Danmarks Statistik
MODELGRUPPEN

Arbejdspapir*

Michael Osterwald-Lenum

08. juli 2016[1]

Principper bag dataarbejdet omkring ADAMs data-bank

Resumé:

Papiret skitserer nogle af de fundamentale principper, der ligger bag det løbende dataarbejde omkring vedligeholdelse og fornyelse af ADAMs databank.

{printet 2017-02-24 18:23}

{dokumentnavn: MOL08716_v0r07.docx}

[1] Revideret senest d. 01. august 2016 for indhold. Nedenstående forbehold på forsiden er tilpasset, så citation uden henvendelse til DST er ok, per 24. februar 2017.

MOL08716

Nøgleord: databank, principper

Indledning

Fra tid til anden er der behov for at minde os selv om de principper der ligger bag det løbende dataarbejde for at vedligeholde og forny systemerne der benyttes ved produktionen af ADAMs databank. Nærværende papir er tænkt som en hjælp i disse situationer.

I følgende kapitel omtales forskellige "principper for dataarbejdet", som typisk bygger på indarbejdede "traditioner" fra modelgruppens arbejde med at etablere, vedligeholde og omlægge såvel databankernes indhold, som systemerne der benyttes til at styre og automatisere disse processer.

Nærværende forfatter har bidraget til at formulere disse traditioner som principper, og kan derfor kun undtagelsesvist betragtes som ophavsmand til traditionerne bag disse principper.

Læseanbefalinger

For små databanker kan opdateringsarbejdet udføres af nogle få medarbejdere, hvis ikke en enkelt medarbejder. For medarbejdere i sådanne miljøer anbefales læseren at springe andet punkt under "Rammen…" over og fortsætte direkte til "Principper…".

For medarbejdere ved større databanker anbefales det ikke at srpinge noget punkte over.

Rammen om arbejdet og dets organisering

0. ***Databanken*** til en makroøkonometrisk model
    a. omfatter et antal ***variabler***, for hvilke der i databanken er samlet observationer dækkende en sammenhængende periode,
    b. bliver opdateret et fast antal gange om året,
    c. frigives mindst to gange under hvert opdaterings forløb; *den første foreløbige version*, og *den endelige version*,
    d. tilbydes brugere med særligt akutte behov, flere gange under opdateringsforløbet, svarende til hver gang banken på et væsentligt område er opdateret.

1. ***Variablerne*** dokumenteres med en
    a. ***variabel liste,***[2] og en
    b. ***ligningsbrowser***.[3]

[2] For hver ***variabel*** fastlægges en ***betegnelse*** (evt. i kodeform), en ***kort tekst*** med "definition" eller indholdsopsummering, en ***måleenhed*** (evt. inkl. prisniveau), en ***kildeangivelse*** (evt. beregningsformel), og en ***central identitet*** som den indgår i.

[3] ***Ligningsbrowseren*** er et (pt. html-baseret) system, der danner opslag, baseret på variabellistens oplysninger, og knytter oplysning om variablen er endogen eller eksogen i forhold til en

c. ***Definition***: For en databank knyttet til en model, er der mindst to væsensforskellige
    i. kilder til definition af den enkelte variabel:
        1. ***kilde-baseret definition***: fastlæggelse af en databank-variabels værdier ved indholdet af en given kildes afgrænsninger, og kilde-variabel, og
        2. ***økonomisk teoretisk begrebs baseret definition***: fastlæggelse af en databank-variabels værdier ved en teoretisk variabel, eller sammenhæng, hvis værdi (af praktiske årsager) tages fra den givne kilde.
    ii. En konsekvens af hvilken kilde til definition der er for den enkelte variabel, er i hvilket omfang eventuelle identiteter mellem kildevariabler nedarves til databankens variabler.
        1. Eksempelvis vil der mellem kilde-baserede variabler kunne nedarves sammenhænge fra nationalregnskabet, som ikke nødvendigvis vil nedarves hvis variablerne er økonomisk teoretisk definerede, hvor den valgte kilde kun er en approksimation til det teoretiske begreb, der ønskes målt.

d. ***Variabelnavnes systematik*** (***nomenklatur***) er vigtig.
    i. Fastlæggelsen af en systematisk nomenklatur bag dannelsen af variablers navne er uhyre vigtig, dels for brugerens evne til at læse en ligning hvori der optræder mange variabler, og tjekke at det er de rette variabler der står på de rette steder i ligningerne, dels for at en bruger kan blive fortrolig med et (for sit arbejde) passende udvalg af variabelnavne, så antallet af opslag holdes nede.
    ii. Den enkelte nomenklatur er et kompromis mellem det i natursprog sigende (lange) navn (der bliver ubekvemt i ligninger med mange forekomster af variabler), og det for skrivning bekvemme i kortere "koder".

2. ***Overordnet organisation*** af arbejdet og ansvar for de enkelte dele.
    a. ***Databankdirektøren*** (DD) har det overordnede ansvar for databanken, for planlægning af opdateringsprocessen, for at opdateringsprocessen glider (sørger for at "stafetten" passerer til rette vedkommende i opdaterings flowet og de næste varskos) og for uddelegeringen af opklaringsopgaver, og orientering af kontorchef om forløbet og eventuelle ekstraordinære mandskabsbehov.
        i. DD holder forskellige ***log***-filer, der opsamler ***proces metadata og udfordringer***.
            1. Udfordringer opdeles i dem der skal løses inden næste opdatering af den foreløbige databank, inden databanken bliver erklæret endelig (dvs. i løbet af en opdateringsrunde), dem der skal rettes op ved dannelsen af næste databank (i samme re-

---

given modelversion, oplyser variablens eventuelle definition i modellen, og giver links til andre variabler, i hvis modelligninger variablen optræder.

gime), der der skal rettes op ved næste modelversion, og dem der skal rettes op ved næste NR-hovedrevision (eller lignende).

ii. DD udarbejder en samlet ***opdateringsplan*** til hvert opdateringsforløb, dels med leverancedatoer for alle hovedleverancer, dels med centrale brugeres behov for opdateringsindsats, dels en moduloversigt med angivelse af moduler og modulansvarlige. Desuden med en oversigt over deltagernes ferieplaner.

b. ***Modulansvarlige***. Alle ***moduler*** (for definition se under principper for arbejdet) har én person som modulansvarlig. Som hovedregel bør der være en backup-person, som er i stand til at gennemføre opdateringen af modulet, når den modulansvarlige er fraværende.

i. Den modulansvarlige en ansvarlig for opdateringen af sine moduler. Dvs. for at

1. hente de fornødne kildedata,
2. bearbejde disse,
3. tjekke deres konsistens og umiddelbare troværdighed, og
4. danne skøn, hvor en kildes data ikke foreligger når modulet skal bruges til opdatering af databanken.

ii. For hvert modul bør der foreligge

1. en driftsvejledningsfil (***!!SeHer.!!***-fil),
2. en gendannelseskommandofil (***datop***/***datarev***-fil, og evt. en separat ***til_obk***-fil), der sætter DD i stand til at genkøre modulet, såfremt dette bliver nødvendigt uden inddragelse af den modulansvarlige,
3. en fil med observationer af fejl (***kox***-fil) og
4. en ***log***-fil, der opsamler hvornår noget er sket, og løbende kommentarer til driften af modulet.

c. ***Udførende***. For visse moduler er arbejdet så velbeskrevet, ukompliceret og regelmæssigt, at det kan uddelegeres til andre, f.eks. studentermedhjælp. For mere komplekse moduler kan andre inddrages.

d. ***Generelt***.

i. For essentielle moduler, der har korte deadlines er det fastansatte, der er modulansvarlige.

Principper for arbejdet og dets organisering

0. ***Genbrug*** er en ædel kunst. Vi ”betaler” for lånet ved at omtale hvor ideerne og kodestumperne kommer fra.

1. (Løsnings-) ***Kode*** skrives gerne i kommandofiler, så en beregning kan gentages, og gode løsninger deles med andre.

   a. Kode skal så vidt muligt være selvdokumenteret – forstået på den måde at "meningen" med afsnit af kode fremgår af ledsagende kommentar strenge.
   b. Kode formuleres, så en opgaves dele kan gentages for alle elementer i en liste.

2. ***Versionsstyring*** er central. ***Hovedprincippet er at vi til enhver tid kan beskrive hvordan vores uddata er fremkommet fra kildedata.*** Derfor gemmer vi indholdet af biblioteker inden større ændringer i kildedata eller kommandofiler gennemføres.
   a. Vi gemmer databankerne med passende mellemrum undervejs i opdateringsprocesser, så fejlopdateringer kan omgøres uden synderlige omkostninger i egen eller kollegaers tid.
   b. Databankdirektørens (DD) anvisninger efterleves uden bøvl.
   c. Skriveadgang til databanken begrænses af den enkelte til det fornødne, og DD er altid orienteret inden. Orienteringspligten er skærpet under opdateringsperioden op til første foreløbige databank.

3. ***Modulopbygning prioriteres***.
   a. Det samlede opdaterings dataflow nedbrydes i en række moduler, hvis indbyrdes afviklingsrækkefølge klarlægges. Dette overblik kaldes dataflowdiagrammet.
   b. Afgrænsningen af hvad der hører med i det enkelte modul er et kompromis mellem alt hvad der naturligt hører sammen i forhold til den enkelte datakilde, og alt hvad der kan samles og aftestes uden inddragelse af andre datakilder.
   c. Databankens overholdelse af en række testbetingelser tjekkes på forskellige steder i det samlede dataflow. Hensigten med disse tjek er at tilvejebringe sikkerhed for at betingelserne for at efterfølgende moduler kan afvikles allerede er opfyldt, samt at udpege de variabler og betingelser, der endnu ikke er fuldt opdaterede hhv. rummer variabler der ikke er det.
   d. For hvert modul skal der foreligge mindst to filer:
      i. ***SeHer***-filen, der giver den helt præcise instruks til opdateringen af modulet. Typisk kaldt (en variant af) ***!!SeHer.!!***.[4]
      ii. Genkørsels kommandofilen, der kan afvikles når et tidligere modul er blevet genkørt, således at alle afledte variabler bliver korrekt opdateret. Denne kaldes typisk (en variant af) ***datop.cmd***.
      iii. Genoverførsels kommandofilen, der afvikles for at overføre modulets data til den fælles databank. Typisk kaldt (en variant af) ***tilobk.cmd***.
   e. Opdateringen af det enkelte modul kan opdeles i fire dele:
      i. indhentning, og evt. transformation, af kildedata til en fast inddata standard,

[4] Dette sikrer at ***SeHer***-filen på en windows pc står øverst i en alfabetisk ordnet liste over filer i et (under-) bibliotek.

   ii. transformation af inddata på standard form til databank variablers form
   iii. konsistenstjek, og opklaring af årsager til afslørede inkonsistenser eller mangler; plots af variablers værdier kan også afsløre overraskende udviklinger, der kan give anledning til yderligere undersøgelser og orienteringer,
   iv. informering af DD af at modulet er opdateret, og eventuelle problemer med kilder, eller overgange til databank variabler.

f. Efter alle moduler, der bidrager med kildedata og nært afledede variabler, er kørt, køres en række moduler, der danner alle de variabler, der ikke kan afledes før; således at alle de variabler der skal indgå i databanken, kommer med.
   i. det såkaldte ***hoved***-program, rummer kommandoer der tilføjer værdier til de resterende variabler.
   ii. Skal databanken benyttes af en model, så kan der være en kommandofil (typisk kaldt ***tabel.cmd***) der afleder såkaldte tabel-variabler.[5]
   iii. Er en ny modelversion under udarbejdelse, så vil dataopdateringssystemet bag den aktuelle databank typisk passe til en gældende modelversion, og de supplerende data behov fra den nye modelversion tilføjes med en kommandofil (typisk med navnet ***estbk.cmd***), der afvikles efter ***hoved***-programmet.

4. ***Central fastsættelse af værdierne af fælles processtyrings variabler og tegnstrenge***. Styring af opdateringsperiode, andre periodesætninger af betydning for arbejdet på tværs af moduler, og fælles databanknavne samt stier.
   a. Opdateringsperioden styres ved en række globalt satte variablers indhold; typisk ved variabler med tekststrenge, der angiver start- eller slutperiode. Herved sikres at opdateringsperioden synkroniseres på tværs af forskellige modulers kommandofiler.
   b. Baggrunds variabler med oplysning om basisår, seneste år med tal fra endeligt NR, etc. sættes ét sted.
   c. Hovedopdateringsdatabankens navn, og sti sættes ét sted.

5. ***Variablers opdaterings status*** skal være klar
   a. Ved opdaterings start overskrives databankens (tidsrække-) variabler for opdateringsperioden med systemets værdi for ”missing”.
   b. En variabels seneste opdatering kan aflæses af en dato, gemt som metadata.

6. Et ***udtrækssystem*** der lister hvilke variabler (der er modelvariabler), der forsat har ”missing”-værdi for forskellige perioder benyttes til at tjekke hvilke variabler der endnu ikke er opdaterede. Specielt når der af disse generelle lister kan uddrages lister for hvert modul, bliver det let for

[5] Dvs. variabler, der indgår som venstreside variabler i en modelfil efter AFTER-sætningen.

DD at henvende sig til de modulansvarlige, for at få hvert modul tilstrækkeligt opdateret for at den samlede proces kan forløbe glat og smidigt.

   a. Erfaringen er at der løbende sker ændringer i datakilder, og delprogrammer for at holde det enkelte modul "på sporet", og derfor kan nogle variabler falde ud af eller undgå at komme ind i det enkelte modul, selvom den modulansvarlige synes at opdateringen af modulet er overstået.

7. Afsluttende ***testbetingelser*** gennemføres.
   a. Når databanken er forbundet med en model, tjekkes alle modellens ligninger der forventes at holde uden residualer, og for ligninger med residualer listes prints med tidsrækker af residualer.
   b. Prisindeks variabler tjekkes for at overholde basisår-værdien, evt. nul-udviklinger eller ikke-positive værdier.
   c. Mængdevariabler tjekkes for særlige værdier: nul-udviklinger, eller ekstreme værdier.
   d. Nationalregnskabsidentiteter tjekkes i løbende priser og, så vidt meningsfuldt også, i foregående års priser.

8. Under ***fejlsøgning af kode***, f.eks. på grund af kommandofilers nedbrud eller resultatfejl er det nyttigt med ***"XX"-kommandoen*** til at spore hvor programmet kører "i skoven".
   a. Hvis software systemet (ikke er et kompileret sprog, men) fungerer så hver linjes kommandoer afvikles fuldt ud inden næste linje påbegyndes, så kan man under fejlsøgningen af en kommandofil indsætte en invalid kommando forskellige steder.[6]
   b. Hvis afviklingen af systemet, på et sted med ***xx***-kommandoen, giver mulighed for at vælge mellem at stoppe afviklingen helt og exitte helt, stoppe kommandoflowet (men bevare adgang til alle de banker der er åbne på det sted i kommandoflowet, etc.), og blot fortsætte (på trods af den opståede fejl), så giver den anden mulighed adgang til at undersøge de åbne databankers tilstand og sammenligne med hvad man ville forvente, hvilket kan være overordentlig hjælpsomt.

9. ***Fejlsøgning af data***.
   a. Kildenær fejlsøgning.
      i. Er der tale om et datasæt med en høj grad af indre sammenhæng, er et udvalg af konsistenstjek vigtige at gennemføre tidligt i processen efter indlæsning for at tjekke dels at indlæsningen er gået godt, og dels at datasættet overholder de begrebsbestemte identiteter.
      ii. Det er nyttigt at sætte de helt centrale aggregater i forhold til hinanden og se på grafer heraf, som overordnet tjek af manglende værdier, eller overraskende udviklinger.

[6] I software systemet AREMOS fungerer eksempelvis kommandolinjen "XX;" sådan.

iii. Sammenligning af den indlæste kilde med den seneste endelige databanks tilsvarende indlæste kilde, er typisk godt brugt tid.
iv. konsistenstjek kan defineres og afvikles med kontrolkommandofiler med uddata i tekstfiler med faste navne, der viser aggregats værdi, kontrolberegnet aggregat, summeret nedefra, og deres forskel. Alle værdier vises for opdateringsperiode og en passende forhistorie.

b. Resultatnær fejlsøgning.
   i. Struktureret udvælgelse er en central metode til at reducere antallet af variabler, hvis værdier skal tjekkes.
   ii. Print af variablers tidsrækker, der inkluderer såvel opdateringsperioden som et passende udvalg af dens forhistorie, kan med fordel vises sammen med de tilsvarende variablers værdier i den seneste endelige databank, og evt. med deres differens (absolut og ændringen i deres absolutte differens i forhold til den tidligere databanks tilsvarende værdi).
   iii. Visualisering er en stærk fejlsøgningsmetode, når et mindre antal variablers værdier skal fejlsøges.
   iv. Transformation af variablers værdier kan være nyttig. Eksempelvis kan dannelse af nogle forhold mellem udvalgte variabler give et meget mere nuanceret billede end et print/plot af den utransformerede variabels værdier. Division med bnp i løbende priser kan skalere andre variabler der er opgjort i løbende priser, så deres "variationsspektrum" kommer bedre frem. Et andet eksempel er logaritmisk transformation af en variabel der udviser eksponentiel vækst over tid.

c. I opdateringsperioden, hvor vi frigiver foreløbige versioner af databanken, annoncerer vi hvilke (nye) kilder der er inddraget og hvilke områder af variabler, der kun kunne opdateres ved ekstraordinære skøn.
   i. Vi er glade for feedback fra vores brugere om at overraskende udviklinger (eller mangler) er observeret, da vi derved kan forbedre vores systems tilstand og funktion.
   ii. Vi tilstræber generelt en åbenhed omkring svagheder i vores datagrundlag, så vores brugere bl.a. har mulighed for at undgå overfortolkninger af tallenes præcision, hvor vi har kendskab til at de måtte være svagere end normalt.
   iii. Vi tilstræber at vores brugere har så præcis en forventning til hvornår den endelige version frigives, som muligt, under hensyntagen til vores behov for at kunne absorbere uventede forsinkelser.

10. ***Dokumentation***. Helt overordnet kan man sige at dokumentationen af dataopdateringssystemet bør være flerdelt. Dokumentationens form og indhold er et kompromis mellem ønsket om at formidle formål med og forståelser af hvorfor bearbejdningen er tilrettelagt på den valgte måde,

og behovet for kort og præcis beskrivelse, der let lader sig vedligeholde.

a. ***Den overordnede struktur***, med nedbrydning af dataflowet i moduler, dokumenteres i en samling dataflow-diagrammer.
b. De enkelte moduler dokumenteres på en måde der er hensigtsmæssig at opdatere og holde opdateret.
    i. ***Transparens***. Eksterne databankbrugere skal have mulighed for at forstå databankens dataindhold og dens dannelse fra kilderne, alene på grundlag af den almene adgang til ADAMs modelgruppepapirer, så *alle moduler dokumenteres i modelgruppe papirer*.
    ii. Overordnet er det ønskeligt at et moduls dokumentations papir, formidler den ***logik*** der ligger bag bearbejdningen af kildedata, og dermed også hjælper en fremtidig medarbejder til at kunne foretage de tilretninger af modulet, der bliver nødvendige når enten kilden ændrer sig, virkeligheden ændrer sig, eller modellens ændrede behov kræver det.
c. For ***omfattende kilder***, såsom nationalregnskabet, der trækkes på i forskellige moduler, kan det være en fordel at have tværgående dokumentationer der knytter sig til kilden; derved aflastes den enkelte modulansvarlige, så kilde-dokumentations ansvaret kan tilfalde den medarbejder, der har den fornødne indsigt.
d. Øvrig mere detaljeret dokumentation fremgår af de ***SeHer***-filer, ***datop***-filer, ***log***-filer og ***kox***-filer, der findes under de enkelte moduler, eller hos DD, herunder ***proces-log***-filer.

Litteratur

< pt. tom >

Bilag: effektivt arbejde med store, evt. fejlbehæftede, datasæt

Kildedata til visse moduler er så omfattende at mere generelle principper for dataarbejde må bringes i anvendelse.

Definition: flade datafiler, er filer med linjer med fast længde (dvs. et fast antal tegn inden linjeskift-tegnet; en linjes indhold kaldes også en ”record”), og en fast struktur, der præciseres ved et såkaldt record-layout (dvs. hvilke tegnafsnit der angiver værdien af det tilsvarende ”felt”; felt 1 tager sin værdi af tegnene 01-12 i en given linje, etc.).

En flad datafil kan importeres i en MS-Excelprojektmappe, og herefter nemt forædles så den passer til den ønskede anvendelse.

Ofte vil nye variablers værdier kunne defineres ved en lille overgangstabel, der anviser hvorledes den nye variabels værdi skal fastlægges på grundlag af

værdierne af andre felter i den enkelte record. Eksempelvis bliver aggregering over nogle koder meget nem, på denne måde.

Princippet her er at ***bevare de oprindelige records***, og så kobles nye felter med koder på, der gør det muligt at summere (et værdifelts indhold) over samtlige records for at bestemme værdifeltets sum knyttet til hver af et nyt felts forskellige værdier (ud af dets værdi-mængde).

Fordelen ved denne tilgang er at hvis der dukker overraskelser op (og det gør der over tid!), så er det altid muligt at genfinde de præcise records der er ansvarlige for de summerede beløb, og kildedata leverandøren kan lettere involveres i opklaringen, da denne vil føle sig ansvarlig for de leverede records (men ikke overfor modtagerens bearbejdninger af disse records).