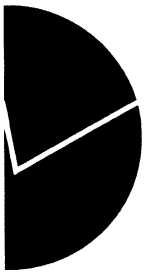

Dennis Fredriksen

Datagrunnlaget for modellen MOSART, 1993

Forskningsavdelingen / Seksjon for offentlig økonomi og personmodeller

Innhold

1. Innledning

Dette notatet dokumenterer datagrunnlaget for mikrosimuleringsmodellen MOSART med 1993 som basisår. I tillegg skal notatet understøtte framtidige oppdateringer av dette datagrunnlaget. En tidligere versjon av datagrunnlaget med 1989 som basisår er dokumentert i Fredriksen (1992) og Fredriksen (1993). Notatet inneholder også en kort omtale av en kobling av Arbeidskraftundersøkelsene opp mot de samme registrene som er brukt til datagrunnlaget for MOSART.

1.1. Modellen MOSART og behov for data

MOSART er en mikrosimuleringsmodell som starter med et utvalg av befolkningen i et basisår og simulerer det videre livsløpet for hvert enkelt individ i dette utvalget. Begivenheter som simuleres er inn- og utvandring, død, ekteskap, fødsler, skolegang, uførhet og yrkesdeltaking. Den samme yrkesdeltakingen vil bestemme pensjonsutbetalingene fra folketrygden for alders- og uførepensjonister. For hvert år som simuleres legges nye individer i form av nye årskull og nye innvandrere til utvalget modellen startet med. Resultatet av simuleringen blir en modellpopulasjon med livshistorien for hvert individ med hensyn til de begivenheter som er simulert. Styrken i mikrosimulering ligger i at modellkonseptet gjør det mulig å se de ulike kjennetegnene i sammenheng på individnivå. Dette fordrer imidlertid gode mikrodata som kan beskrive både utvalget modellen starter med og de begivenheter hvert enkelt individ kan oppleve.

Fordelen med å starte simuleringen med et utvalg av befolkningen med sine *faktiske* kjennetegn, er at disse kjennetegnene vil ha klar innflytelse på samfunnsutviklingen i flere tiår framover. To trivielle eksempler er at kjønn og alder (på et gitt tidspunkt) er fastlagt ved fødselen, og dette har sterk betydning for befolkningsutviklingen i de voksne årskullene. Nå kan en framskriving av befolkningen etter kjønn og alder skje med aggregerte data, hvor hver kombinasjon av kjennetegn er representert. Det er først når vi kombinerer mange kjennetegn, og da gjerne kompliserte sådanne, at mikrodata og mikrosimulering blir viktig.

Et godt eksempel i så måte vil være utdanning, hvor de fleste personer eldre enn 25-30 år vil beholde det samme utdanningsnivået ut livet. Arbeidsstyrkens utdanning vil dermed i lang tid framover være bestemt av hva som har skjedd i utdanningssystemet de siste tiårene. Et annet sentralt eksempel er at hittil opptjente pensjonsrettigheter i folketrygden vil påvirke tilleggspensjonene i kanskje 60-70 år framover. Dette lar seg bare beskrive tilfredsstillende gjennom individuelle pensjonspoenghistorier.

En modell som MOSART vil bygge på mange datakilder/grunnlag. Når vi refererer til «datagrunnlaget for MOSART», sikter vi til det datamaterialet som beskriver utvalget modellen starter simuleringen med. Primært omfatter dette status for hvert individ i basisåret for simuleringen. Noen av disse opplysningene ligger nær opp til å være forløpsdata, i den forstand at de følger hvert individ over en rekke år. Spesielt vil hittil opptjente pensjonsrettigheter bestå av årlige opplysninger for pensjonsgivende inntekt (=arbeidsinntekt) for hvert år fra 1967 og fram til basisåret. Samtidig bør/må de parametrene som brukes i simuleringen beregnes på grunnlag av faktiske begivenheter hentet fra forløpsdata. Generelt finnes det lite av gode forløpsdata. Datagrunnlaget for MOSART er derfor utvidet slik at det ligger til rette for forløpsanalyse med diskret tid med kalenderåret som tidsenhet. Utvalget er representativt for befolkningen i Norge for årene 1964-1993, men de fleste kjennetegnene foreligger kun fra 1985 og framover.

En mer utførlig beskrivelse av MOSART finnes i Fredriksen og Spurkland (1993).

1.2. Sammendrag

I den form modellen har fått omfatter datagrunnlaget for MOSART om lag 12 prosent av befolkningen med opplysninger hentet fra registre som omfatter hele befolkningen. Resultatfilene vil bestå av individer med faktiske opplysninger. Selv om fødselsnummer (eller tilsvarende

identifikasjon) er fjernet, er settet av opplysninger såvidt omfattende at bakveisidentifikasjon er mulig. Utfra gjeldende regler skal det derfor før arbeidet med koblingene starter, sendes en melding til Datatilsynet om oppretting av personregister. Se for øvrig Statistisk sentralbyrå (1994). Utfra vårt skjønn gjelder dette ikke koblingen av data fra Arbeidskraftundersøkelsene mot registerstatistikk, spesielt fordi filene er anonymisert før kobling.

I denne versjonen er basisåret 1993, det vil si at vi har brukt befolkningen per 1.1.1994 (≈31.12.93).Valg av basisår avhenger av aktualiteten på de ulike datakildene, og spesielt utdanning og pensjonspoeng setter visse grenser her. For disse to datakildene vil det være en klar fordel å få med data for året etter basisåret, slik at utdanning og inntektsstatus kan tilordnes korrekte verdier også i basisåret (se avsnitt 2.2 og 2.4). Dette gjør at basisåret blir eldre fordi vi må vente på ytterligere en årgang filer, men vi har likevel valgt denne tilnærmingen. Se for øvrig beskrivelsen av de ulike datakildene i kapittel 2 for valg av basisår, framdriftsplan og begrunnelser.

Grunnlaget for utvalget og de demografiske kjennetegnene er Personregisteret, som er nærmere omtalt i avsnitt 2.1. Utvalget ivaretar bestemte relasjoner mellom personer. Blant annet er ektefeller som bor sammen, eventuelt samboere med felles barn trukket ut sammen. Utvalget er supplert med personer som har utvandret eller dødd, inkludert avdøde ektefeller av enker og enkemenn. Størrelsen på utvalget i denne versjonen er 12,11 prosent, fordelt på 12 underutvalg à 1 prosent, samt to testutvalg på 0,1 og 0,01 prosent av befolkningen. Egenskapene ved utvalget er nærmere omtalt i kapittel 3.

Andre opplysninger enn de demografiske kjennetegnene er hentet fra blant annet utdanningsstatistikken, trygdestatistikken, den rekonstruerte folketrygdbasen i Rikstrygdeverket, skattestatistikken og Statens pensjonskasse. Datakildene og noen av deres eventuelle svakheter er omtalt i avsnitt 2.2-2.6. Vedlegg C og D gir en oversikt over filidenter for datakildene og for edb-programmene som er benyttet i koblingsarbeidet. Flesteparten av datakildene ligger knyttet opp mot Statistisk sentralbyrås stormaskin (comparex-maskinen på Kongsvinger), og uttaket av data har derfor skjedd her. Selve koblingsarbeidet har imidlertid foregått på en unix-maskin. Det er grunn til å tro at stormaskinen vil være faset ut innen neste oppdatering av datagrunnlaget for MOSART, og det er derfor lagt liten vekt på dokumentasjonen av koblingsprogrammene, da dette trolig likevel vil være avlegs innen neste oppdatering.

Resultatet av koblingene er organisert som 'flate' filer tilknyttet arbeidsstasjonen 'johansen'. De ulike typer filer omfatter oversikt over personene i utvalget, mødre, fedre, forløpsdata, pensjonspoeng og tre filer med data som ikke benyttes i simuleringsmodellen. De førstnevnte spesifiserte filene er splittet opp på sine respektive underutvalg. Filene er bygd opp slik at de skal være lette å lese, og ukomprimert legger de beslag på anslagsvis 1100 Mb. Innholdet i de ulike filene og bruken av disse er omtalt i kapittel 4, mens vedlegg A og B gir henholdsvis en fil- og variabelbeskrivelse.

Framskrivingen av nivået på arbeidsstyrken i MOSART er basert på begreper og tall hentet fra Statistisk sentralbyrås arbeidskraftundersøkelser, forkortet AKU. Dette er en intervjubasert undersøkelse, og AKU vil derfor ha andre definisjoner av bakgrunnsvariabler som utdanning og uførhet. For å redusere problemer med inkonsistenser har vi derfor koblet AKU mot de samme registrene som brukes i datagrunnlaget for MOSART, og dette er dokumentert i avsnitt 2.7 og vedlegg A.8.

1.3. Endringer siden forrige datagrunnlag

Modellen MOSART brukes i offentlig planlegging, og for å være relevant må modellens parametre og datagrunnlag ikke være for gamle. Det er derfor nødvendig å oppdatere datagrunnlaget med jevne mellomrom, og foreløpig har vi lagt oss på et intervall på 4 år. Forrige versjon av datagrunnlaget

hadde 1989 som basisår og er som nevnt dokumentert i Fredriksen (1992) og Fredriksen (1993)[1]. Denne oppdaterte versjonen av datagrunnlaget har 1993 som basisår, det vil si at utvalget er trukket på grunnlag av befolkningen per 1.1.1994. Oppdateringen medfører at basisåret flyttes til etter de store endringene på arbeidsmarkedet og i utdanningssystemet i årene 1988 til 1990.

Datagrunnlaget har for øvrig gjennomgått en rekke mindre og større forbedringer, forenklinger og utvidelser. Stratifiseringen av utvalget er betydelig forbedret ved at alder er delt inn i ett års aldersgrupper, noe som er viktig for overganger i utdanningssystemet, inn i arbeidsmarkedet og til alderspensjon. Utvalgsplanen er forenklet ved at avdøde ektefeller er inkludert direkte i utvalget slik at vi slipper å operere med såkalte 'supplerende' personer. Ingen personer er ekskludert fra utvalget på grunn av manglende opplysninger omkring for eksempel ektefelles identitet. Vi har antatt at det er bedre å imputere manglende informasjon på ett kjennetegn framfor å gjøre hele utvalget skjevt.

Datagrunnlaget er anonymisert i den forstand at fødselsnumrene har blitt erstattet av et vilkårlig identifikasjonsnummer som ikke kan rekonstrueres. Det er dermed ikke mulig å koble til andre opplysninger. Imidlertid oppdateres datagrunnlaget for MOSART såvidt ofte at andre opplysninger kan innarbeides i framtidige versjoner av datagrunnlaget.

Opplysningene omkring husholdning er utvidet, og i tillegg til ektepar har vi fått med samboende foreldrepar. Utvalget inneholder også opplysninger om barn/foreldre bor sammen, men både barn og foreldre er nødvendigvis ikke med i utvalget.

Vi har fått med utdanningsdata for året etter basisåret, som gir oss en oversikt over grunnskoleelever også i basisåret for simuleringen, jamfør avsnitt 2.2. Attføring er inkludert i datagrunnlaget fra og med 1989, og imputert for årene 1985-1988. Avtalefestet pensjon (afp) er tatt med, men er helt klart underrepresentert i utvalget, trolig fordi vesentlige grupper av afp faller utenfor den statistikken vi benytter.

Vi har fått med inntektsdata for året etter basisåret, noe som forenkler bruken av datagrunnlaget i simuleringsmodellen, jamfør avsnitt 2.4. Fra skattestatistikken har vi tatt med opplysninger om nettoformue ved statsskatt og ligningsverdi av bolig. Formålet med disse formuesopplysningene er å forbedre simuleringen av formue i startåret for simuleringen ved senere utvidelser av MOSART. Etter planen vil datagrunnlaget også inneholde opplysninger fra Statens pensjonskasse, men disse opplysningene var ikke ferdig tilrettelagt da dette notatet ble sluttført.

1.4. Andre datagrunnlag

Tilretteleggingen av datagrunnlaget for MOSART krever anslagsvis 4-6 månedsverk hvert 4. år, men vi får likevel ikke med oss alle de opplysningene vi kanskje kunne ønske oss. Som en løpende prosess bør derfor andre datakilder vurderes som aktuelle kandidater som datagrunnlag for MOSART. For øyeblikket ser vi ingen slike, men noen av datakildene kan bli av interesse på lengre sikt.

Det er også et moment at mye av ressursbruken går på trekkingen av utvalget, tilretteleggingen av data for simuleringsmodellen og forløpsanalyse, samt dokumentasjon. Dette er arbeid som nesten uansett valg av datakilde vil påløpe prosjektet. En egnet strategi vil være å arbeide for at de ulike registerstatistikkene blir best mulig med hensyn til innhold, kvalitet og dokumentasjon. Med et slikt utgangspunkt vil selve koblingsarbeidet ikke være særlig ressurskrevende.

[1] Det finnes ytterligere en eldre versjon av datagrunnlaget med 1987 som basisår. Denne versjonen inneholdt imidlertid kun status i basisåret, og da kun for demografiske kjennetegn og utdanning. Se Andreassen et al (1993) for detaljer.

Interjubaserte utvalgsundersøkelser
Per i dag omfatter en standard simulering i MOSART 1 prosent av befolkningen, og vi regner med at denne utvalgsstørrelsen vil stige. Dermed vil de fleste utvalgsundersøkelser ha for små utvalg til at de vil være av interesse som modellpopulasjon i MOSART. Videre vil de store undersøkelsene ha en tendens til å bli oppdatert for sjelden, som for eksempel Folke- og boligtellingen (FoB) og Familie- og yrkesundersøkelsen (FoY). Dette er svakheter som må tilskrives kostnadsrammene for disse undersøkelsene, og kan neppe endres. Derimot vil flere av undersøkelsene ha utvalg som ikke dekker hele befolkningen, eksempler er Arbeidskraftundersøkelsene som kun intervjuer personer i yrkesaktiv alder (16-74 år) og FoY som kun dekker hver femte fødselskohort i aldersgruppen 20-45 år. Også dette er svakheter som ekskluderer undersøkelsene som aktuelle modellpopulasjoner for MOSART.

KIRUT
KIRUT er en database som dekker 10 prosent av befolkningen med detaljerte forløpsdata for «Klientstrømmer Inn i, Rundt i og Ut av Trygdesystemet». Imidlertid mangler KIRUT befolkningen over 67 år (de er uansett inne i trygdesystemet), utvalget er individbasert (mangler ektefelle) og har heller ikke pensjonspoengrekker. KIRUT har dermed ingen relevans som modellpopulasjon for MOSART både fordi utvalget mangler store befolkningsgrupper og fordi de viktigste opplysningene ikke er der (pensjonspoeng).

Andre registerkoblinger
I Statistisk sentralbyrå pågår en rekke prosjekter for å koble registre for statistikkformål, blant annet for yrkesdeltaking og for trygd. Per i dag er imidlertid ikke status for disse prosjektene slik at de utgjør noen mulig modellpopulasjon. På et senere tidspunkt kan de imidlertid bli det, eller i det minste være viktige dataleverandører til datagrunnlaget for MOSART. Et av disse prosjektene skal koble personstatistikk som et grunnlag for å levere trygde- og livsforløp til Folke og boligtellingen år 2000, samt tjene som grunnlag for statistikk og forløpsanalyser innen trygd, se Statistisk sentralbyrå (1995A) for mer detaljer. Videre pågår det arbeid med et «Personregnskap» ved Seksjon for arbeidsmarkedsstatistikk, som kan/skal levere opplysninger om yrkesdeltaking basert på registre.

2. Datakilder

Kapittel 2 omtaler de datakildene som inngår i datagrunnlaget for MOSART, hvilke data vi har hentet ut og egenskaper/svakheter ved disse dataene.

2.1. Personregisteret

Grunnlaget for trekkingen av utvalget og de demografiske kjennetegnene er Personregisteret i Norge, som danner basis for det meste av befolkningsstatistikken og blant annet innkrevingen av skatter. Samtlige personer som har vært bosatt i Norge etter 1964 er tildelt et fødselsnummer som entydig identifiserer denne personen. Registeret inneholder opplysninger om blant annet hvem foreldrene er, ekteskapelig status, hvem eventuell ektefelle er og (skattemessig) bosted. Oppdatering skjer ved innrapportering av fødsler, endringer i ekteskapelig status, flytting og dødsfall. Kvaliteten på innrapporteringene er noe variabel for begivenheter knyttet til flytting og ekteskapelig status. Spesielt kan to ektefeller i ett «ektepar» ha forskjellig opplysninger omkring sin ekteskapelig status, for eksempel at den ene er separert mens den andre er skilt.

Data fra Personregisteret for befolkningen per 1. januar i ett valgt år er tilgjengelig noen måneder senere, og legger dermed ingen beskrankninger på valg av basisår sammenlignet med de andre datakildene. Dataene foreligger i en standardversjon i Statistisk sentralbyrå. Trekkingen av utvalget kan med hell starte tidlig, slik at innhentingen av de fleste dataene er ferdig når de siste datakildene foreligger. Arbeidet kan dermed avsluttes raskt etter dette, og datagrunnlaget vil få bedre aktualitet. Enkelte personer kan av ulike årsaker få byttet fødselsnummer, og i prinsippet er det mulig å koble sammen disse endringene. Vi har ikke forsøkt dette i denne eller tidligere versjoner, slik at enkelte

personer vil mangle opplysninger bakover i tid. Med det koblingsopplegget som brukes nå, hvor fødselsnummeret erstattes av et én-entydig identifikasjonsnummer før kobling, burde det ikke by på tekniske problemer å hente ut opplysninger for en person ved hjelp av mer enn ett fødselsnummer.

Opplysningene vi har hentet ut er basert på situasjonsuttaket fra Personregisteret for bosatte personer per 1.1 for hvert år fra 1986 til 1994, samt en fil over ikke bosatte personer per 1.1.1994. Hovedtyngden av informasjonen som er tatt ut er status ved utgangen av 1993, gitt ved situasjonsuttaket per 1.1.1994 for bosatte og ikke bosatte (døde, utvandret). Disse statusopplysningene er brukt til å finne hver persons fødselshistorie, ved å koble barn mot sine foreldre. Det vil si at vi har fått med samtlige fødsler etter 1950, men gradvis færre før dette.

Tabell 1. Folkemengden per 31.12. Tusen personer

År	MOSART	NOS	Avvik
1985	4154	4159	-5
1986	4180	4176	5
1987	4205	4198	7
1988	4227	4221	6
1989	4241	4233	8
1990	4256	4250	6
1991	4277	4274	4
1992	4300	4299	1
1993	4325	4325	0

Kilde: NOS Befolkningsstatistikk.

Tabell 2. Folkemengden etter kjønn og alder, 31.12.1985. Tusen personer

	Menn		Kvinner	
Alder	MOSART	Avvik vs NOS	MOSART	Avvik vs NOS
Alle	2052	-4,7	2102	-0,7
0-4	129	-0,9	124	-0,2
5-9	133	-0,3	127	-0,1
10 - 14	158	-0,3	152	0,2
15 - 19	172	0,5	163	-0,4
20 - 24	164	0,4	155	-0,4
25 - 29	159	-0,5	153	-0,3
30 - 34	158	0,3	150	0,4
35 - 39	163	-0,5	153	-0,3
40 - 44	134	-0,2	127	-0,0
45 - 49	103	0,0	101	-0,0
50 - 54	95	0,0	95	-0,1
55 - 59	101	-0,5	104	-0,2
60 - 64	108	-0,2	116	-0,1
65 - 69	97	-0,6	113	0,2
70 - 74	76	-0,6	98	0,5
75 - 79	53	-0,0	79	0,8
80 - 84	29	-0,8	54	-0,1
85 - 89	13	-0,2	28	0,1
90+	5	-0,2	11	-0,0

Kilde: NOS Befolkningsstatistikk.

Adresseopplysninger er brukt til å identifisere hvem som formelt sett bor sammen gitt at de har en tilknytning fra før, i denne versjonen omfatter dette ektepar, medforeldre og foreldre-barn. Merk at spesielt «barn» vil være registrert bosatt hos sine foreldre helt til de er ferdig med skolegangen. Situasjonsuttakene før 1.1.1994 er brukt til å finne avdøde ektefeller av enker og enkemenn, og samtidig hente ut opplysninger om endringer i ekteskapelig status og flytting til og fra Norge.

Tabell 1, 2 og 3 gir en viss oversikt over befolkningstallene og bevegelser i folkemengden sammenlignet med befolkningsstatistikk fra Norges Offentlige Statistikk (NOS). Tallene fra datagrunnlaget er blåst opp med utvalgsprosenten, og testutvalgene er utelatt (se kapittel 3). Avvikene bakover i tid for befolkningstallene er relativt små, i størrelsesorden 1-2 prosent. Avvikene knytter seg ikke til noen bestemt aldersgruppe, jamfør tabell 2. Dette antyder at populasjonen er representativ bakover i tid.

Når det gjelder bevegelser i folkemengden, tabell 3, har vi kun hentet ut data for de personene hvor vi forventet å finne data utfra årstall for siste begivenhet med hensyn til dødsfall, inn- og utvandring og endringer i ekteskapelig status. Dette reduserte omfanget av edb-jobbene, men har trolig bidratt til å gi en feilvurdering av noen bevegelsene. Antall dødsfall i 1987, inn- og utvandring bakover i tid og skilsmisser og separasjoner er undervurdert. Antall giftermål blir overvurdert. En del personer har uoppgitt dato for ekteskapelig status, og spesielt for unge gifte personer er dette opplagt ikke et årstall før 1964 (jamfør vedlegg B). Koblingen burde derfor muligens ha skjedd for totalbefolkningen, og ikke bare de som angivelig skal være berørt av begivenheter et bestemt år.

Tabell 3. Bevegelser i folkemengden. Ulike begivenheter i tusen

	1985	1986	1987	1988	1989	1990	1991	1992	1993
Fødte									
- MOSART	51,4	52,3	53,6	58,2	59,6	60,7	60,4	59,5	58,7
- NOS	51,1	52,5	54,0	57,5	59,3	60,9	60,8	60,1	59,2
- Avvik	0,2	-0,2	-0,4	0,7	0,3	0,2	-0,4	-0,6	-0,5
Døde									
- MOSART	43,9	40,0	44,2	45,4	44,5	45,8	44,3	45,3	46,5
- NOS	44,4	43,6	45,0	45,4	45,2	46,0	44,9	44,7	46,6
- Avvik	-0,5	-3,6	-0,8	0,0	-0,7	-0,2	-0,6	0,5	-0,1
Innvandret									
- MOSART	16,6	23,0	28,8	28,5	25,1	24,6	25,1	25,6	29,6
- NOS	21,9	24,2	31,1	30,0	25,8	25,5	26,3	26,7	31,7
- Avvik	-5,2	-1,2	-2,4	-1,4	-0,7	-0,9	-1,2	-1,1	-2,1
Utvandret									
- MOSART	8,7	9,4	12,8	18,2	24,4	24,1	19,6	17,5	17,6
- NOS	15,6	16,7	17,4	19,8	27,3	23,8	18,2	16,8	18,9
- Avvik	-7,0	-7,4	-4,6	-1,6	-2,9	0,3	1,4	0,7	-1,3
Giftermål									
- MOSART	21,2	21,3	22,4	22,8	22,2	23,0	20,7	20,9	19,5
- NOS	20,2	20,5	21,1	21,7	20,8	21,9	20,0	19,3	19,5
- Avvik	0,9	0,8	1,4	1,1	1,5	1,0	0,8	1,6	0,0
Skilsmisser									
- MOSART	7,5	7,3	7,7	8,2	8,8	9,4	9,4	9,4	10,6
- NOS	8,1	7,9	8,4	8,8	9,2	10,2	10,3	10,2	10,9
- Avvik	-0,6	-0,5	-0,7	-0,5	-0,4	-0,8	-0,8	-0,8	-0,3
Separasjoner									
- MOSART	6,0	9,4	10,4	11,1	11,8	12,1	11,7	12,1	11,1
- NOS	10,2	10,2	10,7	11,4	12,5	12,9	12,4	12,2	12,0
- Avvik	-4,2	-0,7	-0,3	-0,3	-0,7	-0,8	-0,8	-0,1	-1,0

Kilde: NOS Befolkningsstatistikk

Likevel burde dataene ligge til rette for analyser av fruktbarhet og med noe forsiktighet av dødelighet og bevegelser i ekteskapelig status. Inn- og utvandring virker derimot noe mer problematisk, spesielt for tidligere år.

2.2. Utdanningsstatistikken

Grunnlaget for utdanningsopplysningene i datagrunnlaget er filene «befolkningens høyeste utdanning», dokumentert i Vassenden (1993). Dette er årlige filer som for hvert år siden 1985 for hver bosatt person viser utdanningsnivå og -aktiviteter per 1.10 samme år. Grunnlaget for filene er folke- og boligtellingene fra 1970 og 1980, som omfattet hele befolkningen og som hentet inn opplysninger om utdanning. Videre blir filene oppdatert hvert år ved at alle utdanningsinstitusjoner i Norge innrapporterer sine elever og studenter og hvilke eksamener de eventuelt fullfører til Statistisk sentralbyrå. Dette har gitt oss et utdanningsregister av relativt god kvalitet som omfatter hele befolkningen.

Utdanningsstatistikken for et skoleår vil foreligge sent høsten året etter. Skal koblingsarbeidet være avsluttet ved slutten av år 't', vil år 't-1' være det siste året man kan bruke. Elever som går i grunnskolen blir kun registrert ved at de fullfører grunnskolen, og da året etter. Det kan derfor være hensiktsmessig å bruke år 't-2' som basisår, slik at man vet hvilke 15-17 åringer som går i grunnskolen i basisåret.

Noen problemer er det med utdanningsregisteret, og blant annet utdanning tatt i utlandet kommer i svært liten grad med. Dette gjelder personer som har innvandret etter 1980, selv om noe av dette er rettet opp ved senere intervjuundersøkelser. Videre har dette berørt nordmenn som tar utdanning i utlandet, men her foreligger det planer om å utnytte opplysninger fra Statens lånekasse for utdanning.

Et annet «problem» er at beskrivelsen av utdanning er svært detaljert med en 8-sifret tallkode som beskriver både utdanningens art og nivå. Dette er helt uhensiktsmessig for en simuleringsmodell for utdanning. Det er heller ikke mulig å bruke 'n' første siffer av koden, da dette gir svært grove inndelinger på enkelte områder, og samtidig urimelig findelte grupper for andre nivåer og utdanningstyper. Vi har derfor laget vår egen aggregering av utdanning, som gjør at et begrenset antall utdanningsgrupper etter vårt skjønn gir en best mulig beskrivelse av strømmer gjennom utdanningssystemet og av utdanning som forklaringsvariabel for yrkesdeltaking.

Utdanningens nivå er beskrevet ved 1-års klassetrinn, mens utdanningens art blant annet skiller mellom tekniske fag, helsefag og økonomi-administrasjon på alle nivåer. Vedlegg B inneholder en mer detaljert beskrivelse av utdanningskoden i MOSART. For mer spesielle formål, samt for å holde muligheten åpen for endringer i aggregeringen av utdanning, har vi på egne filer beholdt en ikke-aggregert versjon av utdanningsopplysningene (vedlegg A.6).

Noen utdanningskjennetegn er rettet opp. Grunnskole blir ikke registrert som utdanningsaktivitet. Personer til og med 17 år som har fullført grunnskole, har fått imputert at de gikk i grunnskolen fra de var 14 år til og med året før de fullførte. Vi har hentet inn hvilke 15-17 åringer som fullførte grunnskole i 1994. For personer som vil fullføre grunnskole etter 1994 har vi ikke fått med opplysninger om grunnskole, og dette gjelder noen av de som var 15 år i 1993, de fleste 14-åringene og alle under 14 år. Utdanning for innvandrere hentet inn ved intervjuundersøkelse mangler klassetrinn. Vi har her tilordnet det høyeste klassetrinnet de kan ha innenfor det nivået utdanningen er plassert på (stort sett spenn på to og to år).

Personer som i ett år ikke er bosatt vil mangle utdanningsopplysninger i dette året, og for utdanning som er rettet opp er eldre opplysninger feil. Vi har derfor rettet opp forløpshistoriene for utdanning ved å ta utgangspunkt i fullføringsår for hvert års utdanning. Det medfører at det har oppstått endringer i utdanningene der vi omkoder utdanningens klassetrinn (hjelpepleiere, sykepleiere).

Tabell 4 og 5 oppsummerer noen grove tall for utdanningsvariablene. Avviket i utdanningsnivå i tabell 4 er relativt lite og kan forklares med relativt enkle forhold. NOS Utdanningsstatistikk bygger på befolkningen per 1.10, mens tallene fra MOSART er per 31.12 samme år. Det forklarer både avviket i befolkningstallene og at antallet med uoppgitt utdanning i 1993 er noe for høyt (de to siste månedskullene 16-åringer samt nye innvandrere har ikke kommet med på utdanningsfilen). Utdanningsnivået er noe for høyt i 1986, og det kan tilskrives at en del utdanninger har blitt tilbakedatert som en følge av den nevnte intervjuundersøkelsen blant innvandrere.

Antallet elever og studenter avviker systematisk fra NOS Utdanningsstatistikk, og det skyldes i all hovedsak at NOS Utdanningsstatistikk angir antallet elever og studenter mens MOSART angir antallet personer under utdanning. Forskjellen oppstår ved at noen personer følger flere utdanningsaktiviteter, og dermed går igjen flere ganger i tabellen i NOS Utdanningsstatistikk. Resultatet av koblingen burde ligge til rette for analyser av utdanningsoverganger.

Tabell 4. Befolkningen etter utdanningsnivå. Tusen personer

	1986			1993		
Utdanning	MOSART	NOS	Avvik	MOSART	NOS	Avvik
I alt	3301	3307	-6	3436	3447	-9
Uoppgitt	84	96	-12	98	94	4
Grunnskole	1195	1200	-5	987	994	-7
Videregående skole	1556	1550	6	1725	1729	-4
Høyere utdanning	467	461	6	626	630	-4

Kilde: NOS Utdanningsstatistikk.

Tabell 5. Antall elever og studenter utover grunnskole. Tusen personer

År	MOSART	NOS	Avvik
1985	296	300	-4
1986	307	312	-5
1987	301	308	-7
1988	316	323	-7
1989	360	371	-9
1990	380	393	-13
1991	395	410	-15
1992	410	423	-13
1993	416	429	-13

Kilde: NOS Utdanningsstatistikk.

2.3. Trygdestatistikken

Rikstrygdeverket fører register over alle som mottar en ytelse fra folketrygden, og i prinsippet er disse opplysningene elektronisk tilgjengelig. Statistisk sentralbyrå bruker disse dataene til statistikkformål, se vedlegg C for filreferanser. Den ene kilden vi bruker er «statistikkgrunnlag 1 fra Rikstrygdeverket», som blant annet omfatter alle pensjonister i folketrygden. Videre bruker vi register over «attføringspengetilfeller», men disse dataene er bare tilgjengelig fra 1989 og framover. Kvaliteten på dataene kan være noe varierende, men er trolig det nærmeste vi kan komme gode forløpsdata for trygd.

Filene er tilgjengelig ikke lenge etter årsskiftet, og vil ikke legge noen skranker på valg av basisår sammenlignet med utdanningsstatistikken og pensjonspoengene.

Som nevnt kan kvaliteten variere noe, og vi har derfor rettet opp noen antakelige feil og imputert attføring på grunnlag av beregnet uførepoeng. Spesielt kan det virke som om noen personer faller ut av registrene i et enkelt år, for så å dukke opp som klient året etter, uten at det er grunn til å tro at dette er en reell bevegelse. Den første opprettingen består i å sette alle personer 70 år og eldre til å være alderspensjonister. Videre blir alle personer som er 66 år og eldre og allerede inne i trygdesystemet satt til å være alderspensjonister fra 67 år. Disse to opprettingene berører anslagsvis 3000 personer per år, av en totalbestand på 600 000 alderspensjonister.

Opplysninger om attføring inneholder starttidspunkt, og dette er brukt til å tilbakedatere attføringstiltak, spesielt til årene før 1989 hvor vi mangler slike data. Videre inneholder pensjonspoengrekkene opplysninger om beregnet uførepoeng (bup, se avsnitt 2.4), og det er grunn til å tro at disse er korrelert med tidspunktet for når uførheten inntrådte. Bup'ene inntreffer gjerne før uførepensjonen tilstås, og typisk i året før man kommer inn på attføring (der vi kan kontrollere dette). Videre virker det som om bup'ene fanger opp at personer er i trygdesystemet, der det foreligger en lang klientkarriere, men hvor personen mangler trygdedata i et enkeltstående inneklemt år. I noen tilfeller fortsetter bup'ene å løpe i ett år etter at uførepensjonen/attføring opplagt er avsluttet fordi vi kan observere en høy pensjonsgivende inntekt. Imidlertid er dette få tilfeller, og det er heller ikke lett å sette en inntektsgrense (også en del uførepensjonister vil ha inntekter over 200 000 kroner i året). Vi har derfor imputert at personer som oppebærer bup i et enkelt år (indikert av høyeste uføregrad, se avsnitt 2.4), og ikke mottar andre ytelser, er attføringsklienter fra og med det andre året med bup.

Ansvaret for attføring er delt mellom Rikstrygdeverket og Arbeidsmarkedsetaten, og det er noe uklart om våre data fanger opp de ordningene som faller under sistnevnte etat.

NB! I opprettingen har vi ikke tatt hensyn til om personen er bosatt, død eller utvandret, og spesielt berører dette attføring og alderspensjon.

Opprettingene gjør at datamaterialet må brukes med en viss forsiktighet. Spesielt mangler vi alle attføringstilfeller som ble avsluttet før 1989 og som ikke ledet fram til en uførepensjon. Videre er det ingen strømmer ut av alderspensjon. Materialet kan trolig brukes til analyse av overganger til attføring/uførhet for alle årene. Imidlertid kan det ikke brukes til analyser av overganger ut av attføring eller mellom attføring og uførhet for årene før 1989.

Tabell 6 og 7 gir en oversikt over trygdedataene i MOSART sammenlignet med statistikk fra Rikstrygdeverket (Trygdestatistisk årbok). Det største absolutte avviket knytter seg til antall alderspensjonister, men er samtidig det avviket som bekymrer minst. Avviket kan nesten i sin helhet forklares med alderspensjonister som oppebærer pensjon ut måneden de dør (desember) og alderspensjonister bosatt i utlandet, hvor også noen mangler ordinært fødselsnummer. Avviket i antall uførepensjonister skyldes at vi har omklassifisert uførepensjonister som fyller 67 år i desember til å være alderspensjonister, samt at vi har plassert «ventetid» inn under attføring. I tillegg kommer noen uførepensjonister som har utvandret eller som har dødd mot slutten av året, og derfor oppebærer pensjon ved årsskiftet. Antall attføringspengetilfeller er systematisk overvurdert i forhold til Rikstrygdeverkets statistikk, og differansen utgjøres langt på vei av ventetid for attføring og to tredeler av de med attføring imputert fra beregnede uførepoeng (i 1993). Det kan likevel være grunn til å tro at imputert attføring likevel gir en enklere og bedre beskrivelse av gangen inn i trygdesystemet enn de registrerte attføringstilfellene.

Et vesentlig poeng er at avvikene er stabile over tid, noe som tyder på at datamaterialet gir en relevant beskrivelse av strømmene i systemet. Dette understrekes av tabell 7, hvor avvikene i nye uførepensjonister er gjennomgående lite, spesielt for de siste årene.

Tabell 6. Pensjonister ved utgangen av året. Tusen personer

	1985	1986	1987	1988	1989	1990	1991	1992	1993
Attføringspengetilfeller									
- MOSART	38	43	42	53	62	67	73	74	70
- Rikstrygdeverket	32	34	33	39	48	53	60	66	64
- Avvik	6	9	9	14	14	14	7	8	6
Uførepensjonister									
- MOSART	185	191	202	211	222	228	232	229	225
- Rikstrygdeverket	188	194	207	217	228	234	239	236	232
- Avvik	-3	-3	-5	-6	-6	-6	-7	-7	-7
Etterlattepensjonister									
- MOSART	40	40	38	37	35	35	34	33	32
- Rikstrygdeverket	41	41	40	38	36	36	35	34	33
- Avvik	-1	-1	-2	-1	-1	-1	-1	-1	-1
Avtalefestet pensjon									
- MOSART	-	-	-	-	1	1	1	1	2
- Rikstrygdeverket	-	-	-	-	-	-	-	2	4
- Avvik	-	-	-	-	-	-	-	-1	-2
Alderspensjonister									
- MOSART	561	570	583	591	599	605	610	613	613
- Rikstrygdeverket	570	576	588	595	606	613	616	621	624
- Avvik	-9	-6	-5	-4	-7	-8	-6	-8	-11

Kilde: Rikstrygdeverket, Trygdestatistisk Årbok.

Tabell 7. Nye uførepensjonister. Tusen personer

År	MOSART	Rikstrygdeverket	Differanse
1986	24,3	25,7	-1,4
1987	33,2	35,2	-2,0
1988	30,5	31,5	-1,0
1989	34,0	31,5	2,5
1990	29,2	29,6	-0,4
1991	24,9	25,5	-0,6
1992	19,4	19,9	-0,5
1993	17,9	18,2	-0,3

Kilde: Rikstrygdeverket, Trygdestatistisk årbok.

2.4. Den rekonstruerte folketrygdbasen i Rikstrygdeverket

Rikstrygdeverket fører register over opptjeningen av pensjonsrettigheter i folketrygden, og disse og tilhørende data er tilgjengelig for statistikkformål. Vi har hentet data fra den «rekonstruerte folketrygdbasen», og hentet ut opplysninger knyttet til pensjonspoeng, beregnet uførepoeng, uføregrad, pensjonsgivende inntekt og yrkesstatus i form av selvstendig/ansatt. I tillegg finnes tidsserier for skattekommune og omsorgspoeng i denne databasen. Siden pensjonspoengene brukes til å beregne pensjonsytelser, er det grunn til å tro at dataene er av god kvalitet.

Pensjonspoengene beregnes på grunnlag av skatteligningen, og data for ett år vil dermed først foreligge tidligst høsten etter. Man bør også forutsette at det går noen måneder før dataene er tilgjengelige i registerform. I denne versjonen fikk vi med pensjonspoeng for 1994 for et uttak som skjedde i september/oktober 1995, hvor vi antok at vi kun ville få med data til og med 1993. Alternativt kan det siste pensjonspoenget hentes fra skattestatistikken (avsnitt 2.5), men vil da først foreligge ved årsskiftet året etter. Valg av basisår blir dermed som for utdanning, år 't-1' hvis aktualitet er viktig, år 't-2' hvis kvalitet på kjennetegnene er viktig (jamfør diskusjonen nedenfor). Spesielt hvis man skal bruke beregnede uførepoeng til å imputere attføring (jamfør avsnitt 2.3), så er det ekstra året spesielt viktig.

Tidsseriene for pensjonspoeng er kun tilgjengelig i Rikstrygdeverket, men dette uttaket er en relativt grei affære. Jobben skal være innrapportert til Seksjon for helse- og sosialstatistikk innen november året før uttaket skal skje, slik at denne seksjonen kan få videresendt (en samlet) bestilling fra Statistisk sentralbyrå til Rikstrygdeverket. Leveringstid fra Rikstrygdeverket vil avhenge av arbeidsbelastningen der, men har kun tatt 1-2 måneder de to gangene vi har gjort dette. Se vedleggene for detaljer.

Pensjonsgivende inntekt er summen av lønnsinntekt, næringsinntekt og ytelser som ledighetstrygd, sykepenger og betalt svangerskapspermisjon, og kan derfor brukes som et kjennetegn for yrkesdeltaking. En svakhet er at strømmer inn i og ut av yrkesdeltaking målt på denne måten er avhengig av data for inntektsåret før og etter. For eksempel vil man oppebære inntekt det året man slutter å arbeide, og det er kun ved å se at neste års inntekt er null at man på en fullgod måte kan identifisere at personen har sluttet å jobbe dette året. Det er derfor svært hensiktsmessig å ha inntektsdata for året etter basisåret for datagrunnlaget. Alternativt må dette simuleres i modellen under innlesingen av det utvalget MOSART starter opp med.

En annen svakhet ved pensjonsgivende inntekt er at denne kun er definert for personer i alderen 17 til 69 år, mens yrkesdeltaking normalt angis for aldersgruppen 16 til 74 år. I tillegg kan bevegelser i inntektsstatus ikke defineres tilfredsstillende for 17- og 69-åringer, siden vi ikke kjenner arbeidsinntekten for disse personene for henholdsvis året før og året etter.

Pensjonspoengene i datagrunnlaget er beregnet utfra pensjonsgivende inntekt, slik at eventuelle tillegg for beregnet uførepoeng og omsorgspoeng må gjøres separat i simuleringsmodellen.

Tabell 8. Pensjonsgivende inntekt. Tusen personer, 1993-kroner

	Personer med pensjonsgivende inntekt		Gjennomsnittlig inntekt, 1993-kroner	
År	MOSART	Avvik fra Rikstrygdeverket	MOSART	Avvik fra Rikstrygdeverket
1985	2270	-36	148400	300
1986	2320	-49	153300	1400
1987	2364	-18	156600	0
1988	2379	-29	155500	100
1989	2372	-14	152300	0
1990	2371	-11	155500	400
1991	2387	-9	157100	400
1992	2381	-13	160600	600
1993	2397	-14	161700	400

Kilde: Rikstrygdeverket, Trygdestatistisk årbok.

Tabell 8 gir en oversikt over pensjonsgivende inntekt i datagrunnlaget sammenlignet med statistikk fra Rikstrygdeverket (Trygdestatistisk årbok). Når det gjelder gjennomsnittlig inntekt er avvikene

små. Avvikene i antall personer med pensjonsgivende inntekt er også relativt små, men øker noe bakover i tid utover det som ligger i grensene for utvalgsusikkerhet. En mulig forklaring kan være at inn- og utvandringshistoriene i MOSART blir svake når man går bakover i tid, noe som vil gi at noen personer er registrert bosatt i utlandet når de ikke gjør det. Avvikene er likevel såpass små og stabile at dette ikke burde være noen grunn til å bruke inntektsopplysningene til analyser av bruttostrømmer på arbeidsmarkedet og av nivå på arbeidsinntekt.

2.5. Skattestatistikken

Planlagte utvidelser av MOSART går i retning av å bygge ut modellen til å dekke et fullstendig inntekts- og formuesregnskap for personer. Det vil da være nødvendig å kjenne fordelingen av formue i basisåret, og i mangel av gode data må dette simuleres inn. Denne simuleringen vil trolig gå lettere og bedre hvis man har fordelingen av noen formueskomponenter i tillegg til kjønn, alder, sosio-økonomisk status, inntekt, bostedskommune med videre. Deler av selvangivelsene blir registrert elektronisk, og vil dermed være tilgjengelig for hele befolkningen. Vi har tatt med opplysninger om nettoformue ved statsskatt og ligningsverdi av bolig for å kunne lette simuleringen av fordelingen av formue i basisåret.

Det finnes ingen lett tilgjengelige statistikker å sammenligne koblingsresultatet med, blant annet fordi regnskapsenheten i formuesstatistikken gjerne er husholdning. Imidlertid virker det som om gjennomsnittlig nettoformue vi har lagt inn stemmer med formuesbegrepet i Inntekts- og formuesundersøkelsen (IF) for 1993. Bruk av dataene forutsetter imidlertid langt mer inngående sammenligninger med IF.

2.6. Statens pensjonskasse

En annen utvidelse av MOSART går i retning av å legge inn tjenestepensjoner, blant annet fordi disse ordningene inngår i skattepolitikken og vil utgjøre en vesentlig del av inntektene for pensjonister. Spesielt gjelder dette statspensjonene, hvor eventuelle reduksjoner i folketrygdens ytelser delvis vil bli veltet over på en annen statlig konto. Vi arbeider derfor med å legge inn opplysninger fra Statens pensjonskasse, også fordi data herfra er relativt lett tilgjengelig. Da dette notatet gikk i trykken, var ikke opplysninger herfra ferdig tilrettelagt. Det er derfor vanskelig å trekke noen konklusjoner om kvaliteten på dataene. Opplysningene som er tatt med går på hvem som er ansatt i staten og deres tjenestetid og hvem som mottar en statspensjon og (brutto)ytelsen de får.

Problemer med dataene knytter seg til at mange tjenesteforhold ikke er registrert elektronisk, og dermed i praksis utilgjengelig. Dette knytter seg spesielt til en periode på 1980-tallet hvor opplysninger om statspensjoner skulle arkiveres lokalt hos den enkelte arbeidsgiver. Opplysninger om tjenestetid må derfor beregnes utfra første startår og siste sluttår i staten, uten å vite noe om det er opphold i ansettelsesforholdet mellom disse to årene. Videre vil en del personer som har sluttet å jobbe i staten ikke bli registrert. Bruken av dataene vil derfor forutsette relativt omfattende sammenligninger med annen statistikk sammen med informasjon fra Statens pensjonskasse.

2.7. Arbeidskraftundersøkelsene

Statistisk sentralbyrås Arbeidskraftundersøkelser (AKU) er basert på intervjuer av et tilfeldig utvalg av befolkningen i en utvalgt undersøkelsesuke hver måned. I offentlig planlegging vil det være begreper og tall for yrkesdeltaking fra AKU som er relevant. Sett i forhold til bruken i MOSART er dette problematisk av flere grunner. AKU er lite egnet til å beregne yrkesdeltaking på individnivå når tidsenheten er kalenderåret, se Fredriksen og Spurkland (1993) for hvordan dette er løst i MOSART.

Andre problemer knytter seg til at AKU registrerer en del bakgrunnskjennetegn på andre måter enn i registerstatistikken, og dette kan gi vesentlige konsistensproblemer. Dette knytter seg til at AKU i liten grad fanger opp at en del personer kombinerer arbeid og utdanning eller trygd. Videre vil AKU

rapportere status på intervjutidspunktet, mens vi ofte er interessert i status ved utgangen av året. Dette slår sterkt ut på antallet mødre med barn under ett år. Vi har derfor sett det nødvendig å koble AKU mot de samme registrene som vi bruker i MOSART, for å kunne beregne yrkesdeltakingen for ulike befolkningsgrupper på en tilfredsstillende måte. Vedlegg A.8 gir en beskrivelse av den koblede filen, mens Statistisk sentralbyrå (1995B) gir en oversikt over AKU.

Tabell 9-11 viser hvordan yrkesdeltakingen varierer etter bakgrunnsvariabler avhengig om disse er hentet fra AKU selv eller fra registre. Yrkesprosentene i tabellene angir hvor stor andel som er i arbeidsstyrken (sysselsatt eller arbeidssøker) ifølge AKU. Avvikene i tabell 9 skyldes i all hovedsak at personer som kombinerer pensjon og arbeid, oppfatter (deltids)jobben som hovedaktivitet, og dermed ikke blir spurt om de også mottar en pensjon. Yrkesdeltaking blant pensjonistene ligger anslagsvis 4-5 ganger høyere enn det AKU rapporterer, og forskjellen blir større hvis vi trekker inn arbeidstid. I gjennomsnitt ligger den faktiske yrkesprosenten i denne gruppen på rundt 20 prosent, mot et anslag på 4 prosent i AKU.

Avvikene i tabell 10 skyldes trolig studenter og elever som i perioder jobber heltid, for eksempel feriene, ikke blir spurt om andre aktiviteter som skolegang. Også her blir forskjellen påtakelig større hvis vi ser på arbeidstid. I tillegg til trygdestatus og skolegang har vi trukket inn barnetall og utdanningsnivå fra register. Avvikene her skyldes ikke «svakheter» i AKU, snarere tvert imot. Imidlertid gjør disse dataene det mulig å beregne yrkesprosenter med forklaringsvariabler som er mer konsistente med MOSART.

Tabell 11 oppsummerer problemene med at pensjonsgivende inntekt gjennom et helt år ofte avviker fra yrkesdeltaking i en tilfeldig uke samme år. En god del av problemene kan knyttes til personer som begynner eller slutter å ha pensjonsgivende inntekt et enkelt år. Videre kan problemene knyttes til grupper som kan tenkes å bare jobbe deler av året, typisk studenter og elever og kanskje pensjonister. En del eldre registrerte arbeidsløse vil også motta dagpenger uten at de oppfatter seg som arbeidssøkere når de blir spurt av AKU. Disse avvikene kan rettes opp ved å beregne sannsynligheten for at en person som har pensjonsgivende inntekt også skal være yrkesaktiv i AKU avhengig av relevante bakgrunnskjennetegn som kjønn, alder, skolegang, pensjon og om personen har begynt/sluttet samme år.

Et mer substansielt problem knytter seg til personer som ikke har inntekt, men allikevel oppgir at de er yrkesaktive. En del av disse vil være personer med inntekter som vi ser bort fra i MOSART (for små, for ustabile), samt en del arbeidssøkere uten rett til ledighetstrygd. En supplerende forklaring kan være selvstendig næringsdrivende som ikke tar ut lønn. En annen forklaring kan være at intervjuobjektet oppgir som hovedaktivitet det som forventes av omgivelsene, typisk at man er i arbeid.

Tabell 9. Yrkesdeltaking og trygd i 1993. Tusen personer, prosent

			Pensjonist i AKU			
	Alle		Nei		Ja	
Trygdestatus i register	Antall	Yrkes-prosent	Antall	Yrkes-prosent	Antall	Yrkes-prosent
Alle	3125	68.2	2602	81.1	522	3.9
Ikke pensjonert	2495	81.3	2459	82.4	36	8.9
Nye pensjonister	53	43.5	30	72.2	23	5.6
Etterlatte	27	63.6	23	72.8	4	6.0
Alderspensjonister	286	5.8	18	43.4	268	3.3
Delvis uføre	48	50.7	27	80.4	21	12.6
Helt uføre	175	6.2	17	44.2	158	2.1
Attføring	41	27.2	28	36.7	13	6.7

Kilde: Arbeidskraftundersøkelsene 1993.

Tabell 10. Yrkesdeltaking og skolegang, 1993. Tusen personer, prosent

Utdannings-aktiviteter i register	Alle		Student/elev i AKU: Nei		Ja	
	Antall	Yrkes-prosent	Antall	Yrkes-prosent	Antall	Yrkes-prosent
Alle	3125	68.2	2752	73.8	372	26.6
Ingen aktiviteter	2544	72.2	2525	72.6	18	19.7
Er elev/student	338	38.5	65	90.1	273	26.3
Ble elev/student	90	66.1	60	85.1	30	27.6
Kandidat	100	68.9	64	90.1	36	31.2
Avbrøt utdanning	53	69.1	39	85.3	15	27.4

Kilde: Arbeidskraftundersøkelsene 1993.

Tabell 11. Yrkesdeltaking og pensjonsgivende inntekt, 1993. Personer 18-68 år. Tusen personer, prosent

Sosial status	Alle		Yrkesdeltaking gitt ved pensjonsgivende inntekt: I arbeid		Begynner		Slutter		Utenfor	
	Antall	Yrkes-prosent	Antall	Yrkes-prosent	Antall	Yrkes-prosent	Antall	Yrkes-prosent	Antall	Yrkes-prosent
Alle	2787	75	2221	89	64	51	68	32	435	12
Under utdanning	456	58	358	66	32	43	10	32	56	21
Pensjonister	401	23	114	63	7	26	34	18	245	4
Andre	1931	90	1749	96	25	69	23	51	134	23

Kilde: Arbeidskraftundersøkelsene 1993.

3. Trekking av utvalg

I prinsippet kunne datagrunnlaget for MOSART bygd på hele befolkningen, noe vi har avstått fra. Den viktigste grunnen til å bruke utvalg er at edb-utstyret vi bruker (foreløpig) ikke ville kunne håndtert datamengden for totalpopulasjonen på en hensiktsmessig måte. Overgang til ny teknologisk plattform (unix) kan endre denne konklusjonen. En klar ulempe med utvalgsundersøkelser er at datagrunnlaget får færre observasjoner. Med anslagsvis 12 prosent av befolkningen i utvalget utgjør dette ikke noe vesentlig problem i forhold til andre svakheter ved modellen.

En klar svakhet med å bruke et utvalg når relasjoner mellom personer skal opprettholdes, er at utvalget ikke kan oppdateres ved å *føye* til nye data. Grunnen til dette er at noen personer i utvalget vil ha giftet seg med personer utenfor utvalget. Dermed må man også trekke et nytt utvalg og hente inn alle opplysningene på nytt. Hensyn til personvernet taler også mot et totalregister lagret slik at nye opplysninger kan legges til ved hjelp av fødselsnummer. Det er uklart hvor stor vekt dette bør tillegges, jamfør koblingene i Statistisk sentralbyrå (1995A).

3.1. Krav til utvalgets sammensetning

Utvalget bør så langt som mulig tilfredsstille følgende tre krav:

(1) Tilfeldig og uveid
(2) Representativt
(3) Opprettholde relasjoner mellom personer

At utvalget er tilfeldig og uveid (1) betyr at samtlige personer i populasjonen à priori hadde samme sannsynlighet for å komme med i utvalget. Etter simuleringen kan utvalgstallene 'blåses opp' til

totaltall ved å multiplisere hvert individ med den *samme* faktoren, som blir den inverse av utvalgsprosenten.

Mikrosimulering sikrer konsistens i personregnskapet, for eksempel at det er like mange menn som kvinner som bor sammen som ektepar, foreldre eller samboere. I MOSART kan hvert individ tilhøre forskjellige typer enheter, og hvert individ kan bytte enhet (giftermål, skilsmisse, flytting). Med ulik oppblåsingsfaktor for ulike personer, vil vi uvilkårlig få ulike oppblåsingsfaktorer for ulike personer i samme enhet. Dermed vil konsistensegenskapene være tapt. Kravet om uveidhet er derfor ufravikelig.

At utvalget er mest mulig representativt vil si at oppblåste tall fra utvalget forventningsmessig ligner mest mulig på tallene for totalbefolkningen. Kravet om representativitet (2) kommer nødvendigvis ikke i konflikt med kravet om tilfeldighet eller uveidhet. Utvalgets representativitet kan forbedres ved å forhåndsstratifisere, det vil si at man på forhånd bestemmer seg for å trekke et bestemt antall personer fra gitte grupper, dog slik at hver person fortsatt har samme sannsynlighet for å komme med i utvalget. Med kravet om uveidhet vil antallet personer som trekkes ut fra hver gruppe være gitt ved utvalgsprosenten multiplisert med antallet personer i gruppen i totalpopulasjonen.

Ønsker man å forhåndsstratifisere etter flere kjennetegn, er dette teknisk enklest å gjennomføre for den simultane fordelingen over kjennetegnene. Det vi si at hver kombinasjon av kjennetegn utgjør en gruppe, og at det da bare er denne ene skranken man tar hensyn til i trekkingen. Imidlertid får man fort det problemet at antallet personer i utvalget fra hver gruppe blir mindre enn én hvis antallet kjennetegn/inndelinger blir stort. I tillegg blir avrundingsfeilene påtakelige. Med stratifisering etter den simultane fordelingen er det derfor begrenset hvor mange kjennetegn man kan trekke inn, og man må også prioritere mellom kjennetegnene[2].

I forhåndsstratifiseringen har vi lagt vekt på kjennetegn som har stor forklaringskraft og som er stabile (eller forutsigbare) over tid. Dette gjelder spesielt kjennetegnene kjønn og alder. Videre har vi trukket inn forventet gjenstående antall fødsler for kvinner, da dette har vesentlig innflytelse på befolkningsutviklingen framover. Viktige variabler for (rest)fruktbarheten er alder, antall barn og i noe mindre grad yngste barns alder.

Det er viktig å opprettholde relasjoner mellom personer i utvalget (3), for eksempel mellom ektefeller. Dette kravet kommer fort i konflikt med kravet om tilfeldighet (1) og i enda større grad representativitet (2). Problemet bunner i at flere typer relasjoner fort gir meget store klynger av personer som må trekkes ut sammen. Det blir dermed færre trekkeenheter, og det blir flere kjennetegn for hver enhet som blir viktige i stratifiseringen, for eksempel begge ektefellers alder. De eneste relasjonene som er håndterbare er slike som definerer lukkede celler. Man kan for eksempel velge enten ektepar/sambopar, husholdninger, mor/barn eller far/barn[3].

I denne versjonen av datagrunnlaget har vi brukt par som bor sammen som trekkeenhet, for øvrig har vi trukket individer. Første gruppe av par er definert ved at vi har koblet ektefeller og sjekket om de

---

[2] Det er for eksempel mulig å stratifisere slik at man (kun) tar hensyn til de marginale fordelingene over kjennetegnene. Eksempelvis kan utvalget få riktig antall gifte menn og gifte kvinner i hver aldersgruppe, uten at antallet ektepar i ulike alderskombinasjoner behøver å bli det. Problemet med marginal stratifisering er at man må ta hensyn til flere skranker samtidig, men det skal være mulig å gjennomføre dette på tilnærmet nivå. Fordelen er at man kan trekke inn flere kjennetegn i stratifiseringen, siden avrundingsfeilene blir eliminert. Vi har ikke forsøkt dette her.

[3] For eksempel vil relasjonen foreldre-barn kunne gi grupper av personer som vil omfatte store deler av den norske befolkning. Hvis vi starter vilkårlig med nestor i en familie må vi ta med alle etterkommere denne har. For samtlige av disse personene, må vi ta med alle medforeldre disse har til sine barn. I den grad noen av disse medforeldrene har egne foreldre/besteforeldre i live, må vi ta med disse og alle deres etterkommmere. På denne måten kan man nøste opp store befolkningsgrupper. Satt på spissen er alle i slekt med alle.

har samme adresse. For de som ikke har ektefelle, har vi koblet foreldre med felles yngste barn og sjekket om disse har samme adresse. Par som bor sammen, men som ikke er gift (med hverandre) eller har felles barn har vi ikke fått med.

3.2. Trekkeprosedyre

Basis for trekkingen av utvalget er situasjonuttaket fra Personregisteret per 1.1.1994 og en antakelse om at denne omfatter alle personer bosatt i Norge. Samtlige personer (fødselsnumre) fra situasjonsuttaket er representert på én og bare én record på grunnlagsfilene for trekkingen. Dermed kan utvalget stratifiseres ved fortløpende å telle opp antall personer på grunnlagsfilene. Hver record kan imidlertid gjerne inneholde flere personer, for eksempel et ektepar. Trekkingen bør da foregå på grupper av records som inneholder samme antall personer, slik at man får kontroll med utvalgsstørrelsen.

En del opplysninger omkring personen utgjør grunnlaget for trekkingen av utvalget, og disse informasjonene hentes derfor inn allerede på dette stadiet. I tillegg til kjønn og fødeår, omfatter dette opplysninger omkring ekteskap, ektefelle, foreldre og 'bosted'.

Trekkingen skjer først ved å sortere grunnlagsfilene med hensyn på stratifiseringsvariablene i prioritert rekkefølge. Med to variabler som henholdsvis antar verdier (A, B, ...) og (1, 2, ...) blir rekkefølgen av observasjonene:

A1, A1, ... A1, A2, A2, ... A9, B1, ... B9, C1 ...

Blant de 'n' første personene, hvor 'n' er den inverse av utvalgsprosenten, trekker vi en tilfeldig person. Deretter tar vi med hver 'n'-te person i utvalget etter den først uttrukne. Svakheten oppstår når et intervall på 'n' personer inneholder personer med ulike kjennetegn. Det blir dermed tilfeldig hvilket kjennetegn personen i utvalget fra dette intervallet vil få. Jo flere variabler man stratifiserer etter og jo mindre utvalgsstørrelsen er, jo mer påtrengende blir disse feilene. Feilene vil også bli større for de variablene som kommer lenger ut i sorteringen, fordi antallet ganger kjennetegnet må bytte verdi vil øke (A→B bytter verdi en gang, 1→2 en gang for hver bokstav A, B, C ....).

En feil som blir spesielt påtrengende oppstår der sorteringen bytter verdi for en overordnet sorteringsvariabel (her A→B), fordi spranget i verdi kan bli stort for de underordnete variablene (her 1→9). Dette er avhjulpet ved å sortere vekselsvis stigende og synkende for underordnete variabler, noe som også reduserer antallet ganger variabelen må bytte verdi:

A1, A1, ... A1, A2, ... A9, B9, ... B1, C1, ...

Hovedutvalget består av 12 prosent av befolkningen fordelt på 12 underutvalg à 1 prosent. Hovedutvalget henter ut hver 8 1/3 person[4] fra grunnlagsfilen, og hver tolvte av disse fordeles på hvert sitt underutvalg. Det vil si at hvert underutvalg består av hver hundrede person fra grunnlagsfilen. Skal man kombinere underutvalg lønner det seg å ta underutvalg med maksimal avstand. For eksempel bør ett 4-prosent utvalg bestå av underutvalgene 1/4/7/10, 2/5/8/11 eller 3/6/9/12. Ett slikt kombinert utvalg vil bestå av hver 25.person, og vil ha de samme stratifiseringsegenskapene som om det var trukket som et 4-prosent utvalg. Med 12 underutvalg (i motsetning til 10), er det mulig å kombinere utvalgsstørrelser som 2, 3, 4 og 6 prosent med denne egenskapen.

I tillegg til hovedutvalget er det trukket to små testutvalg som består av henholdsvis 1 og 0,1 promille av befolkningen. Disse er hentet ut som henholdsvis hver 1000 og 10000 person, dog slik at ingen av disse er med i hovedutvalget.

---

[4] Det vil si at vekselsvis hver 8.nde, hver 8.nde og hver 9.nde person kommer med.

Stratifiseringen har skjedd for seks ulike grupper etter følgende kjennetegn:

(i) Samboende ektepar, subsidiært samboende foreldrepar, bosatt i Norge, er stratifisert etter kvinnens alder, mannens alder, om de er gift, antall barn (0,1,2+) og yngste barns alder.
(ii) Øvrige menn bosatt i Norge er stratifisert etter egen alder, antall barn (0,1,2+), ekteskapelig status (ugift, enkemann, andre) og yngste barns alder. Ektefeller/medforeldre som ikke er bosatt i Norge er inkludert i utvalget.
(iii) Øvrige kvinner bosatt i Norge er stratifisert etter egen alder, antall barn (0, 1, 2+), ekteskapelig status (ugift/enke/andre) og yngste barns alder. Ektefeller/medforeldre som ikke er bosatt i Norge er inkludert i utvalget.
(iv) Par hvor ingen av partnerne er bosatt i Norge, er stratifisert etter reg.status (død, utvandret), årstall for endring i reg.status og fødselsår, for den av partnerne som sist hadde endring i reg.status.
(v) Øvrige menn som ikke er bosatt i Norge er stratifisert etter reg.status, årstall for endring i reg.status og fødselsår.
(vi) Øvrige kvinner som ikke er bosatt i Norge er stratifisert etter reg.status, årstall for endring i reg.status og fødselsår.

Situasjonsuttak bakover til 1.1.1986 er koblet til utvalget. Avdøde ektefeller av enker og enkemenn i utvalget er lagt til utvalget. Tilsvarende blir døde personer fjernet fra utvalget, hvor den gjenlevende ektefellen fortsatt bor i Norge og er enke/enkemann og ikke er med i utvalget. Samtidig hentes inn opplysninger for ekteskapelig status og reg.status (bosatt i Norge, død, utvandret). I tillegg til å sikre oss forløpsdata fra personregisteret (flytting, ekteskap), sparer det oss for en runde med Folkeregisteret/Statens datasentral for å hente ut fødselsnumre for avdøde ektefeller av enker og enkemenn.

4. Organisering av data

Fødselsnumrene for samtlige personer i utvalget omtalt i kapittel 3 er lagt ut på en egen liste, og vi har brukt denne listen til å hente ut opplysninger fra de ulike registrene omtalt i kapittel 2. Etter uttak er opplysningene først anonymisert, se avsnitt 4.2, og deretter koblet sammen. Resultatet er deretter fordelt på ulike filer i to dimensjoner av praktiske årsaker. For det første er opplysningene fordelt på sine respektive underutvalg, slik at man kan forholde seg til mindre datamengder om gangen, og samtidig ha fordelene av at dataene har tilnærmet samme stratifiseringsegenskaper og at par befinner seg på samme fil. Videre er opplysningene fordelt på ulike typer filer, da det kan være mer hensiktsmessig å fordele visse opplysninger på flere records av ulik type, snarere enn en lang record med variabel lengde. Spesielt gjelder dette opplysninger om hvert enkelt barn, hvor antallet barn kan variere tildels betydelig fra person til person. Videre gjelder dette forløpsdata, i vår sammenheng som statusopplysninger for hvert enkelt år. Dataene er organisert rundt følgende filer (referanser til vedlegg i parentes):

(i) Utvalgsfil med en record for alle personer og alle opplysninger i basisåret (A.1).
(ii) Mødrefil med en record for hvert barn kvinnen er juridisk mor til (A.2).
(iii) Fedrefil med en record for hvert barn mannen er juridisk far til (A.2).
(iv) Forløpsfil med en record for hver person for hvert år 1985-1993 (A.3).
(v) Pensjonspoengfil med en record for hver person for hvert år 1967-1993 (A.4).
(vi) Bostedskommune med en record for hver person med kjent bosted i Norge (A.5).
(vii) Ikke aggregerte utdanningsdata med en record for hver person for hvert år 1985-1993 (A.6).
(iix) Ikke aggregerte attføringsdata med en record for hver klient for hvert år 1989-1993 (A.7).

Filene (i)-(v) er splittet opp sine respektive underutvalg, mens filene (vi)-(iix) ikke er splittet opp på sine respektive underutvalg. Opplysningene i de sistnevnte filene inngår ikke i nåværende versjon av

MOSART, og behøver derfor ikke ta plass på de mest tilgjengelige filene. Dataene er minst mulig endret i forhold til rådataene, med unntak av de opprettingene som er omtalt i kapittel 2.

4.1. Kvalitetskontroll

Resultatet av koblingene er sjekket fortløpende for inkonsistenser og feil, samt avvik fra offisiell statistikk. Noen av disse sammenligningene er gjengitt i kapittel 2, og sjekk av pekere er omtalt i avsnitt 4.3. Ved bruk av filene til forløpsanalyser tilrådes imidlertid at det tas ut aggregert statistikk for de begivenhetene og variablene som inngår i analysen, og at dette så langt som mulig sjekkes mot tilgjengelig statistikk. Spesielt opplysninger for år bakover i tid kan være skjevt representert, blant annet fordi noen personer får byttet fødselsnumre, noe vi ikke fanger opp.

4.2. Anonymisering

Alle fødselsnumre er fjernet fra filene før koblingene, og i stedet erstattet med et vilkårlig nummer som én-entydig identifiserer hver person i utvalget. Det skal da i prinsippet være umulig å tilbakeføre opplysninger fra datagrunnlaget eller simuleringen til konkrete personer. Dette understreker at de personene vi simulerer er hypotetiske, vi har bare lånt et begrenset sett av faktiske kjennetegn fra et utvalg av befolkningen. Samtidig har det anonymiserte identifikasjonsnummeret gjort det mulig å opprettholde pekere mellom ulike personer i utvalget, så som samboende par, ektefeller og foreldre-barn. Identifikasjonsnummeret gjør det også mulig å fordele opplysninger om den enkelte person på ulike filer og ulike records.

4.3. Innlesing av utgangspopulasjonen

Innlesing av utgangspopulasjonen bør ta utgangspunkt i utvalgsfila (vedlegg A.1), da denne raskest gir tilgang på hvilke personer som er bosatt i basisåret og opplysninger om disse personene i samme år. I tillegg må avdøde ektefeller leses inn, da pensjonsrettigheter kan gå i arv. Noen opplysninger må hentes inn fra de andre filene, men med utgangspunkt i utvalgsfila kan dette begrenses til de personer dette gjelder. Blant annet må pensjonspoengene hentes fra pensjonspoengfila (vedlegg A.4), og i den grad simuleringen bygger på kovariater som beskriver forhistorien til et individ, må disse hentes fra forløpsfila (vedlegg A.3) eller lignende.

I en del tilfeller skal konkrete personer kobles sammen i utvalget, blant annet gjelder dette par og enker/enkemenn mot avdøde ektefelle. Dette gjøres lettest ved å vente til den personen som har høyest identifikasjonsnummer er lest inn, og deretter koble ved å søke etter partneren i listestrukturen i simuleringsmodellen. Det er sjekket at bosatte samboende par finner hverandre ved denne metoden.

I en del tilfeller vil gifte personer mangle opplysninger om hvem ektefellen er eller være bosatt alene. Videre vil en del samboende par bestå av to personer som før var gift med hverandre, men som nå er separert eller skilt. I tillegg vil en del «ektefeller» ha forskjellig ekteskapelig status, for eksempel at den ene er gift, mens den andre er separert. De fleste av disse tilfellene kan tilbakeføres til manglende innrapporteringer til Personregisteret eller at ektefellen ikke har formalisert et samlivsbrudd med en separasjon. Dette er ikke rettet opp i datagrunnlaget, og må rettes opp i initieringen av simuleringsmodellen.

Avdøde ektefelle kan være ukjent eller mangle i utvalget, og man kan da i noen tilfeller bruke opplysninger om medforelder til yngste barn. Vanligvis er dette avdøde ektefelle, men i noen tilfeller dreier dette seg om nåværende samboer. Utover dette må koblingen skje ved å trekke en tilfeldig død person som ligner de opplysninger som finnes (motsatt kjønn, dødsår, gift, alder mv). Ett visst antall døde personer utover avdøde ektefeller bør/må derfor leses inn. For nye enker/enkemenn etter 1985 er de fleste avdøde ektefeller kjent, mens medforelder for enker/enkemenn fra før 1985 er kjent i omlag halvparten av tilfellene.

En del relasjoner må opprettes syntetisk, og blant annet gjelder dette foreldre-barn. Tar man utgangspunkt i foreldrene, må koblingen skje med utgangspunkt i mødre- og fedrefila (vedlegg A.2), hvor det finnes opplysninger om hvert av barna og om de formelt sett er bosatt hos foreldrene. Antallet barn på hvert alderstrinn i utvalget er tilnærmet likt med antallet barn de voksne har.

4.4. Forløpsanalyse

Forløpsfila (vedlegg A.3) er bygd opp med en record for hver person for hvert år med statusopplysninger for (utgangen av) dette året, samt noen endringsindikatorer. Forløpsanalyse kan hensiktsmessig bygge på å sammenligne to påfølgende records fra denne fila, for å fastslå hvilke personer som inngår i risikomengden, eventuelle begivenheter og for å hente ut riktige kovariater. Spesielt i de tilfeller man bruker rekursiv simulering med kalenderåret som enhet, gir opplegget mulighet for å hente inn kovariat fra rett år. For eksempel bør dødelighet baseres på kjennetegn fra året før, mens arbeidsinntekt i nåværende versjon av MOSART henter alle kovariater fra samme år.

Fila inneholder personer som ikke er bosatt, og for en del av disse vil variablene være av vekslende kvalitet. I en del tilfeller er data imputert på grunnlag av data for tidligere år, og det betyr for eksempel at personer som har dødd i 1985 står med svært få opplysninger på fila. Ellers vil kapittel 2 inneholde en vurdering av hvilke begivenheter som kan analyseres med utgangspunkt i dette datamaterialet.

Referanser

Andreassen, Leif, Truls Andreassen, Dennis Fredriksen, Gina Spurkland og Yngve Vogt (1993): *Framskriving av arbeidsstyrke og utdanning*, Rapporter 93/6, Statistisk sentralbyrå.

Fredriksen, Dennis (1992): Datagrunnlaget for trygdemodellen MOSART-T, Interne Notater 92/7, Statistisk sentralbyrå.

Fredriksen, Dennis (1993): Dokumentasjon av input til MOSART, Notater 93/42, Statistisk sentralbyrå.

Fredriksen, Dennis og Gina Spurkland (1993): *Framskriving av alders- og uførepensjon ved hjelp av mikrosimuleringsmodellen MOSART*, Rapporter 93/7, Statistisk sentralbyrå.

Statistisk sentralbyrå (1994): Håndbok i datasikkerhet og fysisk sikring, Statistisk sentralbyrås håndbøker.

Statistisk sentralbyrå (1995A): Sluttrapport fra forprosjektet til prosjektet «Trygd-fobhistorie», Notater 95/22.

Statistisk sentralbyrå (1995B): *Arbeidsmarkedsstatistikk 1994, Hefte I Hovedtall*, Norges offentlige statistikk C 261.

Vassenden, Elisabetta (1993): Befolkningens høyeste utdanning, revidert dokumentasjon, Notater 93/15, Statistisk sentralbyrå.

Vedlegg A. Filbeskrivelser

Samtlige filer er direkte tilgjengelige fra arbeidsstasjonen 'johansen', og inntil videre ligger de på harddisken /ssb/johansen/d1, men kan senere bli overført til et masselager. De fleste filene ligger i ascii-format med en linje for hver record. Alle kolonner vil inneholde ett og bare ett 'tall'. Spesielt vil uoppgitte tall være angitt med eget siffer, ofte null, jamfør vedlegg B. Med unntak av ikke aggregerte utdannings- og attføringsdata, vedlegg A.6 og A.7, vil alle kolonner i alle filer være separerert med minst en blank. De mest sentrale filene er splittet opp på sine respektive underutvalg, referert nedenfor som sample_*ii*, hvor *ii* angir utvalgsnummer 01-12, samt testutvalgene 98 (stort) og 99 (lite), hvor 98 og 99 også angir utvalgsnummeret. Fila *sample_size* under hvert delutvalg inneholder en linje med et desimaltall som angir utvalgsstørrelsen for dette delutvalget. Antallet records refererer til anslag for hvert underutvalg, hvor testutvalgene vil ha henholdsvis 10 og 1 prosent av denne størrelsen igjen. Variabler skrevet i kursiv er mer utførlig beskrevet i vedlegg B.

A.1. Utvalgsfil

Datatype:	Ascii-format
Filident:	/ssb/johansen/d1/mosart/input24/population/sample_*ii*/persons
Enhet:	En record per person, status for 1993
Antall:	59400
Sortering	Identifikasjonsnummer

Felt	***Innhold***
1 - 7	*Identifikasjonsnummer*
9 - 10	*Utvalgsnummer*
12	*Kjønn*
14 - 17	Fødselsår
19	*Reg.status*
21 - 24	Årstall for siste endring i reg.status
26	*Ekteskapelig status*
28 - 31	Årstall for siste endring i ekteskapelig status
33 - 34	Ektefelles *utvalgsnummer*
36 - 42	*Ektefelles identifikasjonsnummer*
44 - 47	Ektefelles fødselsår
49	1: Paret har samme adresse, 0: Ellers
51	Hvem peker partner til, 1: ektefelle, 2: medforelder, 0: Ellers NB! Partner er nødvendigvis ikke med i utvalget (berører ikke-samboende «par»)
53 - 54	Partners *utvalgsnummer* (berører ikke-samboende «par»)
56 - 62	*Partners identifikasjonsnummer*
64 - 67	Partners fødselsår
69 - 70	*Antall barn ifølge personregisteret*
72 - 73	*Antall barn beregnet ved å koble barn-foreldre*
75 - 76	*Yngste barns alder*
78 - 79	Mors *utvalgsnummer*
81 - 87	Mors *identifikasjonsnummer*
89 - 92	Mors fødselsår
94	Mors *reg.status*
96	1: Personen bor sammen med mor, 0: Ellers
98 - 99	Fars *utvalgsnummer*
101 - 107	Fars *identifikasjonsnummer*
109 - 112	Fars fødselsår
114	Fars *reg.status*
116	1: Personen bor sammen med far, 0: Ellers

118	*Status for utdanningsaktiviteter*
120 - 123	*Igangværende utdanning, fagfelt+klassetrinn, MOSART-kode*
125 - 128	*Høyeste fullførte utdanning, fagfelt+klassetrinn, MOSART-kode*
130 - 133	Årstall for fullføring av høyeste fullførte utdanning
135	*Trygdestatus*
137 - 139	*Trygdetype*
141 - 144	Årstall for når personen ble pensjonist mv
146	1: Personen har vært ufør, 0: Ellers
148	*Yrkesdeltaking målt ved pensjonsgivende inntekt*
150	*Stabilitet i yrkesdeltaking*
152 - 154	Antallet år i siste status for yrkesdeltaking
156	*Klasse for pensjonsavgift*
158 - 165	Pensjonsgivende inntekt i 1993-kroner
167 - 175	Nettoformue stat
177 - 183	Ligningsverdi av bolig
185	1: Ansatt med rett til statstjenestepensjon, 2: Tidligere ansatt, 0: Ellers
187 - 190	Antatt første år personen ble ansatt i staten, 0: Ukjent/ikke ansatt
192 - 195	Antatt siste år personen jobbet i staten, 0: Ukjent/fortsatt ansatt/ikke ansatt
197	1: Mottar pensjon fra statens pensjonskasse, 0: Ellers
199 - 204	Ytelse fra Statens pensjonskasse før samordning med folketrygden

A.2. Mødrefil, fedrefil

Datatype: Ascii-format

Filident: /ssb/johansen/d1/mosart/input24/population/sample_*ii*/mothers
Enhet: En record for hver fødsel kvinnen har hatt
Antall: 28400
Sortering Mors identifikasjonsnummer × barnets fødselsår

Filident: /ssb/johansen/d1/mosart/input24/population/sample_*ii*/fathers
Enhet: En record for hvert barn mannen er far til
Antall: 27400
Sortering Fars identifikasjonsnummer × barnets fødselsår

Felt	***Innhold***
1 - 2	Barnets *utvalg*
4 - 10	Barnets *identifikasjonsnummer*
12	Barnets *kjønn*
14 - 17	Barnets fødselsår
19	Barnets *reg.status*
21 - 22	Mors *utvalgsnummer*
24 - 30	Mors *identifikasjonsnummer*
32 - 35	Mors fødselsår
37	Mors *reg.status*
39	1: Mor har felles adresse med barnet, 0: Ellers
41 - 42	Fars *utvalgsnummer*
44 - 50	Fars *identifikasjonsnummer*
52 - 55	Fars fødselsår
57	Fars *reg.status*
59	1: Far har felles adresse med barnet, 0: Ellers

A.3. Forløpsfil

Datatype: ASCII-format
Filident: /ssb/johansen/d1/mosart/input24/population/sample_*ii*/events
Enhet: En record per person per mulig år i perioden 1985-1993
Antall: 412500
Sortering Identifikasjonsnummer × år

Felt	***Innhold***
1 - 7	*Identifikasjonsnummer*
9 - 10	*Utvalgsnummer*
12	*Kjønn*
14 - 17	Fødselsår
19 - 22	År
24 - 26	Alder ved utgangen av året
28	*Reg.status*
30 - 33	Årstall for siste endring i reg.status
35	*Ekteskapelig status*
37 - 40	Årstall for siste endring i ekteskapelig status
42 - 43	*Antall barn*
45 - 46	*Yngste barns alder*
48 - 51	*Igangsværende utdanning, fagfelt+klassetrinn, MOSART-kode*
53 - 56	*Høyeste fullførte utdanning, fagfelt+klassetrinn, MOSART-kode*
58 - 61	Årstall for fullføring av utdanning
63	*Trygdestatus*
65 - 67	*Trygdetype*
69 - 72	Årstall for når personen ble pensjonist mv
74	1: Personen har vært uør, 0: Ellers
76	*Yrkesdeltaking målt ved pensjonsgivende inntekt*
78	*Stabilitet i yrkesdeltaking*
80 - 82	Antallet år i siste status for yrkesdeltaking
84	*Klasse for pensjonsavgift*
86 - 93	Pensjonsgivende inntekt i 1993-kroner

A.4. Pensjonspoengfil

Datatype: ASCII-format
Filident: /ssb/johansen/d1/mosart/input24/population/sample_*ii*/pension_points
Enhet: En record per person per forekommende år i perioden 1967-1994
Antall: 599000
Sortering Identifikasjonsnummer × år

Felt	***Innhold***
1 - 7	*Identifikasjonsnummer*
9 - 10	*Utvalgsnummer*
12 - 15	År
17 - 21	Pensjonspoeng, desimalformat
23 - 27	Beregnet uførepoeng, desimalformat
29 - 31	Høyeste uføregrad i prosent
33 - 40	Pensjonsgivende inntekt i 1993-kroner
42	*Klasse for pensjonsavgift*

A.5. Bostedskommune

Datatype: ASCII-format
Filident: /ssb/johansen/d1/mosart/input24/population/common_data/kommune
Enhet: En record per person med kjent bostedskommune
Antall: 523740
Sortering Identifikasjonsnummer

Felt	*Innhold*
1 - 7	*Identifikasjonsnummer*
9 - 10	*Utvalgsnummer*
12 - 15	Bostedskommune

A.6. Ikke aggregerte utdanningsdata

Datatype: ASCII-format
Filident: /ssb/johansen/d1/mosart/input24/population/common_data/ukodet
Enhet: En record per person per år 1985-1994 hvor utdanningsopplysninger finnes
NB! For 1994 gjelder dette bare fullføring av grunnskole for 15-17 åringer.
Antall: 3694133
Sortering Identifikasjonsnummer × år

Felt	*Innhold*
1 - 7	*Identifikasjonsnummer*
8 - 11	År
12 - 15	Fullføringsår for utdanning
16 - 17	Klassetrinn for høyeste fullførte utdanning
18 - 23	Utdanningens art for høyeste fullførte utdanning
24 - 25	Klassetrinn for igangværende utdanning
26 - 31	Utdanningens art for igangværende utdanning

A.7. Ikke aggregerte attføringsdata

Datatype: ASCII-format
Filident: /ssb/johansen/d1/mosart/input24/population/common_data/attforing
Enhet: En record per person per år 1989-1994 hvor attføringsdata finnes
Antall: 50282
Sortering Identifikasjonsnummer × år

Felt	*Innhold*
1 - 7	*Identifikasjonsnummer*
8 - 11	År
12	0: Fortsatt under attføring ved årsskiftet, 1: Attføring avsluttet i løpet av året
13 - 14	Hjemmel for attføring, jamfør dokumentasjon av inputfiler
15 - 18	Startår for attføringstiltaket

A.8. Yrkesdeltaking

Datatype: SAS-data
Filident: /ssb/johansen/d1/mosart/analysedata/aku93/aku.ssd01
Enhet: En record per «intervju»
Antall: 89371 intervjuer
Sortering Ingen

Kode etter variabelnavn angir SAS-format der dette ikke er numerisk ($ = tekst). Posisjon refererer til recordbeskrivelsen for AKU, spørsmål til spørreskjemaet for AKU, se Statistisk sentralbyrå (1995B).

Variabel	*Innhold*
Faktor, N	Faktor for veid oppblåsing av observasjonene, posisjon 207 - 214 i AKU
Kjonn,$1	Kjønn, posisjon 26 i AKU
Alder	Alder ved utgangen av 1993, posisjon 24 - 25 i AKU
Hstat1, $1	Hovedstatus I, posisjon 28 i AKU, 1: Sysselsatt, 2: Arbeidssøker, 3: Ikke yrkesaktiv
Hstat2, $1	Hovedstatus II, posisjon 29 i AKU
Hstat3, $1	Hovedstatus III, posisjon 30 i AKU
Hsv, $1	Hovedsaklig virksomhet, spørsmål 17 eller 45 for yrkesaktive, ellers spørsmål 54 omkodet til samme svarkoder som for spørsmål 17 og 45
Timer_a	Avtalt arbeidstid, posisjon 97-100 i AKU
Timer_u	Utført arbeidstid, posisjon 132-136 i AKU
A_ekt, $1	Samlivsstatus, spørsmål 71 i AKU, 1: Ugift, 2: Gift, 3: Samboer, 4: Før gift
A_barn	Antall barn under 16 år, posisjon 226 - 227 i AKU
A_yngst	Yngste barns alder, posisjon 226 - 227 i AKU
Utd_ha, $1	Høyeste allmennutdanning fullført, posisjon 178 i AKU
Utd_ti, $1	Har/har ikke tatt tilleggsutdanning, posisjon 179 i AKU
Utd_t2, $2	Tilleggsutdanning, spørsmål 70 i AKU
Utd_t3, $2	Tilleggsutdanning, posisjon 182 - 183 i AKU
Barn	Antall barn ved utgangen av året
Yngst	Yngste barns alder ved utgangen av året
Iguart, $2	Igangværende utdanning, art, MOSART-kode, se vedlegg B
Iguklt, $2	Igangværende utdanning, klassetrinn, MOSART-kode, se vedlegg B
Hfuart, $2	Høyeste fullførte utdanning, art, MOSART-kode, se vedlegg B
Hfuklt, $2	Høyeste fullførte utdanning, art, MOSART-kode, se vedlegg B
Hfuaar	Fullføringsår for høyeste fullførte utdanning, se vedlegg B
Studstat, $1	Status for skolegang, se vedlegg B
Try93, $1	Trygdestatus ved utgangen av 1993, se vedlegg B
Tryt93	Uføregrad/attføringstiltak ved utgangen av 1993, se vedlegg B
Try92, $1	Trygdestatus ved utgangen av 1992, se vedlegg B
Tryt92	Uføregrad/attføringstiltak ved utgangen av 1992, se vedlegg B
Trystat2	Uførhet/attføring, 1: Ingen tiltak, 2: blir pensjonist, 3: Etterlattepensjonist, 4: Alderspensjonist, AFP, 5: Delvis ufør, 6: Helt uføre, 7: Attføring
Istat	Yrkesdeltaking gitt ved pensjonsgivende inntekt, se vedlegg B
Itype, $1	Klasse for pensjonsgivende inntekt, se vedlegg B

Vedlegg B. Variabelliste

Nedenfor følger en forklaring av de variablene som er tildelt egne koder med videre innenfor dette dataprosjektet.

Identifikasjonsnummer
Vilkårlig heltall som alene én-entydig identifiserer hver person i utvalget. Verdiene går fortløpende fra 1 til 719 484. For pekere til partner, ektefelle, barn og foreldre vil '0' angi manglende opplysning. Det er satt av 7 felter til identifikasjonsnummer, da det muliggjør samme type utfiler fra simuleringsmodellen hvor nye identifikasjonsnummer fort kan overskride 999 999 (men neppe 9 999 999). Vi har ikke inkludert fødselsår og kjønn i identifikasjonsnummer, da det forenkler behandlingen av dataene, selv om det fører til at filene blir noe større.

Utvalgsnummer
Hvert underutvalg er nummerert fortløpende fra 1 til 12, miniutvalgene er nummerert henholdsvis 98 (1 promille) og 99 (0,1 promille). '0' angir manglende opplysning.

Kjønn
1: Mann, 2: Kvinne

Årstall
Årstall er angitt med fire siffer. '0' angir at årstallet er ukjent, for årstall fra personregisteret vil dette normalt tilsvare begivenheter for ekteskap og inn- og utvandring før 1964.

Reg.status
1: Bosatt i Norge, 2: Død, 3: Utvandret, 0: Ukjent

Ekteskapelig status
1: Ugift, 2: Gift, 3: Enke/enkemann, 4: Skilt, 5: Separert, 0: Ukjent

Ektefelle, partner
Ektefelle peker til nåværende ektefelle for gifte og siste ektefelle for før gifte. Partner peker til den personen vedkommende bor sammen med, hvis de enten er gift eller har barn sammen.

Antall barn, alder yngste barn
Der opplysninger mangler er '0' satt inn.

Status for skolegang
1: Ingen utdanningsaktiviteter, 2: Er under utdanning, 3: Har begynt utdanning i år, 4: Kandidat, 5: Har avbrutt utdanning i år.

Utdanning, klassetrinn
00: Uoppgitt/ingen utdanningsaktivitet, 09: Grunnskole, 10-19: Angir normert studietid inkludert 9-årig grunnskole

Utdanning, fagfelt
Tall i parentes angir hvilke klassetrinn det enkelte fagfelt går over for respektive utdanningsaktiviteter og utdanningsnivå. Se for øvrig Andreassen et al (1993) for detaljer.

Lavere nivå
01: Ingen utdanningsaktiviteter (00,-), 02: Uoppgitt utdanning (-,00), 03: Grunnskole (09-10)
Videregående skole

04: Gymnas (10-12,12), 10: Andre fag (10-12), 11: Husstellinje, befalsskole, folkehøyskole (10), 21: Økonomi og adm. (10-12), 22: Industri og håndverk (10-12), 23: Hjelpepleie (11)
Høyskoleutdanninger
40: Andre fag (10-18,12-18), 51: Ingeniører (11-16,12-16), 52: Økonomi og adm. (13-16), 53: Sykepleie (13-16,15-16), 54: Undervisning (13-18, 13-16+18),
Universitetsfag
71: Ex.phil (13), 72: Andre fag (19), 81: Humaniora (13-19), 82: Samfunnsfag (13-19), 83: Naturfag (13-19), 91: Jurister (13-19,15+17+19), 92: Sivilingeniører (13-19, 17+19), 93: Leger (13-19,18-19), 94: Tannleger (13-19,17+19)

Trygdestatus
1: Ikke pensjonist, 2: Under attføring ved årslutt, 3: Uførepensjonist, 4: Alderspensjonist, 5: Etterlattepensjonist, 6: AFP-pensjonist

Trygdetype
- Attføring - 10: Imputert attføring, 11: Ventetid for uførepensjon, 12: Medisinsk attføring, 13: Yrkesrettet attføring, 14: Andre årsaker.
- Uførepensjon - Tall angir uføregrad i prosent
- Alderspensjon - 0: Imputert alderspensjon, 50: Delpensjonist, 100: Full pensjon

Kronebeløp
Alle kronebeløp er oppgitt i faste 1993-kroner deflatert ved Statistisk sentralbyrås konsumprisindeks.

Yrkesdeltaking gitt ved pensjonsgivende inntekt
J - har pensjonsgivende inntekt større enn 1000 Nkr, N - Ellers, x - Vilkårlig hvilken verdi.
J/N/x-koder i parentes angir inntekt året før, inneværende år og neste år.

1: Er i arbeid (JJJ), 2: Begynner å arbeide (NJJ) 3: Slutter å arbeide (JJN), 4: Ikke i arbeid(xNx), 5: Tilfeldig inntekt (NJN).

Antall år i siste status for yrkesdeltaking
Angir antall år i siste status for yrkesdeltaking, hvor kode 1 og 2 er slått sammen (yrkesaktiv) og kode 3, 4 og 5 er slått sammen (yrkespassiv), og hvor første år i ny status starter med verdi 1.

Stabilitet i yrkesdeltaking
Lavere kode går foran høyere kode

1: Stabilt yrkesaktiv, dvs minst 5 sammenhengende år som yrkesaktiv, 2: Ny yrkesaktiv,minst 5 sammenhengende år som yrkespassiv før dette, 3: Ustabilt yrkesaktiv, alle andre yrkesaktive, 4: Stabilt yrkespassiv, dvs minst 5 års sammenhengende år som yrkespassiv, 5: Helt ny yrkespassiv, minst 5 sammenhengende yrkesaktive år før dette og blitt yrkespassiv samme år, 6: Ny yrkespassiv, minst 5 sammenhengende yrkesaktive år før dette, 7: Ustabilt yrkespassiv, alle andre yrkespassive

Klasse for pensjonsavgift
1: Inntekt som ansatt utgjør minst 50 prosent av pensjonsgivende inntekt, 2: Inntekt som ansatt viktigst, 3: Inntekt som selvstendig næringsdrivende viktigst, 4: Inntekt som selvstendig i jordbruk, fiske og skogbruk viktigst

Vedlegg C. Filreferanser

Filreferanser nedenfor er fra comparex-maskinen på Kongsvinger. Enkelte av filene vil være dokumentert på egne filer på «området» 'ssb1.dokument.record', hvor seks første siffer i tredje ledd angir navnet på dokumentasjonfila. For eksempel er bosatte per 1.1.1994 dokumentert under 'ssb1.dokument.record.i459a8'.

Situasjonsuttak fra Personregisteret

Bosatte per 1.1.1994	PL214.S0108.I459A8A7.G93MC.V00
Ikke bosatte per 1.1.1994	PL214.S0108.I459D1A2.G9400.V00
Bosatte per 1.1.1993	PL214.S0108.I459A8A7.G92MC.V00
Bosatte per 1.1.1992	PL214.S0108.I459A8A7.G91MC.V05
Bosatte per 1.1.1991	PL214.S0108.I459A8A7.G90MC.V04
Bosatte per 1.1.1990	PL214.S0108.I459A8A7.G89MC.V00
Bosatte per 1.1.1989	PL214.S0108.I459A8A7.G88MC.V00
Bosatte per 1.1.1988	PL214.S0108.I459A8A7.G87MC.V00
Bosatte per 1.1.1987	PL22.S0108.I459A8A8.G86MC.V00
Bosatte per 1.1.1986	PL22.S0108.I459A2A3.G85MC.V00

Utdanning

Bosatte per 1.10.1985	PL213.S4368.I654A1A1.G8500.V01
Bosatte per 1.10.1986	PL213.S4368.I654A1A1.G8600.V01
Bosatte per 1.10.1987	PL213.S4368.I654A1A1.G8700.V00
Bosatte per 1.10.1988	PL213.S4368.I654A1A1.G8800.V00
Bosatte per 1.10.1989	PL213.S4368.I654A1A1.G8900.V01
Bosatte per 1.10.1990	PL213.S4368.I654A1A1.G9000.V00
Bosatte per 1.10.1991	PL213.S4368.I654A1A1.G9100.V00
Bosatte per 1.10.1992	PL213.S4368.I654A1A1.G9200.V00
Bosatte per 1.10.1993	PL213.S4368.I654A1A1.G9300.V00
Bosatte per 1.10.1994	PL213.S4368.I654A1A1.G9400.V00

Attføring

Attføringspengetilfelle i 1989	PL217.S4169.I984A1A1.G8900.V00
Attføringspengetilfelle i 1990	PL217.S4169.I984A1A1.G9000.V00
Attføringspengetilfelle i 1991	PL217.S4169.I984A1A1.G9100.V01
Attføringspengetilfelle i 1992	PL217.S4169.I984A1A1.G9200.V00
Attføringspengetilfelle i 1993	PL217.S4169.I984A1A1.G9300.V00

Pensjonister

Pensjonister i 1985	PL217.S4169.I560A1A1.G8500.V00
Pensjonister i 1986	PL217.S4169.I560B2A1.G8600.V00
Pensjonister i 1987	PL217.S4169.I560C2A1.G8700.V01
Pensjonister i 1988	PL217.S4169.I560C2A1.G8800.V00
Pensjonister i 1989	PL217.S4169.I560C2A1.G8900.V00
Pensjonister i 1990	PL217.S4169.I560C2A1.G9000.V00
Pensjonister i 1991	PL217.S4169.I560C2A1.G9100.V00
Pensjonister i 1992	PL217.S4169.I560C6A1.G9200.V00
Pensjonister i 1993	PL217.S4169.I560D2A1.G9300.V00

Pensjonspoeng m.v.

Dataene er hentet ut fra den «rekonstruerte folketrygdbasen» i Rikstrygdeverket ved program 'R0029906' (EAm/HaR 23.10.85, upublisert notat, Rikstrygdeverket).

Ligningsdata

Ligningsbånd for 1992	PL244.S3515.A891J3A4.G9200.V00
Ligningsbånd for 1993	PL244.S3515.A891J3A4.G9300.V00
Ligningsbånd for 1994	PL244.S3515.A891J3A4.G9400.V00

Statens pensjonskasse

Opplysningene er hentet ut av Trond Nystad, Statens pensjonskasse.

Arbeidskraftundersøkelsene

AKU for 1.kvartal 1993	P6245.S0211.B050N8A8.G93K1.V00
AKU for 2.kvartal 1993	P6245.S0211.B050N8A8.G93K2.V00
AKU for 3.kvartal 1993	P6245.S0211.B050N8A8.G93K3.V00
AKU for 4.kvartal 1993	P6245.S0211.B050N8A8.G93K4.V00

Vedlegg D. Oversikt over edb-arbeidet

Edb-arbeidet er delt mellom comparex-maskinen på Kongsvinger og arbeidsstasjonen 'johansen' tilknyttet MOSART-prosjektet. Arbeidet på comparex-maskinen er begrenset til det som var nødvendig å gjøre der, det vil si å ta ut data fra de filene som er knyttet opp mot comparex-maskinen. Dataene er deretter overført til arbeidsstasjon, hvor koblinger, opprettinger og tilrettelegging har foregått. Nedenfor omtales de edb-programmene som er kjørt for å koble utgangspopulasjonen, med en kort omtale av innholdet i hver jobb og hvor de finnes lagret.

D.1. Uttak av data på comparex

Nedenfor gis en oversikt over de edb-programmene som er brukt til å trekke utvalget og hente ut data på comparex-maskinen på Kongsvinger. Trekkingen av utvalget ligger på «området» 'o320dff.tryg93.utvalg', og omfatter i tillegg noen tabelluttak for kontroll av resultatfilene. Uttaket av data ligger på «området» 'o320dff.tryg93.data'. Navnene i parentes angir programmets filnavn på respektive områder. Programmer som henter ut data for koblingen av AKU mot registre ligger på «området» 'o320.dff.tryg93.aku' (ellers ikke omtalt).

Innlesing av bosatte personer per 1.1.1994 (utvalg.lesbos)
Først leses inn bosatte personer per 1.1.1994, og samtlige personer fordeles på henholdsvis en mannsfil og en kvinnefil. I tillegg opprettes en barnefil med opplysninger omkring foreldre.

Innlesing av ikke-bosatte personer per 1.1.1994 (utvalg.lesibos)
Videre leses inn ikke-bosatte personer per 1.1.1994, og disse fordeles på de samme filene som i forrige trinn.

Kobling av menn og barn (utvalg.mbmenn) og kvinner og barn (utvalg.mbkvin)
Barnefila sorteres på fars fødselsnummer, og opplysningene flettes mot mannsfilen som gir mennene opplysninger om antall barn og alder på yngste barn. I den grad mannen ikke har en ektefelle, knyttes han opp mot mor til sitt eget yngste barn. I den grad far og barn (formelt sett) bor sammen legges denne informasjonen tilbake på barnefila.

Tilsvarende jobb utføres for kvinner og barn.

Svakhet! Fars og mors reg.status (og fødselsår) skulle vært overført til barnefila.

Kobling av ektefeller og foreldre (utvalg.mmk)
Mennene sorteres først på ektefellens/yngste barns mor sitt fødselsnummer. Deretter flettes mennene mot sine kvinner med kvinnens og sitt eget fødselsnummer som koblingsnøkkel. Det betyr at det kun er menn og kvinner som gjensidig peker til hverandre som blir par. Etter/i løpet av koblingen fordeles samtlige personer på seks datasett som utgjør grunnlagsfilene for trekkingen av utvalget. Dette omfatter henholdsvis:

(i) Par hvor begge ektefeller/foreldre er bosatt i Norge (per 1.1.1994),
(ii) Øvrige menn bosatt i Norge, og inkluderer eventuelle partnere som ikke er bosatt i Norge.
(iii) Øvrige kvinner bosatt i Norge, og inkluderer eventuelle partnere som ikke er bosatt i Norge.
(iv) Par hvor ingen av partnerne er bosatt i Norge.
(v) Øvrige menn som ikke er bosatt i Norge.
(vi) Øvrige kvinner som ikke er bosatt i Norge.

I tillegg opprettes det henholdsvis en enkefil og enkemannsfil for å kunne sjekke om eventuell gjenlevende ektefelle av døde personer (fortsatt) er enker eller enkemenn.

Svakhet! Par hvor kun den ene parten peker til den andre burde vært etablert som par, dette gjelder spesielt hvor kun den ene ektefellen har utvandret.

Trekking av utvalg (utvalg.utvbp, .utvm, .utvk, .utvibp, .utvibm, .utvibk)
Trekkingen av utvalget gjøres med seks separate jobber med relativt lik oppbygging. Først legges alle personer fra grunnlagsfila ut på en trekkefil med identifikasjon av moder-recorden (fødselsnummer) og de variablene som skal brukes i stratifiseringen. Dernest sorteres trekkefila på stratifiseringsvariablene. Trekkeprosedyren er bygd opp ved at første person trekkes tilfeldig blant de 8-9 første personene, og deretter inkluderes hver 8 1/3 person i utvalget. Personene fordeles fortløpende på underutvalg, og det opprettes også to miniutvalg til testformål. Til slutt flettes utvalget tilbake mot grunnlagsfila og de aktuelle personene legges ut på utvalget.

NB! Ikke samboende par er splittet opp og trukket som enkeltpersoner.

Ordning av utvalget (utvalg.utvalg)
De ulike delutvalgene slås sammen til et felles utvalg. Samtidig opprettes det en 'stamme' som inneholder meldinger om reg.status og ekteskapelig status sortert på fødselsnummer og årstall for begivenhetene. Stammen brukes senere som grunnlag for å bygge opp en fil med forløpsdata for ekteskap, barn, utdanning og trygd.

Kobling mot eldre situasjonsuttak fra personregisteret (utvalg.soek93, .soek92, .soek86)
Utvalgsfila og stammen kobles mot eldre situasjonsuttak fra personregisteret (først 1.1.1993, deretter bakover til 1.1.1986). De ulike programmene har relativt lik oppbygging, men de eldre situasjonsuttakene har noe ulik recordbeskrivelse, samt at det første søket (soek93) ikke behøver å teste for manglende opplysninger mellom det aktuelle situasjonsuttaket og de yngre situasjonsuttakene.

Programmene er relativt komplekse, men skal kun fjerne døde personer hvor eventuell gjenlevende ektefelle er enke eller enkemann (men ikke med i utvalget), legge til avdøde ektefelle av enker og enkemenn i utvalget, samt tilføye opplysninger om siste ektefelle for før gifte og endringer i reg.status og ekteskapelig status. Dette gjøres ved å koble eldre situasjonsuttak mot personer som påfølgende (eller senere) år har hatt endringer i reg.status eller ekteskapelig status. På den måten kan eksisterende opplysninger og tilhørighet til utvalget revurderes.

Feil ! Personer med uoppgitt årstall for endring i ekteskapelig status skulle automatisk vært med blant de som ble sjekket mot eldre situasjonsuttak.
Svakhet! Hele utvalget burde vært sjekket mot de eldre situasjonsuttakene, men begrensninger ved comparex-maskinen gjorde at vi droppet dette.

Overfoering av utvalget (utvalg.utvalg2)
Programmet legger utvalgsfila ut på en ascii-fil som kan overføres til pc/unix, samt oppretter en fil med fødselsnummer for uttak av data på comparex.

Overfoering av stammen (utvalg.stamme)
Programmet legger stammen ut på en ascii-fil som kan overføres til pc/unix.

Overfoering av fødsler (utvalg.fodsler)
Programmet starter med å identifisere hvilke fødselsmeldinger hvor enten barnet, moren eller faren er med i utvalget. Disse recordene tas så ut og tillegges opplysninger om reg.status og fødselsår for barnet og eventuelt moren og faren. Til slutt legges dataene ut på en ascii-fil som kan overføres til pc/unix.

Innlesing av utdanningsdata (data.bhu93, .bhu94)
Programmet leser inn fila med befolkningens høyeste utdanning, henter ut opplysninger for utvalget og legger dataene ut på en ascii-fil som kan overføres til arbeidsstasjon. Det siste programmet leser kun inn 15-17 åringer som har fullført grunnskole i 1994.

Innlesing av attføringsdata (data.attf93, .att91)
Programmet leser inn fila med attføringsklienter, skiller ut «siste» tiltak, henter ut opplysninger for utvalget og legger dataene ut på en ascii-fil som kan overføres til arbeidsstasjon.

Innlesing av pensjonister (data.gr1d93, .gr1d92, .gr1d91, .gr1d86)
Programmet leser inn fila med pensjonister (GR1), henter ut opplysninger for utvalget og legger dataene ut på en ascii-fil som kan overføres til arbeidsstasjon.

Utsendelse av fødselsnumre til Rikstrygdeverket (utvalg.rtvfnr)
Programmet overfører listen med fødselsnumre til en ascii-fil som kan videresendes til Rikstrygdeverket (driftskontoret står for oversendelsen).

Innlesing av pensjonspoeng fra Rikstrygdeverket (data.rtvdata)
Programmet leser inn fila med pensjonspoeng mv fra Rikstrygdeverket, slår sammen pensjonsgivende inntekt fra ulike kilder, definerer viktigste inntektstype, og legger resultatet ut på en fil som kan overføres til arbeidsstasjon.

Innlesing av ligningsdata (data.skatt)
Programmet leser inn ligningsbåndet, og henter ut opplysninger om nettoformue stat og boligformue, og legger resultatet ut på en fil som kan overføres til arbeidsstasjon.

Utsendelse av fødselsnumre til Statens pensjonskasse (utvalg.spkfnr)
Programmet overfører listen med fødselsnumre til en ascii-fil som kan videresendes til Statens pensjonskasse (driftskontoret står for oversendelsen).

Innlesing av data fra Statens pensjonskasse (data.spkdata)
Programmet vil lese opplysninger fra Statens pensjonskasse.

D.2. Tilrettelegging av data på unix/arbeidsstasjon

Nedenfor gis en oversikt over de edb-programmene som er brukt til å ordne utvalget og dataene på unix/arbeidsstasjon. Samtlige filer er overført med et standard overføringsprogram som ligger på nettet. Jobbene ligger på området '/ssb/johansen/d1/mosart/input31/population/common_data'. Programmer som kobler AKU mot registre ligger på området '/ssb/johansen/d1/mosart/input31/population/common_data/akudata' (ellers ikke omtalt).

Innlesing av utvalget og anonymisering (lesutv.sas)
Programmet leser inn utvalgsfila som er overført fra comparex, og legger den ut som en sas-fil. Samtidig legges alle fødselsnumre ut på en egen fil for anonymisering. Anonymiseringen skjer ved å tilordne hvert fødselsnummer et tilfeldig tall, og deretter sortere på dette tilfeldige tallet. Første person får identifikasjonsnummer 1, andre person får identifikasjonsnummer 2 og så videre. Fødselsnummeret sammen med det anonyme identifikasjonsnummeret og utvalgsnummeret legges ut på en egen fil for å kunne anonymisere andre data også. Alle fødselsnumre på utvalgsfila blir tilordnet sine anonymiserte identifikasjonsnumre i denne jobben.

Innlesing av fødselsmeldinger (lesfods.sas)
Programmet leser inn fødselsmeldingene som er overført fra comparex, og legger den ut som to sas-filer for henholdsvis mødre og fedre. Opplysninger om foreldre legges inn på utvalgsfila.

Opplysninger om bostedskommune legges ut på egen resultatfil, se A.5. Alle fødselsnumre blir anonymisert.

Innlesing av stammen (lesstamme.sas)
Programmet leser inn stammen som er overført fra comparex, og legger den ut som en sas-fil. Samtidig legges opplysninger fra utvalgs-, mødre- og fedrefila til stammen. Fila organiseres som en forløpsfil med en record for hver person denne kan ha opplysninger om seg i perioden 1985-1993. Alle fødselsnumre bli anonymisert.

Oppretting av ektefelle identer (reputv1.sas)
Pekere fra avdøde ektefeller til gjenlevende ektefeller rettes opp i denne jobben.

Nullstilling av utdanningsdata (stammeutd.sas)
Programmet setter default-verdier på utdanningsvariablene på forløpsfila.

Innlesing av utdanningsdata (bhu93.sas, bhu92.sas, bhu91.sas)
Programmene leser inn utdanningskjennetegn, aggregerer disse og legger opplysningene ut på forløpsfila. For 1992- og 1993-årgangen legges også opplysninger til utvalgsfila. Ukodet versjon av utdanning legges ut på egen fil.

Oppretting av utdanningsnivå (rephfu.sas, rephfu2.sas)
Programmet tilbake- og framdaterer utdanningsopplysninger der disse er manglende eller inkonsistente med senere opplysninger. For eksempel vil personer i utgangspunktet mangle utdanningsopplysninger i de årene de er utvandret og enkelte innvandrere har i egne undersøkelser fått fastlagt utdanningsnivået uten at tidligere filer er rettet opp. Den første jobben retter opp forløpsfila, mens den andre jobben overfører opplysningene til utvalgsfila.

Imputering av grunnskole (repigu.sas, repigu2.sas)
Programmet imputerer grunnskole som skoleaktivitet for personer 14 år og eldre opp til året før de har fullført grunnskole (der dette er kjent, opplysninger for 1994 hentes inn separat). Personer som ikke har fullført grunnskole før 1.10.1994 har ikke fått slike opplysninger imputert. Den første jobben retter opp utvalgsfila. Den andre jobben overfører opplysningene til utvalgsfila sammen med opplysninger om studentstatus, det vil si om personen er under utdanning, har begynt eller har fullført/avbrutt en utdanning.

Nullstilling av trygdedata (stammetry.sas)
Programmet setter default-verdier for trygdestatus, trygdetype og trygdeår.

Innlesing av attføringsdata (lesattf93.sas)
Programmet leser inn attføringsdata og legger disse ut på forløpsfila.

Innlesing av pensjonistdata (lesgr1.sas)
Programmet leser inn pensjonistdata, retter opp trygdetype og sammenholder med attføringsdata, før disse opplysningene legges ut på forløpsfila.

Innlesing av pensjonspoeng (lesrtv.sas)
Programmet leser inn pensjonspoeng mv, deflaterer kronebeløp og definerer bruttostrømmer i pensjonsgivende inntekt. Pensjonsrettighetene legges ut på egen fil, mens deler av inntektsdataene legges til forløpsfila og utvalgsfila.

Oppretting av pensjonistdata (reptry.sas)
Programmet tilbakedaterer attføringsopplysninger og imputerer antatt attføring på grunnlag av opplysninger om beregnet uførepoeng. Videre sjekkes alle personer 67 år og eldre, og omgjøres til

alderspensjonister hvis opplysningene tilsier det. Trygdestart og om personen har vært uføre beregnes. Opprettede trygdedata og nye opplysninger overføres til utvalgsfila.

Innlesing av ligningdata (lesskatt93.sas)
Programmet leser inn opplysninger om formue og legger disse ut på utvalgsfila.

Innlesing av data fra Statens pensjonkasse (lesspk.sas)
Programmet vil lese inn data fra Statens pensjonskasse.

Ordning av utvalget (ordnutv)
Programmet legger utvalgsfila ut på en ascii-fil, og dette er et av sluttproduktene omtalt i A.1.

Ordning av mødrefil, fedrefil (ordnfods.sas)
Programmet legger mødre og fedrefila ut på en ascii-fil, og dette er et av sluttproduktene omtalt i A.2.

Ordning av forløpsfil (ordnstamme.sas)
Programmet legger forløpsfila ut på en ascii-fil, og dette er et av sluttproduktene omtalt i A.3.

Ordning av pensjonspoengfil (ordnpoeng.sas)
Programmet legger trygdedataene ut på en ascii-fil, og dette er et av sluttproduktene omtalt i A.4.

Ordning av ikke aggregerte utdanningsdata (ordnutd.sas)
Programmet legger den ukodete versjonen av utdanning ut på en ascii-fil, inkludert opplysninger om 15-17 åringer som har fullført grunnskole i 1994. Se vedlegg A.6.

Ordning av ikke aggregerte attføringsdata (ordnattf.sas)
Programmet legger den ukodete versjonen av attføring ut på en ascii-fil, se vedlegg A.7.

Utkommet i serien Notater fra Forskningsavdelingen

94/11 E. Holmøy og B. Strøm: Virkningsberegninger på MGS-5, 1991-versjonen

94/12 K.Ø. Sørensen: En databank med fylkes-fordelte nasjonalregnskapstall

94/13 B. Holtsmark: Tjenesteytende virksomhet i Norge. Revidert versjon, august 1994

94/15 T. Eika, S.I. Hove og L. Haakonsen: KVARTS i praksis. Macro-systemer og rutiner

94/17 E. Bowitz og I. Holm: Nye relasjoner i MODAG, januar 1994. Teknisk dokumenta-sjon

94/18 Y. Vogt: Innføring i FAME

94/22 M.W. Arneberg: LOTTE-TRYGD. Teknisk dokumentasjon

95/5 D. Fredriksen: MOSART Teknisk dokumentasjon

95/7 K. Olsen: Nytte- og kostnadsvirkninger av en norsk oppfyllelse av nasjonale utslipps-målsettinger

95/15 T. Karlsen: Opptimal karbonbeskatning og virkningen på norsk petroleumsformue

95/17 Å. Cappelen, T. Skjerpen og J. Aasness: Konsumetterspørsel, tjenesteproduksjon og sysselsetting. En mikro til makroanalyse

95/24 H.T. Mysen: Nordisk energimarkeds-modell. Dokumentasjon av delmodell for energi-etterspørsel i industrien

95/26 I. Aslaksen, T. Fagerli og H.A. Gravningsmyhr: Produksjon og konsum i husholdningene

95/29 B.E. Naug: Eksport- og importlikninger i KVARTS

95/31 B.E. Naug: Etterspørsel etter arbeidskraft - en litteraturoversikt

95/35 T.J. Klette: Vekst og produktivitet i norsk industri. Hovedrapport fra et NFR-prosjekt

95/40 L. Lerskau: Oversikt over konjunktur-indikatorer i databasen NORMAP og FAME

95/46 B.E. Naug: Estimering av eksport-relasjoner på disaggregerte kvartalsdata

95/47 K. Moum: Beregning av bruttoproduksjon og eierinntekt i boligsektoren i nasjonal-regnskapet - noen metodiske synspunkter

95/52 T. Kornstad: Simulering av konsum og arbeidstilbud i et livsløpsperspektiv

95/56 A. Langørgen: Faktorer bak kommunale variasjoner i utgifter til sosialhjelp og barnevern

95/58 T. W. Karlsen: Energimarkedet fra 1973 og fram mot 2010

96/3 I. M. Smestad: Valg under usikkerhet: En analyse av eksperimentdata basert på kvalitative valghandlingsmodeller

96/8 B. Lian og K. O. Aarbu: Dokumentasjon av LOTTE-AS

96/9 D. Fredriksen: Datagrunnlaget for modellen MOSART, 1993

Statistisk sentralbyrå

Oslo
Postboks 8131 Dep.
0033 Oslo

Telefon: 22 86 45 00
Telefaks: 22 86 49 73

Kongsvinger
Postboks 1260
2201 Kongsvinger

Telefon: 62 88 50 00
Telefaks. 62 88 50 30

ISSN 0806-3745

Statistisk sentralbyrå
Statistics Norway