

Rutiner för datauttag i SESIM

Resultat från simuleringar i SESIM kan presenteras på ett antal olika sätt vilka är lämpliga i olika analysituationer. Enklare analyser kan göras direkt via SESIM:s windows-gränssnitt med hjälp av olika ”statistikfönster”. Dessa inkluderar univariata statistikor (medelvärde, summa, standardavvikelse, min, max o.s.v.), kernelskattningar av en variabels täthetsfunktion samt frekvenstabeller. Några exempel på dessa statistikfönster ges i nedanstående bild. Godtyckligt många instanser av varje typ av statistikfönster kan vara öppna samtidigt.

The screenshot displays the SESIM software interface with several windows open:

- Control Centre:** Contains control options such as "Save output files", "Weight (100%)", "Sample size %", and "Run system". It also includes fields for "My parameter Excel file" and "Base parameter Excel file".
- Univ: i_age (Age):** A window showing a table of statistics for the variable "i_age (Age)" at "Time 2000".
- Kernel:** A window showing a kernel density estimate plot for "i_age" with "Auto min" and "Auto max" options checked.
- Editor:** A window with a "Print" button and a timestamp "2002-11-01 15:58:58".
- Statistical Summary:** Displays "Year: 2000", "Individuals: 112406", and "Households: 59054".
- Status:** A list of tasks such as "1 Set status", "1 SocMod", "1 Wealth_Housing_Capital", etc.
- Freq: i_agegrp5 (Five year age groups):** A window showing a frequency table for "i_agegrp5 (Fi)" at "Time 2000".

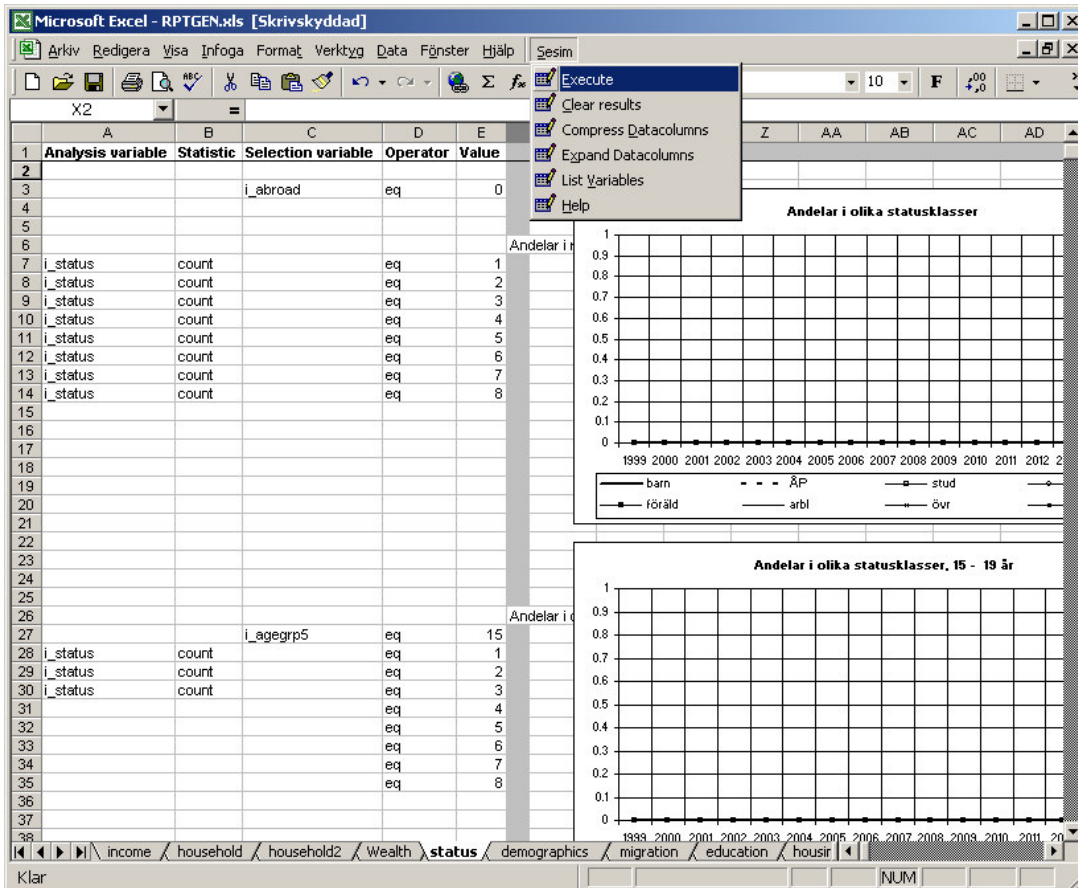
Time	Mean	Sum milj	Std	Min	Max
2000	39.9	380.8	23	0	102

Time	0	5	10	15	20
2000	471410	599505	633760	534735	523515

I många analysituationer är det en fördel att kunna återupprepa samma analys, t.ex. då beräkningsförutsättningarna på något sätt har ändrats. I en sådan situation är det olämpligt att använda SESIM:s gränssnitt eftersom det kräver en del manuellt arbete varje gång.

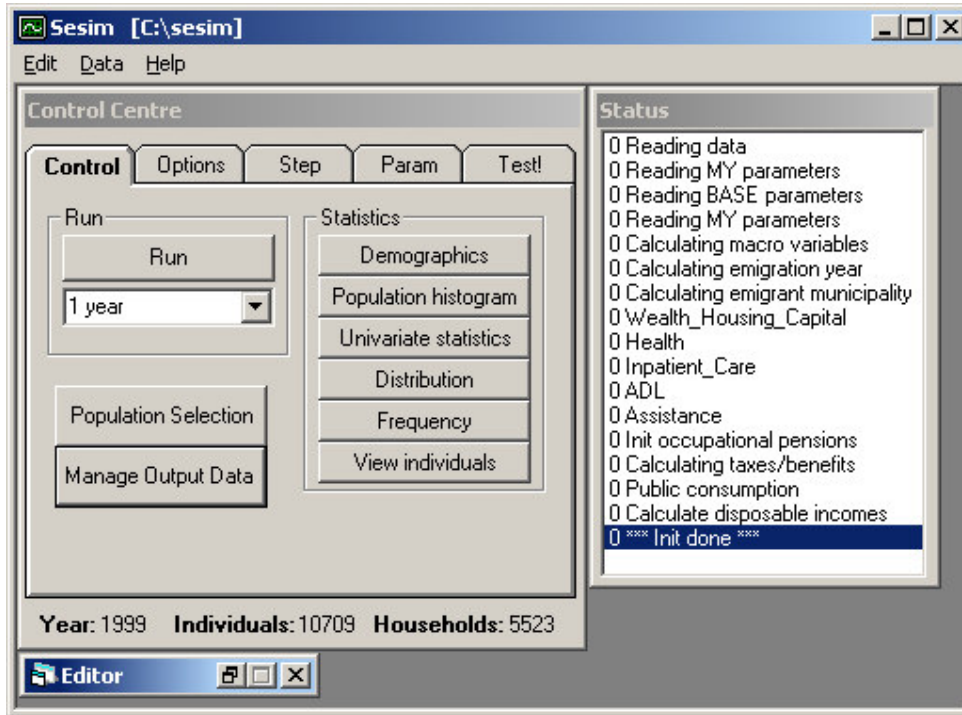
Vid en upprepad analys kan då istället t.ex. SESIM:s rapportgenerator användas. Rapportgeneratoren är en Excelapplikation i vilken databeställningar enkelt kan formuleras och sedan hämtas från SESIM vid exekvering. Nedan visas ett exempel på hur man från rapportgeneratoren beställer en beräkning av antalet

individer, boende i Sverige (i_abroad eq 0) som finns i respektive status (1, 2, ..., 8). På raderna 27-35 upprepas analysen men med tilläggs villkoret att individerna dessutom ska vara mellan 15 och 19 år gamla (i_agegrp5 eq 15).

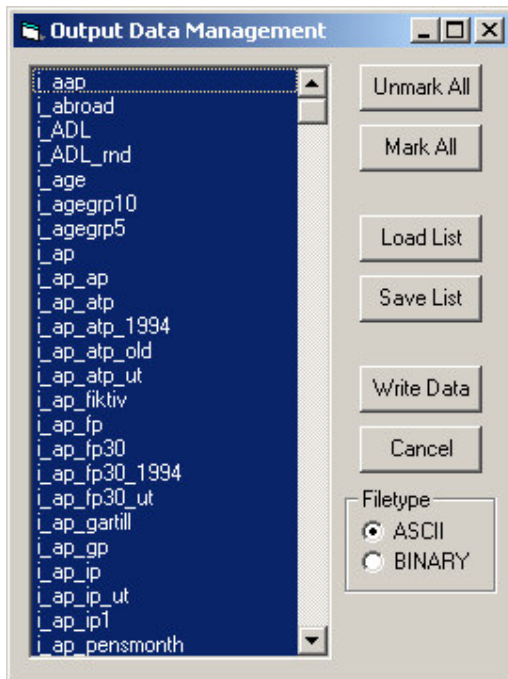


I rapportgeneratorm kan alltså olika databeställningar färdigställas i förväg och sedan enkelt uppdateras vid behov. Resultaten från simuleringarna kan sedan bearbetas och presenteras med hjälp av Excels inneboende funktionalitet. I exemplet ovan visas t.ex. några diagram över individers fördelningar på olika statusklasser. Diagrammen innehåller inga data men uppdateras så fort resultaten skrivs till kalkylarket från SESIM.

Störst flexibilitet i analyserna ges genom att göra uttag av data på individ-/hushållsnivå från SESIM för vidare analyser i ett separat statistiskt programpaket (t.ex. SAS, R eller Stata). Data hämtas genom att klicka på knappen "Manage Output Data" (se bilden nedan) efter det att SESIM har initierats.



I det formulär som öppnas (se bilden nedan) visas SESIM:s variabelinnehåll. Samtliga variabler är markerade i utgångsläget. Om inte alla variabler ska skrivas till fil är det enklast att radera alla selektioner (Unmark All) och sedan markera de relevanta variablerna. För att underlätta hanteringen av utdata kan variabelistor sparas till fil (Save List) och sedan läsas tillbaka (Load List) nästa gång ett datauttag ska göras.



När rätt variabler är selekterade återstår att välja utdataformat och att klicka på knappen "Write Data". Data skrivs då omedelbart till fil. Filens namn beror på vilket utdataformat som valts. Om formatet är ASCII ges

utdatafilen namnet SESIM_YYMMDD_HHMMSS.txt och om formatet är BINARY skapas två filer; SESIM_YYMMDD_HHMMSS_header.txt och SESIM_YYMMDD_HHMMSS_data.bin. För båda dessa format kommer sedan varje års simuleringsdata att adderas till filerna. Detta görs sist i SESIM:s simuleringssekvens. Filerna hamnar automatiskt i katalogen ...\\sesim\microdata.

Filerna i ASCII-format kan importeras direkt till de statistiska programpaket som kan hantera ASCII-data. Den första raden innehåller variabelernas namn. Det bör observeras att dessa filer blir mycket stora och att det kan ta SESIM mycket lång tid att skapa dem under simulering.

Filerna i BINARY-format skrivs betydligt snabbare från SESIM. Det är därför rekommenderat att använda detta format om urvalet är stort eller om antalet valda analysvariabler är stort. Filerna måste dock bearbetas innan de kan importeras till ett godtyckligt statistiskt programpaket. Bearbetningsprogrammet¹ har implementerats i det statistiska programpaketet R (se <http://www.r-project.org/>). R är gratis och tillgängligt enligt "GNU General Public License".

För analyser som utförs i SAS ges i appendix 2 exempel på programstruktur för inläsning av data från SESIM. Programmet hanterar båda typerna av dataformat.

¹ Koden återfinns i sin helhet i appendix 1.

Appendix1 – R-kod

Koden nedan transformerar utdata från SESIM i binärt format till ASCII-format. Filen SESIM_YYMMDD_HHMMSS_header.txt är en textfil som innehåller information om vilken data som finns i filen SESIM_YYMMDD_HHMMSS_data.bin och hur den ska läsas.

Programmet skapar tre textfiler i ASCII-format; idata.txt för individvariabler, hdata.txt för hushållsvariabler och mdata.txt för makrovariabler. Skulle någon av dessa variabeltyper saknas bland de variabler som selekterats för analys skapas inte motsvarande textfil.

När nya analysdata ska bearbetas behöver endast programmets tre första rader (vars första tecken inte är kommenteringstecknet #) ändras. Här anges de aktuella sökvägarna och filnamnen.

Read_binary_data.r

```
#####  
# Program för inläsning av binära data från SESIM.  
# Efter inläsning skrivs data till ASCII-fil.  
#####  
  
### Byt katalog och läs in filer  
setwd("C:\\\\sesim\\microdata")  
headerfile <- read.delim("SESIM_040527_160105_header.txt", header = FALSE, sep = " ")  
binfile <- file("SESIM_040527_160105_data.bin", "rb")  
  
### Beräkna simuleringens första och sista år  
minyear <- min(headerfile$V1)  
maxyear <- max(headerfile$V1)  
minyear; maxyear  
  
### Beräkna antal individ-, hushålls- samt makrovariabler  
firstyear <- subset(headerfile, V1 == minyear)  
temp <- toupper(substr(as.character(firstyear$V2), 1, 1))  
nivars <- length(temp[temp == "I"])  
nhvars <- length(temp[temp == "H"])  
nmvars <- length(temp[temp == "M"])  
nivars; nhvars; nmvars  
  
### Sammanställ variabelnamn  
ivarnames <- as.character(firstyear$V2[toupper(substr(as.character(firstyear$V2), 1, 1)) == "I"])  
hvarnames <- as.character(firstyear$V2[toupper(substr(as.character(firstyear$V2), 1, 1)) == "H"])  
mvarnames <- as.character(firstyear$V2[toupper(substr(as.character(firstyear$V2), 1, 1)) == "M"])  
ivarnames; hvarnames; mvarnames  
  
### Sammanställ typ av variabel (individ, hushåll eller makro)  
vartypes <- toupper(substr(as.character(headerfile$V2), 1, 1))  
table(vartypes)  
  
### Sammanställ datatyp (long, integer, byte eller double)  
datatypes <- toupper(substr(as.character(headerfile$V4), 1, 3))  
table(datatypes)  
  
rowcounter = 1  
  
### För varje år...  
for (year in minyear:maxyear){  
  
  if (exists("idata")) rm(idata)  
  if (exists("hdata")) rm(hdata)  
  if (exists("mdata")) rm(mdata)  
  
  ### För varje variabel respektive år...  
  for (var in 1:(nivars + nhvars + nmvars)){  
  
    nobs <- headerfile$V3[rowcounter]  
    cat(year, " ", nobs, "\n")  
  
    ### Datainläsning betingat på datatyp  
    data <- switch(datatypes[rowcounter],  
      LON = readBin(binfile, integer(), nobs , size = 4),
```

```
INT = readBin(binfile, integer(), nobs , size = 2),
BYT = readBin(binfile, integer(), nobs , size = 1),
DOU = readBin(binfile, double(), nobs , size = 8)

### Den inlästa vektorn vertikalstackas till rätt matris
if (vartypes[rowcounter] == "I"){
  if (!exists("idata")) idata <- data else idata <- cbind(idata, data)
}

if (vartypes[rowcounter] == "H"){
  if (!exists("hdata")) hdata <- data else hdata <- cbind(hdata, data)
}

if (vartypes[rowcounter] == "M"){
  if (!exists("mdata")) mdata <- data else mdata <- cbind(mdata, data)
}

rowcounter = rowcounter + 1
} # end for

### Ett års data inläst --> skriv till fil
### Om första året så skriv kolumnrubriker
if (year == minyear){
  if (nivars > 0) {
    txt <- cbind("year", t(ivarnames))
    write.table(txt, file = "idata.txt", row.names = F, col.names = F, quote = F)
    output <- cbind(rep(year, times = dim(idata)[1]), idata)
    write.table(output, file = "idata.txt", row.names = F, col.names = F, quote = F, append = T)
  }

  if (nhvars > 0) {
    txt <- cbind("year", t(hvarnames))
    write.table(txt, file = "hdata.txt", row.names = F, col.names = F, quote = F)
    output <- cbind(rep(year, times = dim(hdata)[1]), hdata)
    write.table(output, file = "hdata.txt", row.names = F, col.names = F, quote = F, append = T)
  }

  if (nmvars > 0) {
    txt <- cbind("year", t(mvarnames))
    write.table(txt, file = "mdata.txt", row.names = F, col.names = F, quote = F)
    output <- cbind(rep(year, times = dim(mdata)[1]), mdata)
    write.table(output, file = "mdata.txt", row.names = F, col.names = F, quote = F, append = T)
  }
}
else {
  if (nivars > 0) {
    output <- cbind(rep(year, times = dim(idata)[1]), idata)
    write.table(output, file = "idata.txt", row.names = F, col.names = F, quote = F, append = T)
  }

  if (nhvars > 0) {
    output <- cbind(rep(year, times = dim(hdata)[1]), hdata)
    write.table(output, file = "hdata.txt", row.names = F, col.names = F, quote = F, append = T)
  }

  if (nmvars > 0) {
    output <- cbind(rep(year, times = dim(mdata)[1]), mdata)
    write.table(output, file = "mdata.txt", row.names = F, col.names = F, quote = F, append = T)
  }
}
}

close(binfile)
```

Appendix 2 – SAS-kod

Nedanstående kod är kopierad från dokumentet *S:\Projekt\BabyBoom\Dokument\Validering\valideringsrapport.sas* vilket används för att producera valideringsrapporter inom BabyBoomprojektet.

De tre första raderna kan behöva ändras då nya data ska bearbetas. Makrovariabeln FILEPATH ska alltid ange sökväg till datafilerna. Om binära data används ska makrovariabeln TXT_BIN sättas till BIN varvid SAS läser in filerna idata.txt och hdata.txt som skapats av R. Om ASCII-data skrivits från SESIM ska TXT_BIN sättas till en valfri textsträng annan än BIN vilket gör att SAS istället läser in den aktuella ASCII-filen. Makrovariabeln FILENAME sätts då till den aktuella filens namn.

Oavsett vilken typ av indatafiler som används genererar programmet SAS-datasetet sesimdata vilket innehåller både individ- och hushållsvariabler.

```
*****
Program för sammanställning och presentation av simuleringsresultat
från SESIM speciellt med avseende på moduler inom BabyBoom-projektet.
*****;

*** MAKROVARIABLER ***;
%let TXT_BIN = BIN; *** BIN gör att idata.txt och hdata.txt används ***;
%let FILEPATH = C:\sesim\microdata;
%let FILENAME = SESIM_040512_150818.txt; *** ASCII flat-file ***;

%macro datasource(txtbinary);

    *** Data exporterat som binärdata från SESIM och sedan inläst
        och transformerat i R. ***;
    %if &TXTBINARY = BIN %then %do;

        proc import datafile = "&FILEPATH.\idata.txt"
            out = idata replace dbms = dlm;
            delimiter = " ";
        run;

        proc import datafile = "&FILEPATH\hdata.txt"
            out = hdata replace dbms = dlm;
            delimiter = " ";
        run;

        proc sort data = idata out = itemp;
            by year i_hhnr;
        run;

        proc sort data = hdata out = htemp;
            by year h_hhnr;
        run;

        data sesimdata;
            merge itemp htemp (rename = (h_hhnr = i_hhnr));
            by year i_hhnr;
        run;

    *** Data exporterat som ASCII flat-file från SESIM ***;
    %end;
    %else %do;

        PROC IMPORT OUT = sesimdata
            DATAFILE= "&FILEPATH.\&FILENAME"
            DBMS=DLM REPLACE;
            DELIMITER='20'x;
            GETNAMES=YES;
            DATAROW=2;
        RUN;

    %end;
%mend;
%datasource(&TXT_BIN);
```