Interne notater

STATISTISK SENTRALBYRÅ

90/13

15. mars 1990

MIKROSIMULERINGSMODELLER OG MODELLBEFOLKNING.

REISERAPPORT FRA STUDIETUR TIL STATISTICS CANADA, OTTAWA, 27/11-1/12 1989.

Av Olav Ljones

SAMMENDRAG

Reisen til Statistics Canada var en del av en besøksavtale for ansatte i Statistisk sentralbyrå og varte en uke (27. nov - 1.des. 1989). Statistics Canada ga inntrykk av å være veldrevet med høyt faglig nivå på flere områder.

Statistics Canada har en forskningsavdeling (Analytical Studies Branch) hvor det er en gruppe (Social and Economic Studies) som arbeider med simuleringsmodeller for demografi, arbeidsmarked og offentlig politikk. Modellene som de arbeider med har en del til felles med norske tilnærminger på det samme området. De har en statisk mikromodell som regner skatter på tilsvarende måte som den norske modellen LOTTE. Den canadiske modellen skiller seg imidlertid fra den norske ved at den i datagrunnlaget har med langt flere kjennetegn ved inidividene, og at det er inkludert langt flere politikkvariable. De har inkludert både direkte og indirekte skatter, sammen med regler for arbeidsløshetsstønad og alderspensjon. Arbeidet med å tilrettelegge primærdata fra ulike kilder til en modellbefolkningen er en viktig del av arbeidet. Det legges også vekt på å sikre konsistens med nasjonalregnskapstall. De har hittil ikke lagt vekt på å bygge inn økonometriske atferdsrelasjoner.

De har også utviklet demografiske og økonomiske forløpsmodeller som inkluderer yrkesdeltaking og inntekt. Grunnmodellen - DEMOGEN - brukes til analyse av samfunnskonsekvenser av demografiske utviklingstrekk, og som førmodell for modeller som behandler inntekt, trygd, helse og eldreomsorg. I denne modellen inngår demografiske rater som brukes til å simulere kjennetgen ved individer fra en fødselskohort. Modellen utnytter demografiske rater og andre "overgangssansynligheter" fra mange kilder.

I arbeidet med å utvikle mikrosimuleringsmodellene benytter en stormaskin til tilrettelegging av primærdata. Når disse er klargjort utnytter en intensivt ulike teknikker til å presse data og modeller sammen slik at de kan utvikles for personlige datamaskiner (PC). Alle modellene er utviklet for PC-bruk.

1. INNLEDNING.

Reisen var kommet i stand etter kontakt mellom sjefene i Statistics Canada, Ivan Fellegi og i Norge, Gisle Skancke. Ansvaret for gjennomføringen av reisen hadde metodegruppa ved Ib Thomsen. Vi ble under besøket mottat av direktør G.J. Brackstone (sammen med Susan Cepella, B. Prigly og D. Binder) for en generell orientering om Statistics Canada (SC). Etter den generelle orientering var det lagt opp til individuelle program. Avdelingen for internasjonale kontakter hadde lagt forholdene godt til rette for studiebesøket med et velorganisert program.

Statistics Canada er plassert i den offentlige forvaltning på en noe tilsvarende måte som i Norge. Statistics Canada er imidlertid ikke plassert under Finansdepartementet, men under et departement for næring og forskning.

Statistics Canada har for tiden ca 4400 ansatte, men er inne i en periode med kutt i antall stillinger. De har regionale kontorer (provinsene) som har 500-600 ansatte. I tillegg kommer intervjustaben.
Organisasajonskartet viser en organisasjon med mange fellestrekk med vår, med blant annet en egen systemenhet (Informatics). Fra denne kan de andre avdelingene "kjøpe" tjenester. Enkelte avdelinger har imidlertid også sin egen systemekspertise.

Av prioriterte oppgaver for Statistics Canada ble det trukket fram forbedring (redesign) av det som tilsvarer vårt bedriftsregister. Statistics Canada gjennomfører folketellinger hvert 5 år. Den neste skal de ha i 1991. Den forrige var i 1986. Den skulle etter normalt tidsskjema vært gjennomført i 1984, men ble avlyst men så likevel gjennomført to år etter i 1986.

SC har lagt ned mye arbeid i revisjon av sitt publikasjonsprogram og er kjent for å ha enkelte meget vakre publikasjoner. De har også gjort endringer i retten som de enkelte kontorer har til å disponere overskudd fra salg. Dette har ledet til en større interesse for å gjøre publikasjonene attraktive i markedet.

SC har en relativt sterk oppdeling i kontorer og avdelinger og sett utenfra kan det se ut som det er mange atskilte, men beslektede enheter. Et inntrykk som en kan sitte igjen med, er at dette av og til kan resultere i en noe sterk bås-oppdeling. På den annen side kan dette gi klare ansvarsforhold, noe som kan gi et klart resultatansvar.

Maskinutstyret i SC består av en kombinasjon av stormaskin og PC. Det var også tatt i bruk nettverk.

2. FORSKNINGSVIRKSOMHETEN I "ANALYTICAL STUDIES BRANCH".

Jeg brukte mye av tiden til besøk i avdelingen; "Analytical Studies Branch". Dette er en forskningsavdeling som er underlagt "National Accounts and Analytical Services". Denne avdelingen har en del fellestrekk ved vår egen forskningsavdeling. Den enheten som jeg i første rekke konsentrerte meg om var "Social and Economic Studies", som ledes av Michael Wolfson, har omtrent 12 ansatte. Den arbeider med å utvikle ulike typer av simuleringsmodeller for analyser av inntekter, skatter, trygder, offentlige utgifter etc. Det legges også stor vekt på modellering av demografiske forhold, som for eksempel ekteskap og skilsmisse.

Ved siden av denne enheten besøkte jeg også en enhet for "Business and Labour Market Analysis." som arbeider med studier av inntektsfordeling og med paneldata for bedrifter. Det var også en enhet for "Environment and Natural Resources". Denne var relativt ny, og jeg fikk ikke anledning til å sette meg noe nærmere inn i hva denne enheten arbeidet med. Det var også en enhet for "Population Studies". Denne hadde bare en person tilsatt og var av mindre interesse. Arbeidet med befolkningsframskrivinger av tradisjonell type, var plassert i en helt annen avdeling, "Census and Demographic Statistics".

Avdelingen la såvidt jeg kunne se vekt på å holde en rimelig høy forskningsmessig profil, og mange av medarbeiderne hadde doktorgrad. De hadde sin egen publikasjonsserie "Research Paper Series." For å styrke miljøet og utvide den forskningsmessige utnyttelsen av data som SC samler inn var det flere universitetsfolk knyttet til avdelingen som deltidsansatte. For de prosjekter og modeller som jeg satte meg inn i, hadde avdelingen selv ansvaret for systemutvikling. De regnet det som nærmest uaktuelt å basere seg på tjenester fra den sentrale systemenhet. Modellutvikling og systemutvikling virket fullt integrert i prosjektarbeidet sammen med det øvrige arbeid. Medarbeiderne var såvidt jeg forstod gjennomgående rekruttert med økonom/statistiker bakgrunn. For de som særlig arbeidet med systemoppgaver, var det behov for interesse og legning for systemutvikling.De som ble ansatt fikk så god opplæring i de mest brukte programspråk (SAS og C). Det var ikke organisert noen egen systemenhet på Analytical Studies Branch. For enkelte spesiellt vanskelige oppgaver som de ikke hadde forutsetninger for å løse selv, kjøpte de tjenester utenfra.

Ved siden av forskningsrettet arbeid av meritterende karakter, bestod arbeidsoppgavene av systemutvikling og datatilrettelegging. De hadde drøftet noe om de skulle innføre ulike karriereveier for de som hadde forskningsoppgaver og de andre, men ennå ikke kommet til

en avlaring av dette.

I enheten for Social and Economic Studies, ble jeg godt mottatt og fikk mange og nyttige samtaler med Wolfson og hans medarbeidere. Gruppen har mange ulike varianter av mikrosimuleringsmodeller. De er enten statiske som vår LOTTE, eller de baserer seg på simulering av livsløp, slik som vi har tenkt i MOSART.

Felles for alle modellene er at de nær fullstendig er programmert av medarbeiderne i gruppen, og at de lages for bruk på PC. Dette gjøres blant annet med tanke på at modellene skal kunne selges eller overleveres til eksterne brukere.

3. STATISK SIMULERINGSMODELL FOR OFFENTLIG POLITIKK

3.1. Beskrivelse av den statiske politikk simuleringsmodellen - SPSD/M (Social Policy Simulation Database/Model)

Dette er en relativt nyutviklet modell med tanke på en bred bruk blant mange typer av brukere - også utenfor SC. De har til forskjell fra oss, ikke noe løpende ansvar for å delta i skatteberegninger for Finansdepartement eller parlament. De hadde til nå ikke hatt noe offisielt ansvar for modellutvikling, siden Finansdepartementet har sine egne modeller. De modellene som SES har utviklet har således hatt som siktemål å gi modellkunnskap til institusjoner og personer utenfor departementet. SPSD/M er utarbeidet med tanke på åpen distribusjon til alle ved kommersielt salg. De hadde gjennomført markedsundersøkelser for å utrede hvordan en modell skal se ut når det gjelder brukervennlighet. Det var imidlertid litt uklart for meg om også Finansdepartementet nå hadde tatt i bruk modellen.

Modellen har mange interessante sider for vårt videre utviklingsarbeid, slik at sider ved modellen vil bli gjennomgått relativt detaljert. Modellen er laget for salg og er skrevet for PC. Det finnes en miniversjon som er en slags demonstrasjonsversjon. Denne selges for 100 dollar. Jeg fikk med en slik demonstrasjonsutgave av modellen som gave.

Modellen består av to deler, SPSD/* som er en database og modelldelen SPS*/M.

3.2 Modellbefolkning - SPSD

Modellbefolkningen (databasen) bygger både på ulike intervjuundersøkelser på utvalgsbasis, på registeropplysninger og på nasjonalregnskapsdata. Individdata blir anonymisert slik at databasen kan brukes fritt. Databasen bygger på fire ulike datakilder og bruker flere metoder for å bringe disse sammen til ett

"syntetisk" mikrodatamateriale.

De fire datakildene som modellbefolkningen bygger på er:

1. Survey of Consumer Finances (SCF).
Dette er en inntektsundersøkelse med opplysninger om mange inntektskilder, men den mangler opplysninger om f.eks. arbeidsløshet, skattefradrag og forbruksutgifter.

2.Taxation Statistics.
Dette er basert på et tre prosent utvalg av selvangivelser. (Dette er det samme datamaterialet som brukes i Finansdepartementets skattemodell).

3. Unemployment Insurance.
Dette er et ett-prosent utvalg av arbeidsløshetsdata fra Department of Employment and Immigration. Data blir hentet fra et administrativt dataregister. (Data blir også brukt av departementet selv i en egen mikrosimuleringsmodell).

4. Family Expenditure Survey (FAMEX).
Dette er en forbruksundersøkelse med forbruksdata for ialt 50 varegrupper. Utvalget består av ialt 10000 husholdninger.

I tillegg kommer en femte datakilde, nemlig tall fra nasjonalregnskapet.

Disse datakildene blir ved ulike "statistiske metoder og koblinger" bearbeidet og stilt sammen som ett syntetisk sammenkoblet datasett. Det kan, når vi skal vurdere valg av metode, være verdt å minne om at de i Canada ikke har et sentralt og gjennomgående personnummersystem. I tillegg til det, kunne en merke at selv når det finnes brukbare koblingsmuligheter ved å bruke et skatteidentifikasjonsnummer, så er det betydelige legale/politiske restriksjoner på direkte koblinger av individ-data. Gjennom den statistiske kobling som de foretar i SPSD, kan de imidlertid frigjøre datasettet også for eksternt bruk.

Følgende metoder er i bruk:

1. Randomisering.
Dette er en metode som brukes for å tilsløre enkelte av individkjennetegnene som hvis de stilles sammen, kan gi mulighet for personidentifisering. Metoden har bare som formål å gjøre datasettet tilgjengelig for allmennheten. (En form for kryptering av personkjennetegn.)

2.Integrated Weighting.
Dette er en metode som skal redusere "bias", og en tvinger ved dette utvalgene til å stemme med kjente kontroll-totaler. Det gjøres for eksempel ved å justere utvalgsvektene slik at utvalget stemmer med kjente befolkningstotaler (befolkning etter kjønn og alder).

Denne metode anvendes i Canada på et problem som er mye til felles med problemstillinger som vi har sett på i forbindelse med LOTTE. Metodegruppa hos oss (Heldal) har arbeidet med dette. Det er i diskusjonen av vårt utredningsarbeid pekt på at slike metoder kan bidra til å svekke utvalgets egenskaper og slik sett øke feilen for andre variable enn de som det veies for. Det var derfor av interesse å se noe nærmere på de vurderinger som de i Canada hadde gjort av metodene. Den person som hadde arbeidet mest med dette, George Lemaitre hadde for tiden permisjon. Jeg fikk imidlertid med meg endel litteratur og notater om dette. (Vil delvis bli omtalt senere i rapporten.)

3. Stochastic Imputation.
Dette er en metode for å å generere syntetiske kjennetegnsverdier for individer på et grunndatasett. Kjennetegnsverdiene blir hentet fra andre utvalg. Det finnes flere teknikker for dette, og den metoden som her ble brukt baserer seg på tilfeldig trekking fra det andre datasettet. Denne metoden brukes når det andre datasettet foreligger som frekvensfordelinger, og ikke som egentlige individdata. (Ved individata brukes Categorical Matching, se nedenfor.)

4. Micro Record Aggregation.
Dette er en metode som brukes til å lage syntetiske individ-records fra et totalregister, ved å slå sammen (lage et gjennomsnittsindivid) av for eksempel grupper av høyinntektstakere. En bruker totaltellinger fra registerne og ved denne metoden lager en syntetiske høyinntektstakere som dermed ikke lar seg identifisere, men som ligner på virkelige mikrodata. Dette blir en form for anonymisering av individdata.

5. Categorical Matching.
Dette er en syntetisk koblingsmetode som ikke bygger på tilfeldig trekking, men hvor metoden genererer deterministisk hvilke personer fra to uavhengige utvalg som skal sies å være samme person, dvs. som skal kobles sammen til en syntetisk person. Metoden bygger på at det i begge datasettene finnes en del felles opplysninger - samtidig som noen opplysninger bare finnes i hvert av utvalgene. I begge datasettene klassifiseres individene i grupper, basert på viktige kjennetegn, som boligtype/eierforhold, sysselsettingsstatus, inntektsklasse (?). Etter denne inndeling i delutvalg (bin), bestemmes en algoritme som innenfor hvert delutvalg finner, hvilke personer som er mest like. Disse individer sies så å være ett og samme individ.

Denne metoden brukes bl.a. til å slå sammen utvalgene for Inntektsundersøkelsen og Forbruksundersøkelsen i databasen SPSD.

6.Conversion.

Dette er en metode som brukes for å justere for underrapportering av enkelte kjennetegnsverdier. Det dreier seg for eksempel om underrapportering av arbeidsløshetsstønad og ulike andre former for "welfare-benefits". Metoden bygger på en modell for partielt frafall, slik at for enkelte individer rettes det opp fra null til positive verdier på f.eks arbeidsløshetsstønad. Det forandres ikke på kjennetegns-verdier for de som allerede har positive verdier.

7. Justering av framskrivingsfaktorer.
I tillegg til disse metodene har en også mulighet til å sikre konsistens i forhold til nasjonalregnskapsdata for senere år enn basisårer, ved justering av framskrivingsfaktorene. Dette er en metode som brukes ved beregninger for senere år enn det som datasettet skriver seg fra.

Basisdatasettet er hentet fra Survey of Consumer Finances (SCF). Denne undersøkelsen inneholder opplysninger om alle personene i husholdningene og har opplysninger om, demografi, arbeidsstyrkestatus, inntekter, kjennetegn ved boligen. Utvalget er på 36000 husholdninger som tilsvarer 98000 individer. Etter hele tilrettelegging en, ender en opp med en modellbefolkning som inneholder:
I alt 98000 individer fra ialt 38000 husholdninger, med kjennetegn som:

- Husholdningens sammensetning.
- Individuelle sosiale kjenntegn.
- Inntektskomponenter fra markedet.
- Inntektskomponenter, overføringer fra det offentlige.
- Inntektskomponenter, skattereduksjoner (spesielle skatteregler for bestemte sosiale grupper.)
- Inntektskomponenter, utgifter som går til fradrag i inntekten.
- Inntektskomponent, skatter.
- Kjennetegn ved boligen.
- Arbeidsløshetsdata.
- Forbruksutgifter fordelt på varegrupper.

I alt er det 350 variable som innngår i databasen.

3.3. Mikrodata og inntektsregnskapet.

Av flere grunner er det ønskelig å ha et individmateriale som er konsistent med inntektsregnskapet i nasjonalregnskapet. En av grunnene er at modellen skal kunne brukes til simuleringer av effekter av ulike former for indirekte skatter. Den metoden som er benyttet er beskrevet i Adler og Wolfson (1988)

Datamaterialet (modellbefolkningen) inneholder ved siden av opplysninger om inntekter og skatter, husholdningens

utgifter fordelt på ialt 50 poster. Dette er utgifter til ulike varegrupper, samt renter, skatter etc. For å få en komplett regnskapsoppstilling for den enkelte husholdning, er det nødvendig å inkludere en post for formuesendringer.

Budsjettandelene (forbruksutgiftene) i mikromodellen er avstemt slik at de stemmer med nasjonalregnskapet. I Adler og Wolfson (1988) drøftes nærmere framgangsmåten for å bygge bro mellom mikrodatamaterialet som inneholder inntekter og forbruksutgifter, og nasjonalregnskapets inntektsregnskap og forbrukstall.

Selv om en i arbeidet med å bygge bro mellom mikro- og makro-data, tar nasjonalregnskapet som gitt, blir likevel først trinn i prosessen at en ut fra nasjonalregnskapet lager en makro tabell som nærmer seg de begreper som finnes i et individuelt inntektsregnskap. En slik korrigert nasjonalregnskapstabell, skal en så kunne framstille som en aggregat tabell fra mikrodataene. For å få til det, gjennomføres det også korreksjoner i mikrodataene.

For det første er det nødvendig å sikre lik sektoravgrensing, for eksempel lik behandling av personer på institusjon og av private organisasjoner. En foretar derfor i makrotallene korreksjoner (oppsplitting) slik at det er de samme sektorer som dekkes. Det er videre behov for en korreksjon av enkelte utgifts- og inntektsposter, siden regnskapsprinsippene er forskjellige. Dette gjelder for eksempel opplysninger om inntekter fra bolig, utgifter til varige forbruksgoder, privat bruk av offentlige tjenester (som er gratis eller sterkt subsidierte) og størrelser som er knyttet til pensjonssparing.

Som en uavhengig modell i forhold til mikrosimuleringsmodellen SPSD/M er det utviklet en kryssløpsmodell COMTAX for å estimere sisteleddseffekten av ulike indirekte avgifter som er lagt på andre størrelser enn konsum. Ratene fra kryssløpsmodellen brukes som input i mikrosimuleringsmodellen.

3.4 Politikkvariablene og modellsimuleringene.

Modellen beregner for hvert individ; skatter, overføringer og disponibel inntekt. Beregningsresultatene kan så analyseres og det kan beregnes aggregat-størrelser og fordelingsbeskrivelser. For de indirekte skatter er det en egen prosedyre.

I programmet legges det inn faktiske regler for skatter og overføringer for historiske år. Modellen kan brukes til å simulere effektene av endringer i regler for beregningsåret. Disse regelendringene spesifiseres eksogent av brukeren.

I modellen legges det inn detaljerte regler for:
- overføringer
- fastlegging av skattegrunnlag
- fradragsregler etc.
- skattesatser for de ulike skatter
- progresjonsgrenser bunnfradrag etc.
- regler for arbeidsløshetsstønad
- overføringer til ulike familietyper, (herunder det som tilsvarer barnetrygd)
- kompensasjonsregler for indirekte skatter
- alderspensjon
- garanterte minsteinntekter
- ulike former for sosiale støtteordninger (kan være regionale dvs. provins avhengig)

Modellen likner på vår modell LOTTE. En av forskjellene ligger i at de i Canada har inkludert flere regler for overføringer, f.eks alderstrygd og arbeidsløshetsstønad.

Modellen for indirekte skatter bygger på at det for hver enkelt husholdning i utvalget (modellbefolkningen) er estimert buddsjettandeler for rundt 40 varegrupper. Gjennom en kryssløpsberegning har en beregnet indirekte skatters andel av sluttkonsumet av disse 40 varegruppene, slik at ikke bare sisteleddsavgifter lar seg beregne. Dette betyr at en får beregnet proveny av de indirekte skatter samt budsjettkonsekvenser (kompensasjonsbeløp for hver enkelt). I den canadiske modellen er det ikke lagt inn estimerte etterspørselselastisiteter, noe som i prinsippet vil være mulig. I vår modell INSIDENS har det imidlertid vært estimerte priselastisiteter. Selv om disse har vært med i modellen har bruken av INSIDENS ofte vært begrenset til en modellversjon hvor en ikke aktiverer etterspørselselastisitetene.

Den nåværende fulle versjon av modellen koster CAN$ 5000. Modellen er pr sommeren 1989 solgt til 15 brukere. I prisen er ikke inkludert brukerstøtte. Slike avtaler vil eventuelt komme i tillegg.

3.5 EDB løsninger og datatilrettelegging.

I arbeidet med SPSD/M er det foretatt valg, som har gjort det mulig å presse en relativt stor modell og datamasse inn på PC format. Datasettet slik det foreligger som "flat file" fra stormaskinene (SAS-program), vil kreve 55 Mbytes. Inkluderer en de bearbeidinger som skjer i modellen, vil en ha en samlet størrelse på 119 Mbytes. Målet for systemarbeidet, var å få plass på en standard PC, som på det tidspunktet innebar en harddisc på 20 MB. For å få dette til valgte en å:

1. Ordne filen slik at husholdningsdata som er felles for alle medlemmene i en husholdning bare lagres en gang.

2. Data som er like for flere husholdninger, lagres bare en gang.

3. Intensiv sammenpressing av filer ved bruk av binærkoder. (Eks kjønn med to verdier lagres i en bit.)

4. Inntekts- og skattedata ble lagret i forskjellige record-typer, som bare ble tilordnet et individ for positive verdier på beløpene. Dette reduserer størrelsen betydelig da svært mange av inntektspostene er null for store andeler av utvalget. Det samme gjelder data for arbeidsløshet, som bare gjelder en liten andel.

5. Minimal permanent lagring av mellomregninger og avledede størrelser. Mange avledede tall må regnes ut hver gang de skal brukes.

Etter at denne kompresjon var gjennomført, var det mulig å presse hele datasettet inn på 5.25 MB. Med tillegg for noen andre data (framskrivingsforutsetninger) (0.75 MB) og selve programmet og programkode (skrevet i C) 1 MB, krever hele modellen slik den nå foreligger 7MB.

Det er også lagt vekt på å ordne datasettet slik at en kan bruke deler av det; henholdsvis 5%, 10%, og 25% som representative utvalg av totaldatasettet. (Den modellen jeg har fått med meg, inneholder 5% utvalget).

I SC var det gjennomgående slik at modellene var utviklet for kjøring og bruk på personlige datamskiner PC (IBM-kompatible). I arbeidet med å generere modellbefolkningen (SPSD) brukes det imidlertid stormaskiner (main-frame) og programmet SAS. Etter at de ulike veiinger, og statistiske og kategoriske koblinger er gjennomført, klargjøres data for PC.

Stort sett var som tidligere nevnt alt systemarbeid utført av de økonomer som var ansatt i gruppa. Filosofien var klar - at det var bedre å bruke økonomer med interesse for innholdet i modellen som også var interessert i å lære seg det nødvendige programverktøy, framfor å bruke ekstern systemhjelp. Selv om de skulle være ansatt i gruppa var Wolfson ikke sikker på om det var nødvendig å ansette rene systemeksperter. Enkelte veldefinerte problemer som var for vanskelige til at de kunne løse dem selv, hadde de satt bort til eksterne eksperter (jeg opfattet de da slik at de hadde kjøpt tjenester i markedet.)

Modellene er i liten grad bygd opp med menyvalg. Dette gjelder også den kommersielt anlagte SPSD/M. Jeg fikk prøve modellen selv under oppholdet og mitt inntrykk er at selv om skjermbildet er komplisert og det er mye pirkete parameterutfylling, lar det seg gjøre å finne fram ved bruk av manualen. Arbeidet med parameterutfylling kan

for den enkelte bruker standariseres, slik at ofte brukte parametersett var lett å bruke. En fordel ved å legge liten vekt på brukermenyene var såvidt jeg forsto, at en dermed fikk mer fleksible modeller. Strategien var klar - at en først skulle programmere innmaten i modellen, og deretter hvis det var regningssvarende legge på en innpakning som gjorde det mulig å menystyre hele eller deler av modellen. Det virker på meg som det av oss i SSB har blitt fulgt en noe annerledes strategi, ved at en starter med menyene og rammen, og deretter lager innmaten. (LOTTEII, MOSART OG LOTTO/ODIN er muligens eksempler på dette ?)

Modellen er utviklet på en Compaq Deskkpro 386/25. Det er imidlertid ikke nødvendig med denne maskintype. SPSD/M må kjøres på en IBM kompatibel maskin (MS DOS, versjon 3.0 eller høyere). Brukere som ønsker å forandre programkoden (C), må bruke en "text-editor" og Microsoft C kompilator versjon 5.1.

En full gjennomkjøring av modellen med hele datasettet bruker 8 minutter på de raskeste maskinene. Mer kompliserte simuleringer som også krever marginalskatteberegninger, gir lengre kjøretid (eksempler på 24 minutter). Andre og langsommere maskiner gir lengre kjøretid, 2 til 3 timer for ordinære AT maskiner ble nevnt.

Siden alle beregningene foretas i "floating point", vil en maskin som er utstyrt med en numerisk co-processor gjennomføre beregningene tre ganger så fort som en maskin uten dette. Maskinenes klokkefrekvens har også betydning for hastigheten. (Compaq maskinen som brukes mest i SC har en frekvens på 25 MHz).

I programpakken SPSD/M er det inkludert følgende program:

- Spreadsheet interface tools; som sikrer et grensesnitt mellom SPSD/M og regneark for å bearbeide og sammenligne standardtabeller som er beregnet ved en eller flere modellkjøringer.

-Program verktøy (programming tools); et brukerprogram som gjør det enklere å finne en programstreng (grep), sammenligning av to tekstfiler (fcomp), og en versjon av Unix utility programmet "make".

Det har hittil ikke vært etterspørsel etter modellversjoner som er tilrettelagt for stormaskin, selv om det i prinsippet er mulig å overføre modellen til alle maskiner som har en C kompilator. Modellen har vært overført til IBM system 3090, og det har vært arbeidet med å lage en versjon for UNIX minimaskiner.

3.6 En sammenligning av mikoromodeller for PC og stormaskin.

Wolfson var med som ekspert i en arbeidsgruppe under Research Council i USA, som skal evaluere mikrosimuleringemodeller for "social policy", herunder skattesimuleringsmodeller.

(Medlemmer av denne komiteen er blant annet; Michael Wolfson, Gary Burtless (Brookings Inst.), Robert Moffit, Paul Cotton (Fulcrum Techn.),George Sadowsky (Northwestern Univ.), Robert D. Strauss (Carnegie-Mellon Univ.), Thomas Grannemann, Eric Hanusheh.
Komiteens adresse: Panel to Evaluate Microsimulation Models for Social Welfare Programs
National Research Council.
Commission on Behavioral and Social Sciences and Education.
2101 Constitution Avenue
Washington DC 20418-
Sekretær: Constance Citro)

Denne komiteen har laget en god del nyttig bakgrunnsmateriale som kan ha interesse for oss og NORAS skatteforskningsprogrammet.

Et grunnlagsdokumentene for denne arbeidsgruppen, omhandler edb spørsmål. I dette interne arbeidsdokument. er det foretatt en evaluering av mikrosimuleringsmodeller, med en sammenligning av SPSD/M med modellen TRIM. TRIM er den offisielle skattemodellen i Finansdepartementet i USA. TRIM er en mikrosimuleringsmodell som ligner på SPSD/M, bortsett fra at den er skrevet for stormaskin.

Sammenligningen tar for seg:

Modellkjøring.
SPSD/M kjøres interaktivt, mens TRIM kjøres batch.

Tid og kostnader.
Sammeligningen baserer seg på "wall clock time" dvs. tid for å gjennomføre hele beregningen. Konklusjonen er at SPSD/M kan opereres raskere og billigere enn stormaskin modellen TRIM.

Utviklingsmuligheter.
Programmeringsverktøyene som utvikles for PC er langt mer rettet mot analyse og brukerutvikling, enn det som skjer for stormaskin. Her er utviklingen mer rettet mot hvordan en kan operere mange brukere på en effektiv måte til samme tid.

Overførbarhet.
Dette blir tillagt vekt og konklusjonene er at SPSD/M har et bedre utgangspunkt, siden de har valgt programspråket C.

Modell-parametre.
Både SPSD/M og TRIM bruker et stort antall parametre for en kjøring (over 400).

Teknologisk utvikling.
Det har vært en rivende utvikling i tilbudet av maskiner og programvare. Denne utvikling regner en med vil fortsette. Spørsmålet som bør stilles er hvilke konsekvenser denne utvikling bør ha for strategivalg med hensyn på videreutvikling av mikrosimuleringsmodeller. For oss som i beskjeden grad har tatt i bruk de nye tilbud av maskiner og programvare som er blitt utviklet de siste årene, blir også spørsmålet hvordan vi skal tilpasse oss til den allerede eksisterende teknologi når en samtidig tar hensyn til framtidsvyene.

Konklusjonene er ikke oppsiktsvekkende:
-det vil være lett å skaffe seg rimelig datakraft for krevende modeller.
-intern minnene vil være store nok til å operere kompliserte simuleringsmodeller på en enklere måte enn idag, ved reduksjon i innlesning og lagringsprosedyrer.
-det vil fortsatt være behov for data-kompresjon
-bedre lagringsmedium vil være godt tilpasset behovet for omfattende modellbefolkninger.

Framtidsvyer.
Den teknologiske utvikling ligger vel til rette for videreutvikling av skatte-simuleringsmodeller av statisk type (LOTTE). Program og maskiner vil imidlertid også gjøre dynamiske simuleringsmodeller lettere å utvikle.

Den viktigste utviklingslinjen ligger i videreutvikling av modeller for skrivebordsmaskiner (desktop)
Slike maskiner vil ventelig være i stand til å håndtere store datamengder og gjennomføre kompliserte simuleringer. Det regnes med at på mellomlang sikt vil det riktige være å utvikle modeller med tanke på denne maskintype.

3.7. Atferd i mikrosimuleringsmodeller.

Spørsmålet om inkludere atferdsrelasjoner i simuleringsmodeller ble endel diskutert. Wolfson var noe skeptisk til å inkludere atferd i de modellene han arbeidet med, ikke fordi han ikke så verdien av det, men nok mer fordi han var redd for å spre ressursene for mye. Han viste forøvrig til arbeidet i US: ekspertkomiteen. Her er det skrevet flere notater som drøfter dette, blant

annet et av Gary Burtless, Brookings Inst. Endel av sysnpunktene går på at det i og for seg er viktig å ta hensyn til atferdsendringer som en del av politikksimuleringene, men at en gjennomgang av den økonometriske litteratur indikerer at det er mange uløste problemer og at estimatene av sentrale elastisiteter ikke virker robuste. Det argumenteres derfor for økt økonometrisk innsats, men at en også vil ha bruk for simuleringsmodeller hvor en ikke pretenderer å inkludere atferd. I disse notatene drøftes foruten atferd i arbeidsmarkedet også atferdsrelasjoner for helse, pensjon, og finansiell atferd.

I Statistics Canada er det skrevet et notat av professor Lars Osberg som også drøfter dette med atferd i simuleringsmodellene (Osberg 1986). Osberg mener at selv om den statiske arbeidstilbudsmodellen har mange svakheter har det skjedd såpass mye på det økonometriske området at det bør være en klar strategi å inkludere arbeidstilbudsrelasjoner i skattesimuleringsmodeller. Anbefalingene går på å inkludere tilbudselastisiteter og prøve ut ulike verdier på elastsisitetene. Det synes klart at et komplett estimerings og simulerings opplegg som i GATO ikke er vurdert. (Det er åpenbart viktig å få sider ved GATO modellen bedre beskrevet i den internasjonale litteratur.)

Osberg foreslår at de som har ansvaret for å utvikle skattesimuleringsmodeller skal tilpasse seg ulike brukerbehov. Noen brukere og noen problemstillinger er mest tjent med simuleringsmodeller som er enkle i den forstand at de ikke inkluderer atferd. Andre mer langsiktige og dyptpløyende utredningsoppgaver krever derimot modeller som bygger på økonometrisk fastlagte atferdsrelasjoner. Det kan imidlertid fort bli slik at de kortsiktige oppgaver og de enkle modeller får en så stor etterspørsel at det blir lite tid til utvikling av atferdsbaserte modeller. Osberg foreslår derfor at Statistics Canada skal prise de populære tjenester på de enkle modellene, slik at etterspørselen dempes og at det kan skaffes midler fra salgsinntektene til mer langsiktig modellutvikling. Dette er en strategi som også vi bør vurdere å følge.

3.8 Vurderinger.

Det er naturlig å sammenligne SPSD/M med LOTTEII. Modellene har mye til felles med detaljerte skattesimuleringer sammen med et datamateriale. I SPSD/M har de imidlertid gått lenger i flere retninger.

De har inkludert simuleringer av flere regelendringer i modellen. Dette gjelder f.eks. det som tilsvarer arbeidsløshetsstønad og alderspensjon. Vi har også disse inntektskomponentene med i LOTTE og kan rett fram simulere konsekvensene for skatt og disponibel inntekt av

skatteregelendringer. I SPSD/M er det imidlertid også lagt inn parametre som bestemmer selve bruttooverføringen fra det offentlige. I LOTTE har vi ikke ferdiglaget slike rutiner. Spørsmålet er om vi kan klare å utvikle en algoritme som ut fra person/husholdningskjennetegn kan regne ut størrelsen på inntektsoverføringen slik at vi for eksempel kan se hvordan reguleringer av Folketrygdens grunnbeløp (G) eller andre typer av politikkendringer, virker på proveny og på inntektsfordelingen. Tilsvarende algoritme kan lages for arbeidsløshet. I den canadiske modellen har de lagt inn som parametre en rekke størrelser som beskriver både arbeidsløshetsnivå og regler som bestemmer arbeidsløshetstrygd og utbetaling til den enkelte. (I den canadiske modell finnes grenser for antall uker, krav til tidligere inntekt, utbetaling i forhold til tidligere inntekt.)

I den norske Levekårsundersøkelsen, finnes det noen opplysninger om arbeidsløshet. De er imidlertid neppe tilstrekkelige til å gi grunnlag for en eksakt utregning av arbeidsløshetsstønad. Det er mulig at vi etterhvert gjennom SOFA registeret direkte kan framskaffe tilstrekkelig med opplysninger til at vi kan simulere fordelings og provenyeffekter av nivåendringer på arbeidsløsheten og av regelendringer for arbeidsløshetsstønad.

Andre overføringer som er med i SPSD/M er overføringer til familien. I LOTTE har vi med noen slike. Det som foreligger ferdig, er endringer i skattebehandlingen av slike overføringer. Vi har imidlertid ikke brukt modellen til simuleringer av regelendringer utover dette. Det kunne for eksempel ha vært spørsmål om å beregne fordeling og provenyeffekter av et høyere barnetall. Vi måtte da ha utviklet en algoritme som kan simulere en ny fordeling av barn/foreldre og deretter foreta simuleringer av fordelings og provenyeffekter gitt regelverket.

SPSD/M opereres på en måte som har mye til felles med LOTTE. En legger inn parametre som sikrer framskriving av beløpsverdier basert på pris/lønnsvekster og en foretar beregning av ett referansealternativ. En kan foreta beregninger på deler av utvalget, f.eks bestemte sosioøkonomiske grupper. Resultatene kan enten tas ut ved 10 standardtabeller, eller en kan spesifisere egne tabeller ved å bruke et innebygget tabellprogram (X-tab). Resultatene kan også legges ut til videre analyse og bearbeiding i SAS eller regneark.

SPSD/M er utviklet med tanke på å ta vare på to typer av endringer i skatteregler. Det kan for det første gjøres ved å endre verdien på allerede eksisterende parametre i modellen. Dette kan gjøres uten å kjenne selve programmet, dvs bare ved å endre parameterverdien, mens selve

programmet er ukjent. Denne type simuleringer er derfor referert til som black-box tilnærming. Den andre typen av simuleringer krever endringer som forutsetter helt nye inntektsbegrep eller nye skatteformer. Dette krever programendringer og denne tilnærming er omtalt som glass box. En har her benyttet programverktøy som er tilgjengelig under UNIX C (make og makefile). Vi har også denne muligheten i vår modell, selv om det ikke er formalisert på samme måte. Behovet for det kan imidlertid melde seg når /hvis modellen spres til eksterne brukere.

I SPSD/M er det ved hjelp av kryssløpstabeller beregnet effektive indirekte skattesatser. Disse anvendes sammen med en vektor for hvert hushold. COMTAX (the commodity tax input/output model) regner altså indirekte skatter som er betalt av bedriftene om til effektive "omsetningsavgifter" på sisteledd. For oss vil det være et spørsmål om hvordan best utnytte kryssløpsmodellene MODIS/MODAG.

Det er mange felles trekk i modelltilpasningnen noe som gjør at vi har noe å lære av den canadiske tilpasningen:

-Vi må videreutvikle våre EDB løsninger i retning av økt PC bruk.
-Vi må i SSB ha en samlet strategi for å utvikle en modellbefolkning. Dette krever koordinering av datainnsamling og sammenstilling av individdata i en modellbefolkning, som også er rimelig konsistent med nasjonalregnskapet. (Mer om dette i kap.6.)
-Lage en felles modell for direkte og indirekte skatter.
-Utvide simuleringsmodellen med flere politikkvariable enn skatter. I første rekke ulike overføringer fra det offentlige.

4. DYNAMISKE MODELLER - FORLØPSMODELLER.

4.1. Innledning.

I gruppen "Social and Economic Studies" arbeidet de med flere prosjekter som er rettet mot forløpsmodellering. Modellene tar sikte på å belyse anvendte problemstillinger for eksempel knyttet til omlegginger av trygdesystemet, helse og eldreomsorg. Modellene har sålangt ikke blitt utviklet som ferdige salgbare produkter, men har gradvis utviklet seg fra manuell trekking av livsløp basert på estimerte overgangssannsynligheter, til relativt brukervennlige programpakker. Arbeidet bygger mye på å utnytte tilgjengelige resultater fra ulike typer av forskningsrapporter som.

4.2 DEMOGEN - En demografisk simuleringsmodell med yrkesdeltaking og inntekt.

Arbeidet med dynamiske simuleringsmodeller startet med modellen DEMOGEN i 1983 som et utredningsarbeid for en parlamentskomite. Effekter for samfunnet av skilsmissene, var en av de første problemstillinger som det ble arbeidet med. Som en del av modellarbeidet er det gjennomført demografiske analyser, basert på hazardratemodeller. Dette kan en se på som en måte å estimere parametre i en grunnleggende atferdsmodell. Med basis i denne modelltype beregnes det ettårige overgangssannsynligheter. De dynamiske simuleringsmodellene er basert på trekking av overganger/ tilstander i ettårige tidsrom.

De første modellutgavene var nærmest basert på håndtrekking av tilstander. Etterhvert er det utviklet flere modellversjoner skrevet for PC. Et felles programspråk for alle modellene er programmet C (eller utgaver av dette).

Modellen DEMOGEN benytter ikke en minipopulasjon som utgangspunkt for simuleringene. Det er derimot modellen som simulerer komplette livsløp basert på de innleste demografiske parametre (sannsynligheter for demografiske begivenheter). Disse demografiske overgangssannsynligheter er bestemt ut fra faktiske begivenheter. Den populasjon som vi får, ved å simulere en kohort basert på konstante demografiske rater svarer ikke til noen faktisk eksisterende befolkning. En kan si at hvis de demografiske rater er perioderater fra observasjonsåret, vil den simulerte befolkning på en måte tilsvare den stabile befolkning en får ved disse ratene. Det er derfor ikke lett å foreta en empirisk evaluering av denne type modell-befolkning ved å sammenligne med observerte befolkningstall.

I modellen leses det inn sett av overgangssannsynligheter (betingede).

Følgende prosesser/begivenheter inngår:

Begivenhet	**Overgang betinget av**
Død	alder,kjønn
Første ekteskap	alder
Ektemanns alder	konas alder ved første ekteskap
Fruktbarhet	alder, paritet
"Custody"	ekteskapelig status
Barn; fraflytting	alder, paritet
Skilsmisse	alder, ekteskapets varighet, barn tilstede i hushold,giftermålsalder.
Gjengifte	alder, kjønn, skilt eller enke
Andre ektefelles alder	alder, kjønn, ektesk.status
Utdanning nivå og fagfelt	alder, kjønn
Yrkesdeltaking (hvert år)	alder, kjønn, ektsk.status, barnetall, utd.nivå, varighet i yrke,
Arbeidsinntekt	alder, kjønn, yrkeshistorie.

Simuleringen skjer ved at en lager livsforløp ett for ett. En lager et par - mann og kvinne. Det første som en bestemmer, er alder ved død, basert på generering av tilfeldige tall som kan gjenskape fordelingen av samlet levealder. Deretter bestemmes alder ved første giftermål. Sannsynligheten for giftermål avhenger av alder. Metoden er laget slik at ikke alle blir gift. Hvis kvinnen blir gift, trekkes et tall fra en fordeling som bestemmer fødselsåret til mannen. I utgangspunktet bestemmes forløpet til en kohort av kvinner og ikke-gifte menn. Gifte menns fødselsår vil avvike stokastisk fra dette.

Strukturen i modellen er bygd opp omkring sekvensielle

begivenheter. Etter død er det ingen flere begivenheter, skilsmisse kan bare komme etter giftermål osv. Fruktbarhet er imidlertid bare avhengig av alder og paritet og ikke av ekteskapelig status. Ekteskapelig status bestemmes av en relativt detaljert modell hvor det er estimert hazardrate-modeller basert på Family history survey 1983. Dette er en retrospektiv undersøkelse for yrkeshistorie, giftermål og fruktbarhet. Det er lagt ned relativt mye arbeid i denne estimeringen (Rowe 1986, og upublisert manuskript).

Arbeidsinntekten blir bestemt ved at en først bestemmer antall år under utdanning, samt bestemmer utdanningens fagfelt. Yrkesdeltakingen bestemmes år for år basert på estimerte overgangsrater inn og ut av yrkeslivet. Også disse ratene for bevegelsene inn og ut av yrkeslivet er estimert ved hjelp av data fra den retrospektive familie og yrkesundersøkelsen (Family history survey). En har forsøkt å bygge opp en tilnærmet økonometrisk yrkesdeltakingsmodell. Det ser imidlertid ikke ut som en har ambisjoner om å inkludere skatter i denne modellen (Picot 1986). Selve arbeidsinntekten bestemmes ut fra de simulerte opplysningene om utdanning og yrkesforløp, ved ytterligere en stokastisk simulering (Kennedy 1986).

4.3 LIPPS - en inntekt og pensjonsmodell. (Lifetime Income and Pension Policy Simulation).

Denne modellen bygger på de livsforløp som er simulert i DEMOGEN. Modellen simulerer skatter og trygdepremier som betales, og stønader som hvert enkelt individ får fra de offentlige trygdeordninger. Disse beregningene bygger på simuleringer av tilstander som bestemmer skatter som skal betales og stønader som mottas, uten at modellen inkluderer påvirkning på atferden fra skatter og pensjoner.

I modellen LIPPS bygger beregningen på forutsetninger om:
- demografi, basert på resultatene fra DEMOGEN.
- trygderegler og andre politikkvariable
- makroøkonomisk utvikling (priser og inntektsvekst)

Dette modellsystemet er blant annet brukt til en analyse av en omlegging av trygdesystemet, som skal gi trygderettigheter for husarbeid, særlig omsorgsarbeid. Modellen viser at ikke alle konsekvenser av slike trygderformer er like lett å gjennomskue uten modellsimuerlinger (Wolfson 1988).

4.4 CEPHID - eldre, helse inntekt og demografi (Canada's Elderly -Projecting Health, Income and Demography.

Dette er en mikrosimuleringsmodell som bygger på DEMOGEN. Modellen inneholder en relativt grov demografisk modell.

Den inkluderer imidlertid en relativt detaljert beskrivelse av familietilknytning.

I denne modellen er det lagt mye vekt på modellering av faktorer som kan påvirke helse, faktiske sykdommer og behandling og behandlingsutgifter. (Jfr Wolfsons bidrag til IARIW konferansen 1989). Den modellen som de har utviklet kan få en lang rekke anvendelser.

Modellen simulerer:
- ekteskapshistorie
- fruktbarhetshistorie
- yrkesdeltaking
- inntekt

Etter at disse kjenntegnene er simulert, blir det simulert kjennetegn ved individene som er av betydning for helse. Eksempler på slike faktorer er vekt og røyking. Etter det simuleres det sykdomstilfeller for et utvalg /grupper av sykdommer. Neste trinn i simuleringen er simulering av den medisinske behandling. Ut fra denne simuleres dødsfall, basert på behandlingsavhengige dødsfallsrater. For hvert individ kan en simulere offentlige utgifter til pensjon og helsepleie.

Også denne modellen som er laget for PC bruk, er tilpasset brukernes behov. Modellen slik jeg så den inneholder ikke menystyring. Den er som andre PC-modeller interaktiv, og med utfyling av mange parameterkort.

Modellen kan brukes til å simulere effektene av ulike politikkpakker som for eksempel gir forskjellig vekt til kurativ og forebyggende innsats.

Statistics Canada arbeider lite med framskrivinger eller prognoser, som en erklært politikk. Modellene som SES arbeidet sammen med de demografiske framskrivinger var likevel et velegnet modellverktøy for å beskrive konsekvensene for offentlige utgifter av eldrebølgen. I en artikkel om dette av Ivan Fellegi (1988) anvendes flere av modellene.

5. BEDRIFTSKATT.

Det ble liten tid til å drøfte arbeid med bedriftsskattemodeller men de har ikke hatt mye arbeid på dette felt, men regner nå å starte et prosjekt om bedriftsskatt. Wolfson har selv arbeidet noe med dette (Wolfson 1987) og foretatt en analyse på et mikromateriale, av om bedriftskattene bidrar til en økende andel store bedrifter. Wolfson ser på effektive skattesatser (empiriske) og hvordan disse varierer med bedriftens størrelse. Det er en tendens til at de effektive skatter først øker med størrelsen for så å avta for de aller største bedrifter.

Jeg ble kjent med at det hadde vært en konferanse om "Tax Microsimulation Modelling for Business", 11-12 mai 1989 arrangert av Department of the Treasury, Internal Revenue Service, USA. Konferansepapirene fra denne konferansen inneholder interesante arbeider av interesse for oss i vårt arbeid med mikro modeller for bedriftsskatter. konferanserapporten.)

6. OM STATISTISK KOBLING OG EN SKISSE TIL EN NORSK MODELLBEFOLKNING.

6.1 Statistisk kobling.

I Canada er det strenge regler for indivikobling samtidig som mulighetene er mindre enn i Norge siden de ikke har et gjennomgående personnummersystem. I arbeidet med modellbefolkningen SPSD er det tatt i bruk flere metoder som kan ha interesse for oss selv om vi kanskje på noen punkter er bedre rustet til direkte koblinger enn det de er.

Et syntetisk individmateriale basert på en statistisk kobling inneholder ikke like mye informasjon som et komplett individmateriale. Det kan derfor være nyttig å trekke fram noe av det canadiske arbeid med evaluering av metodene, med tanke på en eventuell anvendelse hos oss.

Problemstilingen kan beskrives som følger. Vi kan tenke oss at det er tre variable (X,Y og Z) og to datamaterialer (A og B).(Egentlig er det flere kjennetegn dvs kjennetegns-vektorer.) I materiale A, finnes X og Y, mens det i materiale B finnes X og Z. Problemet består da i å etablere en file C, som en individfile for individene som er med i file A, hvor det finnes opplysninger om (X,Y og Z)

Det kan være grunn til å skille mellom materiale som er totaltellinger og materiale som er basert på utvalg av populasjonen. Hvis det er totaltellinger, og vi som i Norge har personnummer er koblingen kurant. I Canada har en av og til totalmateriale uten å ha personnummer. Da blir opppgaven å utvikle metoder hvor en har som ambisjon å finne kombinasjoner av kjennetegn som så langt som mulig tilsvarer det en får ved en reell kobling (tilsvarende den en får med personnummer.) Jeg besøkte også en enhet i SC som arbeidet med dette for statistikk formål. Her etablerte de en individfile med selvangivelsesopplysninger for hele skattyterpopulasjonen i Canada. Samlede selvangivelsesopplysninger forelå på dataregistrert for alle, ikke bare for et utvalg.

Jeg vil i det følgende se litt på den problemstilling vi har når vi har to utvalgsundersøkelser som inneholder noe felles informasjon (X) og noe atskilt informasjon (henholdsvis Y og Z). Noe av den felles informasjon kan

for Norge være registerkjennetegn som vi kan hente ut gjennom personnummer og totalregistre.

Noen ganger kan en imidlertid også ha en situasjon hvor det i en tredje datafile (D), finnes samtidige observasjoner av Y og Z. Dette kunne for eksempel skrive seg fra et langt mindre utvalg eller en tidligere undersøkelse. En står da overfor et imputeringsproblem. Problemet vil bestå i å etablere sammenhenger mellom X,Y og Z i det komplette materiale, som kan brukes til imputering (predikering) av verdien i de datamaterialene som ikke er komplette (A og B).

Når vi har konstruert den syntetiske individfilen C kan det være flere typer av feil.

-skjevheter i fordelingen av Z
-skjevheter i den simultane fordelingen av (X,Z)
-skjevheter i den simultane fordelingen av (X,Y,Z)

I Canada er det arbeidet noe med Monte Carlo simuleringer for å studere feil ved ulike metoder for statistisk kobling.

Det kan bli skjevheter i fordelingen av kjenntegnverdiene Z som etableres på file C, for eksempel hvis metoden for statistisk kobling er slik at ett individ fra file B kan kobles til flere individer på file A. Det kan derfor være hensiktsmessig å pålegge koblingen en restriksjon slik at en unngår dette. Den metoden som er valgt i konstruksjonen av SPSD, gir tilnærmet denne egenskap. Metoden er flertrinns, og går i korthet ut på å sortere både file A og B i grupper etter kjennetegn i X vektoren, untatt ett av kjennetegnene X_1 Innenfor hver av gruppene kobles filene A og B ved å sortere etter størrelsen på X_1 De personer som har høyest verdi i hver av filene, kobles så sammen, osv. Hvis filene ikke har lik størrelse (antall personer) trengs en prosdyre som sikrer det. Enten kan en duplikere individ i den filen med færrest personer, eller en kan fjerne individer fra den største filen. Prosedyrene for dette må gjennomføres slik at en ikke får skjevheter i marginalfordelingene.

I metoden nevnt foran, bruker en bare en sorteringsvariabel X_1. Det er også metoder som bruker flere slike variable, hvor koblingsmetoden på en eller annen multivariat måte finner de makkere i de to filene A og B som passer best. Slike metoder kan gi et bedre resultat, men også være mer ressurskrevende. I Canada har en som nevnt arbeidet noe med evaluering av slike metoder. I følge resultatene fra denne evaluering, er den metoden som benyttes i SPSD en god metode. Egenskapene til en slik kobling basert på rangnummer etter sortering, avhenger av korrelasjon mellom sorteringsvariabelen X_1 og de kjennetegn som er unike for hver av filene A og B. I SPSD bruker en total inntekt som sorteringsvariabel slik at det er viktig

for metodens egenskaper om en i den koblede filen begrenser seg til analyse av kjennetegn som er sterkt korrelert til inntekt. Selv om den eksisterende metode i SPSD falt heldig ut, er det metoder som antagelig vil kunne gi ennå bedre resultater. Det som nevnes er såkalte log-lineære koblingssmetoder som det anbefales å arbeide videre med.

6.2 Statistisk kobling hos oss ?

Jeg skal kort se noe på muligheten for å utnytte statistisk kobling og behovet for det ved utvikling av en modellbefolkning for Norge.

Vi tar utgangspunkt i at vi ønsker en modellbefolkning som bygger på Inntektsundersøkelsen (IU), dvs at vi skal ha selvangivelsesopplysninger for utvalget. For et begrenset utvalg eksisterer det både forbruksskjema og IU opplysninger. Dette er paneldelen på ca 200 husholdninger. For dette utvalget har vi mulighet til å sammenligne forbruksutgifter fra FU med inntektsbegrep fra IU.Dette gir flere muligheter til blant annet å utprøve ulike koblingsmetoder.

En skisse av innholdet i en norsk modellbefolkning.

	Inntekts-undersøkelsen.	Forbruks-undersøkelsen	Levekårs-undersøkelsen.
Utvalgsstørrelse:			
hushold:	4500	1500	
personer:			5102
forbruk			
panel:	200	200	
Kjenne-tegn:			
Pensjg.innt	x	x	x
Netto inntekt	x	x	x
Skatteklasse	x	x	x
Kjønn/alder	x	x	x
Barn	x	x	x
Ektesk. status	x	x	x
Sosioøk status	x	x	x
Skatt	x	x	x
Andre register Kjenntegn:			
Utdanning	x	x	x
Studielån	x	x	x
Trygder	x	x	x
Arbeidsløsh.	x	x	x
Sosialhjelp	x	x	x
Detaljert			
inntektspsost	y		y
Fradrag	y		y
Formue	y		y
Forbruksutgift			
ialt		z	
varegruppe		z	
boligtype		z	z
Branntakst		z	
Anskaffelsesverdi		z	z
Gjeldsrente bolig		z	z
Husleie		z	z
Fritidshus			
-utgifter		z	
verdi, anskaffelses		z	z
Privatbil		z	z
-kjøpsverdi		z	
-årgang		z	
-kjørelengde		z	
-herav yrke		z	
Varige gjenstander:			
-campingtilh.		z	(z)
-motorsykkel		z	(z)
-båt		z	(z)
-bil		z	(z)
-TV		z	
-Video		z	(z)
-fryser		z	(z)
-oppvaskmaskin		z	(z)
-vaskemaskin		z	(z)
Inntkg arbeid		z	z
Arbeidsreise		z	z
Timelønn			z
Skiftarbeid			z
Omsorgsarbeid			z
Næring		z	z
Yrke		z	z
Bosted	x	z	z
Mottar:			
-hjemmehjelp			z
-hjemmesykepleie			z
Arv/gave			z
Sykdom/helse			z
Helseutgift		z	
Barnehage		z	z

Dette kan være en løs skisse over variabeltyper som kan inngå i en slik modellbefolkning. Utvalget av variable er basert på et raskt skjønn av hva som kan ha interesse ved politikksimuleringer. Utvalget av bakgrunnsvariable kan gjøres større.

Utvalgsplanen for Inntektsundersøkelsen er slik at den har felles utvalg med Levekårsundersøkelsen og for en liten del også med Forbruksundersøkelsen (paneldelen). Det betyr at vi kan si at LU og IU utgjør en personkoblet file. Vi må imidlertid ta hensyn til at LU er en personundersøkelse og at vi må lage en forbindelse mellom personopplysninger i LU og husholdningsoplysninger ellers. Dette har vi noe erfaring med fra et arbeid med å overføre boligverdiopplysninger fra LU inn i LOTTE utvalget.

Den statistiske kobling kan gjennomføres ved først å tredoble forbruksutvalget (lage tre identiske hushold av hvert hushold). Neste trinn er å gruppere etter bakgrunnskjennetegn, som alder, familietype, bosted, sosioøkonomisk status, boligtype osv. Antall grupper skal imidlertid ikke i første omgang gjøres for stort. Deretter sorteres husholdene etter pensjonsgivende inntekt. Siden det er flere personer kan en starte med å sortere etter samlet inntekt, men eventuelt utvikle en metode som bygger på en multivariabel metode hvor en utnytter opplysninger om de enkelte husholdningsmedlemmers inntekt.

Neste trinn bør være en avstemming mot makrotall. Første bør en gjennomgå individrecordene og etablere et personlig innteksregnskap som er tilnærmet slik at sum over alle individer gir tall som begrepsmesig er konsistent med inntektsregnsakpet (eventuelt i modifisert utgave). Konsistens med nasjonalregnskapsutviklingen sikres ved valg av framskrivingsvekter.

Neste trinn blir å videreutvikle den modelldel som nå ligger i LOTTE II. Første ledd i dette bør være å utvikle en modul for indirekte skatter. Neste trinn blir en del for overføringer (alderstrygd uførepensjon, arbeidsløshetstrygd, sosialhjelp)

Det vil ikke for alle disse overføringene være et tilfredsstillende datamateriale til å presist regne ut hvordan regelverksendringer slår ut. Det kan derfor være behov for tilleggsdata og grove tilnærminger. Det vil dermed være begrenset hvilke typer av politikksimuleringer en kan foreta. Dette gjelder i prinsippet også for skatter, siden vi der er avhengig av at grunnlaget for skatteberegningen, lar seg konstruere fra eksisterende selvangivelsesopplysninger.

Når det gjelder forbrukstall, er det et spørsmål om en skal basere seg på "rå-data" fra FU, eller om en skal

basere seg på modellestimerte budsjettandeler. Det vil også være mulig å benytte ulike former for estimerte atferdselastisiteter for eksempel etterspørselselastsisteter.

Ved siden av arbeidet med å utvikle modellbefolkning og en politikk-simuleringsmodell uten atferd (av typen LOTTE eller SPSD/M), er det viktig å videreføre arbeidet med atferdsanalyser. Vi har idag betydelig økonometrisk kompetanse på arbeidstilbud og forbruk med videreutviklingsarbeid i flere retninger. En viktig del av strategien for en modellbefolkning, bør være å sikre bedre innsamling av primærdata og å sikre at de blir tilrettelagt bedre for økonomteriske analyser og simuleringsmodeller(blant annet ved registerutnytting). En viktig del av strategien vil også være fortsatt satsing på mikroøkonometrisk arbeid.

7.ANDRE INNTRYKK.

Det ble også noe tid til besøk ved ulike andre statistikk kontorer og noen seminarer. Fra disse noen spredte inntrykk.

Besøk hos John M. Leyes (Director Small area and administrative division.)
Han fortalte om arbeidet med skattebånd og skattestatistikk i Canada. De hadde totalfiler for alle skattebetalere for hele Canada. 17 mill. skattebetalere. De har et omfattende selvangivelseskjema som i sin helhet bli dataoverført. I tillegg til dette ønsker de å koble til registeropplysninger fra andre kilder. (Ulike overføringer (Family allowances) Old age security, unemployment insurance, sosial assistance etc). De har imidlertid ikke noe felles id-nummersystem. De har derfor utviklet ulike andre metoder som gjør det mulig å foreta tilnærmede koblinger for statistisk bruk. De bruker her adresseinformasjon. På denne måten kobler de 90 prosent direkte. Gifte kobles også ved bruk av adresser og etternavn. Det gjennomføres også da tester på aldersforskjeller etc.
Ved siden av de statistiske problemer som selvfølgelig oppstår når en skal foreta datakobling mellom administrative datafiler når det ikke finnes et felles personnummer-system, syntes det også som de var underlagt politiske restriksjoner med hensyn på kobling. Også i Canada finnes det et "Datatilsyn" som vokter på personvernet.

Arbeidet med disse datafilene ledet fram til inntektsfiler for hele Canada. Filene var også organsisert med tanke på longitudinale analyser. Filene ble derimot ikke brukt til å lage tall for skatter betalt. Dette ble hentet fra andre kilder (summariske oppgaver fra skatteoppkreverne). Til tross for alt arbeidet med statistisk kobling var det

kort produksjonstid på å få filene klare.

Besøk hos Richard Veevers, Project Manager, Special Surveys Group.
Han ledet arbeidet med en arbeids- og inntektsundersøkelse som siktet mot å kartlegge alle arbeidsforhold i løpet av ett kalenderår. Spørreskjemaet var elegant utformet og virket godt gjennomtenkt. Dataene ble blant annet brukt til longitudinelle analyser av endringer i arbeidsmarked og inntektsfordeling. (Jfr analysen til Picot)

Seminar om "Good jobs bad jobs". (Picot)
Picot hadde brukt den longitudinelle databasen (jfr møtet med Veevers). Han hadde funnet at det hadde vært en sterk vekst i jobber med lav eller høy inntekt.
Midten var på en måte blitt borte. Han hadde studert inntektsfordelingen men ved relativt primitive metoder.

Møte med Janice McMechan. Bussines and Labour market analysis.
Hun arbeidet med en databank for bedrifter. Denne databanken ga mulighet til mikroanalyser av individata. Det var engasjert eksterne forskere som fikk arbeidsplass i SC til å arbeide med individata. De fulgte da de personvernregeler som gjelder for ansatte i SC. En av de tilknyttede var professor John Baldwin.
Han arbeidet med et prosjekt hvor de skulle følge bedrifter over livsløpet. Et av momentene som vi samtalte om var hva en skal definere som enheten ved longitudinelle analyser av bedrifter, foretak og konsern. Hans tanke var at den fysiske bedrift er det som er lettest å definere over tid.

LITTERATUR

Adler, Hans og M.C. Wolfson (1988): "A Prototype Micro-macro Link for the Canadian Household Sector." Review of Income and Wealth, No 4, Dec 1988.

Armstrong, J. (1989): "An Evaluatioin of Statistical Matching Methods." Draft, Social Survey Methods Division, Statistics Canada, Ottawa.

Cotton, Paul og George Sadowsky (1989): "Future Computing Environments for Socioeconomic Microsimulation." Unpublished mimeo.

Fellegi, Ivan (1988): "Can we afford an aging society?". Canadian Economic Observer, October 1988.

Kennedy, B. (1986) : "The LIPPS Earnings Module". Mimeo, Statistisc Canada, Social and Economic Studies.

Lemaitre, G. og J. Dufour (1987): " An Integrated Method for Weighting Persons and Families." Survey Methodology, Dec. 1987, Vol 13.No 2 (pp199-207)

Osberg, Lars (1986): "Behavioural response in the context of sosio-economic microanalytical simulation" Statistics Canada, Analytical Studies Branch, Research Paper Series no 1.

Picot, Garnett (1986): " Modelling the lifetime employment patterns of canadians." Statistics Canada, Analytical Studies Branch, Research Paper Series no 4.

Rowe, Geoff (1989): "Event history analysis of marriage and divorce in Canada.". Draft, October 1989. Statistics Canada, Social and Economic Studies.

Wolfson, Michael C.(1988): "Homemaker Pensions and Lifetime Redistribution", Review of Income and Wealth, No3 1988.

Wolfson, Michael C. (1987): "Notes on Corporate Concentration and Canada's Income Tax." Statistics Canada, Analytical Studies Branch, Research Paper series No.8.

Wolfson, Michael C. (1989): " The CEPHID Project,. Canadas Elderly -Projecting Health, Income, and Demography." Working Paper No 1 -- A Project Proposal, Mimeo February 6, 1989, Statistics Canada, Social and Economic Studies Division.

Wolfson, Michael C. (1989b): "A system of health statistics. Toward a new conceptual framework for integrating health data." Paper IARIW, conference Lahnstein aug 20-26, 1989.

Wolfson, Michael C. og Stephen Gribble, Michael Bordt, Brian Murphy og Geoff Rowe (1989): "The Social Policy Simulation Database and Model- An Example of Survey and Administrative Data Integration." Draft, February 3, 1989. Statistics Canada, Social and Economic Studies Division.

Wolfson, Michael C. (1989a): "Divorce, homemaker pensions and lifecycle analysis." Population Research and Policy Review 8:25-24.