- och integrationsaspekterna kommer med i bilden. Detta gäller alla områden, från innehåll till datamodellering och teknisk infrastruktur.

9.2 Analys

Syftet med analysprocessen är att skapa underlag för utformning och realisering av systemet genom att analysera och formalisera alla krav som kan ställas på det. Metoden är att modellera, dvs. identifiera, diskutera och dokumentera systemets objekt och processer.

Objektmodelleringen identifierar systemets utdata och indata och beskriver sambanden mellan dem.

Processmodelleringen visar systemets processer och beskriver ordningsföljd och beroenden mellan dem. I detta steg beskrivs hur uppgifterna skall samlas in. Här dokumenteras också hur användarna av systemet skall kunna arbeta, eller vill arbeta. Detta görs genom att beskriva *användningsfall* enligt en standardiserad mall. Identifiering och beskrivning av användningsfall är en central del av varje SiP-projekt.

Ett IT-system för registerstatistik skiljer sig från ett IT-system, som skall stödja en undersökning med egen datainsamling, främst när det gäller modellering av datainsamlingsprocessen. I stället för att modellera ett ram- och urvalsförfarande gäller här att besluta hur väl de register som utgör de huvudsakliga datakällorna täcker behovet av utdata (se avsnitt 4.1.1). Om ytterligare data behövs måste dessa kompletterande datakällor identifieras och dokumenteras. Det gäller här att modellera de selektions- och matchningsförfaranden som skall göras för att skapa det aktuella registret.

I denna fas skall också utredas hur samverkan med andra system skall ske. Den primära samverkan som måste klargöras är inom registersystemet: vilka beroenden som finns till och från andra register, hur eventuella konflikter mellan olika informationskällor skall hanteras, etc. Dessutom måste också registrets roll som källa för andra register inom registersystemet klargöras.

En komplikation som är typisk för statistiska register är hantering av *tidsdimensionen* (Se avsnitt 4.4.3). En regelbundet återkommande undersökning med egen insamling av data kommer (förenklat) att skapa antingen en fil eller databastabell per undersökningstillfälle, eller en fil/databastabell med en generationsangivelse per undersökningstillfälle. Ett statistiskt register kan, om det bakomliggande registret tillåter det, i princip förändras kontinuerligt i form av tillägg, borttagande eller uppdatering av uppgifter. En av analysprocessens viktigaste uppgifter är att tydligt belysa vilka önskemål och behov av tidsangivelser som finns för fasta uttag såväl som för kända och möjliga uppdrag. Därifrån skall bedömas vilka typer av information som behövs i detta register: det aktuella beståndet, beståndet vid vissa eller godtyckliga tidpunkter, årgångsinformation, händelser, historik, longitudinell information.

Analysprocessen skall också visa hur detaljerad information som skall finnas tillgänglig för olika delar av systemet. Här måste också dokumenteras med vilken frekvens det är möjligt, önskvärt och meningsfullt att uppdatera det statistiska registret från det bakomliggande registret.

Det bör också understrykas att analysfasen skall ge en helhetssyn. Detta innebär att man för det egna registret t.ex. inte får glömma behoven från en befintlig eller möjlig uppdragsverksamhet, men det innebär också att man hela tiden måste lyfta blicken och se till krav och behov från det totala registersystemet, inte bara ”det egna” registret.

9.3 Design

Syftet med designprocessen är att utifrån de krav som ställs på systemet ta beslut om:

- teknikval för den totala systemlösningen
- eventuell uppdelning i delsystem
- utformning av databas (eller databaser)
- arkitektur för varje applikation eller bearbetning.

9.3.1 Teknikval

En viktig del av utformningen av ett statistiskt register är dess samverkan med olika källregister. Då det är ett primärt statistiskt register som skall utformas blir den tekniska samverkan med det administrativa registret särskilt viktig. Den är beroende av flera faktorer, t.ex.:

- lagringstekniken för det administrativa registret: kan data överföras direkt eller måste de konverteras/exporteras?
- uppdateringsfrekvens: om data skall överföras sällan, kanske en gång per år, kan relativt enkla manuella lösningar tillämpas, medan en högre frekvens kräver mera automatisering och tillförlitlighet hos både avsändare och mottagare.
- datamängder: överföring av större mängder kräver naturligtvis mera av den tekniska lösningen. En liten mängd kan överföras via e-post, diskett eller liknande, medan stora mängder kan kräva välplanerade speciallösningar.
- säkerhet/intrång: alla mikrodata som avser enskilda objekt är strikt konfidentiella och måste hanteras i enlighet med SCB:s säkerhetsföreskrifter.
- säkerhet/sårbarhet: detta avser dels riskerna för att förlora data vid överföring genom t.ex. mindre lämpliga rutiner eller tekniker, dels risker förknippade med förlust av kompetenser på båda sidor.
- kommunikationsteknik: det finns ofta ett antal möjliga tekniker att tillgå vid varje tillfälle. Vid valet av teknik skall stor hänsyn tas till de tidigare nämnda faktorerna.

Förutom att det aktuella registret skall samverka med ett eller flera källregister måste det också tekniskt samverka väl med andra register inom registersystemet. Teknikval och lösningar som tidigare gjorts i närliggande register kommer alltså att starkt påverka lösningar för det nu aktuella registret.

Analysen och värderingen av dessa krav, önskemål och möjligheter kan medföra väsentliga skillnader i behov av arbetsinsats i följande utvecklingssteg.

9.3.2 Databasutformning

Utformning av ett systems databaser är kritisk för hur väl det kommer att fungera. Det innebär att den alltid bör genomföras av systemutvecklare med stor erfarenhet av datamodellering.

Databasutformning för ett IT-system för registerstatistik skiljer sig i grunden inte från utformningen för andra IT-system för statistikproduktion. Systemet skall optimeras för:

- användbarhet, användarvänlighet,
- tillförlitlighet och prestanda för laddning av data, nya värden och ändrade värden,
- bearbetning, dvs. rättning, granskning, härledning, m.m. och
- uttag, dvs. tabellframställning, fasta uttag, uppdrag.

Detta ställer ofta motstridiga krav på databasens utformning och resulterar inte sällan i att systemet delas upp i flera databaser. Något förenklat brukar man tala om indatabas, bearbetningsdatabas och utdatabas där var och en är optimerad för olika användning.

- En *indatabas* bör kunna ta emot all slags uppdateringar, inklusive helt felaktiga indata. Indatabasen har därför normalt endast ett minimum av konsistensregler. Data kontrolleras och rättas innan det förs vidare.
- *Bearbetningsdatabasen* är oftast normaliserad, dvs. inga data dubbellagras. Detta är en modell som är optimerad för att utföra granskning och rättning på mikronivå med minimal risk för att införa inkonsistenta data.
- En *utdatabas* skall vara optimerad för de vanligast förekommande slagen av uttag och bearbetning. Sedan den laddats med data skall den inte ändras. Innehållsmässigt är således en utdatabas liktydig med den färdiga datamatrixen. Här kan man utan risk tillåta dubbellagring om det skulle underlätta för användarna.

Synonyma termer:

Indatabas =
Inmatningsdatabas =
Laddningsdatabas =
Input database

Utdatabas =
Uttagsdatabas =
Output database

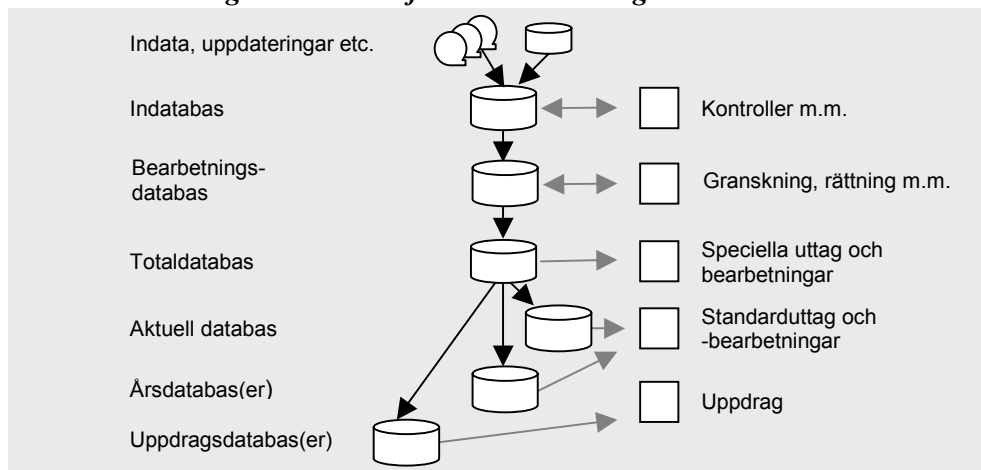
I denna fas skall lagringsformat för alla variabler fastställas. Detta innebär i de flesta fall inget problem, men för kopplingsvariabler och i synnerhet för sådana som kan användas för länkning mellan olika register är det mycket viktigt att följa fastlagda standarder, t.ex. lagring av personnummer med 12 tecken.

Tidsdimensionen, som diskuteras i avsnitt 9.2 ovan, ställer stora krav på databasutformningen. Om systemet har höga krav på tidsangivelser för varje ändring och lång historik, riskerar en databas som klarar alla slags uttag att både bli svåransvänd och lågpresterande. I sådana fall kan man behöva dela upp utdatabasen ytterligare, t.ex. i följande delar:

- en totaldatabas eller historikdatabas,
- en eller flera databaser för fasta tidpunkter som t.ex. årsskiftet,
- en databas för aktuellt läge.

På motsvarande sätt kan en omfattande uppdragsverksamhet baserad på registret medföra behov av att skapa en eller flera särskilda uppdragsdatabaser. Att genomföra ad-hoc-uppdrag och fasta uttag på samma databas kan i olyckliga fall medföra störningar, vilket naturligtvis bör undvikas.

Bild 9.1 Olika slag av databaser för ett statistiskt register



Kravet på samverkan inom registersystemet innebär givetvis en begränsning av val av databashanterare. En programvara som i vissa avseenden har överlägsna prestanda, eller på annat sätt passar väl för ett visst register kanske ändå inte kan väljas, om den samverkar dåligt – eller inte alls – med de mjukvaror som används av närliggande register inom registersystemet.

Modellering för registersystemet – inte registret

Vi har genomgående understrukt vikten av att se till helheten, registersystemet, i stället för till det enskilda registret. Man kan sträcka sig längre än så och hävda att registersystemet är en central del av det totala statistikdatalagret, *the Statistical Data Warehouse*.

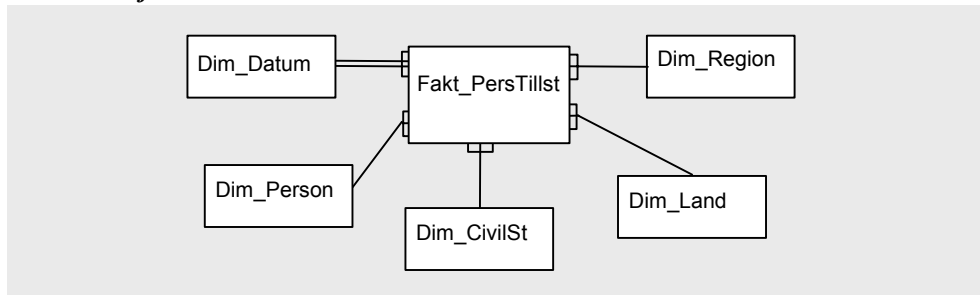
Termen *Data Warehouse* används främst inom företagsvärlden för att beteckna ett sammanhållet, konsistent och kvalitetssäkrat datalager, oftast avseende en lång tidsperiod för uttag och analys av i första hand ekonomiska data.

Datalagret innehåller typiskt stora datamängder fördelade på ett stort antal informationsslag. Vid utformningen av en eller flera databaser för ett Data Warehouse strävar man efter att minimera behovet av matchningar mellan mera än två databastabeller samtidigt, eftersom sådana matchningar dels är logiskt och syntaktiskt svåra för användarna, dels ger sämre svarstider genom att vara tunga för databashanteraren.

Det finns ett antal etablerade tekniker för modellering av ett Data Warehouse. Den vanligaste är förmodligen den som kallas *stjärnmodellen*. Den är avsedd för effektiva ”analytiska” bearbetningar av stora mängder data, t.ex. tabellframställning.

I en stjärnmodell finns de värden och uppgifter man är intresserad av i en eller flera faktatabeller kopplade till dimensionstabeller, vilka, något förenklat, kan ses som metadatatabeller.

Bild 9.2 Stjärnmodell



Exemplet i bild 9.2 låter oss på ett effektivt sätt ta fram grundläggande befolkningsstatistik för godtycklig tidpunkt. Spridningsvariablerna finns i dimensionstabellerna:

- län, kommuner i (Dim_Region),
- födelseland i (Dim_Land),
- civilstånd i (Dim_CivilSt),
- personnummer, födelsedatum, kön i (Dim_Person)
- och redovisningstidpunkter i (Dim_Datum).

Faktatabellen Fakt_PersTillst innehåller inga ”egna” data, utan utgör bara en kopplingspunkt genom att visa giltiga kombinationer av dimensioner¹. För registersystemet innebär detta resonemang att uttagsdatabaser för ett enskilt register inte bör modelleras separat, utan som en del av det totala statistikdata-lagret.

Härledda variabler och vyer

Så länge systemutformningen befinner sig i den konceptuella, eller logiska, skedet behandlas härledda variabler på samma sätt som andra. En viktig del av den fysiska databasutformningen är att besluta om vilka härledda variabler som faktiskt skall lagras permanent och vilka som skall kombineras och beräknas vid behov. Det senare alternativet kan göras helt ad-hoc eller förberedas för större användarvänlighet genom att skapa *vyer*². Fördelen med en vy är att den

¹ Exemplet visar två kopplingar mellan faktatabellen och dimensionstabellen Dim_Datum. Detta gör det möjligt att låta faktatabellen innehålla en rad per förändring i stället för en rad per möjligt datum. Kopplingarna uttrycker att dimensionstabellens *Datum* skall ligga mellan *FrånDatum* och *TillDatum* i faktatabellen.

² En databas-vy är en ”virtuell databastabell”, dvs. ett fördefinierat uttryck som för användaren ser ut och används som en databastabell, men som i praktiken utför operationer på en eller flera bakomliggande databastabeller. *Exempel:* Databastabellen A innehåller uppgifter om en rad inkomstslag, tjänstgöringsgrader osv. Från uppgifterna i A kan vi härleda variabeln heltidslön, som vi lägger i vyn B. Användarna kan sedan med ett enkelt anrop hämta heltidslönen från B, men varje gång kommer beräkningen att utföras på A.

är lätt att förstå och enkel att använda. En nackdel kan vara att den kan kräva höga maskinprestanda för att fungera bra.

En avvägning mellan att lagra en härledd variabel fysiskt i en databastabell eller virtuellt i en vy måste ta hänsyn till vilken påverkan det kan få på prestanda, vilka risker en eventuell dubbellagring kan medföra och hur frekvent variabeln kommer att användas.

Prestanda i en vy är beroende av mängden data (antalet objekt), antal påverkade databastabeller och komplexiteten i det bakomliggande uttrycket. Om en variabel med en enkel beräkning kan härledas från ett fåtal variabler i en eller två databastabeller i samma databas ger det som regel inte nämnvärt sämre prestanda att använda en vy än att först lagra den nya variabeln i en databastabell.

Om variabeln däremot härleds från en annan databas, som kanske dessutom finns på en annan server, kan prestanda försämrast kraftigt vid stora datamängder. Är det då en variabel som kommer att användas frekvent kan det vara nödvändigt att lagra den härledda variabeln lokalt.

Vyer bör i de flesta sammanhang betraktas som vilken databastabell som helst och också dokumenteras på samma sätt. Ett undantag kan vara sådana vyer som skapats endast för att förbereda återkommande matchningar mellan databastabeller. Matchningar diskuteras i avsnitt 9.6.2.

9.3.3 Applikationsarkitektur

I denna fas skall också utredas vilket behov det finns av utveckling av skräddarsydda applikationer. Ett register med mycket frekvent uppdatering av indata behöver en automatiserad process för överföring och laddning, medan ett register med oregelbunden uppdatering eller t.ex. årlig uppdatering kan nöja sig med manuella rutiner. På samma sätt kan standardiserade, ofta återkommande standarduttag eller uppdrag förenklas genom att utveckla en applikation som kan hanteras av ”vem som helst” i stället för att behöva upprepa snarlika sekvenser av handgrepp för varje uttag eller uppdrag.

9.4 Fysisk realisering

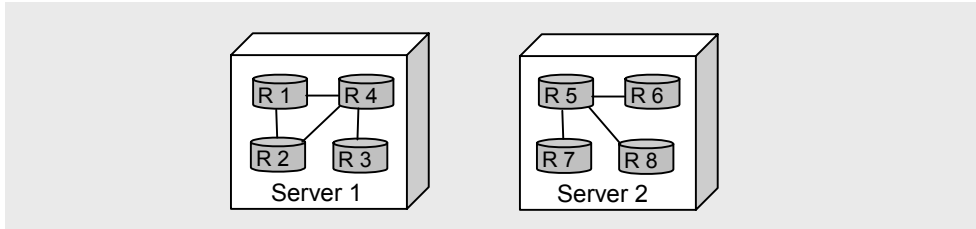
Den fysiska realiseringen innebär i huvudsak att skriva nödvändig programkod och etablera nödvändiga databaser för att systemet skall kunna tas i bruk. Vid såväl programmering som skapande av databaser skall man använda rekommenderade verktyg och följa etablerade standarder.

9.4.1 Fysisk optimering för registersystemet

Förutom att de samverkande databaser som tillsammans utgör den fysiska implementeringen av registersystemet skall modelleras för samverkan måste de även fysiskt samverka väl. Alla databaser i registersystemet kan inte läggas i en och samma fysiska databas, inte ens i samma server. Den fysiska samverkan kommer alltid att påverkas av om en matchning sker inom en server eller mellan servrar. Verktygen för att låta användaren uppfatta en grupp servrar som en sammanhållen logisk enhet förbättras hela tiden, men hela tiden kvarstår faktum att en server arbetar många gånger snabbare internt än vad kommunika-

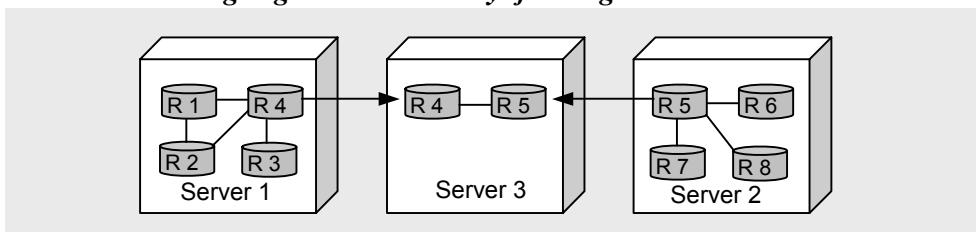
tionslänkarna mellan servrar klarar av. Därför måste en bedömning göras av vilka register i registersystemet som har de största behoven av samutnyttjande och därför bör lokaliseras till samma server. I exemplet i bild 9.3 samutnyttjas informationen i registren R1–R4 resp. R5–R8 regelbundet och databaserna läggs därför på två servrar.

Bild 9.3 Samutnyttjade register läggs på samma server



En fördelning av register enligt exemplet betyder inte att det är omöjligt att samutnyttja registren R4 och R5, men det gör det något mera komplicerat och det tar längre tid. En metod för att behålla samutnyttjandet inom exemplrets två servrar, och samtidigt förbättra samutnyttjandet mellan registren R4 och R5 är att införa en tredje server som innehåller kopior av dessa register. En kontrollerad dubbellagring på detta sätt kan hanteras på flera sätt. Det säkraste är förmodligen att använda databashanterarens inbyggda s.k. replikeringsteknik. I exemplet i bild 9.4 kommer varje förändring i register R4 i Server 1 att vid förutbestämda tidpunkter, t.ex. varje timme, varje natt eller varje vecka, också genomföras i register R4 i Server 3. Samma gäller för register R5 i Server 2 respektive Server 3.

Bild 9.4 Dubbellagring av vissa samutnyttjade register



9.4.2 Indexering

Vid all bearbetning och uttag av data från datalager baserade på relationsdatabaser är en korrekt och lämplig *indexering* a och o för prestanda. Utan lämpliga index kommer databashanteraren t.ex. vid varje matchning mellan två databastabeller att alltid utvärdera varje möjlig kombination innan den gör sitt val – även då användaren bara vill matcha en rad i varje tabell. Med index på primärnyckel resp. främmande nyckel kan den däremot utnyttja denna information för att direkt hämta de två matchande posterna.

Indexering av databastabellerna på primärnyckel och främmande nycklar sköts normalt automatiskt av det verktyg man använder då en databas skapas. I en bearbetningsdatabas är oftast dessa index de enda som behövs för att skapa en bra miljö för att ändra, lägga till eller ta bort data. Ytterligare index på andra variabler kan rentav försämra prestanda för dessa arbetsuppgifter.

Ett index är kopplat till en eller flera kolumner i en databastabell. Det är en intern datastruktur som med olika tekniker låter databashanteraren hitta en enskild uppgift utan att behöva söka sekventiellt

I en uttagsdatabas däremot, där ändringar är sällsynta och väl planerade, och där sökningar och matchningar skall ske även på andra variabler än nycklarna, kan man få stora prestandavinster genom väl genomtänkt indexering. Detta sker till priset av sämre prestanda vid uppdateringar.

9.5 Övriga faser i systemutvecklingen

Då systemet är konstruerat och all programmering är genomförd återstår fortfarande att genomföra och dokumentera ett antal steg eller faser:

- test,
- driftsättning,
- underhåll och vidareutveckling,
- avveckling.

Särskilt underhållet och vidareutvecklingen kan kräva särskild uppmärksamhet, då dels strukturen på de administrativa källregister som används kan ändras och dels kan nya källor tillkomma.

9.6 Användning av register

Att använda registerbaserade system för framställning av statistiska tabeller eller annan bearbetning ställer inga speciella krav jämfört med användning av andra system, utom möjligen i två avseenden. De registerbaserade systemen har ofta stora datamängder och de används ofta för sambearbetning, matchning mot andra datakällor.

9.6.1 Stora datamängder

Att bearbeta eller göra uttag från stora datamängder kan innebära vissa problem. Tiden för en bearbetning ("körning") har oftast inte ett linjärt samband med datamängdens storlek. Tiden är beroende av hårdvara, mjukvara, konfiguration och programmeringsmetod. Begränsningar i minneskapacitet kan t.ex. i svåra fall ge ett exponentiellt samband mellan datamängd och bearbetningstid. Ofta är dock användarens arbetssätt den viktigaste faktorn för att få bra prestanda. Vid måttligt stora datamängder kan användaren ofta prova sig fram med hjälp av testkörningar på ett sätt som inte är möjligt då datamängderna är stora. Ett projekt som genomfördes under 2001-2002 tog fram en rad rekommendationer för arbetssätt vid stora datamängder. Projektrapporten finns på SCB:s intranät³.

9.6.2 Matchning

Att matcha information från flera register eller mellan ett register och en urvalsundersökning är vanligt förekommande arbetsuppgifter. På konceptuell nivå kan det förefalla som en ganska enkel sak, förutsatt att nödvändiga kopplingsvariabler finns tillgängliga.

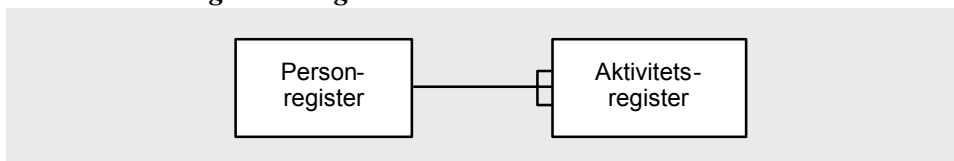
³ IT → Rapporter → IT-rapporter → Hantering av STORA datamängder

När vi kommit så långt att de konceptuella registren blivit databastabeller kommer matchningen (som ofta kallas *join* med SQL-terminologi) att bli ett handgripligt problem. Ett av exemplen i avsnitt 4.5.4 lyder:

Ett personregister med personnummer som primärnyckel kan matchas mot ett aktivitetsregister där förvärvsaktiviteter identifieras av tre variabler: Personnummer, arbetsställesnummer och organisationsnummer. Personnummer i Aktivitetsregistret är främmande nyckel vid matchningen.

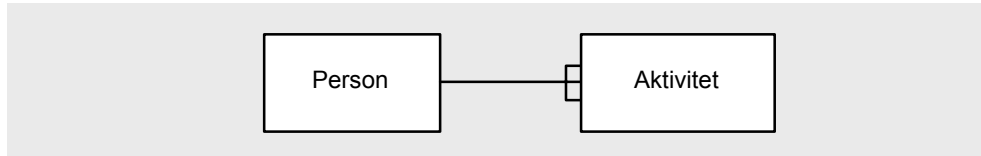
En person (ett personnummer) kan förekomma på endast ett ställe i personregistret, men på ett eller flera ställen i Aktivitetsregistret. Vi får alltså ett-*till-många-förhållande* mellan objekten. Situationen kan beskrivas grafiskt på följande sätt:

Bild 9.5 Matchning av två register



Vid den fysiska implementeringen kan vi först förutsätta att de uppgifter vi vill hämta finns i en enda databastabell i personregistret (kallad *Person*) och en enda tabell i Aktivitetsregistret (*Aktivitet*). Fortfarande tycks situationen vara en ganska lätthanterlig matchning mellan två databastabeller (tvåtabells-join):

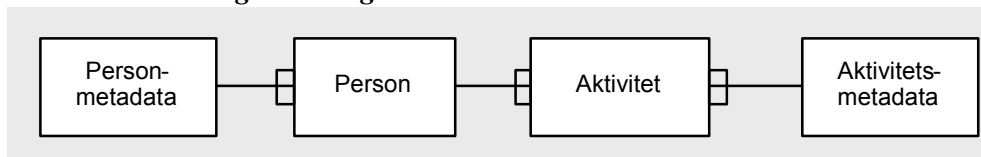
Bild 9.6 Relationen Person - Aktivitet



Detta kan utföras direkt i SQL-språket. Flera användarverktyg, t.ex. SAS och SPSS, ger bra tekniskt stöd för enklare matchningar utan att användaren behöver skriva SQL-kod, förutsatt att nycklarna har angetts på ett korrekt sätt då databastabellerna skapades. För de flesta användare och användningar är dessa verktyg att föredra framför att skriva direkt SQL, eftersom det stöd man får minskar riskerna för glömda kopplingar, andra misstag och felskrivningar.

Oftast innehåller dessa databastabeller emellertid endast koder eller värden (indelningsvariabler, aggregeringsvariabler), men i många fall vill vi få klartexter som finns lagrade i metadatatabeller. Genast kompliceras bilden i vårt enkla exempel till att kräva fyra databastabeller (fyrtabells-join):

Bild 9.7 Matchning av två register med metadata



I ett verkligt fall skulle bilden troligen kompliceras ytterligare genom att t.ex. lagringen av de aktivitetsvariabler vi är intresserade av finns utspridd på fler än en databastabell, att vissa variabler skall aggregeras i samband med matchningen, osv.

En ytterligare, vanligt förekommande, komplikation är hanteringen av ”icke-träff” (*mismatch*, se avsnitt 4.5.3). I vårt exempel kan det finnas objekt i Aktivitetsregistret som av någon anledning saknar eller har felaktigt personnummer, och i personregistret finns med säkerhet objekt som saknar motsvarighet i Aktivitetsregistret. Standardläget i SQL, såväl som i alla användarverktyg (SAS, SPSS, etc.) är att endast fullständiga träffar väljs ut. Om vi vill välja samtliga rader från databastabellen *Person* och komplettera med uppgifter från *Aktivitet* när sådana finns, men annars lämna tomt (*null-värde*), måste detta särskilt anges med hjälp av SQL-funktionen *outer join*. Syntaktiskt – och logiskt – är användning av *outer join* begränsad, t.ex. i kombination med aggregeringar och med fullständiga träffar (*inner join*).

Bild 9.8 Inner och outer join

Inner join – välj endast rader där nr i Person matchar nr i Aktivitet::

```
select Ålder, Lön
  from Person
  join Aktivitet
    on Person.nr=Aktivitet.nr
```

Outer join – välj alla rader i Person, komplettera med matchande information från Aktivitet där sådan finns:

```
select Ålder, Lön
  from Person
 left outer join Aktivitet
    on Person.nr=Aktivitet.nr
```

Redan vårt förenklade exempel visar dock på att den tekniska, eller syntaktiska, svårighetsnivån snabbt stiger. Det gör att det ibland kan vara svårt att avgöra om det resultat man åstadkommer med en avancerad matchning verkligen är korrekt. En slarvigt och felaktigt skriven SQL-kod kan i olyckliga fall ge ett resultat som vid hastigt påseende ser rätt ut – svaret har rätt antal rader och värden i rätt storleksordning – men ändå är felaktigt. Det är därför nästan alltid att föredra att genomföra en uppgift med hjälp av flera enkla, kontrollerbara steg med temporär lagring av mellanresultaten mellan varje steg i stället för att lösa allt med ett avancerat jättesteg. Det viktiga är inte att skriva eleganta kommandon, utan det viktiga är att vara säker på att resultatet blir korrekt.

En annan metod för att förenkla komplicerade bearbetningar är att förbereda vanliga matchningar genom att skapa vyer (se sid. 181). Med en vy kan ett komplext SQL-uttryck döljas så att det för användaren ser ut som en enkel databastabell som innehåller alla intressanta variabler. I vårt exempel kan vi en gång för alla skapa vyn *PersonAktivitet* som innehåller vår fyratabells-join. Varje gång ett uttag behöver göras använder vi sedan vyn i stället för de fyra tabellerna. De datamatiser för ”kombinationsobjekt” som diskuteras i avsnitt 6.2 (se t.ex. bild 6.7) kan med fördel implementeras i form av vyer som kan användas för uttag av vägda värden.

Förutom de syntaktiska problemen att verkligen åstadkomma vad vi avser att göra tillkommer ofta innehållsmässiga problem att välja rätt rader från databas-

tabellerna vid matchning. I vårt exempel kan vi anta att Aktivitetsregistret innehåller en tidsdimension. En syntaktiskt korrekt matchning skulle ändå kunna ge ett nonsens-resultat om man glömmer att ange vilken tidpunkt, intervall, etc., som avses.

9.6.3 Rätt verktyg för uppgiften

I detta kapitel har vi i huvudsak utgått från att data finns lagrat i en relationsdatabas. På en teknisk nivå måste i dessa fall SQL användas för att nå data. Detta får dock inte tolkas som att alla användare skall utföra all bearbetning med ren SQL direkt i databaserna. Det är i stället ofta lämpligt att använda ett verktyg som ger ett användarvänligare gränssnitt (klicka, välj från lista, dra och släpp, osv.) för att hämta och manipulera data och som ger ett bra stöd för matchningar.

Dessutom är det inte alltid säkert att det är effektivast att låta databashanteraren utföra allt arbete. Databashanteraren är optimerad för att arbeta *mängdorienterat*. Detta är dess normala, mycket effektiva arbetssätt. Det är möjligt att tvinga databashanteraren att arbeta *sekventiellt* för att klara sådana arbetsuppgifter som kräver detta, t.ex. urvalsdragning, men med det arbetssättet (stega sig fram med en *cursor*) blir den ineffektiv. I sådana fall och i många fall när statistiska bearbetningar, t.ex. skattningar med vikter, skall göras är det lämpligt att använda t.ex. SAS.

Dess värre (eller kanske dess bättre ...) finns det många situationer när det inte är klart vilket som är det bästa eller lämpligaste verktyget. Användningsområdena för de vanligaste verktygen överlappar ibland kraftigt. Det kan vara möjligt att utföra uppgiften på kanske tre, fyra olika sätt med lika många verktyg. Med vilket verktyg uppgiften löses snabbast och säkrast i det enskilda fallet beror ofta mera på användarens erfarenhet än på verktygets objektiva lämplighet, men nackdelen är att lösningen riskerar att bli personberoende. För att få viss hjälp med att mera objektivt välja verktyg kan man utnyttja informationen på SCB:s intranät⁴.

⁴ IT → IT-metod → Rätt verktyg → Systemutveckling / Statistikprocessen

Kapitel 10. Integritetsskydd

När registersystemet förbättras för att underlätta sambearbetningar, som syftar till att höja statistikens kvalitet, måste även integritetsskyddet förstärkas. Detta skydd kan förstärkas genom att:

1. Förekomsten av *klartext* minimeras.
2. Förekomsten av identifierande variabler som *personnummer* och *fastighetsbeteckning* minimeras.
3. Risken skall minimeras, för att uppgifter avseende enskilda personer eller företag kan härledas ur de *statistiska tabeller* som finns i SCB:s publikationer eller offentliga databaser.
4. Innan forskare efter prövning får möjligheter att analysera datamatriser med *mikrodata* skall dessa datamatriser bearbetas så att risken för röjande av enskilda minimeras.

Integritetsskyddet kan enligt punkterna 1-2 ovan ökas på två sätt, dels genom att antalet databaser med klartext respektive personnummer minimeras och dels genom att antalet personer inom SCB som har tillgång till databaser med klartext respektive personnummer minimeras. Genom dessa åtgärder förstärks också skyddet för personer med spärmarkering (s.k. hemlig adress) ytterligare.

Uppgifter som lämnas till SCB för framställning av statistik omfattas av statistiksekretessen i sekretesslagen. Vidare finns särskilda regler om skydd mot kränkning av enskilds personliga integritet genom behandling av personuppgifter i personuppgiftslagen som anger när behandling av personuppgifter är tillåten. För känsliga personuppgifter gäller ännu strängare regler.

Lagstiftningen innebär i princip att ingen skall ha behörighet till mer data än vad man behöver i tjänsten.

10.1 Förekomst av klartext

Namn, adress och andra uppgifter i klartext, som avser personer eller företag, bör endast finnas i basregistren. Det är endast personal som arbetar med intervjuer och enkäter som skall ha tillgång till namn och adressuppgifter. *Alla* register utom basregistren bör vara helt anonymiserade genom koder.

De administrativa uppgifter som levereras till SCB kan innehålla namn, adress och andra textuppgifter då den administrativa myndigheten behöver dessa för sin administration. Även löneregister från arbetsgivare kan innehålla uppgifter i klartext. I samband med de inledande registerbearbetningarna bör dessa uppgifter ersättas med koder.

Ett fiktivt exempel får illustrera detta:

Bild 10.1 Ett administrativt källregister – obearbetade indata

PersonNr	Namn	Gatuadress	Postnr	Arbetsplats	Tjänstebenenämning	TNS-kod	Mlön	Tjmf
5678901234	Pson Per	Lillgatan 7	111 11	SCB Stockholm	Avdelningsdirektör	4321	18340	0,60
6789012345	Ason Eva	Storgatan 2	777 77	SCB Örebro	Avdelningschef	1234	45780	1,00

Sekretesslagen
9 kap. 4 §

Personuppgiftslagen
1998:204

Den första bearbetningen efter mottagningskontrollen bör vara att personnummer ges korrekt SCB-format, all klartext ersätts med koder och att statistiskt onödiga variabler stryks. Dessutom kan statistiskt viktiga variabler importeras och härledda variabler bildas.

I detta exempel räknas faktisk månadslön, *Mlön*, om till motsvarande heltidslön, *Hlön*, med tjänstgöringens omfattning, *Tjomf*.

Bild 10.2 Motsvarande statistiska register – bearbetade data i indatabasen

PersonNr	BoKommun	AE-id CfarNr	ArbKommun	YrkeSSYK	UtbSUN	Mlön	Tjomf	Hlön
195678901234	0180	12345678	0180	2222	1234567	18342	0,60	30570
196789012345	1880	23456789	1880	3333	7654321	45780	1,00	45780

I detta exempel skulle det vara lämpligt att ha tillgång den ursprungliga tjänstbenämningen för t.ex. framtida ändrade yrkesklassificeringar. Det enklaste sättet är att de administrativa källorna tas bort från datasystemet och arkiveras i låst utrymme när det statistiska registret är färdiggranskat och färdigbearbetat. Framtida kontroller blir då möjliga, utan att de administrativa uppgifterna finns tillgängliga via datasystemet.

10.2 Förekomst av identiteter

Förutom att behörigheterna till att arbeta med mikrodata bör vara väl kontrollerade och förekomsten av klartext bör vara minimerade bör även objektidentiteter användas restriktivt.

Personnummer

Kan vi minska användningen av personnummer? Personnummer finns i de administrativa källor som levereras till SCB. Dessa personnummer förekommer inom administrativa individ-, aktivitets-, företags- och fastighetsregister. Man bör begränsa antalet personer som arbetar med att ta emot administrativa data.

De administrativa uppgifterna överförs efter mottagningskontroll och vissa bearbetningar till indata-, bearbetnings- och totaldatabaser där personnummer kan ha en viktig roll som kopplingsvariabel vid olika bearbetningar. Antalet personer som arbetar med dessa databaser bör också begränsas.

Efter att alla registerbearbetningar har genomförts är det lämpligt att skapa en utdatabas, som kan användas av en större krets av personal som arbetar med uppdrag, rapporter etc. Utdatabasen bör därför vara avidentifierad.

Den nya statistiklagstiftningen gör det möjligt att lämna ut avidentifierade uppgifter försedda med löpnummer för senare uppdatering. Sådana löpnummer kan även användas för känsliga register inom SCB som idag innehåller personnummer. En förutsättning för detta är att nycklarna mellan person- och löpnummer är väl skyddade. En mycket liten krets av personal skulle ha tillgång till dessa nycklar, och den större krets av personal som arbetar med registerbearbetningar och matchningar skulle endast ha tillgång till databaser med löpnummer. Inom Danmarks statistik har man börjat ersätta personnummer med löpnummer för att öka integritetsskyddet.

Skyddad identitet

Cirka 8 000 – 9 000 personer har idag skyddad identitet. Personer med skyddad identitet har genom statistiksekretessen ett starkt skydd inom SCB. Inom Befolkningsregistret har man särskilda rutiner för behandling av uppgifter som avser personer med spärrmarkering. Dessa innebär bl.a. att uppgift om adress, församling och fastighet tas bort ur Befolkningsregistret. Dessa uppgifter sparas i en databastabell med mycket begränsad behörighet. Personer med spärrmarkering ingår vidare inte i SCB:s urval från befolkningsregistret. Genom att minimera användningen av personnummer och klartext stärks skyddet för dessa personer ytterligare.

Avidentifierade register

Register som används av många bör avidentifieras. När register lämnas ut till forskare skall man tänka igenom vilka variabler som bör strykas. I personregister bör namn, adress, personnummer och fastighetsbeteckning tas bort. I företagsregister kan även näringsgren, storleksklass och region vara känsliga variabler som gör det enkelt att identifiera stora företag.

10.3 Röjanderisk i tabeller

Om man med hjälp av statistiska tabeller kan härleda egenskaper av känslig natur hos enskilda personer eller företag, innebär publicerandet av dessa tabeller risk för röjande. Ett röjande innebär dels att man kan identifiera objektet och dels att man kan bedöma variabelvärden för dessa enskilda objekt.

För urvalsundersökningar är risken för röjande som regel låg, men i tabeller som baseras på total- eller registerbaserade undersökningar måste man alltid kontrollera tabellerna så att inte egenskaper hos enskilda objekt kan röjas.

10.3.1 50/90-regeln för summor och medelvärden

I tabellen nedan finns fiktiva uppgifter som avser företagen inom en viss bransch i en viss kommun.

Tabell 10A. Företag efter storleksklass inom bransch Y i kommun X år 2003

	Antal företag	Antal anställda	Omsättning 1000-tals kr	Lönesumma 1000-tals kr	Genomsnittlig omsättning per ftg	Årsmedellön per anställd
Antal anst	(1)	(2)	(3)	(4)	(5) = (3)/(1)	(6) = (4)/(2)
0-9	4	25	14 125	5 900	3 531	236
10-99	2	75	46 750	17 350	23 375	231
100-	1	205	130 550	50 400	130 550	246
Totalt	7	305	191 425	73 650	27 346	241

Två rader i tabell 10A röjer uppgifter för enskilda företag, dels den rad som avser ett enda företag som kan identifieras med kommun, bransch och storleksklass och dels raden med två företag med mellan 10-99 anställda.

Personer som känner till det ena av dessa två företag kan genom subtraktion härleda uppgifter om det andra företaget. Uppgifterna i kolumn (2) – (6) på

dess två rader måste därför undertryckas. Undertryckta celler skall då endast innehålla symbolen “.”.

Enligt de regler som anges i SCB (2001c) får inget objekt bidra med 50 % eller mer till cellvärdet. Inte heller får två objekt tillsammans bidra med 90 % eller mer av cellvärdet. Detta innebär att alla uppgifter i tabell 10A, utom raden med företag med 0-9 anställda, är för känsliga för att kunna publiceras.

10.3.2 Basen skall vara minst tre observationer

Enligt tröskelvärdesregeln i SCB (2001c) måste en cell innehålla minst 3 objekt för att kunna publiceras. Detta innebär att den vita cellen i tabell 10C nedan inte bör publiceras. Däremot kan hela tabell 10B publiceras då det inte går att härleda någon information om enskilda personer.

Det räcker inte att undertrycka den vita cellen i tabell 10C, då man kan räkna ut innehållet i cellen med hjälp av tabell 10B och de icke-vita cellerna i 10C. Man måste då undertrycka fler celler eller, vilket bör vara enklast i detta fall, slå samman åldersklasserna 16-24 och 25-34 år.

Tabell 10B. Bransch Y i kommun X
Anställda efter ålder och kön

Ålder	Kvinnor	Män	Totalt
16-24	2	7	9
25-34	3	9	12
35-44	30	54	84
45-54	44	67	111
55-64	5	84	89
Totalt	84	221	305

Tabell 10C. Bransch Y i kommun X
Medellön per månad efter ålder och kön

Ålder	Kvinnor	Män	Totalt
16-24	15 033	16 483	16 161
25-34	16 550	17 933	17 588
35-44	18 200	19 650	19 132
45-54	20 108	21 367	20 868
55-64	20 008	20 933	20 881
Totalt	19 173	20 488	20 126

10.3.3 Røjanderisken får inte vara för hög

Tabellen nedan anger antal personer dömda för brott i en viss tätort ett visst år. Vissa celler i denna tabell med känslig information innehåller små antal.

Tabell 10D. Tätort Z år 2003

Antal personer dömda för brott efter ålder och kön

Ålder	Kvinnor	Män	Båda könen
16-24	0	7	7
25-34	1	7	8
35-44	2	17	19
45-54	3	2	5
55-64	2	5	7
Totalt	8	38	46

Med endast informationen i denna tabell, kan man inte härleda någon känslig uppgift för någon enskild. Det är först i kombination med antalet personer i tätorten uppdelad efter ålder och kön som känsliga uppgifter kan röjas.

Med tröskelvärdesregeln skulle man kunna tro att de vita cellerna i tabellen inte bör publiceras. Men detta är en felaktig tolkning av regeln, då det är *basen* för tabellcellerna som skall vara minst tre observationer. Basen är totalantalet personer i befolkningen i respektive cell för denna tätort.

Tabell 10D ovan, är egentligen en del av en större, fullständig frekvenstabell där befolkningen delas upp efter tre variabler samtidigt: kön, ålder och grad av brottsbelastning som mäts med kategorierna någon gång dömd/aldrig dömd.

I den fullständiga tabellen delas befolkningen samtidigt in efter variablerna kön, ålder och brottsbelastning. Det är i denna fullständiga tabell som rójanderiskerna syns.

Tabell 10E. Tåtort Z år 2003, befolkningen efter kön, ålder och brottsbelastning

Kön	Ålder	Dömda	Ej dömda	Total befolkning	Rójanderisk
Kvinnor	16-24	0	2	2	0,00
	25-34	1	1	2	0,50
	35-44	2	28	30	0,07
	45-54	3	41	44	0,07
	55-64	2	5	7	0,29
Män	16-24	7	0	7	1,00
	25-34	7	1	8	0,88
	35-44	17	65	82	0,21
	45-54	2	65	67	0,03
	55-64	5	3	8	0,63
Båda könen	16-64	46	211	257	0,18

Av den fullständiga tabellen framgår vad som är olämpligt att publicera. Tröskelvärdesregeln med minst tre som bas, innebär att de vita cellerna inte bör publiceras. Träffar man en kvinna 16-24 år, så kan man dra slutsatsen att hon inte är dömd. Detta är visserligen inte känslig information, men för kvinnor i åldern 25-34 skyddar tröskelvärdesregeln rójandet av känsliga uppgifter.

Däremot så ger varken tröskelvärdesregeln eller 50/90-regeln något skydd för rójande av känsliga uppgifter för de två yngsta åldersklasserna bland männen i Z. Det behövs då en tredje regel:

– *Om rójanderisken är hög, bör uppgifterna inte publiceras.*

I tabellen ovan har rójanderisken i tabellens högra kolumn beräknats genom att ange andelen dömda för varje kategori. Med hjälp av variablerna ålder och kön säger denna risk hur lätt det är att gissa rätt när det gäller denna känsliga information om enskilda individer.

De vita cellerna som inte bör publiceras enligt tröskelvärdesregeln kan slås samman genom att åldersklasserna görs större: 16-34 år. Men för männen hjälper det inte att slå samman dessa åldersklasser, rójanderisken är fortfarande hög. Det bästa i detta fall, är att inte dela in materialet efter för många variabler samtidigt. För tåtorten Z skulle man endast redovisa dömda efter kön.

Det finns för närvarande ingen regel som säger vad som är en rimlig rójanderisk, utan detta får bedömas från fall till fall.

Tabeller baserade på statistiska register kan göras mycket detaljerade. Man kan samtidigt dela upp efter många variabler och variablerna kan ha många klasser.

Men röjanderiskerna kan då bli stora. Därför måste man alltid granska de tabeller man skapar för att minimera röjanderiskerna.

10.4 Röjanderisk i mikrodata

För närvarande lämnas i vissa fall avidentifierade mikrodata ut till forskare och myndigheter för forskningsändamål. De som tar emot dessa mikrodata har då samma juridiska skyldigheter som SCB, för att skydda enskilda personer och företag så att känsliga uppgifter inte röjs.

I framtiden kommer forskare även att kunna analysera avidentifierade mikrodata från sin egen arbetsplats genom att analyser görs tillgängliga via Internet. Mikrodata finns i detta fall kvar inom SCB.

Den finns viss risk för att forskare, som inte har SCB:arnas vaksamhet mot röjanderisker, kan komma att publicera tabeller i forskningsrapporter som röjer uppgifter avseende enskilda personer eller företag.

Det finns tre sätt att minska riskerna:

- Var restriktiv med vilka variabler som skall ingå i forskarens datamaterial.
- Indelningsvariabler bör grupperas med så få kategorier som möjligt, t.ex. i stället för församlingsindelning innehåller datamaterialet endast uppgift om A-region.
- Forskaren får tillgång till ett urval från den aktuella registerpopulationen. Detta bör alltid vara en lämplig inledning på samarbetet. Om forskaren först får lära känna datamaterialet, kan man sedan göra en större datamatrix med endast de variabler som behövs.

I SCB (2003a) beskrivs hur mikrodata skulle kunna bli tillgängliga för forskare utan att mikrodata lämnar SCB:s datasystem. Forskarna skulle från sin ordinarie arbetsplats via Internet få tillgång till avidentifierade data som finns på SCB. Med SAS eller SPSS kan varje forskare gör sina bearbetningar på SCB:s datorer. Endast skärmbilder skickas till forskaren förutom resultatfilerna med tabeller och andra analysresultat.

Kapitel 11. Samordning och sam användbarhet

SCB:s datalager är fysiskt och organisatoriskt starkt decentraliserat med mer än 1 000 databaser lagrade på ett stort antal servrar. En förutsättning för en effektiv statistikproduktion är att de register som ingår i datalagret är innehållsmässigt och tekniskt samordnade. Onödiga särlösningar, som annars kan ställa till oreder när data tolkas, matchas eller jämförs, kan då undvikas.

En viktig del av det registerstatistiska metodarbetet består av arbetet med att strukturera och förbättra helheten. Här ingår det långsiktiga arbetet med att bevaka och påverka tillgången på administrativa data för statistiska ändamål. Dessutom bör systemet utvecklas genom att register skapas för nya statistikbehov. Samordning syftar till att minimera dubbelarbete, förenkla samutnyttjande av data och att öka konsistensen mellan olika register. Vi skiljer på fyra slag av samordning:

- organisatorisk,
- teknisk,
- innehållsmässig och
- metodologisk.

Sam användbarhet ingår i de kvalitetskomponenter som diskuteras i SCB (2001a). Genom innehållsmässig och metodologisk samordning kan sam användbarheten förbättras. Även SCB:s urvalsundersökningar bör vara samordnade med registersystemet.

11.1 Organisatorisk samordning

Vid undersökningar med egen datainsamling kan man avgränsa arbetet så att man i huvudsak tänker endast på en undersökning eller produkt i taget. Det samlade ansvaret för hela produkten kan då finnas på ett ställe i organisationen. Även nationalräkenskapssystemet ställer krav på samordning, men registersystemet kräver dessutom samarbete kring gemensamma källor och variabler.

För registerbaserade undersökningar är det viktigt med en organisation som stödjer ett omfattande samarbete mellan de som ansvarar för registersystemets olika delar. Cheferna har en betydelsefull roll för all samordning, bl.a. för att registersystemets olika delar skall kunna samverka. Men det räcker inte med en lämplig formell organisation, utan verkskulturen måste vara sådan att samarbete över organisatoriska gränser uppfattas som naturligt.

De som arbetar med registerstatistik måste ha ett gemensamt synsätt. Man får inte vara för snävt avgränsad till att endast tänka på det egna registret, utan man måste även förstå sin roll inom systemet. Man har dels skyldigheter att stötta andra registers verksamhet och dels möjligheter att använda systemet för att utveckla nya tillämpningar. Både i avsnitt 4.2 och avsnitt 5.4.6 diskuteras detta.

För att utveckla en sådan förståelse krävs utbildning, gemensamma seminarier, gemensamma projekt och uppdrag men även att den interna rörligheten uppmuntras. Det är en fördel om det finns ett stort antal personer med erfarenhet från flera register inom systemet.

11.1.1 En gemensam enhet med övergripande ansvar för systemet

I avsnitt 5.1.1 påpekas att det registerstatistiska arbetet bedrivs på tre nivåer:

1. *Arbetet på systemnivå.* Arbetet innebär att strukturera och samordna ett stort antal register till ett fungerande system av statistiska register. Systemet som helhet skall utvecklas för att möjliggöra ny och förbättrad statistik.
2. *Arbetet med att skapa ett statistiskt register inom det befintliga systemet.*
3. *Arbetet med en undersökning baserad på ett statistiskt register.*

Det måste finnas en gemensam organisatorisk enhet som ansvarar för arbetet på nivå 1. Arbetet på nivåerna 2 och 3 görs på respektive registerenhet, men det övergripande ansvaret för samordningen inom registersystemet kan endast ligga hos en särskild enhet. Det lokala ansvaret respektive det centrala ansvaret måste vara tydligt reglerat mellan registerenheterna och den gemensamma registerenheten.

Övergripande ansvar för kontakterna med registerlämnande myndigheter, forskarkontakter och uppdrag som berör flera delar av registersystemet, samarbete med andra statistikansvariga myndigheter kring registerfrågor och andra länders statistikbyråer är arbetsuppgifter som kan vara lämpliga för en sådan organisatorisk enhet.

Eftersom metadata avseende olika källor används vid arbetet med att skapa statistiska register, måste metadata vara allmänt tillgängliga. Den gemensamma registerenheten bör därför övervaka tillgången på metadata och att metadata-systemet utvecklas så att arbetet med att skapa register underlättas.

Integritetsskyddet, som diskuteras i kapitel 10, är särskilt betydelsefullt för registersystemet. Den gemensamma registerenheten bör därför övervaka och understödja arbetet kring integritets- och röjandeskydd. Eftersom uppgifter i tabeller som kan innebära röjanderisker samutnyttjas inom SCB, är det viktigt att sekretesskyddet följer med när tabelluppgifter lämnas från en registerenhet till en annan enhet inom SCB.

11.1.2 Samgranskning

I avsnitt 5.7.6 behandlas granskningsarbetet vid registerbaserade undersökningar. Granskningsarbetet delas där upp i två delar: *mottagningskontroll* av data från varje administrativ källa och *konsistenskontroll* där mikrodata från olika källor samgranskas.

Flera av de kvalitetsindikatorer (indikator 3.1b, 5h och 5l) som diskuteras i kapitel 7 bygger på att administrativa data jämförs med uppgifter från undersökningar med egen datainsamling. Detta innebär att samgranskning av mikrodata från flera källor som avser samma objekt är ett viktigt arbetsmoment som bidrar till statistikens kvalitet. För att undvika dubbelarbete bör denna samgranskning ske i nära samarbete mellan alla de som arbetar med de inblandade undersökningarna.

Samgranskningen kan antingen organiseras som gemensamma projekt eller genom att särskilda organisatoriska enheter skapas med ansvar för denna samgranskning.

SCB har under 2004 inrättat en särskild enhet med övergripande ansvar för registersamordning

11.1.3 Integrationsregister och uppdrag

Idéer till nya integrationsregister får inte bli konflikthanledningar utan bör ses som nya samarbetsmöjligheter. På samma sätt bör uppdrag som berör flera registerenheter ses som medel att utveckla gemensam kompetens. Det är viktigt att kompetens och erfarenheter från olika statistiska register tas till vara vid arbetet med att utveckla nya tillämpningar och vid arbetet med ämnesövergripande uppdrag.

Nya integrationsregister innebär att vissa organisatoriska frågor blir aktuella. För att undvika konflikter bör regelmässigt samarbete ske mellan berörda enheter. Ingen bör ha monopol på att använda vissa data. Kostnadsdebitering internt och externt skall följa gemensamma och kända regler. I samband med att integrationsregistret Louise byggdes upp, diskuterades de principer som bör gälla kring samarbetet mellan berörda registerenheter när ett nytt integrationsregister skapas. Dessa principer redovisas i SCB (1999).

11.2 Teknisk samordning

Detta avsnitt har skrivits av
Lars-Göran Lundell

Den tekniska samordningen är av naturliga skäl nästan uteslutande IT-relaterad. En del samordningsfrågor har därför redan tagits upp i kapitel 8 och 9.

Den utan tvekan viktigaste enskilda faktorn för teknisk samordning av ett registersystem är ett väl fungerande, enhetligt och standardiserat *metadata-system*. Metadata kan tjäna flera syften med olika krav. För detta ändamål är det viktigt att metadata lagras centralt, är lättillgängligt och innehåller strikt formaterade data.

De som utvecklar ett nytt registerbaserat IT-system måste ta hänsyn till och anpassa sig till metadata systemet och sträva efter att bygga in metadata som en integrerad del. Om möjligt skall befintliga metadata utnyttjas. Nyskapade metadata skall dokumenteras enligt standard. På samma sätt skall system eller delsystem som utnyttjar information lagrad i registersystemet utvecklas för att söka, bearbeta och hämta data med hjälp av metadata.

Standardiserad *namnsättning* och enhetligt *lagringsformat* för centrala variabler ger viktig samordning. I första hand skall variabler som används som länkar vara standardiserade. Till exempel skall variabeln *personnummer* alltid heta "PersonNr" och lagras som "character(12)", dvs. 12 tecken. Ett GD-beslut om standardiserad namnsättning och lagring av personnummer och ett tiotal andra centrala kopplingsvariabler, eller länkar, togs 1995-12-12.

Kopplingen mellan variabelnamn och *variabeldefinition* skall vara entydig inom registersystemet. Ett variabelnamn som förekommer i flera register skall alltid ha samma definition. Motsatsen skall heller inte förekomma: variabler har samma definition, men olika namn i olika register. Detta innebär även att samma koder skall ha samma betydelse i olika register, dvs. att t.ex. kön kodas på samma sätt. När det gäller register som avser olika tidsperioder och longitudinella register behövs en definitionsdatabas enligt principerna i avsnitt 8.4.2.

Sam användbarhet ställer också krav på samordning av den *fysiska åtkomsten* till data i samverkande register. Detta betyder inte att alla register nödvändigt-

vis måste lagras i samma typ och samma version av databashanterare. Det är dock ett minimikrav att register som skall sam användas – i praktiken hela registersystemet – måste kunna samverka med vanliga bearbetnings- och analysverktyg för att kunna sambearbetas med det aktuella verktyget. Det är en fördel om åtminstone en del av sam användningen kan ske direkt i de inblandade databaserna. Därmed kan man nästan aldrig göra ett oberoende val av mjukvara för lagring av ett enskilt register.

En standardisering av *verktyg* för bearbetning och uttag och helst verktyg som kan utnyttja metadata på ett effektivt sätt är önskvärd. Bearbetningar skall ge samma resultat oavsett vem som gör bearbetningarna – standardiserade verktyg är ett medel som bidrar till detta.

11.3 Innehållsmässig samordning

För att statistiska uppgifter från olika undersökningar skall vara jämförbara, fordras att populationer och variabler i olika undersökningar *definieras* på ett konsistent sätt. Dessutom krävs att registren *bearbetas* med syftet att skapa konsistens mellan registerpopulationer och variabler i olika statistiska register. Den innehållsmässiga samordningen bygger på följande:

- standardiserade *populationer*,
- standardiserade *variabler* och
- standardiserade *grupperingar* av viktiga klassifikationer.

11.3.1 Samordning och standardisering av populationer

I avsnitt 5.4 och 5.5 beskrivs arbetet med att skapa registerpopulationer och avgränsa objekt. Alla populationer skall skapas med hjälp av basregistren. En viktig uppgift för basregistren är att skapa *standardiserade populationer*, som sedan används för att skapa registerpopulationer i systemets övriga register.

Dessa standardiserade populationer skall ha hög kvalitet med avseende på täckning och viktiga indelningsvariabler. De skapas därför *efter* den tidsperiod eller tidpunkt de skall avse. Man väntar tills i stort sett all administrativ information har hunnit inrapporterats och bearbetats. Därefter får registret som innehåller den standardiserade populationen inte ändras. Om sedan andra register enbart använder sådana standardiserade registerpopulationer kommer registerstatistiken att bli konsistent med avseende på objektmängder.

Registersystemet borde bygga på sju olika standardiserade registerpopulationer, varav endast tre finns hösten 2003:

1. Befolkningen vid varje årsskifte, *denna finns idag*
2. Alla personer som tillhört befolkningen ett visst år, årgångsversion
3. Aktiviteterna ett visst år, årgångsversion, *denna finns idag*
4. Företagen vid varje årsskifte
5. Alla företag som varit aktiva någon gång ett visst år, årgångsversion
6. Fastigheterna vid varje årsskifte, *denna finns idag*
7. Fastigheterna ett visst år, årgångsversion

Den standardiserade population som används mest idag är befolkningen den 31/12 år t , enligt den regionala indelningen den 1/1 år $t+1$. Fastigheterna vid årsskiftet avser förhållandena den 1/1 varje år.

Flera register utgår från Befolkningsregistrets standardiserade population som avser den 31/12 varje år: *Utbildningsregistret*, *Inkomst- och taxeringsregistret*, *Sysselsättningsregistret* och *Yrkesregistret*. Alla dessa register har samma registerpopulation och dessutom är alla variabler som hämtats från Befolkningsregistret identiska: *Ålder*, *kön*, *civilstånd*, *födelseland* samt *län*, *kommun*, *församling* och *fastighet* där respektive person är folkbokförd. Alla tabeller som delas in efter några av dessa variabler kommer att ha samma fördelning i alla dessa register – statistiken från dessa register uppfyller alltså höga krav på konsistens och sam användbarhet.

Inom den ekonomiska statistiken är däremot konsistensproblemen större. Företagsundersökningarna är till stor del baserade på egen datainsamling där man utgår från en rampopulation som bildas under november före den avsedda undersökningsperioden, den så kallade SAMU-populationen.

I avsnitt 6.3.1 – 6.3.2 beskrivs allmänt hur ramproblemen uppstår och i exemplet nedan beskrivs hur statistiken avseende energiföretag påverkas av inkonsistenser med avseende på population.

Intäkter och kostnader för energiföretag undersöks både inom Företagsstatistiken (FS) och Energistatistiken (EN). Samgranskning har visat att olika organisationsnummer ingår i undersökningarnas rampopulationer. Vissa orgnr ingår i båda, andra endast i den ena.

Intäkter 1998 enligt:			Intäkter 1999 enligt:			Kostnader 1998 enligt:			Kostnader 1999 enligt:		
	FS	EN		FS	EN		FS	EN		FS	EN
Endast FS	32%	0%	Endast FS	16%	0%	Endast FS	32%	0%	Endast FS	18%	0%
Båda	68%	76%	Båda	84%	96%	Båda	68%	84%	Båda	82%	96%
Endast EN	0%	24%	Endast EN	0%	4%	Endast EN	0%	16%	Endast EN	0%	4%
Summa	100%	100%	Summa	100%	100%	Summa	100%	100%	Summa	100%	100%

Eftersom ramproblemen är stora vill man i många undersökningar använda en mer aktuell version av Företagsregistret. Då olika undersökningar kan bygga på Företagsregistret vid olika tidpunkter kommer undersökningarnas populationer att vara olika. Även näringsgren för samma företag kan vara olika, vilket gör att t.ex. populationen av energiföretag kan vara olika i olika undersökningar. I avsnitt 11.5 beskrivs hur man med nya metoder för konsistenta skattningar, som har utvecklats vid den holländska centralbyrån, kan minska dessa problem.

Problemen i exemplet ovan beror på att man använder Företagsregistret från olika tidpunkter. Dessutom finns inkonsistenser med avseende på företagspopulation som beror på att man använder register som inte har avstämts mot Företagsregistret. Energistatistikens, inkvarteringsstatistikens, jordbruksstatistikens, och utbildningsstatistikens företagpopulationer överensstämmer inte med Företagsregistrets företag inom motsvarande branscher. Man använder olika källor och samverkar inte när det gäller utbyte av information som avser avgränsningen av dessa delar av Företagsregistret. Detta kan delvis bero på att dessa undersökningar eller produkter finns på andra avdelningar än Företagsregistret, organisatoriska gränser kan ha försvårat samarbetet.

En ytterligare orsak till inkonsistenta populationer är att olika register inte har samma objektavgränsningar, t.ex. så bygger Aktivitetsregistret och Företagsregistret på arbetsställen som har avgränsats på något olika sätt.

11.3.2 Samordning och standardisering av variabler

När det gäller den innehållsmässiga delen är det även viktigt med samordning avseende variabler. Ett viktigt samordningsinstrument är Klassifikationsdatabasen (KDB) som innehåller statistiska klassifikationer som används inom SCB och det svenska statistiska systemet. Här ingår klassifikationer och nycklar mellan olika versioner av en klassifikation, som har fastställts som standard i Sverige eller internationellt. Ett exempel på en sådan klassifikation är näringsgren, som anges enligt standard för svensk näringsgrensindelning (SNI).

Inom Registerprojektet (2001) påbörjades en första inventering av variabler med utgångspunkt från Befolkningsregistrets variabler. Jämförelser gjordes med motsvarande variabler i en rad andra register och ett antal olikheter upptäcktes. Detta ledde till ett beslut om samordning och standardisering av vissa viktiga variabler (GD-beslut 2000-04-03).

De variabler som är aktuella för standardisering är sådana som återkommande används inom flera register. Den registerenhet som importerar en variabel från en registerlämnande myndighet eller själva skapar en härledd variabel skall ha ansvaret för denna på SCB. Det är alltså de som arbetar med ett primärt statistiskt register som ansvarar för registrets lokala primära variabler. Ansvaret omfattar bearbetningar, namnsättning och dokumentation av dessa variabler. På motsvarande sätt ansvarar de som skapar viktiga härledda variabler för att dessa namnsätts och dokumenteras.

Inom den ekonomiska statistiken beror vissa inkonsistenser på skillnader i variabeldefinitioner och mätmetoder. Exemplet nedan ger en beskrivning av hur statistiken som avser energiföretag påverkas av inkonsistenser med avseende på variabler. De uppmätta värdena för enskilda energiföretags intäkter och kostnader kan skilja sig mellan de två undersökningarna (FS och EN).

Näringsgren	Intäkter och kostnader	
Företagsstatistiken (FS) redovisar enligt huvudsaklig SNI, dvs. stympar den flervärda SNI-variabeln. Energikostnader och energiintäkter kan då redovisas på "fel" bransch	De företag som ingår med samma organisationsnummer i både Företags- och Energistatistiken:	
	Intäkter 1999	Kostnader 1999
	enl. FS: 108 miljarder	enl. FS: 64 miljarder
Energistatistiken (EN) redovisar enligt verklig SNI, dvs. företag med verksamhet i fler branscher redovisar endast energikostnader och -intäkter till energibranschen	enl. EN: 103 miljarder	enl. EN: 51 miljarder
	Företaget med störst skillnad:	Företaget med störst skillnad:
	EN – FS = 2 miljarder	FS – EN = 6 miljarder

11.3.3 Samordning och standardisering av grupperingar

Många variabler måste klassindelas eller grupperas. Dessa klassindelningar och grupperingar måste göras enligt samordnade regler, annars blir inte tabeller från olika källor jämförbara. T.ex. åldersklasser respektive inkomstklasser bör ha

bildats på samma sätt i olika register och urvalsundersökningar. De stora klassifikationerna näringsgren, sektortillhörighet, utbildning och yrke måste grupperas innan man kan bilda statistiska tabeller. Att gruppera dessa variabler kräver goda kunskaper om variablerna. De som ansvarar för dessa variabler inom SCB bör därför skapa ett antal standardiserade grupperingar som kan passa för olika behov. Både grova indelningar och finare indelningar behövs för olika syften. De grupper som skapas skall också ges standardiserade namn.

Tabellerna nedan visar hur skillnader i grupperingar kan försämrast statistikens sammanvändbarhet. I tabell 11C används *nästan* samma gruppering av variabeln näringsgren, men dessa små skillnader gör att man inte kan dividera lönesummorna med antal sysselsatta för grupperna C-E och H+O+P.

Tabell 11C. Två sätt att gruppera variabeln näringsgren

Sysselsatta (AKU), 1000-tal efter näringsgren, 2001		Lönesumma enligt Kontrolluppgiftsregistret (KU), miljarder kr efter näringsgren, 2001	
A+B jordbruk, skogsbruk, jakt, fiske	96,2	A+B jordbruk, skogsbruk, jakt, fiske	7,6
C-E utvinning, tillverkning; energi- o vattenförsörjning	776,7	C+D utvinning av mineral, tillverkningsindustri	195,1
28-35 verkstadsindustri	376,4	E+90 energi- o vattenförsörjning, avfallshantering	9,8
F byggindustri	231,5	F byggindustri	51,7
G+I handel; transport, magasinering; kommunikation	803,8	G+I handel; transport, magasinering; kommunikation	162,8
G parti- och detaljh; rep av fordon, hushållsartiklar	520,4		
H+O+P personliga och kulturella tjänster; renhållning	332,9	H+Oexkl90+P personliga och kulturella tjänster	47,1
J+Kexkl73 kreditinstitut, fastighetsförvaltn, företagstjänster	600,8	J+Kexkl73 kreditinstitut, fastighetsförvaltn, företagstjänster	157,4
L+Q civila myndigheter, försvar; internat. organisationer	231,7	L+Q civila myndigheter, försvar; internat. organisationer	57,1
M+73 forskning o utveckling; utbildning	379,6	M+73 forskning o utveckling; utbildning	75,9
N enh för hälso- och sjukvård, socialtjänst; veterinärer	777,9	N enh för hälso- och sjukvård, socialtjänst; veterinärer	146,5

I tabell 11D har variabeln sektortillhörighet grupperats på olika sätt. Anmärkningen efter AKU-tabellen visar att detta är en svår variabel. Det är därför stor risk att olika personer inom SCB skulle gruppera sektortillhörighet i grupperna statlig, kommunal och privat sektor på olika sätt. På samma sätt som standardiserade populationer och variabler skall ha lättillgänglig dokumentation skall även standardiserade grupperingar vara lättåtkomliga för alla inom SCB.

Tabell 11D. Två sätt att gruppera variabeln sektortillhörighet

Förvärvsarbetande 16+ år (RAMS) efter sektor, 2001		Sysselsatta (AKU), 1000-tal efter yrkesställning och sektor 2001	
Statlig förvaltning	202 244	Anställda i statlig sektor	222,2
Statliga affärsverk	6 109	Anställda i kommunal sektor	1 079,2
Primärkommunal förvaltning	800 939	Anställda i privat sektor	2 507,9
Landsting	235 471	Företagare	410,2
Övriga offentliga institutioner	15 410	Medhjälpare	13,7
Aktiebolag ej offentligt ägda	2 089 748	Fr.o.m. år 2001 tillämpar AKU standarden för institutionell sektorindelning 1999. Största förändringen är att försäkringskassorna redovisas som statliga (tidigare kommunala) och affärsverken som privata (tidigare statliga). Fr.o.m. år 2001 redovisas kyrkan som privat (tidigare kommunal).	
Övriga företag, ej offentligt ägda	327 993		
Statligt ägda företag och organisationer	168 100		
Kommunalt ägda företag och organisationer	73 194		
Övriga organisationer	171 871		

Behandlingen av tidseriebrott måste också samordnas. Försäkringskassorna, affärsverken och kyrkan måste omklassificeras samtidigt i alla register och undersökningar.

11.4 Samordning av statistiska metoder

Ett gemensamt synsätt och en gemensam terminologi är viktiga för att registerstatistiken skall kunna utvecklas. Detta diskuteras i kapitel 4 och i kapitel 5-7 beskrivs den registerstatistiska metodologin och där påpekas att även metodiken måste vara gemensam.

"*Alla skall använda basregistren*" (avsnitt 5.4.4) och "*alla skall stötta basregistren*" (avsnitt 5.4.6). Metodiken som bygger på dessa principer bidrar till att systemets populationer blir konsistenta. När man sambearbetar flera register är man beroende av att varje register har bearbetats, granskats och dokumenterats så att sambearbetningarna underlättas och resultatet av bearbetningarna kan kvalitetssäkras.

Beräkningar skall göras med syftet att skattningar skall bli korrekta, och beräkningarna kan behöva samordnas så att skattningarna från olika register blir konsistenta. Om t.ex. det finns partiellt bortfall i en variabel som finns i flera register bör bortfallet behandlas på samma sätt i de inblandade registren. Detta beskrivs i avsnitt 6.4.5.

Synen på registerstatistikens kvalitet bör vara likartad inom olika registerenheter. Samma principer för kvalitetssäkring bör användas och relevanta kvalitetsindikatorer bör dokumenteras på det sätt som diskuteras i kapitel 7.

11.5 Samanvändbarhet

Den holländska centralbyrån har utvecklat ett nytt synsätt där statistikens konsistens prioriteras. Detta beskrivs i Netherlands Official Statistics (2000). Statistik från olika källor kan bli konsistent enligt den holländska metodiken, dvs. få hög sam användbarhet, genom att:

- skapa konsistens med avseende på *population* (avser både avgränsningen av objekt och objektmängd),
- skapa konsistens med avseende på *variabler* och
- använda kalibreringsmetoder som ger konsistenta *skattningar*.

För att statistiken skall vara konsistent kan man inte enbart begränsa diskussionen till registerstatistik, utan alla typer av undersökningar, urvals-, total- och registerbaserade undersökningar skall samordnas.

Konsistens med avseende på population

För att uppnå konsistens med avseende på population så samordnas objektmängderna för alla undersökningar som avser en viss objekttyp och en viss referenstid. Detta görs genom att man skapar en gemensam registerpopulation för alla dessa undersökningar. För att den gemensamma registerpopulationen skall ha så hög kvalitet som möjligt bör populationen skapas efter att all relevant administrativ information har hunnit komma in till statistikbyrån.

Konsistens med avseende på variabler

Definitioner av variabler skall vara samordnade. Besläktade variabler samgranskas och samma rättelser och imputeringar görs i alla undersökningar.

Begreppet *konsistens* avser överensstämmelse:

Två undersökningar är konsistenta om de avser samma population, objekten har samma mätvärden på gemensamma variabler och ger samma skattningar.

Begreppet *samanvändbarhet* avser att skattningar från olika undersökningar kan användas tillsammans. För att t.ex. en kvot skall vara meningsfull måste täljare och nämnare vara sam användbara.

Konsistenta undersökningar ger sam användbara skattningar.

Kalibreringsmetodik för konsistenta skattningar

Genom upprepade kalibreringar kan konsistenta skattningar göras för alla ingående undersökningar. Objekten i den gemensamma populationen delas upp i en rad block där alla objekt i varje block har värden på samma variabler. Därefter bildas skattningar för varje block. Man utgår från ett block med registerinformation för alla objekt, därefter bildas skattningar för det kvalitetsmässigt näst bästa blocket som är konsistenta med skattningar från registret. Skattningarna för det andra blocket bildas med den kalibreringsmetodik som beskrivs i avsnitt 6.3.4. I nästa steg bildas skattningar för det kvalitetsmässigt sett tredje blocket som är konsistenta med skattningar från de föregående blocken, osv. Skattningsmetodiken beskrivs i Houbiers m.fl. (2003).

Exempel – konsistent företagsstatistik

Vi illustrerar den holländska metodiken för samordnad och konsistent statistik med ett företagsexempel. Syftet är att visa hur inkonsistenser uppkommer mellan undersökningar som görs vid olika tidpunkter. Om administrativa data som beskriver förändringar i populationen rapporteras med stor tidsfördröjning kommer ramfelen att vara stora hos de undersökningar som görs tidigt (beskrivs i avsnitt 6.3.1). Med den holländska metodiken kan man bilda reviderade skattningar för alla undersökningar som utgår från ett slutligt årgångsregister.

Alla undersökningar i exemplet avser år 2, och för detta år görs en rad undersökningar vid olika tidpunkter. Först görs kvartalsvisa urvalsundersökningar, som bygger på en rampopulation (SAMU) som skapades under november år 1. Denna undersökning avser industrins produktion uttryckt som förädlingsvärde. Skattningarna görs med stöd av variabeln *antal anställda*, en hjälpvariabel som finns i urvalsramen.

Därefter görs i början av år 3 en totalundersökning med egen datainsamling, som bygger på den SAMU-population som skapades under november år 2. Även här undersöks förädlingsvärdet hos företagen. Slutligen görs en registerbaserad undersökning som bygger på data från Kontrolluppgiftsregistret. Dessa uppgifter som avser år 2 blir tillgängliga först under våren år 3.

Vissa administrativa källor som är viktiga för att avgöra vilka företag som har varit aktiva och företagens näringsgren kommer in med eftersläpning. Man kan därför först under hösten år 3 skapa den standardiserade population som avser årgångsregistret för år 2. Arbetsgången är följande:

- Med urvalsundersökningarna, som visas i bild 11.1, görs preliminära skattningar som avser industrins förädlingsvärde under år 2. Med lika sannolikheter väljs 3 företag av totalt 5. Vikterna kalibreras med *Antal anställda* som hjälpvariabel. Företaget med identiteten *JE 5* behandlas som bortfall.
- Med totalundersökningen, som visas i bild 11.2, görs preliminära skattningar som avser näringslivets förädlingsvärde för olika SNI under år 2.
- Med årgångsregistret som grund skapas en konsistent datamatrix för år 2. Denna visas i bild 11.3 och 11.4.
- Slutligen görs konsistenta skattningar för år 2. Dessa finns i tabell 11.E.

Bild 11.1 Urvalsram och urvalsundersökning avseende industrin, SNI D, år 2

A. SAMU november år 1			B. Urvalsundersökning, industrins produktion, kvartalsvis år 2				
Ftg id	SNI	Ant anst	Förädlingsvärde, milj kr				
			kv 1	kv2	kv 3	kv 4	
JE 1	A	5					
JE 2	D	210	25	25	22	25	
JE 3	D	40	15	12	7	14	
JE 4	D	120	2	ej svar	ej svar	ej svar	
JE 5	D	15					
JE 6	D	30					
JE 7	E	55					
JE 8	E	70					
JE 9	F	90					
JE 10	F	5					
JE 11	G	340					
JE 12	G	15					
JE 13	G	10					
JE 14	H	60					
JE 15	K	20					
JE 16	K	5					

C. Vikter för de olika skattningarna			
	d_i	kv 1: w_i	kv 2-4: w_i
JE 2	5/3	1,09	1,18
JE 4	5/3	1,34	1,39
JE 5	5/3	1,63	0,00

D. Skattat förädlingsvärde	
kv 1	50,7
kv 2	46,2
kv 3	35,7
kv 4	49,0
Summa år 2	181,7

Med lika sannolikheter väljs 3 företag av totalt 5. Vikterna w_i fås genom att d_i kalibreras med *Antal anställda* som styrvillkor på det sätt som beskrivs i avsnitt 6.3.4.

JE 5 behandlas som bortfall.

SAMU-ramen i bilden nedan har ändrats: Företaget *JE 5* har upphört under år 2 och ingår inte, medan *JE 17* har tillkommit. Två företag är bortfall i totalundersökningen. Värden för dessa imputeras med kvoten *Förädlingsvärde/Anställd* som beräknas med de företag som svarat inom samma bransch. Det imputerade värdet för *JE 10* beräknas genom: $10 \cdot 26 / 80 = 3,25$ miljoner kronor.

Bild 11.2 Rampopulation och totalundersökning, hela näringslivet, avser år 2

A. SAMU november år 2			B. Totalundersökning, år 2		C. Skattningar för år 2		
Ftg id	SNI	Antal anställda	Förädlingsvärde, milj kr		SNI	Ant anst	F-värde
			Före imputering	Efter imputering			
JE 1	A	5	2	2	A	5	2
JE 2	D	205	97	97	D	405	192
JE 3	D	45	21	21	E	125	193
JE 4	D	110	52	52	F	90	29,3
JE 6	D	25	12	12	G	355	121
JE 17	D	20	10	10	H	70	60
JE 7	E	60	93	93	K	25	21,6
JE 8	E	65	100	100	Summa	1075	618,9
JE 9	F	80	26	26			
JE 10	F	10	ej svar	3,25			
JE 11	G	330	112	112			
JE 12	G	20	7	7			
JE 13	G	5	2	2			
JE 14	H	70	60	60			
JE 15	K	15	13	13			
JE 16	K	10	ej svar	8,61			

Skattningarna i tabell C ovan har bildats genom att summera värdena efter imputering för respektive bransch i datamатriserna A och B.

Skattningarna för SNI D är inkonsistenta, då urvalsundersökningen ger årssumman 181,7 milj kr medan totalundersökningen ger 192 milj kr.

Under hösten år 3 skapas en standardiserad företagspopulation, där alla företag som varit ekonomiskt aktiva någon gång under år 2 ingår. För företagen i denna population samlas mikrodata från alla fyra undersökningar i samma datamatrix.

Datamatrixen i bild 11.3 är alltså fullständigt *konsistent med avseende på population*. Företagspopulationen har ändrats: dels har företaget *JE 18* som startats under året tillkommit, dessutom har näringsgren för företaget *JE 14* ändrats från H till K. Datamatrixen i bild 11.3 är alltså *konsistent med avseende på indelningsvariabeln näringsgren*. De preliminära och felaktiga SNI-koder som finns i bild 11.1 och 11.2 har rättats för alla undersökningar. I årgångsregistret finns dessutom en variabel *vikt*, som anger hur lång tid under året som företaget varit verksamt. Datamatrixen innehåller också lönesummor baserade på administrativa data från Kontrolluppgiftsregistret.

Vissa problem finns kvar i datamatrixen: Mätvärdena som avser förädlingsvärde för företaget *JE 4* är olika i urvalsundersökningen och totalundersökningen. På grund av bortfall saknas två svar i totalundersökningen och dessutom saknas två värden på grund av ramfel i totalundersökningen, som bygger på SAMU från november år 2.

Bild 11.3 Årgångspopulation och datamatrix med alla undersökningar för år 2

Företagsregistrets årgångspopulation				Urval, förädlingsvärde år 2, milj kr					Totalus år 2	Adm data
Ftg id	SNI	Ant anst	vikt, tid av år	kv 1	kv2	kv 3	kv 4	År 2	F-v, milj kr	Lön, milj kr
JE 1	A	5	1						2	0,3
JE 2	D	205	1	25	25	22	25	97	97	41,8
JE 3	D	45	1						21	9,2
JE 4	D	110	1	15	12	7	14	48	52	22,4
JE 5	D	15	0,25	2	saknas	saknas	saknas	2	saknas	0,9
JE 6	D	25	1						12	5,1
JE 17	D	20	0,5						10	2,0
JE 7	E	60	1						93	13,1
JE 8	E	65	1						100	14,2
JE 9	F	80	1						26	14,5
JE 10	F	10	1						bortfall	1,8
JE 11	G	330	1						112	51,8
JE 12	G	20	1						7	3,1
JE 13	G	5	1						2	0,8
JE 18	G	10	0,5						saknas	0,9
JE 14	K	70	1						60	13,5
JE 15	K	15	1						13	2,9
JE 16	K	10	1						bortfall	1,9

I bild 11.3 finns en inkonsistens som avser variabeln *förädlingsvärde*, på grund av att urvalsundersökningarna har gett årsvärdet 48 milj kr för *JE 4*, medan totalundersökningen har gett värdet 52 milj kr. Vi antar att totalundersökningen är mer tillförlitlig och ändrar därför kvartalsvärdena för *JE 4*. De värden som saknas i totalundersökningen ersätts med imputerade värden som baseras på kvoten *Förädlingsvärde/Lönesumma* som beräknas med de företag som svarat inom samma bransch. För *JE 5* som upphört anges värdet 0 för kvartal 2-4 och värdet 2 för hela året. Efter dessa ändringar är datamatrixen färdig.

Bild 11.4 Färdigbearbetad datamatrix med alla undersökningar för år 2

Företagsregistret				Urval, förädlingsvärde år 2, milj kr					Totalus år 2	Adm data
Ftg id	SNI	Ant anst	vikt, tid av år	kv 1	kv2	kv 3	kv 4	År 2	F-v, milj kr	Lön, milj kr
JE 1	A	5	1						2	0,3
JE 2	D	205	1	25	25	22	25	97	97	41,8
JE 3	D	45	1						21	9,2
JE 4	D	110	1	16,3	13,0	7,6	15,2	52	52	22,4
JE 5	D	15	0,25	2	0	0	0	2	2	0,9
JE 6	D	25	1						12	5,1
JE 17	D	20	0,5						10	2,0
JE 7	E	60	1						93	13,1
JE 8	E	65	1						100	14,2
JE 9	F	80	1						26	14,5
JE 10	F	10	1						3,25	1,8
JE 11	G	330	1						112	51,8
JE 12	G	20	1						7	3,1
JE 13	G	5	1						2	0,8
JE 18	G	10	0,5						2,01	0,9
JE 14	K	70	1						60	13,5
JE 15	K	15	1						13	2,9
JE 16	K	10	1						8,61	1,9

Datamatriken i bild 11.4 ovan är fullständigt konsistent med avseende på *population* och *variabler*. Nästa steg är att göra konsistenta *skattningar* för alla undersökningar. Antal årsanställda skattas med hjälp av vikterna som anger hur stor del av året som företaget varit aktivt (se avsnitt 6.2.10). Lönesummor och förädlingsvärden skattas genom att bilda summor för varje SNI. De kvartalsvisa urvalsundersökningarna kalibreras om (vikterna justeras), så att skattningarna överensstämmer med årsundersökningen. Sedan kan relationstal bildas, t.ex. produktiviteten mätt som förädlingsvärde per årsanställd. I detta relationstal kommer täljare och nämnare från olika undersökningar. Det är då en stor fördel att undersökningarna har gjorts konsistenta på mikronivå i datamatriken ovan.

Tabell 11E. Konsistenta skattningar som bygger på den konsistenta datamatriken

Näringsgren	Antal årsanställda	Lönesumma milj kr	Förädlingsvärde milj kr	Förädlingsvärde per anställd milj kr	Årslön per anställd 1000-tals kr
A	5,0	0,3	2,0	0,4	60,0
D	398,8	81,4	194,0	0,5	204,1
E	125,0	27,3	193,0	1,5	218,4
F	90,0	16,3	29,3	0,3	181,1
G	360,0	56,6	123,0	0,3	157,2
K	95,0	18,3	81,6	0,9	192,6
Summa	1 073,8	200,2	622,9	0,6	186,4

Exemplet visar principerna för hur man kan skapa konsistent företagsstatistik. Men detta är endast ett första steg mot konsistent och sam användbar statistik. Antal årsanställda i företagsregistret måste också vara konsistent mot Sys sättningsregistret och lönesummorna måste även vara konsistenta mot Kon trolluppgiftsregistret med avseende på företagspopulation och näringsgren. I PLÖS-rapporten, SCB (2004), ges en aktuell beskrivning av de samordningsproblem som finns inom den ekonomiska statistiken.

Kapitel 12. Visionen om ett effektivt system

Varje kapitel i denna bok innehåller förslag till förändringar. Nya synsätt, nya termer och nya metoder behövs för att registersystemet och registerstatistiken skall kunna utvecklas och fungera på ett ännu bättre sätt än idag. Vi ger här en sammanfattning av våra olika förslag som kräver förändringar av det nuvarande arbetssättet.

Kapitel 1

För att registerstatistiken skall kunna utvecklas behövs en *registerstatistisk teori*, där systemtänkandet är centralt. Ett väl fungerande registersystem är grunden för effektiv produktion av samhällsstatistik.

Kapitel 2-3

Registersystemets struktur presenteras. Den *registermodell* som är resultatet av Registerprojektets arbete (2001) är ett viktigt redskap för att sprida förståelse för registersystemet. De *administrativa källornas* betydelse anges och de fyra *basregistrens* roll definieras.

Kapitel 4

Olika *synsätt* diskuteras. Förståelsen för registersystemets behov och möjligheter måste spridas ytterligare så att alla som arbetar inom registersystemet förstår sin roll och sitt ansvar. Registerstatistiken har varit eftersatt på grund av att det förhärskande synsättet inte har uppmärksammat registerstatistikens behov.

Ett antal *registerstatistiska termer* införs, utan bra termer blir registertänkandet diffust och erfarenhetsutbyte försvåras.

Kapitel 5

Metodiken för att skapa register struktureras. Ett antal principer presenteras, t.ex. "*alla skall använda basregistrens standardiserade populationer*" och "*alla skall stötta basregistren så att de får så hög kvalitet som möjligt*". Det är viktigt att skilja mellan undersökningar med olika metodik, dels urvals- och totalundersökningar och dels registerbaserade undersökningar.

Datainsamlingsfasen i undersökningar med egen datainsamling motsvaras av olika slag av registerbearbetningar som görs för att skapa ett statistiskt register. Dessa registerbearbetningar måste studeras ur statistikvetenskaplig synvinkel.

Kapitel 6

Skattningsmetoder behövs även inom registerstatistiken. Att enbart summera räcker inte alltid. Det traditionella synsättet inom registerstatistiken är att inga skattningsmetoder behövs, men samma statistiska synsätt som idag finns hos de som arbetar med skattningsproblem för urvals- och totalundersökningar behövs också inom registerstatistiken.

Den information som finns i *flervärda variabler* kan tas till vara med hjälp av särskilda skattningsmetoder. Genom olika skattningsmetoder kan problem som beror på *övertäckning*, bortfall eller att *värden saknas* samt *tidsseriebrott* åtgärdas, så att kvaliteten förbättras.

Kapitel 7

Särskilda *kvalitetsindikatorer* som passar för registerstatistiken behövs. Ett antal indikatorer presenteras där man för ett visst register väljer ut de som är viktiga för just det aktuella registret. Dessutom bör man göra en *helhetsbedömning* av kvaliteten.

Kapitel 8

Registerstatistiken behöver metadatasystem där stora mängder *formaliserade metadata* kan bearbetas med hjälp av lämpliga IT-verktyg. Ett *kalendarium* med viktiga förändringar och en *definitionsdatabas* är viktiga delar i ett sådant metadatasystem.

Kapitel 9

Registerstatistikens IT-system behöver dokumenteras på ett annat sätt än den traditionella *systemdokumentation* som avser undersökningar med egen datainsamling.

Data Warehouse-tekniken kan bli ett hjälpmedel för att få:

- effektivare registerhantering,
- mer och enhetligare metadata,
- enklare och säkrare uttag,
- bättre översikt över systemets innehåll.

Kapitel 10

Väl fungerande rutiner för *integritetsskydd* är en mycket viktig del av register-systemet. Man bör överväga om användningen av personnummer kan minskas. Rutiner för *röjandeskydd* skall alltid ingå i arbetet med att presentera och sprida tabeller baserade på registerstatistik.

Kapitel 11

Idag finns brister som beror på bristande samordning. Genom att alla använder *standardiserade populationer*, *standardiserade variabler* och *grupperingar* samt metodik som ger *konsistenta skattningar*, kan många av dessa brister avhjälpas.

Vad behövs ytterligare?

Inom SCB bör det finnas *kompetens för att analysera* stora mikrodatabasmaterial. Det är särskilt analyser av *longitudinella register* som kräver särskild kompetens. En sådan analyskompetens är viktig för att kunna utveckla ny registerstatistik. Dessutom kan forskare som arbetar med SCB-material göra misstag eller komma till kontroversiella resultat. Då måste man inom SCB kunna göra egna analyser.

Även urvalsundersökningar och register bör samverka på ett annat sätt än vad som är vanligt idag. När registeranalyser visar på något intressant bör man kunna göra *riktade urvalsundersökningar* mot just de intressanta delgrupper som har lyfts fram genom registeranalyserna.

Bilaga 1. Riktlinjer för statistiska register på SCB

Enligt GD-beslut 2001-05-07 skall följande allmänna riktlinjer gälla:

1. SCB skall öka användningen av administrativa uppgifter med syfte att minska uppgiftslämnarbördan och för att kunna producera ny statistik utan att uppgiftsbördan ökar.
2. I alla undersökningar skall övervägas om och vilka registerdata som kan användas. Registerdata kan användas istället för att fråga, för att kontrollera kvalitet och som hjälpinformation för att reducera urvals- och bortfallsfelen i urvalsundersökningar.
3. SCB skall samarbeta med myndigheter som lämnar administrativa register och vid behov aktualisera behoven av statistiska variabler eller anpassning av administrativa uppgifter för att förbättra det statistiska värdet av informationen. Kontakterna skall samordnas när flera avdelningar och program är inblandade. Kvalitetsinformation skall återföras till berörda myndigheter.
4. Mottagna administrativa data skall bearbetas på ett enhetligt sätt. På motsvarande sätt som urvalsundersökningar görs enligt en vedertagen metodik, skall även SCB:s statistiska register, som bygger på administrativa material, skapas enligt vedertagna principer. Samarbete över avdelnings- och programgränser kring registerstatistiska metodfrågor skall därför ske regelmässigt.
5. All statistik skall vara kvalitetsdeklarerad och väl dokumenterad, vilket också gäller bakomliggande register. Register som ingår i registersystemet skall vara dokumenterade i Metadok och det skall finnas en fullständig SCBDOK-dokumentation.
6. Definitioner av objekt, objektsmängder och variabler skall vara samordnade. Det skall finnas standardiserade variabler som är särskilt namnsatta, definierade och lätt tillgängliga. De får inte användas med avvikande definition.
7. Varje statistiskt register skall innehålla de kopplingsvariabler som krävs för att kunna länka objekten till respektive basregister.
8. Registren utgör merparten av SCB:s omfattande mikrodatalager. Detta skall förvaltas på ett sådant sätt att kraven på sekretess och integritetsskydd upprätthålls.
9. Om behov finns av att inrätta ett permanent integrationsregister skall en arbetsgrupp tillsättas med uppgift att fastställa om ett nytt register är den bästa lösningen, klarlägga ansvarsområden, skyldigheter och rättigheter för de program som berörs.
10. SCB:s registersystem tillhör det statistiska systemet och skall kunna användas som ett effektivt alternativ för SAM-myndigheter och andra till att själva bygga upp och underhålla omfattande databaser med registerinformation

Bilaga 2. Referenser

- American Statistical Association (1996): *What Is a Survey?* Broschyr utgiven av sektionen för Survey Research Methods.
- Biemer, P., Lyberg, L. (2003): *Introduction to Survey Quality*. Wiley.
- Bolin, E., Carlsson, B., Johansson, L-A., Näsholm, H., Svalstedt, I-B. (1996): *Dataregistrering av dödsorsaksstatistiken*. Slutrapport 1996-01-26, SCB.
- Danmarks Statistik (1994): *Personstatistik i Danmark - Et registerbaseret statistiksystem*.
- Danmarks Statistik (1995): *Statistics on Persons in Denmark – A register-based statistical system*. Eurostat.
- Ekberg, J. (1995): *Høj kvalitén på befolkningsstatistiken*. Ekonomisk debatt 1995:4.
- Eurostat (1997): *Proceedings of the Seminar on the Use of Administrative Sources for Statistical Purposes*, Luxembourg, 15-16 January 1997.
- Greijer, Å. (1995): *Uppskattning av övertäckningen i RTB avseende utlandsfödda med hjälp av AKU*. Metodrapport från BoR-avdelningen 1995:3, SCB.
- Greijer, Å. (1996): *Övertäckningen i Registret över totalbefolkningen - en studie av postreturer*. Metodrapport från BoR-avdelningen 1996:7, SCB.
- Greijer, Å. (1997a): *Skattning av övertäckningen i folkbokföringen med hjälp av SCB:s inkomstregister*. Metodrapport från BoR-avdelningen 1997:11, SCB.
- Greijer, Å. (1997b): *Uppskattning av övertäckningen i folkbokföringen med hjälp av AKU, 1997*. Metodrapport från BoR-avdelningen 1997:12, SCB.
- Granquist, L., Arvidson, G., Elffors, C., Norberg, A., Lundell, L-G. (2002): *Guide till granskning*. CBM 2002:1, SCB.
- Holt, D. (2001): *Comment to Platek and Särndal*. Journal of Official Statistics, Vol. 17, No. 1, 2001, pp. 55-61.
- Houbiers, M., Knottnerus, P., Kroese, A.H., Renssen, R.H., Snijders, V. (2003): *Estimating consistent table sets: position paper on repeated weighting*. Statistics Netherlands , Discussion paper 03005, 2003.
- Hörngren, J. (1992): *The Use of Registers as Auxiliary Information in the Swedish Labour Force Survey*. R&D Report, Statistics Sweden 1992:13.
- Jeppesen, J. (1973): *Statistisk metode*. Akademiskt forlag, Köpenhamn.
- Johansson, D. (1997): *The Number and the Size Distribution of Firms in Sweden and Other European Countries*. IUI Working Paper no 483, 1997.
- Johansson, D. (2001): *The Dynamics of Firm and Industry Growth – The Swedish Computing and Communications Industry*. Kungliga tekniska högskolan, TRITA-IEO R 2001:05.

- Lindström, H. L. (1999): *Kvalitetssäkring i register för statistikproduktion med administrativt underlag*. Rapport från Registerprojektet, November 1999, SCB.
- Lundin, J. (1999): *Underlag till en kvalitetsdeklaration av FDB*. Företagsregistret, SCB.
- Lundström, S. och Särndal, C-E. (2001): *Estimation in the presence of nonresponse and frame imperfections*. Statistics Sweden 2001.
- Nanopoulos, P. (2001): *Comment to Platek and Särndal*. Journal of Official Statistics, Vol. 17, No. 1, 2001, pp. 77-86.
- Netherlands Official Statistics (2000): *Integrating administrative registers and household surveys*. Volume 15, Summer 2000.
- Nilsson, Å. (1995): *Brister i folkbokföringen - övertäckningen bland utomnordiska invandrare*. Metodrapport från BoR-avdelningen 1995:2, SCB.
- Pettersson, H., m.fl. (1993): *KVADD 1 och 2. Quality in Statistics based on Administrative Data*. SCB, rapporter från V-avdelningen juni 1993.
- Platek, R. och Särndal, C-E. (2001): *Can a Statistician Deliver?* Journal of Official Statistics, Vol. 17, No. 1, 2001, pp. 1-20.
- Qvist, J. (1999): *Täckningsproblem i Registret över totalbefolkningen RTB*. R&D Report, Statistics Sweden 1999:1.
- Registerprojektet (2001): *Det statistiska registersystemet – Utvecklingsmöjligheter och förslag. Slutrapport från Registerprojektet*. R&D Report, Statistics Sweden 2001:1.
- Rosén, B. (1990): *Om justering för undertäckning vid undersökningar med urval i "rum och tid"*. R&D Report, Statistics Sweden 1990:3.
- Ryen, L. (1999): *Upprättande av ett händelsekalendarium*. Registerprojektet, SCB, 1999.
- Sandgren, P. (1993): *Kvaliteten i statistik på administrativt insamlade data*. SCB, Utvecklingsavdelningen 30 december 1993.
- Sandgren, P., Wallgren, A., Wallgren, B. (1994): *Att mäta sysselsättning med skatteadministrativa kontrolluppgifter*. SCB, Utvecklingsavdelningen 15 april 1994.
- SCB (1998): *Att köpa statistiska undersökningar*.
- SCB:s vetenskapliga råd (2000): *Förslag till rekommendationer från VR:s möte 2000-11-08*
- SCB (1999): *Slutrapport från LOUISE-projektet*. SCB 1999.
- SCB (2001a): *Kvalitetsbegrepp och riktlinjer för kvalitetsdeklaration av officiell statistik*. MIS 2001:1, SCB.
- SCB (2001b): *SCB-data för forskare, individbaserade register med innehållsbeskrivning*. SCB:s hemsida www.scb.se

SCB (2001c): *Statistisk röjandekontroll av tabeller, databaser och kartor*. CBM, SCB september 2001.

SCB (2002a): *Företagsenheten i den ekonomiska statistiken*. Bakgrundsfakta till Ekonomisk statistik, 2002:3.

SCB (2002b): *Dokumentation av FRIDA – inkomståret 2000*. Programmet för Ekonomisk Velfärdsstatistik, SCB.

SCB (2003a): *Slutrapport från mikrodatabprojektet*. SCB, september 2003.

SCB (2004): *PLÖS – Samordning av produktion, löner och sysselsättning*. Bakgrundsfakta till Ekonomisk statistik, 2004:11.

SOU 2001:34: *Behovet av ekonomisk statistik*. Utredningen om översyn av den ekonomiska statistiken.

SOU 2002:99: *Barns skador i Sverige*. Barnsäkerhetsdelegationen.

Sundgren, B. (1992): *Databasorienterad systemutveckling*. Studentlitteratur.

Wallgren, A. och Wallgren, B. (1983): *Undersökningar och statistik*. Natur och Kultur.

Wallgren, A. och Wallgren, B. (1986): *Företagets informationssystem – Statistisk analys med företagets administrativa data*. Matematiska institutionen, Universitetet i Linköping, LiU-MAT-R-86-1.

Wallgren, A., Wallgren, B. (1998): *Att länka ett system av tidsserier – Omräkning av Arbetskraftsundersökningarna 1987-1992*. Bakgrundsfakta till Arbetsmarknads- och Utbildningsstatistiken 1998:2, SCB.

Wallgren, A., Wallgren, B. (1999): *Administrative Registers in an Efficient Statistical System – How Can we Use Multiple Administrative Sources?* Statistics Sweden and Eurostat.

Wass, C. (2002): *Tidsseriebrott I Utbildningsregistret*. Bakgrundsfakta till Arbetsmarknads- och Utbildningsstatistiken 2002:1

Wilén, A. och Johannesson, I. (2002): *Ett nytt system för Registret över totalbefolkningen – Fler möjligheter och bättre kvalitet*. Bakgrundsfakta till Befolknings- och Velfärdsstatistik 2002:2, SCB.

Voss, G. (1994): *En förbättrad datainsamlingsmetod använd i undersökningen om specialskolornas intäkter och kostnader budgetåret 1992/93 – Ett metodförsök*. SCB-rapport 20n juni 1994.

Bilaga 3. Registerstatistiska termer

Term	Förklaring	sida
Adjungerad variabel	Härledning av variabler genom adjungering. Varje objektinstans i ett register kan knytas till högst en objektinstans i ett annat register. Genom denna relation kan variabler i det andra registret adjungeras till objekten i det första registret.	64, 67, 89, 101
Administrativt informationssystem	Ett informationssystem som skall uppfylla administrativa syften. Exempel: Ett företags lönesystem, samhällets folkbokföringssystem och olika skattesystem.	22, 23
Administrativt register	Register som används i ett administrativt informationssystem.	10-12, 17, 34, 37-40
Aggregerade data	Avser sammanfattande uppgifter, makrodata, för grupper av objekt och brukar presenteras som innehållet i statistiska tabeller.	56, 58
Aggregerad variabel	Härledning av variabler genom aggregering. Varje objektinstans i det första registret kan knytas till många objektinstanser i det andra registret. Man aggregerar värdena för de mikroobjekt i det andra registret som är kopplade till respektive makroobjekt i det första registret.	64, 67
Aggregeringsvariabel	De egentliga statistiska variablerna används då datamatriken analyseras och beskrivs. Vissa variabler, indelningsvariablerna, används för att dela in observationerna i redovisningsgrupper och för varje grupp beräknas beskrivande mått för andra statistiska variabler, <i>aggregeringsvariablerna</i> .	66, 112, 113, 169, 185
Aktuellt bestånd	Aktuellt bestånd/stock är den registerversion som uppdaterats med all tillgänglig information för att avse nu aktiva/levande objekt.	28, 59
Avidentifiera	Genom att utesluta alla identifierande variabler i ett register skapas en avidentifierad datamatrik.	49, 55, 72, 189, 190, 192, 193
Basregister	Ett statistiskt register av stor betydelse för hela register-systemet. Ett basregisters huvudsakliga uppgift i systemet är att avgränsa viktiga objektmängder och att ha länkar till andra basregister.	16, 17, 25, 27-30, 32-38, 44, 48, 58, 59, 67, 69, 70, 77-81, 85, 87, 97, 101, 163, 175
Bearbetningsdatabas	Databas som utformats för effektiv granskning, rättning och annan bearbetning på mikronivå	179, 183
Bestånd vid viss tidpunkt	Beståndet/stocken vid en viss tidpunkt är den version av ett register som uppdaterats för att beskriva objektmängden vid tidpunkten.	28, 60
Bortfallskorrektion	Skattningsmetod som syftar till att minska bortfallsfel.	76, 77, 138-142, 145, 146, 149, 151, 152
Bruttofel	Andelen felklassificerade i registret.	106, 107

Datamatrix	Statistiska uppgifter, mikrodata, som ordnats så att matrixens kolumner innehåller mätvärden på variabler och raderna i matrisen innehåller observationer för objekten. En datamatrix är: - En uppsättning mikrodata som skall användas för vissa statistiska bearbetningar - Innehåller huvudsakligen statistiska variabler - Kan vara avidentifierad - Kan komma från en urvals-, total- eller registerbaserad undersökning - Kan lagras i en databastabell, men kan även skapas vid beräkningstillfället	54-58
Definitionsdatabas	Formaliserade metadata för att hantera stora mängder av variabeldefinitioner som ändras över tiden.	174, 175, 196, 207
Demografisk händelse	Objekt föds, flyttar, förändras eller upphör att existera. Dessa fyra olika typer av händelser kallar vi för de demografiska händelserna, det är dessa som förändrar objektmängden.	59, 60
Dokumentation	Arbetet med att skapa metadata.	17, 34, 65, 71, 73, 88, 94, 95, 99, 109, 161, 166-168, 170-176, 199, 200
Enhetsavgränsning	Definition av objekttyp.	61, 156
Enkel variabel	En variabel som inte är sammansatt av flera variabler. T.ex. Ålder på en person. Jämför med sammansatt variabel.	63
Envärd variabel	En variabel som antar endast <i>ett</i> värde för varje objektinstans.	63, 65, 114, 115
Flervärd variabel	En variabel som antar <i>flera</i> värden för <i>vissa</i> objektinstanser.	63, 65, 69, 101, 110, 112, 114-132, 165, 169
Flödesvariabel	Anger summor för olika tidsperioder t.ex. löneinkomst i tkr under ett år.	62, 132
Främmande nyckel	Främmande nycklar (referensvariabler) används för att beskriva relationer mellan olika objekt.	65-67, 183, 185
Granskning	Grunddata kontrolleras för att upptäcka och åtgärda fel eller misstänkta fel.	34, 40, 46, 52, 64, 68, 70, 81, 85, 89-101, 108, 109, 135, 137, 161-165, 179, 195
Grupperad variabel	Härledning av variabler genom gruppering av värden och klassindelning. En kvalitativ variabel grupperas. T.ex. kan Sveriges kommuner delas in i fyra grupper: Skogslänens inland, övriga Norrland, Storstadsregioner, Övriga kommuner.	56, 63, 199, 200
Historikregister	Register som innehåller information om alla demografiska händelser som inträffat för ett objekt. Ett objekt som har varit med om t.ex. tre händelser återfinns på tre rader i datamatriken etc.	60

Hjälpinformation	Registervariabler som används vid estimationsfasen i en urvalsundersökning.	18, 19, 140
Händelseregister	Händelseregistret för en viss period, är det register som innehåller information om alla demografiska händelser, som inträffat under perioden. För varje slag av händelse skapas ett register.	60
Härledd variabel	Ny variabel bildas med hjälp av befintliga variabler. Härledning kan ske genom gruppering/klassindelning, beräkningar, adjungering eller aggregering. Vi skiljer även på lokala och importerade härledda variabler	40, 41, 44, 63-65, 67, 70, 87-89, 92, 95, 101-107, 109, 165
Härledda objekt	Härledda objekt skapas genom registerbearbetningar som använder befintlig information om relationer. Exempel: Personer som är folkbokförda på samma lägenhet bildar bostadshushåll.	40, 41, 70, 82, 84, 86, 96, 97, 164
Identifierande variabel	Identifierande variabler används för att entydigt identifiera objekt. Motsvarande IT-term är primärnyckel.	65-67
Importerad härledd variabel	En variabel som hämtas från de härledda variablerna i ett annat statistiskt register.	65, 87
Importerad primär variabel	En variabel som hämtas från de primära variablerna i ett annat statistiskt register.	65, 87
Importerad variabel	En variabel som hämtats från ett annat statistiskt register	65, 67, 71, 105, 144, 172
Imputera	Saknade värden ersätts med imputerade värden som beräknas med en sannolikhetsfördelning eller en modell.	66, 90, 95, 98, 101, 107, 137, 141-146
Imputeringsfel	Imputerat värde avviker från det korrekta okända värdet.	143
Indatabas	<i>Inmatningsdatabas, laddningsdatabas.</i> Databas som utformats för att effektivt ta emot material, oavsett kvalitet	179
Indelningsvariabel	Indelningsvariabler används för att dela in observationer i redovisningsgrupper. Kallas även spridningsvariabel.	56, 66, 67, 78, 113, 134, 165, 166, 169, 185, 193, 197
Integrationsregister	Genom att sammanföra eller integrera information från flera register skapas ett integrationsregister. Integrationsregister består <i>uteslutande</i> av variabler från andra statistiska register.	15, 29, 35, 44, 47, 53, 58, 59, 67, 69, 70, 72, 76, 77, 82, 85, 173, 196
Integrerad datainsamling	En administrativ myndighet samlar in uppgifter, som inte används administrativt, åt en statistisk myndighet. Kontrolluppgiften är ett exempel på detta, där samlar Skatteverket även in vissa uppgifter åt SCB.	31, 73, 83
Integritetsskydd	Uppgifter som lämnas till SCB omfattas av statistiksekretessen i sekretesslagen. Vidare finns särskilda regler om skydd mot kränkning av enskilda personliga integritet genom behandling av personuppgifter i personuppgiftslagen som anger när behandling av personuppgifter är tillåten. För känsliga personuppgifter gäller ännu strängare regler.	17, 188-193
Intressepopulation	Avser populationen i den teoretiska frågeställningen.	74, 75

Join	Matchning av databastabeller. Se matchning.	185, 186
Kalendarium	Förändringar bör noteras och lagras, så att det blir lätt att få information för att bedöma jämförbarheten över tiden. Ett <i>kalendarium</i> är ett IT-system med formaliserade metadata där man kan söka information med tid, register och variabel som sökvillkor.	171-173, 176
Kalibrering	Skattningsmetod som använder vikter, där vikterna justeras med syftet att minska effekterna av olika felkällor.	133, 135, 136, 150-152
Klassifikation	En kvalitativ variabel bildas genom att objekt klassificeras, dvs. indelas i ett antal grupper, sådana variabler kallas därför för klassifikationer. Exempel: Näringsgren SNI, yrke SSYK.	34, 62, 106, 147, 152, 153
Klassifikationsfel	Felaktigt värde på en kvalitativ variabel. Se även nettofel och bruttofel.	106, 107, 165, 166, 169
Klassifikationsdatabas	Formaliserade metadata för att hantera alla koder för de viktiga statistiska standarder/klassifikationer, samt nycklar mellan olika versioner av dessa standarder.	174, 199
Klassindeldad variabel	Härledning av variabler genom gruppering av värden och klassindelning. En kvantitativ variabel klassindelas, t.ex. ålder och inkomst.	56, 63
Kodning	Ibland används uppgifter i textform för att skapa statistiskt användbara variabler. Genom en kodningsprocess överförs information i delvis ostrukturerad text till helt strukturerade variabelvärden. Sifferkoder används för att ange olika kategorier för en kvalitativ variabel t.ex. 1 för man, 2 för kvinna.	40, 107, 108
Kodningsfel	Felaktiga värden uppstår under kodningsprocessen.	156, 165
Kohort	En grupp av objekt som har sitt ursprung från en viss period. Longitudinella register kan analyseras genom att olika kohorter följs över tiden.	13, 14
Kombinationsobjekt	Skattningsmetod för flervärda variabler där varje kombination av objekt och värde på den flervärda variabeln motsvarar ett kombinationsobjekt i en ny datamatrix med vikter.	116, 117, 119, 120, 122, 123, 125-127, 130, 131, 152-155, 186
Kombinationsundersökning	Undersökning där total-, urvals- och registerbaserad undersökningsmetodik används för olika delar av populationen.	69
Kommunikationsvariabel	Kommunikationsvariabler som namn, adress och telefonnummer används då SCB behöver kontakta objekt i samband med enkät eller intervju.	65, 67
Konceptuell modell	En konceptuell modell beskriver systemets objekttyper och relationer oberoende av fysiska/tekniska restriktioner.	24-27, 61
Konsistens	Konsistens betyder överensstämmelse. Undersökningar är konsistenta om de avser samma population och objekten har samma mätvärden på gemensamma variabler och ger samma skattningar.	34, 45-47, 57, 88, 124, 126, 144, 145, 149, 150, 156, 159, 194, 197, 198, 201

Konsistenskontroll	Genom samgranskning av mikrodata från flera undersökning kan inkonsistenser avseende objekt och variabler upptäckas. Se även samgranskning.	92, 99, 162, 163, 195
Kopplingsvariabel	En länk mellan två register består av en eller flera kopplingsvariabler. En kopplingsvariabel består av information som identifierar enskilda objekt eller grupper av objekt.	16, 25, 29, 30, 32, 33, 38, 44, 47, 66, 67, 79, 122, 164, 165, 179, 184, 189, 196
Kvalitativ variabel	Mäter egenskaper hos objekt som anger kategoritillhörighet t.ex. kön.	62, 63, 101, 102, 105, 106, 141-144
Kvalitetssäkring	Kvalitetssäkring syftar till att undersöka, åtgärda och förebygga kvalitetsbrister som kan uppstå i olika moment i arbetet med ett register.	17, 21, 40, 54, 71, 161, 166, 201
Kvantitativ variabel	Mäter egenskaper hos objekt som kan beskrivas med talvärden t.ex. ålder hos en person uttryckt i år.	62, 63, 91, 100-102, 104, 111
Källregister	Med källregister avser vi både de administrativa källorna och de SCB-register som används för att skapa ett nytt register.	39, 41, 58, 71, 73, 74, 88, 90, 137, 144, 170, 178, 184, 188
Lokal primär variabel	En primär variabel som direkt har hämtats från ett administrativt register kallas lokal primär variabel.	64, 65, 67, 87, 90, 199
Lokal variabel	En variabel som, för första gången inom registersystemet, bildas inom ett visst register är lokal variabler i det aktuella registret.	65
Lokalt härledd variabel	En härledd variabel som har bildats i det aktuella registret.	65, 67, 87
Longitudinell kvalitet	Vilka jämförelser kan göras på mikronivå över tiden?	109, 167, 168
Longitudinellt register	Ett integrationsregister där register för olika tidpunkter/perioder har sammanförts så att man kan följa samma objekt över tiden.	13, 15, 47, 53, 58, 107, 108, 173, 174, 196, 207
Länk	Består av en eller flera kopplingsvariabler som identifierar objekt. Om sådana variabler har hög kvalitet och finns i flera register är de lämpliga för att användas som länkar när flera register matchas. En länk består av information som identifierar enskilda objekt.	15-17, 25, 26, 28, 30-37, 45, 47, 51, 67, 179
Länkning av tidsserier	Skattningsmetod vars syfte är att öka jämförbarheten över tiden.	147-155
Makrodata	Avser sammanfattande uppgifter för grupper av objekt och brukar presenteras som innehållet i statistiska tabeller.	12, 56, 58, 170
Makrometaddata	Metaddata för makrodata. Beskriver innehållet i statistiska tabeller.	170
Matchning	Länkarna i två eller flera register jämförs. Resultatet blir träff eller icke-träff (mismatch).	15, 29, 34, 38, 40, 45, 65-67, 76, 77, 79, 84, 85, 89, 177, 180, 182-187, 189
Matchningsfel	Uteblivna eller falska träffar vid matchning. Se även mismatch.	164, 165

Metadata	Den information som behövs för att kunna använda och tolka statistik. Beskriver data genom att ange definitioner av objekt och variabler samt metodik och kvalitet.	17, 39, 44, 45, 71, 162, 166, 170-175, 185, 195-197
Mikrodata	Avser uppgifter för enskilda objekt	12, 38, 39, 55, 58, 71, 100, 170
Mikrometadata	Metadata för mikrodata. Beskriver innehållet i datamatriser med mikrodata.	170
Mismatch	Se matchning	40, 64, 68, 71, 77, 85, 137, 164, 165, 167, 186
Modellering	I IT-sammanhang ett formaliserat hjälpmedel vid utveckling av system. Används för att minska risker och utvecklingskostnader.	61, 176, 177, 179, 180
Modellfel	Härledda variabler kan skapas med en modell. Om modellens värde avviker från korrekt värde uppstår modellfel.	102, 143, 165
Mottagningskontroll	Granskning av administrativa data i register som kommer utifrån till SCB.	40, 70, 71, 99, 162, 163, 171, 189, 195
Målpopulation	Intressepopulationen avser populationen i den teoretiska frågeställningen. <i>Målpopulationen</i> avser den operationaliserade populationen, den teoretiska intressepopulationen har översatts till en konkret och undersökningsbar population.	74-78, 111, 133, 169
Mätvärde	Uppmätt värde på en variabel hos ett visst objekt.	40, 55, 61, 62, 66
Naturlig slumpvariation	Skattningar baserade på register kan påverkas av naturlig slumpvariation, varför skillnader och förändringar inte skall tolkas okritiskt.	169
Nettofel	Ett mått på de systematiska fel som beror på klassifikationsfel.	106, 107
Null	Null-värde. Databasterm för <i>värde saknas, tom cell</i> . Skall inte förväxlas med noll eller blank.	186
Nyckel	Se länk	65-67, 183, 185
Objekt	En population består av ett antal objekt eller enheter.	12, 15-18, 23, 24, 61
Objektavgränsning	Definition av objekttyp i ett register eller en population.	20, 28, 40, 44, 45, 67, 70, 81, 82, 85-87, 161, 162, 164, 199
Objektbortfall	Objekt saknas helt i registret. Se även undertäckning.	52, 137
Objektinstans	”Populationen U består av N element/enheter/objekt. I läroböcker i statistik används oftast termerna element eller enheter, inom SCB är termen objekt vanligast. Motsvarande term vid konceptuell modellering av ett IT-system, är <i>objektinstans</i> . Med den terminologin motsvaras en population av personer av objekttypen person där varje enskild person i populationen är en objektinstans. Termen objekt används ofta som synonym till objektinstans	61-64, 66, 78, 84

Objektmängd	När vi beskriver ett register utan att avse en viss undersökning använder vi begreppet <i>registrets objektmängd</i> . Men varje statistiskt register skapas för en eller flera huvudsakliga användningar eller undersökningar. Det är därför vanligt att registrets objektmängd överensstämmer med huvudundersökningens <i>population</i> .	12, 17, 25, 28, 30, 32, 34, 37, 40, 44, 55, 59, 60, 67, 69, 73-78, 80, 81, 85, 161, 163, 164, 169, 197, 201
Objektselektion	Se selektion	15, 73
Objekttyp	Det slag av objekt som ett register eller en population består av. Exempel: Person, hushåll, arbetsställe, fastighet.	24-28, 30, 32, 33, 35, 38, 45, 61, 62, 64-66, 74, 82-85, 87, 101, 115, 130, 156, 161, 167, 201
Observation	Alla mätvärden för ett visst objekt. Observation kallas även post	55, 56, 66
Partiellt bortfall	Vissa variabelvärden saknas för objekt som finns i registret.	52, 77, 90, 95, 110-112, 137, 142-145, 165, 169, 201
Population	Avgränsning i tid och rum av en mängd objekt av samma typ med vissa egenskaper. En population avgränsas för att passa syftet med en viss statistisk undersökning.	12, 30, 51, 61, 73-80, 164, 167, 197-199, 201
Populationsavgränsning	Synonym till populationsdefinition.	45, 80, 92, 127, 134
Post	Observation	39, 54, 66
Postbeskrivning	Enkel form av metadata, där variabler beskrivs	39, 90, 91, 98, 171
Primärnyckel	IT-term för identifierande variabel	65, 66, 183, 185
Primär variabel	Variabel som har hämtats från ett administrativt register. Vi skiljer på lokala och importerade primära variabler.	65, 87, 88
Primärt statistiskt register	Statistiskt register som direkt baseras på åtminstone en administrativ källa.	58, 59, 67, 69, 70, 75, 77, 82, 85, 170, 178, 199
Rampopulation	Avser den objektmängd som ramen och ramförfarandet leder till.	74, 80, 133, 134, 198, 202, 203
Referensvariabel	Referensvariabler (främmande nycklar) används för att beskriva relationer mellan olika objekt. Vid matchning av register som innehåller uppgifter om olika objekt ger referensvariabler träff mellan relaterade objekt.	65-67
Register	En fullständig förteckning över objekten i en viss definierad objektmängd, på grund av kvalitetsbrister kan vissa objekt saknas. Objekten skall vara identifierbara så att registret kan uppdateras och utökas med nya variabelvärden för respektive objekt. Register: - Består ofta av många delregister och registerversioner - Består av alla objekt som tillhör en viss definierad objektmängd/population - Objekten kan identifieras - Innehåller flera slag av variabler (se avsnitt 4.5.4), både statistiska variabler och andra - Lagras i en eller flera databaser	12, 55, 58

Register

– <i>Administrativt register</i>	Register som används i ett administrativt informationssystem	10-12, 17, 34, 37-40
– <i>Aktuellt bestånd</i>	Aktuellt bestånd/stock är den registerversion som uppdaterats med all tillgänglig information för att avse nu aktiva/levande objekt.	28, 59
– <i>Basregister</i>	Ett statistiskt register av stor betydelse för hela registersystemet. Ett basregisters huvudsakliga uppgift i systemet är att avgränsa viktiga objektmängder och att ha länkar till andra basregister.	16, 17, 25, 27-30, 32-38, 44, 48, 58, 59, 67, 69, 70, 77-81, 85, 87, 97, 101, 163, 175
– <i>Bestånd vid en viss tidpunkt</i>	Beståndet/stocken vid en viss tidpunkt är den version av ett register som uppdaterats för att beskriva objektmängden vid tidpunkten.	28, 60
– <i>Historikregister</i>	Register som innehåller information om alla demografiska händelser som inträffat för ett objekt. Ett objekt som har varit med om t.ex. tre händelser återfinns på tre rader i datamatriken etc.	60
– <i>Händelseregister</i>	Händelseregistret för en viss period, är det register som innehåller information om alla demografiska händelser, som inträffat under perioden. För varje slag av händelse skapas ett register.	60
– <i>Integrationsregister</i>	Genom att sammanföra eller integrera information från flera register skapas ett integrationsregister. Integrationsregister består uteslutande av variabler från andra statistiska register.	15, 29, 35, 44, 47, 53, 58, 59, 67, 69, 70, 72, 76, 77, 82, 85, 173, 196
– <i>Källregister</i>	Med källregister avser vi både de administrativa källorna och de SCB-register som används för att skapa ett nytt register.	39, 41, 58, 71, 73, 74, 88, 90, 137, 144, 170, 178, 184, 188
– <i>Longitudinellt register</i>	Ett integrationsregister där register för olika tidpunkter/perioder har sammanförts så att man kan följa samma objekt över tiden.	13, 15, 47, 53, 58, 107, 108, 173, 174, 196, 207
– <i>Primärt statistiskt register</i>	Statistiskt register som direkt baseras på åtminstone en administrativ källa.	58, 59, 67, 69, 70, 75, 77, 82, 85, 170, 178, 199
– <i>Statistiskt register</i>	Register som har bearbetats för statistiska ändamål. Genom att bearbeta administrativa register skapas ett statistiskt register.	11-13, 16, 17
– <i>Årgångsregister</i>	Den registerversion som innehåller alla objekt som har funnits någon gång under ett visst år. Objekt som tillkommer eller upphör under året skall ingå med angivna händelsedatum.	28, 60, 77, 93, 101, 114, 132, 134, 197, 202, 204
Registerbaserad undersökning	Statistisk undersökning utan egen datainsamling. Istället används befintliga administrativa eller statistiska register.	12, 13, 15, 19, 20, 51-54, 68-71, 74, 170-172, 176, 194, 195
Registerhållningsenkät	En enkät som skickas till de objekt i ett basregister där uppgifter saknas eller kan misstänkas vara inaktuell.	79, 81, 83, 93, 164
Registerpopulation	Avser objektmängden i det register som har skapats för den aktuella undersökningen, dvs. den population man faktiskt undersöker.	71, 74-78, 80, 94, 97, 101, 106, 112, 126, 133, 134, 159, 162-164, 167-169, 197, 198, 201
Registerstatistik	Statistik som framställs genom registerbaserade undersökningar	10-13
Registerstatistiska bearbetningar	De bearbetningar som görs för att skapa ett statistiskt register. Bearbetningarna består bland annat av kontroll av grunddata, åtgärder pga. saknade värden, samkörningar och selekteringar, bildande av härledda objekt och härledda variabler.	17, 40, 70

Registersystem	För att mikrodata skall kunna samutnyttjas effektivt skall de statistiska registren infogas i ett gemensamt och samordnat registersystem.	11, 17, 24-27
Registrets objektmängd	Avser objektmängden i ett register när man inte avser en viss undersökning.	69, 70, 73, 75
Relation	Mellan olika objekt av samma objekttyp kan det finnas relationer av olika slag, t.ex. förälder – barn. Mellan objekt av olika objekttyp kan det också finnas relationer av olika slag, t.ex. anställd – arbetsgivare. Då många relationer är juridiskt viktiga finns uppgifter om relationer ofta i administrativa register.	24-30, 35, 41, 64-66, 74, 115, 185
Relationsobjekt	En relation mellan två objekt kan ses som ett relationsobjekt. Det är nödvändigt att objektifiera en relation när det finns variabler som måste knytas till kombinationer av objekt snarare än till enskilda objekt. Exempel: Relationen person anställd av företag, är en statistiskt viktig relation som kan beskrivas med variabler som yrke, lön etc. Då samma person kan ha flera anställningar är dessa variabler inte personvariabler.	26, 29, 31, 115, 116, 152
Relevansfel	Om populationens definition och/eller variabelernas definitioner inte är adekvata för undersökningens syften, så innehåller undersökningens resultat relevansfel.	19, 68, 75, 118, 160
Replikering	Automatiserad, regelbunden kopiering av data mellan servrar	183
Röjanderisk	Om man med hjälp av statistiska tabeller kan härleda egenskaper av känslig natur hos enskilda personer eller företag, innebär publicerandet av dessa tabeller risk för röjande. Ett röjande innebär dels att man kan identifiera objektet och dels att man kan bedöma variabelvärden för dessa enskilda objekt.	52, 190-193, 195
Samanvändbarhet	Skattningar från olika undersökningar kan användas tillsammans.	88, 164, 165, 194, 196, 198-201
Sambearbetning	En registerbearbetning där <i>flera</i> register bearbetas.	34, 40, 47, 53, 76, 184, 188, 201
Sammansatt variabel	En variabel som är sammansatt av flera variabler. T.ex. Telefonnummer är sammansatt av riktnummer och abonnentnummer.	63
Samgranskning	Genom samgranskning av mikrodata från flera undersökning kan inkonsistenser avseende objekt och variabler upptäckas. Se även konsistenskontroll.	99, 195, 198, 210
Selektion	Ett nytt register skapas genom att välja vissa objekt och vissa variabler ur ett befintligt register. Vi skiljer på objektselektion, där vissa objekt selekteras och variabelselektion, där vissa variabler selekteras	15, 40, 51, 63, 78, 89, 110
SiP	Systemutveckling i Praktiken – SCB:s rekommenderade systemutvecklingsmodell	175-177

Skattning	Med ett befintligt statistiskt register kan tabeller med skattningar produceras och publiceras. Termen skattning används allmänt för urvalsundersökningar, men bör även användas inom registerstatistiken. Även här måste man skilja på det verkliga värdet och den skattning som registret ger.	11, 18-20, 51, 52, 69, 71, 98, 110-112, 198, 201, 202
Skattningsmetod	Den metod som används för att beräkna skattningar. Inom registerstatistiken är sättet att skapa ett register den skattningsmetod som avgör vilka skattningar man får. För att åtgärda problem med flervärda variabler, övertäckning, saknade värden samt tidsseriebrott kan ytterligare skattningsmetoder användas.	71, 110, 111
Spridningsvariabel	Se indelningsvariabel	66, 181
SQL	Standardiserat, mängdorierat språk för kommunikation med de flesta relationsdatabaser och vissa andra datakällor	54, 175, 185-187
Standarder	Vissa klassifikationer används som statistiska standarder, dvs. som enhetliga variabeldefinitioner. Se klassifikation	34, 174, 179, 182
Standardiserad gruppering	Många variabler måste klassindelas eller grupperas. Dessa klassindelningar och grupperingar måste göras enligt samordnade regler, annars blir inte tabeller från olika källor jämförbara.	197, 200
Standardiserad population	En viktig användning av basregistren är att skapa standardiserade populationer, som sedan används för att skapa registerpopulationer i systemets övriga register. Dessa standardiserade populationer skall ha hög kvalitet med avseende på täckning och viktiga indelningsvariabler. De ska-pas därför efter den tidsperiod eller tidpunkt de skall avse.	17, 28, 47, 77, 197, 198, 200, 202, 206, 207
Standardiserad variabel	Variabler av grundläggande betydelse som används inom flera register. Det skall finnas ett tydligt ansvar för dessa variabler med avseende på definitioner, namngivning och dokumentation.	17, 34, 65, 88, 197, 199, 207
Statistisk variabel	En variabel är en mätbar egenskap hos objekt och definieras genom vilken objekttyp som har egenskapen t.ex. inkomst för personer eller inkomst för hushåll är två olika variabler.	60, 62, 66, 67, 83, 87, 88, 109
Statistiskt informationssystem	Ett informationssystem som skall uppfylla statistiska syften. Till skillnad från enstaka undersökningar så generar ett statistiskt (informations)system ständigt nya data.	11, 22-24
Statistiskt register	Register som har bearbetats för statistiska ändamål. Genom att bearbeta administrativa register skapas ett statistiskt register.	11-13, 16, 17
Stjärnmodell	En typ av uttagsdatabas som bl.a. är användbar för lagring av longitudinella register	180, 181
Stockvariabel	Anger egenskapen vid en viss tidpunkt t.ex. ålder hos en person vid en viss tidpunkt.	62, 132
Stympad flervärd variabel	Endast det viktigaste värdet för en flervärd variabel används.	115, 165

Summerbar variabel	Se kvantitativ variabel.	62
Surveystatistik	Statistik baserad på urvals- eller totalundersökningar.	10, 13, 19-22, 68, 134, 156
Tabell	Blanda inte samman (statistisk) tabell som innehåller makrodata med databastabell som kan innehålla mikrodata. I en statistisk tabell består både för- och överspalterna av variabelklasser. I en databastabell är förspalten oftast objekt och överspalten variabler.	56, 112-114
Teknisk variabel	Variabel för intern registeradministration	66, 141
Tidsreferens	Variabel som anger tidpunkt för händelser som berör objekt eller uppdatering av register.	35, 40, 66, 67, 70, 74, 78, 163, 167
Tidsseriekvalitet	Vilka jämförelser kan göras över tiden på aggregerad nivå?	129, 140, 167-169
Totalundersökning	Undersökning med egen datainsamling, där mätvärden för alla objekt i populationen samlas in.	10, 12, 13, 17, 18, 68
Tvärsnittskvalitet	Vilka jämförelser kan göras inom samma register för en viss tidpunkt/period?	168, 169
Täckningsfel	Skillnaden mellan målpopulation och registerpopulation. Se även under- och övertäckning.	75, 78, 81, 163, 164
Undertäckning	Objekt som tillhör målpopulationen saknas i registret.	52, 68, 75, 76, 79, 80, 95, 101, 112, 133, 134, 136, 137, 163, 164
Uppräkningstal	I datamatriser för urval får varje observation i urvalet representera många observationer i populationen. En betydelsefull del av arbetet med urvalsundersökningar är att beräkna dessa vikter eller uppräkningstal	56, 66
Urvalsram	Se rampopulation	17, 18, 28, 31, 74, 203
Urvalsundersökning	Undersökning med egen datainsamling, där mätvärden för ett sannolikhetsurval av populationens objekt samlas in.	10, 12, 13, 18, 20, 22, 68, 69
Utdatabas	<i>Uttagsdatabas</i> . Databas som utformats för att ta emot färdigbearbetade data och fungera effektivt för uttag	179, 189
Variabel	En variabel är en mätbar egenskap hos objekt.	55, 61-69
– <i>Adjungerad variabel</i>	Härledning av variabler genom adjungering. Varje objektinstans i ett register kan knytas till högst en objektinstans i ett annat register. Genom denna relation kan variabler i det andra registret adjungeras till objekten i det första registret.	64, 67, 89, 101
– <i>Aggregerad variabel</i>	Härledning av variabler genom aggregering. Varje objektinstans i det första registret kan knytas till många objektinstanser i det andra registret. Man aggregerar värdena för de mikroobjekt i det andra registret som är kopplade till respektive makroobjekt i det första registret.	64, 67
– <i>Aggregeringsvariabel</i>	De egentliga statistiska variablerna används då datamatriken analyseras och beskrivs. Vissa variabler, indelningsvariablerna, används för att dela in observationerna i redovisningsgrupper och för varje grupp beräknas beskrivande mått för andra statistiska variabler, <i>aggregeringsvariablerna</i> .	66, 112, 113, 169, 185
– <i>Enkel variabel</i>	En variabel som inte är sammansatt av flera variabler. T.ex. Ålder på en person. Jämför med sammansatt variabel.	63

–	<i>Envärd variabel</i>	En variabel som antar endast <i>ett</i> värde för varje objektinstans.	63, 65, 114, 115
–	<i>Flervärd variabel</i>	En variabel som antar <i>flera</i> värden för <i>vissa</i> objektinstanser.	63, 65, 69, 101, 110, 112, 114-132, 165, 169
–	<i>Flödesvariabel</i>	Anger summor för olika tidsperioder t.ex. löneinkomst i tkr under ett år.	62, 132
–	<i>Grupperad variabel</i>	Härledning av variabler genom gruppering av värden och klassindelning. En kvalitativ variabel grupperas. T.ex. kan Sveriges kommuner delas in i fyra grupper: Skogslänens inland, övriga Norrland, Storstadsregioner, Övriga kommuner.	56, 63, 199, 200
–	<i>Hjälpinformation</i>	Registervariabler som används vid estimationsfasen i en urvalsundersökning.	18, 19, 140
–	<i>Härledd variabel</i>	Ny variabel bildas med hjälp av befintliga variabler. Härledning kan ske genom gruppering/klassindelning, beräkningar, adjungering eller aggregering. Vi skiljer även på lokala och importerade härledda variabler	40, 41, 44, 63-65, 67, 70, 87-89, 92, 95, 101-107, 109, 165
–	<i>Identifierande variabel</i>	Identifierande variabler används för att entydigt identifiera objekt. Motsvarande IT-term är primärnyckel.	65-67
–	<i>Importerad härledd variabel</i>	En variabel som hämtas från de härledda variablerna i ett annat statistiskt register.	65, 87
–	<i>Importerad primär variabel</i>	En variabel som hämtas från de primära variablerna i ett annat statistiskt register.	65, 87
–	<i>Importerad variabel</i>	En variabel som hämtats från ett annat statistiskt register	65, 67, 71, 105, 144, 172
–	<i>Indelningsvariabel</i>	Indelningsvariabler används för att dela in observationer i redovisningsgrupper. Kallas även spridningsvariabel.	56, 66, 67, 78, 113, 134, 165, 166, 169, 185, 193, 197
–	<i>Klassifikation</i>	En kvalitativ variabel bildas genom att objekt klassificeras, dvs. indelas i ett antal grupper, sådana variabler kallas därför för klassifikationer. Exempel: Näringsgren SNI, yrke SSYK.	34, 62, 106, 147, 152, 153
–	<i>Klassindelad variabel</i>	Härledning av variabler genom gruppering av värden och klassindelning. En kvantitativ variabel klassindelas, t.ex. ålder och inkomst.	56, 63
–	<i>Kommunikations-variabel</i>	Kommunikationsvariabler som namn, adress och telefonnummer används då SCB behöver kontakta objekt i samband med enkät eller intervju.	65, 67
–	<i>Kopplingsvariabel</i>	En länk mellan två register består av en eller flera kopplingsvariabler. En kopplingsvariabel består av information som identifierar enskilda objekt eller grupper av objekt.	16, 25, 29, 30, 32, 33, 38, 44, 47, 66, 67, 79, 122, 164, 165, 179, 184, 189, 196
–	<i>Kvalitativ variabel</i>	Mäter egenskaper hos objekt som anger kategoritillhörighet t.ex. kön.	62, 63, 101, 102, 105, 106, 141-144
–	<i>Kvantitativ variabel</i>	Mäter egenskaper hos objekt som kan beskrivas med talvärden t.ex. ålder hos en person uttryckt i år.	62, 63, 91, 100-102, 104, 111
–	<i>Lokal primär variabel</i>	En primär variabel som direkt har hämtats från ett administrativt register kallas lokal primär variabel.	64, 65, 67, 87, 90, 199
–	<i>Lokal variabel</i>	En variabel som, för första gången inom registersystemet, bildas inom ett visst register är lokal variabler i det aktuella registret.	65
–	<i>Lokalt härledd variabel</i>	En härledd variabel som har bildats i det aktuella registret.	65, 67, 87
–	<i>Primärnyckel</i>	IT-term för identifierande variabel	65, 66, 183, 185
–	<i>Primär variabel</i>	Variabel som har hämtats från ett administrativt register. Vi skiljer på lokala och importerade primära variabler.	65, 87, 88

– <i>Referensvariabel</i>	Referensvariabler (främmande nycklar) används för att beskriva relationer mellan olika objekt. Vid matchning av register som innehåller uppgifter om olika objekt ger referensvariabler träff mellan relaterade objekt.	19, 68, 75, 118, 160
– <i>Sammansatt variabel</i>	En variabel som är sammansatt av flera variabler. T.ex. Telefonnummer är sammansatt av riktnummer och abonnentnummer.	63
– <i>Standardiserad variabel</i>	Variabler av grundläggande betydelse som används inom flera register. Det skall finnas ett tydligt ansvar för dessa variabler med avseende på definitioner, namngivning och dokumentation.	17, 34, 65, 88, 197, 199, 207
– <i>Statistisk variabel</i>	En variabel är en mätbar egenskap hos objekt och definieras genom vilken objekttyp som har egenskapen t.ex. inkomst för personer eller inkomst för hushåll är två olika variabler.	60, 62, 66, 67, 83, 87, 88, 109
– <i>Stockvariabel</i>	Anger egenskapen vid en viss tidpunkt t.ex. ålder hos en person vid en viss tidpunkt.	62, 132
– <i>Stympad flervärd variabel</i>	Endast det viktigaste värdet för en flervärd variabel används	115, 165
– <i>Summerbar variabel</i>	Se kvantitativ variabel.	62
– <i>Teknisk variabel</i>	Variabel för intern registeradministration	66, 141
– <i>Tidsreferens</i>	Variabel som anger tidpunkt för händelser som berör objekt eller uppdatering av register.	35, 40, 66, 67, 70, 74, 78, 163, 167
– <i>Uppräkningstal</i>	I datamatriser för urval får varje observation i urvalet representera många observationer i populationen. En betydelsefull del av arbetet med urvalsundersökningar är att beräkna dessa vikter eller uppräkningstal	56, 63
– <i>Viktgenererande variabel</i>	Olika registervariabler som kan användas för att skapa de vikter som behövs för att göra skattningar för flervärda variabler.	117
Variabelselektion	Se selektion	15, 89, 110
Vikt	Se uppräkningstal.	56, 111, 112
Viktgenererande variabel	Olika registervariabler som kan användas för att skapa de vikter som behövs för att göra skattningar för flervärda variabler.	117
Vy	Databasterm för anpassad presentation av innehållet i databastabell(er).	57, 181, 182, 186
Värdemängd	Mängden av de värden en variabel antar eller kan tänkas anta för någon objektinstans.	63, 174
Årgångsregister, även årgångsversion	Den registerversion som innehåller alla objekt som har funnits någon gång under ett visst år. Objekt som tillkommer eller upphör under året skall ingå med angivna händelsedatum.	28, 60, 77, 93, 101, 114, 132, 134, 197, 202, 204
Årsskiftesversion	En registerversion som avser den 31/12 eller 1/1. Passar för årlig stockstatistik.	77
Övertäckning	Registret innehåller objekt som inte tillhör målpopulationen.	29, 75, 76, 81, 82, 110, 133-136, 150, 163, 173, 206

Samhällsstatistik framställs genom i huvudsak tre typer av statistiska undersökningar: *urvalsundersökningar* och *totalundersökningar*, där någon statistikansvarig myndighet ansvarar för datainsamlingen, samt *registerbaserade undersökningar* som använder redan befintliga administrativa data.

Författarna beskriver, föreslår och diskuterar registerstatistiska metoder för att omforma administrativa data till statistiska register. De betonar ett systemtänkande i metoder och tillvägagångssätt då registersystemet bildas.

ISSN: 0283-8680