Historiske massekilder

Erfaringer med edb-behandling af lensregnskaber

Hans Jørgen Marker

Fortid og Nutid december 1993, hefte 4, s. 265–283.

Ældre regnskabsmateriale er ofte omfattende og uoverskueligt at arbejde med. Det synes nærliggende at anvende moderne edb-teknik på at håndtere de mange enkeltoplysninger i regnskaberne. På baggrund af sit arbejde med edb-behandling af regnskaberne for to jyske len (Dronningborg og Kalø) fra 1600-tallets første halvdel gør Hans Jørgen Marker rede for nogle af de overvejelser, man bør gøre sig forud for sådanne undersøgelser, og giver i et tillæg praktiske eksempler for den, der selv vil give sig i kast med et tilsvarende arbejde. Hans beretning om de mange mere eller mindre fejlslagne forsøg undervejs illustrerer samtidig edb-teknikkens kolossale udvikling i løbet af den sidste halve snes år.

Hans Jørgen Marker, f. 1950, cand.mag. i historie og matematik, Aarhus Universitet. Universitetslektor (fra 1993 arkivar) ved Dansk Data Arkiv i Odense siden 1984.

Der er mange eksempler på kildegrupper, som indeholder materiale, der har en ensartet og gentagen struktur. Sådant materiale ville man i dag vælge statistiske metoder til at overskue. Der findes nogle kilder af denne type, som er frembragt, da statistiske metoder ikke var til rådighed. Hvis man anvender statistiske metoder på lidt ældre kilder, støder man ofte på en række praktiske og metodiske problemer. Nogle af disse vil jeg her forsøge at belyse.

Kilder fra førstatistisk tid er aldrig frembragt med statistik for øje. Grunden til, at disse kilder eksisterer, er altså, at de er frembragt med et andet formål. Dette andet formål skal man derfor tage i betragtning i sin omgang med kilderne. Disse overvejelser bliver ikke mindre komplekse, hvis man ønsker at sammenføre oplysninger fra kilder, som er frembragt med forskellige formål. Det sidste er en øvelse, som man kun bør indlade sig på med stor forsigtighed.

For at karakterisere hvilken type af kilder, jeg har i tankerne, vil jeg anvende begrebet *massekilder*. Ved massekilder forstår jeg kilder, hvori det er mere hensigtsmæssigt at behandle enkeltoplysningerne under ét ved hjælp af faste procedurer end at behandle dem enkeltvis. Hensigtsmæssigheden har to sider:

1. Arbejdets omfang kan medføre, at det er hensigtsmæssigt at anvende en formaliseret arbejdsform. Det gælder, når antallet af enkeltoplysninger er så stort, og når enkeltoplysningerne er tilstrækkelig ensartede til, at den tid, der medgår til at udvikle og afteste de procedurer, der skal analysere data, er rigeligt indvundet under anvendelsen af procedurerne.

2. Logisk konsistens er et andet vigtigt aspekt. Når en person udfører en række operationer, som indeholder et vist moment af skøn, er det en oplagt mulighed, at personens skøn vil være forskelligt til forskellige tidspunkter. Det vil medføre, at nøjagtig det samme spørgsmål kan få forskellige svar til forskellige tider. Derved kan der blive introduceret falske tendenser i materia-

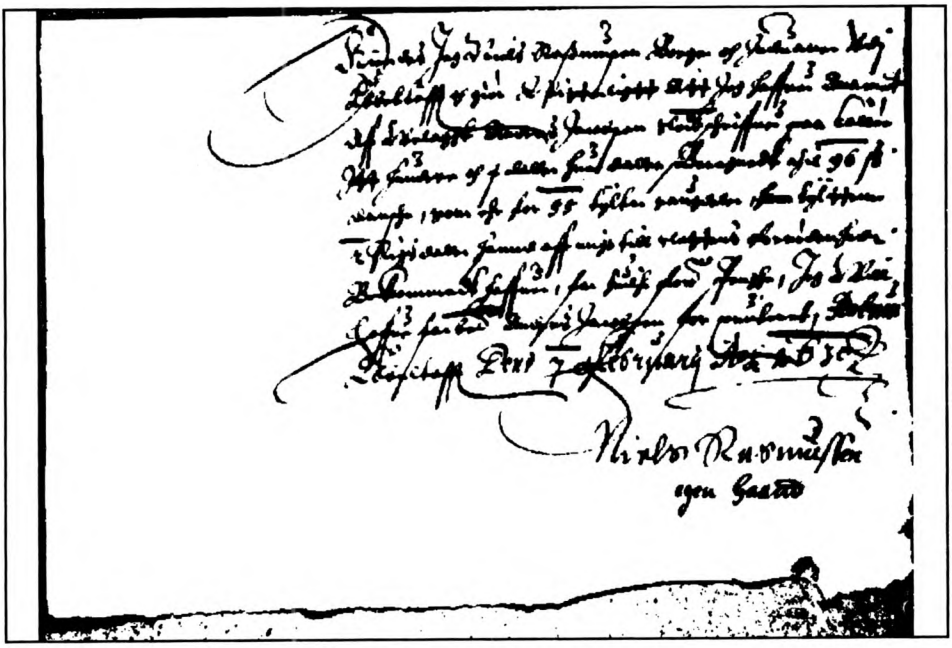

Fig. 1. Kvittering i Kalø lensregnskab 1629/30. Indscanning.

let. Problemet med logisk konsistens er mangetydigt. Der består absolut den mulighed, at man opnår et rigtigere resultat med en massekilde, hvis de mulige tolkningsmåder alle er repræsenteret, end hvis der er indført en konsistens, som kilden egentlig ikke danner grundlag for. Det er i det hele taget et problem ved anvendelsen af kvantitative metoder, at resultaterne kan fremtræde som mere sikre, end de er. Det er derfor en vigtig forpligtelse i præsentationen af resultaterne at gøre opmærksom på den usikkerhed, de er behæftet med.

Som massekilder her er defineret, ligger det i selve definitionen, at det er hensigtsmæssigt at skabe maskinlæsbare datamaterialer på grundlag af dem. De analyser, som massekilderne egner sig til, er netop sådanne, som understøttes effektivt af edb-værktøjer. Ved skabelsen af edb-læsbare datamaterialer løber man imidlertid ind i det kompleks af metodeproblemer, som en sådan maskinlæsbargørelse indeholder. Metodeproblemerne samler sig især omkring den fortolkning, som er en simpel følge af, at kilden gengives på et andet medium end det, originalen befinder sig på. Da datamaterialet nødvendigvis udgør en fortolkning af kilden, bør det være udstyret med en sådan beskrivelse, at almindelige krav til videnskabelighed indfris.[1]

Ved siden af metodeproblemerne bør man ved skabelsen af et datamateriale holde sig dets eventuelle genanvendelighed for øje. Genanvendelighed viser sig i praksis ikke kun at være genanvendelighed for andre brugere end dataproducenten. Den mest sandsynlige genbruger vil ofte være dataproducenten selv. Datamaterialer, der skabes med henblik på en specifik analyse, er i almindelighed redigerede med henblik på den analyse, de skal understøtte. Dette vil ofte begrænse deres anvendelighed til andre formål.

Kiendes Jeg Niels Rasmusen Borger och Induaner Vdj Ebbeltofft och giør Witterligtt Att Jeg haffuer Anamit Aff Welaggt Anders Jacobsen slodtschriffuer paa Calløe Jtt Hundrer och x daller huer daller Beregnedt thil 96 ß. dansche, som ehr for 55 tylter saugdeler huer tylttenn 2 Rigsdaller hannd aff mig till slottens fornødenhed Bekommedt haffuer, for huilche for(skrevne) Penghe, Jeg Will Haffue forbe(meldte) Anders Jacobsen for quiteret, Actum Ebbeltofft Dend 7 Februarij A(nn)o 1630
Niels Rasmussen
egen Haand

Fig. 2. Transskription.

Fortolkning

Når man overfører information fra et medium til et andet, sker der i denne proces en fortolkning af informationen. Fortolkningen kan være rent teknisk, som når en talværdi udtrykt i højere og lavere magnetiseringsniveauer på en diskette overføres til prikker og ikke prikker på en CD-ROM. Hvis man derimod transformerer nogle farveklatter på et stykke papir til elektrokemiske aktiviteter i et menneskes hjerne, og disse aktiviteter videre transformeres til mekaniske påvirkninger af et computertastatur, er sammenhængen mellem blækklatterne og tasteanslagene langtfra triviel. Ved transformationen fra trykt eller håndskrevet dokument til edb-læsbart materiale sker der fortolkning i to trin. Først vælger mennesket en forståelse af tegnene på papiret, dernæst tilrettelægges denne forståelse for computeren. Det første trin i fortolkningen kan man kalde læsning. Det andet trin af fortolkningen vil jeg kalde skrivning. Kilder kan ikke anvendes uden læsning. Derfor bliver læsningen ofte overset i teoretiske diskussioner af skabelsen af maskinlæsbare data. Maskinlæsbargørelsen diskuteres i en sprogbrug, der kun tager hensyn til skrivningen. Af denne årsag mener nogle, at det er muligt at tale om en kildegengivelse uden fortolkning.

I figur 1 ses en af de mulige gengivelser af en del af en kilde.[2] I denne er læsningen udskudt og erstattet af en maskinel reproduktion. En scanner har fortolket små arealer af papiret og har på denne baggrund besluttet, om de skulle være sorte eller hvide. Enhver vil nok medgive, at informationsindholdet i gengivelsen afviger noget fra informationsindholdet i originalen. Bedre tekniske hjælpemidler vil naturligvis kunne reducere denne forskel meget. Ikke desto mindre vil der stadig være tale om en reproduktion, som vil afvige fra kilden selv. Hvis kildens farve og skrift er fuldkommen perfekt gengivet, mangler vi måske stadig gengivelse af papirkvaliteten. Det skal dog også anføres, at man i nogle tilfælde kan opnå en læselighed af det scannede, som ikke er til stede i originalen.[3] Det kan for eksempel ske derved, at man forskyder kontrasten, således at udtværet blæk bliver hvidt, mens den kun lidt mørkere skrift bliver sort. Scanning er ikke nogen perfekt gengivelse, men alligevel er der visse muligheder for at udsætte gengivelsen for læsning på tilsvarende måde som originalen. Den har den fordel, at den er ret hurtig og billig at lave. Modsat har den den ulempe, at den kun giver meget begrænsede muligheder for edb-baseret analyse.

En anden mulig gengivelse ses i figur 2. I denne gengivelse, som man kan kalde transskription, er der tabt en hel del information i forhold til originalen. Der er tabt mere end ved indscanning. Hele den fysiske fremtrædelse af teksten er gået tabt. På den anden side er der foretaget en læsning af kilden, hvilket gør den mere tilgængelig. Man kan vel også sige, at en del af den infor-

kvittering$1/Niels/Rasmussen/borger og indvåner/55 tylter/saugdeller/110 rdlr/7 februari 1630 / Anders Jacobsen

Fig. 3. Tilrettelagt input.

mation, der gik tabt ved indscanningen, er overført i kraft af læsningen. Imidlertid sker overførslen af information nu i mennesketolket form. Det er nok vanskeligere i reproduktionen end i forlægget at se, at det overstregede ord efter *saugdeler* er *huer*. Det kan videregives i transskriptionen. På den anden side er læseren af transskriptionen totalt afskåret fra at forkaste den fortolkning, der er udtrykt i transskriptionen, eller at finde grundlag for en alternativ fortolkning uden at gå tilbage til kilden. Transskriptionen åbner visse muligheder for edb-baseret analyse, men den tilgængelige software understøtter kun få typer af analyser tilfredsstillende. Transskriptionen understøtter naturligvis fremfinding ved hjælp af tekstsøgning. Til tekstfremfinding ville det dog være mere hensigtsmæssigt at have en transskription med normaliseret stavemåde. Til kvantitative analyser, hvori denne oplysning indgår, egner transskriptionen sig meget dårligt. Der skal almindeligvis programmeres så meget, at det vil være enklere at indtaste oplysningen igen.

Transskriptionen kan udbygges med markup,[4] som afmærker bestemte betydningskategorier i teksten. Markup retter sig dog mere mod tekstanalyse end mod statistisk analyse. Der er megen tale om markup som udgangspunkt for konstruktion af strukturerede datamaterialer. Der har også tidligt været gjort forsøg hermed,[5] men det har drejet sig om mammutprojekter, der i karakter afveg meget fra enkeltforskerens arbejde med sit lille, private datamateriale. Hvis markup skal være relevant for almindeligt forskningsarbejde, mangler der stadig nogle redskaber.

I figur 3 ses et meget fortolket format. Efter læsningen af kilden er de »væsentlige« oplysninger blevet skrevet i et format, der kan læses af et program. Formatet vil være afhængigt af programmet. Der sker naturligvis en tilsvarende forenkling af informationen i kilden, hvis dataindtastningen er programunderstøttet, og indtastningen sker i et indlæsningsskærmbillede. I et udtog som det her viste er fortolkningen færdig, og edb-behandlingen kan begynde. I og med at den »væsentlige« information er uddraget, er en hel del triviel og »betydningsløs« information bortkastet. Mange af de naturlige analyser, denne kilde kan udsættes for, har fuldt tilstrækkelig information i uddraget. Alligevel er der noget i udtrykkene »væsentlig« og »betydningsløs«, der virker skræmmende. Problemet er, at den, der skabte datamaterialet, for altid har bortkastet den information, som han ikke selv finder relevant for sin analyse. F.eks. har han ikke medtaget, at Anders Jacobsen var slotsskriver på Kalø. Det ved han så udmærket, for Anders Jacobsen går igen på mange kvitteringer, men den næste, der vil bruge materialet, har muligvis ikke denne viden. Materialet er i denne udgave uegnet til en undersøgelse af stillingsbetegnelser og navneskik. Den, der vil studere stillingsbetegnelser og navneskik på grundlag af Kalø lensregnskab, er altså henvist til at skabe et nyt datamateriale. I teorien skulle dette nye materiale kunne sammenføres med det økonomisk orienterede materiale, der ville blive resultatet af en kodning som i figur 3. Derved ville der blive skabt et mere komplet billede. Jeg kender dog ingen eksempler på, at to uafhængige datamaterialer skabt over samme kilde er blevet sammenført. Når man anvender

tilrettelagt input, foretages der ofte en tolkning eller normalisering samtidig med indtastningen. Normaliseringen sker ved, at man ud over at læse kilden vælger en bestemt gengivelse af det læste, f.eks. laver »Rasmusen« om til »Rasmussen«. Det kan ofte være enklere at gengive normaliserede former end at finde den præcise læsning; på den anden side kan man reducere informationstabet ved at medtage oplysningerne i deres oprindelige form sammen med den normaliserede form.

Der er ikke ved nogen af de nævnte gengivelser tale om, at datamaterialet træder i stedet for kilden på en måde, der gør originalen overflødig. Det mener jeg nu heller ikke, er noget ideal for kildeudgaver.

Prisdata fra tidlig nyere tid

Det, jeg interesserer mig for her, er prisdata fra tiden før Enevælden. En begrundelse for at interessere sig for disse data er, at der findes mange af dem. I det hele taget er det kendetegnende for 16. og 17. århundrede, at der findes masser af eksakte, kvantitative oplysninger i kilderne, men meget få aggregerede størrelser. Der er altså priser, men ikke prisstatistik. Imidlertid er netop de aggregerede værdier meget centrale for vor tids forståelse af samfundsudviklingen.

Det er elementært fristende at forsøge at bruge de mange tal til at skabe de aggregerede størrelser, som datiden ikke har skabt. I mange sammenhænge viser et sådant projekt sig dog at være meget omfattende; ikke mindst er det vanskeligt at sikre og kontrollere resultaternes repræsentativitet. (Her skal det dog i parentes bemærkes, at nutidens offentlige statistik ikke altid lader sig sidestille med Cæsars hustru. Vor tids statistik kan være mere problematisk, end den udgiver sig for at være.) Det, der kan lade sig gøre på baggrund af kilderne, er at skabe aggregerede data, der har udsagnskraft over for et veldefineret sagsforhold i datiden, f.eks. prisudviklingen for udspisningen på et bestemt len eller lignende. I hvilket omfang, sådanne rækker lader sig generalisere, er straks mere diskutabelt.

Nu kan den afgrænsede problematik naturligvis i sig selv være af interesse. Videre er mangelen på gode prisrækker så påtrængende, at forskningen betjener sig af rækker af meget ringe almengyldighed såsom de sjællandske kapitelstakster for rug[6] eller, især i ældre tid, sølvværdien i mønterne. I nogle tilfælde forfalder man til en rent anekdotisk anvendelse af priser.[7] Man bruger et isoleret kildested, der angiver en pris, som et generelt udtryk for prisen på den pågældende vare. Som alternativ til sådanne løsninger er det af nogen interesse at have prisrækker af ganske vist begrænset, men dog veldefineret repræsentativitet.

Fra 16. og 17. århundrede er der bevaret såvel offentlige som private regnskaber. Endvidere findes der en del taksationer, der er fremstillet med det formål at beskrive eller påvirke prisdannelsen. Endelig findes der tilfældige prisoplysninger i breve, dagbøger etc. I denne buket af muligheder har jeg valgt at interessere mig for regnskabsmaterialet, herunder specielt lensregnskaberne. De priser, jeg arbejder med, er sådanne, som har været anvendt i virkelige transaktioner. Det vil sige, at det er priser, der har været lagt til grund i situationer, hvor en bestemt, nærmere angivet varemængde har ændret ejerforhold. Denne type priser er absolut dominerende i lensregnskaberne. Yderligere er den ene part i transaktionen altid lenet, hvilket giver nogen konstans i det niveau, hvorpå der handles, i det mindste når man ser på samme vare. Endelig er regnskaberne reviderede i samtiden, hvilket giver en rimelig grad af troværdighed af de meddelte oplysninger.

De data, der indgår i mit datamateriale, stammer fra flere kilder. De to tredjedele er dog fra Kalø og Dronningborg lens regnskaber. Datamaterialet er arkiveret i Dansk Data Arkiv som DDA-1066: *Priser og lønninger fra Østjylland 1571–1661*. Som arbejdet er skredet frem, er undersøgelsen blevet udvidet, men titlen rammer stadig det centrale. Den overvejende del af oplysningerne i materialet er fra Østjylland og fra første halvdel af 17. århundrede. Der forekommer priser i materialet fra andre dele af landet end Østjylland. De ældste priser er fra 1487, og de yngste er fra 1660. Det er planen i senere faser af arbejdet at udbygge materialet med flere priser og lønninger. Bestræbelsen er dels at dække større dele af landet, dels at få en bedre dækning af 16. århundrede. Fra 16. århundrede er der næsten ikke bevaret lensregnskaber, hvilket gør det vanskeligere at samle et sammenhængende materiale.

Dataindlæsningen

Gengivelsen i figur 3 er den type indlæsningsformat, der anvendes i programpakken *Kleio*. Kleio er et databaseprogrammel for historikere. Det er udviklet i et samarbejde omkring Max-Planck-Institut für Geschichte i Göttingen. Kleio udmærker sig ved at være grimt og utilgængeligt og at kunne alting. Programpakken kan i Danmark erhverves fra Dansk Data Arkiv. Da jeg grundlagde mit datamateriale, eksisterede Kleio ikke. Jeg var derfor ikke stillet over for afgørelsen af, om Kleio var et egnet redskab til mit formål. Det, jeg havde til disposition til indlæsningen af mit materiale, var en 64 kB CP/M 2.2 computer. CP/M betyder Control Program for Microcomputers. Dette styresystem blev produceret af firmaet Digital Research. CP/M 2.2 var det mest udbredte styresystem for microcomputere i tiden, før pc'erne kom frem, og 64 kB RAM var almindeligvis den største bestykning med intern hukommelse for disse maskiner. De mindste maskiner, der sælges i dag, har 2 MB RAM. 2 MB er 2048 kB eller 32 gange mere end de største CP/M-maskiners interne hukommelse. Til den anvendte computer havde jeg en højniveausprogcompiler, hvormed jeg kunne skrive indlæsningsprogrammer. Disse programmer var helt enkle og styrede alene feltlængderne. Data blev lagret på disketter som lister i et tegnformat. Med datidens diskettestørrelser betød det anvendelsen af omkring 80 disketter. Samtidig med indlæsningen blev der taget kopier af de filmsider, hvorfra oplysningerne stammer. Disse kopier har vist sig at være uvurderlige under datarensningen.

Originalfilerne blev dannet i to formater, et for priser og et for lønninger. Målet var i dataindlæsningsfasen at tilvejebringe så mange oplysninger som muligt, frem for at strukturere oplysningerne til analyse. Med den viden om datamaterialet, jeg nu har, og med de redskaber, der i dag står til rådighed, ville jeg nok vælge et indlæsningsformat, som i højere grad ville tilgodese begge hensyn. I dag har man et meget større spektrum af redskaber til rådighed for tilrettelæggelsen af en sådan dataindskrivning, end man havde i begyndelsen af 1980'erne. Den oprindelige prioritering var imidlertid at få mange oplysninger lokaliseret, læst og indskrevet. Det var endvidere en eksplicit forudsætning, at fortolkningen af de indlæste oplysninger hovedsagelig skulle lægges i datarensningsfasen.

Rensning og normalisering

Fra disketterne blev data flyttet op til en mainframe. På mainframen var SAS til rådighed. SAS, Statistical Analysis System, er en udbredt programpakke til dataanalyse og statistik. Jeg ville i teorien have kunnet gennemføre data-

rensningen på mainframen i det, der blev den endelige arbejdsform. Nu er det desværre sådan, at mainframes er nogle uvenlige og langsommelige bæster, som giver lange svartider. Det fortrin, som mainframen måtte have i regnekraft, sættes til i kampen med dens umulige brugergrænseflade og øvrige obstruerende udenværker. Disse lidet tilfredsstillende arbejdsforhold førte mig ud i en række eksperimenter med placeringen af data. Af disse har nok især de fejlslagne interesse for læseren. I arbejdet med massedata kan man bruge lang tid på at opdage en fejl, især hvis fejlen består i, at den valgte arbejdsmåde vil tage for lang tid.

Den del af data, der stammer fra Kalø len, blev samlet i et SAS-datasæt på mainframen. Da jeg i nogen tid havde forsøgt at gennemføre datarensningen i SAS på mainframe, blev data overført til en OSIRIS-fil. OSIRIS var navnet på en ikke længere eksisterende statistikpakke. Det filformat, som OSIRIS anvendte, har dannet grundlag for det format, som dataarkiverne anvender til opbevaring af deres datamaterialer. OSIRIS-filen blev tilgået direkte ved hjælp af programmer, som blev skrevet til formålet. Dette projekt kunne muligvis have været gennemført. Imidlertid bliver et programmeringsprojekt af den størrelsesorden, der her var tale om, let afsporet. Dette projekt blev drejet i retning af et forsøg på at skabe forbindelse mellem programpakken Kleios systemfiler og Dansk Data Arkivs arkiveringsformat OSIRIS. Disse to verdener viste sig dog at være uforligelige. Men jeg har fået at vide, at Kleio i dag indeholder visse rester fra mit opgivne programmeringsprojekt.

I et prismateriale fra den periode, der her er tale om, vil der være elementer af ensartethed og elementer af uensartethed. Det vil ydermere være sådan, at når man behandler materialet ud fra metoder, der forudsætter ensartethed, vil de uensartede træk være dem, der er mest iøjnefaldende, fordi de kræver mest hensyntagen ved behandlingen. Omvendt vil de ensartede træk ved materialet være meget indlysende, når man behandler materialet med individuelt rettede metoder, idet de ensartede dele af materialet bringer arbejdet ind i trivielle gentagelser. Da jeg altså havde foretaget to forsøg på at gennemføre arbejdet med datamaterialet med metoder, der forudsatte ensartethed, var jeg helt overbevist om, at datamaterialet var for uensartet til at kunne behandles hensigtsmæssigt med disse metoder. Det næste forsøg blev derfor at indlæse materialet i en database med henblik på at behandle enkeltoplysningerne enkeltvis.

Valget af database var i nogen grad styret af, at programmellet skulle været et, jeg kendte og havde til rådighed. Det var også af betydning for valget, at jeg ønskede at anvende et produkt, hvorfra data på enkel vis kunne flyttes til regneark og tekstbehandling. I 1987 var dette et mere snærende krav end i dag, hvor dataudveksling mellem forskellige programmer er temmelig godt understøttet. Et af de mulige valg var at bruge WordPerfect, PlanPerfect og DataPerfect.

Efter indlæsning i DataPerfect blev prisoplysningerne placeret i en simpel flad fil. De normaliserede størrelser blev beregnet og indlæst manuelt for hver enkelt oplysning. Denne proces var meget arbejdsintensiv, hvilket var årsagen til, at jeg opgav den, da den var 40% gennemført. Problemerne er almindelig kendte for historikere, der beskæftiger sig med mønt og mål fra nyere tid. Møntforholdene ændrede sig igennem perioden, og det kan i perioder være vanskeligt at afgøre, hvad betegnelserne dækker over.[8] Det sidste gælder også for de angivne mål.

De enkeltoplysninger, som blev færdig datarenset og normaliseret i databasen, er for en dels vedkommende spredt ud igennem materialet, men alle priser

Label	Variabelnavn	Værdi
År	AAR	1630
Identifikation	IDENT	K30R87
Nummer fra indlæsningen	NUMMER	2237
Normaliseret betegnelse	NORMBET	Rug
Normaliseret enhed	ENHED	tdr
Normaliseret mængde	TOT–MAEN	0.000
Samlet pris i skilling	TOT–SK	0.000
Normaliseret enhedspris	ENH–PR	0.000
Betegnelse som indlæst	OPR–BET	Rug og mel
Mængde som indlæst	OPR–MAEN	15.5 læst 8.5 td 1.5 skæp 1 fk
Enhedspris som indlæst	OPR–E–PR	tønden 3 rdl
Samlet pris som indlæst	OPR–T–PR	1421 rdl 15 sk
Køber	KOEBER	Bønderne
Købers hjemsted	KOEB–HJM	KL
Sælger	SAELGER	KL
Sælgers hjemsted	SAEL–HJM	KL
Kilde	KILDE	
Kommentar	KOMMENT	p. opboren aff bønderne
	OTP01	1421
	OTP02	rdlr
	OTP03	15
	OTP04	sk
	OTP05	
	OTP06	
	OTP07	
	OTP08	
	OTP09	
	OTP10	
Bemærkninger til totalpris	OTPKOM	
Totalpris i skilling, beregnet	NOTP	136431.00
	OTM01	15.5
	OTM02	læster
	OTM03	8.5
	OTM04	tdr
	OTM05	1.5
	OTM06	skp
	OTM07	1
	OTM08	fjk
	OTM09	
	OTM10	
Bemærkninger til total mængde	OTMKOM	
Målesystem i mængdeangivelsen	MAALSYST	rug
Total mængde, beregnet	NOTM	473.719
	EHP01	1
	EHP02	tdr
	EHP03	3
	EHP04	rdlr
	EHP05	
	EHP06	
	EHP07	
	EHP08	
	EHP09	
	EHP10	
	EHP–PS01	3
	EHP–PS02	rdlr
	EHP–PS03	
	EHP–PS04	
	EHP–PS05	
	EHP–PS06	
	EHP–MS01	1
	EHP–MS02	tdr
	EHP–MS03	
	EHP–MS04	
Bemærkninger til enhedspris	EHPKOM	
Prisdelen af enheds-pris i skilling	EHP–PT	288.000
Mængdedelen af enhedspris	EHP–MT	1.000
Normaliseret enhedspris	NEHP	288.000

Fig. 4. En enkeltoplysning fra datamaterialet.

fra 1632 og frem blev færdiggjort med disse metoder.

Valget af DataPerfect som database var i særlig grad årsag til, at dette eksperiment blev fejlslagent. Imidlertid har materialet også haft en tur i Borlands Paradox, inden jeg besluttede mig for at sende det tilbage til SAS. Paradox kunne nok have løst opgaven, idet programmeringssproget i Paradox effektivt understøtter generel behandling af data. Microsoft Excel har også fået lov til at snuse til datamaterialet. Da den seneste version af dette program understøtter krydstabuleringer, kunne det jo være.... Excel ville dog kun acceptere halvdelen af datamaterialet, og efter at have kørt i det meste af en weekend på en 20 MHz 386'er med 4 MB RAM opgav Excel totalt at lave en krydstabel af to

variable i et materiale med 4.500 enkeltoplysninger.

Da jeg således næsten var nået halvvejen med datarensningen, indså jeg sidst i efteråret 1992, at den enkeltobservationsbaserede metode ikke egnede sig til mine data. Imidlertid har de mange forskelligartede eksperimenter, som datamaterialet har været udsat for, været mig til nytte i andre sammenhænge. Den indvundne afklaring af forskellige metoders og redskabers anvendelighed til forskellige typer af datamaterialer kan nok også være af en vis generel interesse. Omvendt må man sige, at med den viden og de redskaber, der i dag står til rådighed, er den rimelige tid for normaliseringen af et materiale af denne karakter og størrelse de ca. 200 timer, som jeg har lagt i mit datamateriale siden efteråret 1992.

Datamaterialet blev derfor udskrevet på en listeform, der omtrent svarede til det oprindelige indlæsningsformat, dog med opretholdelse af resultaterne af den foretagne manuelle normalisering. Det listeformede datamateriale blev indlæst i SAS. Valget af SAS kan muligvis have et element af tilfældighed over sig. Tilfældigheden består i, at jeg i detaljer gennemskuede, hvorledes jeg i SAS skulle få normaliseret de tekststrenge, som de oprindelige oplysninger består af, mens det samme ikke gik op for mig i Paradox. Placeringen af det rensede datamateriale i SAS med henblik på analysen har derimod ikke det samme præg af tilfældighed. De relevante analyser, som man kan underkaste et materiale som det foreliggende, er langt bedre understøttet af SAS end af noget databaseprogram.

Den maskinelle normalisering er nu gennemført for mængde, enhedspris og totalpris. I alle tilfælde er den grundliggende metode den samme. Først er tekststrengen opdelt i enkelte ord. Dernæst er disse ord normaliseret, således at første ord er et tal, andet ord en betegnelse udtrykt i et kontrolleret ordforråd, tredje ord igen et tal, osv. Ved kontrolleret ordforråd forstås et ordforråd, der ikke indeholder synonymer, og i dette tilfælde kun har ét tal, flertal. Der er således kun ét ord for tønde, td., tønder, etc., nemlig *tdr.* I intet tilfælde var antallet af ord større end ti, men det er naturligvis noget, man må undersøge separat.

I figur 4 ses en oplysning fra datamaterialet i dets endelige form. I første kolonne af udskriften står *Label*, betegnelsen for den pågældende variabel. Anden kolonne indeholder variablens navn i programmet og sidste kolonne variablens værdi. Lensregnskabsåret 1630/31 er forkortet til 1630. Identifikationen er en kodet henvisning til kilden: *K30* angiver Kalø lens regnskab 1630/31. *R87* angiver folio 87 af regnskabet. I normaliseret betegnelse er »Rug og mel« blevet betegnet som *rug*. Køberen er angivet som *Bønderne*. OTP01-OTP10 er opdelingen af oprindelig totalpris i enkelte ord. OTM01-OTM10 er opdelingen af oprindelig totalmængde på enkelte ord. EHP01-EHP10 er opsplitningen af oprindelig enhedspris. Til beregning af mængderne er variablen MAALSYST indført. Denne variabel opdeler materialet i hovedgrupper, inden for hvilke der som regel gælder det samme målesystem. Sådanne grupper er rum-, længde- og vægtmål. Nogle varegrupper med særlige målesystemer behandles separat, herunder kornsorterne, papir, brænde, etc. Prisdelen af enhedsprisen er overført til et nyt sæt mellemvariable EHP–PS01 – EHP–PS06, ligesom mængdedelen er flyttet til EHP–MS01 – EHP–MS04. EHP–PS og EHP–MS anvendes ved automatiske kontroller. I perioden 1602–1610 er det f.eks. meget almindeligt, at rdlr (rigsdaler) er angivet som dlr (daler). Det kan imidlertid let kontrolleres af et program. Når blot man har mængde, enhedspris og totalpris, kan man lade programmet sammenligne totalprisen med produktet af en-

```
2000401071700287000100121MARIANE-C.              KAARUP              3
2000401071700287000100123367       004000410  010101      3
```

Fig. 5. Gennemkodede data. Eksemplet er fra DDA-1447: Folketællinger fra Odense 1875–1911, produceret og arkiveret af Per Boje, Historisk Institut, Odense Universitet.

hedspris og mængde. Hvis sammenligningen ikke stemmer, kan programmet beregne sammenligningen ud fra antagelsen, at dlr skal være rdlr, mk (mark) skal være rmk (rigsmark), etc. Hvis sammenligningen passer under denne antagelse, kan man lade programmet indføre rettelserne og indskrive en kommentar herom. En sådan kontrol er udført på datamaterialet. Flere kontroller af denne type er mulige, f.eks. for de forskellige tøndeantal på læsterne, der kan forekomme i kornmålene.

Ved siden af automatiske kontroller er der den mulighed at udskrive de fejlagtige enkeltoplysninger og behandle dem med individuelle metoder.

De kodninger, som er foretaget i figur 4, er trivielle og kan formentlig opløses uden bistand fra en kodenøgle. Kodning kan være drevet betydeligt videre, som det ses i figur 5. Denne type data kan kun fortolkes ved hjælp af en kodenøgle. Fordelene ved at placere data i et sådant format er, at materialet er meget let at analysere i en statistikpakke, og at det ikke kan indeholde ambivalenser. Det største problem er nok, at tolkningerne næsten udelukkende er lagt forud for datamaterialets tilblivelse, så det kan være problematisk at dokumentere tolkningerne. Egentlig er gennemkodede data ikke principielt forskellige fra data, som er normaliseret til et kontrolleret ordforråd. Forskellene er hovedsagelig, at nødvendigheden af dokumentation er mere indlysende for gennemkodede data, og at gennemkodede data sparer plads og giver bedre svartider ved visse typer software. Den løsning, som i dag forekommer naturlig, er at operere med såvel en oprindelig som en kodet gengivelse af den samme oplysning. Den kodede værdi kan hensigtsmæssigt konstrueres maskinelt ud fra værdien i den oprindelige (og eventuelt andre variable). Derved opnås, at kodningen bliver konsekvent, og at forkert kodning kan omgøres på enkel vis. Det ligger inden for mine overvejelser at indføre mere kodning i mit datamateriale.

Dataanalyse

I 1989 lavede jeg en lille analyse af prisudviklingen i 1640'erne, som er publiceret andetsteds.[9] Fra dette arbejde og fra et større, upubliceret arbejde[10] kan udledes følgende generelle beskrivelse af den analyse, som materialet nu er ved at gennemgå.

Til studier af prisudviklingen er det hensigtsmæssigt at konstruere prisindekser. Til at udarbejde prisindeks har man brug for en vægtningsmængde, således at de enkelte varer kan vægtes i forhold til hinanden. Eller mere populært: Når man vil lægge sild og øl sammen, lader det sig kun gøre, fordi begge varer kan udtrykkes ved deres værdi i penge, f.eks. skilling. Man skal dog vide hvor mange tønder sild og hvor mange tønder øl, der skal indgå i regnestykket, for at resultatet har mening. De mængder, der skal anvendes i indeksberegningen, kaldes indeksmængder. Hvert enkelt tal i listen kaldes den pågældende vares vægt.

Det er min opfattelse, at man, hvis man har et tilstrækkeligt stort og varieret prismateriale, vil kunne bruge selve materialet til at konstruere vægtningsmængden. Hvis materialet er tilstrækkeligt stort og varieret, vil nemlig de mængder, der forekommer i materialet, afspejle den samlede omsætning i

År	Antal observationer	Største værdi	Mindste værdi	Samlet mængde	Vægtet enhedspris
1602	1	128.000	128.000	637.453	128.000
1604	1	96.000	96.000	55.000	96.000
1607	1	48.848	48.848	2283.000	48.848
1608	3	144.000	80.000	656.922	120.432
1609	2	144.000	88.354	241.000	91.817
1610	5	96.000	70.400	1650.438	83.133
1611	6	101.750	54.370	1009.688	69.708
1612	6	148.000	70.469	4214.500	93.390
1613	2	88.000	80.000	186.906	82.571
1614	1	66.606	66.606	1572.000	66.606
1615	5	148.000	63.513	2119.906	139.395
1616	3	160.000	128.000	1310.375	152.416
1617	5	140.000	80.051	2023.000	107.721
1618	2	136.500	79.947	1258.374	95.738
1619	7	384.000	72.011	2262.125	101.137
1620	5	96.000	75.257	1919.234	83.885
1621	4	89.600	75.200	850.000	85.252
1622	1	128.000	128.000	261.828	128.019
1623	6	256.000	168.000	1246.922	223.442
1624	3	282.000	256.000	486.438	274.903
1625	2	176.000	160.000	400.000	171.997
1626	6	192.000	192.000	738.990	192.008
1627	3	384.000	192.000	287.000	193.338
1628	2	192.000	128.000	122.188	190.429
1629	9	352.000	288.000	180.313	294.045
1630	7	352.000	192.000	1091.969	255.728
1631	6	288.000	160.000	1137.906	179.706
1632	3	352.000	160.000	148.000	170.378
1633	7	240.000	160.000	356.719	195.407
1634	9	256.000	192.000	575.063	202.966
1635	8	252.000	156.000	616.500	171.406
1636	12	192.000	160.000	1253.156	166.598
1637	8	192.000	168.000	1178.375	187.927
1638	9	192.000	144.000	1020.156	167.421
1639	6	272.000	192.000	664.969	200.061
1640	4	192.000	144.000	1134.375	169.489
1641	2	192.000	192.000	99.906	192.000
1642	3	192.000	176.000	356.750	178.781
1643	2	192.000	160.000	235.594	180.854
1644	1	160.000	160.000	134.172	160.000
1645	4	192.000	159.680	592.234	186.543
1646	2	192.000	160.000	184.000	166.348
1647	4	208.000	160.000	1807.725	200.781
1648	3	224.000	192.000	193.500	194.315
1649	1	256.000	256.000	876.813	256.000
1653	3	144.000	144.000	8.000	144.000
1655	4	72.000	68.000	738.188	69.387
1660	1	224.000	224.000	132.000	224.000

Fig. 6. Gennemsnitlige årspriser for rug, pris pr. tønde i skilling (med tre decimaler).

det beskrevne univers. Hvis man arbejder med et enkelt regnskab, som er komplet bevaret, er forudsætningen triviel. Man har alle transaktionerne og derfor en fuldstændig beskrivelse af omsætningsmængden. Hvis man har priser, som er taget ud af en større sammenhæng (f.eks. vareomsætningen i Østjylland) bliver det mere vanskeligt at afgøre, om uddraget er repræsenta-

tivt. Vi kender jo netop ikke den samlede størrelse og sammensætning af vareomsætningen i Østjylland. Problemet er nært beslægtet med traditionelle statistiske problemer som f.eks. at tælle torskene i Nordsøen. Det er værd at lægge mærke til, at udfaldet af overvejelserne over en given datamængdes repræsentativitet er afhængigt af, hvilken brug man vil gøre af resultatet. Har man skaffet sig rimelig baggrund for at mene, at man har et prismateriale, der afspejler den samlede omsætning i det undersøgte område i den undersøgte periode, vil man kunne opdele det i kortere, overlappende tidsrum, sammentælle de samlede mængder for hvert af disse og bruge summerne som vægtningsmængder for de enkelte tidsrum. Perioderne skal være så korte, at omsætningen kan anses for at have en konstant sammensætning inden for hver af dem. Overlappet mellem dem skal anvendes til at sætte periodeindekserne sammen til et længere prisindeks. En sådan teknik vil jeg betegne som intern vægtning.

Ved siden af intern vægtning kan der være muligheder for ekstern vægtning. F.eks. kan man med en udspisningstakst vise udviklingen i fødevarepriserne. Det, der behøves for at konstruere en ekstern vægt, er en samlet liste over varemængder. Når man anvender en sådan liste som vægt, må man derefter forholde sig til, hvad det er, man har fået indekseret.

Med materialet i dets nuværende tilstand er det meget enkelt at konstruere årsgennemsnitspriser og foretage beregninger på grundlag af dem. Hvor SAS med noget besvær lod sig overtale til at medvirke til normaliseringen af data, er programmet på hjemmebane i sorteringer og beregninger.

Materialet indeholder 501 forskellige varer, der giver anledning til 4.033 årsgennemsnitspriser. Udskrevet bliver det en tabel på ca. 120 sider. I figur 6 er vist årsgennemsnitspriserne for rug efter 1600.

Som tabel 6 fremtræder, er der ingen tydelige tegn på manglende normalisering. En værdi, der giver anledning til mistanke, er prisen for 1655. Men faktisk var rugpriserne så lave det år. Kapitelstaksten for Århus stift var 4 mk/td,[11] hvilket er i nydelig overenstemmelse med de værdier, jeg har i alle fire enkeltoplysninger fra 1655. Tabeller som den foreliggende er et redskab i datarensningen. I mange tilfælde kan de afsløre systematiske fejl i normaliseringen. I så fald er den procedurebaserede arbejdsform den enkeltobservationsbaserede langt overlegen. En systematisk fejl kan oprettes på en eftermiddag, når det, der skal gøres, blot er at rette og køre nogle procedurer. Hvis hundreder af enkeltobservationer skal opspores og rettes, som det kan være tilfældet ved enkeltobservationsmetoder, bliver fejlretningen mere arbejdskrævende.

Kildetyper og dataformater

Det er indlysende, at hver enkelt type kodning egner sig bedre til nogle datamaterialer end til andre. De parametre, der styrer egnetheden, er graden af gentagelse i kildens struktur, datamaterialets samlede størrelse og den tilsigtede analyse. Det har også betydning for valget af dataformat, om kilden gengives i sin fulde ordlyd, eller om der foretages uddrag fra den. Modsat det, der i visse kredse har været hævdet, mener jeg ikke, at totale kildeudgaver skal fremhæves som et altoverskyggende ideal. Det gælder lige fuldt i dag som for ti år siden, at edb-anvendelse i historiefaget skal være resultatrettet, og der er andre relevante resultater end skabelsen af genanvendelige datamaterialer. Når man foretager ekstrakter fra en kilde, kan det ofte forekomme, at uanset, at kildens struktur er meget varieret og

kompleks, er ekstrakterne af ensartet struktur og derfor velegnede til indlæggelse i et databaseformat. Normalt vil markup eller tilsvarende tekstbaserede formater være mest velegnede til kilder med en meget kompleks eller varierende struktur. Denne egnethed er dog under forudsætning af, at det materiale, som tænkes behandlet, er af en overkommelig størrelse. Et særligt og vel egentlig uløst problem rejser sig ved meget store og samtidig meget komplekse kilder. Som eksempel kunne man tænke sig brevvekslingen mellem et ministerium og omverdenen i en årrække. En sådan kilde kunne tænkes på forhånd at være maskinlæsbar, men det er ikke indlysende, hvilke redskaber der hensigtsmæssigt kan tages i anvendelse for at analysere den. Problemet er, at selv om samlingen af breve er meget interessant, er de fleste af de enkelte breve isoleret set mindre betydningsfulde. Når der er adskillige tusinder breve, kan man ikke investere flere minutter i hvert. Jeg er blevet fortalt, at der i efterretningsvæsenerne er udviklet programmel med henblik på at løse problemer af denne type, men det er mig ikke bekendt, at sådant programmel er alment tilgængeligt.

Valg af dataformat

Det er altså min påstand, at der ikke findes nogen ufortolket omdannelse af en kilde til et datamateriale. Datamaterialet er en behandling af kilden, og behandlingen bliver produceret med et formål for øje. I valget af dataformat er det således vigtigt, at man gør sig klart, hvilke konsekvenser valget har. Det har betydning for hvilken informationsmængde, der vil være til stede, når datamaterialet underkastes analyse. Valget har også konsekvenser for hvilke vanskeligheder, bestemte analysetyper vil møde.

En transskription i en statistikpakke fordrer mange linier programkode, inden der kommer resultater frem. Omvendt vil en gennemkodet udgave af en kilde ikke levne meget til tekstorienterede analyser.

Ressourceproblematikken skal også tages i betragtning ved valg af dataformat. I edb-baserede projekter fra 1960'erne og 1970'erne var det meget eksplicit, at økonomien var medbestemmende i undersøgelsens design. Datidens edb-projekter var ofte meget kostbare, og edb-ressourcer var erkendt som knappe ressourcer. Det var altså naturligt, at man overvejede lager- og kørselsøkonomi i forbindelse med historiske edb-projekter, ligesom man gjorde det i forbindelse med al anden edb-anvendelse. I dag er edb-ressourcerne nærmest uendelige i sammenligning med de ressourcer, der kunne disponeres over for to årtier siden. Derfor bliver ressourceovervejelserne ikke taget fuldt så alvorligt i dag. Imidlertid er arbejdstid også i dag en virkelig knap ressource. Ikke mindst når det drejer sig om store materialer, herunder sådanne materialer som produceres med henblik på kvantitativ analyse, må dataformatet vælges med skyldigt hensyn til tidsforbruget. Selv ved et materiale af en så relativt beskeden størrelse som 9.000 enkeltoplysninger vil to minutter brugt på hver enkeltoplysning blive til 300 timer. Hvis man bruger 15 minutter på at lave markup af hver enkeltoplysning, bliver der ved 9.000 enkeltoplysninger disponeret 2.250 timer. Det er et ret omfattende projekt, og de kvantitativt orienterede analysemuligheder er begrænsede. Hvis et projekt modsat beskæftiger sig med syv breve, er markup-transskription muligvis netop den ideelle måde at gengive data.

Helt centralt er det dog, at man gør sig klart, at man ved frembringelsen af et datamateriale på grundlag af en kilde giver en fortolkning af kilden.

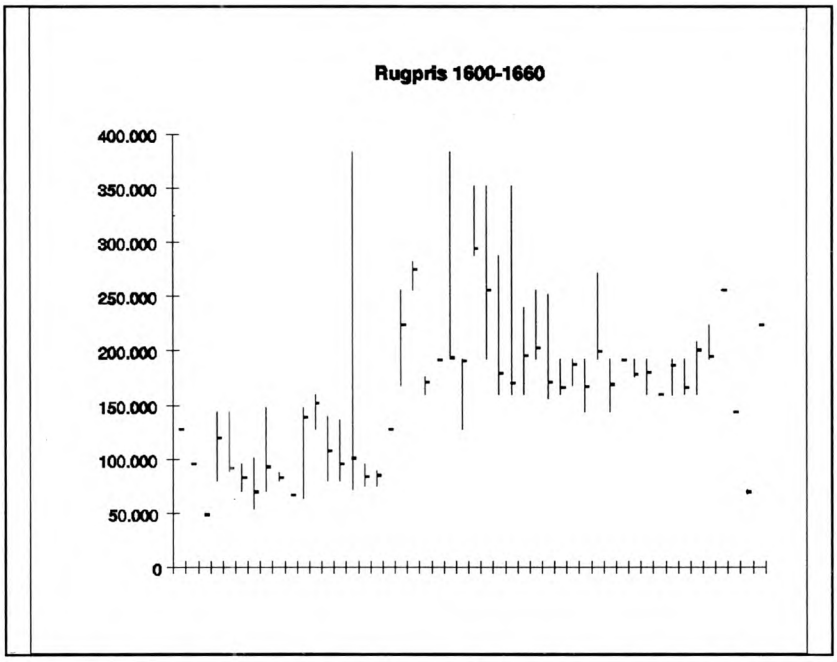

Fig. 7. Grafisk fremstilling af rugprisudviklingen på basis af figur 6. De sorte firkanter markerer årets gennemsnitspris, mens de lodrette linjer anskueliggør spændet mellem årets højeste og laveste pris. Prisskala: Antal skilling pr. tønde (med tre decimaler).

Denne fortolkningsproces giver anledning til samme niveau af metodeovervejelser som enhver anden historievidenskabelig arbejdsproces. Det er centralt, at den, der læser fortolkningen, kan skelne mellem de forskellige grader af fortolkning, der er foretaget. Især vil det øge et materiales anvendelighed, hvis tolkninger, der går videre end almindelig transskription, er holdt adskilt fra sådanne fortolkninger, som kun består i transskription. Det kan naturligvis diskuteres, hvor stor værdien af bogstavret transskription er i forhold til gengivelse med normaliseret stavemåde. Det er en diskussion, som ikke er begrænset til processer, hvori maskinlæsbargørelse indgår. Når man har at gøre med databaseagtige datamaterialer, kan hensynet til adskillelse af transskription og videre tolkning gennemføres ved at have flere felter til den samme informationsenhed, et felt med informationen transskriberet og andre felter med videre grader af fortolkning. Det er indlysende, at klar adskillelse af forskellige fortolkningsniveauer har værdi, hvis datamaterialet skal anvendes igen af en anden end den, der har produceret det. Mindre indlysende er det måske, at det også for den, der producerer materialet, er værdifuldt at have muligheden for at skelne mellem forskellige fortolkningsniveauer. Det, man i den forbindelse skal erindre sig, er, at arbejdet med at producere et datamateriale tager tid. Mens arbejdet står på, erhverver man sig viden om kilden. Hvis den viden, man har ved slutningen af arbejdet, skal komme fortolkningen af de først behandlede enkeltoplysninger til gode, er det af stor værdi, at fortolkningsniveauerne kan adskilles. En anden faktor er, at mennesket i udøvelsen af skøn inddrager faktorer uden for den foreliggende problemstilling. Derfor kan det valg, som samme person foretager i samme skønssituation, være for-

PRISLOEN.PDA

År 1642 Identifikation D42R48 Nummer 66145

Normaliserede størrelser:

Betegnelse Bly Enhed skppd

Antal enheder 4,750 Total pris 56,401 rdlr.

Enhedspris 12,000 rdlr.

Oprindelige størrelser:

Betegnelse Bly

Antal enheder 4.5 skipped 4 lispd

Enhedspris pr skippd 12 rdl

Total pris 56 rdl 1.5 ort 2.5 sk

Øvrige oplysninger:

Køber DBL Hjemsted DBL

Sælger Mads Hansen Hjemsted Randers

Kilde DBL

Kommentar nr 15.

Fig. 8. Indlæsningsskærmbillede i DataPerfect. Inputfelter er angivet ved understregning.

skelligt på forskellige tidspunkter. Ved adskillelse af forskellige fortolkningsniveauer kan man opnå, at ensartede skøn ligger så tæt i tid som muligt og derfor har en rimelig chance for at møde primærundersøgeren i det samme temperament. Her kunne naturligvis også argumenteres, at det ved en statistisk orienteret analyse kan være en fordel, hvis samme fortolkningsniveau når primærundersøgeren med stor tidsafstand, således at primærundersøgerens humørsvingninger kan udbalanceres statistisk.

Tillæg: For de teknisk interesserede

Oprindeligt indlæsningsformat

Den oprindelige indlæsning skete til en listeform, hvor en typisk oplysning for priser kunne være:

3
D11R11
Rug
pr tønde 5 rigsmark
28 tønder
35 daler
Jørgenn Schriffuer
DBL
Randers
DBL

DBL

De enkelte felter i posten er adskilt med linieskift. Felterne er:

Postnummer på disketten.
Identifikation: D for Dronningborg len (K for Kalø len), 11 som de to sidste cifre af årstal, R for regnskab (B for bilag), 11 for filmbillede inden for den pågældende enhed.
Oprindelig betegnelse
Opgivet enhedspris
Oprindelig mængde
Oprindelig samlet pris
Køber
Sælger (Dronningborg len forkortet til DBL, Kalø len til KL)
Købers hjemsted
Sælgers hjemsted
Kommentar
Kilde

Datarensning i DataPerfect

Data var placeret i DataPerfect som en almindelig flad fil, der blev tilgået i et skærmbillede som Figur 8.

Oplysningerne fra den oprindelige datafil blev for de flestes vedkommende placeret under linien *Oprindelige størrelser:*. Identifikation blev dog placeret i øverste linie. Løbenummeret fra disketten er indeholdt i *Nummer*. *År* er for de fleste enkeltoplysningers vedkommende automatisk beregnet under indlæsningen ud fra *Identifikation*. 1640 i *År* henviser til regnskabsåret 1. maj 1640 til 30. april 1641.

Indlæsning i SAS

Ved indlæsning i SAS var rækkefølgen af felterne:

År	Det beregnede handelsår
Ident	Identifikation
Nummer	Enkeltoplysningsnum
Norm–bet	Normaliseret betegnelse
Enhed	Mængdeenhed for de normaliserede størrelser
Ant–enh	Manuelt beregnet total mængde
Tot–skil	Manuelt beregnet totalpris i skilling
Norm–e–pr	Manuelt beregnet enhedspris i skilling pr. enhed
Opr–bet	Betegnelse som indlæst
Opr–mngd	Mængde som indlæst
Opr–e–pr	Enhedspris som indlæst
Opr–t–pr	Totalpris som indlæst
Koeber	Køber
Koeb–hjm	Købers hjemsted
Saelger	Sælger
Sael–hjm	Sælgers hjemsted
Kilde	Kildeangivelse
Komment	Kommentar
$$$	Skilletegn til brug for SAS indlæsningsprogrammet

Nedenfor ses to af enkeltoplysningerne i indlæsningsformatet:

År	1630	1630
Ident	D30R11	K30R87
Nummer	57080	2237
Norm–bet	Rug	Rug
Enhed	tdr	tdr
Ant–enh	0.000	0.000
Tot–skil	0.000	0.000
Norm–e–pr	0.000	0.000
Opr–bet	Rug	Rug og mel
Opr–mngd	110 tønder	15.5 læst 8.5 td 1.5 skæp 1 fk
Opr–e–pr	pr tønde 2 rdl	tønden 3 rdl
Opr–t–pr		1421 rdl 15 sk
Koeber	Borgerskabet	Bønderne
Koeb–hjm	Randers	KL
Saelger	DBL	KL
Sael–hjm	DBL	KL
Kilde	DBL	
Komment		p. opboren aff bønderne
$$$	$$$	$$$

De to enkeltoplysninger er henholdsvis fra Dronningborg og Kalø lensregnskaber. Den manuelle normalisering er ikke gennemført for nogen af dem.

Datarensning i SAS

Som eksempel på proceduren vil jeg beskrive normaliseringen af totalprisen. Det vil fremgå for den programmeringskyndige, at SAS er et temmelig kluntet programmeringssprog. Når det alligevel kan være rimeligt at anvende SAS frem for at programmere alting op fra grunden i et højniveausprog, skyldes det, at SAS grundlæggende har styr på sine dataformater. Denne datasikkerhed ville man være nødt til selv at skabe, hvis man havde at gøre med et smartere programmeringssprog. Til syvende og sidst gælder det dog, at man kan gøre alt i alting. Valget af redskab indebærer derfor et betydeligt subjektivt element.

Normaliseringen af totalprisen er opdelt over to SAS-programmer, eller i SAS-terminologi, to datatrin. Det første program starter med en almindelig præambel. Derefter er der nogle almindelige omkodninger af konstaterede fejl i materialet:

```
libname u1066 »c:\data\u1066« ;

data u1066.pris2 ;
 set u1066.priser ;
  array otp $ otp01-otp10 ;

  format otpkom $40. ;
  label otpkom = »Bemærkninger til total-
                  pris« ;

  slut = 10 ;
  /* generelle omkodninger */
  if opr–t–pr = »(8.5 mark)« then opr–t–pr =
                »8.5 mark« ;
  if opr–t–pr = »dbl« then opr–t–pr = »« ;
  if opr–t–pr = »4 tønder« then do ;
    opr–maen = opr–t–pr ;
    opr–t–pr = koeber ;
    koeber = »« ;
  end ;
```

Ved hjælp af ordren »libname« og i nogle tilfælde også med »filename« sker sammenkoblingen mellem SAS-sproget og den måde, hvorpå det aktuelle styresystem behandler filer og kataloger. Resten af programmet er uafhængigt af det styresystem, hvorpå man afvikler det. I SAS skelnes mellem datasæt, som er SAS-systemfiler, og filer, som er alt muligt andet, fortrinsvis tekstfiler. Datatrinnet i SAS indledes med ordren »data«. Efter »data« angiver man navnet på det datamateriale, man vil have udskrevet, outputdatasættet. Derefter angiver man med ordren »set« det datamateriale, man vil have læst, inputdatasættet. Nedenunder kommer så alle de ordrer, man vil have udført. Datatrinnet i SAS virker almindeligvis på den måde, at der læses en enkeltoplysning fra inputdatasættet, derpå udføres alle ordrerne i datatrinnet på denne enkeltoplysning, og endelig udskrives enkeltoplysningen i den skikkelse, den derved har opnået, til outputdatasættet. Derefter læses en ny enkeltoplysning, og det samme gentager sig, indtil der ikke er mere at læse. Der er mange muligheder for variere dette standardforløb.

Nu er det så tiden at opdele den oprindelige totalpris. Dette er godt understøttet af SAS med en funktion ved navn scan:

```
/* jeg har checket at ord 11 aldrig findes */
do i = 1 to 10 ;
 otp{i} = scan(opr–t–pr,i,' ') ;
 end ;
```

Så kommer en lang række individuelle omkodninger af enkeltoplysninger. Data er nemlig netop så specielle, at en given metode altid rammer skævt for nogle egenskaber ved materialet:

```
select(nummer) ;
 when(2263) do ;
  otp07 = »« ;
  otp01 = »10« ;
  otpkom = trim(otpkom)||»Rdlr rettet til
           10 « ;
  end ;
 when(3103) do ;
  otpkom = »Kan det mon passe« ;
  do i = 3 to 7 ;
   otp{i} = »« ;
   end ;
  end ;
```

En vigtig logisk konstruktion i SAS er »select«. Med denne ordre styrer man programforløbet i overensstemmelse med værdien af den variabel, hvis navn er angivet i parentes efter »select«. Når variablen har en værdi, der modsvares af den værdi, der findes efter en 'when'-sætning, udføres de ordrer, der står efter det pågældende »when«. Hvis variablen har en værdi, hvortil der ikke svarer noget »when«, udføres det, der står efter »otherwise«.

Det fortsætter på denne måde et par sider, så det er der ingen grund til at besvære læseren med.

```
 when(65031) do ;
  otp07 = »0.5« ;
  otp08 = »alb« ;
  end ;
 otherwise ;
 end ;
```

Nu normaliseres ordene ét for ét:

```
 /* ret fejlkoder */
 if otp01 = »8o« then otp01 = »80« ;
 if otp01 = »9o« then otp01 = »90« ;
 if otp01 = »55..5« then otp01 = »55.5« ;
 if otp01 = »97.5.5« then otp01 = »97.5« ;
 /* beregn værdier */
 sted = 1 ;
 link division ;
 rykv = 1 ;
 do while(rykv = 1) ;
 rykv = 0 ;
 sted = 2 ;
 if otp02 = »5« then do ;
  otp01 = »49.5« ;
  sted = 2 ;
  link rykven ;
  end ;
 link divadd ;
 link moentkod ;
 end ;
```

Det ses, at nogle af indtastningerne er foretaget af urutineret arbejdskraft. I adskillige tilfælde er bogstaverne *o* og *l* brugt for tallene *0* og *1*.

Ved hjælp af en loop-variabel, rykv, styres, at processen gennemkøres en gang til, hvis der er mulighed for, at der er rykket noget nyt hen på den plads, der undersøges. Hvis man f.eks. har udtrykket »1 1/2 slet daler« bliver det først omformet til »1.5 slet daler« og dernæst til »1.5 sldlr« . Endelig vil loopet køre en gang til for at konstatere, at »sldlr« tilhører det kontrollerede ordforråd.

Subrutiner udtrykkes i SAS med link. Subrutinen »division« laver ord, der indeholder / (brøkstreg), om til decimalbrøker:

```
division:
 if index(otp{sted},'/') > 0 then do ;
  dividend = scan(otp{sted},1,'/') ;
  divisor = scan(otp{sted},2,'/') ;
  otp{sted} = dividend / divisor ;
  divf = 1 ;
```

```
  end ;
 return ;
```

Subrutinen »divadd« gør brug af »division«, men lægger resultatet af divisionen til det tal, der står foran, og rykker rækken af ord en plads til venstre. Subrutinen »moentkod« er den, der indfører det kontrollerede ordforråd for møntbetegnelser:

```
moentkod:
 if otp{sted}='0rt' then otp{sted} = »ort« ;
 if otp{sted} = »a« or otp{sted} = »al« then
                  otp{sted} = »alb« ;
 if otp{sted}='cou.' then do ;
  otp{sted} = »cdlr« ;
  link rykven1 ;
  end ;
 if otp{sted}='courantd' then otp{sted} =
               »cdlr« ;
 if otp{sted}='curantdl' then otp{sted} =
               »cdlr« ;
 if otp{sted}='d.slet' then otp{sted} = »sldlr« ;
 if otp{sted}='dal' then otp{sted} = »dlr« ;
 if otp{sted}='dale' then otp{sted} = »dlr« ;
 if otp{sted}='daler' then otp{sted} = »dlr« ;
```

Det fortsætter nu med mange flere af denne type trivielle omkodninger. De sidste af disse omkodninger ser på næste ord i rækken, da det ofte er kvalificerende for det foregående: »daler slet«, »daler in specie«, »skilling lybsk«, etc.:

```
if index(upcase(otp{sted+1}),'SLE') > 0 or
 index(otp{sted+1},'sld') > 0 or otp{sted+1}=
       »s.« or
 otp{sted+1} = »sl.« or otp{sted+1} = »s.mø«
 then do ;
  do i = 2 to sted by 2 ;
   if (otp{i} = »dlr«) or (otp{i} = »mk«) then
                do ;
    otp{i} = »sl«||otp{i} ;
    end ;
   end ;
   link rykven1 ;
   if otp{sted+1} = »mønt« or otp{sted+1} =
                   »?)« then do ;
    link rykven1 ;
    end ;
   end ;
  end ; /* på < sted – 2 */
 return ;
```

Når de ti ord er gennemløbet, skrives et datamateriale, som nu i OTP01-OTP10 indeholder normaliserede ord på de lige pladser og tal på de ulige pladser. Det næste program indeholder så beregningen af mønterne. Denne omregning har jeg valgt at udtrykke i et langt loop:

```
do i = 1 to 5 ;
 if otp{2*i} ne »« then do ;
  select(otp{2*i}) ;
   when(»alb«) do ;
    notp=notp+otp{2*i-1}/3 ;
    end ;
   when(»cdlr«) do ;
    notp=notp+80*otp{2*i-1} ;
    end ;
   when(»cmk«) do ;
    notp=notp+20*otp{2*i-1} ;
    end ;
   when(»dlr«) do;
    if aar < 1546 then
     put »Daler i « aar » i nummer « nummer
          opr–t–pr ;
    if aar > 1545 and aar < 1567 then
     notp=notp+48*otp{2*i-1} ;
    if aar > 1566 then
     notp=notp+64*otp{2*i-1} ;
    end ;
```

Herefter fortsættes med de øvrige møntenheder, der findes i materialet.

```
 when(»rdlr«) do;
   if aar < 1602 then
    put »Rigsdaler i « aar » i nummer « num-
         mer opr–t–pr ;
   if aar > 1602 and aar < 1609 then
    notp=notp+66*otp{2*i-1} ;
   if aar = 1609 then
    notp=notp+68*otp{2*i-1} ;
   if aar > 1609 and aar < 1616 then
    notp=notp+74*otp{2*i-1} ;
   if aar > 1615 and aar < 1618 then
    notp=notp+80*otp{2*i-1} ;
   if aar > 1617 and aar < 1620 then
    notp=notp+84*otp{2*i-1} ;
   if ((aar > 1619 and aar < 1624)
    or (aar > 1624 and aar < 1875)) then
    notp=notp+96*otp{2*i-1} ;
   if aar = 1624 then
    notp=notp+100*otp{2*i-1} ;
   end ;
```

Til slut er der en kontrolsætning, som ikke bør give noget output. Så længe der kommer noget ud af denne put-sætning, mangler der noget af normaliseringen.

```
    otherwise do ;
     put »ukendt enhed « otp{2*i} » i år « aar » i
          nummer « nummer opr–t–pr ;
    end ;
   end ; /* på select */
  end ; /* på if */
 end ; /* på do i */
```

Noter

De undersøgelser, der er redegjort for i denne artikel, er udført med bistand fra Statens Humanistiske Forskningsråd og fra Rigsarkivet, som har ordnet og nyfotograferet lensregnskabsserierne for Kalø og Dronningborg len, umiddelbart før jeg skulle bruge dem.

1. Se herom Herbert Reinke, Kevin Schürer og Hans Jørgen Marker: Information Requirements and Data Description in Historical Social Research. A Proposal, *Historical Social Research* 43/43, Köln 1987.
2. Rigsarkivet (RA), Regnskaber 1559–1660, Lensregnskaber, Kalø 1629/30, udgiftsbilag s. 163.
3. Manfred Thaller, udg.: *Images and Manuscripts in Historical Computing,* St. Katharinen 1992, særlig Pedro Gonzales: The Digital Processing of Images in Archives and Libraries, s. 97–122.
4. Se f.eks. C.M. Sperberg-McQueen & Lou Burnard, udg.: *Guidelines for the Encoding and Exchange of Machine-Readable Texts*, Chicago/Oxford 1990.
5. Et særlig berømt projekt af denne type var Alan Macfarlanes arbejde med Earls Colne. Se Alan Macfarlane: *Reconstructing Historical Communities*, Cambridge 1977. Datamaterialet fra dette arbejde er vist gået tabt. Manfred Thaller: The Need for Standards, *Modelling Historical Data*, udg. Daniel I. Greenstein, St. Katharinen 1991, s. 8.
6. E. Ladewig Petersen: *Fra standssamfund til rangssamfund 1500–1700*, København 1980, (*Dansk social historie* 3), s. 248–249.
7. Udtrykket stammer fra M.I. Finley: *The Ancient Economy*, London 1973, s. 25.
8. Hans Jørgen Marker: Sletdalerbegrebet i første fjerdedel af 17. århundrede, *Historie, Jyske Samlinger*, Ny række 15, 1985, s. 633–640.
9. Hans Jørgen Marker: Danish Prices in the 1640'ies, *L'ordinateur et le métier d'historien*, udg. Bernard Lavalle, Bordeaux 1990, s. 35–44.
10. Hans Jørgen Marker: *En modelanalyse til belysning af den indflydelse som udviklingen i priser, lønninger og skatter udøvede på pengeafkastet af en østjysk højadelig godsbesiddelse i perioden fra kornprisfaldet 1618–20 frem til årene forud for Karl-Gustav-krigene 1657–60*, upubliceret afhandling, Historisk Institut, Aarhus Universitet 1982.
11. Troels Dahlerup: Om tienden, *Fortid og Nutid* 29, 1981, s. 8.