SYMPOSIUM I ANVENDT STATISTIK

2017

Syddansk Universitet
Institut for Virksomhedsledelse og Økonomi

Danmarks Statistik
DST Survey

SYMPOSIUM I ANVENDT STATISTIK

23.-24. januar 2017

Redigeret af Peter Linde
på vegne af organisationskomiteen

Støttet af SAS Institute Inc.

SDU, Institut for Virksomhedsledelse og Økonomi Danmarks Statistik

Forord

Det er symposiets formål at fremme information om såvel anvendt statistik som statistisk databehandling. Symposiet er tværfagligt med særlig vægt på metodik, formidling og fortolkning af statistiske analyser. I år er Institut for Virksomhedsledelse og Økonomi, Syddansk Universitet vært for symposiet, hvilket vi gerne vil takke for. Symposiet arrangeres af Danmarks Statistik og Institut for Virksomhedsledelse og Økonomi, SDU og den faglige forening Symposium i Anvendt Statistik er ansvarlig for det faglige program og økonomien.

Denne publikation indeholder foredragene fra det 39. Symposium i Anvendt Statistik. Dette års indlæg kommer fra mange forskellige fagområder og lægger vægt på forskellig metoder og problemstillinger. Som det er normalt ved videnskabelige indlæg, er bidragsyderne ansvarlige for indholdet af indlæggene, og spørgsmål herom kan rettes direkte til forfatterne.

Med symposiet tilstræbes det at skabe et forum for tværfaglig inspiration og kritik blandt andet for at udbygge kommunikationen mellem personer, der arbejder med beslægtede metoder inden for forskellige fagområder.

Peter Linde, Organisationskomiteen

ISBN 978-87-501-2268-5

Trykt hos Litotryk i 125 eksemplarer

Organisationskomiteen for Symposium i Anvendt Statistik 2017

Lisbeth la Cour
Økonomisk Institut
Copenhagen Business School
Porcelænshaven 16A
2000 Frederiksberg
llc.eco@cbs.dk

Peter Linde
DST Survey
Danmarks Statistik
Sejrøgade 11
2100 København Ø
pli@dst.dk

Anders Milhøj
Økonomisk Institut
Københavns Universitet
Studiestræde 6
1455 København K
Anders.Milhoj@econ.ku.dk

Esben Høg
Matematiske Fag
Aalborg Universitet
Fredrik Bajers Vej 7
9220 Aalborg Ø
esben@math.aau.dk

Gorm Gabrielsen
Institut for Finansiering
Copenhagen Business School
Solbjerg Plads 3
2000 Frederiksberg
stgg@cbs.dk

Anders Holm
Sociologisk Institut
København Universitet
Øster Farimagsgade 5
1014 København K
ah@soc.ku.dk

Helle M. Sommer
SEGES
Landbrug & Fødevarer
Axeltorv
1609 København V
hmso@seges.dk

Niels Kærgaard
Fødevare- og Ressourceøkonomi
Københavns Universitet
Rolighedsvej 25
1958 Frederiksberg
nik@life.ku.dk

Mogens Dilling-Hansen
Institut for Økonomi
Århus Universitet
8000 Århus C
dilling@econ.au.dk

Klaus Rostgaard
Statens Serum Institut
Artillerivej 5
2100 København Ø
KLP@ssi.dk

Jørgen Lauridsen
Økonomisk Institut
Syddansk Universitet
Campusvej 55
5230 Odense M
jtl@sam.sdu.dk

Kaare Brandt Petersen
SAS Institute
Købmagergade 7-9
1050 København K
Kaare.Brandt@sdk.sas.com

Birthe Lykke Thomsen
Det Nationale Forskningscenter for Arbejdsmiljø
Lersø Parkallé 105
2100 København Ø
blt@arbejdsmiljoforskning.dk

Indholdsfortegnelse

Økonomi og Samfund

Statistisk Metode

Familial Risk and Heritability of Cancer Among Twins. The Nordic Twin Cancer Study

Jacob Hjelmborg; Nordic Twin Study of Cancer (NorTwinCan) Collaboration, Epidemiology, Biostatistics and Biodemography, Institute of Public Health, University of Southern Denmark;

Our aim is to estimate familial cancer risk essential to cancer risk etiology and prediction. The prospective study of 80 309 monozygotic and 123 382 same-sex dizygotic twin individuals (N = 203 691) within the population-based registers of Denmark, Finland, Norway, and Sweden were followed up a median of 32 years between 1943 and 2010. The main outcome was incident cancer for 41 main (ICD-10) cancer sites. Time-to-event analyses were used to estimate familial risk (risk of cancer in an individual given a twin's development of cancer) and heritability (proportion of variance in cancer risk due to inter-individual genetic differences) with follow-up via cancer registries. Statistical models adjusted for age and follow-up time, and accounted for censoring and competing risk of death.

In this long-term follow-up study among Nordic twins, there was significant excess familial risk for cancer overall and for specific types of cancer, including prostate, melanoma, breast, ovary, and uterus. This information about hereditary risks of cancers may be helpful in patient education and cancer risk counseling (1. JAMA 2016). Concordance for different cancers was highly common suggesting further investigation.

Further, the combined cohort has a great potential for untangling complicated exposures for cancer: As a particular instance we show that tobacco exposure causes lung cancer even when adjusting for genetic factors. Interactions between genes and environmental exposure in the development of lung cancer are not supported from the largest twin cohort study with longest follow-up ever. Familial effects have decreased influence with increasing age (2. Thorax 2017).

BIAS I OVERLEVELSESANALYSE MED UOBSERVEREDE CENSUREREDE OBSERVATIONER

SÖREN MÖLLER

OPEN - Odense Patient data Explorative Network, Odense Universitetshospital og Klinisk Institut, Syddansk Universitet, Odense

BAGGRUND

Overlevelsesanalyse tager normalt udgangspunkt i, at stikprøven selekteres fra en kohorte, og at individer indgår i stikprøven uafhængig af, om de oplever hændelsen eller ej.

I nogle situationer er det dog lettere (eller sværere) at inkludere individer, der har oplevet hændelsen end individer, der ikke har oplevet hændelsen, for eksempel fordi den præcise afgrænsning af kohorten i risiko, som individerne udtrækkes fra, er uklar.

Fraværet af disse observationer betegnes i litteraturen som **højretrunkering** [1, 2]. En typisk situation, hvor problemet opstår i litteraturen, er studier af AIDS, hvor personer kommer i risiko, når de bliver HIV-inficeret, men først bliver inkluderet i studiet når de får konstateret AIDS, som er hændelsen, der undersøges, og hvor de så retrospektivt får bestemt deres risikotid [3, 4].

Typisk diskuteres situationer, hvor alle individer, der ikke har oplevet en hændelse er trunkeret. I denne fremstilling er vi dog mest interesseret i den situation, hvor vi har observeret en brøkdel af de censurerede individer, men hvor vi ikke ved hvor stor en del af populationen i risiko det drejer sig om. Vi er motiveret af følgende to eksempler:

Motiverende eksempel A [5]. Stofmisbrugere der er døde af en overdosis undersøges med henblik på risikoforskel mellem genetiske strata. Næsten alle døde stofmisbrugere i en periode er inkluderet i studiet, da de er blevet undersøgt på et retsmedicinsk institut, derimod er populationen af levende stofmisbrugere ukendt, og det er kun en brøkdel af disse, der er medtaget i stikprøven. Dette bevirker at en ukendt, men ret stor, andel af de narkomaner der stadig er i live er højretrunkerede i studiet.

Motiverende eksempel B. I en kohorte af patienter med en bestemt diagnose, vil vi undersøge risikoen for at blive diagnosticeret med en specifik følgesygdom i forhold til en eksponering. Antager vi at den oprindelige diagnose kan være underdiagnosticeret i populationen, men vil blive diagnosticeret senest samtidigt med følgesygdommens optræden, så vil individer med følgesygdommen være overrepræsenteret i stikprøven i forhold til populationen. Dette vil svare til delvis højretrunkering og vil udløse et bias, hvis der ikke tages højde for denne biaskilde.

Problemstilling. Det spørgsmål vi ønsker at undersøge er, hvor stor effekten af denne delvise højretrunkering er på estimerede hazardrater og hazardratioer, for at kunne vurdere (i det mindste groft), hvor stor en effekt denne biaskilde har på resultater af studier, hvor der forekommer højretrunkering af en del af de individer der ikke har oplevet hændelser.

Date: 16. december 2016.

FORHOLD TIL VENSTRETRUNKERING OG CENSURERING

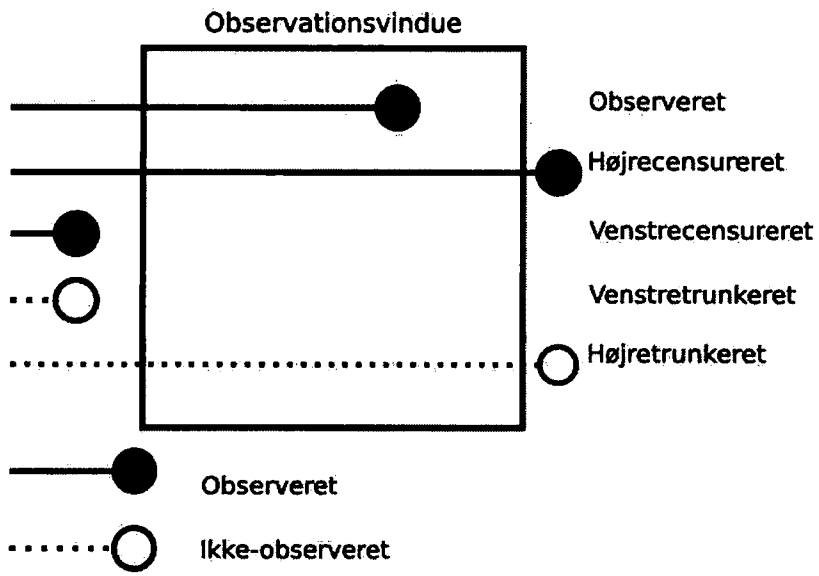
De typiske fejlkilder (se figur 1), der skal tages højde for i overlevelsesanalyse er

- **Højrecensurering:** Individer, der har en hændelse efter observationsperioden, bliver kun observeret som hændelsesfri ved slutningen af observationsperioden.
- **Venstrecensurering:** Individer, der havde en hændelse før observationsperioden, bliver kun observeret som havende haft en hændelse før, men ikke hvornår de havde hændelsen.
- **Venstretrunkering:** Individer, der havde en hændelse før observationsperioden, bliver ikke observeret.

Derimod kigger vi her på fejlkilden, der opstår ved

- **(Delvis) højretrunkering:** (En brøkdel af) de individer, der ikke har en hændelse inden slutningen af observationsperioden, observeres ikke.

I denne fremstilling vil der også være højrecensurerede observationer, men vi antager, at der hverken sker venstrecensurering eller venstretrunkering og vi ser desuden bort fra muligheden for konkurrerende risici.



FIGUR 1. Observationstyper i overlevelsesanalyse

Der findes i litteraturen metoder til at modellere højretrunkerede data, dog primært med fokus på estimering af selve overlevelsesfunktionerne og motiveret af dobbelttrunkerede observationer i astronomisk forskning [6,7], men også med test for forskel i overlevelsesfunktioner motiveret af højretrunkerede AIDS-data [3]. Disse metoder er også blevet implementeret i R [8]. Vi er derimod hovedsageligt interesseret i risikosammenligning mellem forskellige eksponeringsgrupper.

DATAGENERERENDE PROCES

Vi antager en datagenererende proces bestående af

- $X \in \{0,1\}$ Binær kovariat, som indikerer de to grupper vi vil sammenligne hændelsesrisikoen imellem.
- T_{event} Tid for hændelse (død) med fordeling afhængig af X.
- T_{cens} Censureringstid uafhængigt af T_{event} og X.
- $T_{obs} = \min(T_{event}, T_{cens})$
- $I_{event} \in \{0,1\}$ Censureringsindikator $I_{event} = 1$ hvis $T_{event} \leq T_{cens}$ ellers 0.

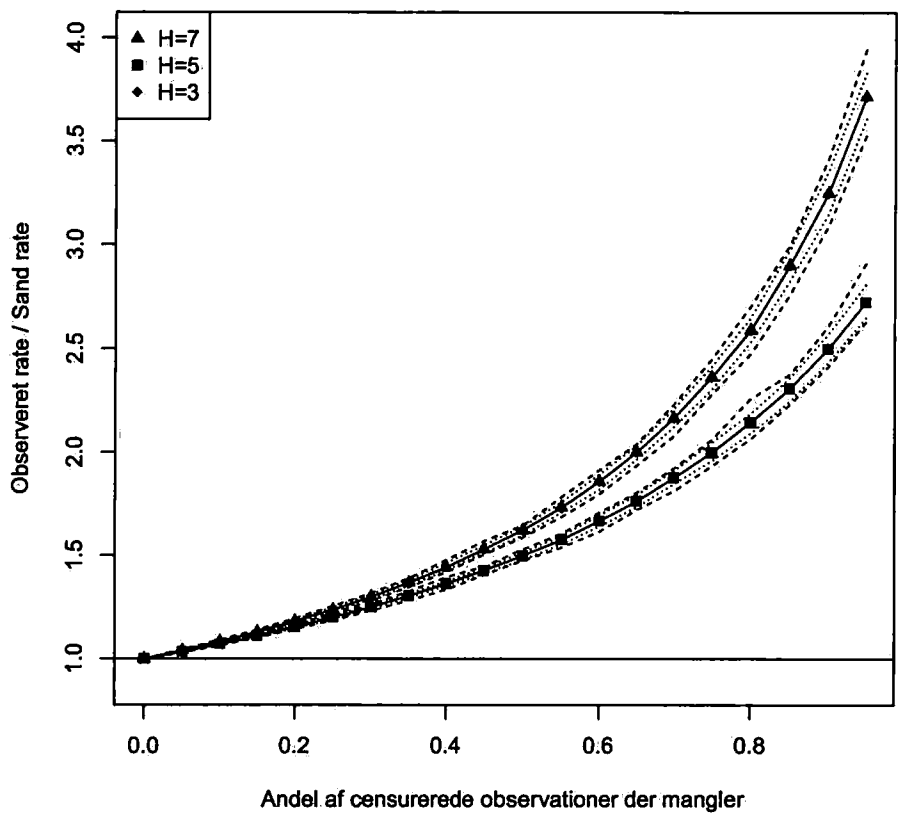
- $I_{obs} \in \{0, 1\}$ Højretrunkeringsindikator $I_{obs} = 1$ for observerede observationer (herunder alle med $I_{event} = 1$) og 0 for højretrunkerede observationer. Uafhængigt af X, T_{event} og T_{cens} betinget på $I_{event} = 0$.
- $P(I_{obs}|I_{event} = 0) = p_{obs}$ Højretrunkeringssandsynlighed

$p_{obs} = 1$ svarer til almindelig overlevelsesanalyse uden højretrunkering, mens $p_{obs} = 0$ svarer til den typiske situation i litteraturen, hvor alle observationer uden hændelse er trunkerede.

Vi er også interesseret i at undersøge en simplere situation med kun én gruppe, svarende til ovenstående med $X = 0$ for alle individer.

SIMULATIONER

I det første simulationseksperiment har vi simuleret eksponentialfordelte hændelsestider med raterne $H = 3, 5, 7$ og exponentialfordelte censureringstider med raten 10. For hver rate og for p_{obs} mellem 0.05 og 1 (i skridt af 0.05) simulerede vi 100 datasæt med 10.000 observationer. I hvert datasæt estimerede vi eksponentialfordelingens rate både med alle observationer taget med (sand rate) og med $1 - p_{obs}$ af de censurerede observationer tilfældig trunkeret (observeret rate). Se figur 2.

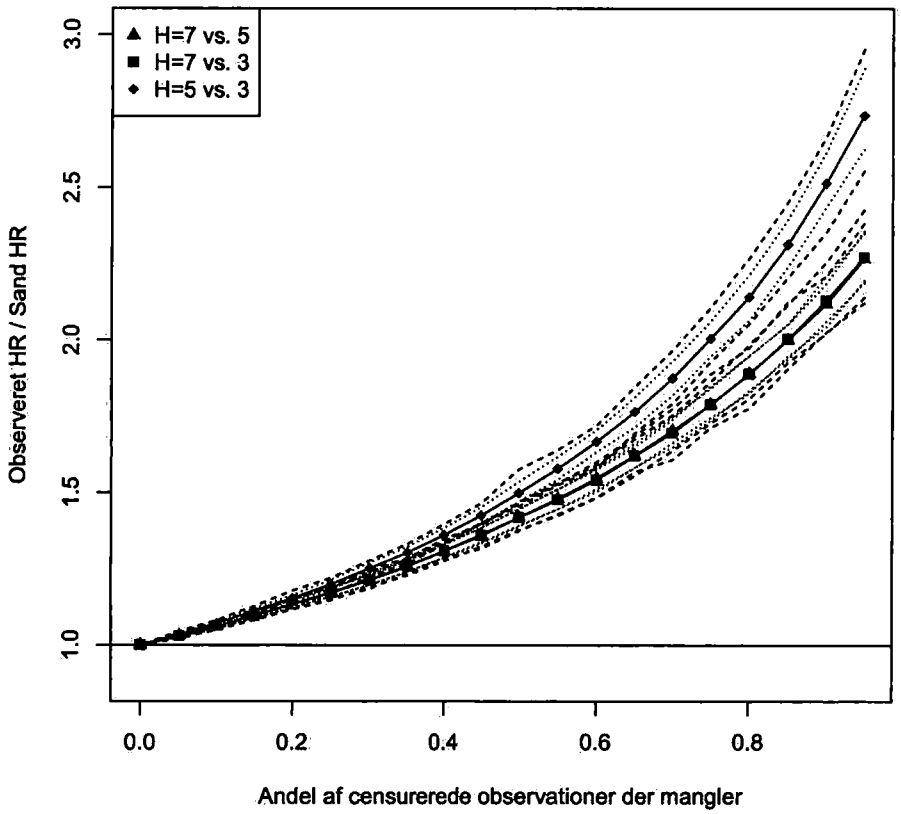


FIGUR 2. Simulerede eksponentialfordelte data: Hazardrate ift. $1 - p_{obs}$

Grafen viser de resulterende forhold mellem observeret og sand rate for forskellige p_{obs} og forskellige sande rater, med maksimum/minimum og empiriske 95%-intervaller markeret. Vi ser som forventet, at den observerede rate stiger, jo flere censurerede observationer mangler. Desuden er denne effekt

stærkere, jo lavere den sande rate er, hvilket er i god overensstemmelse med, at en lavere rate indebærer flere censurerede observationer og derved et større informationstab ved trunkeringen.

Simulerer vi data for to grupper med forskellige hændelsesrater, men med samme censureringsrate og undersøger forholdet mellem den sande hazardratio og den observerede hazardratio får vi resultaterne i figur 3.



FIGUR 3. Simulerede eksponentialfordelte data: HR ift. $1 - p_{obs}$

Grafen viser at jo flere censurerede observationer, der mangler, jo større bliver afvigelsen mellem den sande hazardratio og den observerede hazardratio, og denne effekt er stærkere for observationer med lave hazardrater.

RESULTATER FRA MOTIVERENDE EKSEMPEL A

I [5][1] blev der inkluderet en stikprøve bestående af 274 afdøde narkomaner, der blev undersøgt på retsmedicinske institutter i Danmark i perioden 2006 til 2012. Som sammenligning blev der 2010 til 2012 taget blodprøver af 309 levende narkomaner rekrutteret fra et misbrugsbehandlingssted. Som mulig analysetilgang (som ikke endte med at blive rapporteret i [5]) overvejede vi at undersøge genotypers association med dødelighed ved overlevelsesanalyse med

- død som hændelse

[1]Jeg takker Dorte J. Christoffersen fra Retsmedicinsk Institut på Syddansk Universitet for at måtte bruge hendes data til at illustrere problemstillingen.

- alder ved død som hændelsestid
- censurering af levende narkomaner ved prøvetagning (Da vi ikke ved hvor længe de er i live efterfølgende.)
- de forskellige genotyper som eksponering

Højretrunkering optræder oplagt i situationen i [5], da de levende narkomaner inkluderet kun er en brøkdel af den (ukendt store) kohorte af narkomaner i Danmark. Derimod er de døde narkomaner inkluderet en langt større del af kohorten, da alle undersøgt på retsmedicinske institutter i en periode blev inkluderet. For at undersøge effekten af højretrunkering undersøgte vi hvordan HR for genotyper beregnet ved Cox-regression ville ændre sig i sammenligning med de reelle data, når vi enten duplicerede de censurerede observationer 2 henholdsvis 10 gange eller samplede 2 henholdsvis 10 gange antallet af reelle censurerede observationer fra de censurerede observationer. I tilfældet af samplingen tog vi middelværdien mellem 1000 kørsler.

Situation	$n_{\text{hændelse}}$	$n_{\text{censureret}}$	HR_A	HR_B	HR_C
Reelle data	274	309	0.8474	1.0885	0.7658
2*Duplikeret	274	618	0.8555	1.1239	0.7262
10*Duplikeret	274	3090	0.8673	1.1897	0.6683
2*Samplet†	274	618	0.8611	1.1269	0.7308
10*Samplet†	274	3090	0.8696	1.1900	0.6696

† middelværdi af 1000 simulationer

Resultaterne fra dette eksempel giver et mindre klart billede end de simulerede data, eftersom effekten for HR_A overvurderes i det reelle (trunkerede) datasæt i forhold til de simulerede (potentielt ikke-trunkerede) større datasæt, mens der for HR_B og HR_C observeres et konservativt bias, hvor det mindre trunkerede datasæt giver en HR tættere på 1 end de større datasæt.

Denne forskel kan potentielt skyldes den mere realistiske ikke-konstante baselinehazard i disse reelle data i forhold til de simulerede eksponentialfordelte data, men den kunne også skyldes effekten af, at de tilføjede censurerede observationer er dupliceret/samplet fra de eksisterende censurerede observationer med mindre variabilitet til følge.

Konklusion

Sammenfattende kan vi konkludere at delvis højretrunkering som ventet resulterer i overvurdering af hændelsesrisikoen. Det viser sig også at trunkeringen i de simulerede data resulterer i en overvurdering af risikoforskelle mellem grupper, men mere uklart i de reelle data. Denne effekt vokser jo større en andel af de censurerede observationer, der er trunkerede. Konklusionen må være at det er vigtigt at være opmærksom på denne fejlkilde, da den kan medføre falsk-positive associationer.

Referencer

[1] B. W. Turnbull, *The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data*, Journal of the Royal Statistical Society. Series B (Methodological) **38** (1976), no. 3, 290–295.

[2] J. P. Klein and M. L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd ed., Statistics for Biology and Health, vol. 99, Springer-Verlag New York, 2003.

[3] S. W. Lagakos, L. M. Barraj, and V. de Gruttola, *Nonparametric analysis of truncated survival data, with application to AIDS*, Biometrika **75** (1988), no. 3, 515–523.

[4] J. D. Kalbfleisch and J. F. Lawless, *Regression models for right truncated data with applications to AIDS incubation times and reporting lags*, Statistica Sinica **1** (1991), 19–32.

[5] D. J. Christoffersen, P. Damkier, S. Feddersen, S. Möller, J. L. Thomsen, C. Brasch-Andersen, and K. Brøsen, *The ABCB1, rs9282564, AG and TT Genotypes and the COMT, rs4680, AA Genotype are Less Frequent in Deceased Patients with Opioid Addiction than in Living Patients with Opioid Addiction*, Basic & Clinical Pharmacology & Toxicology **119** (2016), 381–388.

[6] B. Efron and V. Petrosian, *Nonparametric analysis of doubly truncated data*, Journal of the American Statistical Association **94** (1999), no. 447, 824–834.

[7] P. Shen, *Nonparametric Methods for Doubly Truncated Data*, Annals of the Institute of Statistical Mathematics **62** (2010), 835–853.

[8] C. Moreira, J. de Una-Alvarez, and R. M. Crujeiras, *DTDA: An R Package to Analyze Randomly Truncated Data*, Journal of Statistical Software **37** (2010), no. 7.

Danskernes alkohol forbrug

Anders Milhøj
Økonomisk Institut,
Københavns Universitet
anders.milhøj@econ.ku.dk

Resume: I artiklen gennemgås de nyeste tendenser i danskernes alkoholforbrug ud fra de offentligt tilgængelige datakilder - mest oplysninger om Statens indtægter fra alkoholbeskatningen. Dernæst beskrives de oplysninger, der kan opnås fra de spørgsmål om alkoholvaner i den nyeste udgave af European Social Survey, ESS. Til sidst sammenholdes disse oplysninger om alkoholforbrug i ESS med ESS's velkendte variable om lykke og selvvurderet helbred.

Indledning

I aviser og andre medier er det den gængse opfattelse at problemerne med alkohol er stigende. Det trak fx store overskrifter, at den traditionelle gymnasiefest ved skolestart i august 2016 i Ulvedalene i Dyrehaven nord for København endte med ca 10 ambulanceindlæggelser og 180 samaritterbehandlinger. Den almindelige mening er, at de unge drikker mere og mere. Spørgsmålet er om det er rigtigt, at de unge drikker mere og mere?

Mange går videre og hævder at de unge drikker så meget, fordi de ældre så sandelig også drikker for meget og derved er elendige rollemodeller. Spørgsmålet er om det er rigtigt, at de ældre mere og mere?

Hvor meget drikker danskere?

Som bekendt er der skat på alle alkoholiske drikke bortset fra lys øl etc. Det er derfor muligt meget præcist at opgøre, hvor meget alkohol, der lovligt sælges i Danmark.

Der udgives hvert år en rapport om grænsehandlen. Et af elementerne i disse rapporterer er, at der udføres stikprøvekontroller ved grænserne, så man kan se hvor meget alkohol, der indføres fra udlandet. Denne import fra udlandet er lovlig indenfor visse grænser. Tallene for den private import er dog meget usikre, da de er

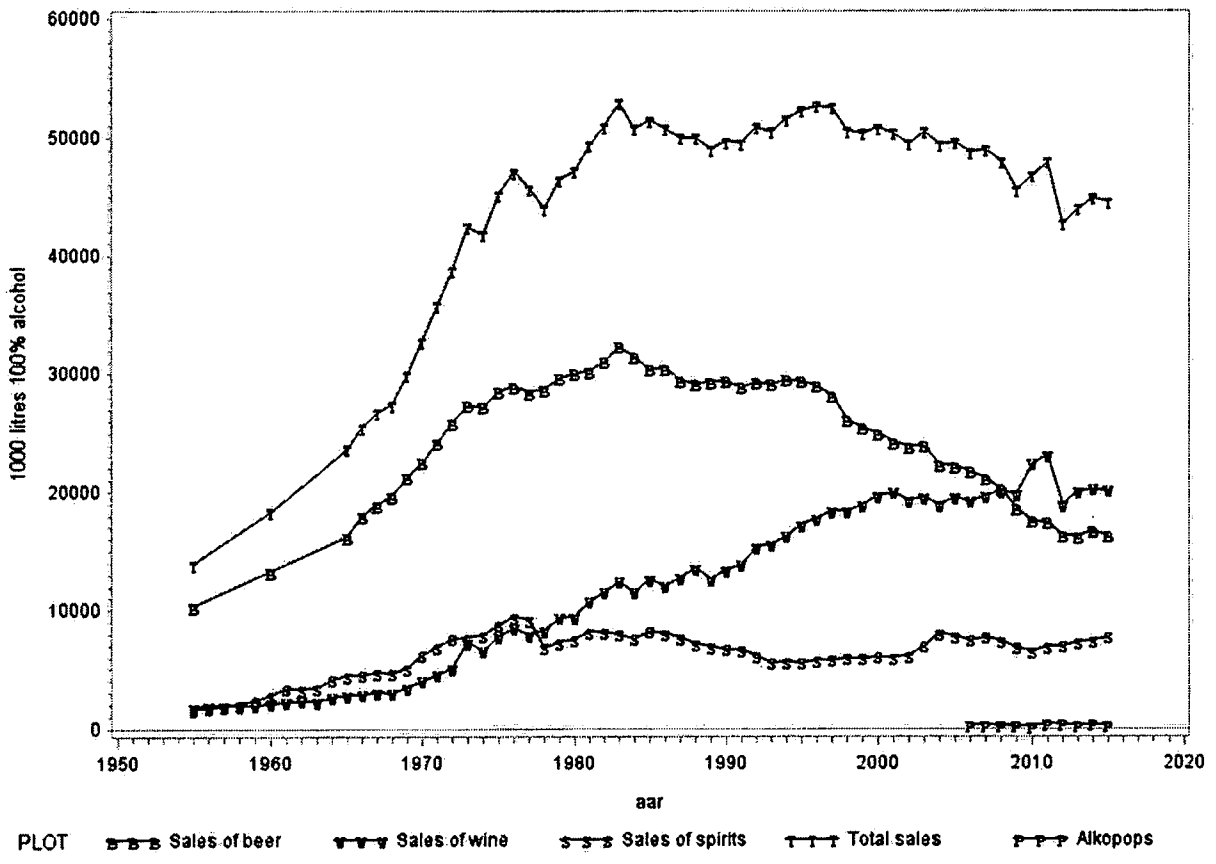
stikprøvebaseret, og da der sikkert er en tendens til underapportering. Der forekommer også decideret smugling fx via lastbiltransporter af andre varer.

Et yderligere problem går den modsatte vej, da det jo er velkendt at nordmænd og svenskere privatimporterer en del alkohol fra Danmark.

Så er der også de mange rejser, som ofte forbigås i debatten. Danskere drikker meget alkohol under ferier i udlandet og udlændinge drikker alkohol under ferier i Danmark. I medierne omtales dette forhold kun, når det gælder unge menneskers ture til Prag eller Sunny Beach.

Selvom det ikke er korrekt at sættes lighedstegn mellem alkoholsalg - og da slet ikke skattebelagt alkoholsalg - i Danmark og så den mængde alkohol danskerne drikker, er disse offentliggjorte lovpligtige tal den bedste kilde!

Det totale afgiftsbelagte alkoholsalg har jf figuren været svagt faldende, dog med fluktuationer siden en top i 1982. Figuren viser det samlede alkoholsalg i Danmark omregnet til liter ren 100% alkohol. Salget i 2016 var 10% lavere end i 1982. I disse næsten 25 år er antal indbyggere i Danmark steget med over 600.000, så gennemsnitsforbruget pr indbygger er altså faldet en del.

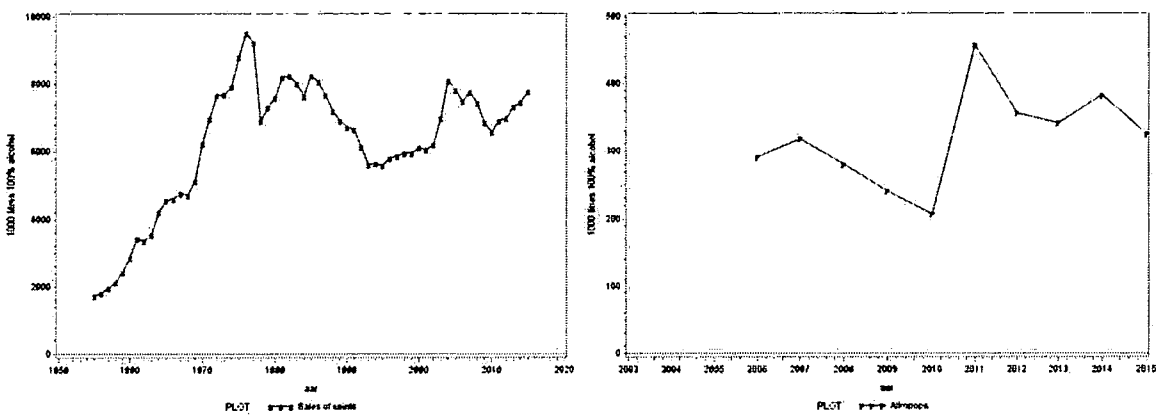


Hvad drikker danskerne?

Der er forskellige skattesatser på øl, vin og spiritus. Det gør det muligt at studere udviklingen i forbruget af disse tre hovedtyper alkohol. Figuren viser, at ølsalget styrtdykker mens vinsalget stiger og stiger. Jeg har tidligere redegjort for, at dette skyldes ændrede vaner (man drikker ikke øl til pastaretter) og ikke ændrede prisrelationer.

Spiritussalget er nogenlunde konstant, når det bedømmes ud fra den første figur. Der er et par knæk i kurven. De skyldes ændrede afgiftssatser på spiritus i forhold til skattesatserne på øl og vin. Den sidste stigning i salget skyldes en kraftig reduktion i skatten på spiritus i 2001, hvor skatten på øl og vin samtidigt blev holdt konstant. Det er bemærkelsesværdigt, at det samlede alkoholsalg ikke steg selvom spirituslaget steg.

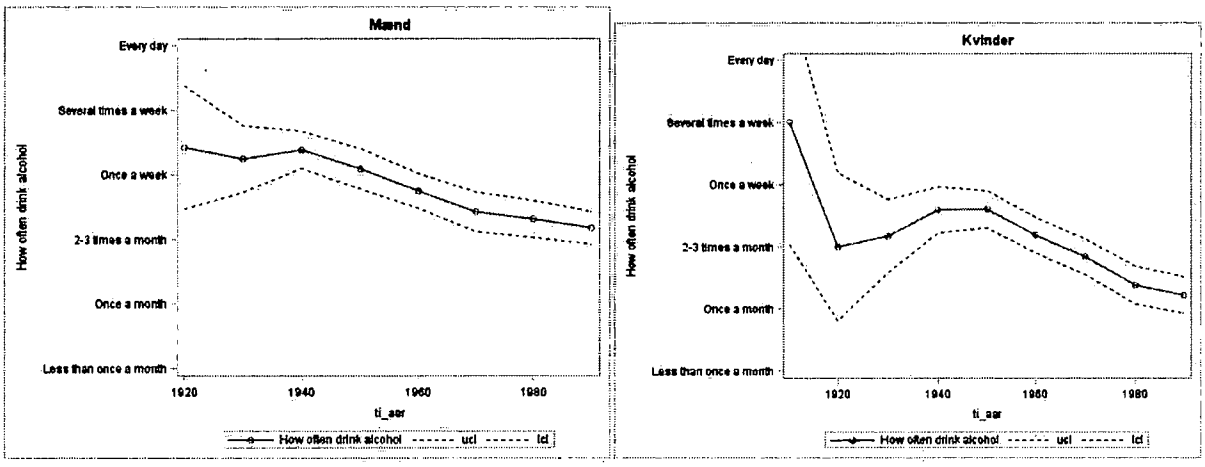
I de senere år har der været en særskat på alkopops og den slags blandingsprodukter. Figuren viser udviklingen i dette salg – igen omregnet til 100% alkohol. Salget er ubetydeligt i forhold til det øvrige - se P-erne i nederste højre hjørne på den første figur.



Ellers viser de mere præcise figurer for salget af spiritus og en del fluktuationer. I spiritussalget kan der med lidt god vilje ses en voksende tendens de sidste 15 år. Denne stigning kan være ungdommens øgede forbrug nyere produkter som af "shots", fx vodka med bolsjesmag. Alkopops salget er meget lavt og er uden betydning for det samlede billeder f danskernes drikkemønster.

Hvem drikker

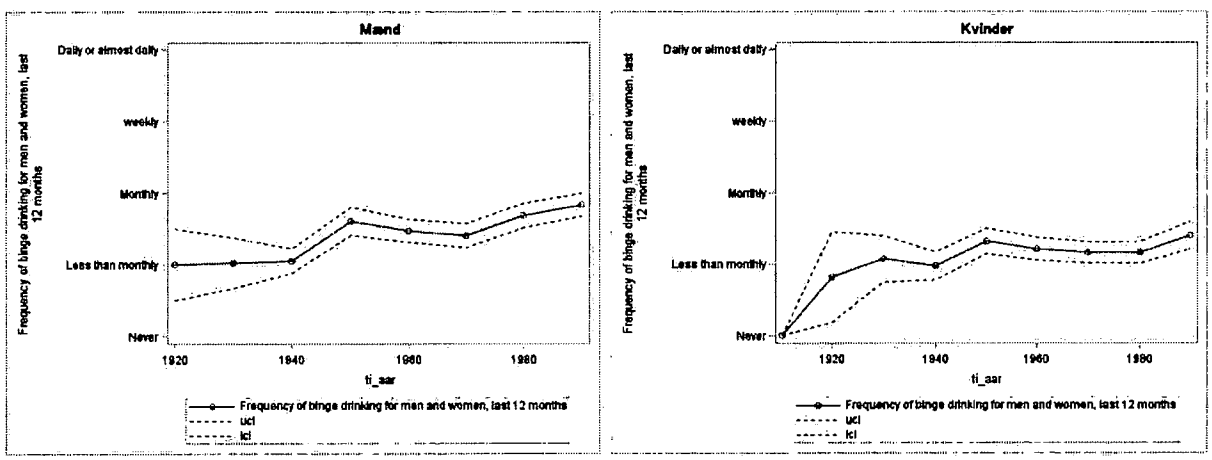
European Social Survey, ESS, indeholder i sidste runde, runde 7 i 2014, et par spørgsmål om alkoholforbrug. Der spørges bla til, hvor ofte, der drikkes alkohol. Figurerne viser svarene opdelt i gennemsnit for 10-års aldersklasser. Bemærk at x-aksen er fødselsår, der skal altså trækkes fra 2014, som er det år, hvor undersøgelsen blev gennemført. De adspurgte er i intervallet 15 til op imod 100 år, med de yngste længst til højre på den vandrette akse.



Der er to ældre damer over 95 i datasættet, længst til venstre den vandrette akse. De drikker ofte, men dette høje tal er naturligvis meget usikkert bestemt. Det ligger mere fast, at yngre drikker sjældnere end ældre i aldersintervallet fra 15 til 75 år både for mænd og kvinder. Kvinder i aldersintervallet 75 til 90 år drikker ikke så ofte som kvinder i 50-erne.

Nu er det jo ikke sikkert, at de nuværende 50-årige har bevaret deres drikkekvaner fra da de var i 20-erne til i dag. Men antages det, viser tallene altså, at ungdommen drikker sjældnere nu end tidligere.

Der spørges også til, hvor ofte respondenten "bingedrikker", dvs drikker mere end fem genstande ved samme lejlighed.



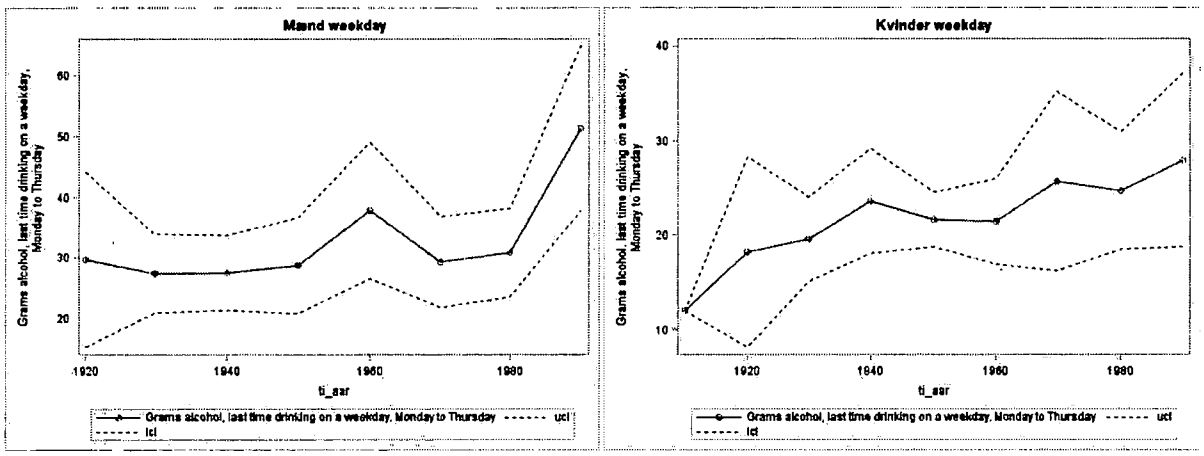
De to ældre damer bingedrikker aldrig! Men ellers viser svarene, at det højst foregår månedligt selv blandt de yngste. Tendensen til, at yngre bingedrikker oftere end ældre, ses også af figurerne, selvom den ikke er så markant.

Hvor meget drikkes der?

Der spørges til, hvor mange gram alkohol, der blev indtaget sidste gang respondenten drak på en hverdag, respektive på en weekenddag. Svarene er omregnet til gram alkohol, hvor en genstand svarer til tolv gram en alkohol. De viste gennemsnit er

gennemsnit, for de der rent faktisk har svaret på spørgsmålet, dvs at respondenter, der ikke drikker, ikke er medregnet i de viste gennemsnit.

Først ses på, hvor meget der drikkes på en hverdag. Her er det markant, at de yngste drikker langt mere på en gang end ældre.

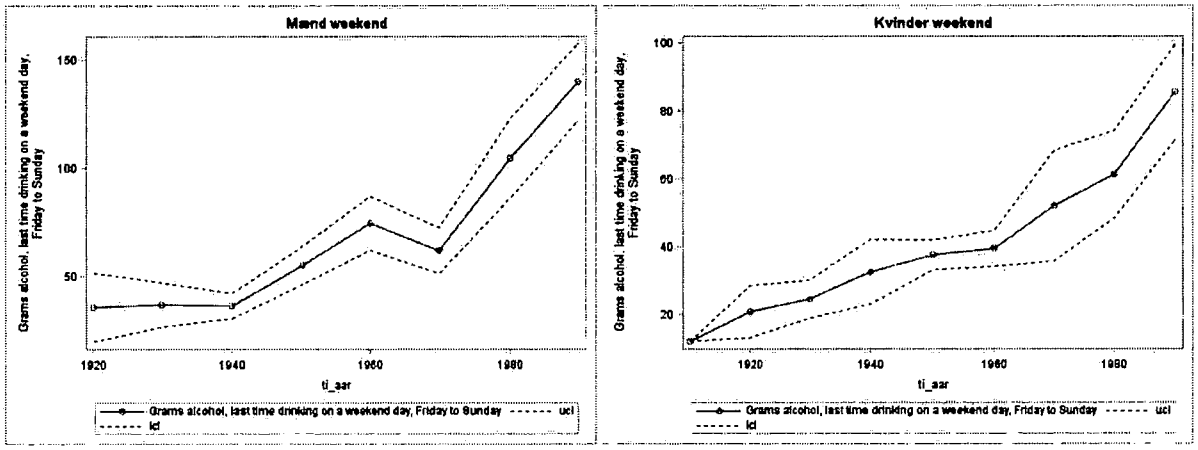


I den mere kuriøse afdeling ses, at de to ældre damer drikker præcis en genstand, dvs 12 gram alkohol, hver gang, de drikker uanset om det er hverdag eller weekenddag.

Gennemsnittet for de alleryngste mænd i intervallet 15 til 25 år er ca 50 gram dvs omkring fire genstande, mens det kun er omkring to genstande for de yngste kvinder.

Tendensen til, at kvinder drikker mere pr gang, når de drikker på en hverdag, jo yngre der er, er åbenlys ud fra den lineære trend på figuren. For mændene er niveauet mere konstant med en top for de ca 50 til 65-årige og så det markante løft for de under 25-årige.

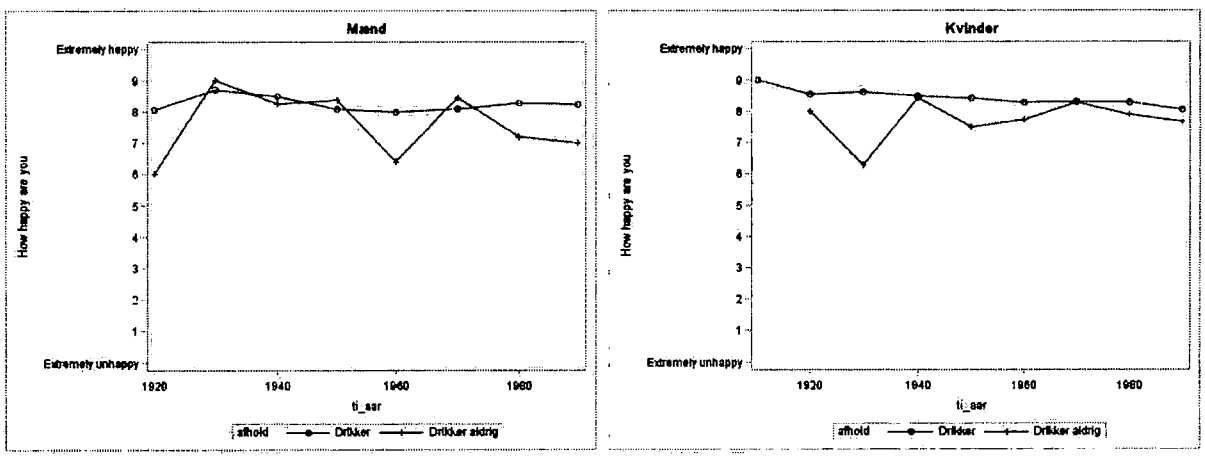
Når det gælder antal genstande, sidst der blev drukket på en weekenddag, er den opgående lineære trend fro begge køn meget dominerende. Det betyder altså, at både mænd og kvinder drikker langt mere pr gang i weekenden end ældre. For mænd i den yngste gruppe er tallet omkring 140 gram alkohol, dvs op imod tolv genstande, mens det "kun" er omkring otte genstande for kvinder i den yngste gruppe.



Bliver vi gladere af at drikke?

ESS stiller også spørgsmål om selvvurderet lykke og selvvurderet helbred.

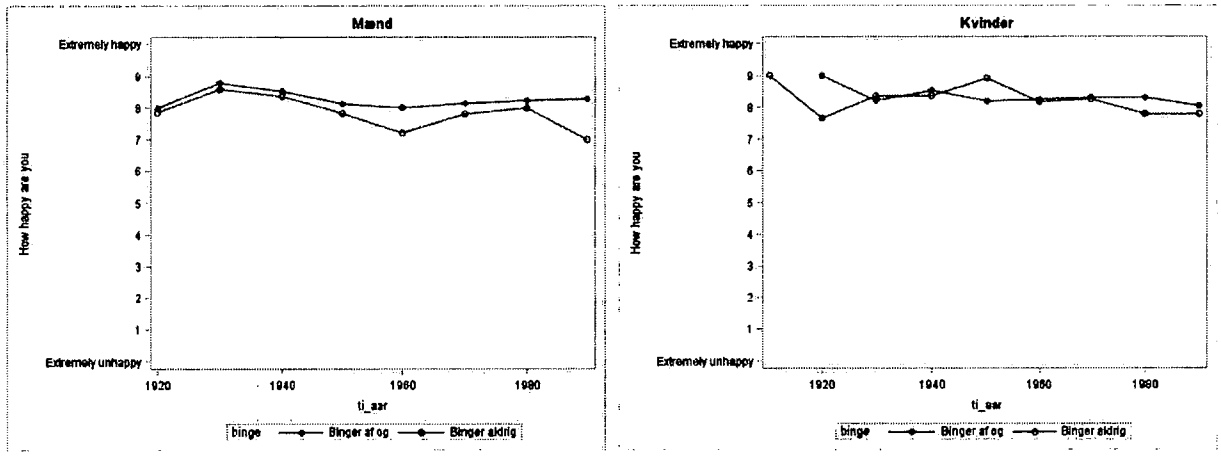
Den selvvurderede lykke er besvaret på en skala fra nul til ti, hvor ti er "extremely happy". På figurerne er den gennemsnitlige lykke for mænd og kvinder tegnet for tiårs aldersklasserne, idet der er delt op alt efter om respondenten er totalt afholdende eller drikker lidt (eller meget)



Igen i den kuriøse afdeling ses, at de to ældre damer, der drikker en genstand om dagen er de lykkeligste i hele datasættet. Mænds lykkefølelse er stort set uafhængig af alderen, mens kvinder ser ud til at blive en anelse lykkeligere jo højere alder, der har - kurven har i al fald en svag nedadgående trend fra ældre mod yngre.

Der er ikke markante forskelle på kurverne for afholdende og drikkende respondenter. Der er langt flere, der drikker i det mindste lidt, end der er totalt afholdende i datamaterialet. Det betyder at punkterne for de totalt afholdende er usikkert bestemt, så denne kurve fluktuerer en del. Det generelle billede er, at de totalt afholdende er en anelse mindre lykkelige end de, der drikker.

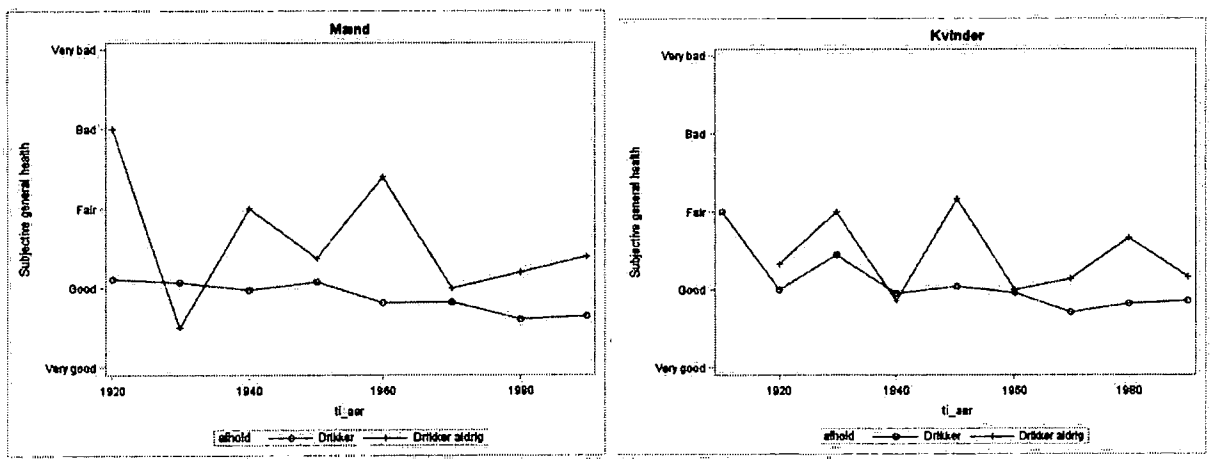
De tilsvarende kurver for respondenter, der bingedrikker (dvs drikker mere end 5 genstande ved samme lejlighed) bare en gang imellem og respondenter, der aldrig bingedrikker giver mere jævne kurver, da der flere, der aldrig bingedrikker, end der er totalt afholdende. Billedet er, at de mænd, der aldrig bingedrikker, er knapt så lykkelige som de, der bingedrikker af og til. Bortset fra de to ældre damer, der aldrig bingedrikker, men som fremstår som de lykkeligste i hele datasættet.



Bliver vi sundere af at drikke?

Svaret er jo klart NEJ i følge alle Sundhedsstyrelsens anvisninger. Ved et tidligere symposieindlæg, Milhøj(1987), er det påvist, at dødeligheden af skrumpelever hænger markant sammen med det totale alkoholforbrug set over en over hundrede års periode.

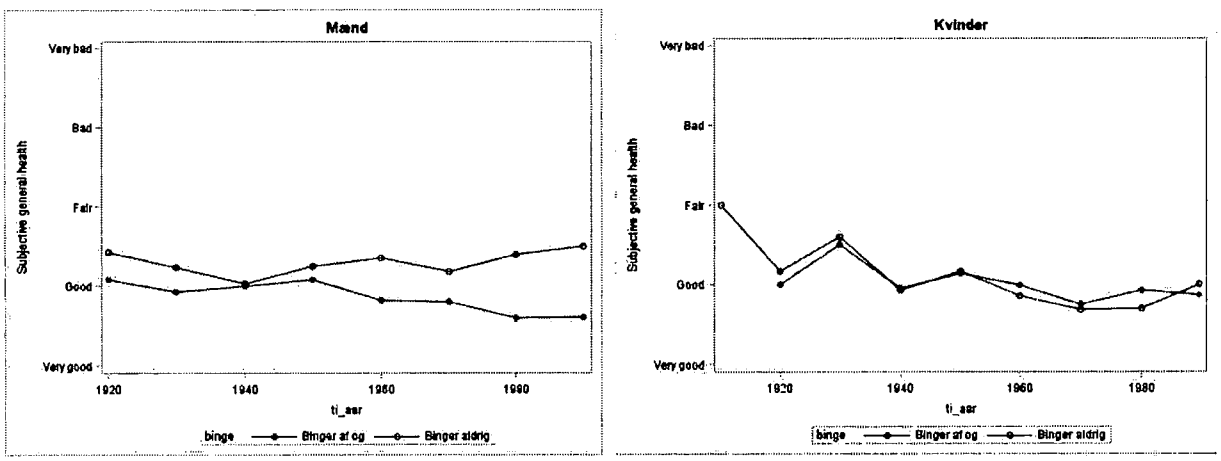
I dette afsnit behandles i stedet den selvvurderede lykke fra ESS datasættet, og her er billedet snarere det modsatte. Men det skal selvfølgelig bemærkes, at de viste grafer på ingen måde underbygger en kausalitet mellem et vist alkoholforbrug og godt helbred. Det kan lige så vel være den modsatte effekt, at det kræver et godt helbred at drikke!



De to meget gamle damer, der drikker moderat, har et "fair" helbred, hvilket jo nok skyldes, at de ofte betænker alternativet. De ældste mænd har et decideret "bad" helbred, hvis de ikke drikker. Hvis de derimod drikker, er helbredet i gennemsnit næsten "good".

For de drikkende opleves helbredet svagt dårligere med alderen, dvs en svag nedadgående tendens fra venstre mod højre på figurerne. Da der er ret få, der ikke drikker, er udsvingene større på kurven for de afholdende. Men tendensen er, at afholdende oplever et dårligere helbred end respondenter, der drikker.

For bingedrikning fås et klarere billede, da der flere, der aldrig bingedrikker, end der er totalt afholdende i datamaterialet.



For kvinder er der absolut ingen forskel at spore mellem de, der bingedrikker af og til og de, der aldrig gør det. For mænd er den tendens at jo yngre mændene er, jo dårligere opleves helbredet bland de, der aldrig bingedrikker, ned blandt de, der bingedrikker en gang imellem. Men denne effekt kan jo som tidligere nævnt sagtens vare kausalitet den modsatte retning.

Referencer

Anders Milhøj, 1987. Sammenhængen mellem det totale alkoholforbrug og antal dødsfald af alkoholrelaterede sygdomme analyseret ved tidsrækkemetoder. *Symposium i Anvendt Statistik, Uni-C, Århus*

Searching for outbreaks of hundred-year-oldness in Denmark

Anne Vinkel Hansen[1,2*], Laust Hvas Mortensen[1,2], Rudi Westendorp[1,2,3]

[1] Statistics Denmark, Denmark
[2] Department of Public Health, University of Copenhagen, Denmark
[3] Center for Healthy Aging, University of Copenhagen, Denmark
* Corresponding author: Statistics Denmark, Sejerøgade 11, 2100 København Ø, Denmark, email: aih@dst.dk, telephone: +45 39175083

Interest in geographical regions with high prevalence of centenarians has been ongoing at least since the start of the 20th century (Poulain et al 2013). The hope is that studying the population of such regions will be a first step on a road to discovery of determinants of longevity and healthy aging. However, living to a 100 is in a sense an arbitrary marker, and a high prevalence of centenarians in a region is not necessarily a region where people are generally healthier at a higher age, or even where they are more likely to reach 80.

The search for high-longevity areas has seen a number of false starts, where attempts at validation of the age of the oldest inhabitants of a newly discovered high-longevity area revealed that they had been misremembering or misreporting (Poulain et al. 2013). Here, time is working for longevity researchers, as centenarians become increasingly more prevalent and public records become better - even when dating back a century. Still, the definition of exactly what is meant by "a high prevalence of centenarians" in a particular study seems to be determined by data availability, as is the choice of where to look for high-longevity regions.

The "blue zone" in Sardinia, Italy, (Poulain et al 2004) a small group of villages with a significantly increased number of hundred-year-olds per 1000 births from 1880 to 1900 is the first of the current generation of well-validated high-longevity areas. The Sardinian blue zone is characterized by high male-to-female rates among centenarians. Other high-longevity areas are the Okinawa region in Japan (Willcox et al 2008), the island of Ikaria in Greece (Panagiotakos et al 2010) and the Nicoya peninsula in Costa Rica (Rosero-Bixby et al 2013).

The search for an explanation of the known "blue zones" is still ongoing. For high-longevity zones in Calabria, Italy, areas with high prevalence of centenarians have been noted to also have a low variety of surnames, hinting at an explanation grounded in inbreeding (Montesanto et al 2007). Topographic factors, particularly altitude and steepness of terrain have been suggested (Pes et al.2013). Life style factors have been investigated with varying results - the population of the Nicoya region have a diet righ in fibres, proteins and trans fats (Rosero-Bixby et al 2013), while Okinawa is the only

blue zone to have a significantly lower caloric intake than the reference population (Willcox et al 2007).

Population and methods

We construct a cohort of men and women born in Denmark 1906-1915 and still alive and resident in Denmark by Jan 1st 1977. Parish of birth is recorded in Statistics Denmark's registry data, and the cohort can be followed from age 71 to death, emigration or age 100.

We search for clusters of centenarians using a Kulldorf spatial scan (Kulldorf 1997) as implemented in the R package SpatialEpi. For each parish, the number of centenarians and the expected number based on number of births by sex and year of birth are assigned to the geographical centroid of the parish. The method then constructs zones by aggregating neighboring areas until a pre-specified proportion of the total population is included. For each zone, the likelihood is computed assuming a Poisson distribution of the number of observed centenarians. The zone most likely to be a cluster, plus any secondary clusters significant at the desired level, are detected by a likelihood ratio test, and significance measures are computed by Monte Carlo simulation.

Results

According to Statistics Denmark (Statbank.dk/HISB3), there were 740,927 live births from 1906 to 1915. Of these, we find 425,791 still alive and resident in Denmark by age 71. In order to make analyses by geographic location of birth, we make exclusions as follows: Those emigrating between age 71 and 100 (n = 792), those lost in registry (n = 752) and those who do not have a parish of birth recorded (n = 51,425). Of the latter group, the majority (n = 43,166) have a record of municipality of birth in place of parish. Of the 372,822 individuals remaining in the study, 4,859 (1.3 %) reach the age of 100. The distribution by age and sex is shown in table 1. As expected, the number of centenarians increases markedly over the period, with a much higher proportion of centenarians among women than men.

Table 1: Distribution by sex and birth year of births, number reaching age 71 and number reaching age 100

	Women				Men			
	Total live births*	Alive age 71	Alive age 100	Centenarians per 100 71-year-olds	Total live births*	Alive age 71	Alive age 100	Centenarians per 100 71-year-olds
Total	361,531	203,811	4006	1.97	379,396	169,011	853	0.50
Birth year								
1906	36,181	18,921	320	1.69	38,036	15,251	57	0.37
1907	36,189	19,214	318	1.66	38,135	15,631	72	0.46
1908	37,405	20,199	372	1.84	38,828	16,382	68	0.42
1909	37,130	20,857	383	1.84	39,171	17,315	83	0.48
1910	36,677	20,754	407	1.96	38,622	17,269	96	0.56
1911	36,135	20,699	410	1.98	37,798	17,398	90	0.52
1912	36,338	21,159	454	2.15	38,321	17,828	111	0.62
1913	35,481	21,046	428	2.03	36,994	17,502	91	0.52
1914	35,785	21,178	442	2.09	37,509	17,542	93	0.53
1915	34,210	19,784	472	2.39	35,982	16,893	92	0.54

References

Kulldorff, Martin. "A spatial scan statistic." Communications in Statistics-Theory and methods 26.6 (1997): 1481-1496.

Montesanto, A., et al. "Spatial analysis and surname analysis: complementary tools for shedding light on human longevity patterns." Annals of human genetics 72.2 (2008): 253-260.

Panagiotakos, Demosthenes B., et al. "Sociodemographic and lifestyle statistics of oldest old people (> 80 years) living in ikaria island: the ikaria study." Cardiology research and practice 2011 (2011).

Pes, Giovanni Mario, et al. "Lifestyle and nutrition related to male longevity in Sardinia: an ecological study." Nutrition, Metabolism and Cardiovascular Diseases 23.3 (2013): 212-219.

Poulain, Michel, et al. "Identification of a geographic area characterized by extreme longevity in the Sardinia island: the AKEA study." Experimental gerontology 39.9 (2004): 1423-1429.

Poulain, Michel, Anne Herm, and Gianni Pes. "The Blue Zones: areas of exceptional longevity around the world." Vienna Yearbook of Population Research (2013): 87-108.

Rosero-Bixby, Luis, William H. Dow, and David H. Rehkopf. "The Nicoya region of Costa Rica: a high longevity island for elderly males." Vienna yearbook of population research/Vienna Institute of Demography, Austrian Academy of Sciences 11 (2013): 109.

Willcox, Bradley J., et al. "Caloric restriction, the traditional Okinawan diet, and healthy aging." Annals of the New York Academy of Sciences 1114.1 (2007): 434-455.

Willcox, D. Craig, et al. "They really are that old: a validation study of centenarian prevalence in Okinawa." The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 63.4 (2008): 338-349.

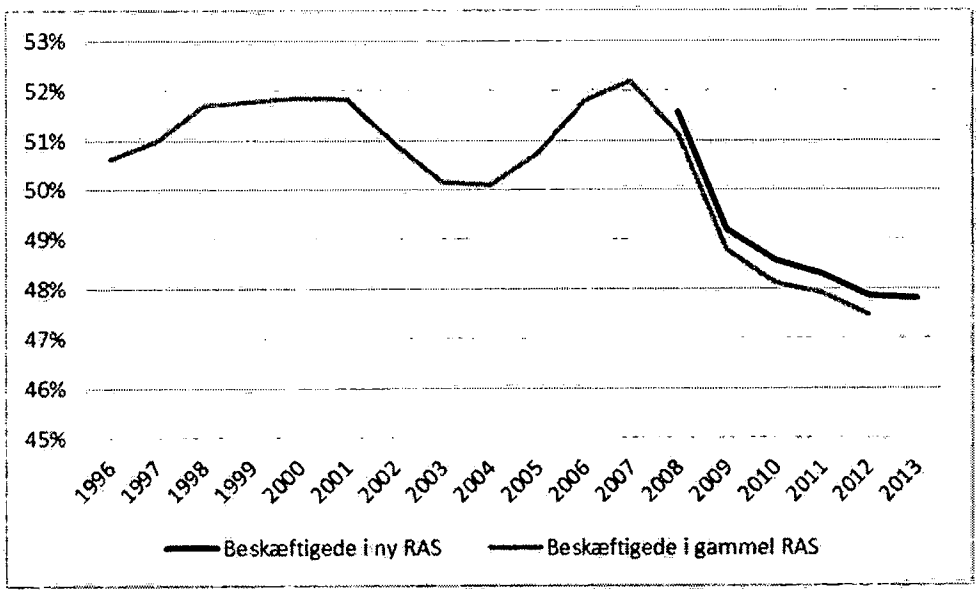
Mikrosimulering af RAS 2000-2007

Jens Clausen, CRT

Center for Regional- og Turismeforskning (CRT) udvikler og vedligeholder Den regionaløkonomiske model SAM-K/LINE for den danske økonomi. Modellen anvender data fra mange forskellige kilder såsom mikrodata på person- og virksomhedsniveau, kommunale nationalregnskaber, ADAM-fremskrivninger, og mange flere. En af de centrale datakilder er RAS-registeret fra Danmarks Statistik, hvorfra bl.a. anvendes variablen for socioøkonomisk status, SOC_STATUS_KODE, som primær kilde til at opgøre befolkningens arbejdsmarkedssituation.

Ved omlægningen af RAS i 2015, som opdaterede registret for årene tilbage til 2008, ændredes datakilder og opgørelsesmetoder så der opstod et databrud i 2008. Der er væsentlige forskelle mellem ny og gammel (fx giver de forskellige antal beskæftigede) så det er ikke umiddelbart muligt at lave sammenhængende serier længere tilbage end til 2008, hvad der giver en ret kort periode som endda indeholder ret store konjunkturændringer. I figur 1 herunder ses andelen af beskæftigede, trukket direkte fra ny og gammel RAS:

Figur 1: Andel hovedbeskæftigede i ny og gammel RAS



Kilde: RAS forskningsregister leveret af Danmarks Statistik. Det skal bemærkes, at der arbejdes med en fejlbehæftet version af RAS2008-2013, som forventes rettet i slutningen af januar 2017.

Det kan ikke ses på figuren, men forskellen i årene 2008-2012, hvor begge tal findes, skyldes dels det at ca 20.000 personer som er beskæftigede i gammel RAS ikke er det i ny RAS og dels at det omvendte gælder for ca det dobbelte antal personer, så der for hvert år i alt er ca 22.000 flere beskæftigede efter revisionen.

For at kunne anvende længere sammenhængende serier i modellen har vi mikrosimuleret flere variabler i RAS 2000-2007, bl.a. SOC_STATUS_KODE og ARBKOM (arbejdsstedskommune). I forbindelse med mikrosimuleringen er der fokuseret på, at fordelingerne kommer til at 'passe' bedst muligt indenfor de kombinationer af alder, uddannelse mv. som anvendes i modellen.

Metode

Formålet er her at danne den 'nye' variabel SOC_STATUS_KODE i persondata for årene 2000-2007 sådan at den kan bruges som udgangspunkt for modelberegninger mm. Som det ses i Tabel A.1 er der en meget stærk sammenhæng mellem nye og gamle værdier, specielt ses det at der for næsten hver eneste værdi af (den gamle) SOCSTIL_KODE findes én værdi af (den nye) SOC_STATUS_KODE som 90%-99% af personerne har. SOC_STATUS_KODE kunne med god rimelighed simpelthen bestemmes som den hyppigste værdi for hver SOCSTIL_KODE, men for at få den rette overordnede fordeling af SOC_STATUS_KODE trækkes der i stedet for hver person en SOC_STATUS_KODE fra den relevante fordeling.

Konkret bliver befolkningen i 2008 opdelt i grupper efter SOCSTIL_KODE samt baggrundsfaktorer som alder, køn, uddannelse og branche (se liste i tabel A.2). Indenfor hver gruppe beregnes en nøgle med sandsynlighederne for at have de forskellige værdier af SOC_STATUS_KODE. Gevinsten ved at generere SOC_STATUS_KODE ud fra flere akser end blot SOCSTIL_KODE er, at fordelingen kommer til at passe bedst muligt indenfor hver enkelt kombination af akser. Befolkningerne i årene før 2008 opdeles derefter i grupper på samme måde og sandsynlighederne fra 2008 bruges til at trække værdier af SOC_STATUS_KODE. Hvis en person i simulationsåret har en kombination af baggrundsfaktorer, som ikke findes i 2008, betinges ud fra så mange af de nævnte variabler som muligt.

Karakteristika for de fleste personer i 2000-2007 findes eksakt i 2008-data (mindst 68% som i 2000) og for næsten alle (mindst 99% som i 2000) findes de centrale karakteristika som anvendes i basismodellen (socstil_kode, branche, alder, uddannelse og køn).

For at forbedre resultatet er det nødvendigt at fordelingen af SOCSTIL_KODE er fastholdt i årene 2000-2008. Som det ses i tabel A.3 og beskrives i den online dokumentation[1], har variablen flere væsentlige databrud i perioden, bl.a. pga. ændrede datakilder og opgørelsesmetoder. For at sikre at alle kategorier af SOCSTIL_KODE så vidt muligt fastholdes i årene 2000-2008, bliver følgende kategorier lagt sammen:

- Arbejdsløshedsforsikrede selvstændige (CRAM selvstændige) og årsafgrænsede selvstændige (AKM selvstændige) samles i én kategori før 2008.
- Brutto- og nettoledige samles i én kategori fra 2008 ligesom de var indtil 2007.
- Personer i tilbagetrækning fra arbejdsstyrken: Denne overordnede gruppering består ifølge statistikdokumentationen af de tre følgende undergrupper: Modtagere af efterløn, modtagere af overgangsydelse (udgår i 2007) samt modtagere af flexydelse (tilkommer i 2008). Disse tre kategorier samles i én i gennem hele perioden.
- De tre kategorier Uddannelsesforanstaltning, Integrationsuddannelse (2001-2007) og Anden/Særlig aktivering (udgår i 2007) samles i én kategori. Dette er gjort ud fra statistikdokumentationens bilag "Foranst_tilstand_socstil.xls", der viser hvordan de forskellige foranstaltninger er sammenflettet.

I hvert år gælder det for et væsentligt antal personer, at de simuleres til at ville have været beskæftigede ifølge ny RAS men ikke var det ifølge observeret gammel RAS. Disse personer kommer derfor til at mangle oplysninger om arbejdssted og beskæftigelse[2] som derfor simuleres. For at sikre sammenhængende oplysninger om arbejdssted, branche mv. gøres dette ved at trække arbejdsstedsoplysningerne fra en tilfældig lignende person. Der trækkes oplysninger fra en tilfældig person som har fået

---

1 http://www.dst.dk/da/TilSalg/Forskningsservice/Dokumentation/hoejkvalitetsvariable/befolkningens-tilknytning-til-arbejdsmarkedet--ras-/socstil-kode

2 Udover arbejdsstedskommune og -adresse mangler bl.a. også branche og funktionskode.

prædikteret samme SOC_STATUS_KODE, bor i samme kommune og er af samme køn.

Simulationsresultater

Der er simuleret for hele perioden 2000-2012. De primære resultater af simulationerne er værdierne af SOC_STATUS_KODE og specielt hvad det betyder for beskæftigelsestallet, som kan ses i Tabel 1 herunder.

Tabel 1: Beskæftigede 2000-2013, observeret og simuleret

	Observeret		Simuleret				Observ.
År	i gammel RAS		Ikke besk	Nye	I alt	Simuleret	i ny RAS
2000	2772868	51,84%	-22984	+33749	+10765	52,04%	
2001	2782306	51,83%	-23684	+35066	+11382	52,04%	
2002	2741386	50,92%	-21969	+37338	+15369	51,21%	
2003	2706434	50,14%	-21862	+38439	+16577	50,45%	
2004	2710462	50,09%	-17541	+37677	+20136	50,46%	
2005	2754646	50,75%	-17029	+37117	+20088	51,12%	
2006	2821641	51,80%	-17121	+35193	+18072	52,13%	
2007	2857565	52,19%	-16941	+33882	+16941	52,49%	
2008	2816542	51,26%	-18626	+44058	+25432	51,72%	51,72%
2009	2698818	48,89%	-17421	+58053	+40632	49,63%	49,34%
2010	2674951	48,26%	-17768	+59970	+42202	49,02%	48,76%
2011	2655765	47,72%	-18167	+62027	+43860	48,51%	48,45%
2012	2659274	47,60%	-17552	+59287	+41735	48,35%	48,03%

Tabel 1 viser, at den simulerede afgang fra beskæftigelse er nogenlunde konstant gennem perioden, mens tilgangen er relativt lav før 2008 og helativt høj efter 2008. Antallet af personer simuleret til at afgå hhv. tilgå gruppen af beskæftigede, afgøres af befolkningens fordeling af baggrundsfaktorer samt overgangssandsynligheder fra gammel til ny RAS2008.

Ligesom simulationen i 2008 rammer meget præcist, ville 2009-2012 også kunne simuleres præcist hvis der var anvendt en simulationsnøgle baseret på de enkelte år. Imidlertid er formålet med at simulere for 2009-2012 at kunne vurdere betydningen af at simulere i andre år end basisåret. Grunden til at tilgangen i årene 2009-2012 simuleres ca 15.000 for højt kan i hvert af årene deles op i to omtrent lige store kilder,

idet ca. 8.000 skyldes ændringer i den gennemsnitlige fordelingsnøgle for den overordnede SOCSTIL_KODE i forhold til 2008 og resten skyldes ændringer i de detaljerede fordelingsnøgler. Dette illustrer, at fordelingsnøglen fra gammel SOCSTIL_KODE til ny SOCIAL_STATUS_KODE ikke er helt konstant efter 2008.

Udover at fordelingsnøglen ikke er konstant over tid, påvirker det også resultaterne, at klassifikationen i RAS ikke ser ud til at være helt konsistent over tid. Dette ses i tabel A.3 hvor der fra år til år er store ændringer i antallet af personer i de enkelte kategorier. F.eks. er der mellem 2007 og 2008 store forskelle i antallet af bl.a. momsbetalere, topledere, lønmodtagere uden nærmere angivelse, uddannelsessøgende og øvrige udenfor arbejdsstyrken. Dette ville i sig selv, dvs. hvis den observerede frodelingsnøgle var konstant over tid, påvirke det simulerede antal af til- og afgange fra beskæftigelse.

Selv om der primært er simuleret for perioden 2000-2007, hvorfra simulationsresultaterne skal anvendes, er det også meget relevant at se på resultaterne for perioden 2008-2012 hvor det simulerede kan sammenlignes med observerede data. Tabel 2 herunder viser hvor godt simulationen rammer det observerede.

Tabel 2: Andel personer med korrekt simuleret SOC_STATUS_KODE og beskæftigelse

	2008	2009	2010	2011	2012
Eksakt simuleret SOC_STATUS_KODE	96,37%	93,04%	92,99%	93,12%	93,57%
Korrekt simuleret beskæftigelse	98,70%	98,00%	97,94%	97,94%	97,98%

Som det ses i tabel 2, rammer simulationerne bedst i 2008 hvor 96,37% er simuleret til den eksakt rigtige SOC_STATUS_KODE, hvad der passer godt med den forventede in-sample præcision på 96,36%. Når træfsikkerheden er lavere i de efterfølgende år, er det simpelthen udtryk for at sammenhængen mellem den gamle SOCSTIL_KODE (med baggrundskarakteristika) og den nye SOC_STATUS_KODE ikke er helt konstant gennem perioden.

I forhold til Den regionaløkonomiske model SAM-K/LINE er det også relevant at have en ide om simulationspræcisionen indenfor cellerne i de mangedimensionale matrix, som modellen arbejder med.

Tabel 3: Simuleret beskæftigelse i modellens celler

	2008	2009	2010	2011	2012
Beskæftigelse, ny RAS	51,72%	49,34%	48,76%	48,45%	48,03%
Beskæftigelse, simuleret	51,72%	49,63%	49,02%	48,81%	48,35%
Gennemsnitlig absolut afvigelse	0,05%	0,46%	0,50%	0,56%	0,53%

I tabel 3 ses det, at den gennemsnitlige simulerede celle-beskæftigelse i hvert af årene efter 2008 er ca. 0,3% for høj. Dette er dog udtryk for at beskæftigelsesandelen har en gennemsnitlig absolut afvigelse på ca. 0,5%.

Tabel A.1

RAS2008

Fordeling af (ny) SOC_STATUS_KODE for hver værdi af (gl) SOCSTIL_KODE

(Hovedrecords, rækkeprocent)

	N	110 Selvstændige (primær status ult. nov.)	120 Medarbejdende ægtefæller	131 Topledere	132 Lønmodtagere på højeste niveau	133 Lønmodtagere på mellemniveau	134 Lønmodtagere på grundniveau	135 Andre lønmodtagere	136 Lønmodtagere u.n.a.	200 Arbejdsløse	311 Støttet beskæftigede uden løn	312 Feriedagpenge	313 Vejledning og opkvalificering	314 Ledighedsydelse	315 Børnepasningsorlov fra ledighed	316 Barselsfravær fra ledighed	317 Sygefravær fra ledighed	318 Kontanthjælp (passiv)	319 Introduktionsydelse	320 Revalidering	411 Førtidspension	412 Efterløn	413 Fleksydelse	414 Folkepension	415 Anden pension	511 Personer under uddannelse	512 Børn og unge	513 Øvrige uden for arbejdsstyrken
Alle	5501426	3.7	0.1	1.7	6.7	9.6	20.0	5.2	4.7	1.2	0.1	0.0	0.5	0.2	0.0	0.1	0.6	1.0	0.0	0.1	3.8	2.2	0.1	14.0	0.2	14.6	7.3	2.4
115 Arbejdsgiver	58416	92	0.0	0.8	1.3	1.2	1.8	0.4	2.0	0.0	0.0	.	0.0	0.0	.	0.0	0.1	0.0	.	.	0.0	0.0	0.0	0.1	0.0	0.0	.	0.1
116 Momsbetaler	135825	96	0.1	0.1	0.3	0.2	0.7	0.2	0.6	0.2	0.0	0.0	0.1	0.1	.	0.1	0.4	0.1	.	.	0.2	0.1	.	1.1	0.0	0.0	.	0.1
118 AKM-selvstændige	14113	41	0.9	0.0	0.2	0.1	0.3	0.1	0.5	1.5	0.3	0.0	0.5	0.7	0.1	1.0	3.9	0.8	.	0.1	1.9	2.0	0.4	17	2.7	3.8	0.0	20
120 Medarbejdende ægtefælle	6292	0.3	92	.	.	0.0	0.1	0.0	0.1	0.0	.	.	0.0	0.0	.	0.9	1.6	.	.	.	0.4	0.0	.	2.1	0.0	0.1	.	2.1
130 Lønmodtager uden nærmere angivelse	260862	0.5	0.1	0.6	0.9	0.8	5.5	1.8	89	0.0	.	0.0	0.0	0.0	0.0	0.1	0.2	0.0	.	0.0	0.0	0.0	.	0.0	0.0	0.2	.	0.1
131 Topleder	92519	0.1	0.0	99	0.1	0.1	0.2	0.0	0.1	0.0	.	.	.	.	.	0.0	0.0	.	.	.	.	.	.	.	.	0.0	.	0.0
132 Lønmodtager på højeste niveau	367961	0.3	0.0	0.1	98	1.0	0.4	0.1	0.5	0.0	.	.	.	.	.	0.1	0.0	.	.	.	0.0	0.0	.	0.0	0.0	0.0	.	0.0
133 Lønmodtager på mellemniveau	524357	0.1	0.0	0.1	0.4	98	0.3	0.1	0.8	0.0	.	.	0.0	.	0.0	0.1	0.0	.	.	.	0.0	.	0.0	0.0	.	0.0	.	0.0
134 Lønmodtagere på grundniveau	1082502	0.2	0.0	0.0	0.1	0.3	98	0.2	0.5	0.1	0.0	0.0	0.0	0.0	.	0.1	0.1	0.0	.	0.0	0.0	0.0	0.0	0.0	.	0.0	0.0	0.0
135 Andre lønmodtagere	274108	0.2	0.0	0.0	0.1	0.1	0.7	98	0.6	0.1	0.0	.	0.0	0.0	.	0.1	0.1	0.0	.	.	.	0.0	.	0.0	0.0	0.0	0.0	0.0
200 Nettoledige	61515	0.0	.	0.0	0.1	0.2	1.9	1.1	0.5	95	0.1	0.0	0.5	.	.	.	0.0	.	.	.	.	0.0	.	.	.	0.0	.	0.1
310 Uddannelsessøgende	162851	0.3	0.0	0.0	0.4	0.4	3.1	1.1	0.8	.	.	.	0.0	.	.	.	0.0	0.0	.	0.0	0.0	.	.	.	0.0	92	.	2.2
317 Beskæftiget uden løn	9695	0.4	.	0.0	0.2	0.5	3.3	1.8	3.5	0.8	58	.	1.0	2.8	.	0.1	5.1	0.7	.	0.4	9.3	.	2.2	0.8	0.5	1.3	0.0	6.9
318 Orlov fra ledighed	1195	1.8	.	0.3	3.2	2.6	2.2	(-)	0.3	.	.	.	.	.	90	.	.	.	.	.	.	.	.	.	.	.	.	(-)
319 Uddannelsesforanstaltning/vejledning og opkvalificering	33470	0.1	.	.	0.0	0.0	0.2	0.1	0.2	5.4	0.2	.	82	0.1	.	0.0	0.1	1.0	0.1	0.3	0.2	.	0.0	0.0	0.0	6.0	.	3.5
322 Barselsdagpenge	5171	0.4	.	.	0.0	0.1	0.1	0.0	4.0	.	.	.	.	.	.	95	0.1	.	.	.	.	.	.	.	.	0.0	.	0.4
323 Sygedagpenge	29246	0.5	0.0	0.0	0.0	0.1	1.0	0.3	1.3	.	0.2	.	0.4	0.0	.	0.0	96	0.0	.	0.0	0.0	.	.	.	0.0	0.0	.	0.6
324 Efterløn	130551	0.5	0.0	0.0	0.1	0.1	0.3	0.1	0.2	.	.	.	.	.	.	.	0.0	.	.	.	0.0	94	2.4	0.0	0.7	0.1	.	1.2
326 Kontanthjælp	52973	.	.	0.0	0.0	.	0.0	0.0	0.1	0.2	0.2	.	1.3	0.0	.	.	0.0	98	.	.	.	.	.	.	.	0.1	.	0.0
327 Revalideringsydelse	4550	0.5	.	.	0.0	0.0	0.3	0.0	0.1	.	2.0	.	15.6	.	.	.	.	0.0	.	81	.	.	.	.	.	0.0	.	0.0
328 Tjenestemandspension	11310	1.2	0.0	0.1	0.4	0.2	0.9	0.4	0.5	.	.	.	.	0.0	.	.	0.1	.	.	.	0.0	0.2	0.0	0.0	63	7.2	.	26
329 Folkepensionist	779306	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.1	.	.	.	.	.	.	.	0.0	.	.	.	.	.	.	98	0.1	1.0	.	0.1
330 Øvrige uden for arbejdsstyrken	155941	3.2	0.1	0.0	0.2	0.3	1.4	0.6	0.5	1.2	.	0.2	0.0	0.0	.	0.0	0.1	0.0	.	0.0	0.0	0.0	0.0	0.2	0.0	15	.	77
331 Førtidspensionist	210846	0.3	0.0	.	0.0	0.0	0.1	0.0	0.2	.	.	.	.	.	.	.	.	0.0	.	.	99	.	.	0.3	0.0	0.0	.	0.0
332 Introduktionsydelse	1151	0.2	.	.	.	.	(-)	(-)	(-)	2.0	0.6	.	4.2	(-)	.	.	.	10	78	.	.	.	.	.	.	1.3	(-)	2.4
334 Ledighedsydelse	7916	.	.	(-)	.	.	0.1	(-)	0.1	.	0.4	.	1.1	98	.	.	.	.	.	.	.	.	.	.	.	.	.	0.1
400 Barn eller ung (d.v.s. under 16 år)	1026784	0.0	.	0.0	0.0	0.0	0.2	0.3	0.5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	60	39	.

Tabel A.2: Prioriteret liste med variabler som bruges til betinget simulation af SOC_STATUS_KODE

Variabel	Klassificering
Branche	Som i BASIS-model (11 hovedbrancher)
Alder	Som i BASIS-model (15 aldersgrupper)
Uddannelse	Som i BASIS-model og Velfærdsversion (32 uddannelsesgrupper)
Igangværende uddannelse	1 = Under uddannelse, 0 = Ikke (2)
Køn	Mand/Kvinde (2)
Bruttoindkomst	Opdelt efter fordeling indenfor ovenstående variabler: 0%-10%, 10%-25%, 25%-50%, 50%-75%, 75%-90%, 90%-100% (6 intervaller)
Arbejdsløshedsforsikring	Som bruttoindkomst (6)
Landsdel	(11 landsdele)
Familietype	Gift/reg, samlevende, enlig, barn (4)
Offentlig pension	1 = Modtager offentlig pension, 0 = Ikke (2)
Topskat	1 = Betaler topskat, 0 = Ikke (2)

Tabel A.3: Oversigt over SOCSTIL_KODE for hovedrecords i RAS2000-2012

SOCSTIL_KODE 2000-2012	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Alle	5349212	5368354	5383507	5397640	5411405	5427459	5447067	5475791	5511451	5534738	5560628	5580516	5602628
115 Arbejdsgiver	75048	73389	72070	70170	69645	68318	67938	66261	58429	54564	53429	51866	51286
116 Momsbetaler	115609	114203	106960	99482	105893	108302	109319	110155	135862	135750	132490	133124	132670
117 CRAM-selvstændige	12356	12295	12101	11832	4445	3471	3277	2897	.	.	.	.	.
118 AKM-selvstændige	7445	9566	7615	7868	7966	7624	7509	8547	14157	14610	14257	15247	14447
120 Medarbejdende ægtefælle	12815	11551	10179	9190	8492	7810	7207	6647	6294	5646	5129	5129	4984
130 Lønmodtager uden nærmere angivelse	353139	347311	352903	474353	435853	451878	487774	459574	261049	252760	225316	207102	200469
131 Topleder	62736	63639	60678	62781	66528	69840	70504	73626	92536	90448	99650	101239	101036
132 Lønmodtager på højeste niveau	318709	330611	332636	319227	330234	329912	342724	347512	368036	386585	574863	592669	603009
133 Lønmodtager på mellemniveau	402077	416994	419932	429867	440904	455492	462297	490780	524430	499455	303598	291256	286956
134 Lønmodtagere på grundniveau	1126402	1114485	1090265	1014132	1024608	1025647	1030990	1054905	1082826	1009791	1028730	1038241	1030409
135 Andre lønmodtagere	286532	288262	276047	207532	215894	226352	232102	236661	274338	250270	238633	237679	235083
200 Nettoledige	118520	110501	119250	147666	134586	107734	80270	59860	43895	95756	88308	84875	98615
201 Bruttoledige	.	.	.	.	.	.	.	.	17647	27755	40523	37374	33550
310 Uddannelsessøgende	120512	121184	132283	139946	140739	137045	125094	131527	163070	186480	202435	223568	235845
315 Flexydelse	.	.	.	.	.	.	.	.	3119	4500	5184	5648	6189
316 Delvis ledighed	.	.	.	.	.	.	.	.	2716	5316	5081	5214	5367
317 Beskæftiget uden løn	15571	15230	15693	12075	10256	10334	8642	10969	9752	10633	14021	17115	15563
318 Orlov fra ledighed	4187	4828	3985	1936	1826	1472	945	798	1197	639	36	34	17
319 Uddannelsesforanstaltning/vejledning og opkvalificering	27691	24024	25794	17499	20582	17417	11865	43800	31259	36109	38327	32726	24911
320 Særlig/aktivering	7921	9580	11583	14686	15087	15392	15880	.	.	.	.	.	.
322 Barselsdagpenge	1836	1613	5805	6895	6627	6607	5974	5086	5173	5491	6507	7525	8142
323 Sygedagpenge	5080	5174	15012	15511	14325	14602	14881	16674	29260	30270	30164	29078	29131
324 Efterløn	156278	160313	163507	175486	172888	150095	138215	138220	127441	121454	114084	103807	94884
325 Overgangsydelse	23311	18216	13438	9265	5483	2324	119	.	.	.	.	.	.
326 Kontanthjælp	50786	50317	52686	54284	57562	55049	44894	39284	53037	60463	64204	74016	82363
327 Revalideringsydelse	23772	22283	22126	21666	19944	18407	16253	4561	4553	3885	2760	1967	1679
328 Tjenestemandspension	2185	10416	10337	16140	18510	16379	14510	12112	11321	11645	11516	11035	11201
329 Folkepensionist	640414	640247	641317	644128	649117	712984	747518	760706	779370	800959	824798	851555	876159
330 Øvrige uden for arbejdsstyrken	128723	131261	125217	121215	125166	124979	124656	123875	155731	167167	173844	164664	172135
331 Førtidspensionist	223131	221163	222359	224198	236146	212477	208508	208596	210922	214341	216017	215980	214040
332 Introduktionsydelse	11857	4836	4259	3152	1984	1188	909	478	1157	1242	1227	1436	1869
333 Integrationsuddannelse	.	5922	6191	4583	2923	1904	1306	2571	2251	2994	.	.	.
334 Ledighedsydelse	.	.	2190	4398	6321	7045	7637	6632	7918	10091	10979	12509	11954
335 Aktivering iflg. kontanthj.statistikregister	.	.	10356	8895	6894	5492	5369	2365	.	.	.	.	.
400 Barn eller ung (d.v.s. under 16 år)	1014569	1028940	1038733	1047582	1053977	1053887	1051981	1050112	1032705	1037669	1034518	1026838	1018665

Regional udvikling og iværksætteri

Analyser af branchespecifikke forskelle baseret på mikrodata

Mogens Dilling-Hansen
Department of Economics and Business Economics
Aarhus Universitet
Mail: dilling@econ.au.dk

Resumé

Denne analyse tager udgangspunkt i den regionale udvikling i Danmark de seneste årtier, hvor det har været tydeligt at der sker en vandring fra land mod de større bysamfund. Spørgsmålet er, om dette skift vil fortsætte eller der har fundet en naturlig udjævning sted? Det eneste fornuftige svar på dette spørgsmål er, at det afhænger af hvordan vilkårene er for dem, der bor i yderområderne ... kan indbyggere i yderområderne af Danmark finde et job, så har de også en mulighed for at blive boende.

De præsenterede resultater er baseret på mikrodata for iværksættere i Danmark i perioden 2001 til 2013, og de registerbaserede data er aggregeret til relevante branche og regionale niveauer, således meso-analyser er mulige. Hovedresultater er, at der skabes signifikant flere nye virksomheder i større byområder end i yderområderne af Danmark, og det er oven i købet sådan, at de særligt interessante 'Pavitt-virksomheder' i særligt omfang startes i byområderne. Eneste trøst er, at overlevelsesraten er lidt højere på landet, og at fremstillingssektoren også for nye virksomheders vedkommende står stærkt i yderområderne.

1 Indledning

De seneste tiår har på mange måder illustreret, at selv om Danmark er lille i udstrækning og relativt homogent på de fleste økonomiske og sociale dimensioner, så er der alligevel markante ændringer i det geografiske billede. Større bysamfund, specielt Hovedstaden, oplever en markant positivt netto tilflytning, og modsat er områder væk fra bysamfundene præget af tilbagegang (eller i hvert fald stagnation) i indbyggertallet. Effekten er til at få øje på når man fokuserer på indkomstudviklingen generelt, og hvor specielt boligmarkedet udvikler sig i to forskellige spor - seriøse diskussioner om boligbobler føres samtidig med at liggetider vokser og priser på fast ejendom falder.

På det lange sigt er regionale flytninger ikke det store problem, så længe de realøkonomiske forhold er på plads, og det bemærkelsesværdige ved den aktuelle situation er, at der ikke er store arbejdsløshedsproblemer i områder med befolkningsmæssig stagnation. På længere sigt er mulighed for at få et job væsentlig, og derfor er spørgsmålet om evt. regionale forskelle i etablering af nye virksomheder et vigtigt emne.

Formålet med denne analyse er at undersøge branchespecifikke og spatiale forhold for nye virksomheder i Danmark. Den overordnede hypotese, der undersøges, er at der med kontrol for brancheforskelle er væsentlige forskelle, og at disse forskelle er specielt markante, når der kigges på 'gode virksomheder'. En 'god virksomhed' er i denne forstand en virksomhed, der dels skaber et væsentligt nettobidrag til samfundet og dels er med til at skabe regional dynamik – de dynamiske forhold, der skabes af teknologiske ændringer, undersøges ved at bruge Pavitts virksomhedstaksonomi. Datagrundlaget er Danmarks Statistiks opgørelse af nye virksomheder i Danmark og analyseenheden er landets 98 primærkommuner. Data på mikroniveau (læs: virksomhedsdata) aggregeres således informationer kan anvendes til de relevante meso-analyser – typisk på delkommune niveau.

2 Teoretiske begrundelser for regionale forskelle i iværksætteri

Nye virksomheder skaber beskæftigelse og vækst, og derfor er ikke alene antallet af nye virksomheder, men også typen af nye virksomheder vigtig for regionen, som virksomheden etableres i. Den samlede teoretiske model skal udover den spatiale dimension også forklare udviklingen i branchestrukturen.

Der findes en række lister over forhold, der giver gode vilkår for start og overlevelse af nye virksomheder, men meget få af disse argumenter er teoretisk funderet. Et typisk eksempel på forhold, der er vigtige for små nystartede virksomheder, bliver undersøgt i verdensbankens 'Enterprise Survey', som afvikles regelmæssigt i perioden 2005-14, se World Bank (2016). Kushnir et al. (2010) bruger dette mikrobaserede survey af virksomheder I 139 lande til at identificere virksomhederne vigtigste hæmsko ('*Percent of firms choosing xxx as their biggest* obstacle') for udvikling af virksomheden. Der er valide svar fra virksomheder i 98 forskellige lande (dog ikke Danmark), herunder både U- og I-lande, og blandt 15 "key obstacles" er de seks vigtigste forhold tydeligt påvirket af at udviklingslande er med i undersøgelserne (rangeret efter betydning)

i. Mangel på elektricitet
ii. Mangel på finansiering
iii. Sædvaner i den uformelle økonomi
iv. Skattetryk
v. Politisk instabilitet
vi. Korruption

Listen af hindringer er følgende, jf. World Bank (2016): "*Access to finance; access to land; business licensing and permits; corruption; courts; crime, theft and disorder; customs and trade regulations; electricity; inadequately educated workforce; labour regulations; political instability; practices of competitors in the informal sector; tax administration; tax rates and transport*".

På trods af at finansieringsforhold er næst vigtigste faktor, så er denne identifikation af globale forhindringer for etablering af nye virksomheder næppe særlig brugbar til forklaring af den regionale eller branchemæssige udvikling i nye danske virksomheder, og den neoklassiske teori giver heller ikke mange svar på hverken antallet af nye virksomheder.

Geroski (1995) identificerer et antal "stylized facts og results", som også stiller spørgsmål ved hvorledes den regionale og branchemæssige struktur ændres

Stylized fact #1:	Start af nye virksomheder (entry) er den vigtigste faktor for brancheudvikling - dobbelt så vigtig som branchepenetration udført af eksisterende virksomheder
Stylized fact #2:	Start af nye virksomheder er vigtig for udvikling af brancherne (within variation), medens cross-section variation ikke er holdbar.
Stylized fact #3:	Start og afgang (entry og exit) af nye virksomheder i brancher er stærkt korrelerede
Stylized fact #4:	Overlevelsesrater for nye virksomheder er meget lave og først efter ca. ti år er nye virksomheder på størrelse med eksisterende virksomheder.
Stylized fact #5:	Start af nye virksomheder er den hyppigste form for entry, men også langt mindre succesfuld end markedspenetration af eksisterende virksomheder
Stylized fact #6:	Der er varierende entry rates over tiden og tilgang ændrer sig over branchens life-cycle.
Stylized fact #7:	Large-scale entry i nye brancher og post-entry penetration i modne brancher er ikke optimalt.
Stylized result #1:	Profit i branchen skaber ikke umiddelbart højere entry rates
Stylized result #2:	Entry barrierer er høje for alle brancher
Stylized result #3:	Hverken profit eller entry barrierer er gode til at forklare start af nye virksomheder
Stylized result #4:	Price-cost margin (profit) påvirkes ikke signifikant af entry
Stylized result #5:	Høj entry rate er stærkt korreleret med højt effektivitet og innovation.
Stylized result #6:	Konkurrence med eksisterende virksomheder i brancher er 'selektiv'.
Stylized result #7:	Priskonkurrence er ikke typisk redskab ved entry.
Stylized result #8:	Virksomhedsstørrelse og alder er væsentligst overlevelses- og vækstfaktor

Konklusionerne i Geroski (1995) er på flere områder nedslående, hvis formålet er at forklare forskelle i regional og branchemæssig udvikling i nye virksomheder. Nye virksomheder starter i alle brancher og regioner; men forklaringen på, hvorfor de opstår, kan ikke findes i de grundlæggende økonomiske rammevilkår – i stedet er det mest sandsynlige udfald, at den nye virksomhed er væk inden for fire år. Dilling-Hansen (2016) viser således baseret på danske iværksætterdata,

- at antallet af nye virksomheder er procyklisk og at niveauet i dag for nye virksomheder stadig ikke er på niveau med tiden før den finansielle krise i 2008
- at nye virksomheder de første fire år af deres levetid har signifikant lavere sandsynlighed for at overleve i forhold til de etablerede virksomheder, og
- at kun godt halvdelen af alle nye virksomheder er tilbage efter de første fire leveår

Omvendt er *Stylized result #5* grundlaget for forklaring af de store brancheforskydninger de seneste tiår: Innovative virksomheder er selve hjørnestenen i økonomisk udvikling, og i stedet for at forklare motiver bag start af nye virksomheder rettes fokus på at identificere de 'vigtige nye virksomheder'. Med andre ord er innovative virksomheder vigtigere for økonomisk udvikling af branche og region end mindre innovative virksomheder.

Pavitt (1984) leverede den nødvendige taksonomi til identifikation af 'vigtige virksomheder' baseret på den traditionelle brancheopdeling, og opdelingen af fremstillingssektoren er senere blevet fuldendt for service sektoren af Miozzo & Soete (2001). Selv om Archibugi (2001) kritiserer taksonomier for at være forsimplende og uden dynamik, så er inddelingen stadig velegnet til at identificere innovative brancher, der kan forklare hvorfor vi har markante skift i den teknologiske udvikling.

Pavitt (1984) og Miozzo & Soete (2001) er grundlaget for en taksonomi af alle virksomheder, som kan anvendes til identifikation af de 'vigtige nye virksomheder', se tabel 1. Anvendelse af Pavitts taksonomi er udbredt ved forklaring af et produkts livscyklus, ved beskrivelse af teknologiske ændringer over tid, til klassifikation af innovationsformer, og teorien bag samspillet/processen mellem de forskellige virksomhedstyper er illustreret i figur 1.

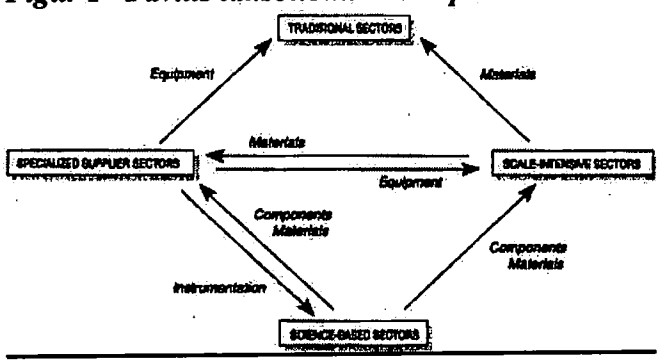
Den teknologiske udvikling i et samfund er ifølge Pavitt (1984) bestemt af en række drivere, som skaber grundlaget for traditionelle produktionsvirksomheder, og disse drivere er virksomheder med særlig fokus på udvikling, innovation, R&D. Figur 1 viser retningen af den dynamiske udvikling, og figuren viser, at det er omfanget af *science-based, scale-intensive* og *specialised equipment firms* som skaber udvikling og vækst i en økonomi. For nye virksomheder er det ensbetydende med, at selv om alle nye virksomheder skaber udvikling og derfor er til fordel for økonomien under ét, så er videntunge virksomheder de vigtigste.

Tabel 1 Pavitts taksonomi for virksomheder

Taksonomi	Pavitt (1984)	Miozzo & Soete (2001)
Supplier-dominated firms	Primære erhverv Tekstilindustrien Service (?)	Trad. service sectors Retail Public services
Science-based firms	High-tech Pharmaceuticals/electronics	Large (innovative) organisations, supermarkets (!)
Scale-intensive firms	Basic materials and durables Consumer goods, automobiles	Network sector Telecom, banking
Specialised equipment firms	Spec. machinery, instruments	Science based - computers

Kilder: Pavitt (1984) samt Miozzo & Soete (2001)

Figur 1 Pavitts taksonomi – samspil mellem virksomhedstyper



Kilde: Pavitt (1984)

For et givet geografisk område betyder det, at udvikling af området afhænger af omfanget af nye virksomheder inden for high-tech området. Det efterfølgende afsnit tester derfor, hvorvidt der er absolutte forskelle i den generelle iværksætteraktivitet og specielt om de "vigtige" brancher også er repræsenteret.

3 Datagrundlag for empiriske analyser

Danmarks Statistik opgør årligt nyetablerede virksomheder fordelt på personligt ejede og selskaber, og analyserne er baseret på Danmarks Statistiks forskerservice ordning, hvor mikrodata for iværksætteri analyseres sammen med oplysninger om regnskabsmæssige og personlige forhold i perioden 2001 til 2013, se Danmarks Statistik (2016). På baggrund af en simpel optælling af nye virksomheder inden for relevante branchegrupperinger og kommunegrupperinger er iværksætteraktiviteten beregnet som andelen af nye virksomheder i forhold til befolkningens størrelse målt i promille. Nedenfor er også vist analyser, hvor aktiviteten er målt i forhold til antallet af personer i den erhvervsaktive alder – resultaterne bliver mere sammenlignelige med internationale opgørelser, hvor iværksætterraten generelt ligger mellem 2 og 6 ‰; men konklusioner er helt upåvirkede af om befolkningen eller de erhvervsaktive bruges som nævner. For 2013 er iværksætterraten hhv. 3.3 og 5.3 ‰ afhængig om befolkning eller erhvervsaktive bruges som nævner, og der er store regionale forskelle i iværksætteraktiviteten; København, Rudersdal og Gentofte kommuner topper listen med et niveau på omkring 5-6 ‰ og Skive, Brønderslev, Struer og Sønderborg kommuner ligger i bunden af listen med et niveau under 2 ‰ (begge tal beregnet i forhold til indbyggertallet).

3.1 Branchestruktur

Tabel viser udviklingen i nye virksomheder fordelt på NACE hovedbrancher. Den procent fordelingen er vist for 2001 og 2013, hvor der i begge år startede omkring 18.500 nye virksomheder.

Tabel 2 Nyetablerede virksomheder i 2001 og 2013 på hovedbrancher

Dansk branchekode 2007, NACE rev. 2	*Beskrivelse*	*Fordeling % - 2001*	*Fordeling % - 2013*
01.00 – 09.99	1. Landbrug og råstofudvinding	0.2	0.1
10.00 – 33.99	2. Fremstillingsindustri	4.0	3.7
41.00 – 43.99	3. Bygge- og anlægsvirksomhed	13.2	12.8
45.00 – 47.99	4. Engros og detailhandel	19.0	13.8
49.00 – 53.99	5. Transport og godshåndtering	6.4	3.8
55.00 – 56.99	6. Overnatning og restaurationer	9.8	7.7
58.00 – 63.99	7. Information & kommunikation	9.2	10.4
64.00 – 66.99	8. Penge- og finansieringsinstitutter	0.8	0.4
68.00 – 68.99	9. Fast ejendom	5.4	4.5
69.00 – 75.99	10. Liberale, vidensk. og tekniske tj.	13.6	17.3
77.00 – 82.99	11. Administrative tj. og hjælpetj.	6.4	9.3
35.00–39.99/84.00-96.99	12. Forsyning og off. Tjenester	11.6	16.1
97.00 – 98.99	13. Medhjælp	0.0	0.0
99.00 – 99.99	14. Organisationer mm.	0.4	0.1
Alle		N=18.391	N=18.510

Noter. Erhvervsstrukturen er baseret på Danske Branchekode 2007, DB07, som er baseret på NACE, rev. 2. Alle registrerede virksomheder i Danmark er tildelt en unikt CVR nummer, jf. https://erhvervsstyrelsen.dk/om-det-centrale-virksomhedsregister-cvr, og det er virksomhedens hovedaktivitet (størst værditilvækst), der bestemmer branche tilhørsforhold.

3.2 Regionale dimension

De 98 kommuner er forskellige på en række dimensioner, og det gælder på en række områder: Geografi, demografi, erhvervsstruktur og økonomi. Der er også en række navne tilknyttet til de enkelte typer af kommuner, men her anvendes den systematiske kategorisering af kommuner baseret på 7 indikatorer:

1) *Urbanisering – befolkning og tæthed*
2) *Center periferi – afstande til typiske arbejdssteder*
3) *Landbrugets betydning – beskæftigelse i primære erhverv*
4) *Udvikling – beskæftigelses- og befolkningsudvikling*
5) *Demografi – andel erhvervsaktive*
6) *Uddannelse – andel med ingen og mindst mellemlange uddannelser*
7) *Økonomi – beskatningsgrundlag*

På baggrund af data fra 2004-07 har Kristensen m.fl. (2006) opdelt kommunerne i fire typer:

a. Bykommuner (35 stk) – typisk københavnske kommuner og større byer (Århus, Odense)
b. Mellemkommuner (17 stk) – ex. Næstved, Horsens, Vejle
c. Landkommuner (30 stk) – ex. Viborg, Herning, Hjørring, Kalundborg
d. Yderkommuner (16 stk) – ex. Lolland, Bornholm, Thisted

Inddelingen af kommuner i disse fire typer af kommuner er særligt velegnet til at teste forskelle i iværksætter aktivitet, fordi områderne har en naturlig spatial afgrænsning således pendling til og/eller etablering af virksomhed uden for regionen ikke er et naturligt alternativ.

4 Empiri – regionale forskelle i iværksætteri

Analyserne er baseret på opgørelse af iværksætteraktiviteten i et område beregnet som antallet af nye virksomheder i forhold til befolkningen i en kommune (målt i promille) og den branchemæssige dimension er her illustreret ved analyser af fremstillingssektoren, engros og detailhandelen samt de udvalgte 'Pavitt-industrier', der på sigt vil give et dynamisk og innovativt erhvervsliv.

Analyserne er baseret på simpel 1-way ANOVA, hvor den afhængige variabel y er de 'i' iværksætterrater i område 'j' (branche, region eller kombinationer af begge):

$$y_{i,j} = \mu_j + \varepsilon_{i,j}$$

Det grundlæggende F-test er relativt robust over for brud på normalitetsantagelsen, men specielt antagelsen om varians homogenitet kan skabe problemer: Hvad gør man med analyser, hvor variansen vokser med gennemsnittet? I det følgende præsenteres Levene's test for varians homogenitet, og selv om testet i flere tilfælde fejler ved stigende middelværdier, så er de grundlæggende test på signifikans ret klare. Tabel 3 viser denne problematik i model 1: Varianserne kan ikke antages ens, og forsøg på at slå kommunerne sammen (model 2) løser ikke problemerne.

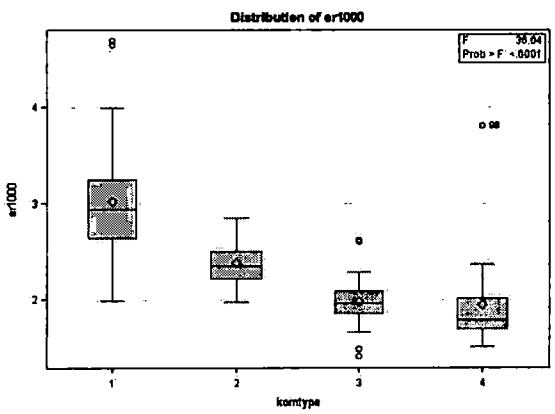
Tabel 3 One-way variansanalyse af iværksætterraten fordelt på kommunetyper, 2013

1-way ANOVA \ Model	1 *Iværksættere pr indbygger*	2 *Kommunetyper (1+2) & (3+4)*	3 *Iværksættere pr 17-64 årig*
$F_{3;94}$ = MSM/MSE	36.4	68.6	29.4
P ($F_{i,j}$ >F*)	<0.0001	<0.0001	<0.0001
R^2	0.54	0.42	0.48
Varianshomogenitet – P	0.03	0.04	0.02
Gennemsnit/(st.dev.)			
Bykommuner (35)	3.76 / (0.76)	3.50 / (0.74)	6.21 / (1.2)
Mellemkommuner (17)	2.96 / (0.24)		4.88 / (0.4)
Landkommuner (30)	2.50 / (0.31)	2.47 / (0.42)	4.20 / (0.6)
Yderkommuner (16)	2.44 / (0.57)		4.23 / (1.2)

Noter. One-way variansanalyse med iværksætterraten som afhængig variable og kommunetype som inddelingskriterium. Test for varianshomogenitet er foretaget ved et Levene's test. Antallet af frihedsgrader for model 2 er hhv. 1 og 96.

Præsentationen af resultaterne i tabel 3, 4 og 5 skal læses fortløbende, og model 1 i tabel 3 viser de basale resultater. Det statistiske test på modellen er baseret på en F-fordelt størrelse med 3 og 94 frihedsgrader, F=MSM/MSE. Modellen er klart statistisk signifikant med den anførte reservation om, at varianserne ikke kan antages ens (signikansniveau på 95%), og tolkningen er lige så klar: Inddelingen i de fire kommunetyper viser markante forskelle i iværksætterraten på tværs af Danmark. Borgere i de større bykommuner har næsten dobbelt så høj tendens til at starte nye virksomheder som borgere i land- og yderkommunerne, og specielt dette resultat er meget robust i de efterfølgende modeller.

Figur 2 Grafisk illustration af forskelle i iværksætteri fordelt på kommunetyper



Model 2 i tabel 3 er baseret på en sammenlægning af kommunetype 1+2 samt 3+4. Figur 2 illustrerer hvorfor denne model ikke er mærkbart bedre end model 1: Variansen inden for kommunetype 2 (mellemkommunerne) er ganske rigtigt mindre, men bykommunerne har både højere middelværdi og varians. Det fremgår også af figur 2, at der i hver kommunegruppe er nogle få markante kommuner med et relativt højt iværksætterniveau – København og Gentofte kommuner er de eneste (by-)kommuner med rater over 4 ‰, og Læsø kommune er den kommune i kategorien 'Yderkommuner', som har det højeste niveau for iværksætteri.

Model 3 i tabel 3 er vist for at illustrere effekten af at anvende 17-64 årige som nævner ved beregning af iværksætteraten. Iværksætterraten vokser, men de fundamentale resultater er fuldstændig i tråd analyserne i dette papir (baseret på normering med befolkningstallet).

Tabel 4 One-way variansanalyse af iværksætterraten – udvalgte brancher

1-way ANOVA \ Model	1 *Fremstilling*	2 *Engros&detail*	3 *Iværksættere med omsætning>100000*
$F_{3,94}$ = MSM/MSE	1.65	14.3	36.0
P ($F_{i,j}$ >F*)	0.18	<0.0001	<0.0001
R^2	0.05	0.32	0.53
Varianshomogenitet - P	0.00	0.54	0.10
Gennemsnit/(st.dev.)			
Bykommuner (35)	0.12 / (0.05)	0.52 / (0.12)	3.02 / (0.6)
Mellemkommuner (17)	0.14 / (0.08)	0.46 / (0.09)	2.38 / (0.2)
Landkommuner (30)	0.12 / (0.05)	0.36 / (0.10)	1.98 / (0.3)
Yderkommuner (16)	0.17 / (0.15)	0.33 / (0.13)	1.94 / (0.5)

Noter. One-way variansanalyse med iværksætterraten som afhængig variable og kommunetype som inddelingskriterium. Test for varianshomogenitet er foretaget ved et Levene test. I tilfælde med få iværksættere i en kommune er denne kommune ikke med i analyserne. Branchen engros & detail indeholder virksomheder baseret på E-handel. 14.754 ud af de 18.510 har en omsætning det første år på over 100.000 kr. Alle modeller fra tabel 3 og tabel 4 (1+2) giver samme resultater for iværksættere med omsætning større end 100.000 kr.

Tabel 4 viser tre interessante forhold. *For det første* (model 1) er det eneste område, hvor der IKKE er statistiske forskelle i iværksætterraten inden for fremstillingssektoren. Der er med andre ord ikke forskel på regionerne, når udviklingen inden for antallet af nye industrivirksomheder betragtes. *For det andet* (model 2) er det generelle mønster med relativt flere nye virksomheder i bykommunerne (og til dels i mellemkommunerne) også at finde inden for engros & detailhandel, og dette resultat er noget i strid med den vanlige opfattelse af, at E-handel og tilsvarende kan startes i alle områder af Danmark. *For det tredje* (model 3) er der heller ikke store forskelle i andelen af 'spekulative' virksomheder uden reel omsætning mellem regionerne; tages der udgangspunkt i de nye virksomheder, der det første år skabte en omsætning på mindst 100.00 kr (ca. 80% af populationen), så er mønstret måske endnu mere klart.

Tabel 5 One-way variansanalyse af iværksætterraten – udvalgte modeller

1-way ANOVA \ Model	1 *Overlevede fra 2001*	2 *Pavitt – snæver definition*	3 *Pavitt – udvidet definition*
$F_{3,94}$ = MSM/MSE	12.8	49.8	32.1
P ($F_{i,j}$ >F*)	<0.0001	<0.0001	<0.0001
R^2	0.29	0.62	0.51
Varianshomogenitet - P	0.001	<0.0001	0.0002
Gennemsnit/(st.dev.)			
Bykommuner (35)	1.11 / (0.28)	0.42 / (0.16)	0.84 / (0.4)
Mellemkommuner (17)	0.93 / (0.16)	0.21 / (0.08)	0.40 / (0.1)
Landkommuner (30)	0.81 / (0.14)	0.15 / (0.06)	0.31 / (0.1)
Yderkommuner (16)	0.84 / (0.20)	0.10 / (0.04)	0.17 / (0.1)

Noter. One-way variansanalyse med iværksætterraten som afhængig variable og kommunetype som inddelingskriterium. Test for varianshomogenitet er foretaget ved et Levene test. I tilfælde med få iværksættere i en kommune er denne kommune ikke med i analyserne. I model 1 er populationen de 5.277 (eller29%) af 18.391 iværksættere fra 2001, der i 2013 stadig var i live. Model 2 (n=1.673) og 3 (n=3.348) er virksomheder, der i Pavitt definition er 'technologi change'-skabende – NACE hovedbrancherne er 20, 21, 26, 28, 29, 30, 51, 61, 62, 63, 72 (snæver) samt 70 & 71 (udvidet definition).

Med udgangspunkt i den teoretiske begrundelse for at kigge med særlig interesse på de videntunge, de specialiserede, de forskningstunge, de effektive virksomheder - alle er kendetegnet ved at være underleverandører til de traditionelle sektorer - er der i tabel 5 vist 2 modeller (model 2 og 3), der tydeligt viser at de mest interessante virksomheder hvad angår fremtidig udvikling også skabes i de større bysamfund. Model 1 i tabel 5 er igen signifikant med højere aktivitet i byerne, men for yderkommunerne er der måske lidt trøst at hente her: Model 1 er virksomheder, der er startet i 2001 og som stadig er aktive i 2013 (ca 29%), og her ses det at selv om modellen stadig er signifikant, så er det dog klart virksomhederne i yderkommunerne, der har den største chance for at overleve.

5 Afrunding

De præsenterede analyser er baseret på nye virksomheder opgjort af Danmarks Statistik og resultaterne er meget markante og bekymrende for fremtidig regional udvikling: Det er i de større bysamfund, det relativt største antal nye virksomheder skabes, og specielt er virksomheder, der normalt anses for at være ekstra vigtige for fremtidig udvikling, i endnu større grad et storbyfænomen. Måske skal vi være mindre bekymrede for fraflytningen og i stedet fokusere på at skabe flere virksomheder gennem det lange seje træk ... det siges jo, at "jyden han æ stærk å sej"!

Referencer

Archibugi, Daniele (2001) Pavitt'S Taxonomy Sixteen Years On: A Review Article, *Economics of Innovation and New Technology*, 10:5, 415-425

Castellacci, Fulvio (2001) Pavitt's Taxonomy Sixteen Years On: A Review Article, *Economics, Innovation, New Technology, vol. 10*, pp. 415-425

Castellacci, Fulvio (2008) Technological paradigms, regimes and trajectories: Manufacturing and service industries in a new taxonomy of sectoral patterns of innovation, *Research Policy 37*, 978-994

Danmarks Statistik (2016), Dansk Branchenomenklatur – dansk branchekode, Danmarks Statistik,http://www.dst.dk/da/Statistik/dokumentation/Nomenklaturer/DB

Danmarks Statistik (2016), *Iværksættere i Danmark – times dokumentation*, Danmarks Statistik, http://www.dst.dk/da/Statistik/dokumentation/Times/undervisningsministeriets-statistikberedskab/ivaerksaetter

Danmarks Statistik (2016), Notat om nye brancher inden for detailhandel via internet, Danmarks Statistik, http://www.dst.dk/ext/erhvervreg/nybrint--pdf

Dilling-Hansen, M. (2016), SMEs: Peter Pan Syndrome or firms not grown up? Creativity, business skills and economic growth of Danish entrepreneurial firms. *(Forthcoming) Athens Journal of Business and Economics, December 2016.*

Geroski, P. (1995), What do we know about entry?, *International Journal of Industrial Organisation*, 13, pp. 421-440.

Kristensen, I. T., Kjeldsen, C. og Dalgaard, T. (2006). *Landdistriktskommuner - indikatorer for landdistrikt*. Danmarks Jordbrugsforskning, Afdeling for Jordbrugsproduktion og Miljø, Tjele.

Kushnir K, Mirmulstein ML, and Ramalho R (2010) *Micro, Small, and Medium Enterprises around the World: How Many Are There, and What Affects the Count?* Washington, DC: World Bank and International Finance Corporation (IFC).
Lipczynski, J., J. Wilson & J. Goddard (2013), *Industrial Organisation*, 4th ed., Pearson.

Miozzo, M. and L. Soete (2001), Internationalization of Services: A Technological Perspective, *Technological Forecasting and Social Change*, Vol. 67, Issues 2–3, 7 June 2001, pp. 159–185

Pavitt, K. (1984), Sectoral Patterns of Technical Change: Towards a Taxonomy and a Theory, *Research Policy, 13(6),* pp. 343-373, 1984.

World Bank (2016), *Enterprise Survey*, The World Bank, 19-Dec-2016, http://data.worldbank.org/data-catalog/enterprise-surveys

Regional Development and the Role of Innovation

Andreas P. Cornett & Nils Karl Sørensen*, University of Southern Denmark*
E-mail: nks@sam.sdu.dk and Cornett@sam.sdu.dk

Abstract:

The purpose of the current paper is to analyze the impact of regional potentials as measured by the innovation performance on the per capita level of income. A very simple model is set up where the level of income is determined by a wide selection of innovation indicators described by data from the regional innovation scoreboard (RIS) published by the European Commission. The statistical material contains 12 normalized scoreboard indicators covering every second year for the period 2007 to 2013. Data are divided by 23 European Union (EU) members where a regionalization is possible. The result is a panel of data with 732 observations.

The setup is applied on several different sub data sets related to the presence of super clusters, presence of cities, and the EU classification of regions into four categories of innovation drivers. Further, a classification relative to the use of EU structural funds is considered and related to endowments by regions.

Overall, the results points toward a quite diversified picture, but in general the population with tertiary education serves as a good indicator along with the variables public-private co-publications and the 'NON R&D Innovation' expenditure i.e. investments in equipment and machinery and the acquisition of patents.

JEL Classification: R11, R12, R58

1. Introduction

Innovation is seen as a major driver of wealth in the European space. The priority and goal of the structural policy undertaken by the European Union is to stimulate and create an atmosphere that stimulates innovation. Innovation performance should then lead to an increase in economic wealth as measured by the income per capita.

However, many nations are quite diversified with regard to the regional distribution of the income. Cornett & Sørensen (2008) showed that the polarization of economic activities could lead to excess growth in some regions, and this could lead to convergence as well as divergence among the regions of Europe. Further, Cornett & Sørensen (2014) showed that the period up to the economic recession in 2008 was characterized by convergence where the period after the recession has been characterized by divergence among many regions. It was found that regions across nations increasingly shared some similar characteristics. These patterns were leading to the appearance of "regional clubs" across nations. Finally, Cornett &

* Department of Business and Economics, Alsion 2, DK-6400 Sønderborg, Denmark. Phone +45 6550 1211 or 1229.

Sørensen (2012a, 2012b) provided a relation the role of innovation on growth based in the Innometrics (2006, 2011) statistical framework. Here the role of cities and some types of clusters were considered. It was found that the possibilities for economic development were closely related to the factor endowment in a given region.

2. Determining the Level of Income by use of Innovation Performance Statistics

In order to analyze the impact of innovation performance as a driver of the Gross Domestic Product (GDP) a very simple approach is adopted. For a region i the GDP can be stated as:

$$GDP_i = f(Z_i) \qquad i = 1, \ldots, n$$

Here n is the total number of regions. Z is a matrix of innovation performance drivers, and it is further assumed that $f' > 0$ so a given innovative performance has a positive impact on GDP. This could be interpreted as a simple production function approach.

Data, variables and Countries considered

The framework is applied on a data set compiled from the *Regional Innovation Scoreboard* (RIS) published by the European Union; see European Commission (2012, 2014a). The present data set reports statistics for the following European Union members primarily collected at the NUTS two level: Belgium, Bulgaria, the Czech Republic, Denmark, Germany, Ireland, Greece, Spain, France, Italy, Hungary, the Netherlands, Austria, Poland, Portugal, Rumania, Slovenia, Slovakia, Finland, Sweden, the United Kingdom and Croatia. In addition, statistics can be found for Norway. In total there are 23 countries with a total of 183 regions. Statistics has been compiled for the years 2007, 2009, 2011 and 2013[1]. The impact of the years has been analyzed by use of dummy variable for the years 2009, 2011 and 2013. On this basis a data panel of 732 observations can be established. Further, data on the GDP by region has been compiled from Eurostat[2].

[1] Future work will also consider the updates for 2015 published in the summer of 2016. Future work will also intend to combine the Innometrics Statistics and the RIS statistics into a single consistent database covering detailed by region the innovation performance from the millennium and forward. This long time period will enable us to consider the issue of regional convergence in a set up robust relative to the recession from 2008 to 2012.

[2] Cornett & Sørensen (2012a) used a data set on innovation performance based on Innometrics (2006). They used innovation statistics from 2004 to analyze the influence of innovation on convergence. The present analysis is a little different because the economies have been in recession for most of the period. As demonstrated in

For the analysis, the GDP per capita is used. The Regional Innovation Scoreboard (RIS) variables are normalized shares and ranges between 0 and 1. Data are divided into a total of a total of 13 "growth drivers". These variables constitute the description of the Z matrix. The variables are summarized in Table 1. A criticism to this approach is that the variables are selected rather ad hoc.

Table 1: Regional Innovation Scoreboard Variables

Variable	***Measurement and rationale for inclusion***
POP with tertiary education	Population with tertiary Education. Number of persons in age class with some form of post-secondary education age group 25–64. *Included as a general indicator of the labour supply of advanced skills.*
PUB R&D expenditure	R&D expenditure in the public and higher education sector. *Induced as an overall driver of technological development.*
BUS R&D expenditure	R&D expenditure in the business sector. *Included as an indicator of formal creation of new knowledge within firms.*
Non R&D innovation expenditure	Non R&D innovation expenditure. *This variable measures investment in equipment and machinery and the acquisition of patents and licenses.*
SME's innovative in house	*Measures the degree to which the SME's has introduced any new or significantly improved products or production processes, have innovated in-house.*
Innovative SME's collaboration	Innovative SME's collaboration with others. *Measures the degree to which SME's are involved in innovation co-operation. Complex innovations often depend on the ability to draw on diverse sources of information and knowledge, or to collaborate on the development of innovation.*
Pub-private co-publications[1]	Number of public-private co-publications (PPC's) This indicator captures the public-private research linkages and active collaboration between the public and the private sector
EPO patents	EPO patens applications. *The capacity of firms to develop new products will determine their competitive advantage. This is approximated by number of patent applications at the European Patent Office.*
Technological innovator	SME's introducing product or process innovations. *New products are the key to innovation in manufacturing activities. Higher shares of technological innovators should reflect a higher level of innovation activities.*
Marketing innovator	SME's introducing marketing or organizational innovations. *Many firms, in particular in the service sectors, innovate through non-technological forms of innovation.*
Employ in med or high tech	Employment in knowledge intensive activities. *The sectors of medicine and high technology use frequently a highly educated and trained workforce.*
Sales new markets	Sales to new to market, and new to firm innovations *Measures the turnover of new or significantly improved products.*

Note: 1) Not available for 2013. **Source:** Own complication from European Commission (2012, 2014b).

Cornett & Sørensen (2014) the rate of convergence by region decreased from 2.63 percent before the recession to 1.60 percent in the period 2008 to 2012.

Table 2: Innovation Drivers of GDP by Regions, Clusters and Cities

	Full data set			Super Clusters			City Dummies			Super Clusters and City Dummies		
	Coefficients	*Error*	*Sig*	*Coefficients*	*Error*	*Sig*	*Coefficients*	*Error*	*Sig*	*Coefficients*	*Error*	*Sig*
Intercept	2507	1733		9078	3461	***	-5610	3798		8101	4922	
POP with tertiary Education	23817	2155	***	19138	2993	***	31347	4293	***	21433	4891	***
PUB R&D expenditure	982	2089		11019	3575	***	7289	4490		25796	5402	***
BUS R&D expenditure	9071	2535	***	16717	4056	***	7695	5714		3568	7419	
NON R&D Innovation exp	16764	2527	***	24622	5030	***	8888	5050	*	22257	6705	***
SME innovating in house	13980	3023	***	12891	4023	***	13106	5305	**	8104	5751	
Innovative SME's collaboration	511	2203		6020	3237	*	4695	4566		5057	5374	
Pub-private co-publications	22012	2638	***	38189	5644	***	31200	6373	***	50430	8596	***
EPO patents	2879	2969		5976	4817		743	6121		21140	8664	**
Technological innovator	2461	3501		6746	4868		1296	6559		11350	7399	
Marketing innovator	9300	2869	***	19523	4495	***	13458	5271	**	27642	6149	***
Employ in med or high tech	9712	2064	***	12695	3556	***	10701	4145	**	15742	5725	***
Sales new markets	4662	2020	***	277	3319		4388	4078		3609	5204	
Year dummy 2009	-2504	802	***	-4180	1283	***	-2951	1524	*	-4850	1828	***
Year dummy 2011	2	808		-81	1269		-735	1549		65	1911	
Year dummy 2013	-1489	1001		-3779	1715	**	-1180	1906		-5366	2577	**
Multiple R		0.83			0.87			0.87			0.90	
R Square		0.68			0.76			0.75			0.80	
Adjusted R Square		0.68			0.74			0.74			0.77	
Standard Error		7542			5851			7727			6324	
Observations		732			188			216			104	
Total of Regions		183			47			54			26	

*Note: *** is strong significance at the 1 % level, ** is significance at the 5 % level, * is weak significance at the 10 % level.*
Source: Calculations based on GDP statistics from EUROSTAT and European Innovation Scoreboard, various volumes.

Estimation Results on the significance of Super Clusters and Cities

Table 2 reports results of four simple OLS regressions on a non–transformed linear model. So no logarithmic transformations etc. have been undertaken[1]. The results focus on the impact of innovation of super clusters and city dummies respectively. The classification with regard to super clusters follows the classification given in Center for Strategy and Competitiveness, CSC (2011) whereas the classification into cities follows the classification in Cornett & Sørensen (2012a).

All innovation scoreboard variables are positive as expected. In general, the power of the models as expressed by the adjusted R^2 and the standard error are satisfactory.

In all cases the variable population with tertiary educations is positively significant and large. So a qualified labor supply is of importance for the generating a high GDP. This is especially tur in the cities. The coefficient of the regression with the super clusters is smaller than for the overall regression. The variables NON R&D innovation expenditure, Public-private co-publications, SME's introducing marketing or organizational innovations and employment in medicine and high technological firms are all also significant for in all the regressions. Finally, the variable NON R&D Innovation expenditure is significant across all categories. These three variables are then the general contributors of innovation behavior to the formation of GDP.

Turning to the process of research and development (R&D), then different patterns are observed. Public R&D is having impact on GDP in the super clusters and also super clusters and large cities. The observed findings are consistent with the analysis undertaken in Cornett & Sørensen (2012a). Here it was also found that the presence of EPO patents has strong influence on economic growth and the process of convergence. This was especially true for the regions with high income. In the present analysis this variable does not perform very well in order to explain the level of GDP. An explanation of the different results could be either the definition of the variables or presence of multicollinearity.

With regard to time effects then the time dummy for year 2009 is negatively significant and for the super cluster and including the cities this is also true for year 2013. The time effects are in general weaker than expected taking the strong nature of the world-wide slowdown of the economy since 2008 into account.

[1] The model has been estimated based on a logarithmic transformed data. However, this did not improve the estimations. The coefficients can in such a case be interpreted as the elasticities or changes of the impact of the scoreboard indicators. The estimated results are available on request to the authors.

Estimation Results on the significance of Innovation Levels

In order to dig deeper into the regional pattern of innovation and growth reported in Figure 3 in the previous section Table 3 provides in depth insight into some of the main drivers. The regions divided into the four classifications given namely the *Innovation Leaders*, *Innovation Followers*, *Moderate Innovators* and the *Modest Innovators*. The division into groups has been obtained from European Commission (2014b) and Nordregio (2016). This classification is an update of the classification given in Cornett & Sørensen (2012a). Figure 1 gives an illustration of the geographical location of the regions in Europe at NUTS 2 level by innovation location. It is observed that the regions with the highest innovation classification also is the regions with the highest income per capita.

Figure 1: Innovation Scoreboard by Regions in Europe 2014.

Source: European Commission (2014b) page 4.

The results reported in Table 3 shows that the share of the population with tertiary education is the variable that is of general significance regardless of the level of innovation. The highest values are

obtained with regard to the innovation leaders and for the Modest Innovators. These two groups are rather different. Also the Pub-private co-publications are significant across all groups, and take the highest for the Moderate and Modest Innovators. The public and private expenditure on R&D follows a rather diversified pattern. For the innovation leaders and the modest innovators the public expenditure on R&D is significant whereas the business expenditure on R&D is significant for the innovation followers and moderate innovators. The explanation of this pattern could be due to differences in development policy. As in the first set of regressions results, the EPO patents perform poor. The performance of the regions can also be split up into the period considered. The growth rate in the Regional Innovation Scoreboard can be calculated. For 2007 to 2013 the growth rate for the Innovation Leaders can be calculated to equal 1.3 percent; for the Innovation Followers it equals 3.9 percent; for the Moderate Innovators it equal 1.8 percent, and finally for the Modest Innovators the growth rate equals -2.2 percent.

This is an interesting result the negative growth rate of the Modest Innovators underlines the increase in the diversity of the regions in Europe. This issue was also addressed with regard to the process of convergence in Cornett & Sørensen (2014).

Innovation Performance and Location

The final issue to be addressed in this section is to analyze how innovation performance is related to the location of a given region. Here it is frequently observed that the more wealthy and central regions move away from the other regions see for example Cornett & Sørensen (2008). One of the results is that the economic crisis has reinforced not only intraregional divergence within countries but also the traditional divide between the stronger Northwest European countries and the South and East of Europe. It is evident that the two variables are related: Strong factor endowments are found in the urban located regions whereas the weak factor endowments are found in the peripheral regions.

Table 3: Innovation Drivers of GDP by Regions, Leaders and Followers

	Innovation Leaders			Innovation Followers			Moderate Innovators			Modest Innovators		
	Coefficients	*Error*	*Sig*	*Coefficients*	*Error*	*Sig*	*Coefficients*	*Error*	*Sig*	*Coefficients*	*Error*	*Sig*
Intercept	5597	6995		-258	5422		-935	3646		-6094	3704	
POP with tertiary Education	42232	8048	***	15320	3490	***	24837	3987	***	34360	6293	***
PUB R&D expenditure	18666	4821	***	2625	3173		-11	4033		11849	6482	*
BUS R&D expenditure	8489	5283		11649	4393	***	17038	5142	***	737	7720	
NON R&D Innovation exp	17538	9125	*	6172	6716		17590	4351	***	12951	4657	***
SME innovating in house	10404	6157	*	12050	4807	**	13758	7337	*	8037	12133	
Innovative SME's collaboration	10844	6173	*	8663	3744	**	4826	4405		1167	7210	
Pub-private co-publications	11451	6108	*	11816	4653	**	27018	5130	***	23398	6602	***
EPO patents	8916	6533		6350	7037		34521	6501	***	8775	9312	
Technological innovator	265	7058		2386	5391		4235	7809		34425	13885	**
Marketing innovator	12841	6947	*	8253	5114		7892	5145		520	7839	
Employment in med or high tech	19027	6456	***	19961	4502	***	2803	3320		3780	5835	
Sales new markets	2880	6039		7134	4522		9128	3079	***	739	4816	
Year dummy 2009	-6621	1874	***	-3752	1377	***	-409	1332		-973	1898	
Year dummy 2011	-2780	2005		-860	1429		517	1331		-1434	1940	
Year dummy 2013	-3981	3141		-209	1926		2810	1775		-1408	2454	
Multiple R		0.81			0.64			0.75			0.84	
R Square		0.65			0.41			0.56			0.70	
Adjusted R Square		0.59			0.36			0.54			0.67	
Standard Error		5923			6666			7139			8132	
Observations		104			204			260			160	
Total of Regions		26			51			65			40	

*Note: *** is strong significance at the 1 % level, ** is significance at the 5 % level, * is weak significance at the 10 % level.*
Source: Calculations based on GDP statistics from EUROSTAT and European Innovation Scoreboard, various volumes.

Table 4: Innovation Drivers of GDP by Regional Urbanization and Development

	FP: Strong absorbers - cities			FP: Strong absorbers - no cities			SF: low users - cities			SF: low users - no cities		
	Coefficients	*Error*	*Sig*	*Coefficients*	*Error*	*Sig*	*Coefficients*	*Error*	*Sig*	*Coefficients*	*Error*	*Sig*
Intercept	12825	9024		10543	5198	**	-6457	3283	*	8892	2154	***
POP with tertiary Education	18852	7882	**	2598	4545		10158	4230	**	12355	2864	***
PUB R&D expenditure	12844	8042		7035	3028	**	556	4457		1627	2938	
BUS R&D expenditure	8391	9633		3212	5445		18722	5324	***	14065	3385	***
NON R&D Innovation exp	13578	9878		25399	5963	***	5106	4183		21945	3224	***
SME innovating in house	14278	7506	**	19713	6298	***	4567	4737		21314	3640	***
Innovative SME collaboration	8039	7779		6884	5127		4020	3946		1430	2434	
Pub-private co-publications	33561	10646	***	2408	4908		26084	6448	***	8189	3312	**
EPO patents	1205	13557		8769	7375		19662	5217	***	10204	4497	**
Technological innovator	18014	10127	*	2664	9189		27	5417		2305	4004	
Marketing innovator	21430	8920	**	15829	7378	**	21331	4546	***	9914	3523	***
Employ in med or high tech	5504	10354		8693	5262		13260	3094	***	8571	2769	***
Sales new markets	7664	6785		1620	6954		4060	3634		356	2332	
Year dummy 2009	3371	2579		-3383	1637	**	-2335	1144	**	-1814	950	*
Year dummy 2011	561	2727		2272	1813		775	1155		1133	956	
Year dummy 2013	742	3687		-262	2129		2069	1505		-1864	1201	
Multiple R		0.72			0.84			0.94			0.84	
R Square		0.52			0.71			0.89			0.71	
Adjusted R Square		0.40			0.61			0.87			0.70	
Standard Error		7625			3926			3897			5501	
Observations		76			60			104			288	
Total of Regions		19			15			26			72	

*Note: Sig is significance. *** is strong significance at the 1 % level, ** is significance at the 5 % level, * is weak significance at the 10 % level.*
Source: Calculations based on GDP statistics from EUROSTAT and European Innovation Scoreboard, various volumes.

Table 4 reports the findings for the resulting OLS regressions on the regional GDP per capita. The two most interesting cases relative to the evidence put forward in Table 5 is found in the first and the last regression. For the FG strong absorbers – cities much fewer variables are significant than for the SF low users – no cities.

Table 5: Classification of regions according to use/absorption of funding and urbanization

	Urban located region	Peripheral located region	Total
"Strong factor endowment" High regional potential NEG / (FG Strong absorber)	19	15	34
"Weak factor endowment" Low regional potential NEG / (SF low users)	26	72	98
Total	45	87	132

In both cases the population share with tertiary education, SME's innovating in house, public-private co-publications and marketing innovator variables are significant. For the SF low users the BUS R&D expenditure, NON R&D Innovation expenditure, EPO patents and employment in medicine and high tech industries are also significant. Consequently, the contribution from innovation to GDP is much diversified for the SF low users – no cites. Relative to funding this implies more open options and possibilities.

To address this issue the regions has been divided on hand by the use/absorption of EU funding and on the other hand by location as expressed by the city classification. The resulting cross-tabulation is found in Table 5. Notice that the total of regions does not sum up to 183 because not all regions have been considered for funding.

3. Conclusion and Perspectives

The purpose of the current project has been to contribute to the understanding of the drivers behind regional economic development and income creation. A central issue is to estimate the impact of regional potentials on this process and to contribute to the development of instruments to cope with the problem, to overcome path dependency in lagging regions. The main focus in this contribution is on the innovations as drivers of GDP growth and income generation.

Overall, the results points toward a quite diversified picture, but in general the population with tertiary education serves as a good indicator along with the variables public-private co-publications and the 'NON R&D Innovation' expenditure i.e. investments in equipment and machinery and the acquisition of patents.

These variables stress the importance of three drivers for the generation of income namely the presence of a well-educated labor force; the dynamics of the corporation between the private and the public sector, and finally the importance of the right equipment and the acquisition of patents. With regard to the type of R&D expenditure – business versus public, the picture is quite diversified, and no overall pattern can be observed. Contrary to what was found in Cornett & Sørensen (2012a) the presence of EPO patents has very limited influence on the level of GDP. An explanation of the this result could be that the EPO variable referees to new patents whereas the acquisition of the existing patents in the well-established process of production embodied in the variable NON R&D innovation expenditure is of higher importance. Overall these results point toward a conclusion that a path dependency characterized by negative downward spiral in lagging regions will extremely difficult to overcome since the most important drivers identified are in increasingly short supply in these regions.

The material presented allows a classification into four categories innovation performers namely the Innovation Leaders, Innovation Followers, Moderate Innovators and Modest Innovators. Comparing the Innovation Leaders with the Modest Innovators it is evident that the number of significant variables is much higher for the former than for the latter. Consequently, the Modest Innovators has fewer innovation parameters. The innovation growth for the Modest Innovators has also been negative over the considered period. This is in line with the decrease in GDP for the same group of regions studies in Cornett & Sørensen (2014). The best innovation performance over the period is observed for the group of Innovative Followers. With an innovation growth rate equal to 3.9 percent this group is catching-up on the group of Innovation Leaders that experienced a growth rate equal to 1.2 percent.

Regions with strong factor endowments are primarily urban whereas regions with weak factor endowments are primarily peripheral. For the strong endowment regions located in the cities there are fewer innovation drivers than with regard to the peripheral regions with weak factor endowments.

Our preliminary results points at least toward three areas to extend our analysis: first of all the application of the NEG concept could be extended, at least with the inclusion of entrepreneurial skills, see figure 1 as a second activator of regional potentials. Secondly the concept of regional factor endowments and potential needs specification and clarification how potentials can be

activated. Last but not least a detailed specification and understanding of the linkages between economic growth, path-dependency and different categories of regions is needed.

References

Center for Strategy and Competitiveness, CSC (2011): "Strong Clusters in Innovative Regions". Stockholm School of Economics. European Commission Entreprise and Industry.

Cornett, A.P. & N. K. Sørensen (2014): "Regional GDP Convergence in the European Regions in the light of the Economic Recession". In: Peter Linde (editor): Symposium i anvendt statistik 2014. Københavns Universitet, Danmarks Statistik, page 111–117.

Cornett, A.P. & N. K. Sørensen (2012a): "Innovation and regional disparities – a survey of regional growth drivers and economic performance" pp. 81-108 in Charlie Karlsson et al. (2012), *Innovation, Technology and Knowledge*, Routledge , Abingdon (UK) & New York (NY).

Cornett, A.P. & N. K. Sørensen (2012b): "Determinants of Convergence and Disparities in Europe: Innovation, Entrepreneurship and the Processes of Clustering". Paper for the special session ' SS- Knowledge, Innovation and regional Growth', 52th Congress of the European Regional Science Association, "Regions in Motion – Breaking the Path" Bratislava, Slovakia August 21 – 25. (Published in ERSA - proceedings of 52th Congress of the European Regional Science Association, Bratislava Slovakia).

Cornett, A.P., & Sørensen, N.K. (2008): "International vs. Intra-national Convergence in Europe – an Assessment of Causes and Evidence", pp. 35-56 in *Investigaciones Regionales* No. 13, Otoño 2008, Asociación Española de Ciencia Regional, Madrid.

Danmarks Statistik, Regional accounts, København.

ESPON (2013): "SGPTD Second Tier Cities and Territorial Development in Europe: Performance. Policies and Prospects" Applied Research 2013/1/11, Final Report Version 30/06/2012.

European Commission (2014a):"Investment for jobs and growth Promoting development and good governance in EU regions and cities", Sixth report on economic, social and territorial cohesion Brüssel 2014.

European Commission (2014b): "Regional Innovation Scoreboard 2014" Luxembourg 2014.

European Commission (2012): "Regional Innovation Scoreboard 2012" Luxembourg 2012.

Eurostat (2016): "Regio Data" Luxembourg. (www.europa.eu).

Eurostat (2012): "Focus on territorial typologies" pp.192-201 in *Eurostat regional yearbook 2012*, Luxembourg.

Innometrics (2011): "European Innovation Scoreboard 2011", Maastricht Economic Research Institute on Innovation and Technology (MERIT) and the Joint Research Centre (Institute for the Protection and Security of the Citizen of the European Commission).

Innometrics (2006): "European Innovation Scoreboard 2006", Maastricht Economic Research Institute on Innovation and Technology (MERIT) and the Joint Research Centre (Institute for the Protection and Security of the Citizen of the European Commission).

Nordregio (2016): "State of the Nordic Region 2016", *Nordregio Report 2016:1* Stockholm, Sweden.

Is obesity epidemic?

Jørgen T. Lauridsen, COHERE, IVØ, SDU, jtl@sam.sdu.dk.

Summary

Alike most countries, Denmark experiences an increasing percentage of obese citizens of all ages. Although obesity is not in a clinical sense infectious, this development has commonly been termed "the obesity epidemic", based on an assumption of a spread of lifestyle behaviour underlying obesity. However, most studies focuses on the epidemic pattern over time as measured by changes in obesity rates. Some studies focus on space, but merely in terms of geographical patterns of exogenous factors underlying obesity; only few consider endogenous spillover without an explicit motivation. The present study adds by considering endogenous spatial spillover of obesity rates as an expression of a spatial epidemic dynamic caused by learning effects. Obesity rates and explanatory variables for a spatial panel of 98 Danish municipalities, observed 2010 and 2013, is exerted to spatially adjusted regression models. A spatial behavioural spillover of an endogenous learning nature is demonstrated, thus indicating a spatial epidemic dynamic behind obesity rates.

Keywords: Spatial econometrics, Obesity, Health Behaviour

JEL classifications: C21, I12, I14

1. Introduction

During the later half of the 20th century, most of the Western world experienced an increase in obesity. It is estimated that, by 2015, more than 700 million people globally would be obese, a problem of epidemic proportions given the role that obesity plays in the development and progression of many chronic diseases, including type 2 diabetes, various types of cancer and heart disease (WHO 2000; Birmingham et al., 2003; Choudhary et al., 2007; Low et al. 2009). According to WHO, about 13 percent of the World adult population (11 percent of men and 15 percent of women) were obese in 2014 (WHO 2015). For Denmark, the figure is slightly higher, namely 15 percent (Ministry of The Interior and Health 2006), and it is even higher for other countries. Thus, for US, the figure is more than one third of the population (Guettabi and Munasib 2014), and it has been estimated that, by 2010, up to one-fifth of the US healthcare expenditures will be allocated to treatment of obesity related diseases (Lakdawalla et al. 2005).

Obesity is of major concern for the individual as well as society. From an individual perspective, obesity involves a substantial loss of life quality as well as healthy life years (Kushner and Foster 2000). From a societal perspective, obesity involves healthcare costs as well as loss of income (McCormick et al. 2007). Thus, there is an

intense focus on evaluation, prevention and intervention directed toward obesity and its consequences (Leeman et al. 2012).

The organisation of the present paper is as follows. Section 2 will be devoted to a discussion of the determinants and geography of obesity. Next, statistical methods are presented in Section 3. Further, Section 3 will briefly motivate and outline the data just, while the results are presented in Section 4. Finally, the study will be concluded in Section 5.

2. Determinants and geography of obesity

From an economic perspective, individuals are assumed to behave rational thus acting optimally in any given situation in order to maximize their personal utility. This holds for obesity as well: A lifestyle implying a critically low weight causes disutility in terms of weakened immune system, fragile bones and tiredness (Brown 2000). Similarly, a lifestyle implying a critically high weight, and eventually obesity, causes disutility in terms of loss of life quality and healthy life years (Pengpid & Peltzer 2014). Thus, everything else kept equal, an optimum weight should be found in between these two extrema.

However, everything is not equal. The individual may be surrounded by a variety of impediments stimulating an unhealthy lifestyle, eventually causing obesity. While traditional efforts to reduce obesity have focused on individual behaviour, there has recently been a shift toward understanding the obesity from an ecological perspective (Egger & Swinburn 1997) with focus on features such as the built environment and cultural and socio-economic influences, which contribute to an 'obesogenic environment' (Swinburn et al., 1999).

The implications of space have been considered in different ways. Mapping and spatial analytical techniques have been used to explore regional variations and spatial clustering of obesity rates investigated geographical variation in overweight and obesity between urban and rural communities, and found a tendency for communities to cluster based on the incidence of overweight and obesity.

As suggested by the ecological perspective (Egger & Swinburn 1997, Swinburn et al. 1999), one consequence of "obesogenic environment" may be that obesity becomes acceptable, potentially implying the presence of learning as a driving force behind obesity. If such learning is present, then it should manifest itself into spatial clustering as suggested by the spatial studies (Pouliou & Elliott, 2009, Abayomi et al., 2009, Lebel et al. 2009, Penney et al. 2013). However, while a spatial spillover of obesity is readily implied, few studies considered this feature. Chen and Wen (2010) applied a multilevel spatial model to two Taiwanese surveys from 2001 and 2005 and provided evidence of endogenous spillover for the last year. Chen et al. (2014) used a geo-referenced cross section of individuals from the city of Indianapolis in Indiana, US to

estimate a spatial model for BMI and found significant indication of endogenous spatial spillover. The present study contributes by hypothesizing and investigating the potential presence of spatial learning spillover within the framework of a spatial endogenous spillover specification, which links the data from two Danish surveys from 2010 and 2013.

Specifically, a regression setup will be applied, including exogenous explanatory drivers of obesity, enhanced with an endogenous spatial spillover and adjustment for correlation across surveys (see section 3 for details). According to suggestions from recent literature, a variety of exogenous variables will be included. Variation in obesity rates across central and peripheral areas may be ambiguous. On one hand, there are indications of higher rates in peripheral areas (Djernæs and Jensen 2013), while they may also be lower due to the less access to unhealthy food (Emond et al. 2011). A similar central-peripheral ambiguity may hold when considering effects of urbanisation and population density (WHO 2010). Next, the age distribution may play a role, as the rate of obesity is increasing with age (Stunkard 1983). Ethnic minorities are known for having less healthy lifestyle, including higher obesity rates (Caprio et al. 2008). Education plays an important role for lifestyle. In particular, persons without education have been pointed to as having less healthy lifestyle and thus a higher propensity for overweight (Devaux et al. 2011). Finally, socially marginalised persons are known for having less healthy lifestyle and thus being more exposed for obesity (Apolloni et al. 2011).

3. Methodology

The point of departure is a linear regression model defined for the N=98 Danish municipalities in a single year by

$$(1) \qquad y_t = X_t\beta + \upsilon_t, \quad \upsilon_t \sim N(0,\sigma^2 I)$$

where X_t is an N by K dimensional matrix of the K explanatory variables, y_t an N dimensional vector of the obesity rates in the municipalities, and β a K dimensional coefficient vector measuring the effects of the explanatory variables on obesity rate. The term υ_t is a residual term, which represents the fertility rates when controlled for the explanatory factors of X_t and may be denoted the residual obesity rate.

Operationally, endogenous (learning) spatial spillover is controlled for by adding the average of y_t in the neighbourhood municipalities (denoted by y_t^W) as an explanatory variable in (1) to obtain the *spatially autoregressive* (SAR) specification (Anselin, 1988)

$$(2) \qquad y_t = y_t^W\lambda + X_t\beta + \upsilon_t,$$

where λ is a parameter specifying the magnitude of spill-over, formally restricted to the interval between (-1) and (+1), but for most practical purposes restricted to be positive.

Alternatively, any kind of spatial clustering, including observed as well as unobserved exogenous spatial spillover, may be controlled for by applying the spatially autocorrelated (SAC) specification (Anselin 1988)

(3) $y_t = X_t\beta + \varepsilon_t,\ \varepsilon_t = \lambda W\varepsilon_t + \upsilon_t.$

The SAC approach will be applied to investigate whether spatial spillover of obesity rates is merely ascribed to exogenous structures rather than being of an endogenous learning nature.

One further methodological problem needs attention for the SAR as well as the SAC specification. While pooled data for T=2 years are applied, the residual obesity rates across years for any municipality are correlated. Also, the variance of the residual obesity rate within each year may potentially vary across years. Thus, between any two years, the covariance of the residual obesity rate reads as

(4) $E(\upsilon_t'\upsilon_s) = \sigma_{ts}^2 \quad t,s = 1,..,T.$

To obtain efficient estimates of β, we apply Feasible Generalised Least Squares (F-GLS) estimation as suggested by Zellner (1962) to obtain Seemingly Unrelated Regression (SUR) estimates for β. By integrating (4) into any of (1) through (3), SUR, SAR-SUR and SAC-SUR specifications are obtained.

4. Data

Table 1. Data applied for the study

Variable	Definition	Mean	SD
Obese[1]	Percentage of population with BMI 30 or over	14.9	2.88
Peripheral[2]	Peripheral municipality	0.15	0.36
Age 24-64[3]	Percentage of population aged 24-64 years	51.8	1.65
Age 65-[3]	Percentage of population 65 years and over	18.8	3.39
Immigrant[3]	Immigrants from non-Western countries per 10,000 inhabitants	265.9	154.6
No education[3]	Percentage of population without professional education	22.5	5.29
Population density[2]	Population per square kilometre	165.0	209.8
Urbanity[3]	Percentage of population living in urban areas	83.1	12.9
Tax base[3]	Income deductible for municipal taxation per inhabitant (1,000 DKK)[2]	165.1	31.5
Service level[3]	Municipal service level	1.00	0.04

Source: [1] The Danish Health Profile (www.danskernessundhed.dk), [2] The Danish Ministry of Taxation (www.skm.dk), and [3] the Key Figure Base (www.im.dk)

The data to be applied are defined in Table 1, which further shows the means by year of the variables. Data were obtained for 98 Danish municipalities in the years 2010 and 2013.

As dependent variable, Obesity specifies the percentage of population, which are obese, i.e., have BMI above 30. The rate is calculated from two independent surveys of 2010 (92,990 respondents) and 2013 (83,990 respondents), which were stratified across the Danish municipalities. The surveys were collected as part of The Danish National Health Profiles 2010 and 2013 by The National Institute of Public Health at University of Southern Denmark.

Obesity rates for the 98 Danish municipalities for 2013 are shown in Figure 1.

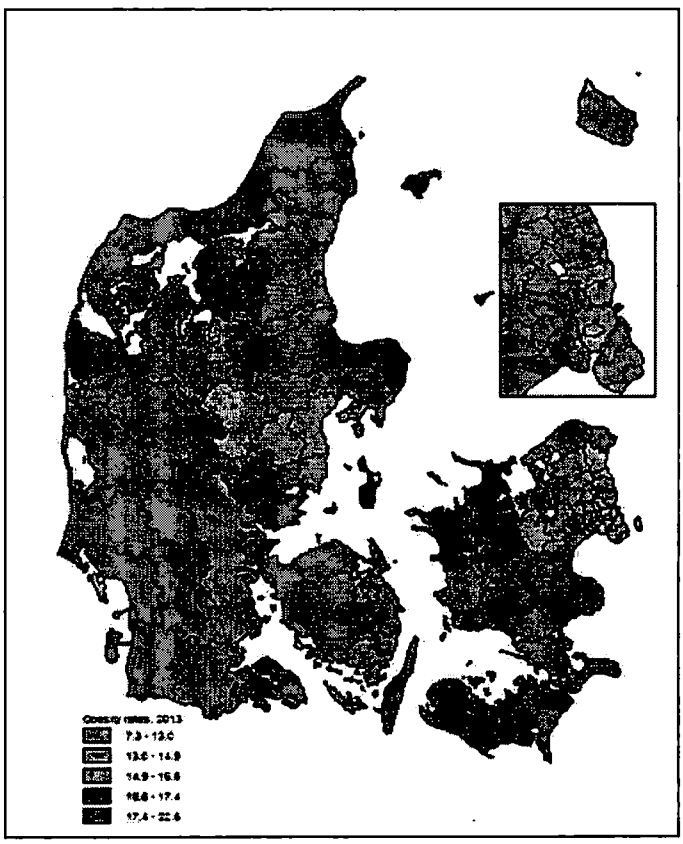


Figure 1. Obesity rates (%) 2013.

As suggested by earlier studies, explanatory variables are made op of nine variables. First, an indicator for the municipality being peripheral is applied. This definition was set by The Ministry of Taxation, which defines a municipality to be peripheral if 1) labour income per inhabitant in 2001-2003 was at most 90 percent of the national level, and 2) the population growth 2000-2005 was less than 50 percent of the national level. The effect of age distribution is captured by two variables measuring the percentages of population in the age groups 24 to 64 years old and 65 years and above. Next, ethnicity was captured as number of immigrants from non-Western countries. Effects of agglomeration and urbanity were measured by two variables, population density and percentage of population living in urban areas. Income level is measured by the municipal tax base. Finally, as a proxy for social marginalisation in the municipality, the municipal service level, i.e., the municipal operating costs per inhabitant divided by social needs, is applied. This figure, which was defined by The Ministry of the Interior, calculates social needs as a weighted sum of several social indicators.

5. Results

Table 2 shows the spatially unadjusted SUR model, the SAC-SUR model specifying spatial spillover to be of an exogenous nature, and the SAR-SUR incorporating endogenous (learning) spatial spillover.

Table 2. SUR, SAC-SUR and SAR-SUR models for obesity rates

Variable	SUR	SAC-SUR	SAR-SUR
Constant	9.64 (6.89)	10.56 (7.25)	4.99 (6.89)
Peripheral	-0.51 (0.39)	-0.49 (0.36)	-0.50 (0.37)
Age 24-64	-0.06 (0.10)	-0.04 (0.11)	-0.04 (0.10)
Age 65-	0.05 (0.06)	0.07 (0.06)	0.04 (0.06)
Immigrant	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
No education	0.32 (0.04)***	0.31 (0.04)***	0.31 (0.04)***
Population density	-0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)
Urbanity	-0.05 (0.02)***	-0.05 (0.02)***	-0.05 (0.02)***
Tax base	-0.02 (0.01)**	-0.02 (0.01)**	-0.01 (0.01)
Service level	6.85 (3.18)**	4.03 (3.16)	6.52 (3.09)**
Year=2013	1.33 (0.21)***	1.29 (0.26)***	1.11 (0.23)***
Spatial lag (λ)	-	0.28 (0.19)	0.18 (0.07)***
LogL	-146.9	-143.3	-143.5
AIC	323.9	316.7	317.0

Note. Figures are regression coefficients with standard errors in parentheses. Significanse indicated by * (10%), ** (5%), *** (1%).

Regarding effects of explanatory variables, the three models are by and large in agreement. High proportions of population without professional education are systematically related to high obesity rates. Furthermore, a high degree of urbanity is connected to low obesity rates. Next, high income levels are systematically related to lower obesity rates. Also, high service levels, which are assumed to approximate the percentage of population with social burdens, are connected with high obesity rates. Next, a time trend is observed, as the obesity rates on average increased with a little more than one percent from 2010 to 1013. Apparently, peripheral areas are not different with respect to obesity rates, and age structure and ethnicity seem unrelated to obesity rates too. Finally, while significant exogenous spatial spillover cannot be reported from the SAC-SUR specification, the significant spillover parameter of the SAR-SUR seems to indicate endogenous learning spillover.

6. Conclusions

The present study adds to existing knowledge by explicitly implementing the spatial dynamics of obesity rates. It is shown to be less probable that underlying clustering of exogenous factors should affect obesity rates. Rather, evidence is provided that a spatial behavioural spillover of an endogenous learning nature is present, thus indicating a spatial epidemic dynamic of obesity rates. Furthermore, the results indicate that obesity rates in Denmark, which has otherwise been increasing in the pass, actually drops slightly from 2010 to 2013, thus indicating that the obesity epidemics may have culminated. Thus, an important policy recommendation would be to consider the small-area spatial learning nature behind obesity, as this pattern may provide a key to the further reduction of obesity rates.

References

Abayomi JC, Watkinson H, Boothby J, Topping J., Hackett AF. 2009. Identification of hot-spots of obesity and being underweight in early pregnancy in Liverpool. Journal of Human *Nutrition* and Dietetics 22, 246–254.

Anselin L. 1988. *Spatial Econometrics: Methods and Models*. Kluwer Academics, Dordrecht

Apolloni A, Marathe A, Pan Z. 2011. A Longitudinal View of the Relationship Between Social Marginalization and Obesity. In Salerno J, Yang SJ, Nau D, Chai S-K (eds.). *Social Computing, Behavioral-Cultural Modeling and Prediction*, pp. 61-68. Heidelberg: Springer.

Birmingham CL, Jones P, Hoffer LJ. 2003. The management of adult obesity. *Eating and Weight Disorders* 8, 157–163.

Brown WJ, Mishra G, Kenardy J, Dobson A. 2000. Relationships between body mass index and well-being in young Australian women. *International Journal of Obesity* 24: 1360-1368.

Caprio S, Daniels SR, Drewnowski A, Kaufman FR, Palinkas LA,_Rosenbloom AL, Schwimmer JB. 2008. Influence of Race, Ethnicity, and Culture on Childhood Obesity: Implications for Prevention and Treatment. *Diabetes Care* 31: 2211-2221.

Chen D-R, Wen, T-H. 2010. Elucidating the changing socio-spatial dynamics of neighborhood effects on adult obesity risk in Taiwan from 2001 to 2005. *Health & Place* 16: 1248-1258.

Chen SE, Florax EJ, Snyder SD. 2014. Obesity and fast food in urban markets: a new approach using geo-referenced micro data. Health Economics 22: 835-856.

Choudhary AK, Donnelly LF, Racadio JM, Strife JL. 2007. Diseases associated with childhood obesity. *AJR American Journal of Roentgenology* 188: 1118–1130.

Devaux M, Sassi F, Church J, Cecchini M, Borgonovi F. 2011. Exploring the Relationship Between Education and Obesity *OECD Journal: Economic Studies* 2011/1: 121-159.

Djernæs SV, Jensen M. 2013. *Flov over din krop? Konstruktionen af den overvægtige person i reklamer og kampagnen Vægtstop*. Speciale. Roskilde: Roskilde Universitet.

Egger, G. & Swinburn, B. (1997) An 'ecological' approach to the obesity pandemic. *British Medical Journal* 315: 477–480.

Emond JA, Madanat HN, Ayala GX. 2011. *Access to healthy and unhealthy food items in urban grocery stores: How store audit data can describe a food environment considering customer ethnicity.* Conference Paper, 139st APHA Annual Meeting and Exposition

Guettabi M, Munasib A. 2014. "Space Obesity": The Effect of Remoteness on County Obesity. *Growth and Change*, 4: 518–548.

Kushner RF, Foster, GD. 2000. Obesity and quality of life. Nutrition 16: 947-952.

Lakdawalla DN, Goldman D, Shang B. 2005. The health and cost consequences of obesity among the future elderly. *Health Affairs* 24(Suppl. 2): W5R30–W5R41.

Lebel, A., Pampalon, R., Hamel, D. & Theriault, M. (2009) The geography of overweight in Quebec: a multilevel perspective. *Canadian Journal of Public Health* 100: 18–23.

Leeman J, Sommers J, Vu M, Jernigan J, Payne G, Thompson D, Heiser C, Farris R, Ammerman A. 2012. An Evaluation Framework for Obesity Prevention Policy Interventions. Preventing Chronic Disease 9: 110322 (9 pages).

Low S, Chin MC, Deurenberg-Yap M. 2009. Review on epidemic of obesity. *Annals of the Academy of Medicine, Singapore* 38: 57– 65.

McCormick B, Stone I, Corporate Analytical Team. 2007. Economic costs of obesity and the case for government intervention. Obesity Reviews 8: 161-164.

Pengpid S, Peltzer K. 2014. Prevalence of overweight/obesity and central obesity and its associated factors among a sample of university students in India. Obesity Research & Clinical Practice 8: e558–e570.

Pouliou T, Elliott SJ. 2009. An exploratory spatial analysis of overweight and obesity in Canada. *Preventive Medicine* 48: 362–367.

Stunkard A. 1983. Human Ageing and Obesity. In Platt D (ed.). *Geriatrics 2*, pp. 436-445. Heidelberg: Springer.

Swinburn B, Egger G, Raza F. 1999. Dissecting obesogenic environments: the development and application of a framework for identifying and prioritizing environmental interventions for obesity. *Preventive Medicine* 29: 563–570.

WHO. 2000. *Obesity: preventing and managing the global epidemic.* Report of a WHO consultation. World Health Organization Tech. Rep Ser. 894, i–xii; 1–253.

WHO. 2010. Urbanization and health. *Bulletin of the World Health Organization*, 88, 241-320.

WHO. 2015. Obesity and Overweight. *Fact Sheet*, N°311. New York: WHO.

Zellner A. 1962. An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests of Aggregation Bias. *Journal of the American Statistical Association* 58: 977-992.

The impact of user charges on the demand for sterilisations.[1]

Christian Kronborg, e-mail: cka@sam.sdu.dk, phone +45 65 50 30 85
Jørgen T. Lauridsen, e-mail: jtl@sam.sdu.dk, phone +45 65 50 21 42

University of Southern Denmark
Department of Business and Economics
Centre of Health Economics Research (COHERE)
Campusvej 55, 5230 Odense M, Denmark

Abstract
In 2011, there were user fees on a variety of treatments in public hospitals, including sterilization. The aim of this study was to explore the impact of user charges in a health care system with universal coverage where general practitioners serve as gate-keepers to specialised health care. For this purpose we used changes in the national health act that introduced and later repealed user charges for specific procedures at public hospitals.

We used register data from the National Registry of Patients and Health Insurance Service Registry. Data on all individuals 25 years of age or older who had undergone sterilisation in the years 2009 to 2012 was the extracted. Data were analyzed using the interrupted time-series analysis.

A total of 5747 women and 19,889 men were included in the analysis. We find a significant reduction in the number of sterilisation treatments at public hospitals in 2011 compared to previous years. In 2012 when the user charges were abolished the number of sterilisations increased, but not to the same level as before the introduction of the user charge. Furthermore, the user charges shifted demand for male sterilisations to private hospitals.

Introduction

User charges are a controversial source for financing of health care services as they influence equity and efficiency (Thomson et al., 2010). It reduces access to health care for poor individuals who may be unable to pay for the services they need. On the other hand, user charges may improve efficiency, because individuals will abandon services of low value, provided, however, they have the information needed to make the right decision, understand it and are able to act rationally based on it.

[1] This is on going work. A previous version in Danish is available at www.sdu.dk: Christian Kronborg, *Brugerbetaling for sterilisation på offentlige sygehuse. Resultater fra et naturligt eksperiment.* COHERE Discussion Papers 2015:3.

It can be argued that user charges in health care systems, e.g. in the UK and Denmark, where citizens register on a list with a general practitioner, who then acts as a gatekeeper to specialised health care result in an efficient use of health care services. Consequently, user charges would be less needed in such systems on efficiency arguments. However, this requires that the gatekeeping general practitioner has the information, including information about the cost of providing a specialised health care service, and that the general practitioner uses this information when referring a patient to e.g. a specialist or hospital department. It is conceivable that the general practitioner will not use the cost information if the cost has no consequence for the patient.

In addition, a user charge may also give information to the provider about how much the patients are willing to pay for the service and whether or not the patients actually value the service. That is, generally, rational consumers will not purchase a product they do not value at least as high as the price.

The aim of this study was to explore the impact of user charges in a health care system with universal coverage where general practitioners serve as gatekeepers to specialised health care. For this purpose we used changes in the national health act that introduced and later repealed user charges for specific procedures at public hospitals.

We exploit two policy changes with regard to user charges for sterilisations at publicly owned hospitals in Denmark. From 1st January to 31st December 2011 patients that underwent a procedure for sterilisation at a public hospital in Denmark had to come up with user charge. Before and after 2011, the procedure was provided free of charge. The same policy changes applied to sterilisation procedures that were carried out at an office-based specialist clinic with a contract with the National Health Insurance except that user charges were not introduced until 1st March 2011. That is, from this date to 31st December 2011 the National Health Insurance did not reimburse sterilisations at private practicing surgeon clinics.

Furthermore, we use the opportunity offered by a rich set of administrative register data on health care utilization at public and private providers in Denmark. The administrative registers cover the entire population of Denmark, and both public and private hospitals are required to provide information about all procedures they carry out on patients to the register. However, physicians and clinics that contract with the National Health Insurance are exempted from this requirement. That may cause some limitations on the interpretation of the results. Consequently, we focus on the response to user charge for sterilisation procedures that are carried out in a hospital setting because information about sterilisations carried out at office-based surgeon clinics having a contract with the National Health Insurance was unavailable.

According to Danish legislation any person aged 25 years or older can be sterilised without permission.[2] However, persons who are mentally ill or in a condition so they do not understand the implication of the procedure must have a permission from a regional council that is appointed by the minister of health. Furthermore, persons below the age of 25 years can be sterilised only after consultation with the regional council if special circumstances apply, e.g. heredity genes that may cause severe illness in prospective children that is considered prudent to prevent or the applicant for sterilisation is considered to be an unfit parent (The Danish Health Act §§105-114).

Male sterilisation by vasectomy is a small operation that is usually done as a day procedure under local anaesthesia (Natarajan and Oakley, 2008). The procedure can be carried out in a physician's office, clinic or a hospital. Normally, the procedure does not require hospitalisation. Female sterilisation is a more invasive procedure in comparison, and it is usually done under general anaesthesia.

The results show that the introduction of the user charges at public hospitals in 2011 caused a pronounced reduction in the frequency of sterilisations in both men and women compared to previous years. Furthermore, the abolition of the user charge in 2012 caused the number of sterilisations to increase compared to 2011, but to a significantly lower level than that in 2009 and 2010. Another finding from the study is that the user charge for sterilisation at public hospitals caused a shift to mainly male sterilisations performed in private hospitals.

Institutional setting

The Danish health care system provides universal coverage to all residents in Denmark. It is mainly financed through taxes, although approximately 17% of the total health care expenditures come from patient out-of-pocket payments, primarily for prescription medicines and adult dental care (Pedersen et al., 2012). Administration of the health care system is decentralised. Five regions own and run their public hospitals and they are responsible for planning and financing office-based health care services such as general practice and medical specialists (e.g. orthopaedic surgeons and gynaecologists). Public hospitals employ medical doctors and specialists, whereas general practitioners (GPs) and office-based medical specialists are self-employed and own their own clinics. GPs and the office-based medical specialists contract with the regional health authorities. The contracts cover reimbursable services and the fee schedules.

The GPs are gatekeepers to the specialised health care services such as elective surgery at public hospitals and consultations with office-based specialists. That is, a referral from a GP allows the patient to receive specialist health care services without out-of-pocket payment at a public hospital or office-based medical specialists who have contract with the regional health authority.

[2] Since September 1, 2014, the age limit has been 18 years.

Most residents (98%) in Denmark are listed with a GP (Pedersen et al., 2012),which entitles them to consult a GP and to have most GP services free of charge. Balance billing applies to patients that are not listed with a GP. Residents in Denmark are free to choose a GP to be listed with. However, usually they choose a GP clinic close to their homes.

A number of private hospitals and private clinics operate in Denmark. The majority of their services are paid for by private health insurance or by the patients themselves. Furthermore, the regional health authorities contract with private hospitals to deliver certain procedures for residents in their region.

Policy changes

From January 1 to December 31, 2011, user charges for in-vitro fertilisation, refertilisation, and sterilisation applied at public hospitals. Furthermore, sterilisations at office-based private practising physicians who contracted with the National Health Insurance were no longer subject to reimbursement from 1th March 2011.

The user charges were implemented to relieve public budgets from the costs of health care. In the spring 2010, the government decided to introduce the user charges. This led to a political process in the parliament where the government introduced the bill in the parliament and the bill was discussed in a parliamentary committee. Furthermore, during the process in the parliament there was a public hearing about the bill. The bill for the introduction of user charges was passed in December 2010.

In August 2011, the prime minister called for a parliamentary election. This led to a change in government. The new government was a coalition of three parties that had voted against the user charges back in 2010. In the November 2011 it began a parliamentary process to abandon the user charges, and in December 2010, the parliament passed the bill to abandon these.

The user charge for treatment at public hospitals was set to equal the diagnosis related group (DRG) charge. In 2011, this was 8,457 DKK for male and 12,984 DKK for female sterilisation, respectively (Indenrigs- og Sundhedsministeriet, 2010).However, from 15th April 2011 the user charge was changed to 900 DKK if the sterilisation was carried out as part of another procedure (Indenrigs- og Sundhedsministeriet, 2011).

For comparison, based on information from private hospitals that provide male sterilisation, the price for this procedure ranged between 3,300 and 12,000 DKK in 2009-2010 and in 2012 whereas it ranged between 3,300 and 7,000 DKK in 2011. The reimbursement for male sterilisation with an office-based medical specialist was 2,292 DKK in 2012.

Methods

Data

For this study, we used data from the National Patient Register (NPR) and the National Health Service Register (NHSR) on male and female sterilisations that were carried out in 2009-2012 (Lynge et al., 2011, Andersen et al., 2011). The NPR includes information on all discharges including outpatient visits of individual patients from all Danish hospitals, whereas the NHSR includes registrations of treatments from office-based private practising medical specialists with a contract with the National Health Insurance.

In the NPR, each record identifies the patient with a unique personal identification number and information about the hospital department, diagnoses, classification of surgical procedures, and date and time of treatment and operation. From the register, we extracted the records on all patients aged 25 years old or older who underwent a procedure with the codes KKFD46 (male sterilisation) or KLGA (female sterilisation), where the contact was classified as an encounter for sterilisation (International Classification of Diseases diagnosis code Z30.2). We categorized hospitals according to public or private ownership based on the hospitals' identification codes, which we linked to the hospital classification that is applied for reporting to the NPR. Records on sterilisations because of medical reasons and sterilisation after permission from the regional council in pursuance of §106 or §107 in the Danish Health Act were excluded, since these procedures were exempted from user charges.

The NHSR includes registration of all invoices that health care providers have sent to the regional health authority in order be reimbursed for their services (Andersen et al., 2011). Each record includes the patient's personal identification number together with information about the type of service (e.g. a consultation or a procedure) and the reimbursement (fee-for-service) that the health care provider received from the regional health authority. We extracted all records where the invoice included reimbursement for a male sterilisation procedure.

A limitation with the NPSR is that it does not include information about the exact date and time of the delivery of the service. Rather it indicates the year and the week number when the provider sent the invoice to the regional health authority. Consequently, we assumed that the health care service was provided to the patient in the same week as the invoice was sent to the region. In addition, we aggregated the week numbers to calendar months. For that purpose the first day in the week decided which month the week number was allocated to. This may have caused some misclassification of the timing of the delivery of these health care services. For example if the first day of the week was the 31th of August, all invoices – and consequently the associated health care services – in that week were considered to have taken place in August even though that some of the services may actually have taken place in September.

In cases where a male was registered with a sterilisation procedure both in the NPR and the NHSR, we used the information from the NPR.

Based on these data extractions, we counted the number of sterilisations by gender and hospital ownership (public or private). To ensure variation for the statistical analyses (see next section), these counts were done on a weekly base for 208 weeks spanning from primo 2009 to ultimo 2012.

Statistical analysis

An interrupted time series (ITS) analysis (Wagner et al., 2002, O'Keeffe et al., 2014, Kontopantelis et al., 2015) was used to assess the impact of the policy changes where user charges were introduced in 2011 and withdrawn in 2012. Specifically, we examined the frequency of female or male sterilisations in public or private sector per week, assuming the following form:

$$Y_t = \alpha_0 + \alpha_1(P_t^{(1)} + P_t^{(2)}) + \alpha_2 P_t^{(2)} + \beta_0 T_t + \beta_1(T_t P_t^{(1)} + T_t P_t^{(2)}) + \beta_2(T_t P_t^{(2)}) + \gamma_1 EXP_t^{(1)} + \gamma_2 EXP_t^{(2)} + \gamma_3 PAY_t + \sum_{j=1}^{12} \delta_j M_t^{(j)} + \epsilon_t$$

where Y_t is the aggregated number of sterilisation in week t, $t = 1, .., 208$, and $T_t P_t^{(i)}$ an interaction term defined as the product of T_t and $P_t^{(i)}$, $i = 1,2$.

Table 1 presents the variables used for the subsequent analysis.

Thus, for any week prior to the intervention, the average number of sterilisations was α_0, with a weekly increase of β_0. For any week during the intervention period, the same figures were $(\alpha_0 + \alpha_1)$ and $(\beta_0 + \beta_1)$, while they were $(\alpha_0 + \alpha_1 + \alpha_2)$ and $(\beta_0 + \beta_1 + \beta_2)$ for the period after the intervention. Thus, α_0 and β_0 measures the level and slope throughout the entire period, while α_1 and β_1 measures the level and slope changes from introducing the intervention, and α_2 and β_2 the level and slope effects from rolling back the intervention. These effects were adjusted for seasonal variation by including the monthly indicators $M_t^{(1)}$-$M_t^{(12)}$ (estimated with the restriction that $\sum_{j=1}^{12} \delta_j = 0$), for level shifts caused by expectations of introduction and removal by including $EXP_t^{(1)}$ and $EXP_t^{(2)}$, and for the level shift caused by reimbursement by including PAY_t .

Finally, as initial Durbin Watson tests based on simple linear regression estimates of the ITS model (not reported here) indicated autocorrelation, it was decided to estimate it with an adjustment for a first order autocorrelation process in the residuals.

Table 1 Variables included in the regression model

STERIL	Weekly number of sterilisations, divided by male / female and public / private
$P^{(1)}$	Indicator for intervention period: 1 for 2011, 0 otherwise
$P^{(2)}$	Indicator for period after intervention: 1 for 2012, 0therwise
T	Time trend on a weekly base (T = 1, .., 208)
$EXP^{(1)}$	Indicator for period when intervention discussed (1 for May to December 2010, 0 otherwise)
$EXP^{(1)}$	Indicator for part of intervention period when new government in power (1 for October to December 2011, 0 otherwise)
PAY	Indicator for period when reimbursement of private fee in play (1 for January and February 2011, 0 otherwise)
$M^{(j)}$	Indicator for month j, j = 1, .. , 12

Results

Data on sterilisation procedures of 9,159 females and 12,084 males were extracted from the National Patient Register. A total of 76 observations of female sterilisation and 29 observations of male sterilisation were excluded because the patients were less than 25 years of age and therefore did not comply with the inclusion criteria. Furthermore, 3,331 observations of female sterilisation and 355 of male sterilisation were excluded because the contact was not classified as an encounter for sterilisation. Five observations of female sterilisations and 11 of male sterilisations were excluded because the procedure was carried out due to medical reasons.

From the National Health Service Register a total of 8,223 observations of male sterilisations were extracted. Five of the observations were excluded because the patients were less then 25 years of age, and 18 observations were excluded because the patients were also observed with a sterilisation in the NPR.

Figure 1 show the number of persons who had a female (part A) or male (part B) sterilisation in a hospital setting by year, month, and hospital ownership, respectively. It also shows that the introduction of the user charge for sterilisations at public hospitals caused a shift in whether the procedure was carried out at a public or private hospital. This shift was particularly pronounced for male sterilisations. In 2011, 53% of all male sterilisations were performed at private hospitals, whereas less than 1% of all sterilisations were performed at private hospitals in 2009-2010. In 2012, 2% of all male sterilisations were performed at private hospitals. Additionally, Figure 3 shows the number of persons who had a male sterilisation in an office-based private practising surgery clinic with a contract with the National Health Insurance.

Figure 1 Number of persons that underwent an operation for female sterilisation in a hospital setting by year, month and hospital ownership 2009-2012.

A) Females **B) Male**

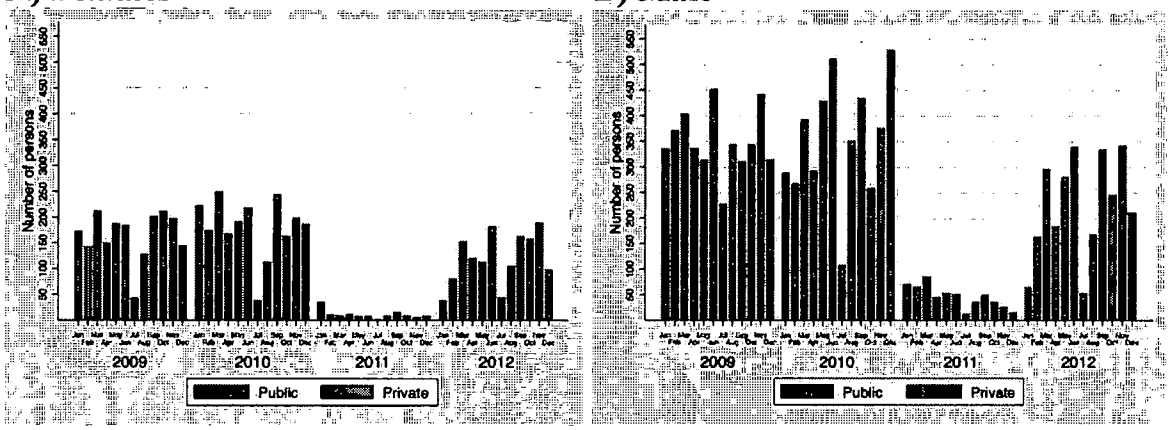


Figure 2 Number of persons that underwent a male sterilisation procedure at an office-based private practising clinic contracting with the National Health Insurance by year and month.

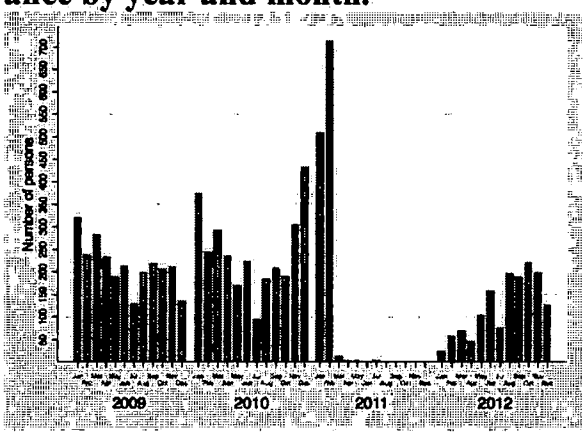


Table 2 shows the estimated ITS models.Turning first to the estimation results for public provider treatments for females, the average number of sterilisations without the intervention were 31.4. The intervention reduced this number with 38.7, i.e., the expected number of sterilisations was practically equal to zero during the intervention period. Thus, the intervention substantially and significantly reduced the demand for sterilisation from public providers. Following the rollback of the intervention, the number of sterilisations increased with 20.5 per week. Thus, although the demand partly recovered after the intervention, it did not fully return to the level before the intervention. Furthermore, without the intervention, the average weekly increase in the number of sterilisations was 0.90. This figure was significant at the 5 percent significance level. The changes in slope caused by the intervention and its rollback are not significant and will thus not be interpreted. Neither the expectation variables nor the payment variable were statistically significant from zero.

Table 2 Interrupted Times Series (ITS) models, estimated with autocorrelated residuals

	Public hospitals		Private hospitals	
	Female	Male	Female	Male
Intercept	31.4 (7.51)***	84.6 (9.16)***	0.02 (0.32)	0.31 (0.61)
$P^{(1)}$	-38.7 (-5.01)***	-85.2 (-4.83)***	-0.11 (-0.96)	7.52 (7.85)***
$P^{(2)}$	20.5 (2.49)**	40.2 (2.16)**	0.09 (0.78)	-6.11 (-6.03)***
T	0.90 (2.12)**	-0.74 (-0.78)	-0.003 (-0.43)	-0.02 (-0.30)
T* $P^{(1)}$	1.08 (0.68)	1.99 (0.54)	0.07 (2.88)***	-0.32 (-1.60)
T* $P^{(2)}$	0.30 (0.19)	0.61 (0.16)	-0.07 (-2.78)***	0.23 (1.14)
$EXP^{(1)}$	4.34 (1.03)	2.83 (0.31)	0.004 (0.07)	0.05 (0.10)
$EXP^{(1)}$	19.16 (1.40)	-5.40 (-0.33)	0.11 (1.06)	-2.87 (-3.21)***
PAY	10.61 (1.40)	21.40 (1.20)	0.32 (2.83)***	-4.36 (-4.56)***
Month indicators included	yes	yes	yes	yes
DW (simple model)	1.32	1.88	2.16	1.66
DW (AC model)	1.75	1.95	2.00	1.98
R^2 (AC model)	0.57	0.52	0.25	0.62

For public provider treatments for males, the average weekly number of sterilisations without the intervention was 84.6. Alike for females, this demand was practically reduced to zero by the intervention, during which the average weekly demand dropped with 85.2 sterilisations. Again, the demand partly recovered after rollback of the intervention. For males, no significant time trend patterns were found. None of the expectation variables or the payment variable was significant.

Regarding private provider treatments for females, the demand for sterilisations is very limited. The weekly number of sterilisations without the intervention was not significantly different from zero, and neither were the effects of the intervention and its rollback. The weekly time trend without the intervention was neither different from zero. Apparently, a significant increase in this trend was caused by the intervention, followed by a similar reduction when rolling back the intervention. Thus, the user charge seemed to shift females gradually toward public providers, and its rollback shifted them gradually away again. None of the expectation variables were significant. However, as expected, reimbursement of the payment had a stimulating effect on demand, as the weekly number of sterilisations increased with 0.3 sterilisations.

Finally, for private provider treatments for males, the average demand was 0.3 sterilisations per week before the intervention. During the intervention, this number rose to 7.5 sterilisations per week, after which it dropped with 6.1 following the rollback of the intervention. Thus, the user charge seemed to shift males towards public providers, and its rollback shifted them away again. The time trends as well as the changes in the time trend caused by the intervention and its rollback were not significantly different from zero. Thus, opposed to females, where the shift towards (and away from) private providers happened gradually, the shift occurred instantaneously for males. This may indicate that females behaved relatively more adaptive in their reaction to the intervention than males, who reacted immediately. Opposed to what was the case for females, however, the opportunity of reimbursement of payment had a negative effect on males, as their demand dropped with 4.4 sterilisations per week. Finally, for males, the entrance of a new government, who was expected to roll back the user payment, had the expected effect, as it led to an instantaneous drop in demand of 2.9 sterilisations per week.

For all models, the Durbin-Watson statistics are close to two and thus indicative of absence of higher order autocorrelation. Thus, the present adjustment for first order autocorrelated residuals seems appropriate. Finally, the R^2 values indicate reasonable model fits for demand for public sterilisations ($R^2 = 0.57$ for females and $R^2 = 0.52$ for males) and for demand for private sterilisations by males ($R^2 = 0.62$), while the model fit is somewhat lower for demand for private sterilisations by females ($R^2 = 0.25$).

Discussion

The results show that the introduction of the user charges for sterilisation at public hospitals in 2011 reduced the frequency of sterilisations in both men and women compared to previous years, and that the abolition of the user charge in 2012 caused the number of sterilisations to increase but to a significantly lower level than that in 2009 and 2010. Furthermore, the user charge shifted demand of male sterilisation to private hospitals.

It is an important strength of this study that it relies on data from administrative registers that are made available for research through Statistics Denmark (Andersen et al., 2011). It is the same registers that are used for the official statistics of the activity in the Danish health care system. However, the study also has some limitations. First of all, the user charge for sterilisation was a temporary policy change although this may not have been realised at the time of its introduction. However, at least from the autumn 2011 it may have been anticipated that the user charge would be abandoned within a relative short time horizon. Therefore, foresighted patient may have postponed their decision to undergo the procedure. The results from the study suggest that foresightedness may have been present. In particular the finding of the considerable increase in the number of male sterilisation in office-based private practising surgery clinic in January and February 2011 may represent such considerations of the patients

or physicians, e.g. general practitioners who refer the patients to the surgery. Consequently, the temporary nature of the policy change may cause threats to the external validity of the study since individuals and institution may not have fully adapted to the new situation (Meyer, 1995)

Another limitation of the study is that the policy change applied to all residents in Denmark. Thus, it was not possible to identify a group of individuals that were not influenced by the policy who were comparable to those who were. Thus, a complete experimental design with an intervention group and a control group was not available. Nevertheless, we exploited the quasi-experimental nature the policy changes introduced and used interrupted time series (ITS) models to analyse the impact of the user charges. ITS analysis is a widely used in studies with observational data where complete randomisation in not possible (Kontopantelis et al., 2015). Furthermore, the ITS analysis makes full use of the longitudinal nature of the data and makes it possible to estimate causal effects of an intervention (Kontopantelis et al., 2015). The ITS analysis requires that the trend is linear, and that there are no external time-varying effects or autocorrelation (ibid.).

From a conceptual perspective sterilisation procedures differ from other types of health care services, since such a procedure does not improve health. This may cause some problems in generalising the results from this study to other types of health care. Within the health economics literature the demand for health care services is considered as a derived demand for health. That is, the purpose of consuming health care is to improve health when an individual experiences deteriorations in health because of illness. Consumption of health care may also include an investment in health in order to prevent deterioration in health in the future. However, a sterilisation is not a health care service in the sense that it is to restore an individual's health because of illness. Rather, it is to control or limit a natural body function. In that sense, the demand for sterilisations is derived from the demand for fertility limitation. That is, individuals derive utility from being able to control their fertility and may achieve more utility from being infertile than from being fertile, all else equal. A sterilisation may then be considered as an investment in health or a particular health state (i.e. being infertile).

Also a sterilisation is a substitute for other types of contraceptives. Some types of contraception, mainly for females, such a contraceptive pills may have adverse effects, which can be avoided if the female or her partner is sterilised. Then, the demand for sterilisation can be interpreted as a demand for improvements in health, since the individual's health is better when sterilised than in a situation where the individual is not and must rely on contraception that have adverse effects.

Sterilisations do not come with the same type of uncertainty as other types of health care services. Where illness is uncertain in the sense that an individual does not know

if he or she is going to be ill in the future with the need for health care, it is fully within the individual's control to decide whether and when to be sterilised.

Furthermore, this study differs from previous studies of the demand for health care as it analyse the impact of user charge on a health care service that an individual (usually) demands once compared to physician visits and outpatient hospital visits in general. Therefore, it has not been possible to analyse how the introduction of the user charge influences individuals' health care utilisation compared to their utilisation in previous years without the user charge.

Conclusion

In 2011, user charges for a sterilisation at public hospitals were introduced. The introduction caused a reduction in the demand for sterilisations and shifted the demand for male sterilisations to private hospitals.

References

ANDERSEN, J. S., OLIVARIUS NDE, F. & KRASNIK, A. 2011. The Danish National Health Service Register. *Scand J Public Health,* 39, 34-7.
INDENRIGS- OG SUNDHEDSMINISTERIET 2010. Bekendtgørelse om betaling for bestemte behandlinger med kunstig befrugtning, refertilisation og sterilisation i det offentlige sundhedsvæsen. BEK nr 1627 af 21/12/2010. *In:* INDENRIGS- OG SUNDHEDSMINISTERIET (ed.). København.
INDENRIGS- OG SUNDHEDSMINISTERIET 2011. Bekendtgørelse om betaling for bestemte behandlinger med kunstig befrugtning, refertilisation og sterilisation i det offentlige sundhedsvæsen. BEK nr 285 af 05/04/2011. *In:* INDENRIGS- OG SUNDHEDSMINISTERIET (ed.). København.
KONTOPANTELIS, E., DORAN, T., SPRINGATE, D. A., BUCHAN, I. & REEVES, D. 2015. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. *BMJ,* 350, h2750.
LYNGE, E., SANDEGAARD, J. L. & REBOLJ, M. 2011. The Danish National Patient Register. *Scand J Public Health,* 39, 30-3.
MEYER, B. D. 1995. Natural and Quasi-Experiments in Economics. *Journal of Business & Economic Statistics,* 13, 151-161.
NATARAJAN, V. & OAKLEY, N. 2008. Vasectomy. *In:* THOMAS, W. G. & SENNINGER, N. G. M. (eds.) *Short Stay Surgery.* Springer Berlin Heidelberg.
O'KEEFFE, A. G., GENELETTI, S., BAIO, G., SHARPLES, L. D., NAZARETH, I. & PETERSEN, I. 2014. Regression discontinuity designs: an approach to the evaluation of treatment efficacy in primary care using observational data. *BMJ,* 349, g5293.
PEDERSEN, K. M., ANDERSEN, J. S. & SONDERGAARD, J. 2012. General practice and primary health care in Denmark. *J Am Board Fam Med,* 25 Suppl 1, S34-8.
THOMSON, S., FOUBISTER, T. & MOSSIALOS, E. 2010. Can user charges make health care more efficient? *BMJ,* 341, c3759.
WAGNER, A. K., SOUMERAI, S. B., ZHANG, F. & ROSS-DEGNAN, D. 2002. Segmented regression analysis of interrupted time series studies in medication use research. *J Clin Pharm Ther,* 27, 299-309.

Social media data as predictors of Mikkeller sales?

Niels Buus Lassen, Dep. of IT Management, CBS
Lisbeth la Cour, Dep. of Economics, CBS
Anders Milhøj, Dep. of Economics, KU
Ravi Vatrapu, Dep. of IT Management, CBS

1. Introduction.

In recent years, social media data such as Twitter, Facebook and Google Trends data have proven promising as predictors for measures of economic outcomes of private firms. The main advantage of using social media data as predictors lies in the speed with which such data can be extracted and employed in the forecasting process. Once a firm has learned how to collect and pre-process their social media data, the information is available almost in real time and this implies that such data in combination with a good predictive model will provide a very useful tool for the management of the firm.

When working with social media data the concept of 'Big data' often comes to people's minds. In our case this is only partly true: we do work with large amounts of social media data, but once they have been pre-processed, we end up as many studies in the literature using quite simple dynamic regression models based on rather few time series observations. Hence the whole distinction between 'tall', 'fat' and 'huge' data as suggested in Doornik & Hendry (2014) becomes of less relevance. Ideally, if we were able to get economic performance data for a firm at a high frequency like the daily frequency, we would move closer to a situation where a more automatic model selection procedure would be relevant.

The novelty of the present paper is a predictive model for the total sales of Mikkeller using data at a monthly level. With these data we are allowed to be more precise when it comes to specification of the lag-structure in the dynamic regression model. Also we look into the importance of the data-preparatory work – in our case an unobserved component filtering of the data prior to regression modeling - on the social data proves to be for the final model and finally, we investigate the predictive power of types of social media data that have not been used as predictors before for a brewing company: Google shopping and YouTube data.

2. Briefly on the existing literature.

The idea of using social media data as predictors for e.g. company sales is not new. When it comes to model building, various experiments have been conducted and a summary of around 40 articles covering the time period 2005 – 2015 can be found in Buus Lassen et al (2017). For the present purpose the most interesting observations from these studies are that 1) almost 50% of the studies use some kind of regression model as their predictive model, 2) the range of social data types studied seem to cover Facebook, Twitter, Google Trends , Instagram, Tumblr, blogs and Youtube.

Theoretically, the argument for considering social data activity as predictors for sales obtains support from e.g. the AIDA model mentioned in Buus Lassen et al (2014). AIDA means *Awareness, Interest, Desire and Action* and refers to stages in a sales process. If social media data help increase the attention or can be considered a proxy for attention towards a product then it may also affect the final decision about buying. It is the general perception that more attention will increase sales even if the attention is negative.

When it comes to the specification of a set of predictive models we follow the literature and limit ourselves to the class of dynamic regression models. In these models we will have sales as our dependent variable and the different social data as suggested regressors. The reason why it is of interest to study social data regressors from different social media and search sources lies in the different ways such media are used. Google searches have proven to be the best social data for predicting sales. We call the Google data unpolished with a good connection to people's brains. Facebook data are polished, because people tend to display success and not failures on this social data. Twitter data are better than Facebook data for sales modeling, because Twitter data are less polished. But Google data are still beating Twitter data for sales modeling, because Google data are unpolished.

When building predictive models, the data frequency is of high importance. The higher the data frequency the more room is given for the researcher to build dynamic regression models with the aim of eventually forecasting economic performance measures such as sales. For many private companies sales data will only be available from the accounting data at a quarterly frequency (official balance sheets). This will limit the number of observations, and characteristics such as seasonal patterns may be more difficult to extract. Monthly data are much better as more observations will usually be available but at this frequency regular patterns over the week will still be impossible to discover. If possible to get, data at a daily frequency would be very well suited from all perspectives but are rarely available. In the present study we are able to work with monthly sales data for Mikkeller, a Danish micro brewery which has activities all over the world.

3. The data and methodology.

In order to build a predictive model for Mikkeller's sales we use data from Mikkellers accounting system combined with Google Trends, Google shopping and YouTube data. The social data has been collected from the free-access numbers available on the respective WEB-pages. We have searched for the word 'Mikkeller'. The free data from Google that we use are indexed such that they will vary between 0 and 100[1]. The time span of the study has been limited by our access to historical sales data and also the frequency of the data is reflecting our access to Mikkeller data. In the end we have a sample of monthly data that covers Januar 2014 to September 2016. Prior to analysis we index the sales data such that the max value becomes equal to 100. This transformation does not affect significance results later in the modeling process.

3.1 Pre-processing methodology

Our first considerations when it comes to data preparatory work concerns whether to use simple transformations of the series or just the raw series themselves. From a graphical inspection of total sales and the log of total sales it seems that using the log of sales may offer a slight statistical advantage as the variance seems more constant over the sample period than for the raw series, see Figures 1A and 1B.

With respect to the sales data we are checking the stationarity properties of the time series by means of an ACF graph. Stationarity is preferable for a regression model although stationarity may be of minor importance when the purpose of the model is forecasting.

The social data may consist of different components that we would expect to have different predictive value. Prior to including our social data time series as explanatory factors in our regression models we have the possibility to split them into a trend component, a seasonal component and an irregular component using classical times series techniques for unobserved components models (ucm). Our prior is that the irregular component will contain the most valuable information for predictive purposes as this component will capture special events, that creates a lot of attention towards the firm and its products. We also estimate models that use the social data in their 'raw' form without the ucm pre-processing for comparison reasons.

3.2 Unobserved Component Models

An unobserved component model, UCM, decomposes the observed series y_t into a sum of many components, as for instance

$$y_t = \mu_t + \varepsilon_t \qquad \mu_t = \mu_{t-1} + \eta_t$$

[1] It is possible to get the actual number of searches but they are not available for free.

Here the series μ_t is understood as the level of the series; but this level is unobserved. Only the series y_t which is affected by some noise or irregularities is observed. This noise series, ε_t, could in technical applications be measuring errors. But in this presentation the series ε_t is used as the irregular component which consists of special events happening to the series at time t which are not a part of the underlying level μ_t. In this paper these irregular components which are estimated for the observed sales series and for the three social data series are used in a usual regression/time series model in order to see if the social data series have any impact to the sales data in a setup where all usual time series variation for each series is accounted for by the unobserved components.

This basic formulation could be extended by trends and seasonality, and various forms for introducing autocorrelation in the model formulation also exist. A trend component has the form:

$$\beta_t = \beta_{t-1} + \xi_t$$

and the seasonal component is defined in a way so it does not affect the level component:

$$S_t = -(S_{t-1} + ... + S_{t-11}) + \zeta_t$$

In total these ideas lead to the model:

$$y_t = \mu_t + \beta_t + S_t + \varepsilon_t + \varphi\varepsilon_{t-1}$$

where also an autoregressive term for the irregular series is included.

All remainder terms, ε_t, η_t, ξ_t and ζ_t, are assumed to be mutually independent white noise series. Their variances could be estimated; the larger this component variance the more volatile the component. But it is also possible to fix this variance to the value zero which gives a constant component, e.g. a model with fixed seasonal dummies is found if $\text{var}(\zeta_t) = 0$. But if $\text{var}(\xi_t) > 0$ the trend is allowed to vary over time which is a very flexible feature!

The parameters of these models, the variances and the AR(1) parameter, and the component values could be estimated by the Kalman filter. This gives an algorithm for successive calculation of the unobserved components at time t conditioned on previous observations y_{t-i} $i = 0, .., t-1$. The Kalman filter is useful if prediction is the purpose of the analysis as the algorithm does not include future observations y_{t+i}. A further smoothing estimation, where all available information is used when estimating the unobserved components at any time t, also exist. In this paper this method will be used.

Our hypothesis when it comes to the UCM components is that they probably will have most potential if the data frequency is high. With our monthly data the idea may still

be applicable but there is a danger that the temporal aggregation level will make it more difficult to find the type of effects we are looking for.

3.3 The regression models

In order to specify a predictive model, we direct our focus to the class of dynamic regression models. Unfortunately, even with monthly data we are left with rather few observations which will limit our possibility to work with both complex lag structures and many non-linear terms like power expressions and interactions.

The primary model equations we use are of the type:

$$y_t = \beta_0 + \gamma y_{t-1} + \beta_1 x1_{t-1} + \beta_2 x2_{t-1} + \ldots + \beta_k xk_{t-1} + \varepsilon_t \quad t = 1, \ldots, T \qquad (1)$$

where y is sales, xj is a social data measure and the sub scripts, t - 1, indicate that only lagged values of sales and social media data are used as predictors. The basic model can be extended by allowing for more than one lag of each xj variable but due to the limited number of observations for the estimation period the more predictors we include the fewer lags we can afford to consider. Also by including the lag of y in the equation we actually consider a longer lag structure of x but with a specific exponentially decaying pattern in the effects over time. Hence our preferred initial specification as provided by equation (1) will include the lag of y and only one lag for each additional explanatory factor. The error term, ε_t, is assumed to fulfill the standard assumptions for OLS estimation.

We also provide empirical evidence based on a model of the type:

$$y_t = \beta_0 + \gamma y_{t-1} + \beta_1 x1_{t-j} + \varepsilon_t \quad t = 1, \ldots, T \qquad (2)$$

where the error term, ε_t, again is assumed to fulfill the standard assumptions for OLS estimation. Choosing to include just the j'th lag may be based on more empirical arguments. Also y_{t-1} may be left out of equation (2) if that seems to make more sense.

It is difficult to judge the predictive performance of a specific forecasting model unless we have some benchmark to compare to. For sales of individual companies there is no general guideline in the literature on how to choose such a model, so we will argue for our choice in the following way: we want a benchmark model that is simple, that seem to capture some of the apparent time series properties in our data and that do not contain exogenous explanatory factors. In this study we will suggest two such model 1) a simple AR(1) model[2] as suggested by the standard identification procedure from classical time series analysis (see Figures 2A and 2B.

[2] Because we have no indication of non-stationarity of our sales series, we go for the AR(1) specification instead of a random walk which is often used as a benchmark in the exchange rate literature.

In addition to this model we also consider a model that includes the first lag of log sales but also a December dummy and a time trend. The December dummy is at first hand negative but as the sales numbers are at the time of production the low December values are due to low sales in January or later months.

Benchmark 1: $y_t = \beta_0 + \gamma y_{t-1} + \varepsilon_t \quad t = 1, ..., T$ (3)

Benchmark 2: $y_t = \beta_0 + \gamma y_{t-1} + \beta_1 D\_December_t + \beta_2 Trend_t + \varepsilon_t \quad t = 1, ..., T$ (4)

In the end, as our model is a forecasting model, we need to split the sample into a training and a test part in order to assess the out-of-sample forecasting properties, see e.g. Hyndman & Athanasopoulos (2014). With our short sample we retain only the last 3 month of the sample for the test part, i.e. July – September 2016. This leaves us with an estimation or training sample of 30 observations: January 2014 – June 2016. We will provide an out-of-sample forecast results based on a series of 1 step ahead predictions for July – September 2016. Evaluations will be based on graphs comparing actual sales to predicted sales for the selected models and also by numerical measures like RMSE and MAE.

4. Descriptive statistics.

In this section we will provide a series of graphs and tests that will help us illustrate the basic time series properties of the data and support us in arguing for the transformations we decide to use in our models. We start by showing the development over time in both the sales and the log of sales, see figures 1A and 1B. The first impression of the development of the series is that there seems to be some indication of an upwards trending behavior. An alternative to this interpretation could be an interpretation of a non-trending series but with a level shift upwards. We will start by considering the first case of an increasing trend as it makes it unnecessary for us to decide on an exact timing of a level shift. But it may be worthwhile to keep this second option in mind for future studies. When comparing the graphs of sales and log of sales it is not clear which one to prefer. In the end we decided to go for the log of sales as this is often done for longer time series where the variation increases with increases in the level. Also focusing on the logs will allow for an interpretation in percentage terms when it comes to the analysis of the regression models.

Figure 1A: Total Sales

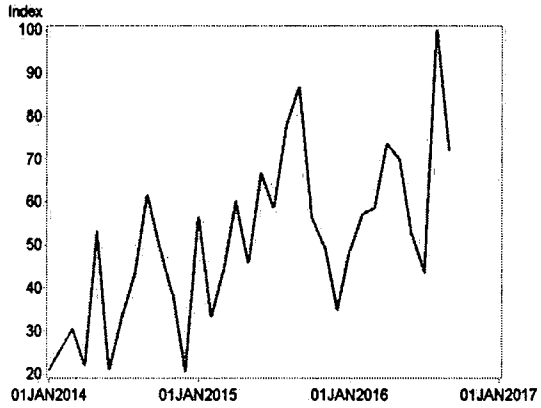
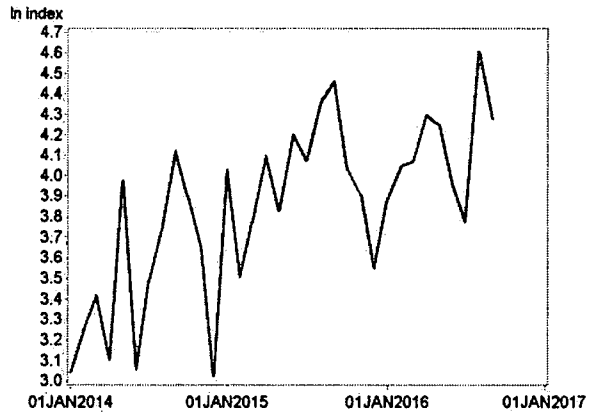


Figure 1B: Log of Total Sales



When taking a closer look at the time series properties of the series it seems that a decision of treating this series as stationary would be a good starting point. The ACF graph clearly supports this conclusion as the 1st order autocorrelation coefficient is 0.41[3].

Figure 2A: ACF for log of Total Sales

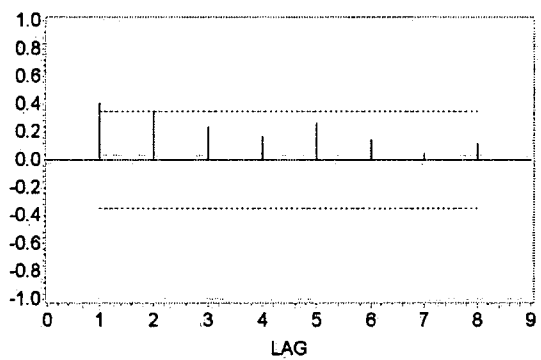
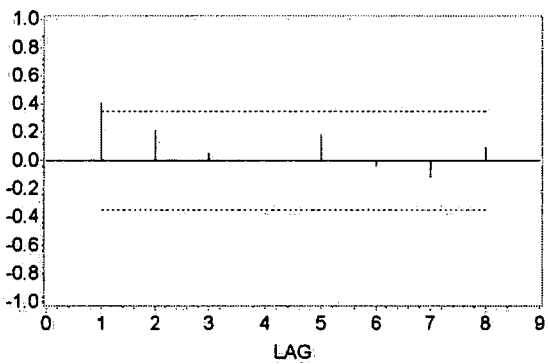
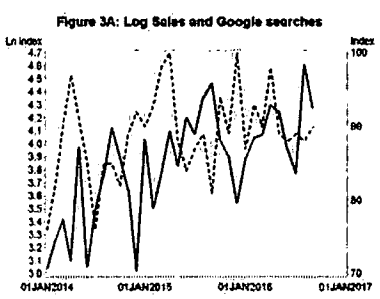


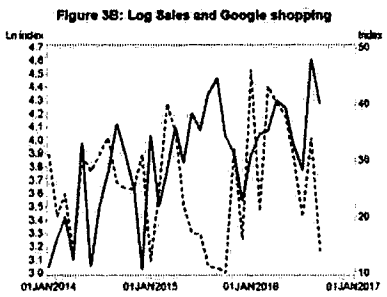
Figure 2B: PACF for log of Total Sales

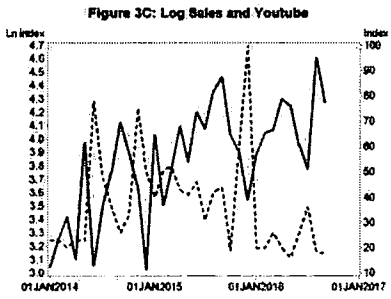


To get a first impression of potential linkages from social data activity to the Mikkeller sales we plot each social data activity in a time series plot together with the log of sales. As seen in figures 3A – 3C none of the social data series seem to follow the sales very closely – not even if some lagging is considered. To gain further knowledge about the correlation behavior we also provide a matrix plot of log sales and the potential regressors. There is for some of the variables a vague pattern that would support a correlation different from zero but the general picture is not too promising.

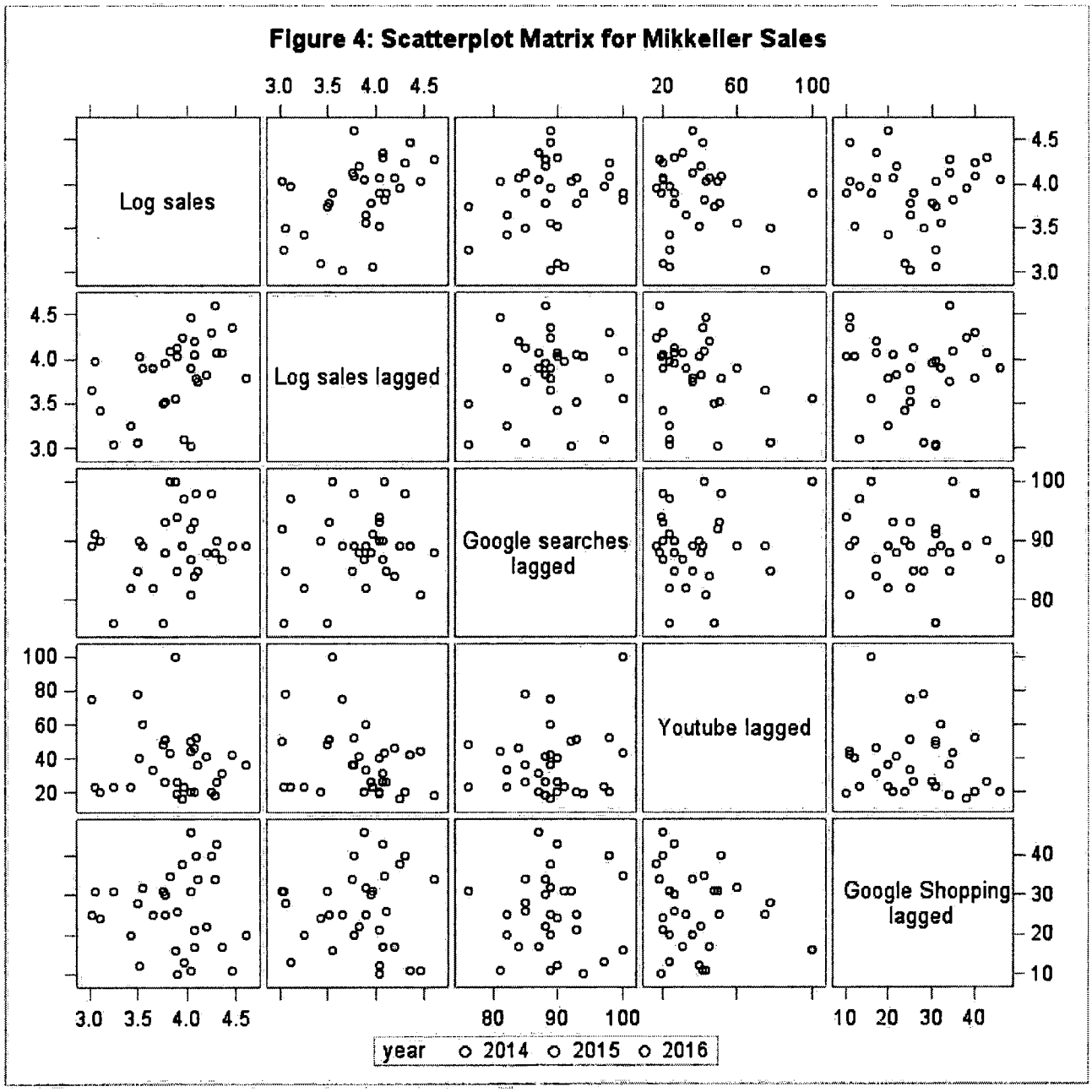
[3] Also an ADF test of non-stationarity of the series supports a conclusion of stationarity. With a trend in the equation of this test we reject at the 10% level the null of a unit root with p-values of 0.001 and 0.059 for zero and 1 lagged differences in the equation, respectively. However with our very short sample period this test may not be too reliable. At least it does not contradict our impression from the ACF graph.







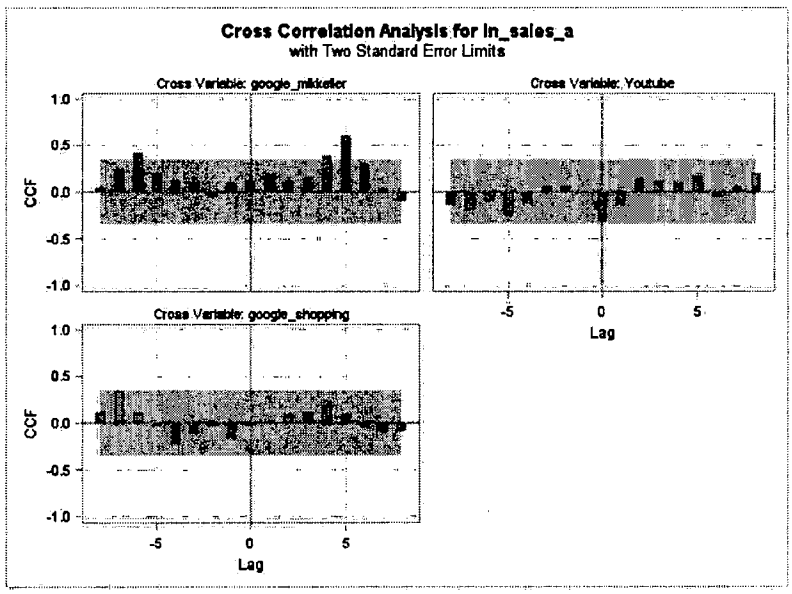
Note: in Figures 3A – 3C the solid curve represents the sales.



Next in order to get some more knowledge about possible lagged effects, in Figure 5 we present the cross correlations between the indexed sales variable and each of the social activity variables. The cross correlations are constructed in such a way that a

positive number, s, on the horizontal axis implies a correlation between sales at time t and the social variable s periods prior to the present one. It seems from these graphs that we cannot expect much explanatory power from including these social media variables as regressors in a predictive model for sales. The only significant spike is for Google searches with a lag of five. If many of the customers buy their quantities at irregularly rather than on a smooth continuous basis it may make sense to include the lag 5 variable in a regression model. It is, however, also possible that this lag becomes significant for more random reasons due to our short sample period. An out-of-sample test of the model may help here.

Figure 5



Finally, we display the summary statistics of sales and the four series of social media attention:

Table 1: Descriptive summary statistics.

Variable	N	Mean	Std Dev	Minimum	Maximum
Log sales	33	3.8423177	0.4212766	3.0264408	4.6051702
Google searches	33	88.9393939	5.9315860	76.0000000	100.0000000
Youtube	33	37.1818182	19.6793905	16.0000000	100.0000000
Google Shopping	33	25.8484848	9.9784806	10.0000000	46.0000000

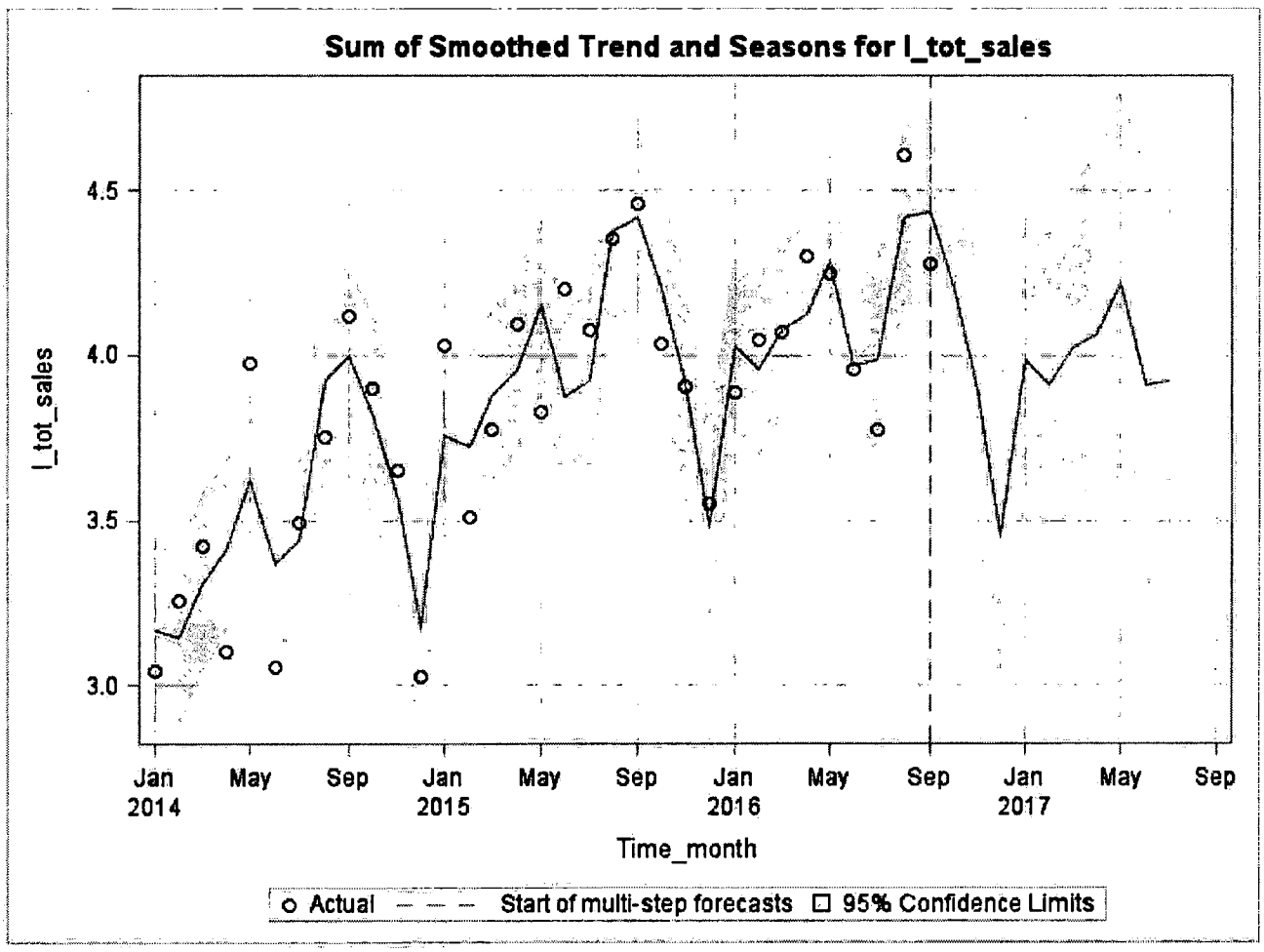
5. Unobserved components models for the sales series and the social media series.

For the log transformed, indexed sales series the resulting model is:

$y_t = \mu_{t-1} + \beta_t + S_t + \varepsilon_t + \varphi\varepsilon_{t-1}$

The variance in the seasonal dummy series is fixed to zero, meaning that the dummy variables are constant. The trend variance varies and this allows the trend to be significantly positive for the first year or two and then the trend is zero. This changing trend is clearly seen at the fit plot even if the seasonal component is also present at the plot.

Figure 6



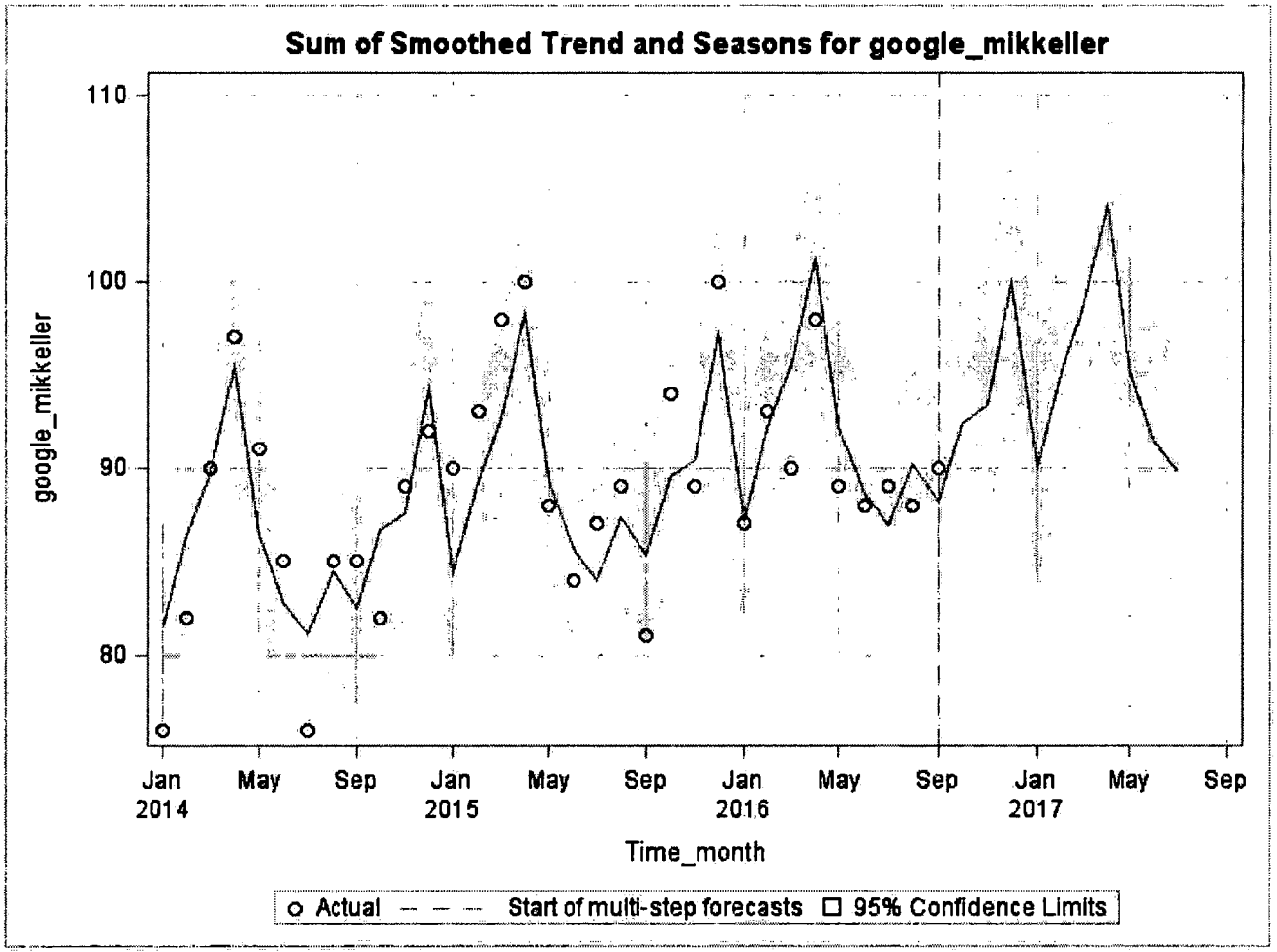
For the Google series the resulting model is:

$y_t = \mu_t + \beta_t + S_t + \varepsilon_t + \varphi\varepsilon_{t-1}$

Again the variance in the seasonal dummy series and now also in the trend series are fixed to zero, meaning that the trend and the seasonal dummy variables are constant.

The series has a constant upward trend at 0.25 each month, which is significantly positive. This means that the interest in google searching for the word "Mikkeller" is steadily increasing. The trend is clearly seen at the model plot in figure 7.

Figure 7



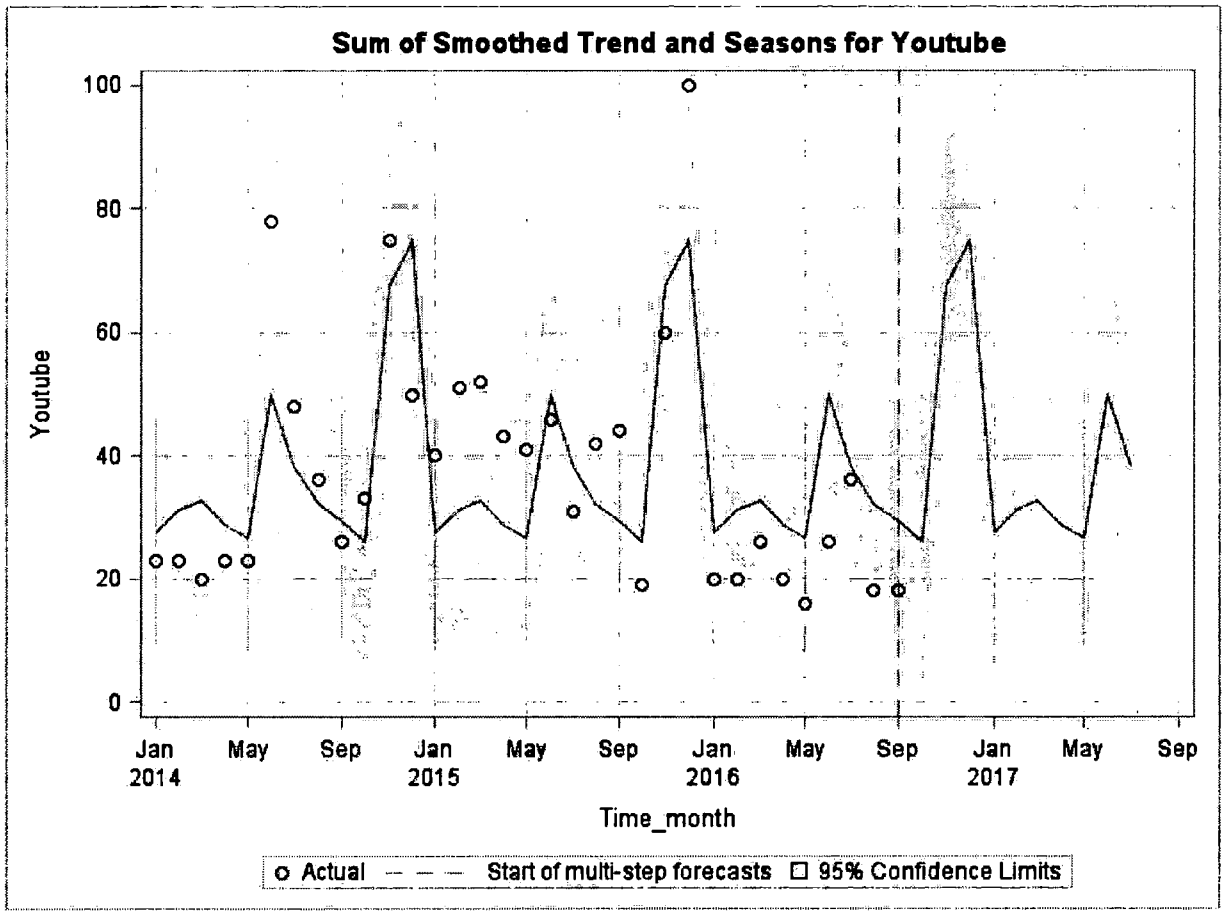
The seasonal component tells that Google searching for the word "Mikkeller" has peaks every year in April and December.

For the Youtube series the AR(1) term is insignificant and no trend is present. The resulting model is

$$y_t = \mu_t + S_t + \varepsilon_t$$

where all variances are fixed at zero. The model plot tells that the series has a clear seasonal component with large values in the months of November and December.

Figure 8



For the series of Google Shopping the resulting model is even more simple as the seasonal component is insignificant.

$y_t = \mu_t + \varepsilon_t$

6. Results of predictive modeling at the monthly frequency.

We now consider various specifications for models that contain social media activity data and/or their lags as explanatory factors as suggested by the main equation (1) but also by the additional equation (2).

As our main purpose it to determine a model that can produce 1 step ahead forecasts out-of- sample, we will as a starting point not allow for contemporaneous regressors in the models. This may be a limitation when we are working with monthly data as social media activity in the beginning of a month may affect sales already in the same month. Another possibility is that it actually takes several months for us to see a reaction if each customer only send an order with months in between.

6.1.1 In sample regression results.

Table 2: Regression results for Log Sales before ucm.

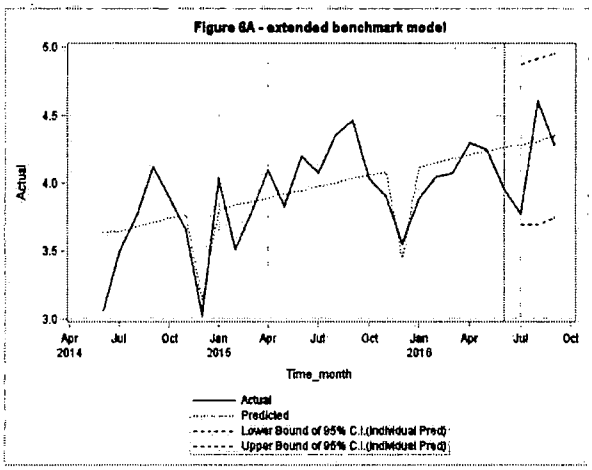
Variables	Benchmark 1 AR(1)	Benchmark 2 extended AR(1)	Model (1) Extended	Model with more lags	Model with lag 5
Intercept	2.57***	3.40***	4.19***	3.69**	1.18**
	(0.77)	(0.61)	(1.38)	(1.47)	(0.55)
Lagged sales	0.34*	0.02	0.03	-	-0.17
	(0.20)	(0.17)	(0.24)		(0.12)
December	-	-0.65***	-0.70***	-	-0.42***
		(0.19)	(0.23)		(0.13)
Trend	-	0.03***	0.03***	-	0.02***
		(0.01)	(0.01)		(0.01)
Google searches, lag1	-	-	-0.01	0.01	-
			(0.01)	(0.01	
Youtube, lag1	-	-	0.00	-0.01*	-
			(0.00)	(0.00)	
Google shop-ping, lag1	-	-	0.00	-0.00	-
			(0.01)	(0.01)	
Google searches, lag2	-	-	-	-0.01	-
				(0.01)	
Youtube, lag2	-	-	-	0.00	-
				(0.00)	
Google shop-ping, lag2	-	-	-	0.00	-
				(0.01)	
Google searches lag5	-	-	-	-	0.03***
					(0.01)
Adj. R square	0.08	0.48	0.43	-0.06	0.79
# observations	25	25	25	25	25

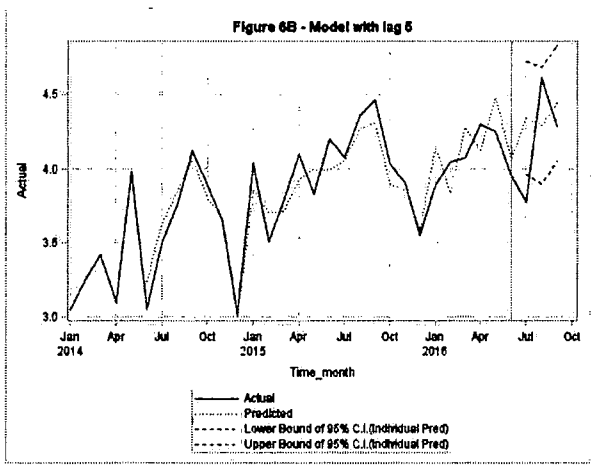
Note: the estimation sample has been restricted such that it is the same for all specifications even though models with fewer lags could have used more observations. Note2: Standard errors in parentheses. Significance at 10%: *, 5%:**, 1%: ***.

The basic massage from the table is that much of the variation in the log of sales can be explained by deterministic terms like a seasonal December dummy and a trend. The table also reveals that it is difficult based on the present sample to find significant effects from the social media activity variables. The most promising suggestion is to include the lag 5 of the Google searches in the model. This factor becomes very significant and also the model overall reaches an R square close to 80% in that case. The economic interpretation of including the lag 5 term is, however, more uncertain but may relate to lagged buying behavior from some customers.

6.1.2 Out-of-sample predictive power?

We predict the log of sales for the time period July 2016 to September 2016. First we show a set of graphs that compares such prediction for the actual values. We show graphs for the extended benchmark model and for the model including lag 5 in the last column of table 2.





From these graphs it is evident that not many of the movements in sales are captured by the benchmark model. Also the confidence bands for the prediction are quite wide and the actual values are inside the band. The 'lag 5'-model are capturing the movements in sales much better in sample but out of sample the big swings in July and August 2016 are not captured very well. The actual value for July is in fact outside the confidence bounds.

In table 3 we show some numerical measures for the forecasting performance of the models from table 2. We have chosen just to focus on a few measures and some of the more commonly used ones: MAE (mean absolute error) and RMSE (root mean squared error)[4].

Table 3: Summary measures on predictive power.

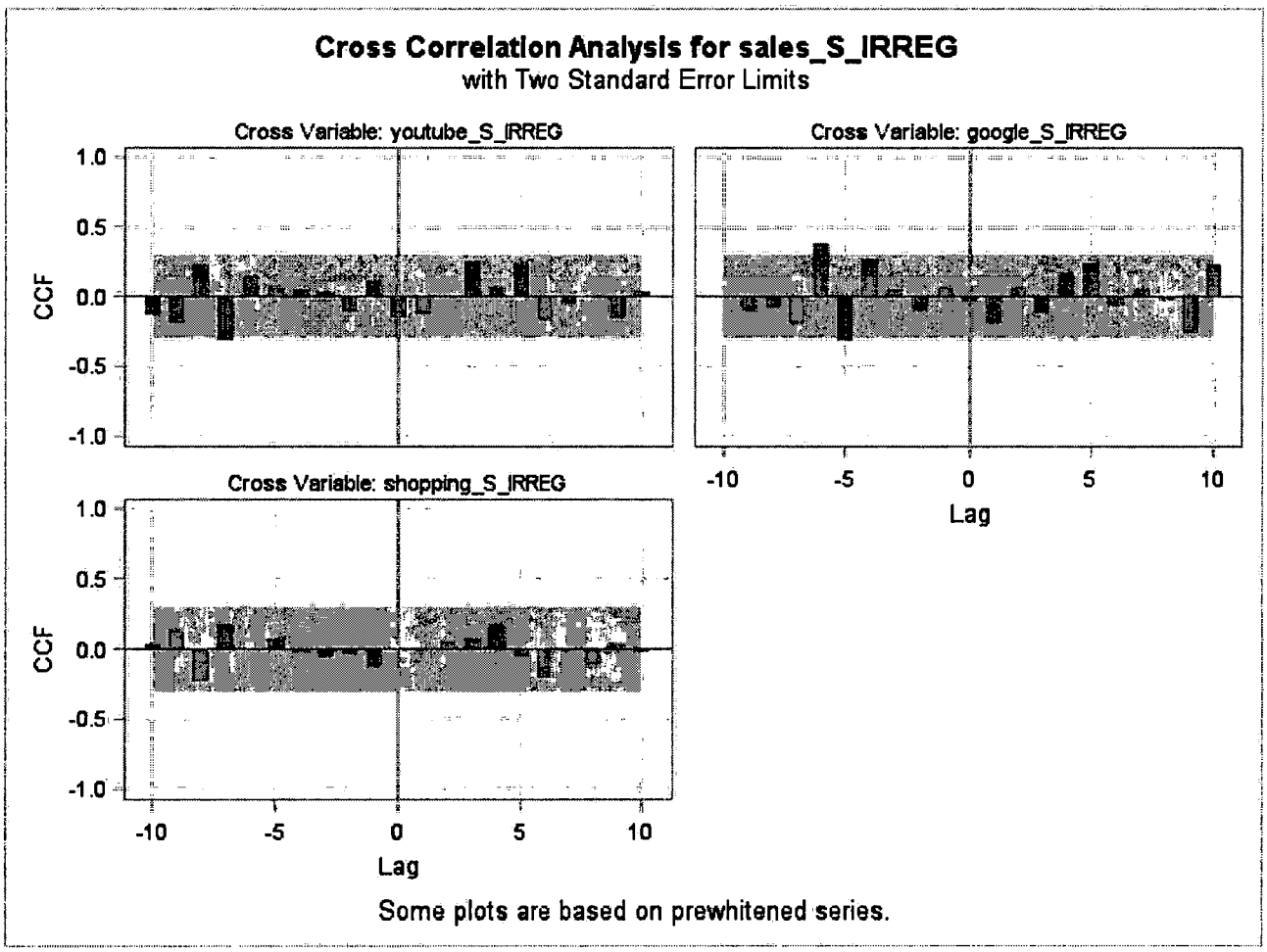
Summary measure	Benchmark 1 AR(1)	Benchmark 2 extended AR(1)	Model (1) Extended	Model with more lags	Model with lag 5
MAE	0.344	0.292	0.315	0.381	0.350
RMSE	0.449	0.342	0.363	0.444	0.385

Note: In all cases the numbers have been calculated based on the 3 months of July, August and September 2016.

[4] For formulas on how to calculate these measures , please consult e.g. Hyndman & Athanasopoulos (2014)

The numbers in table 3 also indicate that the extended benchmark model performs better when it comes to forecasting out-of-sample in our case. The model that performed best in-sample lies in the middle.

The results of this forecasting exercise may be due to the few months on which we base it and may also be a result of the very volatile behavior in this period. Somehow it seems that maybe some customers have decided to postpone their orders from July to August.



6.2. Predicting the irregular component for the sales series by the irregular components for the social media series. Initial investigations.

As an alternative to the attempts elsewhere in this paper, in this section we try to model the relations among the irregular components for all four series. This seems to be a fruitful way to go as the irregular components are cleaned from all trends, level shifts and seasonal variation. The relation between the sales on the left hand side and the three social media series on the right hand side is however insignificant even if lags or leads are considered.

This is easily seen by the cross correlation functions; see the plots for the series S_IRREG above. Two of the irregular components include an autoregressive term by construction which means that these two series are prewhitened before calculating the cross correlations.

It is clear that no significant relations are found at these plot and we will not pursue this modeling further at the present stage.

7. Summary and conclusion

In this paper we have pursued our idea of applying a preparatory ucm model to both regressors and regressand to determine a forecasting model for the monthly sales of the Danish micro brewery Mikkeller. Our modeling attempts were mainly unsuccessful as the ucm modeling did not lead to any significant regression model based on regressors being activity from Google trends, Youtube and Google Shopping. Also when following a more traditional strategy without the preparatory ucm modeling, the benchmark model that contained a trend and a December dummy seemed to perform the best even though we found some support for a lag-5 effect from Google Trends. Much of our lack of success with the present modeling may be due to the fairly short sample period that consisted of monthly data from January 2014 until September 2016. Therefore future studies building on the same idea but with access to longer and maybe even more high frequent sample periods may prove more successful.

8. References

Buus Lassen, N., la Cour, L., Vatrapu, R. (2017), 'Predictive Analytics with Social Media data' in Sloan & Quan-Haase ed. *The SAGE Handbook of Social Media Research Methods*, Chapter 20, pp 328-341

Buus Lassen, N., Madsen, R. and Vatrapu, R. (2014). 'Predicting iPhone Sales from iPhone Tweets', Conference Paper, *2014 IEEE International Enterprise Distributed Object Computing Conference.*

Buus Lassen, N., Vatrapu, R., la Cour, L., Madsen, R. and Hussain, A.(2016), 'Towards a Theory of Social Data: Predictive Analytics in the Era of Big Social Data', in P. Linde (Ed.) *Symposium i Anvendt Statistik*, Page 241-256

Doornik & Hendry (2014). 'Statistical Model Selection with 'Big Data', *Department of Economics Discussion Paper Series*, University of Oxford, #735.

Hyndman, R. J., & Athanasopoulos, G. (2014). *Forecasting: principles and practice*: OTexts: https://www.otexts.org/fpp/

Find en flirt

Sara Armandi, SAS Institute

1 Introduktion

D. 3. april 1973 foretog den amerikanske projektleder Martin Cooper fra data- og kommunikationsvirksomheden Motorola eftersigende det første opkald fra en håndholdt mobiltelefon. Opkaldet gik til rivalen Joel Engel fra Bell Labs, som igennem flere år havde ligget i krig med Motorola om at komme først med den håndholdte mobiltelefon (Sørensen, 2013).

I løbet af få år har mobiltelefonen udviklet sig fra at være for de få og et tegn på velstand, til en næsten uundgåelig allemandseje. Mobilen er vigtig for det moderne menneske i den travle verden, da det både skal være nemt og hurtigt at komme i kontakt med omverdenen. Uden mobiltelefonen ville mange af de menneskelige relationer der eksisterer i dag måske slet ikke være mulige, da de i en verden uden mobiltelefonen ville kræve alt for meget tid at opretholde. Men hvad er det for nogle relationer der eksisterer over telefonen? Nogle relationer eksisterer på mobilen udelukkende af praktiske årsager, mens andre relationer er for hyggens skyld. Et eksempel på en hyggelig relation, som (for mit vedkommende i hvert fald) måske ikke havde eksisteret på samme måde, hvis der ikke havde eksisteret mobiltelefoner er *flirts*.

1.1 Hvad er en flirt?

Der findes ingen præcis definition, men i denne artikel er en flirt betegnet som en person, man finder interessant og umiddelbart har flere, og mere romantiske følelser for, end en ven. En relation der starter som en flirt kan enten flyde ud i sandet, føre til et venskab eller måske endda føre til et kæresteforhold.

Normalt vil der i en periode være intensiveret telefoniskaktivitet med en flirt som følge af den øgede interesse. Der vil i særdeleshed være tale om en øget sms-aktivitet, da sms'er er rimelig uforpligtende, ligesom en flirt er det. Hvis man kunne få indblik i en persons mobilaktivitet, burde det derfor være muligt at få en indsigt i mobilindehaverens relationer, og måske endda i hvem denne person flirter med.

Detaljerede oplysninger om mobilaktivitet er overraskende lettilgængelige, da de fleste mobilfakturaer indeholder udførlig information omkring hvert opkald der er foretaget samt hver enkelt sms der er sendt. Denne type data er oprindeligt kun tiltænkt forbrugeren

som en betalingsopkrævning og oversigt over forbrug. Men en lille smule manipulation gør det muligt at få indsigt i meget personlige oplysninger om relationer mellem mobilindehaveren og kontaktpersonerne. Mobilindehaveren kan dermed selv få en dybere forståelse af relationerne til sine kontakter. På den anden side kan dataet også udnyttes til overvågning af andre, f.eks. hvis man har adgang til sin kærestes eller sin datters telefonregninger[1]. Telefonregningen kan altså afsløre mere end bare hvor mange penge der bliver bruget.

I denne artikel undersøges det, igennem en eksplorativ clusteranalyse af mobiladfærden, om det på baggrund af særlige aktivitetsmønstre er muligt at segmentere kontaktpersoner i forskellige relations-grupper. Det er i særdeleshed af interesse at undersøge, om det er muligt at finde frem til hvilke kontakter der kan kategoriseres som flirts.

2 Teori

Der er mange årsager til at man ønsker at segmentere sit data. Et eksempel kan være en salgsvirksomhed som ønsker at dele kunderne i forskellige grupper som har forskellige profiler, så marketingsstrategier kan optimeres. I denne artikel er formålet rent deskriptivt, da forhåbningen er at få en større indsigt i data. Metoden der anvendes til dette formål er clusteranalyse[2].

2.1 Clusteranalyse

Clusteranalyse er en såkaldt usuperviseret klassificering, hvilket betyder at segmenteringen udelukkende bestemmes på baggrund af det pågældende software samt den algoritme der anvendes. Dette er i modsætning til superviseret klassificering, hvor en bruger kan guide segmenteringen ved at udtage data der er repræsentative for bestemte clustre, og derefter bede softwaren om at bruge dette som referencer for segmenteringen af andet data.

I clusteranalyse segmenteres data således at observationerne internt i en gruppe (kaldet et cluster) er så ens som muligt, samtidig med at de adskiller sig mest muligt fra andre clustre. Segmenteringen er baseret på baggrund af data og bestemmes af ligheder i inputvariablene. Da inputvariablene spiller en væsentlig rolle i clusteranalysen, er udvælgelsen af disse også afgørende. Først og fremmest skal inputvariablene være

[1] Det kan selvfølgelig ikke anbefales at man overvåger nogen uden at de er klar over det.
[2] Teoriafsnittet er primært baseret på Ravenna, A., Truxillo, C. & Wells, C. (2015)

meningsfulde for analysens objektiver, da fortolkningen og forklaringen af de endelige clustre ellers ikke vil give mening. Derudover vil uafhængighed mellem variablene samt et begrænset antal være med til at gør clustrene mere stabile. Endelig bør inputvariablene have lav skævhed og topstejlhed (kurtosis), så man undgår at danne meget små clustre der kun indeholder en enkelt eller meget få observationer.

2.2 *K*-means

En af de mest almindeligt anvendte metoder til clusteranalyse er *k*-means algoritmen. Det er en simpel og hurtig algoritme der skaleres godt til store datasæt. Algoritmen tildeler observationer til centroider og minimerer afstanden fra observationerne til disse clustermidtpunkter.

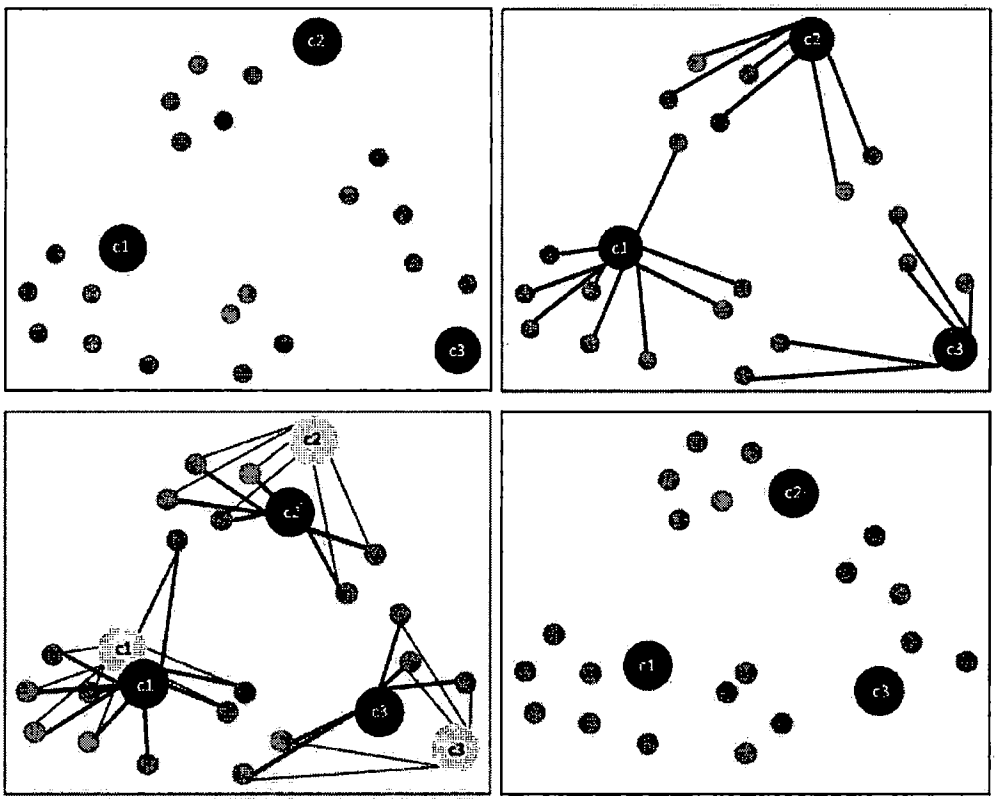
Udover valg af inputvariable, er det første skridt i algoritmen at vælge en passende værdi for *k*, som angiver antallet af clustermidtpunkter kaldet centroider. Man bør vælge *k* til at være konsistent med den naturlige koncentration af observationer, eller med ens analytiske objektiver.

Figur 1 viser *k*-means clusteranalysen i aktion. I dette eksempel vælges først tre tilfældige punkter der anvendes som centroider for observationerne. I denne artikel er den initiale værdi for centroiderne valgt til at have den samme værdi som *k* vilkårlige observationer i data. Derefter tildeles hver observation til centroiden tættest på. Det næste der sker er at kvadratsummen af de tilhørende observationer inden for hvert cluster bestemmes. Denne værdi er givet ved:

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{j=1}^{p}\left(x_{ij}-\bar{x}_{kj}\right)^2 \qquad (1)$$

hvor n_k er antallet af observationer i det k'te cluster og $\bar{x}_{kj}$ er gennemsnitsværdien for ikke-manglende observationer for den j'te variabel i det k'te cluster (ClimServ, 2007). I *k*-means clusteranalysen er det netop værdien (within-cluster sum of squares) i Ligning (1), der forsøges minimeret. Ved hver iteration udregnes en ny lokation for centroiderne, hvilket svarer til den endelige konfiguration af algoritmen. Efter hver iteration bliver den endelige konfiguration sendt igennem samme loop indtil centoriderne er konvergeret til deres endelige positioner.

Figur 1: *K*-means clusteranalyse i aktion



Kilde: http://blog.guillaumeagis.eu/k-means-clustering-apache-mahout/

K-means clusteranalysen kan altså deles op i tre steps:

1. Vilkårlig determinering af initiale centroider
2. Tildeling af nærmeste observationer i clusteret til centroiderne
3. Bestem nye centroider ud fra gennemsnittet af observationer fra dette cluster

De ovenstående steps gentag indtil det maksimale antal iterationer er nået, eller indtil forskellen i kvadratsummen indenfor clustrene i to successive iterationer er lavere end en fastsat grænseværdi.

3 Data

Analysen er baseret på fakturaer over mit eget mobiltelefonforbrug hentet på "Mit Telmore". På Mit Telmore findes månedlige telefonregninger samt uddybende specifikationer for de sidste fem år. For et par år siden var jeg forudseende, og hentede de daværende tilgængelige fakturaer, hvorfor der foreligger fakturaer helt tilbage fra 2008.

Figur 2 viser et eksempel på hvordan data er struktureret i fakturaerne. Ud fra disse er det muligt at danne et datasæt hvor hver observation svarer til en enkelt aktivitet foretaget fra mobiltelefonen. I dette tilfælde dækker en aktivitet enten over et opkald foretaget eller en sms sendt fra telefonen. Der er således udelukkende tale om data der indeholder informationer om udgående aktivitet (fra mobilindehaveren til kontakterne), og der foreligger derfor ingen oplysninger om indgående opkald eller sms'er (fra kontakterne til mobilindehaveren).

Figur 2: Udklip fra faktura over mobiltelefonaktivitet

SMS

Beskrivelse	Antal	Dato	Varighed	Beløb
SMS - 23613652	1	2015-07-01 13:31:28		0,00
SMS - 61332039	1	2015-07-01 12:23:13		0,00
SMS - 61332039	1	2015-07-01 03:49:02		0,00
SMS - 22128081	1	2015-07-01 02:32:53		0,00
SMS - 22128081	1	2015-07-01 02:32:49		0,00
Sub total	**288**			**0,00**

Opkald i Danmark

Beskrivelse	Antal	Dato	Varighed	Beløb
Opkald - 23613652	1	2015-07-19 09:27:21	00:00:50	0,94
Opkald - 23613652	1	2015-07-18 20:19:18	00:01:02	1,58
Opkald - 23613652	1	2015-07-18 16:48:00	00:02:51	2,22
Opkald - 61332039	1	2015-07-18 15:51:05	00:01:55	1,58
Opkald - 23613652	1	2015-07-18 13:50:13	00:04:00	2,86
Opkald - 61332039	1	2015-07-18 12:19:59	00:01:19	1,58

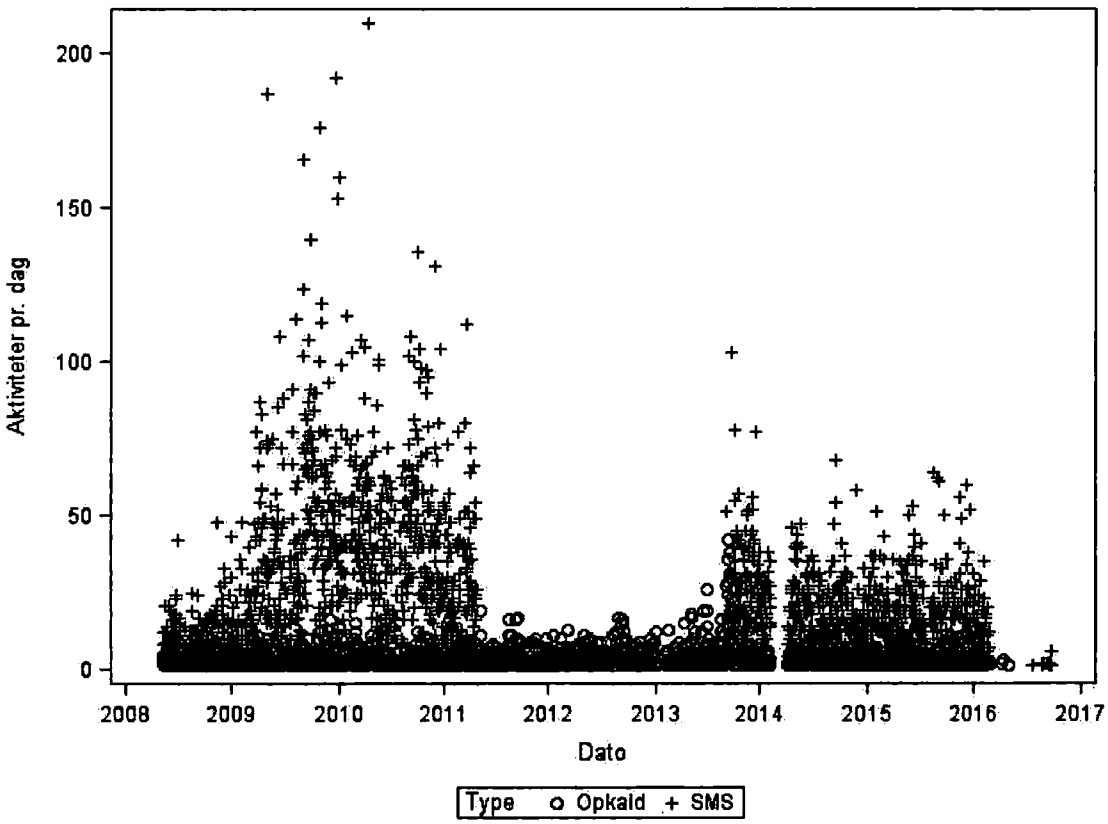
Kilde: Mit Telmore, https://www.telmore.dk/website/mit-telmore

Der er i alt 56.462 observationer, som hver svarer til en enkelt aktivitet. Af disse er 48.433 (85,78%) kategoriseret som sms'er, mens 8.029 (14,22%) er opkald. Den første aktivitet observeres d. 7. maj 2008 kl. 19:50:24 og den sidste observation er fra d. 27. september 2016 kl. 18:19:57. De 56.462 aktiviteter fordeler sig blandt i alt 901 forskellige kontaktpersoner, hvor den højeste aktivitetsfrekvens for en enkelt kontakt er på hele 4.639 aktiviteter (svarende til 8,22% af det samlede antal aktiviteter). Der viser sig dog også at være en masse kontakter, helt præcis 274 (30,55% af alle aktiviteter), hvor der kun har været en enkelt aktivitet i hele perioden.

Figur 3 viser udviklingen i den daglige aktivitet fordelt på de to aktivitetstyper. Det er tydeligt at se, at der er en forskel i frekvenserne på antal daglige opkald i forhold til daglige sms'er. Desuden viser figuren, at der i en periode ikke har været registreret

sms'er. Perioden strækker sig fra d. 28. april 2011 til den 6. september 2013. Årsagen til dette databrud er ukendt, og på trods af utallige samtaler og mailkorrespondancer har Telmore ikke en forklaring på det manglende data.

Figur 3: Daglig telefonaktivitet 2008-2016



Figur 3 viser at antallet af opkald er nogenlunde konstant over hele perioden, selvom det kan være lidt svært at se. I 2013 sker en stigning i det daglige antal opkald, men det aftager igen i 2014. Antallet af sms'er er lavt i 2008 hvorefter det stiger kraftig, for derefter at være faldende over hele perioden (hvis man ser bort fra databruddet). Det relativt lave antal sms'er det første år kan skyldes at jeg i den periode havde en fast kæreste. Vi går fra hinanden, og så stiger sms-intensiteten. Fra 2014 får jeg en ny kæreste, hvor intensiteten igen falder. En anden årsag til at sms-aktiviteten er dalende kan være at der i løbet af de sidste par år er kommet flere og flere alternativer til at sende sms'er, f.eks. Facebook, Messenger, Snapchat, osv.

I starten af 2014 er der slet ingen aktivitet. Dette skyldes at jeg fra februar 2014 og to måneder frem befandt mig i det caribiske øhav, og valgte ikke at anvende min mobiltelefon. Yderligere, er der i starten af 2016 og frem nærmest ingen aktivitet.

Årsagen til dette er, at jeg startede job d. 1. februar 2016 og fik udleveret en mobiltelefon fra arbejdet, hvorefter min egen telefon lige så stille måtte lade sig pensionere.

3.1 Aggregering og rensning

For at tilpasse datasættet til clusteranalysen aggregeres data, således at en observation svarer til en kontaktperson. Således reduceres datasættet fra at indeholde 56.462 observationer til 901 observationer.

Udover aggregering foretages en rensning af datasættet, så irrelevante observationer udelades. Først og fremmest fjernes telefonnumre der er fejlindtastet, hvilket betyder, at numre der indeholder andet end 8 cifre slettes. Der er desuden taget højde for at enkelte kontakter har skiftet nummer undervejs, hvorfor disse er samlet til ét telefonnummer. Dette er dog ikke gjort systematisk for samtlige kontakter, hvorfor enkelte numre nok burde have været slået sammen. Hermed reduceres datasættets størrelse til 891 observationer.

For at begrænse data yderligere betragtes udelukkende alle de kontakter hvor der har været aktivitet i over en uge. Antallet af dage mellem første og sidste aktivitet skal altså være 8 eller der over. Hermed udelades i alt 441 kontakter (49,16% af det samlede antal kontakter)

Da modellen udelukkende regner på de observationer hvor der er værdier for alle inputvariable, kræver det, at der har været både sms- og opkaldsaktivitet med den pågældende kontaktperson. Af denne årsag udelades 202 observationer, da de enten kun har modtaget sms'er eller kun opkald (bl.a. forsvinder hjemmetelefonen i mine forældres hus, da man ikke kan sende sms'er til det nummer). Dette svarer til at vi fjerner 22,52% af det oprindelige antal kontakter, eller i alt 44,89% af de kontakter hvor aktivitetsperioden har været over 7 dage. I alt udføres den endelige clusteranalyse på 248 kontaktpersoner.

3.2 Inputvariable

Som nævnt i teoriafsnittet, er valget af inputvariable afgørende for at lave en fornuftig clusteranalyse. I denne artikel tages udgangspunkt i fire variable der alle er udregnet for hver enkelt kontakt. Disse er valgt på baggrund af den umiddelbare definition af en flirt i Afsnit 1.1, samt på baggrund af de kriterier der er nævnt i Afsnit 2.1.

- Antal dage mellem første aktivitet og sidste aktivitet (aktivitetsperioden).
- Det gennemsnitlige antal sms'er sendt pr. dag på de dage hvor der har været aktivitet. Der er altså ikke tale om et gennemsnit over hele aktivitetsperiode, men i

stedet et gennemsnit over de dage hvor der rent faktisk er aktivitet. Årsagen til at målet er beregnet således, og ikke er beregnet ud fra den fulde aktivitetsperiode, skyldes databruddet i sms-strømmen, som tydeligt fremgår af Figur 3 i Afsnit 3.
- Variansen af antal sms'er sendt pr. dag i hele den aktive periode.
- Det gennemsnitlige daglige antal opkald. I modsætning til det gennemsnitlige antal sms'er er dette et gennemsnit over hele den aktive periode.

Til sidst dannes en variabel der fortæller noget om min relation til 83 af de i alt 248 kontakter. Her kategoriseres kontakterne på baggrund af hvordan jeg definerer mit forhold til dem (Det kan tænkes, at hvis kontakterne selv skulle definere relationen, at de havde defineret den anderledes). Denne variabel skal anvendes til at tjekke om clusteranalysen gør et godt stykke arbejde.

4 Resultater

Den endelige clusteranalyse udføres ved anvendelse af SAS® Visual Statistics softwaren[3]. *K*-means algoritmen benyttes, og herigennem fordeles de 248 kontakter (svarende til 27,65% af alle kontakter i det oprindelige datasæt) i fire forskellige clustre. Fordelingen af kontakterne er vist i Tabel 1. Her ses det, at cluster 1 er klart det største cluster med i alt 138 observationer, mens cluster 0 og 3 kun indeholder hhv. 24 og 19 observationer. Cluster 2 indeholder i alt 67 af de 248 observationer.

Tabel 1 indeholder yderligere seks nøgletal, der giver en indsigt i hvert cluster. Et hurtigt kig på tabellen viser, at cluster 3 klart er det cluster, der skiller sig mest ud. Udover at være det mindste cluster, har det også den klart højeste værdi af *RMS af STD*. Dette er et mål for den geometriske gennemsnitlige afstand mellem observationerne i hvert cluster. Generelt ligger observationerne i cluster 3 derfor mere spredt, hvilket også er tydeligt ud fra *Within-Cluster SS*, svarende til kvadratsummen af de tilhørende observationer inden for hvert cluster bestemt i Ligning (1) i teoriafsnittet. De to tal, *Min. afstand* og *Max. afstand* angiver hhv. den mindste og den største afstand fra centroiden til en observation i clusteret. Igen er det cluster 3 der adskiller sig mest ved at have observationer der ikke ligger meget tæt ved centroiden. *Nærmest cluster* viser tallet for det cluster hvis gennemsnitsværdi er tættest på det betragtede cluster. Både cluster 0 og 1 minder på baggrund af dette mål mest om cluster 2, mens gennemsnitsværdien for cluster 2 er tættest på den for cluster 0. Den sidste kolonne i Tabel 1 vuser den gennemsnitlige afstand mellem clustercentroiderne og de tilhørende observationer. Igen er det cluster 3 der skiller

[3] Alle figurer i denne artikel er lavet i SAS Enterprise Guide 7.1 pga. det sort-hvide format.

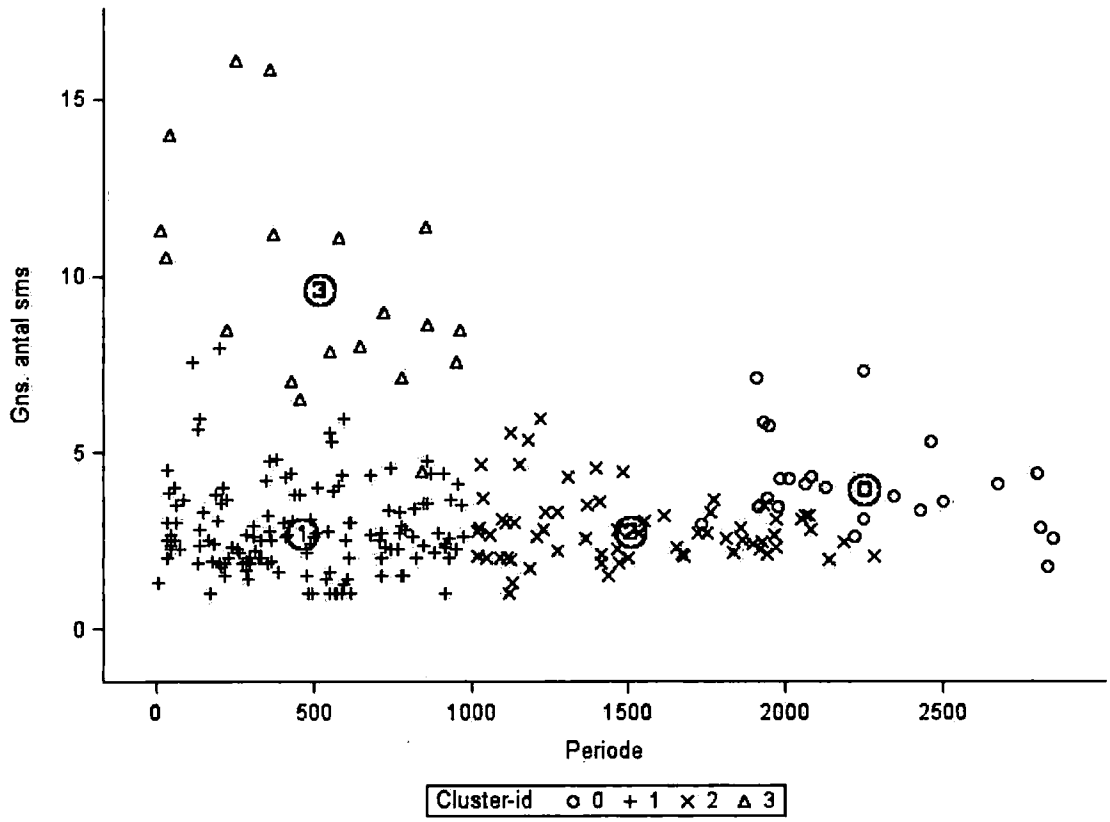
sig ud ved at have en utrolig høj gennemsnitsafstand. Cluster 1 derimod har generelt observationer der ligger meget tættere på centroiden.

Tabel 1: Clusteroversigt

Cluster-id	Obs	RMS af STD	Within-Cluster SS	Min. afstand	Max. afstand	Nærmeste cluster	Centroide-distance
0	24	0,643	38,003	0,422	3,135	2	1,382
1	138	0,417	95,325	0,099	3,805	2	0,538
2	67	0,371	36,408	0,104	1,735	0	1,382
3	19	2,049	302,295	1,213	8,974	1	4,136

Clustre kan også sammenlignes på baggrund af deres opførsel i forhold til inputvariablene. De fire inputvariable kan overføres til en todimensional projektion af hvert cluster i forhold til to udvalgte inputvariable. Disse plot er gode når man leder efter ligheder eller forskelle på clustre. Nedenfor beskrives to af de i alt seks mulige kombinationer af de fire variable.

Figur 4: Plot af inputvariablene *Periode* og *Gns. antal sms*

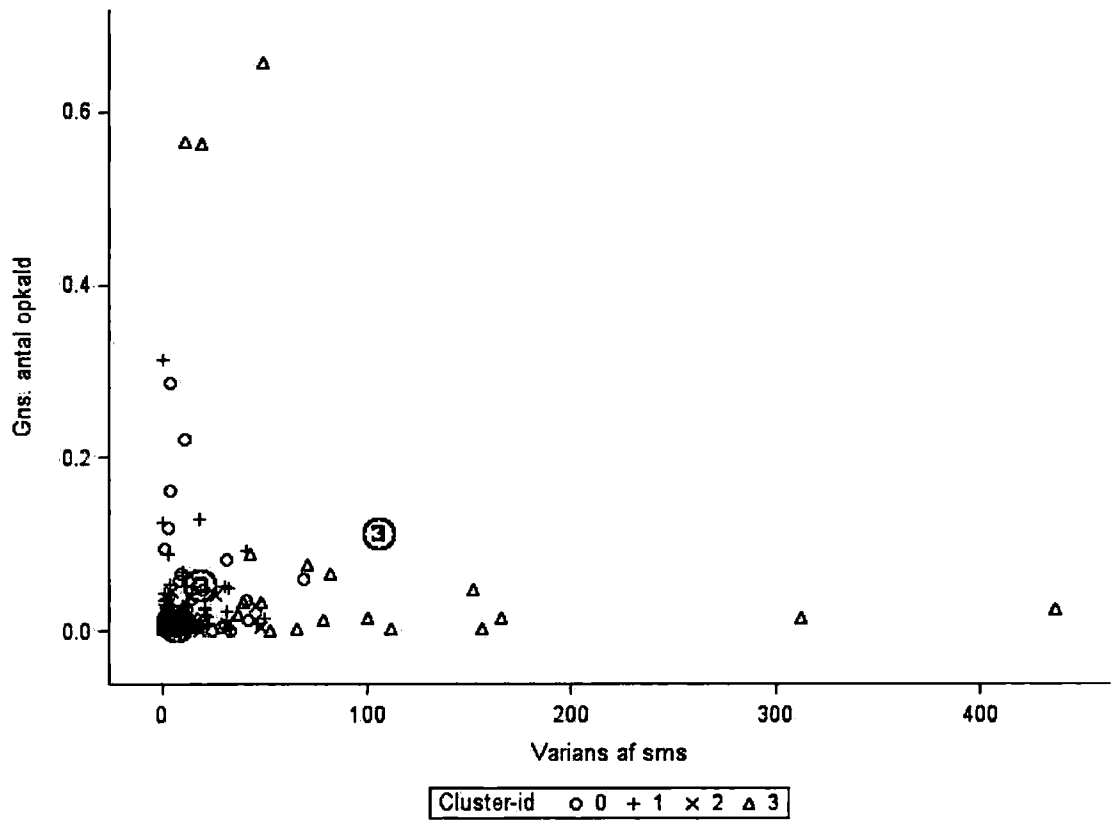


Anm.: Cirklerne med numre angiver clustercentroiderne for hvert af de fire clustre

I Figur 4 er det tydeligt at se, at der er forskel på de fire clustre. Først og fremmes adskiller cluster 3 sig især fra de tre andre da det gennemsnitlige antal sms'er sendt på aktive dage er markant højere. Kontakterne i cluster 3 modtager altså i gennemsnittet flere sms'er pr. dag. Til gengæld er aktivitetsperioden for både cluster 1 og 3 ikke særlig lang i forhold til cluster 2 og især cluster 0. Cluster 0 indeholder kontakter der får gennemsnitligt lidt flere sms'er end kontakterne i cluster 1 og 2, samt en lang aktivitetsperiode.

De sidste to inputvariable er hhv. variansen af det gennemsnitlige antal sms'er samt det gennemsnitlige antal opkald pr. dag over hele aktivitetsperioden. Disse er plottet mod hinanden i Figur 5.

Figur 5: Plot af inputvariablene *Varians af sms* og *Gns. antal opkald*



Anm.: Cirklerne med numre angiver clustercentroiderne for hvert af de fire clustre

Igen kan det ses at cluster 3 adskiller sig iøjefaldende meget fra de 3 andre clustre idet både sms-variansen samt det gennemsnitlige daglige antal opkald ligger højere for disse

kontakter. I forhold til det gennemsnitlige antal opkald for cluster 3, skyldes det primært 3 observationer, der trækker gennemsnittet gevaldigt op. En eksklusion af disse ville reducere værdien af clustercentroiden til et niveau svarende til de tre andre clustre, og endda også til et gennemsnitligt antal opkald lavere end for cluster 0.

Det er svært, ud fra Figur 5, at sige særlig meget om cluster 1 og 2, udover at både varians og gennemsnitlig antal opkald ligger lavt i forhold til cluster 0 og i særdeleshed i forhold til cluster 3.

4.1 Karakteristik af clustrene

Ud fra resultaterne beskrevet ovenfor er det muligt at karakterisere de fire clustre, og forsøge at finde ud af, om et af disse kunne passe på definitionen af en flirt fra Afsnit 1.1.

Tabel 2: Karakteristik af clustre

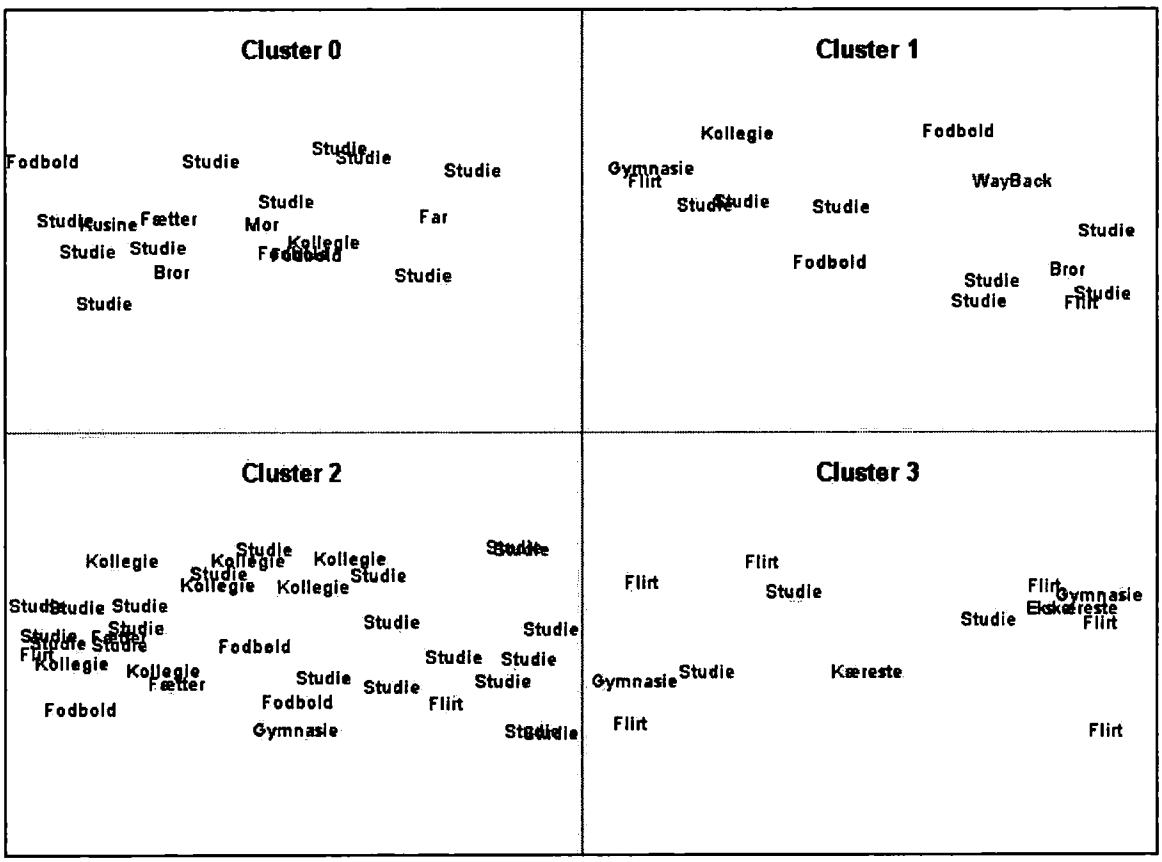
Varlabel	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Periode	Lang	Kort	Middel	Kort
Gns. sms	Lav+	Lav	Lav	Høj
Varians	Middel	Lav	Lav	Høj
Gns. opkald	Middel	Lav	Lav	Høj
Type	Familie	Bekendte	Venner	Flirt

Tabel 1 opsummerer resultaterne observeret i Figur 4 og Figur 5. Cluster 0 karakteriseres som familie, da aktivitetsperioden stort set strækker sig ud over alle årene hvorfra der er data. Cluster 1 og 2 kategoriseres som hhv. bekendte og venner, mens cluster 3 passer med den forventede definition af en flirt.

Dette stemmer godt overens med resultaterne vist i Figur 6. Her ses fordelingen af de 83 kontakter der er blevet tildelt en relation, på de fire clustre. Det ses at cluster 0 indeholder både mor, far og en bror samt et par andre familiemedlemmer. De studie- og kollegie- samt fodboldkammerater der falder i denne gruppe er nogle jeg har haft kontakt til i mange år og nogle af mine rigtig gode venner. Cluster 1 er flygtige bekendtskaber, som der ikke har været særlig meget telefonisk kontakt med. Her ses bl.a. at to kontakter, som jeg ellers havde defineret som flirts, ender i dette cluster. Aktiviteten med disse to flirts har altså været meget lille. Desuden falder min ene bror også i dette cluster, hvilket kan skyldes at han først for nylig er begyndt at bruge sin mobiltelefon. Cluster 2 indeholder en masse kontakter fra kollegie og studie, og viser sig at være gode venner, og især tidligere rigtig gode venner, som jeg af den ene eller anden grund nu har mistet kontakt til.

I cluster 3 falder over halvdelen af de kontakter, som jeg selv havde kategoriseret som flirts. Desuden er både kæreste og ekskæreste røget i dette cluster, som begge på et tidspunkt har været flirts, og som af en eller anden årsag ikke er røget over i det mere familiære cluster, cluster 0. Dette kan skyldes at aktivitetsperioden er relativt kort, samt at antallet af sms'er er højt.

Figur 6: Fordeling af kontakter i clustre efter relation

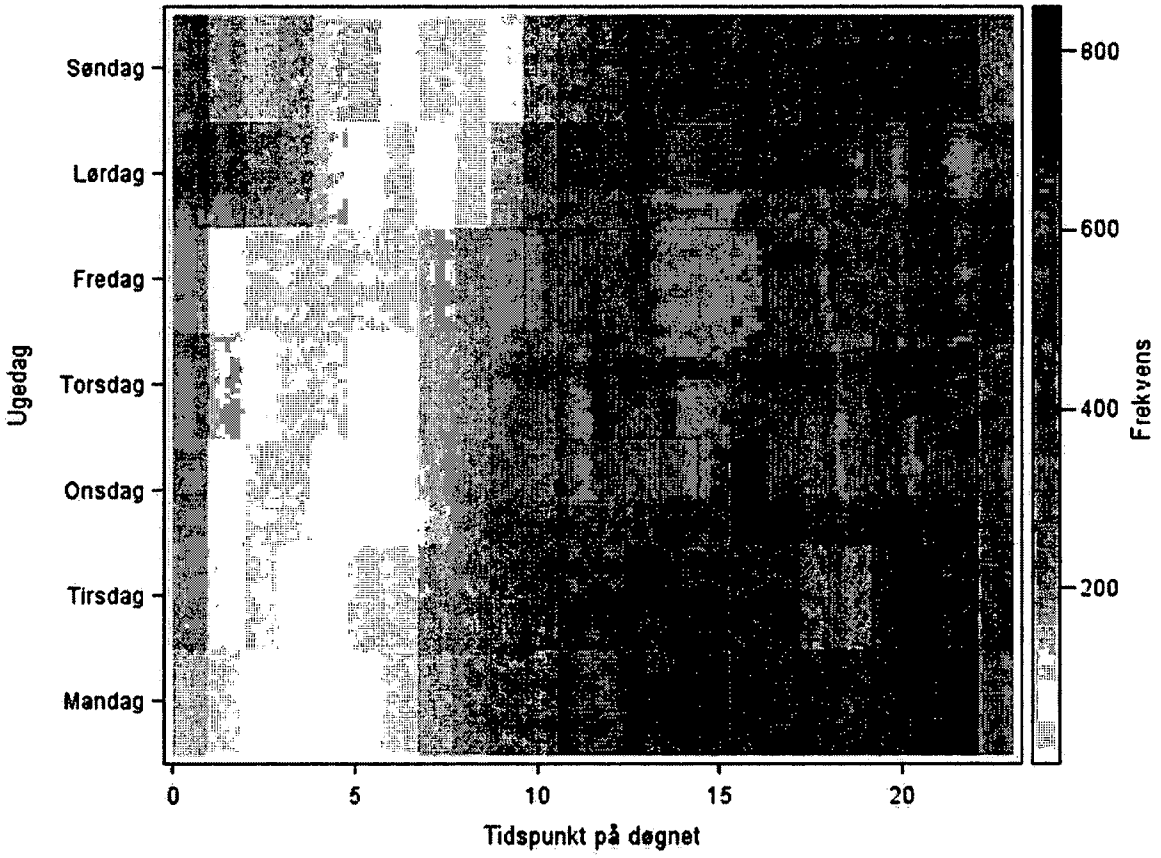


Efter en inspektion af Figur 6 har jeg erkendt, at nogle de kontakter der indgår i cluster 3, som jeg umiddelbart ikke selv havde kategoriseret som flirts, måske er fejlkategoriserede. De har måske i virkeligheden været en form for flirts. Hvis alle 248 kontaktpersoner havde fået tildelt en relations-værdi, ville billedet i Figur 6 nok fremstå endnu tydeligere.

4.2 Kontakttidspunkt

Det kunne tænkes, at det for de fire forskellige clustre er muligt observeres forskellige adfærdsmønstre for andre variable end inputvariablene. Eksempelvis kan aktiviteten være afhængig af ugedag, samt tidspunkt på døgnet.

Figur 7: Frekvensen af aktiviteter i løbet af ugen og døgnet



Figur 7 viser frekvensen af aktivitet (indikeret af farveskalaen), fordelt på ugedag og tidspunkt på døgnet for samtlige 248 kontakter, da det viser sig, at der ikke er den helt store forskel clustrene imellem. Ud fra Figur 7 kan det ses at der generelt ikke er lige så meget aktivitet i nattetimerne fra midnat og frem til omkring kl. 7-8. Til gengæld kan det ses at der er mere aktivitet både natten til lørdag og søndag, men at der til gengæld også generelt går lidt længere tid før der er aktivitet igen lørdag og søndag morgen.

5 Konklusion

Mobiltelefonfakturaer fra en 8-årig periode blev analyseret ved anvendelse af clusteranalyse for at få et indblik i mobiladfærden. Det blev undersøgt, om det på baggrund af et særligt aktivitetsmønster kunne bestemmes hvilken type kontaktperson der var tale om. Helt konkret blev det undersøgt, om det er muligt, ud fra udgående mobiltelefonaktivitet at finde frem til hvilke kontakter der i perioden har været flirts.

Den udegående mobilaktivitet observeres, og der viser sig en tydelig forskel i adfærden overfor forskellige kontaktpersoner. I alt fordeles de 248 kontakter i fire forskellige clustre, hvor især et cluster skiller sig ud fra de andre. På baggrund af centroideværdierne for inputvariablene for clusteret, kategoriseres dette som flirte-clusteret. Dette skyldels at aktivitetsperioden er relativ kort, at det gennemsnitlige antal sms'er sendt pr. dag, samt variansen af denne er meget høj. Det eneste der måske ikke er helt så intuitivt er, at det gennemsnitlige antal opkald pr. dag er rimelig højt i forhold til de andre clustre. Årsagen til dette er, at få enkelte observationer trækker denne gennemsnitsværdi meget op. Disse enkelte observationer er kontakter, hvor forholdet har udviklet sig til at være lidt mere end bare en flirt – i to af tilfældene har flirten udviklet sig til et faktisk kæresteforhold.

For at kunne bruge analysens resultater på fremtidig data over mobilaktivitet, bør datasættet opdeles i hhv. trænings-, test- og valideringsdata, da modellen pt er baseret det fulde datasæt, og dermed er tilpasset den nuværende datastruktur. Generelt bør analysen også komplimenteres af andre former for analyse, da cluseranalysen primært er et godt indledende skridt mod en større prædiktiv modellering.

Så pas på med hvordan du bruger din telefon, og i særdeleshed, hvem der får lov til at betale din telefonregning. Det er nemlig let at observere din adfærd og fra den regne ud hvad du i virkeligheden tænker om kontaktpersonerne på din mobiltelefon.

Referencer

Agis, G. (2015). *K-means clustering with Apache Mahout*, Blog.guillaumeagis.eu, Tilgået 1. januar 2017 fra http://blog.guillaumeagis.eu/k-means-clustering-apache-mahout/

ClimServ (2007). *K-means (centroid) cluster analysis*, IDL Online Help, Tilgået 1. januar 2017 fra http://climserv.ipsl.polytechnique.fr/documentation/idl_help/IMSL_K_MEANS.html

Ravenna, A., Truxillo, C. & Wells, C. (2015). *SAS® Visual Statistics: Interactive Model Building Course Notes*, Cary, NC: SAS® Institute Inc.

Sørensen, L.M. (2013). *Mobiltelefonen fylder 40 år og soler sig i popularitet*, dr.dk, Tilgået 1. januar 2017 fra http://www.dr.dk/nyheder/kultur/medier/mobiltelefonen-fylder-40-aar-og-soler-sig-i-popularitet

Anders Milhøj
Nyheder i SAS Analytics 14.2

Department of Economics, University of Copenhagen
Øster Farimagsgade 5, DK-1353 København K
Anders.Milhoj@econ.ku.dk

I november 2016 blev Analytical Products i den opdaterede version 14.2 sendt på markedet. Denne opdatering indeholder som de mange tidligere opdateringer af de analytiske programpakker inden for statistik, økonometri, operationsanalyse etc. Disse opdateringer er nu løsrevet fra samtidige opdateringer af det samlede SAS-program, så det er stadig Base SAS, version 9.4, der anvendes. Desuden er den gratis SAS-applikation SAS-U opdateret; især er det bemærkelsesværdigt, at der via SAS-U nu også stilles procedurer til rådighed for avancerede ikke-modelbaserede tidsrækkeanalyser.

SAS's nyere analytiske releases

Version 9.4 med Analytical updates 14.1, som udkom sommeren 2015, blev ca 1. december 2016 opdateret til version 14.2. Her omkring Nytår er 14.1 stadig den aktuelle for universitetsansatte og studerende på fx sasdownload.dk, men distributionen til studerende opdateres meget snart til 14.2.

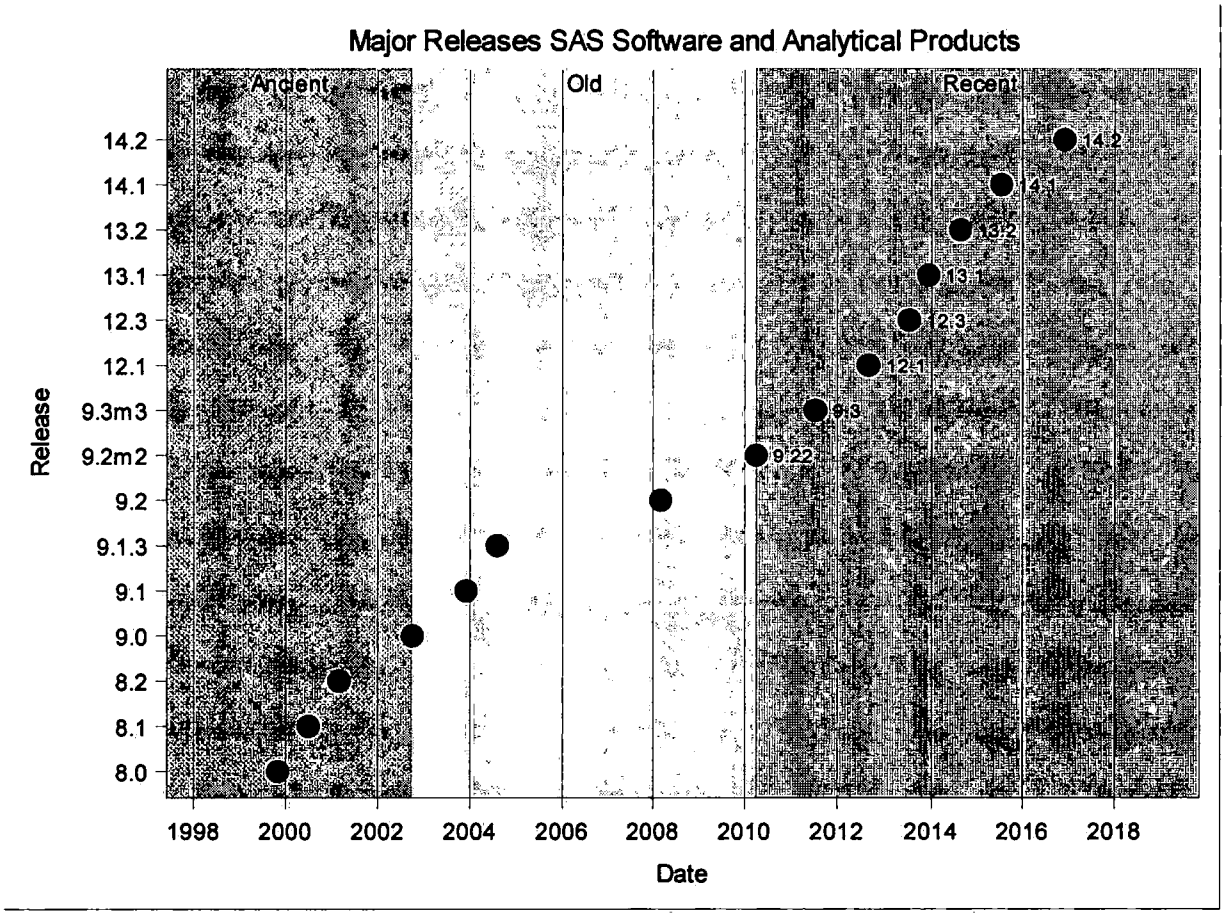
På en SAS blog,

http://blogs.sas.com/content/iml/2013/08/02/how-old-is-your-version-of-sas-release-dates-for-sas-software/

findes der en oversigt over Analytical updates indenfor de sidste år. Det er selvfølgelig i form af et SAS program, der som output viser en graf, som vist her, idet jeg har tilføjet de nyere opdateringer siden 2013. Den lodrette akse er en smule misvisende, da fx springet fra 12.1 til 12.3 udelukkende var vedligeholdelse "maintenance".

I symposieindlægget januar 2016 gennemgik jeg visse af nyhederne i 14.1. I dette indlæg fokuseres på version 14.2, som indeholder en del flere nyheder, end en ændring på første decimal ellers skulle antyde.

Kilderne til disse nyhedsoversigter er er SAS-hjælpen, som kan tilgås af alle - også uden en SAS-installation - via http://support.sas.com/, idet manualerne for SAS-pakkerne STAT, ETS. OR, QC er offentligt tilgængelige for alle.



De mange nyheder kan ses ved at følge linkene:

SAS/ETS 14.2: http://documentation.sas.com/#/?cdcId=etscdc&cdcVersion=14.2

SAS/IML 14.2: http://documentation.sas.com/#/?cdcId=imlcdc&cdcVersion=14.2

SAS OPTGRAPH Procedure 14.2:
http://documentation.sas.com/#/?cdcId=procgralgcdc&cdcVersion=14.2

SAS/OR 14.2: http://documentation.sas.com/#/?cdcId=orcdc&cdcVersion=14.2

SAS/QC 14.2: http://documentation.sas.com/#/?cdcId=qccdc&cdcVersion=14.2

SAS/STAT 14.2: http://documentation.sas.com/#/?cdcId=statcdc&cdcVersion=14.2

men det er jo nok lettere at Google.

Øget tilgængelighed til SAS

Der blev for et par år siden udviklet et nyt interface til SAS kaldet SAS Studio. Det virker som en kombination af det bedste fra en traditionel kode-SAS tilgang og så den fantastiske editor fra Enterprise Guide.

Desuden er der frigivet en "University Edition", som stilles gratis til rådighed for "alle", der påstår, at de studerer ved et universitet. Den markedsføres som SAS-U. Håbet er, at SAS på den måde kan komme længere "ned" i undervisningssystemet, så allerede high school elever, altså gymnasieelever, kan stifte bekendtskab med de mange muligheder i SAS.

Det bedste ved SAS-U er, at den er udviklet som "virtuel applikation", der kan køres på langt flere platforme end det traditionelle SAS-system. Især kan det uden videre anvendes på en Mac! Det er jo en mulighed, der har været efterspurgt længe i universitetsverdenen, hvor Mac's markedsandel kan være over 50%. Men i det store hele virker SAS-U på samme måde som SAS Studio.

SAS-U kan downloades fra

http://www.sas.com/da_dk/software/university-edition.html

Der kræves en virtualization software pakke, se

http://www.sas.com/da_dk/software/university-edition.html#m=system-requirements

Som hovedregel går installationen af SAS-U nemt; den meste tid går med at downloade selve SAS-U, som er en fil på knapt 2 Gb. Den skal blot ligge et sted på brugerens harddisk. Når den først er downloaded skal den tilknyttes den virtuelle boks. Her har jeg oplevet, at visse studerende har haft problemer med at få tilordnet deres egne filreferencer korrekt. Moralen er, at man skal gøre hvad der står i vejledningen - bestemt IKKE, hvad man tror, der står.

En væsentlig ulempe er, at kun SAS-BASE, STAT pakken og IML (matrix-regning) er med i SAS-U. Det skal dog understreges, at selv de esoteriske dele af disse pakker er med i SAS-U, men dog ikke High Performance procedurerne med præfixet HP..., som omtales senere.

Desuden er dele af ETS (økonometri og tidsrækkeanalyse) pakken inkluderet fra og med sommeren 2014. Det er de dele af ETS, der kan betegnes som "data-scientist" procedurer i modsætning til videnskabelige, professionelle procedures. Det betyder, at PROC ESM til forudsigelser og PROC UCM, som er en del af indholdet i vores symposieindlæg dette år om salget af specialøl forudsagt ud fra social media data, er medtaget i SAS-U. Derimod er fx PROC VARMAX og i 14.1 PROC X13 ikke medtaget.

SAS-U afvikles som en applikation i en internet browser. Det betyder, at filer på harddisken på brugerens egen PC skal tilgås på en anden måde en normalt via "delte filer" i den virtuelle boks; men svært er det ikke.

En yderligere ulempe er, at SAS-U i praksis ikke kan håndtere store datamængder med titusindvis af observationer med hundredevis af variable. Det skyldes selvfølgelig, at tilgangen via en internetbrowser ikke kan optimere adgangen til CPU og RAM på samme effektive måde som en traditionel SAS-installation. Et eksempel med den danske del af PISA undersøgelsen med knapt 8000 observationer af 738 variable kører dog uden problemer, men en analyse af det samlede PISA materiale er i praksis umuligt.

SAS og Big Data

Listen over HP (High Performance) procedurer, som kopierer de gængse SAS procedurer, udvides stadigt. De indeholder ikke de mange grafiske faciliteter, for man kan jo ikke tegne millionvis af punkter i et diagram. Men de regner hurtigt, og de udnytter maskinfigurationen fuldt ud. Fx kan de regne multithreadet, hvis maskinen indeholder flere processorer. Det nye fra og med version 9.4 er, at de også stilles til rådighed for almindelige SAS brugere i simple PC installationer. Derved kan visse ekstra raffinementer udnyttes af alle som fx, at der i PROC HPREG, men ikke i PROC REG, stilles en CASS statement til rådighed. Desuden kan programmer afprøves på egen PC før de sendes til produktionskørsler på fjerne servere med stor maskinkraft. Det kræver specielle licenser at udnytte faciliteterne til distribueret kørsel af SAS-programmer.

Disse muligheder for analyser af Big data virker noget fjerne for an almindelig universitetsansat eller student. Men nogen skulle arbejde for at forskere fik mulighed for at afprøve mulighederne i praksis, fx på forskermaskinerne i Danmarks Statistik, hvor datamængderne kan være betydelige.

SAS Viya

I 2016 er en ny form for SAS annonceret med det lidt besynderlige navn SAS Viya. Ideen er at tilbyde SAS's forskellige komponenter på de platforme, hvor data og problemerne befinder sig; i "skyen" eller på engelsk "in the cloud", hvor det så end er. Der er også en øget integration med andre programmeringssprog.

Fx gives der en ramme for fra sin mobiltelefon at kunne afvikle SAS kørsler med data på flere forskellige gigantiske datasæt fordelt på flere servere verdenen over. Data kan være i realtid, så statistiske analyser kan indgå i et løbende workflow. Desuden udnyttes regnekraften på de enkelte servere udnyttes optimalt, ved at data ikke flyttes væk fra den server, de ligger på, medmindre det er strengt nødvendigt. Viya kan tilgå data på "public REST ATIs" dvs på databaser, som stiller deres interne datastruktur til rådighed for udefra kommende brugere.

En del af fleksibiliteten er, at SAS via Viya kan tilgås fra andre cloudbaserede programmer som Python, Java og Lua. Disse programmer er i høj grad open-acces, og der findes mange yderst avancerede programmer til fx tekstgenkendelse, og der kommer nyt i langt kvikkere tempo end SAS's mere veldokumenterede procedurer. Pointen er altså, at de mange mere kvalitative analyser, som ofte foregår i Python, kan udnytte

SAS's styrke inden for datahåndtering og statistiske analyser. Omvendt kan SAS's egne styrker let kombineres med styrkerne i disse øvrige programmer.

Visse dele er tilgængelige nu, mens andre er annonceret til at udkomme i løbet af 2017/18. Interesserede kan følge med på sas.com/viya.

Data scientist

IT-Universitetet i København har i samarbejde med bl.a. SAS fået godkendt sin ansøgning om oprettelse af ny bacheloruddannelse i Data Science. Der optages 50 sommeren 2017, og på sigt skal der optages 100 hvert år.

Uddannelsen skal ikke officielt ses som en konkurrent til fx statistikuddannelserne, da statistikindholdet ifølge studieplanen for den nye uddannelse er yderst begrænset; tilsyneladende kun et kvart år ud af de tre år på en bacheloruddannelse. De eksisterende mere matematiske dataanalyse uddannelse får følgende ord med på vejen:

På tværs af uddannelser inden for matematik tegner sig et billede af uddannelser, der opfylder de grundlæggende krav til en data scientists' kompetencer inden for avanceret matematisk forståelse. Dog er uddannelserne så specialiserede, at den studerende får matematiske kompetencer, der er mindre relevante for en data scientist. På flere af uddannelserne er der mulighed for at supplere matematikken med datalogiske og teknologiske fag, hvorved den studerende bedre opnår en profil som data scientist. Uddannelserne giver dog kun begrænsede kompetencer til at arbejde med ustruktureret data, hvilket er en af de centrale arbejdsopgaver inden for data science, og derudover er kompetencer inden for forretningsforståelse samt formidling af data næsten ikkeeksisterende på uddannelserne.

Det lyder jo i første omgang som om, anvendt statistik, som dette symposium omhandler, ikke bliver så vigtigt i fremtiden. På samme linie er Malou Aamund og Olivia Simson Aamund, der i en kronik i dagbladet Børsen 31/12-2016 priser den nye uddannelse på ITU, idet de bl.a. skriver: *Det giver for eksempel ikke mening at uddanne unge mennesker i færdigheder, som teknologi kan udføre langt bedre.*

På den anden side skriver de dog også: *Fremtidens arbejdsmarked vil nemlig efterspørge lavtuddannet arbejdskraft med sociale kompetencer og dømmekraft samt højtuddannede med dyb kundskab inden for teknologi, kompleks problemløsning og innovation.* Da anvendte statistikere vist ikke er specielt kendte for deres sociale kompetencer og dømmekraft, er der altså trods alt håb for os anvendte statistikere.

STAT nyt

Der er kommet to nye procedurer, der begge kan bruges til beregninger af behandlingseffekter. Man kan sige, at de to procedurer på forskellige måder tilgår det samme problem, nemlig at der i mange analyser er forskelle mellem behandlede og ubehandlede individer, når det gælder en lang række andre variable. Disse andre relevante variable, også kaldes confounding variable, medfører let misvisende resultater, hvis der ikke tages højde for dem i analyserne.

PROC CAUSALTRT beregner effekten af en behandling variabel på en kontinuert eller en diskret resultatvariabel. Fx kan der kompenseres for, at visse sygdomsbehandlinger kun tilbydes patienter, for hvilke man har en forventning om, at behandlingen virker; så er det jo ikke så underligt at de behandlede patienter klarer sig bedre end de ubehandlede. Men når graden af fx tilstødte komplikationer inddrages i analysen, er behandlingen knapt så attraktiv.

Begrebet behandling, bogstaverne treatment TRT i procedurenavnet, kan dække over meget andet. Fx er det ikke samme typer børn, der går i privatskole som i folkeskole; der er forskelle i deres socioøkonomiske baggrundsvariable, så effekten for indlæring af at gå i privatskole kan ikke umiddelbart beregnes. Visse (mandlige virksomhedsejere) hævder, at lønforskellene mellem mænd og kvinder udelukkende skyldes forskelle i branchevalg, arbejdstid og uddannelsesniveau. Men med PROC CAUSALRTRT kan der tages højde for, at et øget uddannelsesniveau måske øger mænds løn mere end det øger kvinders løn. Resultaterne af analyserne videreformidles både i diverse tal, fx gennemsnitlige behandlingseffekter (dvs fx lønforskelle) og i en række intuitive grafer, som fx boxplots.

PROC PSMATCH indeholder en række værktøjer til propensity score analyser. Denne procedurer matcher efter bedste evne de behandlede individer med ubehandlede individer. Fx kan regnskovsarealer, der er fredet, matches med regnskovsarealer, der ikke er fredet på en måde så de matchede arealer er ens i andre henseender. Fx er det et problem, at fredede arealer typisk ligger på bjergsider eller langt fra al transportinfrastruktur, så der ikke kan ryddes skov på dem, selvom de ikke var blevet fredet. Dernæst kan man så statistisk i en model for parvise observationer, men så sandeligt også mere kvalitativt, undersøge forskellen mellem skovbevarelsen i de fredede arealer og det matchede ikkefredede areal. Uden denne matching vil fredning virke meget skovbevarende, da skoven på det fredede areal, som regel ikke umiddelbart kan ryddes. Ved direkte at udpege et matchet individ, er der langt bedre muligheder for konkrete opfølgende undersøgelser og analyser, end hvis det samme problem var løst ved hjælp af fx PROC CAUSALTRT.

De to procedurer PROC CAUSALTRT og PROC PSMATCH demonstreres i det mundtlige indlæg ved eksempler fra den danske del af den nyeste PISA undersøgelse. Spørgsmålet er om motion styrker kompetencerne inden for naturvidenskab.

Desuden er der følgende væsentlige forbedringer i andre procedurer i SAS/STAT 14.2:

Procedurerne FREQ og SURVEYFREQ giver mulighed for endnu flere output tabeller.

Proceduren NLIN giver nu mulighed for estimation af funktioner af parametrene i en ny ESTIMATE statement. I en CONTRAST statement kan der nu testes multiple hypoteser om parametrene.

Proceduren PHREG giver mulighed for tidsafhængige ROC analyser.

I PROC POWER er der flere muligheder styrkeberegninger i generaliserede lineære modeller.

SURVEYIMPUTE proceduren er udvidet med to-trins fully efficient fractional hot-deck imputering.

SURVEYSELECT inkluderer nu også balanceret bootstrap udvælgelse og sekventiel Poisson udvælgelse.

ETS Nyt

I SAS/ETS 14.2 er der introduceret en ny procedure, PROC SPATIALREG. Den analyserer økonometriske modeller for tværsnitsdata, hvor observationerne kan være geografisk korrelerede. Det sker ofte, at observationer geografiske naboer, fx personer der bor tæt ved hinanden, er korrelerede, også selvom der inddrages alle tilgængelige informationer i baggrundsvariablene i fx en regressionsanalyse. Disse korrelationer ødelægger forudsætningerne for de statistiske tests. I visse tilfælde kan den geografiske sammenhæng være af selvstændig interesse, og PROC SPATIALREG indeholder avancerede geografiske modeller, der er konstrueret som generalisering af tidsrækkemodeller, fx de spatial autoregressive moving average (SARMA) modeller.

PROC SPATIALREG demonstreres også ved et eksempel om socioøkonomiske forhold københavnske roder i det mundtlige indlæg.

Desuden er der opdateringer til følgende SAS/ETS procedurer:

HPCDM proceduren

HPSEVERITY proceduren

QLIM proceduren

SEVERITY proceduren

SSM proceduren

TIMESERIES proceduren

Det nye RANDOM statement i PROC QLIM gør det muligt at estimere modeller med stokastiske parametre modeller ud over modeller med tilfældige effekter. Man kan desuden modellere heterogenitet med stokastiske effekter i enkeltligning modeller fx i modeller for panel data, hvor der er parameterheterogenitet mellem enhederne.

I PROC VARMAX er der sket en del. Vektor autoregressive fractional integratet glidende gennemsnit (VARFIMA) er også understøttet - endda med mulighed for eksogene variable (VARFIMAX).

IML nyt

Grænsefladerne mellem SAS/IML 14.2 og R er opdaterede til de seneste versioner af R 3.3.x.

QC nyt

SAS/QC 14.2 omfatter forbedringer af ANOM, CAPABILITY, CUSUM, MACONTROL, RAREVENTS, and SHEWHART procedurerne.

OR nyt

SAS/OR 14.2 inkluderer performance forbedringer i algoritmerne til LP, MILP, NLP problemerne og til netværksanalyser.

OPTGRAPH

Der er desuden kommet en ny OPTGRAPH procedure, der stiller algoritmer inden for grafteori, kombinatorisk optimering og netværksanalyse til rådighed. En oversigt er givet i følgende opremsning

Biconnected components
Centrality metrics
Maximal cliques
Community detection
Connected components
Core decomposition
Cycle detection
Eigenvector problem
Weighted matching
Minimum-cost network flow
Minimum cut
Minimum spanning tree
Reach networks
Shortest path
Graph summary
Transitive closure
Traveling salesman

Men licensen fulgte desværre ikke med OR licensen, så jeg har ikke afprøvet mulighederne endnu.

Daily eating activity of dairy cows from 3D accelerometer data and RFID signals: prediction by random forests and detection of sick cows

Leslie Foldager[1,2], Lars Bilde Gildbjerg[1], Heidi Voss[1], Philipp Trénel[3], Lene Munksgaard[1], and Peter T. Thomsen[1]

[1]Department of Animal Science, Aarhus University, DK8830 Tjele. Email: leslie@anis.au.dk
[2]Bioinformatics Research Centre, Aarhus University, DK8000 Aarhus C
[3]AgroTech, Danish Technological Institute, DK8200 Aarhus N

Abstract

Feed intake is very important for dairy cows and deviation from normal eating behaviour may predict a cow that needs treatment. We used video recordings of dairy cows at the Danish Cattle Research Centre (DKC) combined with data from a neck-collar mounted 3D accelerometer and RFID device from Lyngsoe Systems (Aars, Denmark) to develop a random forests model for predicting daily eating activity. We investigated performance by internal cross–validation and the results indicate that we obtain accurate predictions of daily eating time by the algorithm. Technical challenges are delaying the planned tests on commercial farms. We are therefore currently utilising historical data from DKC to examine the potential of using changes in daily eating time for detection of sick cows.

1 Introduction

Feed intake is very important for dairy cows not only to obtain high milk yield but also to stay healthy. Reduced daily eating time is associated with various diseases and deviation from normal eating behaviour may therefore predict a cow that needs treatment. In modern dairy production, loose housing systems are used and all cows eat from the same feed bunk. Thus from a management perspective, it is fairly easy to keep track of the amount of feed being placed in the feed bunk whereas obtaining knowledge of feed intake at the level of the individual cow is by far more complicated. Nevertheless, many

dairy farmers now take advantage of automatic sensor–based systems to collect massive amounts of data at the cow level. The systems are measuring and monitoring behavioural and production characteristics such as daily activity, number of steps, number of lying bouts, duration of rumination, social activity (via positioning), milk yield, milk composition, somatic cell counts in the milk, and much more. The data may be used to predict certain states such as heat (the short period of the estrous cycle at which insemination should be timed) and diseases or injuries such as lameness. These predictions may guide the farmer to perform management procedures, e.g. insemination or disease treatments. In the present study, we set out to investigate if deviation from normal eating behaviour can predict sick cows.

Lyngsoe Systems (Aars, Denmark) developed a neck–collar mounted 3D accelerometer and radio frequency identification (RFID) sensor device (CowTrack Logger), and a low frequency (LF) field generator (CowTrack Exciter Box) using a single–wire electric loop. This sensor system recorded activity (acceleration) and positioning of the cow at the feed bunk (RFID signal) from which we planned to measure eating time, number of visits at the feed bunk, and daily activity. For the first prototype of the logger, data was manually transferred to a computer via a detachable memory stick. The second prototype data was designed to broadcast data regularly by wireless transmission. As with many electronic devices such as cellular phones, sport watches with GPS tracking and heart rate monitor, and wireless equipment in general, lifetime of the sensor battery is an issue. Optimally, batteries should hold the charge for the lifetime of the animal. Obviously, this will depend on factors like the amount and complexity of calculations done on the sensor, physical size of the sensor and thus the possible size and number of batteries, the technology used for transmission of data, and the amount of data transmitted.

The study was planned to consist of two phases. In Phase 1, the first prototype of the sensor should be tested in a research herd on a limited number of animals for a shorter duration, and a statistical algorithm for estimation of daily eating time from the sensor data should be constructed. In Phase 2, a second prototype should be tested on a larger scale (more animals and longer duration) in a few commercial herds. Moreover, the plan was to investigate if deviations from daily eating activity as determined by the algorithm from Phase 1 can predict cows requiring treatment, and by these means develop methods supporting detection of sick cows.

Due to technical challenges delaying the planned tests in commercial herds, the second phase was recently revised into two subtasks: Phase 2a) solving technical challenges related to the use of the logger in commercial settings; Phase 2b) utilising historical data to examine the potential of using changes in daily eating time for detection

of sick cows. Regarding Phase 2a, the main challenge is that the technology chosen has turned out inappropriate for larger herds. Cut short, the time slot for transmission of data from the individual logger is too short. This results in loss of data and gaps when the onboard memory of the logger is full.

The present paper mainly concerns the construction of a random forests model for estimation of daily eating time in Phase 1 used for monitoring behavioural changes in Phase 2. Performance and influence of changing parameters of the model were investigated by cross–validation. Handling the high computational load and memory requirements for doing this by use of the GenomeDK high performance computing (HPC) cluster hosted at Aarhus University. Moreover, we define alarm limits for prediction of sick cows in Phase 2.

2 Material

The first prototype of the logger was build into a rather large ($185 \times 85 \times 20$ mm) plastic box that was mounted on the neck–collars by use of a leather bag. The logger consisted of an electronic circuit board containing a 3D accelerometer chip, a LF detector tag, and a real–time clock (RTC) block always running by use of a button cell battery on the circuit. To increase the running time the internal rechargeable battery was supplemented by a detachable rechargeable battery pack providing in total approx. 48 hours of runtime. Actually only the extra battery pack was recharged during the present study. Data was stored in text files on a 2 GB detachable USB memory stick and transferred manually to a computer. Two copies of this logger was used for Phase 1 taking place in research stables at Danish Cattle Research Centre (DKC), Foulum.

The second prototype of the logger was smaller and mounted on plastic neck–collars and covered by heat-shrinkable tubing to reduce the risk of it being intertwined in the headlocks at the feed bunk. Presently we do not have the exact size but a guesstimate would be that the plastic box is approx. $80 \times 40 \times 20$ mm without coating. The system was planned to work for a longer period (at least 12 months) and the batteries were not rechargeable. Data was stored on an onboard memory card and transferred on a regular basis by wireless radio communication to a receiver. From the receiver, data was transferred to a server via mobile broadband. Approximately 400 copies of the logger was produced and a LF loop was milled into the feed bunk concrete and covered by epoxy on two commercial dairy herds. The size of the herds were about 150 and 190 lactating cows, respectively. Lyngsoe Systems managed technical installation, maintenance, and data storing.

2.1 Sample for Phase 1

Data from 24 cow/logger combinations were collected in the period from January to April 2014 and synchronised with video recordings. Each session lasted from 21 to 48 hours and sessions using the same cow was separated by at least 14 days and treated as being independent. For ease of presentation, we will refer to this as 24 cows. Video observations were classified per second (continuous focal animal observation) by a trained technician into the states: ears behind feed bunk, ears above feed bunk, eating, other, or view blocked. This classification was dichotomised into *eating* or *not eating*.

Logger data was time–synchronized with the video data by AgroTech, and collapsed to one observation per second by averaging the original data sampled at a frequency of typically 12–14 hertz. The RFID signal was dichotomised: zero if no signals was detected within the second, and one otherwise. In total we have 2,864,478 records (i.e. seconds) from the 24 cows.

In addition to *present time* data points, we also included time–lagged observations with a window size k that was varied between 8 and 128 seconds. That is, when predicting current eating status, we also apply information from the time points $1, 2, \ldots, k$ seconds back in time. The idea being that this could capture changes over time in a way that only requires keeping information for a few minutes in memory as opposed to doing time and power consuming calculations of trends. The effect of lag windows size on prediction performance was investigated with the trade-off in mind that larger windows size is expected to use more power and thus reduce battery life.

2.2 Sample for Phase 2b

Data for one lactation period from 75 dairy cows was obtained from DKC databases, 33 Jersey and 42 Danish Holstein. These cows were controls in another experiment and thus "treated as usual". The cows were followed for at least four weeks (31–429 days, mean=228, median=280) in the lactating period after calving (most of them from day 1 after calving) and allowing for at most a single day gap in data collection. Data "collection" was stopped prematurely for two cows (after day 142 and 173, respectively) due to larger gaps. Three cows have a late entry (day 31, 36 and 65) in terms of registration of eating time and activity. We decided to retain these cows in the study as mounting of sensors well into the lactation may be expected in commercial settings too. Regarding parity, approx. 30% were first lactation cows, 45% were second or third lactation cows, and 25% were in their fourth to sixth lactation. Both breed and parity are likely to influence activity level and eating behaviour but since we are following the course of the

individual cow these factors should not be important for the present study. In total 378 diseases and 169 other non–relevant incidences were registered in the database.

3 Methods

We developed random forests algorithms for prediction of daily eating time using observations at present time and a number of seconds back in time (lag window). Random forests belongs to the field of machine learning and is in essence an ensemble of classification trees or set of decision rules. For a rather technical presentation and introduction to random forests, we refer to Breiman (2001), a paper written by one of the important contributors to the development this method, Leo Breiman (1928–2005). Calculations were carried out using the *randomForest* package (Liaw and Wiener, 2002) in R version 3.1.0 (R Core Team, 2014). The algorithm is outlined in Liaw and Wiener (2002) but we will give a short description of the step stones below. Moreover, guidance and definition of some of the measures returned can be found in Breiman (2003). The R package implements Fortran code originally developed by Leo Breiman and Adele Cutler, see `https://www.stat.berkeley.edu/~breiman/RandomForests/`. The first versions simply provided an R interface to the Fortran programmes while later updates have migrated more and more of this code to C.

For each of a pre–specified number of trees a bootstrap sample is drawn from the original data by sampling with replacement. These bootstrap samples have the same size as the original data but contains on average approx. two thirds of the individual records, since some records are selected more than once and some are not selected at all. This bootstrap sample is used for training an *unpruned* classification tree. At each node of the tree, a number of predictors are chosen at random as candidates for splitting the data present at the current node into two chunks. We used the default setting of *randomForest* for binary classification, which is to pick the square root of the number of predictors at each node. The algorithm then chooses the candidate and cut–point (if continuous) that gives the largest reduction of the Gini index (Breiman *et al.*, 1984), i.e. the *purest* child nodes. Each tree is grown as large as possible. The random selection of candidate predictors at each node protects from overfitting (Breiman, 2001) and no pruning is needed. Once the random forest collection of trees has been obtained, a new record can be run through each decision tree and majority votes used to predict the class of the new data.

We evaluated performance and impact of the lag window size by a *leave–one–cow–out* cross–validation strategy, i.e. in turns preserving data from one cow as test set and

using data from the other 23 cows for training of a random forests model. Performance was assessed by sensitivity, specificity, and average deviation from the daily eating time determined from video recordings.

3.1 Random forests models for Phase 1

Though this is a bit sloppy, we will present a random forests model by the usual formula notation in R, i.e.

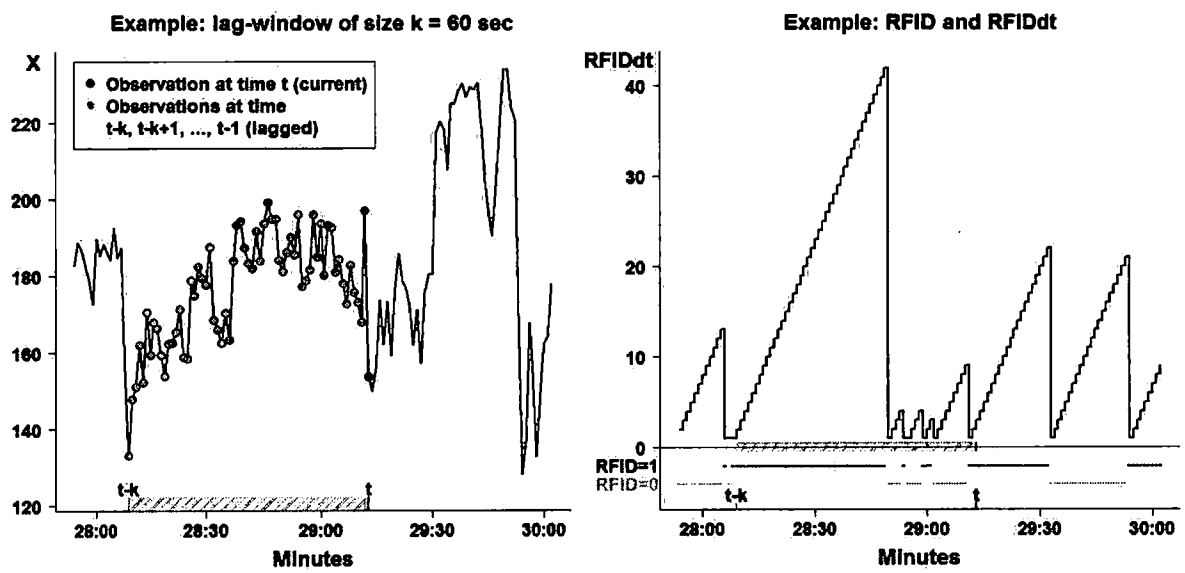
$$Y \sim v_1 + v_2 + \cdots + v_m \ , \tag{1}$$

where Y is some continuous or categorical response of interest and $v_1, \ldots, v_m$ are the m variables used for prediction.

Let (x_t, y_t, z_t) denote the per second averaged 3D–accelerometer data at time t. Let r_t denote the corresponding dichotomised (0/1) RFID signal. The corresponding time–lagged observation will be denoted by e.g. $x_{t-1}, x_{t-2}, \ldots, x_{t-k}$ where t is the present time and k is the lag window size, i.e. seconds back in time. Since the random forests method requires complete data, it is necessary to wait the lag time before starting the prediction but in the big picture, this should not matter. In addition, we keep track of the time spent in the current RFID state (0 or 1) and we will denote this s_t, see Figure 1.

Figure 1

Illustration of the predictors in the model: (x,y,z), RFID, RFIDdt and *lagged* observations.



Note that s_t is at stepwise increasing function counting the number of seconds that the RFID has been zero or one, which is reset to one every time r_t changes from zero to

one or from one to zero. After examining a number of variants we decided to use:

$$Y_{\text{eating},\,t} \sim x_t + y_t + z_t + r_t + s_t + \sum_{j=1}^{k} x_{t-j} + y_{t-j} + z_{t-j} + r_{t-j} + s_{t-j} \ , \qquad (2)$$

where $Y_{\text{eating},\,t}$ is eating status (yes/no or 0/1) at time t obtained from video recordings.

The use of an RFID loop implies extra costs in terms of milling it into the feed bunk concrete, covering the bunk by layers of epoxy to protect the wire from ensilage juice and mechanical damage, power consumption, and additional electronic components. Moreover, it can be technically challenging to adjust the field so that the RFID signals are obtained only when the cow sticks its head through the headlocks and thus holds her head above the feed bunk. In addition to this, other feed bunk systems such as neck rails exist and may introduce other complications when less iron influences the spreading of the electromagnetic field. We therefore also examined a variant without the RFID measures, i.e. a random forests model grown from the accelerometer data alone:

$$Y_{\text{eating},\,t} \sim x_t + y_t + z_t + \sum_{j=1}^{k} x_{t-j} + y_{t-j} + z_{t-j} \ . \qquad (3)$$

3.2 Alarm classification in Phase 2

In the second part of the project, the plan was to predict daily duration of eating for a larger number of cows for a longer period in commercial herds. Moreover, we intended to calculate a measure of daily activity from the accelerometer coordinates. Using the pattern from preceding days, we then hypothesised that deviations for the individual cow could be used to pick sick cows for veterinarian inspection. Below, we define alarm limits that we intended to adapt during Phase 2. Their usefulness is assessed by Phase 2b although this will not proof the usability in the commercial setting.

The alarm limits are based on moving average control charts. Since changes through the lactation is expected, we let the limits vary over time by use of a w days sliding window, whereby limits are calculated from the non–missing observations in the w days preceding the current day. Each limit is adjustable by change of a constant so that the manager of the herd can lower or increase the alarm thresholds and thus the balance between sensitivity and specificity. We assume that data is from the first day after calving but this is easily extended into the period before calving. Only lower alarm limits are used, i.e. triggered by a current day value below the limit determined by preceding days.

Assume that individual daily eating time ($D^{(ET)}$) is predicted from 3-D accelerometer data and RFID signals by use of a random forests model. For simplicity, we will omit indexing of the individual cow. In addition, assume that a daily activity is recorded by

means of an integrated motion index ($D^{(MI)}$) approximated by per second discretisation as described below.

For each of the observed or predicted measures $D^{(V)}$, $V \in \{ET, MI\}$, the current level and evolvement over time is followed for each cow. Let $d = 1,2,\ldots$ denote the so–called *days in milk* (DIM) which is equal to days after calving and let $C_1^{(V)}$ and $C_2^{(V)}$ be constants set by the user. Note that different constants may be used for ET and MI. In principle, V could denote other measures such as number of steps, lying bouts, or standing time. Nevertheless, the alarm classification defined here only uses eating time and motion index.

To determine $D^{(MI)}$, use the average accelerometer coordinates per second that we used as input to the random forests prediction, e.g. $x_t = \frac{1}{n_t}\sum_{j=1}^{n_t} x_{t_j}$, where $t_1, t_2, \ldots, t_{n_t}$ denote sub–seconds. In Phase 1, sub–second frequency typically was 12–14 hertz but from a sub–sampling examination it was determined to use 8 hertz for Phase 2. The motion index is defined as $MI_t = \sqrt{x_t^2 + y_t^2 + z_t^2}$ and the daily integrated motion index can be approximated by

$$D^{(MI)} = \int MI_t dt \approx \sum_{t=1}^{86400} MI_t \ , \tag{4}$$

where 86400 is the number of seconds per day.

3.2.1 Alarms defined by deviation from expected level

Let $E_d^{(V)}$ be the expected level of variable V on day d determined by averaging $D_t^{(V)}$ from the preceding w days, i.e. $t \in \{d-1, d-2, \ldots, d-w\}$. Missing information, e.g. because $d \leq w$, is handled by simply exclusion as follows

$$E_d^{(V)} = \frac{1}{n_d^{(V)}} \sum_{j=1}^{w} 1_{\{D_{d-j}^{(V)} \text{ not missing}\}} D_{d-j}^{(V)} \ , \tag{5}$$

where

$$n_d^{(V)} = \sum_{j=1}^{w} 1_{\{D_{d-j}^{(V)} \text{ not missing}\}} \tag{6}$$

is the number of non–missing measures of $D_t^{(V)}$ during the preceding w days. Here $1_{\{condition\}}$ denotes the indicator function taking the value 1 when the *condition* is met and 0 otherwise. Note that formally the calculation of $E_d^{(V)}$ requires $n_d^{(V)} > 0$ and thus that there are at least one non–missing value. Therefore, $d > 1$ is a necessity.

Daily alarm limits are then calculated as

$$A1_d^{(V)} = \begin{cases} 0 & ,\text{if } n_d^{(V)} < 2 \\ E_d^{(V)} - C_1^{(V)} \frac{s1_d^{(V)}}{\sqrt{n_d^{(V)}}} & ,\text{if } n_d^{(V)} \geq 2 \end{cases} \tag{7}$$

Here $C_1^{(V)}$ is a constant and $s1_d^{(V)}$ is the square root of the unbiased variance estimate determined by the non–missing values from the preceding w days. Note that $n_d^{(V)} \geq 2$ is required for calculation of $s1_d^{(V)}$ and implies the necessity of $d \geq 3$.

Since we are only considering positive measures the limit when $n_d^{(V)} < 2$ will always be fulfilled and may be debated. Nevertheless, we see no obvious alternative except maybe from "not defined" which would be a bit impractical. From *ET* and *MI* alarms, we will define the following three categories for the individual cow on day d: 0 if there are no alarms; 1 if there is one alarm (either *ET* or *MI*); 2 if there are two alarms.

3.2.2 Alarms defined by deviation from expected trend

In addition to the alarms defined in section 3.2.1, we also considered how to capture larger changes in the trend or more specifically sudden drops in the measures. We will refer to the instantaneous trend as the change from last observation divided by the days elapsed. Again, only lower limits are of interest and determined by moving averages via a sliding window over the preceding w days in milk.

Let $\hat{\beta}_d^{(V)}$ denote the maximum likelihood estimate of the slope from a linear regression of $D_d^{(V)}$ against the preceding w days. We are not interested in the intercept from this regression, only the slope. Let

$$\bar{t}_d^{(V)} = \frac{1}{n_d^{(V)}} \sum_{j=1}^{w} 1_{\{D_{d-j}^{(V)} \text{ not missing}\}}(d-j) \ , \tag{8}$$

$$SSD_d^{(V)} = \sum_{j=1}^{w} 1_{\{D_{d-j}^{(V)} \text{ not missing}\}}(d-j-\bar{t}_d)^2 \tag{9}$$

and

$$SPD_d^{(V)} = \sum_{j=1}^{w} 1_{\{D_{d-j}^{(V)} \text{ not missing}\}} D_{d-j}^{(V)}(d-j-\bar{t}_d^{(V)}) \ . \tag{10}$$

Then

$$\hat{\beta}_d^{(V)} = \frac{SPD_d^{(V)}}{SSD_d^{(V)}} \ . \tag{11}$$

Moreover, the standard error of $\hat{\beta}_d^{(V)}$ is

$$\frac{s2_d^{(V)}}{\sqrt{SSD_d^{(V)}}}\ , \tag{12}$$

where

$$s2_d^{(V)} = \sqrt{\frac{1}{n_d^{(V)}-2} SSD1_d^{(V)} - \frac{(SPD_d^{(V)})^2}{SSD_d^{(V)}}}\ . \tag{13}$$

We have to require at least three non-missing observations to determine this standard deviation. We now define alarm limits for the instantaneous trend to be

$$A2_d^{(V)} = \begin{cases} 0 & ,\text{if } n_d^{(V)} < 3 \\ \hat{\beta}_d^{(V)} - C_2^{(V)} \frac{s2_d^{(V)}}{\sqrt{SSD_d^{(V)}}} & ,\text{if } n_d^{(V)} \geq 3 \end{cases} \tag{14}$$

Note that we issue an alarm the second day of lactation if the instantaneous trend is negative, i.e. if the measure the second day is lower than the first day. The instantaneous trend is not defined on the first day. Moreover, $n_d^{(V)} \geq 3$ implies the necessity of $d \geq 4$.

Using the definitions in Section 3.2.1 and Section 3.2.2, we will refer to the two alarms (level or trend) concerning eating as one *type* of alarm and the two alarms on motion index as a second *type*. Ranked by some sort of *severity*, we define a six leveled categorisation: 0 if there are no alarms; 1 if there is one alarm; 2 if there are two alarms of the same *type*; 3 if there are two alarms of different *type*; 4 if there are three alarms; 5 if all four alarms are triggered. Note that for the categories 3, 4 and 5 there will be at least one alarm of each *type* whereas for 0, 1 and 2 there are at most one *type* of alarm.

4 Results

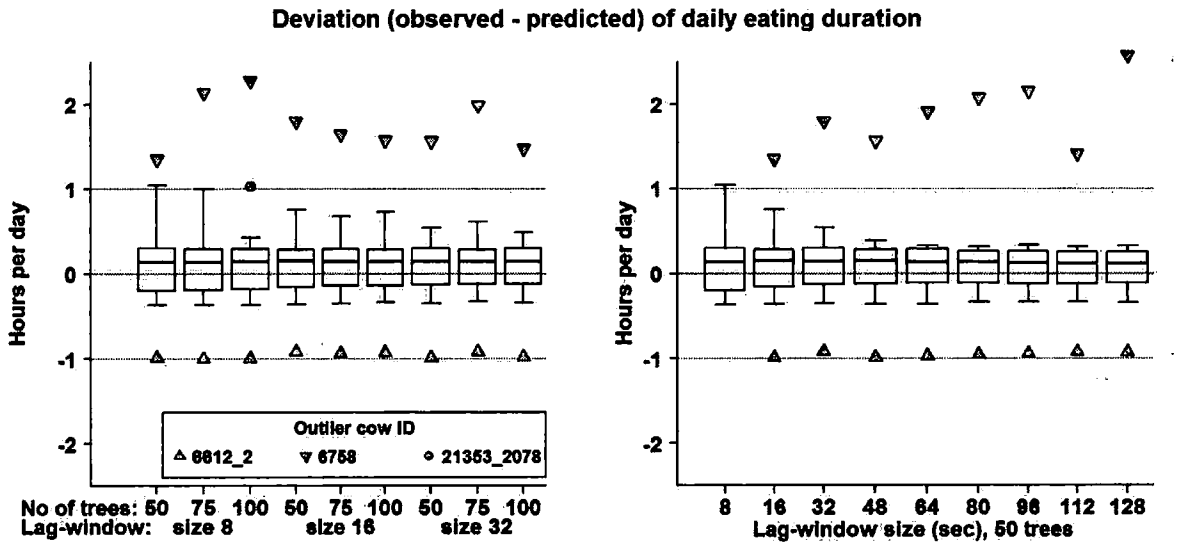
The results shown in Figure 2 and Table 1 are from *leave–one–cow–out* cross–validation, i.e. we reserve observations from one cow as test set and use the rest as training set. The random forest is then estimated by use of the training data and its performance determined by predicting status of the test cow. Differences in performance between the models with and the models without RFID measurements can be seen in Table 1.

Estimated eating time is deviating more markedly for two cow: cow ID 6612_2 that tend to have under–estimated daily duration of eating; and cow ID 6758, which are having over–estimated eating time (see Figure 2). We searched for reasons and stumbled over some strange patterns in the measured accelerometer coordinates for these two cows, see Figure 3. Examining the video recordings around the time points where sudden changes are very clearly seen, we found that the loggers sometimes had been turned

around. Apparently, the logger could be caught in the head lockers when the cow started or stopped eating or searching for food. Figure 2 also shows the typical difference in pattern between periods where the cow is eating and periods with other activities.

Figure 2

Comparing the *leave–one-cow–out* cross–validation results from random forests models (including RFID measures) for varying number or trees and size of lag–window. Note that not all results were ready at time of printing, see Table 1



5 Conclusions and discussion

Feed intake is very important for dairy cows and deviation from normal eating behaviour may predict sick cows. Therefore, we investigated whether a device from Lyngsoe Systems (Aars, Denmark) could be used to estimate eating behaviour. Data were collected from 24 cows and synchronised with video recordings at the Danish Cattle Research Centre (DKC), Foulum. The sensor recorded 3D accelerometer data and radio frequency identification (RFID) signals for positioning of the cow at the feed bunk. Video observations from 21 to 48 hours per cow were classified per second as eating or not eating. Logger data was reduced to per second level by averaging original 12–14 hertz signals.

In Phase 1 of the study, we developed a random forests prediction model to be used for monitoring eating behaviour of dairy cows. Our results show that daily eating time is predicted reasonably well by a random forests algorithm using sensor observations at present time and a number of seconds back in time (lag–window).

Table 1

Cross–validation results, showing median and quartiles ($_{p25}$ *median* $_{p75}$) for the two models with and without RFID measures:

$$Y_{\text{eating},t} \sim x_t + y_t + z_t + r_t + s_t + \sum_{j=1}^{k} x_{t-j} + y_{t-j} + z_{t-j} + r_{t-j} + s_{t-j} \quad \text{(a)}$$

$$Y_{\text{eating},t} \sim x_t + y_t + z_t + \sum_{j=1}^{k} x_{t-j} + y_{t-j} + z_{t-j} \quad \text{(b)}$$

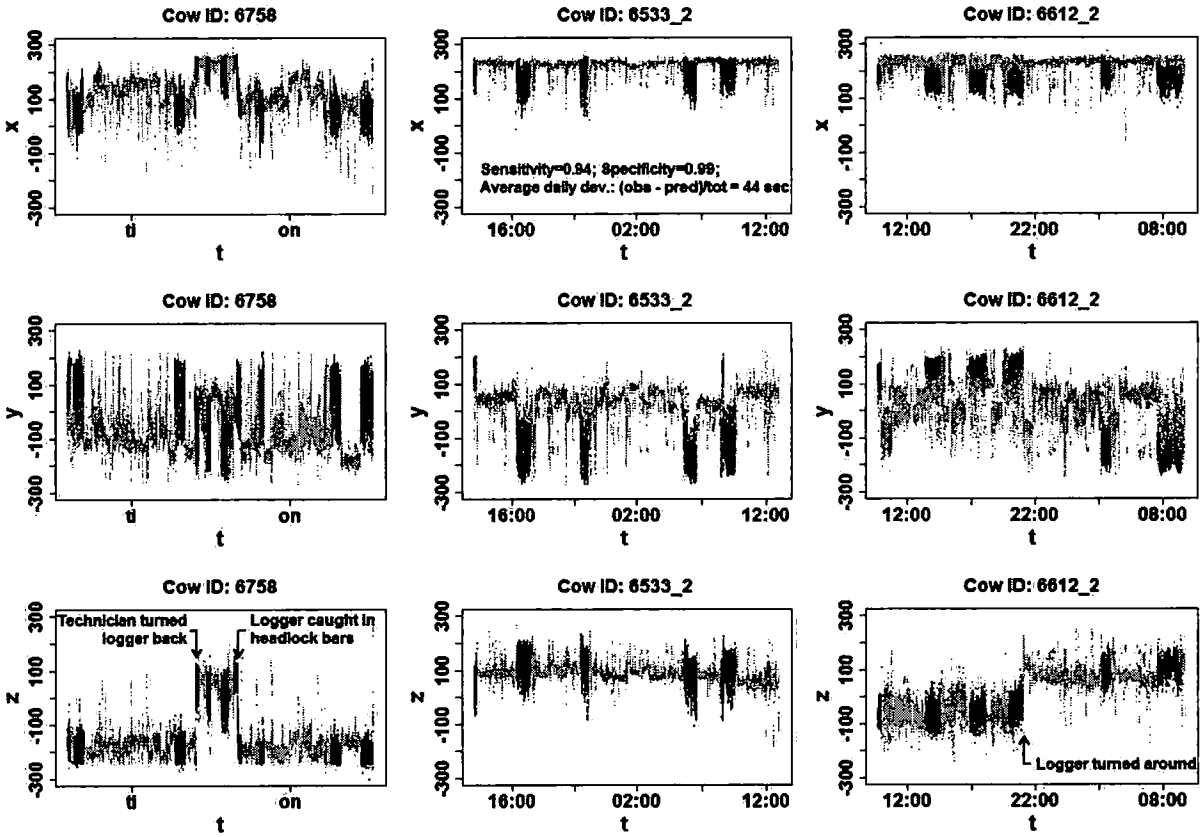
In both cases random forests models with 50 trees and varying sizes, w, of the lag–window.

Model	w	Sensitivity	Specificity	Bal. Accuracy	Eating (hours)
a	8	$_{0.87}$ 0.9 $_{0.93}$	$_{0.98}$ 0.98 $_{0.99}$	$_{0.92}$ 0.94 $_{0.96}$	$_{-0.19}$ 0.14 $_{0.29}$
b	8	$_{0.56}$ 0.76 $_{0.83}$	$_{0.95}$ 0.96 $_{0.97}$	$_{0.77}$ 0.86 $_{0.89}$	$_{-0.4}$ −0.13 $_{0.87}$
a	16	$_{0.86}$ 0.9 $_{0.93}$	$_{0.98}$ 0.98 $_{0.99}$	$_{0.92}$ 0.94 $_{0.96}$	$_{-0.13}$ 0.16 $_{0.27}$
b	16	$_{0.55}$ 0.76 $_{0.84}$	$_{0.95}$ 0.96 $_{0.97}$	$_{0.76}$ 0.86 $_{0.89}$	$_{-0.31}$ −0.12 $_{0.91}$
a	32	$_{0.86}$ 0.9 $_{0.94}$	$_{0.98}$ 0.99 $_{0.99}$	$_{0.92}$ 0.94 $_{0.96}$	$_{-0.11}$ 0.15 $_{0.3}$
b	32	$_{0.54}$ 0.78 $_{0.85}$	$_{0.96}$ 0.96 $_{0.97}$	$_{0.76}$ 0.87 $_{0.89}$	$_{-0.25}$ −0.12 $_{0.95}$
a	48	$_{0.86}$ 0.89 $_{0.93}$	$_{0.98}$ 0.99 $_{0.99}$	$_{0.93}$ 0.94 $_{0.96}$	$_{-0.1}$ 0.16 $_{0.28}$
b	48	$_{0.55}$ 0.78 $_{0.85}$	$_{0.96}$ 0.97 $_{0.97}$	$_{0.76}$ 0.87 $_{0.89}$	$_{-0.3}$ −0.05 $_{1}$
a	64	$_{0.87}$ 0.9 $_{0.94}$	$_{0.98}$ 0.99 $_{0.99}$	$_{0.93}$ 0.94 $_{0.96}$	$_{-0.1}$ 0.15 $_{0.29}$
b	64	$_{0.53}$ 0.79 $_{0.84}$	$_{0.96}$ 0.97 $_{0.97}$	$_{0.75}$ 0.88 $_{0.89}$	$_{-0.28}$ −0.03 $_{1.03}$
a	80	$_{0.87}$ 0.9 $_{0.94}$	$_{0.98}$ 0.99 $_{0.99}$	$_{0.93}$ 0.94 $_{0.96}$	$_{-0.1}$ 0.14 $_{0.27}$
b	80	$_{0.53}$ 0.79 $_{0.84}$	$_{0.96}$ 0.97 $_{0.97}$	$_{0.76}$ 0.88 $_{0.9}$	$_{-0.27}$ −0.03 $_{1.03}$
a	96	$_{0.87}$ 0.9 $_{0.94}$	$_{0.98}$ 0.99 $_{0.99}$	$_{0.92}$ 0.95 $_{0.96}$	$_{-0.11}$ 0.13 $_{0.26}$
b	96	$_{0.53}$ 0.79 $_{0.84}$	$_{0.96}$ 0.97 $_{0.98}$	$_{0.75}$ 0.88 $_{0.9}$	$_{-0.26}$ −0.02 $_{1.06}$
a	112	$_{0.86}$ 0.9 $_{0.94}$	$_{0.98}$ 0.99 $_{0.99}$	$_{0.93}$ 0.95 $_{0.96}$	$_{-0.1}$ 0.13 $_{0.25}$
b	112	$_{0.53}$ 0.79 $_{0.83}$	$_{0.96}$ 0.97 $_{0.98}$	$_{0.75}$ 0.87 $_{0.9}$	$_{-0.26}$ 0.01 $_{1.05}$
a	128	$_{0.86}$ 0.9 $_{0.94}$	$_{0.98}$ 0.99 $_{0.99}$	$_{0.93}$ 0.95 $_{0.96}$	$_{-0.1}$ 0.13 $_{0.25}$

Performance was measured by *leave–one–cow–out* cross–validation, i.e. in turns preserving data from one cow as test set and using data from the other 23 for training of a random forests model. Number of trees only affected results slightly and 50 trees seemed to suffice except maybe for the smallest size of the lag-window. We then examined lag–window sizes from 8 to 128 seconds with 50 trees but results did not change much, and a lag–window of size around one minute seems to be a reasonable choice.

Figure 3

Accelerometer (x,y,z)–coordinate observations for the two most outlying cows (ID 6758 and 6612_2) compared with the cow (ID 6533_2) that had the best prediction of daily eating time in the *leave–one-cow–out* cross–validation. Black points indicate time points where the cow is eating. Sudden changes in patterns were investigated and reasons are noted on the plots.



The results suggest that the first prototype device can be used to estimate eating behaviour of dairy cows with good accuracy. However, a second generation that has been developed needs to be validated on commercial farms. The problem of the logger getting turned by the head lockers were considered during the development of the housing of this second generation and appears from manual observations to be working well in this respect.

We investigated if the RFID part of the system could be left aside. However, results from using accelerometer information alone clearly demonstrates that the RFID measurements carry an important information for the prediction of daily eating time. Therefore, this simplification cannot be recommended on basis of the current results.

Estimating classification rules in the models with larger lag–window size requires a huge amount of memory on the computer. Nevertheless, such potentially heavy demands when building the decision rules should have no practical impact on the com-

puter requirements on the farm. Therefore, requirements to build the models is not a concern as long as we have access to machines that are big enough to do this job. The number of classification rules and amount of data needed to be stored for calculations (lag–window) may however be of practical relevance.

An important question to answer is of course how large impact number of trees and size of lag–window has on the ability to predict sick cows. Unfortunately this important part was delayed due to technical challenges and will not be dealt with by the present project. We are currently utilising historical data from DKC to investigate the potential of predicting sick cows from deviations in daily eating behaviour and activity level. The first results (not shown) indicate that the use of deviations from expected *instantaneous trend* result in far too many alarms.

6 Acknowledgements

We are thankful for the opportunity to use the GenomeDK HPC hub at Aarhus University (`http://genome.au.dk`). We thank technical staff from Dept. of Animal Science, Aarhus University, involved in the collection of data and scoring of videos in Phase 1. We thank the staff at Lyngsoe Systems in Aars (Denmark) and Toronto (Canada) for the development of hardware, firmware, technical solutions on farms, for data acquisition, storage and exchange in Phase 2, and collaboration in the project generally. We thank the Danish Cattle Research Centre (DKC) for the ability to use historical data.

COWTrack is a GUDP project conducted in cooperation between Lyngsoe Systems and Department of Animal Science, Aarhus University with support from The Danish AgriFish Agency, Ministry of Environment and Food.

References

Breiman, L. , Friedman, J.M. , Olshen, R.A. and Stone, C.J. *Classification and Regression Trees.* Chapman & Hall/CRC, 1984. ISBN 0–412–04841–82

Breiman, L. Random Forests. *Machine Learning* 2001; **45**: 5–32. `http://dx.doi.org/10.1023/A:1010933404324`

Breiman, L. Manual for Setting Up, Using, and Understanding Random Forest V4.0. `https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf`

Liaw, A. and Wiener, M. Classification and Regression by randomForest. *R News* 2002; **2**(3): 18–22. `http://CRAN.R-project.org/doc/Rnews/`

R Core Team R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2014. `https://www.R-project.org/`

Development of a predictive algorithm for a pig farming decision support system

Mikkel Boel & Leslie Foldager

Abstract

The development and use of predictive algorithms is a growing field in many disciplines, e.g. science, economics and medicine. It is based on a process of developing and validating an algorithm with respect to predicting a specific outcome, and can depend on one or more methods of e.g. statistical analyses and machine learning. The goal is basically to make predictions about a specific outcome, based on the present state and the correlations to past state(s) and outcomes.

Here we use changes in behaviour and environment during the development over time to predict incidences of tail biting in slaughter pigs. Reconfiguration of explanatory information, compensating for temporal-/seasonal- or group related differences in the data was used to increase general applicability of decision support tools; using the residual information from linear/non-linear models (e.g. of season or day), differencing or simple ratios. Cross-validation was used to examine the validity for usage in a pig farming decision support system. Finally, algorithm performance was evaluated using measures of the predictive capability.

Hvad er din bil værd?

Michael Sperling SAS Institute

Alle der har prøvet at købe ny, brugt bil, har haft overvejelser om mærke, model[1], alder, antal kørte kilometer osv. man ønsker. Venner og bekendte er klar med gode råd og hjemmesider, som eksempelvis www.bilbasen.dk, er også behjælpelig med information om den enkelte bils pris i forhold til lignende biler, men man er alligevel i tvivl. Dette er et forsøg på at komme lidt nærmere, hvad en "fair pris" er, samt hvad der driver denne pris. Proc Mixed i SAS benyttes til at lave en model for udbudspriser på www.bilbasen.dk.

Data

Med brug af SAS, er data hentet fra www.bilbasen.dk i perioden fra august til november 2016. I alt 9.026 forskellige annoncer er hentet med brugte biler af mærkerne: Alfa Romeo, Audi, BMW, Citroën, Kia, Mazda, Nissan, Peugeot, Renault, Skoda, Toyota, Volvo og VW2. Biler produceret før 2005 samt leasingbiler og helt nye biler, som ingen kilometer har kørt, er sorteret fra. Producent (eks: Peugeot) og model (eks: 308) bruges til at danne klynger, og for at sikre en rimelig volumen i hver klynge (eks. Peugeot 308), er modeller med mindre end 50 unikke annoncer sorteret fra. Tilbage er 5.153 annoncer, 3.643 i træningssættet og 1.510 i testdatasættet.

Antal kørte kilometer og bilens pris er begge højreskæve. Proc transreg foreslår en box-cox transformation med lambda lig 0,4 for antal kørte kilometer, mens en logtransformation er anvendt for bilens pris. Efter logtransformationen er udbudspriserne tilnærmelsesvis normalfordelte, og stikprøven repræsenterer det fulde datasæt i forhold til middelværdi, varians, skævhed og kurtosis.

Deskriptiv statistik

Antal kørte kilometer og alderen på bilen må forventes at være væsentlig for bilens pris. Det vil føre for vidt at præsentere alle modeller her, så Peugeot 308 og Audi A3 er udvalgt som eksempler.

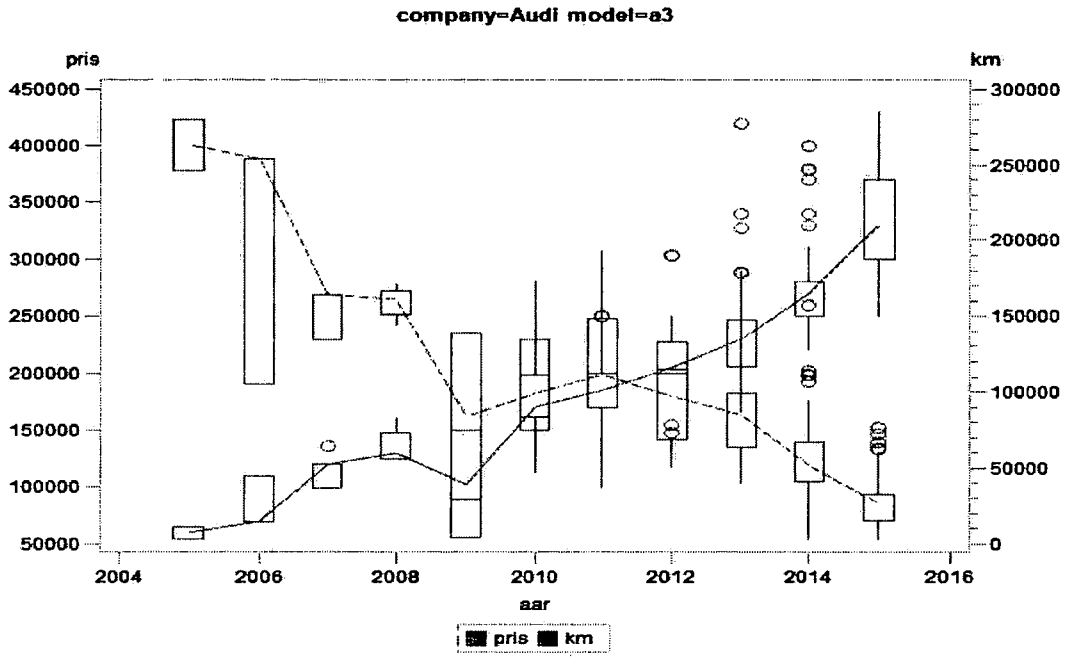
Audi A3

Den blå kurve viser udviklingen i prisen for Audi A3 biler. For biler produceret i 2015 er medianen af udbudspriserne ca. 330.000 kroner, mens samme model produceret i 2005 har en medianpris på ca. 60.000 kroner. Middelværdi og variansen er generelt faldende med bilens alder.

---

[1] "Model" refererer her til bilmodeller (eks. Peugeot 308), mens der andre steder refereres til den statistiske model. Beklager forvirringen

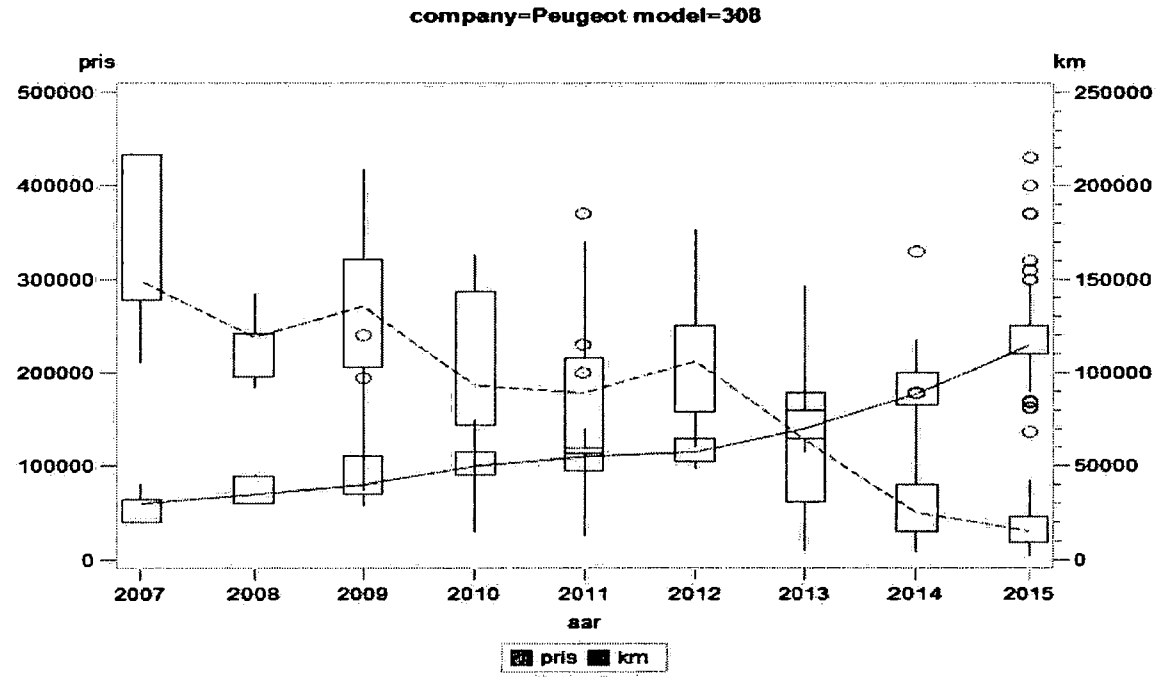
[2] Producenterne er tilfældigt valgt efter forfatterens hukommelse



Den stiplede kurve viser bilernes gennemsnitlige antal kørte kilometer. Bilerne fra 2005 har i gennemsnit kørt 263.000 kilometer, mens Audi A3 biler fra 2015 i gennemsnit har kørt 27.000 kilometer. Over perioden på 10 år kører Audi A3ere i omegnen af 235.000 kilometer og prisen falder ca. 270.000 kroner - begge dele tilnærmelsesvis lineært - der er i hvert fald ikke tydelige eksponentielle udviklinger.

Peugeot 308

Peugeot 308 modeller produceret i 2015 har en gennemsnitlig udbudspris på ca. 229.000 kroner, faldende til gennemsnitlig udbudspris på ca. 60.000 kroner for samme model produceret i 2007. Medianen af antal kørte kilometer, 149.000 kilometer i 2007,



faldende til 15.000 kilometer for Peugeot 308 produceret i 2015. Over perioden på otte år kører Peugeot 308ere i omegnen af 135.000 kilometer (100.000 kilometer færre end Audi A3) og prisen falder ca. 170.000 kroner (100.000 mindre end Audi A3).

På tværs af producenter og modeller, er der variation i antallet af biler, kørte kilometer og udbudspriser, men de to ovenstående eksempler repræsenterer det generelle billede.

Mixed Models

Mixed models (Proc Mixed i SAS) er anvendt, da vi har en kontinuert afhængig variabel, udbudsprisen, som er tilnærmelsesvis normalfordelt, samt at vores data har en hierarkisk struktur, hvor den enkelte observation tilhører en klynge, nemlig modellen og producenten.

I det konkrete eksempel, er der i teorien tre niveauer, producent (f.eks. Peugeot), modellen (f.eks. 308) og den enkelte bil, men alle de objektive egenskaber er på den enkelte bil. Biler varierer imidlertid i værdi ud over de objektive egenskaber, dvs. der er en værdi (brand eller reel) i de enkelte producenter og modeller. Toyota og Volkswagen har gennem de seneste år fået nogle ridser i imaget. Vi har ikke data fra før "Dieselgate", så det kan ikke belyses, men i fremtiden, vil der muligvis komme lignende skandaler, som kan belyse dette lidt bedre. Endvidere kan man forestille sig, at værdiforøgelsen af motorstørrelse af afskrivningen pr. år varierer mellem modeller og producenter. For nuværende betragtes en model med to niveauer.

$$Y_i + X_i\beta + Z_i u_i + \varepsilon_i$$

Hvor β er fixed effects parametre og dermed gennemsnit af Y_i, udbudsprisen, for hele stikprøven. u_i er random variables, dvs. de enkelte bilmodellers afvigelser fra det generelle billede beskrevet af β, og er normalfordelte med middelværdien 0 og variance-covariance matrixen G. ε_i er residualerne, som er normalfordelte med middelværdien 0 og variance-covariance matrix R_i. u_i, og ε_i er uafhængige.

Det er som altid hensigten at finde en simpel model, som beskriver data tilfredsstillende. I modeludvælgelsen er restricted maximum likelihood estimation anvendt med udgangspunkt i en hel simpel model. AIC og BIC er anvendt som de primære mål for modeludvælgelse, men der er også skævet til andre indikatorer, herunder residualplots og parameterestimater.

Den primære model

I den primære model indgår alder, antal kørte kilometer, motorens størrelse, antal kørte kilometer pr. liter samt en række vekselvirkninger. Der er benyttet maximum

likelihood estimation, og satterthwaite-metoden er benyttet, til at beregne antallet af frihedsgrader. Datasættet er forholdsvist stort, hvilket medfører at estimationsmetode (REML vs. ML) samt valg af metode til beregning af frihedsgrader, ikke har den store betydning. Som random effects, er de enkelte modeltyper nestet i producent. Bilmodellerne er nestet ind i producent af hensyn til læsevenlighed, samt fordi det forbedrer performance i Proc Mixed.

```
proc mixed data = work.abt_ic covtest method = ml ;
 class company model;
 model log_price = age|bc_mileage motorsize /
 solution ddfm=sat residual outp=pred_log_price;
 random intercept / subject=model(company) g solution;
 ods output fitstatistics=fits;
 store bilb_abt.mixed;
run;
```

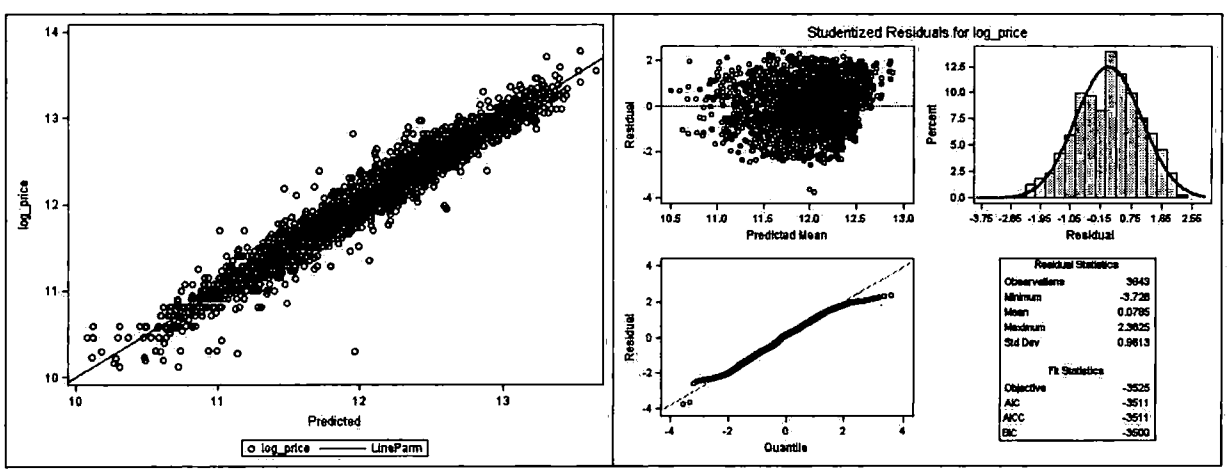
Dele af outputtet (Fit statistics) er gemt i datasættet fits, de prædikterede værdier er gemt i datasættet pred_log_price, mens modellen er gemt i et eksternt bibliotek, således, at den statistiske model kan scores på andre data (testdata), end de data vi har anvendt til at estimere modellens parametre (træningsdata).

Modelevaluering

Med `ods graphics` option samt `residual` i `model`-statementet i Proc Mixed genereres automatisk residualplots, mens følgende kode til anvendes til at danne plot på baggrund af de eksporterede data fra Proc Mixed.

```
proc sgplot data=pred_log_price;
 scatter x=pred y = log_price  ;
 lineparm x=10 y=10 slope=1 ;
run;
```

Residualerne (til højre) er generelt pæne, om end der er lidt tunge haler. Grafen til venstre viser de prædikterede og de faktiske værdier plottet mod hinanden. Som



analytiker er dette tilfredsstillende, men da prisen er logtransponeret, kan selv små afvigelser fra 45-graders-linjen medføre forholdsvis store afvigelser.

Resultater

Fixed effects i modellen er alder, antal kørte kilometer, motorstørrelsen og antal kørte kilometer på literen, samt enkelte vekselvirkninger. Udbudsprisen falder med kilometertælleren og alderen, men forøges med motorstørrelsen og antal kørte kilometer på en liter brændstof. Disse er alle i overensstemmelse med, hvad man skulle forvente.

Solution for Fixed Effects

Effect	*Estimate*	*Std. Err*	*DF*	*t Value*	*Pr > |t|*
Intercept	12.1247	0.07470	47	162.32	<.0001
Age	-0.02309	0.004120	3604	-5.60	<.0001
bc_mileage	-0.00077	0.000067	3605	-11.44	<.0001
age*bc_mileage	-0.00036	0.000015	3604	-24.08	<.0001
Motorsize	0.2991	0.01094	3618	27.33	<.0001

Random effekt estimaterne er også meget intuitive. Alfa Romeo, Audi, BMW, Volkswagen og Volvo udbydes generelt til højere priser for samme tekniske egenskaber, mens Citroën, Kia, Peugeot, Renault og Toyota ligger under gennemsnittet givet alder, forbrug og motorstørelse. Billedet for Mazda, Nissan og Skoda er lidt blandet og afhænger af modellerne.

Ekstreme og indflydelsesrige observationer

Med en analytikers/statistikers øjne, er modellen ganske god. Men hvad med en bilsælger eller en bilkøbers? Hvor god er modellen til at ramme bilernes faktiske udbudspriser? Kan vi bruge modellen til at finde de billige hhv. for dyre biler?

Gennemsnitspriserne, de prædikterede mod de faktiske, afviger generelt kun med få tusinde kroner, og den gennemsnitslige absolutte forskel er godt 10 pct. Det er rimelig pænt (uden at være prangende).

Outliers

For en række biler gælder, at modellens prædikterede værdi afviger en hel del fra den faktiske udbudspris på www.bilbasen.dk. De 10 største afvigere - altså forskellen mellem den prædikterede værdi og den faktiske udbudspris - er følgende:

Company	*Model*	*Age*	*Km*	*Km/L*	*Motor-size*	*Predicted price*	*Actual Price*	*Price diff.*
Audi	a4	1	23.000	19.6	3,00	598.271	899.900	301.629
VW	passat	1	36.000	18.5	2,00	383.965	599.900	215.935
Peugeot	308	5	49.000	17.8	2,00	154.880	369.900	215.020
Audi	a6	2	6.000	18.9	3,00	763.024	969.900	206.876

Company	Model	Age	Km	Km/L	Motor-size	Predicted price	Actual Price	Price diff.
Audi	a6	2	6.000	18.9	3,00	763.024	969.900	206.876
Alfa Romeo	giulietta	6	9.000	13.2	1,75	247.719	429.900	182.181
Alfa Romeo	giulietta	6	8.000	13.2	1,75	250.897	429.900	179.003
Peugeot	308	1	5.000	24.4	2,00	250.950	429.900	178.950
Audi	a6	4	113.000	16.9	3,00	462.949	639.900	176.951
Volvo	xc60	3	6.000	14.7	2,40	665.013	489.900	175.113

Generelt er det de dyrere biler (især Audi A6), hvor den prædikterede og faktiske pris afviger mest. Modellen undervurderer i alle 10 tilfælde.

Enkeltobservationers indflydelse

Med brug af `influence` optionen beregnes bl.a. Cook's Distance, CovRatio og PRESS og med `ods output` fås observationerne ud i et datasæt.

```
model log_price = age|bc_mileage motorsize / influence;
ods output InfluenceStatPanel = influence;
```

Cooks afstandsmål beregnes efter følgende formel, hvor en større værdi er lig med en større indflydelse for den enkelte observation.

$$D(\beta) = \frac{(\hat{\beta} - \widehat{\beta_U})'\widehat{Var}[\hat{\beta}]^{-}(\hat{\beta} - \widehat{\beta_U})}{rank(X)}$$

De fem største værdier af Cooks afstandsmål, altså de fem observationer, som har størst indflydelse på parameterestimaterne, var følgende:

Company	Model	Age	Km	Km/L	Motor size	Price	Cook's D
Citroen	c5	3	199.000	11.9	2,00	29.800	0.06060
Toyota	Avensis	11	293.000	13.9	1,80	25.000	0.04089
Citroen	c5	11	105.000	12.7	1,80	29.000	0.03232
Skoda	octavia	11	220.000	20.0	1,90	30.000	0.02936
Alfa Romeo	Giulietta	6	8.000	13.2	1,75	429.900	0.02157

CovRatio, til at belyse enkeltobservationers indflydelse på præcisionen af parameterestimaterne, beregnes efter følgende formel

$$COVRATIO(\beta) = \frac{det_{ns}(\widehat{Var}[\widehat{\beta_U}])}{det_{ns}(\widehat{Var}[\hat{\beta}])}$$

Hvor $det_{ns}(M)$ er determinanten af den ikke-singulære del af matricen **M**. Nedenfor er listet de fem annoncer, hvor CovRatio er længst fra 1.

Company	Model	Age	Km	Km/L	Motor size	Price	COVRATIO
Citroën	c5	3	199.000	11.9	2,00	29.800	0.8293
Peugeot	308	5	49.000	17.8	2,00	369.900	0.9534

Company	Model	Age	Km	Km/L	Motor size	Price	COVRATIO
Citroën	c5	11	105.000	12.7	1,80	29.000	0.9556
Peugeot	308	7	93.000	13.3	1,60	194.800	0.9695
Kia	ceed	1	123.000	20.4	1,60	84.900	0.9699

Et noget blandet billede af gamle og nye biler af forskellige producenter. Generelt ligger CovRatio tæt på 1. Det bemærkes, at en tre år gammel Citroën har væsentlig indflydelse på både parameterestimaterne og varianserne.

Endelig er Predicted Residual Sum of Squares (PRESS) residualerne beregnet.

$$PRESS_{(U)} = \sum_{i \in U} \hat{e}_{i(U)} \quad , \quad \hat{e}_{i(U)} = y_i - x'_i \hat{\beta}_{(U)}$$

Den prædikterede værdi beregnes uden den enkelte observation for derved at finde den enkelte observations samlede indflydelse på den prædikterede værdi på tværs af alle parametrene. Større værdi er lig med større indflydelse.

Company	Model	Age	Km	Km/L	Motor size	Price	PRESS Res
Audi	a5	7	114.000	13.5	2,00	369.900	1.1102
BMW	520d	5	139.000	18.9	2,00	424.900	1.0723
Audi	a5	8	204.000	14.9	2,70	298.700	1.0541
BMW	520d	6	223.000	19.2	2,00	305.000	1.0368
Volvo	xc60	7	125.000	12.0	2,40	369.800	1.0225

Alternative modeller

Gruppering af producenter

Audi, BMW og Volvo udbydes generelt til højere priser for samme tekniske egenskaber, mens Citroën, Kia, Peugeot, Renault og Toyota ligger under gennemsnittet givet alder, forbrug og motorstørrelse. Det giver anledning til en hypotese om, at inddele bilerne tre grupper, *luksus-*, *billige-* og *mellemklassebiler* i stedet for efter producent. Denne simplificering giver dog en væsentlig dårligere beskrivelse af de 3.643 biler i træningsdatasættet.

Random Slope

Fra tid til anden høres udsagn a la "tyske biler holder prisen meget bedre". Men forholder det sig egentlig således, at prisen falder mere med alderen for nogle typer af biler end for andre? Eller at værdien af en større motor er afhængig af bilens model og producent? Modellen udvides med random effects (random slopes) for alder og motorstørrelsen er tilføjet med en fri kovariansstruktur for de tre random effects.

```
random intercept motorsize age / subject=model(company) GCORR TYPE=UN ;
```

Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Std Err	Z Value	Pr Z
UN(1,1)	model(company)	0.2840	0.07385	3.84	<.0001
UN(2,1)	model(company)	-0.04682	0.02234	-2.10	0.0361
UN(2,2)	model(company)	0.02531	0.008232	3.07	0.0011
UN(3,1)	model(company)	-0.00023	0.002234	-0.10	0.9165
UN(3,2)	model(company)	-0.00041	0.000789	-0.52	0.6036
UN(3,3)	model(company)	0.000549	0.000151	3.64	0.0001
Residual		0.01824	0.000435	41.97	<.0001

Varianserne for de tre random effects er klart signifikante. Der synes altså at være en variation i, hvor meget udbudsprisen for enkelte modeller varierer med motorstørrelse og alder. Kovariansen mellem effekterne er til gengæld overvejende insignifikante, så dermed ingen sammenhæng mellem afskrivningsraten og merværdien af større motor. Det synes derfor fornuftigt at tilføje random effects for alder og motorstørrelse og lade kovariansstrukturen bestå af varianskomponenterne alene (standardindstillingen).

```
random intercept motorsize age / subject=model(company) type=VC;
```

Mens modellens umiddelbare statistiske performance er forbedret, er fortolkningen mere problematisk Det er svært at finde et entydigt billede af de tilføjede random effects for den enkelte producent (endsige nationalitet), hvorfor det er vanskeligt at svare på spørgsmålet, om tyske biler faktisk holder prisen bedre.

Kovariansstruktur for residualerne

Standardindstillingen kovariansstrukturen for residualerne, R_i, er compound symmetry i Proc Mixed. Det er muligt at vælge andre kovariansstrukturer, eksempelvis ustruktureret, autoregressiv og heteroskedastisk compound symmetry m.fl. Dette gav dog ikke en væsentlig bedre model i det konkrete tilfælde.

Kvadratled som forklarende variable

Modellen har vanskeligt ved at prædiktere prisen for især de dyre biler, som ofte er karakteriseret ved at have en stor motor. En udvidelse af modellen med kvadratled gav imidlertid ikke væsentlige forbedringer af modellen forklaringskraft - slet ikke, hvis motorstørrelsen også indgår som random effect.

Testdata

Som nævnt indledningsvist blev data splittet i et testdatasæt og et træningsdatasæt. Efter estimering på træningssættet blev modellen gemt, og denne model er benyttet til at score testdatasættes 1.510 annoncer, for at teste modellens generaliserbarhed.

```
proc plm restore=bilb_abt.mixed;
   show cov Parms;
   score data=TestData out=ScoreResults;
run;
```

Proc Plm benyttes til at score testdata med brug af den model, som blev udviklet på træningsdata og efterfølgende gemt. Akkurat som for træningssættet er udbudspriserne for de dyrere biler som Audi og Volvo vanskelige at prædiktere. Gennemsnitsresultaterne er noget nedslående med relativt store afvigelser fra de prædikterede til de faktiske udbudspriser. Gennemsnitsafvigelsen for mange af de dyre biler er over 100.000 kroner. Modellens præcisions på testdatasættet lader noget tilbage at ønske.

Company	Model	Age	Km	Km/L	Motor size	Predicted price	Price	Price diff.
Audi	a6	3	56.000	15.6	3,00	294.569	799.900	505.331
Audi	a6	4	58.000	15.6	3,00	266.862	719.900	453.038
Audi	a6	3	55.000	15.6	3,00	295.340	729.900	434.560
Audi	a6	3	40.000	16.9	3,00	308.390	699.900	391.510
Audi	a5	1	17.000	16.1	1,80	269.480	649.900	380.420
Volvo	xc60	2	39.000	15.6	2,40	280.966	649.900	368.934
Audi	a6	3	107.000	15.6	3,00	264.506	629.900	365.394
Volvo	v60	2	29.000	55.6	2,40	289.060	649.900	360.840
Volvo	xc60	3	82.000	14.7	2,40	231.833	579.900	348.067
Audi	a6	1	4.000	22.7	2,00	304.058	649.900	345.842

Udstyr

Modellen har sværest ved at prissætte de dyreste modeller, som typisk har en del ekstraudstyr, så måske vil inddragelse af viden om bilernes udstyr have relevans. *Udstyr* indeholder en række prædefinerede udstyrstyper (klimaanlæg, servostyring, armlæn, etc.). Der fremkom i alt 57 dummyvariable, som en ad gangen er tilføjet som forklarende variabel til den primære model. En SAS-macro er anvendt til at loope over 57 forskellige modelspecifikationer med ens grundstamme og tilføjet en enkelt dummy variabel for, om bilen har eksempelvis klimaanlæg.

Nedenstående tabel viser modelformulering samt AIC for modellen sorteret efter AIC. Række 10 er den oprindelige model, hvorved vi har ni modeller, som umiddelbart giver en forbedring af modellen.

Obs	Model	AIC
1	log_price = age|bc_mileage motorsize Regnsensor	-3641.13
2	log_price = age|bc_mileage motorsize Laederindtraek	-3578.98
3	log_price = age|bc_mileage motorsize Xenonlygter	-3540.25
4	log_price = age|bc_mileage motorsize Dellaedersaeder	-3534.24
5	log_price = age|bc_mileage motorsize Aircondition	-3532.55
6	log_price = age|bc_mileage motorsize Laederrat	-3524.70
7	log_price = age|bc_mileage motorsize Automatgear	-3518.87
8	log_price = age|bc_mileage motorsize Tagraeling	-3514.55
9	log_price = age|bc_mileage motorsize Taagelygter	-3511.78
10	log_price = age|bc_mileage motorsize	-3511.39

Regnsensor og xenonlygter synes at have en betydning for bilens pris, hvorfor begge medtages. Målt på AIC og BIC opnår denne model en væsentlig forbedring, i forhold

til den tidligere model. Alle parameterestimater har de fortegn, som umiddelbart forventes og er i øvrigt i samme niveau, som den primære model.

Solution for Fixed Effects					
Effect	*Estimate*	*Std Error*	*DF*	*t Value*	*Pr > |t|*
Intercept	12.1068	0.07261	45.9	166.74	<.0001
Age	-0.02233	0.003999	3595	-5.58	<.0001
bc_mileage	-0.00081	0.000066	3596	-12.43	<.0001
age*bc_mileage	-0.00034	0.000014	3595	-23.89	<.0001
motorsize	0.2848	0.01069	3608	26.65	<.0001
regnsensor	0.06808	0.005689	3596	11.97	<.0001
xenonlygter	0.05097	0.008151	3596	6.25	<.0001

Generelt er estimaterne robuste over for indflydelsesrige observationer, alternative kovariansstrukturer, andre modelspecifikationer for fixed effects og anden stikprøveudtagning. Modellen har gode preskriptive evner, dvs. evner til at beregne, hvor meget mere en bil bør koste, hvis den har eksempelvis 0,2 liter større motor.

Konklusion

Overordnet er det lykkedes at finde en model, som ud fra en analytisk synsvinkel gør det ganske godt. Såvel fixed effects som random effects estimaterne er intuitive og residualerne er tilnærmelsesvist normalfordelte. Modellens evne til at prædiktere en bils faktiske udbudspris - særligt modellens evne til at vurdere out-of-sample annoncer - er dog knapt så imponerende, heller ikke efter dummyvariable for ekstraudstyr som regnsensor og xenonlygter er tilføjet modellen. Parameterestimaterne synes forholdsis valide og robuste, hvilket gør modellen anvendelig i forhold til preskriptiv statistik.

Der er behov for, at få en mere præcis model. En oplagt kandidat til videreudvikling er at inddrage ejerens beskrivelse af bilen, som en forklarende faktor ved hjælp af tekstanalyse, ligesom det kan være interessant om resultaterne er konsistente over tid.

Referencer:

http://www2.sas.com/proceedings/sugi29/188-29.pdf

http://www2.sas.com/proceedings/sugi29/189-29.pdf

http://www2.sas.com/proceedings/forum2007/178-2007.pdf

http://support.sas.com/resources/papers/proceedings12/332-2012.pdf

http://support.sas.com/resources/papers/proceedings13/433-2013.pdf

http://support.sas.com/resources/papers/proceedings14/1869-2014.pdf

Experimental Evidence on Informational Value of Auditor Assurance Reports used for Bank-lending to Small Enterprises

Claus Holm, Department of Economics and Business Economics, Aarhus University. Contact: hoc@econ.au.dk

Jakob Dahl Jensen, Aarhus University

Abstract: We examine how auditors' assurance on financial statements of small enterprises affects bank loans. Most enterprises in Denmark now have the option to choose between two types of audits. Both types are termed reasonable assurance engagements thus enabling the auditor to provide an opinion in a positive form whether the financial statements are free from material misstatement. In an experimental study we examine whether the signal of assurance through "pure audits" is stronger than assurance through "extended reviews" when deciding on bank lending to small enterprises. The subjects participating in the experiment are senior bank personal typically associated as business advisors or managers of customer services. Using a 2*2 between-subjects experimental design we examine the effect on enterprise credit rates by manipulating two factors; namely assurance type and opinion form. We find support for the first main effect (assurance type) which suggests that small enterprises will receive higher credit rating as an effect of audit as compared to an extended review (i.e., associated with lower auditor assurance). When the main effect for the two types of assurance engagements are further examined, simple main effect suggests that the effect is prevalent when the enterprise receives a clean report, but not necessarily when the audit opinion is modified. We also find support for the second main effect (opinion form) which suggests that an enterprise will receive higher credit rating when the auditor's opinion is provided as a clean opinion as compared to a modified opinion (i.e., possibility of a material misstatement in the financial reports). The simple main effects are supported for this factor. Both for extended review and audit the credit rate will be higher in case of clean opinion as opposed to modified audit report. Overall we find no support for interaction effect between the two factors affecting the informational value of auditor reports on small enterprises.

Introduction

In the Danish fairy-tale written by Hans Christian Andersen the ugly duckling turned into a beautiful swan. We explore why the new assurance service for small enterprises in Denmark labelled "extended reviews" could be considered an ugly duckling and offer experimental evidence supporting that some swan-like potential is hidden here.

In Denmark "small enterprises" have the least extensive reporting requirements; the so called class B requirements in the Financial statement Act (2015). This latest version

of the Act now classifies about 95 percent of Danish enterprises as small, while medium sized and large companies all together amounts to the remaining 5 percent. The small enterprises are those enterprises which do not exceed any two of the "44;89;50-threshold" at the balance sheet date of two consecutive financial years (i.e. thresholds of: (i) balance sheet total of DKK *44* million; (ii) revenue of DKK *89* million; and (iii) an average of *50* full-time employees during the financial year). Very small enterprises (*4;8;12*) are exempt from the audit requirement, while the remaining small enterprises can choose between two types of audits: pure audits and extended review (the ugly duckling). Both types belong to the category called "reasonable assurance engagements." In such engagements the auditors' opinion shall express whether the subject matter gives a true and fair view or has been prepared in accordance with the assessment and measurement criteria applied. This is an opinion declared in a positive form.

Small enterprises which are exempt from the audit requirement may voluntary chose to have a limited assurance engagement (a review). In such engagements the auditors' opinion shall express whether, in connection with the work performed, the auditor has come to attention of any matters that give rise to concluding that the subject matter does not give a true and fair view or has not been prepared in accordance with the assessment and measurement criteria applied. Here the auditor's opinion is declared in a negative form. While "pure audits" and "reviews" are recognized engagements in international audit regulation, the "extended review" is a Danish departure classified in regulation-terms as belonging to the review-type engagements, but in the Danish legislation classified as belonging to the audit-type engagements (hence the ugly duckling label used in this paper).

Research Question

The financial audit is considered an indispensable corporate governance mechanism for public companies operating in capital markets (OECD, 2004). The demand for assurance provided through audits can be explained in the context of the agency problems associated with the separation of ownership and control in public companies (Jensen and Meckling, 1976; Fama and Jensen, 1983; Watts and Zimmerman, 1983). In order to lower the information asymmetry between management and absentee owners and thus lowering agency costs, it is paramount to ensure assurance on the company's governance and reporting processes (e.g., Farber, 2005; Satava et al., 2006). While the purpose of the audit as well as the declaration of the responsibilities of the auditor forms part of the disclosed information, the quality of audits is hard to discern from the companies' audit report disclosures (Holm and Zaman, 2012; Francis, 2011).

In Denmark it has been a tradition to require audits of almost all enterprises independent of size. However, in recent years audit exemptions for the smallest enterprises have been introduced (subsequent exemptions in the Financial Statements

Act 2006, 2010 and 2013). While the exemptions were made based on legislators arguments of lowering administrative costs for smaller enterprises, the experience has been that either the audit was kept on a voluntary basis (thus leading to no decrease in costs) or the absence of audit assurance led to new additional costs - for example from bank lenders, which were not satisfied with limited assurance from voluntary reviews or no assurance at all. The risk of business failures, potential for tax evasions and the overall user experience with the audit exemptions of 2006 and 2010 led the government to introduce the option of extended review assurance in 2013. Hence, the aim of the new reporting standard has been to strike a balance between social considerations to make it easier to do business and to maintain a reasonable control, quality and integrity of the information that flows from the small enterprises to business partners, lenders and authorities.

We know very little about how auditors' assurance on financial statements of small enterprises affect bank loan approvals or the terms and conditions of loans once approved. A report from Copenhagen Economics (2014) proposes that small enterprises should deduct increased cost in lending from the cost savings on audit and with support from a Finish study further suggests that the net cost-effect may be negative. Hence, this provides motivation for studies trying to quantify the impact on credit rating from various levels of auditor engagements. We do not assume that the informational content of an auditor's assurance report is sufficient for making loan decisions by lenders by its own. However by considering variations in the signal provided by the auditor we aim to ascertain the existence of differential informational content.

RQ: Is the signal of assurance through pure audits stronger than assurance through extended reviews when deciding on bank lending to small enterprises?

Initially, we may distinguish between three types of assurance engagements. In a **review**, the auditor will by means of performing analyses and conducting inquiries be enabled to state with limited assurance (in Danish: begrænset sikkerhed) that the financial statements are free from material misstatement. In an **extended review**, the review is supplemented (extended) by verification of information by use of external sources, that is: extracts from the Land Registry and the Registry of Persons and Motor Vehicles; bank letters confirming engagements with banks; inquiry of enterprise's attorney of potential disputes, lawsuits and litigation; and verification that the reporting to the tax authorities of tax withheld at source, labour market contributions, payroll tax and VAT is stated in accordance with the bookkeeping records (Bek. 385, 2013). The additional procedures are assumed to allow the auditor to state in a positive form that the financial statements are free from material misstatement. However the level of assurance is lower than in a pure audit. An **audit** requires both more depth and use of audit procedures dedicated toward verification of information, i.e., enabling the auditor to state with reasonable assurance (in Danish: høj grad af sikkerhed) that the

financial statements are free from material misstatement. In this study we will focus on the implication of assurance type on bank lending by considering the two audit alternatives, namely pure audits and extended reviews.

In addition, the signal provided by the auditor's report may likely be influenced by the form of opinion. While most assurance engagements will result in the auditor stating a **clean opinion** (also known as an unmodified report), the audit standard ISA 705 establishes three types of **modified opinions**, namely, a qualified opinion, an adverse opinion, and a disclaimer of opinion. The decision regarding which type of modified opinion is appropriate depends upon: (a) The nature of the matter giving rise to the modification, that is, whether the financial statements are materially misstated or, in the case of an inability to obtain sufficient appropriate audit evidence, may be materially misstated; and (b) The auditor's judgment about the pervasiveness of the effects or possible effects of the matter on the financial statements. In this study we will focus on the implication of opinion form on bank lending by considering the two alternatives which may be stated following both an audit and an extended review. Thus we distinguish between a clean opinion and the type of modified opinion in which the auditor states that there is a possibility of a misstatement in the client's financial records that is "material but not pervasive." In this case, the auditor states that everything in the audit is fairly presented with the exception of a specific item, and why that item is an exception.

Design and method

Participants

In order to elicit information from a pool of subjects with the relevant knowledge and experience in credit assessment and loan approval, the subjects chosen from local bank branches were senior personal typically associated as business advisors or managers of customer services. A total of 63 subjects completed an online survey holding the research instrument after an initial agreement by phone to participate.

Research design and instrument

In order to examine the informational content of different assurance reports, four different versions of the research instrument were distributed randomly to the subjects. The signal provided to the subjects had the same basic form varying only by two factors name the Assurance Type (Audit assurance or Extended Review assurance) and Opinion Form (Clean opinion or Modified opinion). Hence, the research design was set up as a 2*2 between-subjects factorial design. The manipulated (independent variables) were both binary, while the dependent response variable was elicited as the credit rate (credit worthiness) on a 5 point Likert scale with 1 being low and 5 being high. Because we only received one response at both end-points, we truncated the scale and converted the responses to three categorical responses from 1 to 3.

The assurance report follows a set form and order prescribed by audit standards and in the research instrument it holds the following information across the four cases:

- Who is the addressee? Receiver of the report (identical across all cases)
- Classification of type of assurance (audit engagement in case 1 and 3, extended review in case 2 and 4)
- Managements' responsibility for the financial statements (identical across all cases)
- Auditors' responsibility (audit engagement in case 1 and 3, extended review in case 2 and 4)
- Opinion (clean/unmodified opinion in case 1 and 2, modified opinion in case 3 and 4)
- Auditors assessment of the management report (identical across all cases)

The modified opinion is described in identical manner in both the audit and extended review. While a focus on inventory is not part of a regular extended review, the auditor must address this if he becomes aware of potential material influence on the financial statements during the course of his investigation. In order to provide an identical manipulation across the cases, the information in the modified opinion is based on the auditors' assessment that the inventory of the small enterprise is overvalued and therefore there is a need for a one-time depreciation with an after-tax effect on the annual income and equity amounting to DKR 3 Mio.

Hypotheses

We hypothesize directional main effects on credit worthiness based on the two factors in our experiment, namely Assurance Type and Opinion Form:

> H1: The bank credit rating of a small enterprise is higher for an audit assurance than for an extended review assurance engagement

> H2: The bank credit rating of a small enterprise is higher for a clean/unmodified assurance opinion than for a modified opinion

In the absence of prior expectations in relation to conditional relationship between the two main factors we hypothesize the interaction effect to be zero (null-form):

> H3: The effect of Assurance Type on the bank credit rating of a small enterprise does not depend on Opinion Form (no interaction effect).

Results

In table 1 the average on the elicited dependent variable "credit rates" is shown for each of the four cases as well as the averaged totals. Enterprises which have clean audits have the highest credit rates (2.87 on average on 3 point scale) while enterprises with modified opinions derived from the auditors extended reviews have the lowest credit rates (1.47 on average).

Table 1. Average credit rates in 2*2 experimental design.
Case manipulation: Opinion Form*Assurance Type

		Assurance Type		
		Audit	Extended Review	Total
Opinion Form	Clean	CASE 1 2.87 n=15	CASE 2 2.25 n=16	2.55 n=31
	Modified	CASE 3 1.87 n=15	CASE 4 1.47 n=17	1.66 n=32
	Total	2.37 n=30	1.85 n=33	2.10 n=63

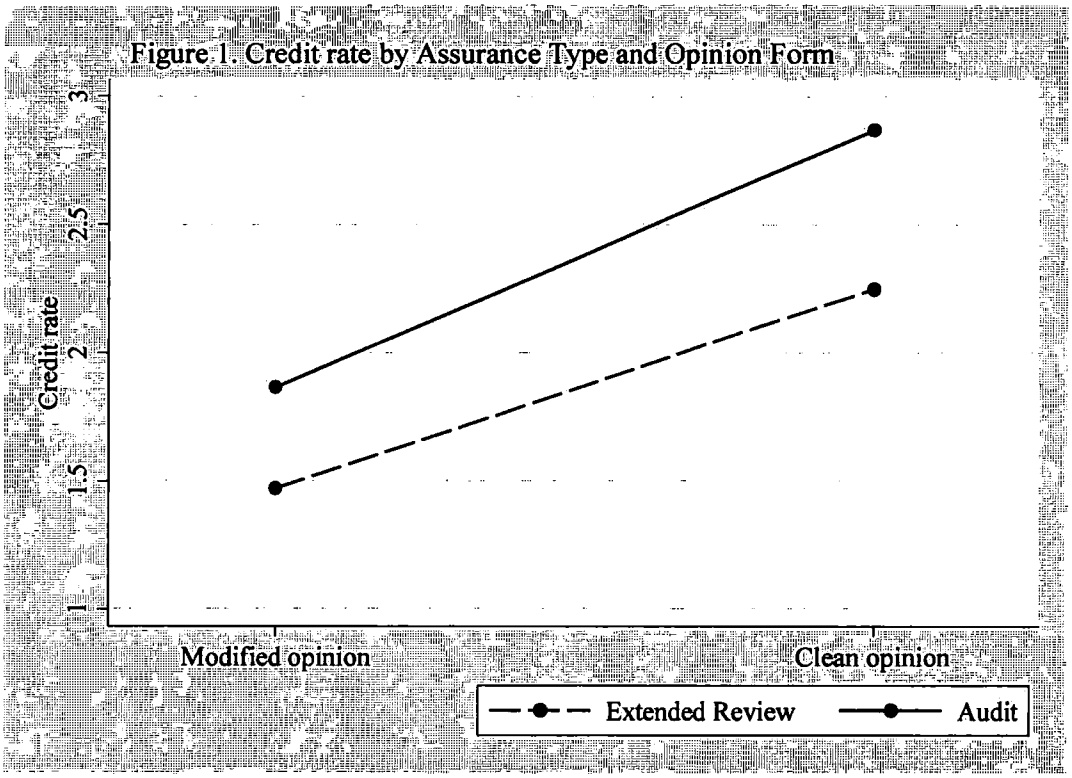
The dependent variable "Credit rate" is measured on ordinal scale 1-3.

First we treat the measure for credit rates as interval scaled. Under the assumption that the distribution of the sample means are normally distributed we run a two-way ANOVA. As shown in table 2, the hypothesized main effects are significant in the predicted directions. That is H1 is supported; enterprises with audits receive a higher credit rate than enterprises with extended reviews (F=14.23; p=0.0004). Likewise H2 is supported; enterprises receiving a clean opinion in the auditor report receive a higher credit rate than enterprises with a modified opinion (F=43.92; p=0.0000). As suggested by the ANOVA in table 3 and parallel lines in figure 1, the data suggests no interaction effect (H3).

Table 2 ANOVA

Source	Partial SS	df	MS	F	Prob>F
Model	16.7266	3	5.5755	19.70	0.0000
H1 AssuranceType	4.0276	1	4.0276	14.23	0.0004
H2 OpinionForm	12.4336	1	12.4336	43.92	0.0000
H3 Interaction	0.1911	1	0.1911	0.67	0.4146
Residual	16.7020	59	0.2831		
Total	33.4286	62	0.5392		

Model summary: n=63, Root MSE= 0.5321, R-squared=0.500, Adj R-squared= 0.475



Next we assume that our dependent variable credit rates is going to be treated as ordinal under the assumption that the levels of credit worthiness have a natural ordering (low, middle, high), but the distances between adjacent levels are unknown. We run an ordered logistic regression with interaction (results parallel to table 2 not shown) and without interaction. In table 3 we first show the regression coefficients in terms of ordered log-odds (logits) and in addition coefficients in terms of proportional odds. The Log likelihood of the fitted model is -45.6 (convergence after the fifth successive iteration of maximum likelihood estimation). The reported R-square is McFadden's pseudo R-squared as applied in STATA.

The ordered logistic regression supports that both AssuranceType (H1) and OpinionForm (H2) are significant factors in explaining credit rate. In table 3, we first report the ordered log-odds (logit) regression coefficients. Standard interpretation of the ordered logit coefficient is that for a one unit increase in the predictor, the response variable level is expected to change by its respective regression coefficient in the ordered log-odds scale while the other variables in the model are held constant. The cutpoints are thresholds used to differentiate the adjacent levels of the response variable credit rate. Cutpoint 1 (0.51) is the estimated cutpoint on the latent variable used to differentiate low credit rate from middle and high when values of AssuranceType and OpinionForm are evaluated at zero. Cutpoint 2 is the estimated cutpoint on the latent variable used to differentiate low and middle credit rate from high when values of the predictor variables are evaluated at zero. Responses that had a value of 4.022 or greater on the underlying latent variable that gave rise to our credit rate variable would be classified as high when values of AssuranceType and OpinionForm are evaluated at zero.

We have condensed a second output table into table 3 thus reporting the proportional odds ratios for the ordered logit model as well. We have tested the proportional odds assumption using Likelihood-ratio test of proportionality of odds. The non-significant result suggest that the assumption holds, hence we interpret the derived odds ratios parallel to those from a binary logistic regression. For AssuranceType, we would say that for a one unit increase, i.e., going from Extended Review to Audit, the odds of high credit rate versus the combined middle and low categories are 7.492 greater, given that all of the other variables in the model are held constant. Likewise, the odds of the combined middle and high categories versus low apply is 7.492 times greater, given that all of the other variables in the model are held constant. For a one unit increase in OpinionForm, i.e., going from modified to clean opinion, the odds of high credit rate versus the low and middle categories of credit rate is 31.617 times greater, given that the other variables in the model are held constant. Because of the proportional odds assumption, the same increase, 31.617 times, is found between low credit rate and the combined categories of middle and high credit rate.

Table 3 Ordered logistic regression (logit coeffients and proportional odds ratios)

	Coef	Std.err	z	P>|z|	95% Conf.Interval		Odds ratio	Std.err
H1 AssuranceType	2.014	0.577	3.49	0.000	0.884	3.144	7.492	4.320
H2 OpinionForm	3.454	0.735	4.70	0.000	2.013	4.894	31.617	23.236
Cutpoint 1	0.510	0.455			-0.382	1.402	0.510	0.455
Cutpoint 2	4.022	0.756			2.541	5.503	4.022	0.756

Model summary: n=63, Log likelihood=-45.6120, LR chi2(2)=41.78, Prob>chi2=0.000, Pseudo R2=0.3142

Finally, we consider contrasts in order to explore the robustness of the main effects. We use the post-estimation procedure for pairwise comparison of marginal means in STATA. The six possible marginal contrasts in the 2*2 design are reported as significant at the 5% level when using pairwise comparison tests not adjusted for the number of comparisons. When adjustment is made to reflect the number of comparisons (using Bonferroni or Sidak type adjustments of the comparison-wise error rate based on the upper limit of the probability inequality) two of the six contrasts revert to insignificant differences. In table 4 the individual contrasts are listed in order of the size of difference between credit rating as measured by the three point interval scale.

Table 4 Test of individual contrasts of marginal means (sorted by size of difference)

1) Is the credit rate higher for clean audit (2.86) than for modified extended review (1.47)? **Yes.** The informational value of an audit receiving clean opinion is expected to be higher both due to the lower assurance provided by extended review and due to the problem identified in the modified opinion. This is a combination of the two main effects.
2) Is the credit rate higher for clean audit (2.86) than for modified audit (1.86)? **Yes.** The informational value of an audit receiving clean opinion is expected to be higher than an audit with identified problem. This is consistent with support for the directional prediction for the main effect OpinionForm.
3) Is the credit rate higher for clean extended review (2.25) than for modified extended review (1.47)? **Yes.** The informational value of an extended review receiving clean opinion is expected to be higher than extended review with identified problem. This is consistent with support for the directional prediction of the main effect OpinionForm.
4) Is the credit rate higher for clean audit (2.86) than for clean extended review (2.25)? **Yes.** The informational value of an audit receiving clean opinion is expected to be higher than for an extended review receiving clean opinion. This is consistent with support for the directional prediction of the main effect AssuranceType.
5) Is credit rate higher for modified audit (1.86) than for modified extended review (1.47)? **No.** The informational value of an audit with identified problem is expected to be higher than for an extended review with the same problem. The test result is inconsistent with the support for the directional prediction of the main effect AssuranceType as well as inconsistent with the absence of an interaction effect.
6) Is credit rate higher for clean extended review (2.25) than for modified audit (1.86)? **No.** No prior expectation has been suggested as to the difference in informational value of an extended review receiving a clean opinion as compared to an audit with identified problem. This is a combination of the two main effects.

Conclusion

Most enterprises in Denmark now have the option to choose between two types of audits. Both types are termed reasonable assurance engagements thus enabling the auditor to provide an opinion in a positive form whether the financial statements are free from material misstatement. In an 2*2 experimental design we examine whether the signal of assurance through "pure audits" is stronger than assurance through "extended reviews" when deciding on bank lending to small enterprises. While "pure audits" and "reviews" are recognized engagements in international audit regulation, the "extended review" is a Danish departure classified in regulation-terms as belonging to the review-type engagements, but in the Danish legislation classified as belonging to the audit-type engagements (an ugly duckling). Our findings suggest that small enterprises will receive higher credit rating as an effect of audit as compared to an extended review. Hence the option for the small enterprise to choose the less expensive extended review comes with a cost of lower credit rating and likely less favourable loan conditions. In a sense this suggests that banks recognise the extended review as an ugly duckling. Our findings also suggest that an enterprise will receive higher credit rating when the auditor's opinion is provided as a clean opinion as compared to a modified opinion (i.e., possibility of a material misstatement in the financial reports). The simple main effects are supported for this factor. Both for

extended review and audit the credit rate will be higher in case of clean opinion as opposed to modified audit report. Hence, the extended review has some swan-like potential. Our findings specifically support that the banks recognise the informational value of extended reviews as higher when the opinion is clean as opposed to identifying a possible material misstatement. Therefore the current challenge is to establish an understanding of the circumstances under which the extended review assurance is sufficient to support the information needs of the external users when making economic decisions contingent on reliance on the financial statements from small enterprises.

References

Copenhagen Economics (2014) 'Effekt af ændret revisionspligt for mindre virksomheder', FSR – Danske Revisorer.

Danish Executive Order on Approved Auditors' Reports (2013) 'Bekendtgørelse om godkendte revisorers erklæringer' (BEK nr 385 af 17/04/2013).

Fama, E. F. and Jensen, M. C. (1983) 'Separation of Ownership and Control', *Journal of Law & Economics,* 26, pp. 301-326.

Farber, D. B. (2005) 'Restoring Trust after Fraud: Does Corporate Governance Matter?', *Accounting Review,* 80, pp. 539-561.

Financial Statements Act (2015) 'Årsregnskabsloven' (LBK nr 1580 af 10/12/2015).

Francis, J. R. (2011) 'A Framework for Understanding and Researching Audit Quality', *Auditing,* 30, pp. 125-152.

Holm, C. and Zaman, M. (2012) 'Regulating audit quality: Restoring trust and legitimacy', *Accounting Forum,* 36, pp. 51-61.

Jensen, M. C. and Meckling, W. H. (1976) 'Theory of the Firm: Managerial Behavior, Agency Costs and Ownership structure', *Journal of Financial Economics,* 3, pp. 305-360.

OECD (2004) Principles of Corporate Governance. *OECD, Paris.*

Satava, D., Caldwell, C. and Richards, L. (2006) 'Ethics and the Auditing Culture: Rethinking the Foundation of Accounting and Auditing', *Journal of Business Ethics,* 64, pp. 271-284.

Watts, R. L. and Zimmerman, J. L. (1983) 'Agency Problems, Auditing, and the Theory of the Firm: Some Evidence', *Journal of Law and Economics,* 26, pp. 613-633.

HECKMAN'S PARADOX AND QUALITY OF EARLY UNIVERSAL INVESTMENTS IN HUMAN CAPITAL: A RESEARCH NOTE

Mogens Nygaard Christoffersen, senior researcher emeritus, SFI

INTRODUCTION

Sociologists and economists as Heckman are concerned about the growing polarization of American society and its implications for productivity; Heckman and colleagues were interested in the effectiveness of early interventions in offsetting these trends. "America has a growing skills problem. One consequence of the skills problem is rising inequality and polarization of society". "Another consequence of the skills problem is the slowdown in growth of the productivity of the workplace" (Heckman, 2008). The argument is that policies that supplement the child rearing resources available to disadvantaged families reduce inequality and raise productivity in US economi.

HECKMAN'S THEORY

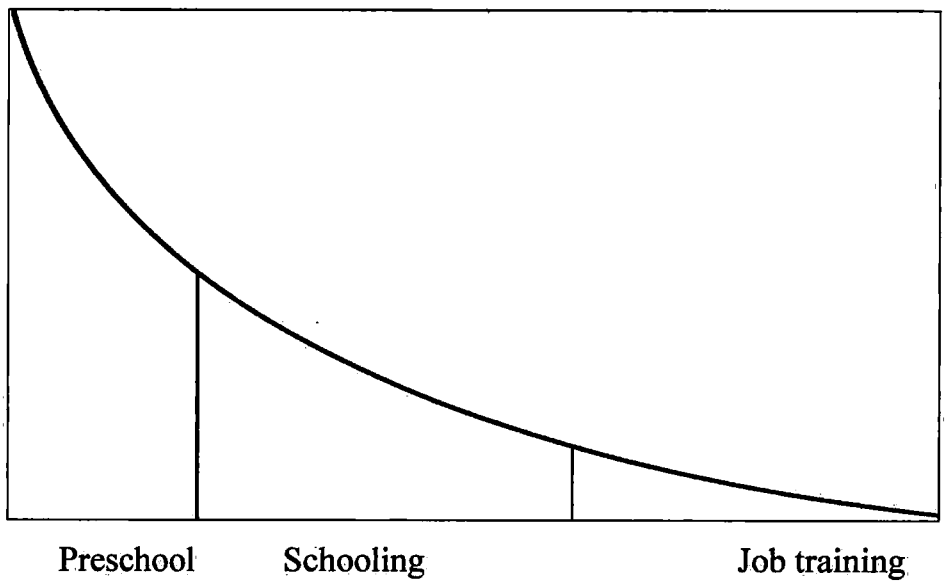
Heckman and Cunha find that about half of the inequality in the present value of lifetime earnings is due to factors determined by age 18 and they formulated a theory about early investment in human capital (Heckman, 2008).

The theory is that there is a dynamic complementarity, or synergy, in early investments compared to investment later in life (Heckman & Masterov, 2007). The dynamic complementary theory in cognitive and non-cognitive development could be illustrated in a simple model. Figure 1 shows the return to a marginal increase in investment at different stages of the life cycle starting from a position of low but equal initial investment at all ages.

Heckman assume that the early investments will give relatively higher adulthood earnings. The other side of the coin is that grossly inadequate investment in cognitive and non-cognitive ability in the early year will have devastating consequences. Heckman argues that the known examples of children who received minimal social and intellectual stimulation during their formative years demonstrated cognitive delays, serious impairments in social behaviour and abnormal sensitivity to stress (Heckman, 2008). Young children late-adopted out of this kind of deprivation care often have persisting cognitive, socio-emotional and health problems, compared to early-adopted children (Dennis, 1973; Kreppner, O'Connor, & Rutter, 2001;

O'Connor, Rutter, Beckett, Keaveney, & Kreppner, 2000; Perry, 2007). In other words, lack of investments increases vulnerability in a child while early investment in a child's cognitive and non-cognitive development creates robustness.

FIGURE 1.
Rate of return to investment in human capital.



Source: (Cunha, Heckman, Lochner, & Masterov, 2006; Heckman, 2008).
Note: Imaginary numbers illustrating the hypotheses

Interventions during preschool or in kindergarten improve cognitive and non-cognitive skills in a lasting way. Earnings and productivity are improved (Heckman, Moon, Pinto, Savelyev, & Yavitz, 2010; Kautz, Heckman, Diris, Ter Weel, & Borghans, 2014).

The dynamic complementary means that capabilities produced at one stage of life raise the productivity of investment at subsequent stages. In order for the early investment to be productive, early investment should be followed up by later investment. The stock of capabilities at stage t+1 is a function of all past investments t=1, 2, 3,...T. Stocks of capabilities acquired by stage t-1 make investment at stage t more productive (Heckman, 2008).

Two equal investments at time t, one investment in children with low capabilities, and a similar investment in children with high capabilities, will consequently result in higher productivity in the latter group than in the former. In

other words, it seems as if an universal investment at time t then will result in a higher inequality at time t+1 between children with low capabilities and children with high capabilities. This constitutes a contrast to focus on policies that supplement the child rearing resources ability to reduce inequality and raise productivity in U.S. economi.

THE PUZZLE

Evidence from enriched preschool programs that target high-risk children from disadvantaged families have shown long-term effects on children's skill formation (Heckman & Masterov, 2007). According to Heckman's theories, early investment in cognitive and non-cognitive developmental abilities will give higher return than investment at a later stage. Skills or abilities at stage t+1 (S_{t+1}) is a positive function of Skills at stage t (S_t) and Investments at stage t (I_t). Investments include all inputs invested in the child including parental and social inputs as for example enriched preschool programs. The technology of skill formation can be written as

$$S_{t+1}=f_t(S_t,I_t).$$

The enrichment programs that are mentioned in the presentation of the theory are for example 'Abecedarian program', or 'High/Scope Perry Preschool program'. The research question is: what will happen if these programs are enlarged to most of the children? Intuitively, such universal investment in an intensive preschool program targeting *all* children is expected to increase inequalities because resourceful children who had been well stimulated at home will gain more human capital by the added universal investment than less resourceful children who had previously been under stimulated. Unexpectedly, some early universal investments in enriched preschool programs increase later productivity more in disadvantaged children than in resourceful children. We will try to give a possibly explanation to this paradoxical situation.

EVIDENCE ON LIFE CYCLE SKILL FORMATION

Heckman and colleagues were inspired of some of these early randomized controlled trails (RCT) with high quality preschool interventions in disadvantaged areas in Michigan, Carolina and Wisconsin (Milwaukee) during the 1960s to 1980s. A comprehensive systematic review of randomised trials showed a remarkable increase in cognitive development in preschool children in the short and long term measures (Zoritch, Roberts, & Oakley, 2000). The review included day-care defined as non-parental day-care without any specification of quality. Only few RCT studies followed the children into adulthood.

Another systematic review of 84 programs included both random assignment and less-rigorous quasi-experimental methods (Duncan & Magnuson, 2013). The study found that programs beginning before 1980 produced larger effect sizes than those that began later. The likely reason for the decline is - according to Duncan & Magnuson - that conditions for children in initial conditions for both the experimental group and the group in these studies have improved substantially (Duncan & Magnuson, 2013). This could be in combination with a high quality in the first enriched programs. A possible decreased engagement among the later preschool teachers compared to the enthusiastic starters, could also play a part.

A third systematic review included 1) only randomly assignments and secondly, 2) only the mentioned high quality preschool interventions e.g. Abecedarian or High/Scope Perry Preschool program (Christoffersen, Højen-Sørensen, & Laugesen, 2014; Nielsen & Christoffersen, 2009). And 3) only studies with at least a cognitive outcome at the age 4 to 5 years old. The systematic search included five studies (Table 1 and Figure 2). The children in one of these studies were followed longterm and shoved lasting positive effects on such outcomes as greater educational attainment, higher earnings, and lower rates of crime (Schweinhart, Barnes, & Weikart, 1993; Barnett & Belfield, 2006; Duncan & Magnuson, 2013).

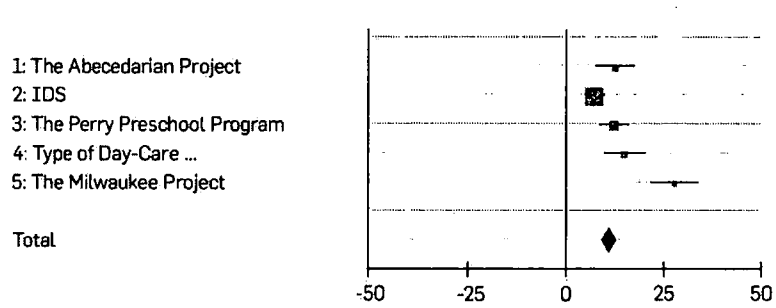
This enrichment program is an example per excellence of an early investment in cognitive and non-cognitive developmental abilities (Heckman & Masterov, 2007). The program is an intensive program that target cognitive abilities as well as socioemotional skills, physical and mental health, perseverance, attention, motivation, and self-confidence. The researchers have described program in details in order to make it possible to repeat it elsewhere (Sparling & Lewis, 1981; Sparling & Lewis, 1984; Bryant, Ramey, Sparling, Wasik, & Graham, 1987; Wasik, Ramey, Bryant, & Sparling, 1990; Hohmann, Banet, & Weikart, 1979; Schweinhart et al., 1993; Weikart, 1972; Weikart, Epstein, Schweinhart, & Bond, 1978; Deutsch, 1966; Deutsch et al., 1971).

The review found remarkable differences in the age 54 months old. The 'Stanford Binet Intelligence Scale' or 'McCarthy Scales of Children's Abilities' showed significant differences in all five studies (Table 1 and Figure 2). The improvements were in average 11.3 (CI 9.5-13.0) as consequence of the mentioned high quality preschool program (Christoffersen et al., 2014; Nielsen & Christoffersen, 2009). The high quality day-care have a well-documented effect on preschool children's cognitive development in disadvantaged children.

A very common problem in these studies is that the alternative treatment of the control groups are inadequate described, but we assume that these disadvantaged

children received a relatively poor treatment. Another problem is the heterogeneity between the samples, which indicate that they may be considered to be drawn from the different population (Table 1).

FIGURE 2. Mean IQ at 54 months of age in high-scope preschool vs. control group. Randomised controlled trials. Systematic review.



Source:(Christoffersen et al., 2014; Nielsen & Christoffersen, 2009)

TABLE 1. Mean IQ at 54 months of age in high-scope preschool vs. control group. Randomised controlled trials. Systematic review.

	High quality preschool			Control group					
Study	Mean IQ	SD	n	Mean IQ	SD	n	weight pct.	Mean differ.	CI 95 %
1. The Abecedarian Project	101,7	11,8	50	89,2	13,4	47	11,8	12,5	(7,5-17,5)
2. IDS	99,2	11,3	260	92	12,4	142	49,5	7,2	(4,7-9,7)
3. The Perry Preschool Program	95,5	11,5	58	83,3	10	65	20,4	12,2	(8,4-16,0)
4. Type of Day-Care	101,4	10,1	61	86,5	10,5	19	10,4	14,9	(9,5-20,3)
5. The Milwaukee Project	121,5	8,5	17	93,9	10,1	18	7,9	27,6	(21,4-33,8)
Total			446			291	100,0	11,3	(9,5-13,0)

Note: Test of heterogeneity ($p<0.0001$). Weighted total (Lipsey & Wilson, 2001).

1: "Abecedarian project" (Ramey & Campbell, 1979; Campbell, Pungello, Ramey, Miller-Johnson, & Burchinal, 2001); 2: "The Institute for Developmental Studies, IDS" (Deutsch, Taleporos, & Victor, 1974; Deutsch et al., 1971); 3: „Perry Preschool Project" (Schweinhart et al., 1993); 4: "Type of Day-

care" (Burchinal, Lee, & Ramey, 1989); 5: "The Milwaukee project" (Garber & Hodge, 1989; Garber, 1988). Source: (Christoffersen et al., 2014; Nielsen & Christoffersen, 2009)

The studies were experimental and only few were reporting long-term consequencies in adulthood such as increased earnings and improved employment (Barnett & Belfield, 2006; Campbell, Ramey, Pungello, Sparling, & Miller-Johnson, 2002; Reynolds et al., 2007; Reynolds, Temple, Robertson, & Mann, 2001; Schweinhart et al., 1993; Olds, 2002).

At first glance, these long-term studies seemed to support Heckman's dynamic complementary theory, where the children exposed to early investment in their cognitive development profited more of an enrichment preschool education than the children in the control group did. The results could also be seen as consequences of an intensive program's dynamic substitute to previous investments.

THE PARADOX

Heckman and colleagues find that "A subsidy for an intensive preschool program for *all* American children would be unnecessary. Most parents are able to provide good care themselves or to obtain surrogate care. Any subsidized program should be targeted carefully." And they consider providing the program only to certain subsets of the poor in the US (Heckman & Masterov, 2007). They conclude that scaling up the model programs to the target population (the most needed) add great value to the output of American society. Additional research is needed in order to better characterize the interaction between US government programs and market-provided child care (Blau & Currie, 2006).

In Denmark, for example, there have been universal subsidised Nurse Home Visiting Program, NHV to *all* newborn children, and likewise subsidised universal preschool, and subsidised universal kindergarten in some decades. In contrast to the vast majority of U.S. states, child-care in the Scandinavia countries are universal available to all children, and child-care are not targeted at the disadvantaged families any more. This is the background for the following research questions: What are the effects of the introduction of universal programs in the early years such as Nurse Home Visiting Program (NHV) or High/Scope Perry Preschool program (HSPP) or similar intensive program with universal subsides? Are these programs increasing polarization in the society or will these programs act as a substitute instead of complementary?

A popular way of studying effects of universal programs is to study so called 'natural experiments' where the program in focus suddenly changes. Table 2 shows the long-term effects of expansion of universal preschool investments. Consulting

state-of–the-art reviews shows that expansion of early education generally benefits disadvantaged children at school entry, adolescents and adults; however the gains may be less pronounced when subsidies increased the use of low quality care (Ruhm & Waldfogel, 2012; Christoffersen et al., 2014; Bauchmüller, Gørtz, & Rasmussen, 2012).

Several North American studies have explored the long-term impact for disadvantaged children (Deming, 2009), while nationwide child care program for children are common in Scandinavia and Europe, studies of universal care programs for preschool children are unusual in U.S. We only found a single study (Herbst, 2013). The study has evaluated the distributional impact of universal investments in children's cognitive and no-cognitive development and employment benefits in US (see Table 2). The child care program aimed at providing children ages 0 to 12 so that mothers could contribute to the nation's war production efforts regardless of family income. The program was heavily subsidized and universal child care program administered throughout the U.S. during World War II and stopped suddenly 1946. The results show that the benefits accrued largely to the most economically disadvantaged children when they became adults (Herbst, 2013).

Some studies of expansion of pre-primary education in India, Uruguay, and Argentine found without exception that the students living in disadvantaged areas gained much more than student from more affluent environments (Hazarika & Viren, 2010; Berlinski, Galiani, & Manacorda, 2008; Berlinski, Galiani, & Gertler, 2009). Some studies of sudden expansions of preschool capacity (see Table 2) that had taken place in a long list of European countries e.g. Norway (Havnes & Mogstad, 2015), Denmark (Bingley & Westergård-Nielsen, 2012) and France (Dumas & Lefranc, 2012).

Some of the studies had shown significant improvements in earnings and employment for persons who have joined the universal preschool interventions compared to the persons who by pure chance did not get the same opportunity (Havnes & Mogstad, 2015; Herbst, 2013; Dumas & Lefranc, 2012; Ruhm & Waldfogel, 2012).

Several studies found that children from immigrant backgrounds increased educational attainments relative to native-born children (Fredriksson et al., 2010; Spiess et al., 2003). Another study of the West Germany setting indicates that universally accessible care can contribute to decrease inequalities across children from different socio-economic backgrounds (Felfe & Lalive, 2013).

A study focuses on an early 1990's reform in Spain, which led to a sizeable expansion of public subsidized full-time childcare for 3-years old. They find a positive effect on children's cognitive development at least among children with less educated

parents and for girls. Effects are driven by girls and disadvantaged children (Felfe, Nollenberger, & Rodriguez-Planas, 2012). But not all universal investment in expansions of day-care capacities showed these long-term results which, for example, a Canadian study revealed (Baker et al., 2008).

Table 2. Long-term effects of sudden expansion of universal preschool.

Study	Cite	Expansion or sudden change	Distributional outcome
(Baker, Gruber, & Milligan, 2008)	Quebec, Canada	The subsidies increased the use of low quality care and level of quality were not maintained while expanding	The program led to significant negative effects on socio-emotional and health outcomes of children under the age of 5, and worse relationship quality and more hostile, less consistent parenting.
(Herbst, 2013)	USA	Expansion from 1943 to 1946, then closed	Distributional effect largely to the most disadvantaged adults
(Fredriksson, Hall, Johansson, & Johansson, 2010)	Sweden	Pre-school expansion between 1967 and 1982	Childcare attendance reduces the gap in language skills between children from immigrant backgrounds relative to native-born children, but no differential effects on inductive skills, or long-run educational attainment.
(Havnes & Mogstad, 2015)	Norway	Preschool expansion in late 1970's maintaining a high quality while expanding	Improvement in educational attainment, labour market participation were largest for children of low-educated mothers. The expiation led families to use child care-centres instead of informal care.
(Bingley & Westergård-Nielsen, 2012)	Denmark	Preschool expansion 1970's and early 1980's	The effect tends to be larger for disadvantaged children, particular daughters of less-educated mothers
(Spiess, Büchel, & Wagner, 2003)	Germany	Expansion during 1984 to 1994	Immigrants attending kindergarten have increased educational attainments compared to immigrants who did not attend kindergarten
(Felfe & Lalive, 2014)	Germany	Three reform: Introduction of center-based care, expansion of enrollments in two stages, maintained a high level while expanding	Effect tends to be stronger for children from disadvantaged background, in particular children of less-educated mothers or children of foreign parents.

(Felfe, Nollenberger, & Rodriguez-Planas, 2015)	Spain	Expansion of public subsidized full-time childcare for 3-years olds between 1990 and 2002.	Improvement in children's reading and math skills at age 15, and grade progression during primary and secondary school. Effects are driven by girls and disadvantaged children.
(Dumas & Lefranc, 2012)	France	Expansion of pre-schooling in the 1960s and 1970s	Wage level effect size is larger for children from disadvantaged socioeconomic background
(Filatriau, Fougère, & Tô, 2013)	France	Early enrolment in pre-elementary school in 1993, children born 1991	Improves literacy and numeracy from third to nine grades which is estimated to increase average monthly wage three years after leaving school
(Hazarika & Viren, 2010)	India	Governmental sponsoring early childhood developmental facilities	Positive effects for 13 to 19 years old, and speeds grade progression conditional on enrolment.
(Berlinski et al., 2008)	Uruguaya	Expansion in the supply of pre-primary education between 1992 and 2004.	Treated children have completed extra years of primary and secondary education. Pre-school exposure has much bigger impact on children whose mother is less educated and among those living outside the relatively more affluent area.
(Berlinski et al., 2009)	Argentina	Expansion of universal pre-primary education between 1993 to 1999 after pre-primary education became compulsory after 1993	Increases average test-scores in third grade in Spanish, mathematics, classroom attention, effort discipline, and participation. Gains from preschool education are bigger for the students living in more disadvantaged municipalities both in Spanish test score and Mathematics..

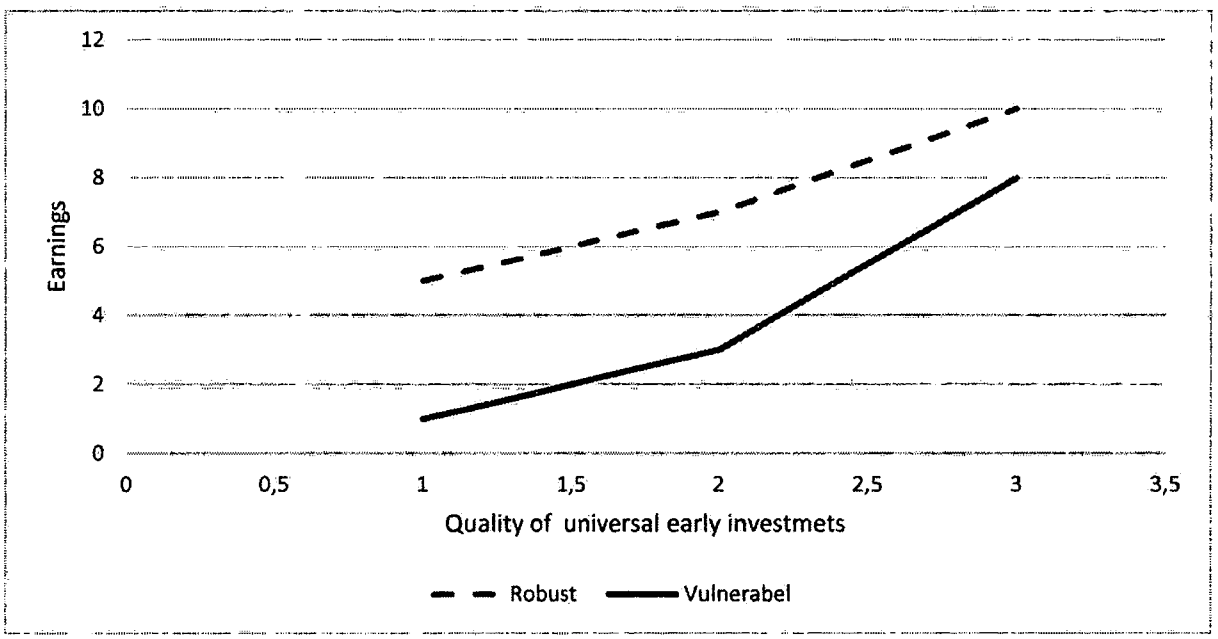
Source: (Bauchmüller et al., 2012; Gupta & Simonsen, 2010; Ruhm & Waldfogel, 2012)

According to the Heckman's dynamic complementary theory, we would intuitively expect an increased inequality as a result of universal early investment in human capital, because "The dynamic complementary theory" suggests that capabilities produced at one stage of life raise the productivity of investment at subsequent stages. When the children are receiving universal preschool programs, it is to be expected from the theory that resourceful children who had been well stimulated at home will gain more human capital than less resourceful children who had previously been under stimulated. Universal Nurse Home Visiting Program and universal preschool programs in the earliest years, such as an up-scaled version of "Abecedarian" or "The High/Scope Perry Preschool Study", would be expected to

increase inequality between children with initial differences in human capital, according to an immediate cognition of "The dynamic complementarity theory".

Contrary to expectations, some of the studies of universal preschool programs showed increased equality where children from the most disadvantaged families gained more than the children from the more affluent families (Havnes & Mogstad, 2015). In a Danish study, girls from families with low-income or mothers with no education or vocational training seemed to gain more than others (Bingley & Westergård-Nielsen, 2012). A French study showed accordingly that children from disadvantaged family background seem to gain more than children from a more affluent background (Dumas & Lefranc, 2012).

FIGURE 3 Earnings and quality of universal early investmets in human capital for robust and vulnerabel children.



Note: Imaginary numbers illustrating the hypotheses.

The principle of increased equality in earnings because of *high quality universal early investment* in human capital is illustrated in Figure 3. The assumption is that poor quality in universal preschool, kindergarten or primary schools will increase inequality in later earnings and productivity.

According to Heckman's theory, we would also expect that children who are exposed to positive health effects of the NHV program in infancy might benefit more from participating in HSPP program than those who did not have access to NHV. Contrary to these expectations a Danish study finds that the two interventions are substitutes instead of the Heckman's dynamic complementary effect (Rossin-Slater &

Wüst, 2015). This means that the children who were participating in the NHV program gained less of participating in the HSPP program than the children who did not participate in the NHV program. The supplementary is the opposite of the complementarity principle. These results cannot be extrapolated to the wider population because the HSPP program at the time was highly selective.

It seems as if Heckman's dynamic complementary theory cannot explain why some universal programs tend to reduce inequality when the universal programs supplement the child rearing resources available to both disadvantaged families and resourceful families. Contrary to expectations, it is often found that universal programs in the early years such as nurse home visitation (Eckenrode et al., 2000; Olds et al., 1998; Olds, Henderson, Tatelbaum, & Chamberlin, 1988; Olds, Henderson, Chamberlin, & Tatelbaum, 1986; Olds, 2007), or preschool program (Weikart, Berrueta-Clement, Schweinhart, Barnett, & Epstein, 1984; Schweinhart et al., 2005), act as a substitute instead of complementary. We have chosen to call it, the Heckman's Paradox.

POSSIBLE EXPLANATIONS

Several processes may explain the paradoxical situation. A reason why we sometimes find the paradoxical results is that less resourceful children gain much more from an early universal high quality program compared to their alternative care arrangements than it is the case among the children living in a more advantaged and stimulating environment. The more resourcefull children already live in a more advantaged and the program only take them to another stimulation environment, instead of taking them to a more stimulating environment, which is the case with the disadvantaged children. There are two problems with this explanation. One problem is that it rejects the heckman's theory about the technology of skill formation.

Meghir and Palme found for example similar results in a universal school reform, which increased final educational attainment and earning. The school reform increased years of education and levelled out income inequalities, because the entire effect is due to the increase in the educational attainment of those with unskilled fathers (Meghir & Palme, 2005). In this study the explanation was fairly simple. The relatively gain of the universal program turned out to be of various consequences in the school population according to the family background. This explanation could not be applied to the mentioned intensive preschool programs which are tagetting all children.

The other problem is that it seems as if ressoucefull children benefit relatively more of low quality day-care programs than more vulnerable children do. The theory

might be too simplistic and not including the dynamic processes in the day-care setting of high and low quality programs.

The research question is to explore under which conditions universal programs targeted towards the early years will tend to reduce inequality or under which conditions it tends to increase inequality in income in adulthood earnings. We suggest that low quality universal programs will increase inequality.

Quality parameters could be *structural* e.g. adult child ratio, educational level, number of children in groups, health and safety regulation or *processual* quality referring to the quality that characterizes the interactions between children and their caregivers. The structural features of child-care will support and facilitate more optimal interactions (Blau & Currie, 2006; Christoffersen et al., 2014; McGurk, Mooney, Moss, & Poland, 1995). The quality of the interaction between adults and children in preschool is seen as the most important factor in stimulating child cognitive and social development (Bronfenbrenner & Bronfenbrenner, 2009). The adults are able to introduce cognitive and social stimulating activities, such as learning games (Sylva, Roy, & Painter, 1980). The quality depends on adult's sensibility to the children's verbal and non-verbal communication. The quality depends on the adult's ability of immediate and appropriate response to the child's communication, and the adult's capacity to introduce learning games adapted to the child's age and development. The adults can elaborate and support the language, the child's sentiments, cognitive, social development and feeling (McGurk et al., 1995). Some studies indicate that the low quality preschool program will demonstrate low sensibility of the children's needs and exercise little contact between the child and the adults.

Consequences of the Heckman's Paradox is that the resourceful children are more robust and less sensible if the universal program have low quality while the disadvantaged children will resign oneself. Contrary to the vulnerable children, the robust children will be very insistent about having contact with the adults (Diderichsen, 1997; Weikart et al., 1984; Howes & Hamilton, 1992; Howes & Hamilton, 1992).

Some studies indicate that low quality universal programs increase inequalities between robust children and disadvantaged children and our hypothesis is that a low quality universal program will increase inequality in adulthood earnings and employment compared to universal programs of high quality. It seems as if both groups of children get the same intervention but this may be wrong. Despite the naming 'universal' programs the effect on disadvantaged and robust children might not be the same.

The sudden investments in expansions of day-care capacities were not always followed by effective restrictions on quality of the increase day-care capacity (Table 2). In the Norwegian case the focus was on quality parameters while in the Canadian case, the expansion considered of family based child care. The impacts of the Quebec's "$ 5 per day childcare" were that children were worse off in a variety of behavioural and health dimensions, ranging from aggression to motor-skills to illness. The professional preschool was not expanded accordingly, and the quality of day-care along the measured dimensions was not improved in the Canadian case (Baker et al., 2008). Although, low quality day-care arrangements may expand inequalities between children, this does not explain why high quality enrichment program decrease inequalities. Only low quality programs seeme to be complementary while high quality programs and more intence programs seemed to be supplementary.

The puzzle will not go away. We are still unable to explain why a low-quality program seems to increase inequality, and a high-quality evelopmental program in the early years universal to all children regardless of family income seems to reduce later income inequalities instead of increasing inequalities as expected from "The dynamic complementary theory in cognitive and non-cognitive development".

ACKNOWLEDGMENTS

Antti Kääriälä, a collegue and scholar at the University of Helsinki, has critised and supported the process with new ideas. The paper has also benefited from inspiration and insightful criticism from senior researcher Karsten Albæk, the Danish National Center for Social Reasearch.

[litteraturlisten kan fås hos forfatteren]

Smart Meter Data Analyse

Alexander Tureczek

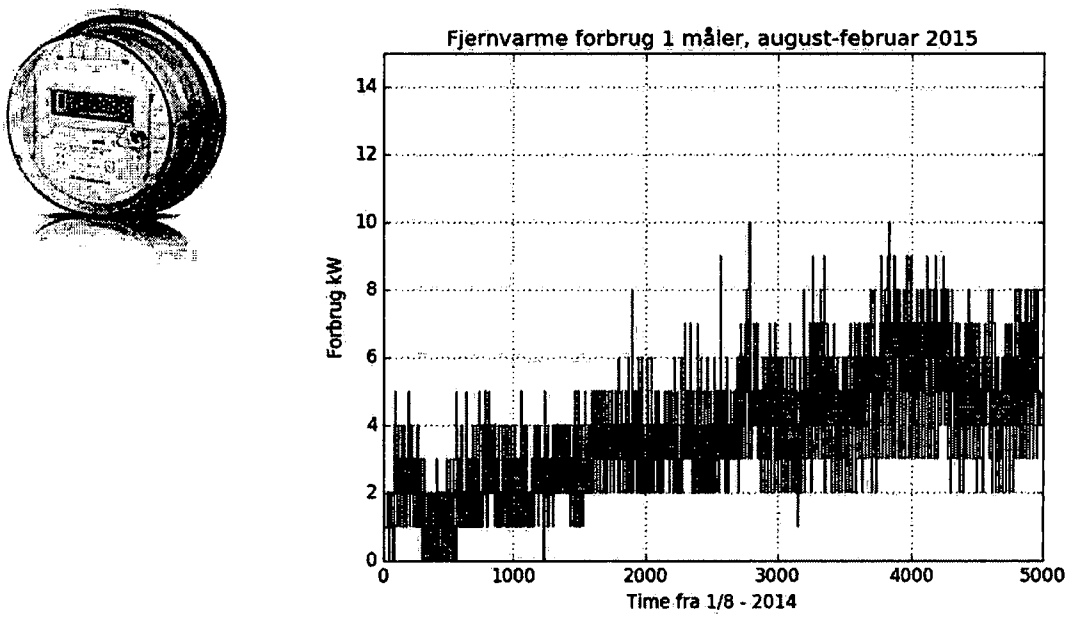
Systems Analysis, DTU Management Engineering

Introduktion

Senest med udgangen af 2020 skal elkunder i Danmark have installeret en smart måler. Jf. Bekendtgørelse om fjernaflæste elmålere og måling af elektricitet i slutforbruget. Målerne rapporterer automatisk elforbruget på kvarters basis til el leverandøren. Data skal indrapporteres til energinet.dk' datahub med time værdier svarende til 8700 time målinger per husstand om året.

Smart Meter & Data

Smart Meter er digitale forbrugsmålere – el, fjernvarme, vand, etc. - som måler husstandens forbrug hver 15. minut. Det giver meget lange og potentiel højfrekvente tidsrækker, figur 1 viser en smart måler og en smart meter tidsrække med time målinger i 7 måneder for en fjernvarmekunde.



Figur 1 - Smart måler (tv.) 7 måneders realiseret forbrug (th.)

Som udgangspunkt bruges målerne til meget præcis forbrugsaflæsning, men måler dataene gemmes og rummer potentiale for optimering af energisystemet, men også for misbrug af personlige data. For at samle inspiration og kortlægge tidligere smart meter

analyse projekter er der gennemført et systematisk review af litteraturen omkring smart meter analyse for elmålere.

Metode & Analyse

Udgangspunktet for det systematiske review af smart meter analyse er Okolis 8 punkt strategi for systematisk litteratur søgning i Information Systems Research. Okolis metode er udviklet til at sikre systematik og reproducerbarhed i litteratur søgninger. Okolis 8 punkter er som følger:

1: Formål: Præcis beskrivelse litteratur søgningens formål.

2: Protokol: nedfældet beskrivelse af hvordan artikler skal vurderes. Dette er særligt vigtigt hvis der er flere personer om litteratursøgningen og vurderingen.

3: Litteratur søgning: Beskrivelse af litterature søgningen, hvor og hvordan.

4: Praktisk screening: eller screening for inklusion, dette er en grov screening af artikler. Artikler vurderes ud fra titel og abstract, i tvivls tilfælde medtages artiklen. Denne proces er for at gøre litteratursøgningen mulig.

5: Screening af Kvalitet: Screening for inklusion, her læses artiklerne og det vurderes om artiklerne bidragers til litteratur søgningens formål. Artikler som bidrager inkluderes i den endelige base af artikler.

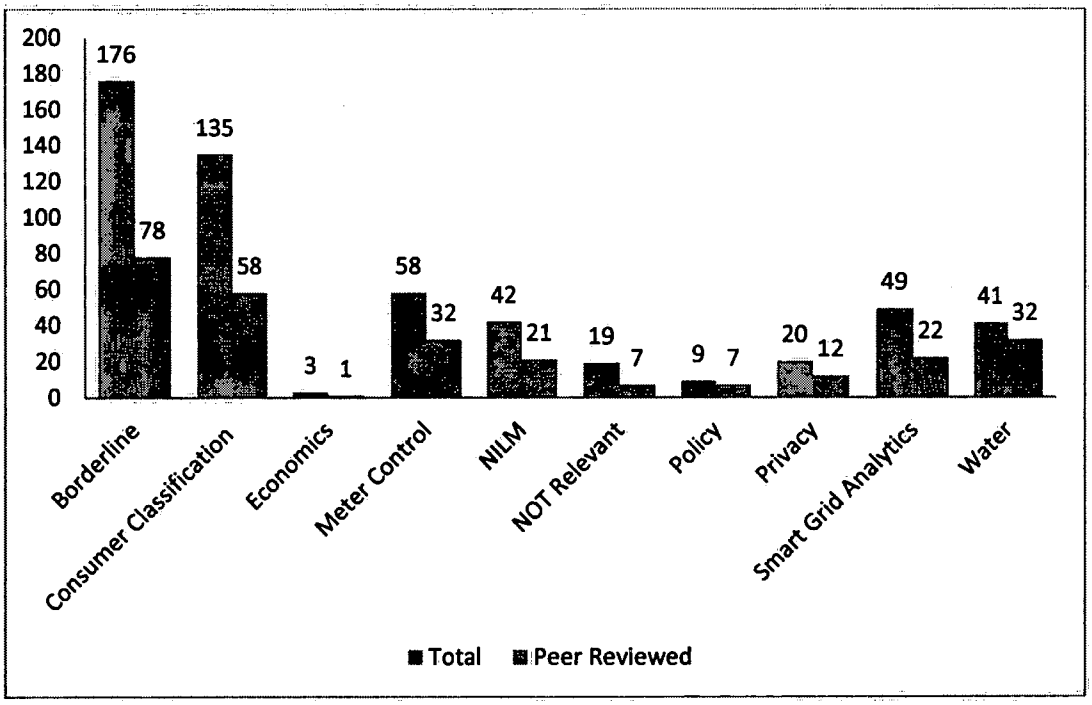
6: Data udtræk: Efter vurdering af de enkelte artikler udvalgt i 5, udtrækkes den relevante information fra artiklerne.

7: Analyse: af data udtrukket i 6.

8: Skriveproces: opsummer og skriv resultaterne af litteratursøgningen.

Formålet med denne litteratursøgning er at danne en basis for hvilke data og analyser som er blevet anvendt til at klassificere elforbrugere med. Thomson Reuters Web of Science (WoS) søgemaskine blev anvendt til at søge på tværs af publishers. WoS giver mulighed for at søge på forskellige parametre, her blev titel og topic anvendt og 27 forskellige søge sætninger blev anvendt med 2099 resulterende artikler.

Screening for titel og abstract reducere antallet af artikler til 552 som er blevet klassificeret i 10 kategorier som kan ses i figur 2. Resten af studiet koncentrere sig om borderline og Consumer Classification.



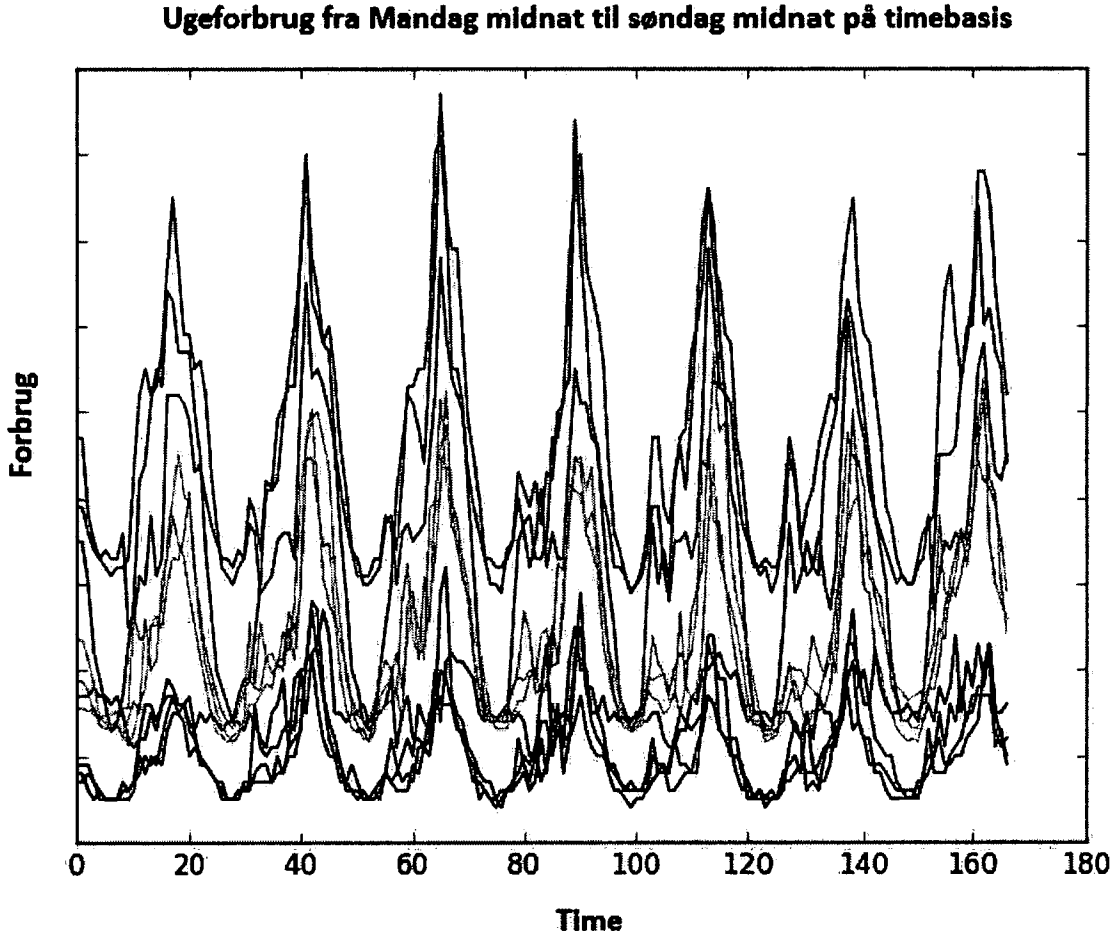
Figur 2 - Overordnede kategorier af smart måler analyser. Borderline angiver artikler med usikkerhed om måler data benyttes. NILM er non-intrusive load monitoring. Not relevant er f.eks. insulin målere.

De 311 artikler i kategorierne Borderline og Consumer Classification bliver reduceret yderligere ved kun at inkludere peer-reviewed artikler. Dertil bliver Borderline revisited da der har været tvivl om anvendelsen af smart måler data. Den endelige inklusion af artikler til kvalitets vurdering er 71. I gennemgangen af de 71 artikler er der fundet 34 relevante artikler som bruger smart måler data til at klassificere elforbrug.

Data udtrækket fra de 34 artikler indeholder bl.a. en oversigt over de enkelte metoder benyttet til klassificering, dimensionering og validering. En vurdering af de enkelte papers data beskrivelse på baggrund af en til artiklen udviklet scoringsmetode. Samt dybere indsigt i datasæt størrelser og herkomst.

I artiklerne er en af de hyppigst brugte klassificeringsmetoder til at klassificere elforbrug K-means. K-means er hurtig og let tilgængelig den er implementeret i de gængse proprietære og open source software pakker. Figur 3 visser klassificering af 10 målere på time basis over en uge. Der er brugt k-means klassificering i python til at opdele i 3 klasser. K-means kan levere gode resultater.

Dagsvariationerne ses tydeligt på grafen samtidig med at der er en tydelig adskillelse mellem grupperne



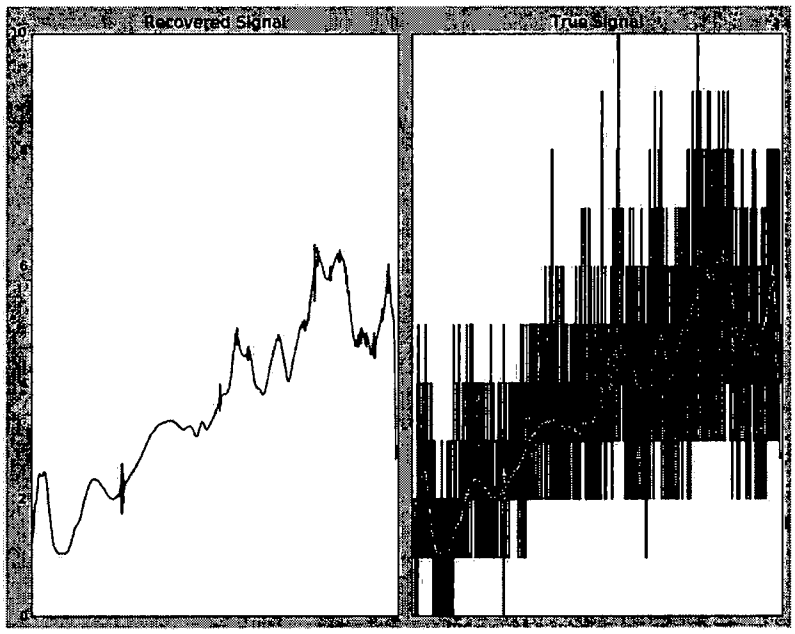
Figur 3 - 10 målere klassificeret i 3 klasser med k-means. (1 uge med time værdier)

Smart Meter data er tidsserier men K-means gør ingen brug af autokorrelation eller sæson information. Det er generelt for de i artiklerne observerede teknikker at der ikke udnyttes tidrække metoder.

Diskussion & konklusion

Et forslag til en metode der ikke er observeret i de undersøgte artikler er Wavelet transformation. Wavelet transformation vil kunne beskrive tidrækken via wavelet koefficienter og derefter klassificere forbruget på baggrund af koefficienterne i

reducerede rum. Figur 4 viser en wavelet transformation til den underliggende funktion på et 7 måneders forbrug.



Figur 4 - Wavelet rekonstrueret signal (tv.) Overlay på oprindeligt signal.

Generelt kan siges om de 34 undersøgte artikler, der er en høj prævalens af metoder fra statistikken i form af moderne klassificering. Men ikke mange forsøg som stikker dybere end at bruge de mest gængse metoder som K-means og hierarkisk klustering. Teknikker som virker godt, men som ikke udnytter viden om det aktuelle data struktur.

Referencer:

Okoli, C., Schabram, K. (2010). "A Guide to Conducting a Systematic Literature Review of Information Systems Research". Sprouts: Working Papers on Information Systems, 10(26). http://sprouts.aisnet.org/10-26

Bekendtgørelse om fjernaflæste elmålere og måling af elektricitet i slutforbruget (2013), https://www.retsinformation.dk/Forms/R0710.aspx?id=160434

Konfidensinterval for prisindeks

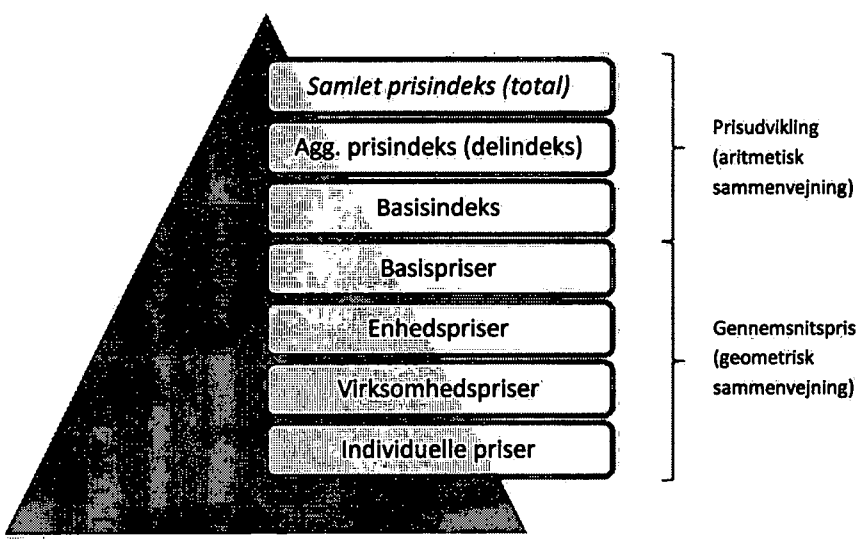
Jakob Holmgaard, Priser og Forbrug, Danmarks Statistik

Resumé:
Papiret skitserer en metode til at beregne konfidensintervaller for traditionelle prisindeks med matchende produkter[1]. Metoden illustreres ved en teoretisk fremstilling og et eksempel med Producentprisindeks for tjenester. Der er tale om en standardmetode, som for eksempel benyttes af det amerikanske Bureau of Labour Statistics.

Indledning:
Usikkerheden på et prisindeks kan opdeles i to dele: (1) Stikprøveusikkerheden, som afspejler, at vi kun har en stikprøve, og (2) den ikke-stikprøverelaterede usikkerhed – for eksempel muligheden for bias pga. manglende besvarelser, ikke-repræsentative produkter eller fejlregistrerede priser. Notatet vil alene behandle stikprøveusikkerheden.

Nedenfor er vist opbygningen af producentprisindeks for tjenester. Man kan lave samme figur og vurdere usikkerheden med samme metode ved Producentprisindekset for byggeri, Producentprisindekset for varer og Forbrugerprisindekset, der alle beregnes efter samme skabelon som Producentprisindekset for tjenester.

Figur 1: Opbygning af Producentprisindeks for tjenester



Elementerne i pyramiden for tjenesternes Producentprisindeks omfatter:

[1] I de tilfælde hvor man kan opstille prisindekset som en hedonisk regression kan man beregne usikkerheden på kvalitetskorrektionen. Dette emne behandles ikke i dette notat, men der henvises til litteraturlisten.

Individuelle priser
Virksomhederne bedes om at udvælge de produkter (typisk 2-3 stk.) som har størst omsætning og som kan følges over tid. Det er vigtigt at sikre en unik beskrivelse af de udvalgte tjenester, så priserne kan opdateres i de efterfølgende perioder.

Virksomhedspriser
De udvalgte virksomheder indberetter én eller flere priser. En virksomhedspris beregnes som det uvejede geometriske gennemsnit af en virksomheds individuelle priser inden for samme enhed. Niveauet med virksomhedspriser er medtaget for at kunne give de indberettende virksomheder forskellige vægte. I praksis, anvender man dog sjældent virksomhedsvægte i producentprisindeks, fordi man mangler oplysninger om omsætningsvægte på dette niveau. I fx Forbrugerprisindekset har man ofte omsætningsoplysninger på forretningsniveauet, som kan sammenlignes med producentprisindeksets virksomhedsniveau.

Enhedspriser
Enhederne består af nogenlunde homogene produkter, som ofte passer med CPA-koden (CPA- Classification of Products by Activity). CPA-koderne er en statistisk klassifikation af produkter med tilhørende aktiviteter. Selvom der er tale om forholdsvis homogene produkter, kan priserne let variere inden for samme enhed. For eksempel er der forskel på en kilometerpris på transportlængden og en timepris på transporttiden, men begge er priser på transport. En enhedspris beregnes som det geometriske gennemsnit af virksomhedspriserne indenfor samme enhed.

Basispriser
En basispris beregnes som det vægtede geometriske gennemsnit af enhedspriserne inden for det samme basisaggregat, og indekseres efterfølgende som basisindeks.

Basisindeks
Basisindeksene er de mest detaljerede prisindeks, og beregnes som nævnt vha. de tilhørende basispriser.

Aggregerede prisindeks (delindeks)
Over basisindeksniveau samles tjenesterne i et hierakisk system med stadig mere aggregerede grupper. De aggregerede delindeks beregnes som et aritmetisk Laspeyres prisindeks[2].

Samlet SPPI
Det samlede Producentprisindeks for tjenester beregnes som et aritmetisk Laspeyres prisindeks, der sammenvejer de aggregerede delindeks[3].

---

[2] I praksis er der tale om et Laspeyres-type prisindeks, da vægtene stammer fra en basisperiode, der ligger før periode 0. Dette emne behandles ikke i dette notat.

[3] Da Producentprisindekset for tjenester endnu ikke indeholder alle tjeneste-brancher, bl.a. mangler de finansielle tjenester, må man betragte det samlede SPPI som ufuldstændigt.

Eksempel på beregning af basisindeks[4]

Det konstruerede eksempel viser beregningen af et basisindeks med virksomheds- og enhedsvægte. Indekset består af to enheder, X og Y, som vejer henholdsvis 0,4 og 0,6, og der indsamles priser fra fire virksomheder A, B, C og D. Virksomhederne indberetter flere priser inden for samme enhed, hvilket er angivet med et vare-løbenummer (vare_lb_nr). Prisrelativet angiver forholdet mellem prisen i periode 1 og periode 0.

Enhed X (vægt 0,4)

Virksomhed	Virksomhedsvægt	Vare_lb_nr	Periode 0	Periode 1	Prisrelativ
Virksomhed A	0,5	A1	6	5	0,8333
		A2	5	6	1,2000
		A3	5	6	1,2000
		A4	4	6	1,5000
Virksomhedspris			4,949	5,733	1,1583
Virksomhed B	0,2	B1	6	5	0,8333
		B2	5	5	1,0000
Virksomhedspris			5,477	5,000	0,9129
Virksomhed C	0,3	C1	6	6	1,0000
		C2	6	5	0,8333
		C3	7	5	0,7143
Virksomhedspris			6,316	5,313	0,8412

Prisrelativet for enhedsprisen på X sammenvejer prisrelativerne på XA, XB, XC:

$$EP^{x}_{0:1} = 1{,}1583^{0{,}5} \cdot 0{,}9129^{0{,}2} \cdot 0{,}8412^{0{,}3} = 1{,}0034 \quad (1)$$

Enhed Y (vægt 0,6)

Virksomhed	Virksomhedsvægt	Vare_lb_nr	Periode 0	Periode 1	Prisrelativ
Virksomhed A	0,6	A1	5	4	0,8000
		A2	4	5	1,2500
		A3	4	5	1,2500
		A4	5	5	1,0000
Virksomhedspris			4,949	5,733	1,0574
Virksomhed B	0,2	B1	4	5	1,2500
		B2	5	6	1,2000
Virksomhedspris			5,477	5,000	1,2247
Virksomhed D	0,2	D1	4	5	1,2500
		D2	5	5	1,0000
		D3	6	5	0,8333
Virksomhedspris			6,316	5,313	1,0137

Enhedsprisrelativ for Y sammenvejer prisrelativerne på YA, YB, YD:

$$EP^{y}_{0:1} = 1{,}0574^{0{,}6} \cdot 1{,}2247^{0{,}2} \cdot 1{,}0137^{0{,}2} = 1{,}0798 \quad (2)$$

[4] Eksemplet er inspireret af eksemplet fra 'Forbruger- og nettoprisindekset – Dokumentation' side 43-44.

Herefter kan basisindekset beregnes som det geometriske gennemsnit af de to enhedsprisrelativer for X og Y ganget med forrige periodes basisindeks, der sættes til 100:

$$\text{Basisindeks}_{0:1} = 1{,}0034^{0{,}4} \cdot 1{,}0798^{0{,}6} \cdot 100 = 104{,}85 \qquad (3)$$

I Bilag A vises en alternativ beregning af basisindekset.

Konfidensinterval for basisindeks:
Det netop gennemgåede regneeksempel på beregningen af et basisindeks kan udvides med et konfidensinterval for basisindekset i ligning (3). I forhold til dette eksempel kan vægtene og stikprøvernes sammensætning betragtes som usikre størrelser, mens priserne må antages at være sikre. Det er svært at vurdere usikkerheden på vægtstrukturen, så fokus vil ligge på den usikkerhed, der er forbundet med stikprøvesammensætningen. For eksempel kunne man havde valgt en anden virksomhed end A, dermed ville de tilhørende priser givetvis afvige, og man ville få et andet basisindeks end de 104,85 i ligning (3).

Udviklingen i basisindekset kan beskrives ved ligning (4):

$$\Delta I_{0:1} = I_1/I_0 = \prod_{i=1}^{k} \left(EP_{0:1}^{i}\right)^{w_i} \qquad (4)$$

Hvor k angiver antallet af enheder.
Det fremgår af ligning (4), at det er prisrelativerne og ikke prisniveauerne, der bestemmer prisudviklingen for basisindekset. For eksempel kunne priserne A1 for virksomhed A under enhed X udskiftes med 50 og 60 i stedet for 5 og 6 uden basisindekset påvirkes, eftersom prisrelativet er det samme, se tabellen nedenfor:

Enhed X (vægt 0,4)

Virksomhed	Virksomhedsvægt	Vare_lb_nr	Periode 0	Periode 1	Prisrelativ
Virksomhed A	0,5	A1	*60*	*50*	0,8333
		A2	5	6	1,2000
		A3	5	6	1,2000
		A4	4	6	1,5000
Virksomhedspris			4,949	5,733	1,1583
Virksomhed B	0,2	B1	6	5	0,8333
		B2	5	5	1,0000
Virksomhedspris			5,477	5,000	0,9129
Virksomhed C	0,3	C1	6	6	1,0000
		C2	6	5	0,8333
		C3	7	5	0,7143
Virksomhedspris			6,316	5,313	0,8412

Til at måle variansen på prisindekset i periode 1 bruger vi variansen på prisrelativerne og ikke variansen på prisniveauerne. Variansen på prisniveauerne er for høj, hvis de produkter, der repræsenterer samme enhed ikke er homogene, for eksempel kan

priserne vedrøre forskellige mængdeenheder. Pointen ligger i, at de anvendte prisobservationer er matchet, så hvis der er en observation i tælleren på en speciel skala, for eksempel decimeter i stedet for meter, er der også netop én observation i nævneren på samme skala. Dermed forkortes skalaeffekten altid væk, så prisindekset og dets varians ikke påvirkes; og variansen på prisrelativerne kan bruges til at beskrive variansen på prisindeksets stigningen fra periode 0 til 1. I bilag B vises, at eksakt varians og hele udfaldsrummet af observationerne er to sider af samme sag.

Prisrelativerne i ligning (4) opfattes som stokastiske variable og enhedsvægtene som konstanter. Da det er nemmere, at angive variansen på en sum end variansen på et produkt, tages logaritmen på begge sider af (4), så det geometriske gennemsnit bliver til et aritmetisk, hvor sumtegnet sigma erstatter produkttegnet pi.:

$$\Delta \ln I_{0:1} = \sum_{i=1}^{k} w_i \cdot \ln\left(EP_{0:1}^{i}\right) \tag{5}$$

Idet vi antager, at usikkerheden på et prisrelativ er ukorreleret med usikkerheden på de andre prisrelativer, er variansen på den logaritmiske indeksændring i ligning (5) givet ved:

$$Var(\Delta \ln I_{0:1}) = \sum_{i=1}^{k} \left((w_i)^2 \cdot var\left(\ln\left(EP_{0:1}^{i}\right)\right)\right) \tag{6}$$

Det tilhørende konfidensinterval for basisindekset i ligning (4) kan beregnes ved hjælp af følgende formel:

$$KI_{0:1}^{BI} = \exp\left(\Delta \ln I_{0:1} \pm 1{,}96 \cdot \sqrt{\sum_{i=1}^{k} \left((w_i)^2 \cdot var\left(\ln\left(EP_{0:1}^{i}\right)\right)\right)}\right) \tag{7}$$

hvor,

- $KI_{0:1}^{BI}$: Konfidensinterval for basisindeks fra periode 0 til 1
- exp: Eksponentialfunktionen
- $\Delta \ln I_{0:1}$: Logaritmen til basisindekset fra periode 0 til 1
- w_i: Vægten til enhed i
- $var\left(ln\left(EP_{0:1}^{i}\right)\right)$: Variansen på logaritmen til enhedsprisrelativet mellem periode 0 og 1

Ved beregning af variansen på logaritmen til enhedsprisrelativet i ligning (7) springes niveauet med virksomhedspriser over, jf. figuren for data-pyramiden. Dette skyldes, at der i praksis er få virksomheder, der indberetter mere end én pris pr. enhed, så datagrundlaget til at beregne variansen for virksomhedsprisrelativerne er meget tyndt. Variansberegningen med virksomhedspriser er dog vist i Bilag C. Nedenfor vises variansberegningen uden virksomhedspriser. Til vurdering af variansen på enhed X's prisudvikling har vi følgende log-prisrelativer:

Enhed X (vægt 0,4): log-prisrelativ

Virksomhed	Virksomhedsvægt	Vare_lb_nr	Prisrelativ	Log-prisrelativ
Virksomhed A	0,5	A1	0,8333	-0,1823
		A2	1,2000	0,1823
		A3	1,2000	0,1823
		A4	1,5000	0,4055
Virksomhed B	0,2	B1	0,8333	-0,1823
		B2	1,0000	0,0000
Virksomhed C	0,3	C1	1,0000	0,0000
		C2	0,8333	-0,1823
		C3	0,7143	-0,3365

Og til vurdering af variansen på enhed Y's prisudvikling har vi følgende log-prisrelativer:

Enhed Y (vægt 0,6): log-prisrelativ

Virksomhed	Virksomhedsvægt	Vare_lb_nr	Prisrelativ	Log-prisrelativ
Virksomhed A	0,6	A1	0,8000	-0,1823
		A2	1,2500	0,1823
		A3	1,2500	0,1823
		A4	1,0000	0,4055
Virksomhed B	0,2	B1	1,2500	-0,1823
		B2	1,2000	0,0000
Virksomhed D	0,2	D1	1,2500	0,0000
		D2	1,0000	-0,1823
		D3	0,8333	-0,3365

Variansligningen ser således ud[5]:

$$\text{Varians} = \frac{\Sigma(x-\bar{x})^2}{n-1} \qquad (8)$$

Bemærk, at der divideres med n-1 og ikke n, da der korrigeres for frihedsgrader ved beregning af variansen på en stikprøve. For store samples gør det ikke den store forskel, men for små samples som i vores tilfælde har det betydning, om man dividerer med n eller n-1. Ved beregning af variansen på en population divideres med n, hvilket kan bruges til at illustrere sammenhængen mellem den eksakte varians og variansen beregnet med bootstrap-teknikken, se bilag B.

Ved anvendelse af ligning (8) kan variansen på log-prisrelativ-observationerne i hver af de to enheder beregnes til:

	VARLPX	VARLPY
Periode 0:1	0,0553	0,0332

[5] Se (Tauqueer Ahmad)

Da hver prisobservation inden for en enhed har vægten 1/n samt samme varians, kan den gennemsnitlige varians for log-prisrelativerne til de to enheder beregnes således:

$$\text{VARGLPX} = \frac{1}{n^2} \cdot n \cdot \text{VARLPX} = \frac{\text{VARLPX}}{n} = \frac{0{,}0553}{9} = 0{,}0061 \quad (9)$$

$$\text{VARGLPY} = \frac{1}{n^2} \cdot n \cdot \text{VARLPY} = \frac{\text{VARLPY}}{n} = \frac{0{,}0332}{9} = 0{,}0037 \quad (10)$$

Begge enheder består af 9 prisobservationer, dvs. n = 9.

I ligning (9) og (10) ovenfor er anvendt regnereglen om, at variansen af en konstant gange en stokastisk variabel er konstanten i anden ganget med variansen til den stokastiske variabel:

$$\text{Var}(a \cdot X) = a^2 \cdot \text{var}(X) \quad (11)$$

Til at beregne den samlede varians anvendes regnereglen i ligning (11) samt regnereglen om, at variansen på summen af to uafhængige stokastiske variable er summen af varianserne:

$$\text{Var}(A + B) = \text{Var}(A) + \text{Var}(B) \quad (12)$$

Den samlede varians beregnes som følger:

$$\sum_{i=1}^{k}\left((w_i)^2 \cdot \text{var}\left(\ln\left(EP_{0:1}^{i}\right)\right)\right) = 0{,}4^2 \cdot 0{,}0061 + 0{,}6^2 \cdot 0{,}0037 = 0{,}0023 \quad (13)$$

Den nedre og øvre grænse for basisindekset i ligning (3) kan udtrykkes ved:

$$\text{Basisindeks}_{0:1}^{\text{nedre}} = \exp\left(\ln(104{,}85) - 1{,}96 \cdot \sqrt{0{,}0023}\right) = 95{,}42 \quad (14)$$

$$\text{Basisindeks}_{0:1}^{\text{øvre}} = \exp\left(\ln(104{,}85) + 1{,}96 \cdot \sqrt{0{,}0023}\right) = 115{,}21 \quad (15)$$

Konfidensinterval for delindeks:

De aggregerede prisindeks (delindeks) beregnes som et aritmetisk Laspeyres prisindeks:

$$P_{0:t}^{LA} = \frac{\sum_{i=1}^{N} p_t^i \cdot q_0^i}{\sum_{i=1}^{N} p_0^i \cdot q_0^i} \quad (16)$$

I praksis omskrives formlen til et vægtet gennemsnit af prisrelativerne, der angiver forholdet mellem det aktuelle basisindeks og basisindekset i referenceperioden. Det resulterende Laspeyres-prisindeks kan herefter udtrykkes som et vægtet gennemsnit af prisrelativerne ved følgende omskrivning:

$$P_{0:t}^{LA} = \frac{\sum_{i=1}^{N} p_t^i \cdot q_0^i}{\sum_{i=1}^{N} p_0^i \cdot q_0^i} = \sum_{i=1}^{N}\left(p_t^i \cdot \frac{q_0^i}{\sum_{i=1}^{N} p_0^i \cdot q_0^i}\right) = \sum_{i=1}^{N}\left(\frac{p_t^i}{p_0^i} \cdot \frac{p_0^i \cdot q_0^i}{\sum_{i=1}^{N} p_0^i \cdot q_0^i}\right)$$

$$= \sum_{i=1}^{N} w_0^i \cdot \left(\frac{p_t^i}{p_0^i}\right) = \sum_{i=1}^{N} w_0^i \cdot \left(\frac{BI_t^i}{BI_0^i}\right) \quad (17)$$

hvor,

$$w_0^i = \frac{p_0^i \cdot q_0^i}{\sum_{i=1}^N p_0^i \cdot q_0^i} \quad \text{og} \quad \sum_{i=1}^N w_o^i = 1 \tag{18}$$

Ved at anvende regnereglen i ligning (11) kan variansen til det samlede prisindeks $P_{0:t}^{LA}$ af den eksakte formel (21) opskrives således:

$$\text{var}(P_{0:t}^{LA}) = \sum_{i=1}^N \left((w_o^i)^2 \cdot \left(\text{var}(\frac{BI_t^i}{BI_0^i}) \right) \right) \tag{19}$$

I formlen for variansen indføres en første ordens Taylor approksimation til logaritmefunktionen med udgangspunkt i 1, hvor funktionen er nul og hældningen 1:

$$\ln\left(\frac{BI_t^i}{BI_0^i}\right) \approx \ln(1) + \ln`(1) \cdot \left(\frac{BI_t^i}{BI_0^i} - 1\right) = 0 + 1 \cdot \left(\frac{BI_t^i}{BI_0^i} - 1\right)$$

$$<=> \frac{BI_t^i}{BI_0^i} \approx \ln\left(\frac{BI_t^i}{BI_0^i}\right) + 1 \tag{20}$$

Approksimationen er god ved små ændringer i basisindeks.

Dermed fås:

$$\text{var}(P_{0:t}^{LA}) \approx \sum_{i=1}^N \left((w_o^i)^2 \cdot \left(\text{var}(\ln\left(\frac{BI_t^i}{BI_0^i}\right) + 1) \right) \right) <=>$$

$$\text{var}(P_{0:t}^{LA}) \approx \sum_{i=1}^N \left((w_o^i)^2 \cdot \left(\text{var}(\ln(BI_t^i) - \ln(BI_0^i) + 1) \right) \right) \tag{21}$$

Variansen til A minus B plus en konstant, hvor A og B er uafhængige stokastiske variable, er summen af varianserne til A og B. Variansen på det samlede prisindeks giver derfor:

$$\text{var}(P_{0:t}^{LA}) \approx \sum_{i=1}^N \left((w_o^i)^2 \cdot \left(\text{var}(\ln BI_t^i) + \text{var}(\ln BI_0^i) \right) \right) \tag{22}$$

Og konfidensintervallet for delindeks kan skrives som:

$$KI_{0:t}^{DI} = \Delta I_{0:t} \pm 1{,}96 \cdot \sqrt{\sum_{i=1}^N \left((w_o^i)^2 \cdot \left(\text{var}(\ln BI_t^i) + \text{var}(\ln BI_0^i) \right) \right)} \tag{23}$$

hvor,

- $KI_{0:t}^{DI}$: Konfidensinterval for delindeks fra periode 0 til t
- $\Delta I_{0:t}$: Delindekset fra periode 0 til t
- w_o^i: Vægten for delindeks i
- $Var(lnBI_i)$: Variansen på logaritmen til basisindeks i

Afprøvning af metoden på empirisk data:

I den følgende gennemgang tages der udgangspunkt i producentprisindekset for tjenester (Service Producer Price Index), som dækker en lang række erhvervstjenester. Som eksempel udvælges vejgodstransport med branchekoden 49410000 i DB07 (Dansk Branchekode 2007), som er opdelt i 7 enheder:

[illegible]	Vejgodstransport i kølevogn
[illegible]	Godstransp. ad vej i tank- eller sættevogne (olieprodukter)
[illegible]	Godstransp. ad vej i tank- eller sættevogne (andet flyd.)
[illegible]	Transport ad vej af containere til kombineret transport
[illegible]	Transport ad vej af tørt styrtgods
[illegible]	Anden vejgodstransport
[illegible]	Udlejning af lastbiler med fører

I eksemplet anvender hver enhed CPA-koder (Classification of Products by Activity). Der anvendes vægten 1/7 til hver enhed. Samlet set for alle 7 enheder indberettes der ca. 70 priser pr. kvartal.

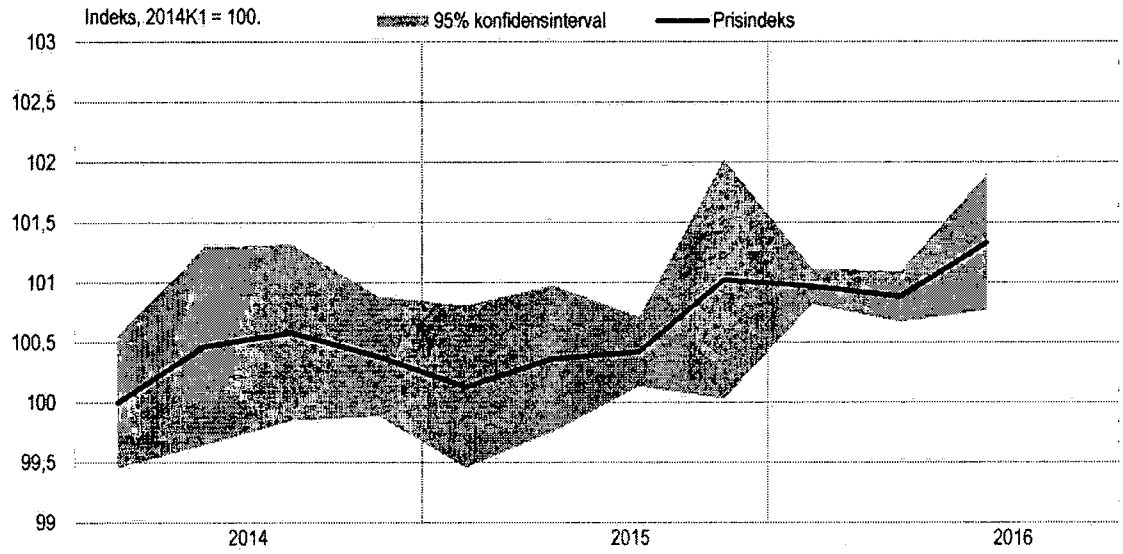
I stikprøven indgår en række virksomheder, som bliver bedt om at udvælge nogle repræsentative produkter (typisk 1-3) med en stor omsætning inden for de 7 enheder. Derudover skal produkternes priser kunne følges over tid, så vi får produktmatch. Desuden bliver virksomhederne hvert kvartal bedt om at opdatere prisen for det seneste kvartal. Nedenfor er vist nogle eksempler på priser for enhed nummer 49411100 – vejgodstransport i kølevogn:

Enhed 49411100 (vægt 1/7)

Virksomhed	Vare_lb_nr	Periode 0	Periode 1	Prisrelativ
Virksomhed A	A1	43,04	43,75	1,0165
Virksomhed B	B1	367	367	1,0000
Virksomhed C	C1	9,25	9,25	1,0000
	C2	9,25	9,25	1,0000
Virksomhed D	D1	6,9	6,9	1,0000
Virksomhed E	E1	183,34	183,34	1,0000
	E2	2.035	2.035	1,0000

I eksemplet er der 5 virksomheder, der tilsammen indberetter 7 priser på enheden. Kun to af virksomhederne indberetter mere end én pris. Ved beregningen af variansen ses der derfor bort fra virksomhedspriser. Variansen beregnes for de 7 prisrelativer inden for enheden uden hensyntagen til hvilke virksomheder der har indberettet prisen. Priserne spænder vidt fra 6,9 til 2.035, så der er oplagt ikke tale om flere bud på det samme produkt. Det skyldes, at priserne vedrører forskellige mængder, for eksempel prisen pr. kilometer eller prisen pr. times kørsel. De fleste priser, der indberettes, er uændret i forhold til forrige periode. Det kan skyldes, at der er tale om listepriser, som kun ændres én gang årligt, eller at virksomheden har indberettet for sent, så prisen er videreført som uændret fra forrige periode. Konfidensintervallet for basisindekset for vejgodstransport er vist i figuren nedenfor:

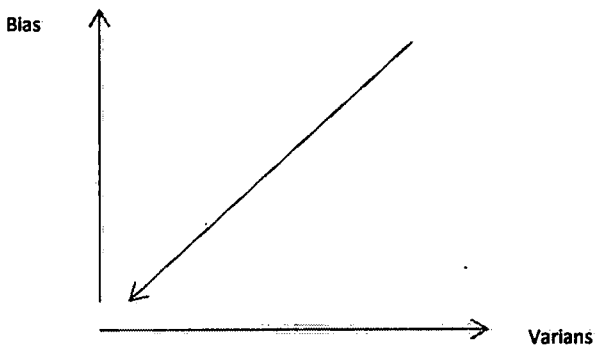
Figur: Konfidensinterval for prisindeks for vejgodstransport:



De mange uændrede priser reducerer den målte varians. Ideelt set skal prisindekset afspejle de faktiske priser og ikke listepriser. Hvis virksomheden via rabatter afviger fra listepriserne for at kunne differentiere priserne for forskellige kundesegmenter, for eksempel store kunder versus små kunder, bør rabatten indregnes. Listepriser fungerer ofte som et forhandlingsoplæg, som kunden kan få rabat i forhold til. Producentprisindekset skal helst afspejle de faktiske priser efter rabat, for eksempel bruges producentprisindekset til at lave deflatorer til nationalregnskabets fastprisberegning.

Nedenstående figur illustrerer den normale stigende sammenhæng mellem varians og bias. Men selvom variansen (konfidensintervallet) er lille, kan bias være stor. For eksempel hvis de udvalgte virksomheder ikke er repræsentative for hele populationen af virksomheder i den pågældende branche.

Figur: Varians og bias



Et ideelt prisindeks vil ligge i punktet (0,0), dvs. ingen bias og ingen varians. Det er vigtigt at undgå bias, så man kan ikke vurdere usikkerheden på et prisindeks alene ud fra konfidensintervallet. Konfidensintervallet bør suppleres af en vurdering af prisindeksets bias.

Andre datakilder:
I nogle tilfælde består de indsamlede priser af gennemsnitspriser, som den indberettende virksomhed selv har beregnet. I disse tilfælde har man information til at beregne prisindekset men ikke til at beregne variansen, så man kan ikke beregne et konfidensinterval for disse priser.

Bilag A: Alternativ beregning af basisindeks
I dette bilag præsenteres en alternativ metode til at beregne basisindekset på, hvor prisrelativet ikke bruges direkte i beregningen.

Enhed X (vægt 0,4)

Virksomhed	Virksomhedsvægt	Vare_lb_nr	Periode 1	Periode 2
Virksomhed A	0,5	A1	6	5
		A2	5	6
		A3	5	6
		A4	4	6
Virksomhedspris			4,949	5,733
Virksomhed B	0,2	B1	6	5
		B2	5	5
Virksomhedspris			5,477	5,000
Virksomhed C	0,3	C1	6	6
		C2	6	5
		C3	7	5
Virksomhedspris			6,316	5,313

Enhedsprisen for X sammenvejer virksomhedspriserne på XA, XB, XC:

$$EP_1^x = 4{,}949^{0{,}5} \cdot 5{,}477^{0{,}2} \cdot 6{,}316^{0{,}3} = 5{,}434 \qquad (A.1)$$

$$EP_2^x = 5{,}733^{0{,}5} \cdot 5{,}000^{0{,}2} \cdot 5{,}313^{0{,}3} = 5{,}452 \qquad (A.2)$$

Enhed Y (vægt 0,6)

Virksomhed	Virksomhedsvægt	Vare_lb_nr	Periode 1	Periode 2
Virksomhed A	0,6	A1	5	4
		A2	4	5
		A3	4	5
		A4	5	5
Virksomhedspris			4,472	4,729
Virksomhed B	0,2	B1	4	5
		B2	5	6
Virksomhedspris			4,472	5,477
Virksomhed D	0,2	D1	4	5
		D2	5	5
		D3	6	5
Virksomhedspris			4,932	5,000

Enhedsprisen for Y sammenvejer virksomhedspriserne på YA, YB, YD:

$$EP_1^y = 4{,}472^{0{,}6} \cdot 4{,}472^{0{,}2} \cdot 4{,}932^{0{,}2} = 4{,}561 \quad \text{(A.3)}$$

$$EP_2^y = 4{,}729^{0{,}6} \cdot 5{,}477^{0{,}2} \cdot 5{,}000^{0{,}2} = 4{,}924 \quad \text{(A.4)}$$

Herefter beregnes basispriserne som det vægtede geometriske gennemsnit af de to enhedspriser for X og Y for en given periode.

Basispriserne for hver af de to perioder beregnes:

$$BP_1 = 5{,}434^{0{,}4} \cdot 4{,}561^{0{,}6} = 4{,}892 \quad \text{(A.5)}$$

$$BP_2 = 5{,}452^{0{,}4} \cdot 4{,}924^{0{,}6} = 5{,}129 \quad \text{(A.6)}$$

Eller:

$$BP_1 = (4{,}949^{0{,}5} \cdot 5{,}477^{0{,}2} \cdot 6{,}316^{0{,}3})^{0{,}4} \cdot (4{,}472^{0{,}6} \cdot 4{,}472^{0{,}2} 4{,}932^{0{,}2})^{0{,}6} = 4{,}892 \quad \text{(A.7)}$$

$$BP_2 = (5{,}733^{0{,}5} \cdot 5{,}000^{0{,}2} \cdot 5{,}313^{0{,}3})^{0{,}4} \cdot (4{,}729^{0{,}6} \cdot 5{,}477^{0{,}2} 5{,}000^{0{,}2})^{0{,}6} = 5{,}129 \quad \text{(A.8)}$$

Basisindekset er som beskrevet det laveste niveau, som der beregnes indeks på, og beregnes som udviklingen i basispriserne ganget med basisindekset i forrige periode (her sat til 100), dvs.

$$\text{Basisindeks}_2 = \frac{5{,}129}{4{,}892} \cdot 100 = 104{,}85 \quad \text{(A.9)}$$

Resultatet er identisk med ligning (3), da der reelt laves samme beregning på to måder.

Bilag B: Eksakt variansberegning vs. variansberegning med bootstrap-teknik

I dette bilag vises to måder, at beregne en eksakt varians på med udgangspunkt i følgende hypotetiske datasæt:

Tabel: Udgangssample

Virksomhed	Virksomhedsvægt	Vare_lb_nr	Pris
Virksomhed A	0,4	A1	4
		A2	3
Virksomhed B	0,6	B1	6
		B2	5

Metode 1:
Hvis der tages udgangspunkt i variansen på hele udfaldsrummet (populationen) kan variansen defineres som[6]:

$$\text{Varians}_{\text{population}} = \frac{\Sigma(x-\bar{x})^2}{n} \quad \text{(B.1)}$$

Ved anvendelse af ligning (B.1) giver variansen på pris-observationerne indenfor hver virksomhed følgende:

$$\text{Var(A)} = \frac{(4-3{,}5)^2+(3-3{,}5)^2}{2} = 0{,}25 \quad \text{(B.2)}$$

$$\text{Var(B)} = \frac{(6-5{,}5)^2+(5-5{,}5)^2}{2} = 0{,}25 \quad \text{(B.3)}$$

Hernæst beregnes den gennemsnitlige varians indenfor hver virksomhed:

$$\text{Var}(\overline{\text{A}}) = \frac{\text{Var(A)}}{n} = \frac{0{,}25}{2} = 0{,}125 \quad \text{(B.4)}$$

$$\text{Var}(\overline{\text{B}}) = \frac{\text{Var(B)}}{n} = \frac{0{,}25}{2} = 0{,}125 \quad \text{(B.5)}$$

For at beregne den vægtede gennemsnitlige varians, skal vi først bruge regnereglen der siger, at variansen på to uafhængige stokastiske variable er variansen på den ene stokastiske variabel plus variansen på den anden stokastiske variabel, hvor det antages, at kovariansen er nul:

$$\text{Var}(A + B) = \text{Var}(A) + \text{Var}(B) \quad \text{(B.6)}$$

Derudover bruges regnereglen om, at variansen til en konstant gange en stokastisk variabel er konstanten i anden ganget med variansen til den stokastiske variabel:

$$\text{Var}(a \cdot X) = a^2 \cdot \text{var}(X) \quad \text{(B.7)}$$

Hermed fås den vægtede gennemsnitlige varians for hele udgangssamplet:

$$\text{Var}(\overline{\text{A}} + \overline{\text{B}}) = W_A^2 \cdot \text{Var}(\overline{\text{A}}) + W_B^2 \cdot \text{Var}(\overline{\text{B}})$$
$$= 0{,}4^2 \cdot 0{,}125 + 0{,}6^2 \cdot 0{,}125 = 0{,}065 \quad \text{(B.8)}$$

Metode 2:
En anden metode til at komme frem til det samme resultat - ligning (B.8), kan fås ved at kombinere priserne i udgangssamplet (bootstrap) i alle tænkelige kombinationer. De to priser fra virksomhed A kan kombineres på $2^2 = 4$ måder med tilbagelægning. Ligeledes kan de to priser fra virksomhed B kombineres på $2^2 = 4$ måder. Dermed kan de fire priser i udgangssamplet kombineres på $4^2 = 16$ måder. De 16 samples er vist i tabellen nedenfor og udgør hele udfaldsrummet:

[6] For at beregne variansen på en population kan man i SAS benytte optionen 'vardef=weight' under proc means eller funktionen 'varians.p' i Excel.

Tabel: 16 samples

Sample	Forretning	Pris	Sample	Forretning	Pris	Sample	Forretning	Pris	Sample	Forretning	Pris
1	A1	4	5	A1	4	9	A1	4	13	A1	4
1	A2	4	5	A2	4	9	A2	4	13	A2	4
1	B1	6	5	B1	6	9	B1	5	13	B1	5
1	B2	5	5	B2	6	9	B2	6	13	B2	5
2	A1	3	6	A1	3	10	A1	3	14	A1	3
2	A2	3	6	A2	3	10	A2	3	14	A2	3
2	B1	6	6	B1	6	10	B1	5	14	B1	5
2	B2	5	6	B2	6	10	B2	6	14	B2	5
3	A1	4	7	A1	4	11	A1	4	15	A1	4
3	A2	3	7	A2	3	11	A2	3	15	A2	3
3	B1	6	7	B1	6	11	B1	5	15	B1	5
3	B2	5	7	B2	6	11	B2	6	15	B2	5
4	A1	3	8	A1	3	12	A1	3	16	A1	3
4	A2	4	8	A2	4	12	A2	4	16	A2	4
4	B1	6	8	B1	6	12	B1	5	16	B1	5
4	B2	5	8	B2	6	12	B2	6	16	B2	5

Hernæst beregnes den vægtede gennemsnitspris for hver af de 16 samples samt den vægtede gennemsnitspris i udgangssamplet. Differencen mellem de to gennemsnitspriser er vist i kolonnen yderst til højre i tabellen nedenfor:

Sample	Gnspris pr. sample	Gnspris i udgangssample	Difference (absolut)
1	4,9	4,7	0,2
2	4,5	4,7	0,2
3	4,7	4,7	0,0
4	4,7	4,7	0,0
5	5,2	4,7	0,5
6	4,8	4,7	0,1
7	5,0	4,7	0,3
8	5,0	4,7	0,3
9	4,9	4,7	0,2
10	4,5	4,7	0,2
11	4,7	4,7	0,0
12	4,7	4,7	0,0
13	4,6	4,7	0,1
14	4,2	4,7	0,5
15	4,4	4,7	0,3
16	4,4	4,7	0,3

Den vægtede gennemsnitlige varians for de to forretningsgrupper beregnes som den gennemsnitlige kvadrerede difference mellem gennemsnitsprisen pr. sample og gennemsnitsprisen i udgangssamplet (kolonnen yderst til højre), dvs.

$$\text{Var}(\overline{A} + \overline{B}) = \frac{(0{,}2)^2 + (0{,}2)^2 + (0)^2 + \cdots + (0{,}5)^2 + (0{,}3)^2 + (0{,}3)^2}{16} = 0{,}065 \qquad \text{(B.9)}$$

Resultatet er identisk med (B.8), da der reelt laves samme beregning på to måder. Variansberegning med bootstrap-teknikken, som anvender en delmængde af de mulige kombinationer på udgangssamplet, kan således opfattes som et specialtilfælde af den eksakte variansberegning, hvor man anvender alle tænkelige kombinationer på udgangssamplet.

Bilag C: Variansberegning med virksomhedspriser

I dette bilag vises beregningen af konfidensintervallet for det konstruerede eksempel, hvor niveauet med virksomhedspriser medtages i variansberegningen.

Variansen på logaritmen til prisrelativet mellem periode 0 og 1 afspejler for enhed X lavet af virksomhed A spredningen i de logaritmiske værdier af de fire prisrelativer, som i nedenstående tabel repræsenterer prisudviklingen på enhed X lavet af virksomhed A. Data fra virksomhed B og C omfatter henholdsvis to og tre priselativer, der også bruges til at vurdere variansen på prisudviklingen for enhed X.

Enhed X (vægt 0,4): log-prisrelativ

Virksomhed	Virksomhedsvægt	Vare_lb_nr	Prisrelativ	Log-prisrelativ
Virksomhed A	0,5	A1	0,8333	-0,1823
		A2	1,2000	0,1823
		A3	1,2000	0,1823
		A4	1,5000	0,4055
Virksomhed B	0,2	B1	0,8333	-0,1823
		B2	1,0000	0,0000
Virksomhed C	0,3	C1	1,0000	0,0000
		C2	0,8333	-0,1823
		C3	0,7143	-0,3365

Og til vurdering af variansen på enhed Y's prisudvikling har vi følgende log-prisrelativer:

Enhed Y (vægt 0,6): log-prisrelativ

Virksomhed	Virksomhedsvægt	Vare_lb_nr	Prisrelativ	Log-prisrelativ
Virksomhed A	0,6	A1	0,8000	-0,1823
		A2	1,2500	0,1823
		A3	1,2500	0,1823
		A4	1,0000	0,4055
Virksomhed B	0,2	B1	1,2500	-0,1823
		B2	1,2000	0,0000
Virksomhed D	0,2	D1	1,2500	0,0000
		D2	1,0000	-0,1823
		D3	0,8333	-0,3365

Ved anvendelse af ligning (8) kan variansen på logprisrelativ-observationerne i hvert strata beregnes til:

	VARLPXA	VARLPXB	VARLPXC
Periode 0:1	0,0593	0,0166	0,0284

	VARLPYA	VARLPYB	VARLPYD
Periode 0:1	0,0166	0,0249	0,0412

Hvert stratum kan opfattes som en stokastisk variabel, som der er n bud på, hvor n angiver antallet af observationer. Der gælder, at hver observation inden for et strata har samme varians. Det følger heraf, at variansen på den gennemsnitlige log-virksomhedsprisrelativ kan beregnes ved at dividere variansen til logprisrelativet med antallet af observationer inden for hvert stratum bestående af enhed og virksomhed. For eksempel beregnes den gennemsnitlige varians for logvirksomhedsprisrelativet XA således:

$$\text{VARGLPXA} = \frac{1}{n^2} \cdot n \cdot \text{VARLPXA} = \frac{\text{VARLPXA}}{n} = \frac{0{,}0444}{4} = 0{,}0148 \qquad \text{(C.1)}$$

Den gennemsnitlige varians på de gennemsnitlige log-virksomhedsprisrelativer er:

	VARGLPXA	VARGLPXB	VARGLPXC
Periode 0:1	0,0148	0,0083	0,0095

	VARGLPYA	VARGLPYB	VARGLPYD
Periode 0:1	0,0041	0,0124	0,0137

Til at beregne den gennemsnitlige varians på log-enhedsprisrelativerne, anvendes regnereglerne i ligning (11) og (12). Det giver følgende udtryk for variansen på det gennemsnitlige log-prisrelativ for enhed X:

$$\text{VARGLPX} = W_{XA}^2 \cdot \text{VARGLPXA} + W_{XB}^2 \cdot \text{VARGLPXB} + W_{XC}^2 \cdot \text{VARGLPXC}$$

$$= 0{,}5^2 \cdot 0{,}0148 + 0{,}2^2 \cdot 0{,}0083 + 0{,}3^2 \cdot 0{,}0095 = 0{,}0049 \qquad \text{(C.2)}$$

I tabellen nedenfor er vist en oversigt over den gennemsnitlige varians på de to log-enhedsprisrelativer:

	VARGLPX	VARGLPY
Periode 0:1	0,0049	0,0025

Den samlede varians beregnes som følger:

$$\sum_{i=1}^{k}\left((w_i)^2 \cdot \text{var}\left(\ln\left(EP_{0:1}^i\right)\right)\right) = 0{,}4^2 \cdot 0{,}0049 + 0{,}6^2 \cdot 0{,}0025 = 0{,}0017 \qquad \text{(C.3)}$$

Den nedre og øvre grænse for basisindekset i ligning (3) kan udtrykkes ved:

$$\text{Basisindeks}_{0:1}^{\text{nedre}} = \exp\left(\ln(104{,}85) - 1{,}96 \cdot \sqrt{0{,}0017}\right) = 96{,}72 \quad \text{(C.4)}$$

$$\text{Basisindeks}_{0:1}^{\text{øvre}} = \exp\left(\ln(104{,}85) + 1{,}96 \cdot \sqrt{0{,}0017}\right) = 113{,}67 \quad \text{(C.5)}$$

Konfidensintervallet er en anelse mindre end i ligning (14) og (15).

Litteraturliste:

Carsten Boldsen (2004): Forbruger- og nettoprisindekset, dokumentation. Danmarks Statistik.

Jakob Holmgaard (2016): Usikkerhed ved opgørelse af udviklingen i boligpriser. Danmarks Statistik. Publiceret som en DST-analyse: http://www.dst.dk/Site/Dst/Udgivelser/nyt/GetAnalyse.aspx?cid=27505

Jakob Holmgaard (2016): Hedonic House Price Index, Prices and Consumption, Statistics Denmark. Intern arbejdspapir.

Lars Hervig Jacobsen og Jakob Holmgaard (2016): Hedonic based price index, Priser og Forbrug, Danmarks Statistik. Publiceret i Symposium for anvendt statistik 2016.

Tanqueer Ahmad: Jackknife and bootstrap methods of variance estimation.

U.S. Bureau of Labor Statistics (BLS), July 2016: Variance estimates for price changes in the Producer Price Index, 2015.

Wingren, Jan Eric (June 2009): Variance analysis for PPI and SPPI. Final technical implementation report. Statistiska centralbyrån.

Valg af kontrol gruppe

Maria Holm, Metode og Analyse – Danmarks Statistik

Indledning

I forbindelse med mit speciale i foråret 2015 blev jeg bekendt med en årelang diskussion af matching og valg ag kontrolgruppe. Specialet omhandlede et forsøg der skulle undersøge husholdningers forbrug af elektricitet, og deres villighed til at rykke forbrug rundt på et døgn. Dette kaldes for 'Demand side management'. Hvis vi skal gå over til brug af vedvarende energi, er det nødvendigt, at forbruget udvikler sig i en mere fleksibel retning. Vind- og vand energi kan indtil videre ikke så godt opbevares og gemmes, derfor skal det bruges når det er der. Forsøgs husholdningerne i vores forsøg, som var et samarbejde mellem SydEnergi og IFRO på KU, modtog enten dagligt, flere gange om ugen eller en gang ugentligt sms besked om, at de skulle rykke forbrug væk fra eller hen til en periode. De modtog også forskellige beskrivelser af motivet for, at gøre dette. Ekesempelvis for, at opnår en pengebesparelse, CO2-besparelse eller en kombination. OPOWER eksperimentet i USA er et kendt eksempel på netop 'Demand side management' og elektricitetsforbrug (Allcott 2011). Husholdningerne blev inviteret tilfældigt, men kunne afvise invitationen, hvilket medfører en risiko for, at dem der valgte at acceptere invitationen har en enten observerbar eller ikke observerbar bagvedliggende karakteristika, der påvirker deres villighed til, at deltage – selektions pres. I min stikprøve efter, at have renset data har jeg 5410 kontroller og 186 forsøgshusholdninger at vælge fra.

På baggrund af følgende test $\frac{\hat{\theta}_{treat}-\hat{\theta}_{control}}{SE(\hat{\theta}_{treat}-\hat{\theta}_{control})}$ hvor $SE(\hat{\theta}_{treat}-\hat{\theta}_{control})=\sqrt{\frac{s^2_{treat}}{n_{treat}}+\frac{s^2_{control}}{n_{control}}}$ kunne jeg konkludere, at på nogle baggrundskarakteristika var de to grupper forskellige, og der var behov for, at matche dem. I de følgende afsnit kommer jeg ind på en kendt diskussion omkring matching. Indeværende artikel handler altså ikke så meget om mit forsøg og tilhørende resultater, men mere om diskussionen bag. I artiklen anvendes forsøg og treatment i flæng.

Valg af kontrolgruppe og forsøgsevaluering

Jeg har i mit speciale taget udgangspunkt i en diskussion omkring matching af kontrol- og interventionsgruppe som startede med Rubin (1974). Rubins model defineres ofte som 'det observerede udfalds' model og er givet ved 0. y_1 eller $y(T)$ og y_0 eller $y(C)$er de to mulige udfald, som er latente hos subjektet. y_1 kunne således være testresultatet hos en elev, som har undergået et særligt tiltag i en folkeskole. y_0 er resultatet hos den **samme** elev, som ikke har undergået et sådant tiltag. Effekten af et tiltag kan da måles som $y_1 - y_0$, dettte kan skaleres op til N elever/forsøg. Den gennemsnitlige effekt ville da være $\frac{1}{N}\sum_{i=1}^{N}[y_i(1) - y_i(0)]$. Da man ikke kan teste et tiltag hos den samme elev, har man således med et manglende data problem at gøre. Det man observerer er i virkeligheden en vægtning af de to mulige udfald.

Lad $w = 1$ hvis en elev har gennemgået et særligt tiltag og lad $w = 0$ hvis denne elev ikke har gennemgået dette. Det observerede udfald vil da være givet ved 1).

1) $y = wy_1 + (1 - w)y_0 \Leftrightarrow y = y_0 + w(y_1 - y_0)$

For at overkomme det manglende data problem kan man udføre et randomiseret eksperiment. Med to individer ser dette sådan ud: gennemsnittet af effekten af en intervention hos individ 1 $y_1(T) - y_1(C)$ og 2 $y_2(T) - y_2(C)$ givet i 2).

2) $\frac{1}{2}[(y_1(T) - y_2(C)) + (y_2(T) - y_1(C))]$

Ligning 2) svarer til at man udførte forsøget to gange på det samme individ. I indeværende forsøg har jeg ikke haft adgang til en kontrolgruppe, blot en stor gruppe af potentielle kontroller. Valget på en endelig kontrolgruppe er baseret på en diskussion mellem (LaLonde 1986), (Dehejia and Wahba 1999) , (Smith and Todd 2005) og til sidst (Sekhon 2008), (King and Nielsen 2016) og (Heckman, Ichimura et al. 1997).

Heckman's selektions model:

Formålet med et forsøg er, at kunne måle den gennemsnitlige 'treatment' effekt τ_{ate} eller under mindre strenge antagelser den gennemsnitlige treatment effect på forsøgsgruppen τ_{att}.

3) $y = y_0 + w(y_1 - y_0)$

Ligning 3) er ligning 2) fra afsnittet ovenfor hvor C er skiftet ud med 0 og T er skiftet ud med 1. Heckman tager derefter forventningen til ligning 3).

4) $E(y|w = 1) - E(y|w = 0) = E(y_0|w = 1 + w(y_1 - y_0)|w = 1) - E(y_0|w = 0 + w(y_1 - y_0)|w = 0) \Rightarrow E(y_0|w = 1 + w(y_1 - y_0)|w = 1) - E(y_0|w = 0)$

Ligning 4) fremkommer fordi $E(w(y_1 - y_0)|w = 0) = 0$. 4) omskrives til 5)

5) $c + E(y_0|w = 1) - E(y_0|w = 0) + \underbrace{(E(y_1 - y_0) - E(y_1 - y_0))}_{\text{indsatte led} = 0} \Rightarrow 6)$

6) $\underbrace{E(y_1 - y_0)}_{\tau_{ate}} + \overbrace{\left\{\underbrace{E((y_1 - y_0)|w = 1)}_{\tau_{att}} - E(y_1 - y_0)\right\}}^{Sorting\ gain} + \underbrace{E(y_0|w = 1) - E(y_0|w = 0)}_{Selection\ bias}$

Det ses altså at ligning forventningen til den observerede udfaldsmodel kan skrives som en kombination af den gennemsnitlige 'treatment' effekt og det der refereres til som 'sorting gain' og selektions bias. Kort forklares den gennemsnitlige 'treatment' effekt på forsøgsgruppen τ_{att}, inden sorting gain og selektion bias beskrives.

Optimalt vil man gerne sikre, at dem, der er i forsøgsgruppen og kontrolgruppen kan sammenlignes. I det tilfælde observerbare og uobserverbare baggrundsvariable er fuldstændigt sammenlignelige hos de to grupper er $\tau_{ate} = \tau_{att}$. Tildeling af status som forsøgsperson er uafhængig af mulige udfald. $w \perp (y_1, y_0)$.

Hvis de grupper ikke længere er sammenlignelige på baggrund af observerbare baggrundsvariable, er det stadig muligt at estimere τ_{att}. Vi så ovenfor at i tilfælde af et

Sorting gain = τ_{ate}- τ_{att} , det vil sige at $w \perp (y_0)$, at der er en ekstra gevinst ved at deltage (Heckman, Ichimura et al. 1996).

Selektions bias derimod kan have tre årsager (Heckman, Ichimura et al. 1996):

- Bias som følge af **non-overlapping support**. Værdier af enkelte variable optræder i kontrolgruppen og ikke i forsøgsgruppen.
- Selvom der er overlapping support, så kan tæthedsfunktionen være forskellig i de to grupper.
- Selection bias – udeladt variabel bias.

Support er i denne sammenhæng det samme som fordelingen af værdier af baggrundsvariable hos kontrol og treatment gruppen. Jo mere overlap jo mere sammenlignelige.

I nedenstående afsnit kommer jeg ind på matching som en del af løsningen på selektions bias. Se her for yderligere diskussion af emnet:

Heckmans selektions model antagelser:
De to ovenstående antagelser kan udvides til, at ved at betinge på en vektor af baggrundsvariable $\boldsymbol{x}$ kan man antage 'ignorability of treatment'. Betinget på $\boldsymbol{x}$ så $w \perp (y_0, y_1)$ eller $w \perp (y_0)$. Det sidste kræver at, der er et overlap af $\boldsymbol{x}$ mellem kontrol- og forsøgsgruppe.

Overlap hænger sammen med propensity scoren. Propensity scoren er egentlig sandsynligheden for tildeling af treatment: Logit propensity score model: $P(Treatment) = p(x) = \frac{\exp(x_i\beta)}{[1+exp(x\beta)]}$ [1]

Overlap defineres som $0 < p(x) < 1$. For at estimere τ_{att} kræver det at, $p(x) < 1$. Det vil sige, at $p(x)$ kan være 0 for nogle værdier af $\boldsymbol{x}$.

SUTVA: Stable unit assumption. Når et individ gennemgår et forsøg smitter dette ikke af på de resterende individer både kontrol og forsøgsindivider.

Ved at udføre et kontrolleret randomiseret forsøg vil man netop kunne opnå **gennemsnitlig balance** mellem kontrolgruppe og forsøgsgruppe, dette er både med hensyn til de observerbare og ikke observerbare baggrundsvariable (King and Nielsen 2016). Alternativt kan man udføre et såkaldt 'fully blocked' (FB) eksperiement. Det sidste er en måde hvorpå man søger at matche på baggrundsvariable, sådan at der er **absolut ingen observerbare forskelle** på kontrol- og forsøgsgruppen. Derudover vil FB opnå en gennemsnitlig balance af de uobserverbare forskelle.

Ikke eksperimentelle metoder
Der er adskillige ikke eksperimentielle metoder inden for økonometri hvis ikke man har adgang til et komplet randomiseret forsøg (Smith and Todd 2005) (Caliendo 2006).

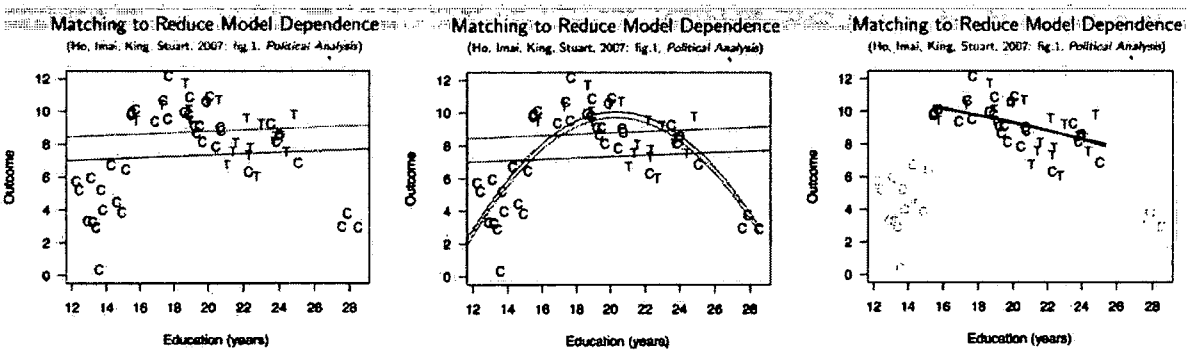
[1] Kan også være en hvilken som helst anden sandsynligheds model

- Før og efter estimator: $Y_{t\_post} - Y_{t_{pre}} = \varphi\left(x_{t_{post}}\right) - \varphi\left(x_{t_{pre}}\right) + \tau * w_{it} + \varepsilon_{t_{post}} - \varepsilon_{t_{pre}}$
  Kræver at individet ikke er blevet påvirket af andet mens tiden er gået.
- Dif in Dif model: $y_{it} - y_{it'} = \varphi(x_{it}) - \varphi(x_{it'})^2 + \alpha * w_i + (u_{it} - u_{it'})$
  Her antages det at forsøg og kontrolgruppe ville følge den samme udvikling, også i fraværet af et forsøg.

Observationsstudie metoder der søger at efterligne eksperimentelle metoder – Matching

Grundlæggende så matcher man en person fra kontrolgruppen med en fra forsøgsgruppen på baggrund af observerbare variable under antagelse af 'ignorability og treatment' (Rosenbaum and Rubin 1985) og (Caliendo 2006).

Matching minimerer risiko for modelafhængighed (King and Nielsen 2016):



Ovenfor ses det at hvis man fjerner kontroller C, som ikke har en tilhørende forsøgsperson T, så vil risikoen for modelafhængighed falde. I det første billede konkluderes det at effekten af et uddannelse får outcome til at stige. Billede 2 viser, at outcome falde.

Mahalonobis metric:

Denne metode matcher kontrol- og forsøgsgruppe med hinanden direkte på baggrund af baggrunds karakteristika. $\text{MD}(X_i, X_j) = \sqrt{(X_i - X_j)^T S^{-1} (X_i - X_j)}$ (King and Nielsen 2016)

Differencen mellem X_i og X_j, hvor I og j er henholdsvis kontrol- og forsøgsgruppe standardiseres for, at teste om denne er langt fra nul. Gary King kritiserer denne standardisering, for det må være op til undersøgeren at fastsætte en rimelig grænse for hvad differencen må være. Se i øvrigt også 'Coarsend matching' (King and Nielsen 2016).

Propensity score matching

Matching på baggrund af propensity scoren er blevet en foretrukken metode at matche grupper, da man ikke skal bekymre sig at matche i mange dimensioner, som det er tilfældet med Mahalonobis metric. Hvis man sammenligner grupper på baggrund af en sandsynlighed, har man dermed skaleret problemet ned til en dimension (Caliendo 2006).

[2] $\varphi(x_{it})$ er en function af de kontrolvariable der kan påvirke den afhænge variabel, nemlig udviklingen over tid.

Nearest Neighbour matching: Denne metode matcher en deltager fra forsøgsgruppen med mulige deltagere fra kontrolgruppen. En passende afstand $|\hat{p}_i - \hat{p}_j| < \delta$ defineres. Man kan vælge en kontrol deltager per deltager fra forsøgsgruppen, eller matche flere. Man kan også matche med tilbagelægning. Det er et valg mellem bias og varians (King and Nielsen 2016)

Caliper matching: Er en variant af nearest neighbour matching. Der defineres en maksimum afstand $|\hat{p}_i - \hat{p}_j| < \delta$ således at dårlige matches smides væk, dermed kan også deltagere fra forsøgsgruppen blive smidt væk.

Stratificering eller interval matching. Overlappet er inddelt i intervaller og estimationen af den samlede effekt er et vægtet estimat af effekten indenfor hvert interval (Smith and Todd 2005) (Dehejia and Wahba 1999).

Kernel matching (Smith and Todd 2005): giver observationer fra kontrolgruppen en vægt.

$$effekt = \frac{1}{n_1}\sum_{i\in I_1}\left\{Y_{1i} - \frac{\sum_{j\in I_0} Y_{0j} G\left(\frac{P_j - P_i}{\alpha_n}\right)}{\sum_{k\in I_0} G\left(\frac{P_k - P_i}{\alpha_n}\right)}\right\}$$ α_n er en 'bandwidth' parameter. $\frac{\sum_{j\in I_0} G\left(\frac{P_j - P_i}{\alpha_n}\right)}{\sum_{k\in I_0} G\left(\frac{P_k - P_i}{\alpha_n}\right)}$ er en vægtning.

Difference in difference matching: Denne metode fjerner en eventuel tidsafhængig uobserverbar variabel.

$$effekt = \frac{1}{n_1}\sum_{i\in I_1\cap S_p}\left\{(Y_{1ti} - Y_{0t'i}) - \sum_{i\in I_1\cap S_p} W(i,j)(Y_{0tj} - Y_{0t'j})\right\}$$

Dif in Dif matching estimatoren er en variant af den kendte dif in dif estimator men her med en vægtet kontrolgruppe differens.

Kombination af matching og analyse

Man kan bruge matching som et såkaldt 'prepsocessing tool' og derefter udfører en analyse ganske som man ville gøre det uden matching (King and Nielsen 2016) (Caliendo 2006).

Valg af matching metode

Dette afsnit tager udgangspunkt i et amerikansk forsøg på, at få udsatte folk i arbejde i projektet der omtales som NSW og fandt sted i midten af 1970. Forsøget giver adgang til en forsøgsgruppe og en kontrolgruppe tilfældigt udvalgt. Forsøget er et eksempel på et fuldstændigt randomiseret forsøg, og er blevet udnyttet til at bedømme alternative de estimatorer, man har gjort brug af i et observationsstudie, nemlig regressionsjustering og en "Two-step-estimator" (LaLonde 1986). I sit evalueringsstudie sammenligner LaLonde resultatet fra det randomiserede forsøg med andre datasæt og surveys på lignende oplysninger. Uden videre vælger LaLonde de observationer fra de andre surveys der balancerer forsøgsgruppen og kontrollerne i de variable LaLonde vurderer mest essentielle. (Dehejia and Wahba 1999) kritiserer LaLonde for, at for subjektiv i son vurdering af hvilke baggrundsvariable, der bør balancere mellem forsøgsgruppen og den endelige kontrolgruppe. Ydermere når LaLonde udvælger de observationer af bpde forsgruppen og kontroller, som han vil bruge til sin analyse er han for lidt restriktiv. LaLonde vælger således kun de

observationer hvor der er observeret en indkomst i 1975, hvor forsøget starter. LaLonde har ikke nok information til, at udelukke eventuelle usædvanlige hændelser, der også kunne påvirke beskæftigelse og løn (Aschenfelters dip) (Angrist and Pischke 2009).

Kort om NSW og andre surveys

Målgruppe: Kvinder der betegnes som AFDC kvinder, forhenværende stofmisbrugere, high-school dropouts,

Formål: Dem der tilnævntes forsøgspersoner var garanteret et job i 9 til 18 måneder. Efter 1978 måltes effekten da på indkomst.

Baggrundsvariable: Køn, alder, etnicitet, drop-outs, civilstatus, information om løn og joberfaring før træning.

I det fulde randomiserede eksperiment balancerer alle baggrundsvariable (LaLonde 1986, s. 606 tabel 1).

Alternative kontrolgrupper er taget fra andre surveys. Kriterierne for, at være med omfatter bl.a. at man ikke var pensioneret på det givne tidspunkt, at man ikke havde et job og lignende kriterier.

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE *PSID* AND THE *CPS-SSA*[a,b]

Name of Comparison Group[d]	Comparison Group Earnings Growth 1975–78 (1)	NSW Treatment Earnings Less Comparison Group Earnings: Pre-Training Year, 1975 Unadjusted (2)	Pre-Training Year, 1975 Adjusted[c] (3)	Post-Training Year, 1978 Unadjusted (4)	Post-Training Year, 1978 Adjusted[c] (5)	Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons Without Age (6)	Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons With Age (7)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78 Unadjusted (8)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78 Adjusted[c] (9)	Controlling for All Observed Variables and Pre-Training Earnings (10)
Controls	$2,063	$39	$−21	$886	$798	$847	$856	$897	$802	$662
	(325)	(383)	(378)	(476)	(472)	(560)	(558)	(467)	(467)	(506)
PSID-1	$2,043	−$15,997	−$7,624	−$15,578	−$8,067	$425	−$749	−$2,380	−$2,119	−$1,228
	(237)	(795)	(851)	(913)	(990)	(650)	(692)	(680)	(746)	(896)
PSID-2	$6,071	−$4,503	−$3,669	−$4,020	−$3,482	$484	−$650	−$1,364	−$1,694	−$792
	(637)	(608)	(757)	(781)	(935)	(738)	(850)	(729)	(878)	(1024)
PSID-3	($3,322	($455	$455	$697	−$509	$242	−$1,325	$629	−$552	$397
	(280)	(539)	(704)	(760)	(967)	(884)	(1028)	(257)	(967)	(1103)
CPS-SSA-1	$1,196	−$10,585	−$4,654	−$8,870	−$4,416	$1,714	$195	−$1,543	−$1,102	−$805
	(61)	(539)	(509)	(562)	(557)	(452)	(441)	(426)	(450)	(484)
CPS-SSA-2	$2,684	−$4,321	−$1,824	−$4,095	−$1,675	$226	−$488	−$1,850	−$782	−$319
	(229)	(450)	(535)	(537)	(672)	(539)	(530)	(497)	(621)	(761)
CPS-SSA-3	$4,548	$337	$878	−$1,300	$224	−$1,637	−$1,388	−$1,396	$17	$1,466
	(409)	(343)	(447)	(590)	(766)	(631)	(655)	(582)	(761)	(984)

[a] The columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is $886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

[b] Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

[c] The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

[d] See Table 3 for definitions of the comparison groups.

Tabel 1: Tabel kopieret fra LaLondes studie af effekten af jobtræning. (LaLonde 1986, s. 610 tabel 5)

Tabel 1 viser resultaterne både med og uden regressionsjustering brugt på LaLondes øvrige kontrolgrupper sammenlignet med resultatet fra det randomiserede forsøg. Bemærk resultaterne kommer fra studiet af effekten på mænds indkomst i 1978. Resultaterne i den

røde boks er forskellige modeller kørt med kontrolgruppen fra det fuldstændigt randomiserede forsøg. I den blå ramme ses effekten på indkomst i 1978 ($re78$), til højre er der kontrolleret for baggrundsinformation og til venstre er der ikke kontrolleret. I den sorte ramme kontrollerer LaLonde også for uobserverbare variable med forskellige variationer af 'Dif in Dif' estimatoren. Den yderste venstre kolonne er de forskellige kontrolgrupper. Det ses at resultaterne er afhængige af valget af kontrolgruppe og også kombineret med valget af model. LaLonde forsøger at afhælpe dette ved brug af Heckmans 'Two-step' estimator (Wooldridge 2010).

TABLE 6—ESTIMATED TRAINING EFFECTS USING TWO-STAGE ESTIMATOR NSW Males

Variables Excluded from the Earnings Equation, but Included in the Participation Equation	Comparison Group	Heckman Correction for Program Participation Bias, Using Estimate of Conditional Expectation of Earnings Error as Regressor in Earnings Equation: Estimate of Coefficient for Training Dummy	Estimate of Expectation
Marital Status, Residency in an SMSA, Employment Status in 1976, AFDC Status in 1975, Number of Children	*PSID*-1	−1,333	−2,357
		(820)	(781)
	CPS-SSA-1	−22	−1,437
		(584)	(449)
	NSW Controls	899	−835
		(840)	(2601)
Employment Status in 1976, Number of Children	*PSID*-1	−1,161	−2,655
		(864)	(799)
	CPS-SSA-1	13	−1,484
		(584)	(450)
	NSW Controls	889	−808
		(841)	(2603)
No Exclusion Restrictions	*PSID*-1	−667	−2,446
		(905)	(806)
	CPS-SSA-1	213	−1,364
		(588)	(452)
	NSW Controls	889	−876
		(840)	(2601)

Notes: The estimated training effects are in 1982 dollars. For the females, the experimental estimate of impact of the supported work program was $851 with a standard error of $317. The one-step estimates from col. 11 of Table 4 were $2,097 with a standard error of $491 using the *PSID*-1 as a comparison group, $1,041 with a standard error of $503 using the *CPS-SSA*-1 as a comparison group, and $854 with a standard error of $312 using the NSW controls as a comparison group. Estimates are missing for the case of three exclusions using the NSW controls since AFDC status in 1975 cannot be used as an instrument for the NSW females. For the males, the experimental estimate of impact of the supported work program was $886 with a standard error of $476. The one-step estimates from col. 10 of Table 5 were $−1,228 with a standard error of $896 using the *PSID*-1 as a comparison group, $−805 with a standard error of $484 using the *CPS-SSA*-1 as a comparison group, and $662 with a standard error of $506 using the NSW controls as a comparison group. Estimates are missing for the case of three exclusions for the NSW males as AFDC status is not used as an instrument in the analysis of the male trainees.

Tabel 2: Heckman's 'Two-step' estimator (LaLonde 1986, s. 616 tabel 6)

Effekten af jobtræning estimeret I en 'Two-step' estimator er estimeret til at være 213. Dette skal sammenlignes med estimaterne fra den gule ramme i Tabel 1. Det ses, at der er en væsentlig forbedring,

Propensity score matching

I deres studie også af NSW jobtræning konkluderer (Dehejia and Wahba 1999) herefter DW, at hvis man udvider informationssættet til også at inkludere, de forsøgs og kontrolpersoner,

hvor man kender deres indkomst i 1974, eliminerer ma risikoen, for at estimere Aschenfelters dip. Resultat fra deres første estimation findes i **Fejl! Henvisningskilde ikke fundet.**.

	A. Lalonde's original sample					C. RE74 subsample (results use RE74)				
	NSW treatment earnings less comparison group earnings 1978		Unrestricted differences in differences: Quasi-difference in earnings growth 1975–1978		Controlling for all variables[f]	NSW treatment earnings less comparison group earnings - 1978		Unrestricted differences in differences: Quasi-difference in earnings growth 1975–1978		Controlling for all variables[f]
Comparison group	Unadjusted[b] (1)	Adjusted[c] (2)	Unadjusted[d] (3)	Adjusted[e] (4)	(5)	Unadjusted[b] (1)	Adjusted[c] (2)	Unadjusted[d] (3)	Adjusted[e] (4)	(5)
NSW	886	798	870	802	820	1,794	1,688	1,750	1,672	1,685
	(472)	(472)	(497)	(488)	(488)	(633)	(636)	(632)	(638)	(640)
PSID-1	−15,578	−8,067	−2,380	−2,110	−1,844	−15,205	−870	−862	216	731
	(913)	(990)	(680)	(748)	(782)	(1150)	(931)	(941)	(866)	(886)
PSID-2	−4,020	−3,492	−1,364	−1,694	−1,676	−3,647	94	721	907	683
	(781)	(935)	(729)	(878)	(886)	(959)	(1042)	(868)	(1004)	(1028)
PSID-3	697	−509	629	−562	−576	1,070	821	1,370	822	825
	(760)	(967)	(757)	(967)	(968)	(902)	(1100)	(897)	(1101)	(1104)
CPS-1	−8,870	−4,416	−1,543	−1,102	−997	−8,498	−8	−76	799	972
	(562)	(577)	(429)	(450)	(452)	(712)	(672)	(537)	(547)	(550)
CPS-2	−4,195	−2,841	−1,848	−1,129	−1,149	−3,822	815	−288	879	700
	(533)	(620)	(469)	(551)	(551)	(671)	(672)	(574)	(664)	(658)
CPS-3	−1,008	−1	−1,204	−263	−234	−635	1,270	−91	1,326	1,326
	(539)	(641)	(592)	(677)	(675)	(657)	(798)	(641)	(798)	(798)

NOTES: Panel A replicates the results of Lalonde (1986, table 5). The estimates for columns (1)–(4) for NSW, PSID-1–3, and CPS-1 are identical to Lalonde's. CPS-2 and CPS-3 are similar but not identical, because we could not exactly reproduce his subset. Column (5) differs because the data file that we obtained did not contain all of the covariates used in column (10) of Lalonde's Table 5.

[a] Estimated effect of training on RE78. Standard errors are in parentheses. The estimates are in 1982 dollars.

[b] The estimates based on the NSW control groups are unbiased estimates of the treatment impacts for the original sample ($886) and for the RE74 sample ($1,794).

[c] The exogenous variables used in the regression-adjusted equations are age, age squared, years of schooling, high school dropout status, and race (and RE74 in Panel C).

[d] Regresses RE78 on a treatment indicator and RE75.

[e] The same as (d), but controls for the additional variables listed under (c).

[f] Controls for all preexistent covariates.

Tabel 3: (Dehejia and Wahba 1999, s. 1056 tabel 2) Effekten af jobtræning kun for mændene.

DW udfører derefter den samme analyse DW udfører derefter den samme analyse men nu matcher de på p(x).

	NSW earnings less comparison group earnings		NSW treatment earnings less comparison group earnings, conditional on the estimated propensity score					
			Quadratic in score[b]	Stratifying on the score			Matching on the score	
	(1) Unadjusted	(2) Adjusted[a]	(3)	(4) Unadjusted	(5) Adjusted	(6) Observations[c]	(7) Unadjusted	(8) Adjusted[d]
NSW	1,794	1,672						
	(633)	(638)						
PSID-1[e]	−15,205	731	294	1,608	1,494	1,255	1,691	1,473
	(1,154)	(886)	(1,389)	1,571)	(1,581)		(2,209)	(809
PSID-2[f]	−3,647	683	496	2,220	2,235	389	1,455	1,480
	(959)	(1,028)	(1,193)	1,768)	(1,793)		(2,303)	(808
PSID-3[f]	1,069	825	647	2,321	1,870	247	2,120	1,548
	(899)	(1,104)	(1,383)	1,994)	(2,002)		(2,335)	(826
CPS-1[g]	−8,498	972	1,117	1,713	1,774	4,117	1,582	1,616
	(712)	(550)	(747)	1,115)	(1,152)		(1,069)	(751
CPS-2[g]	−3,822	780	505	1,543	1,622	1,493	1,788	1,563
	(670)	(658)	(847)	1,461)	(1,345)		(1,205)	(753
CPS-3[g]	−635	1,326	556	1,252	2,219	514	587	862
	(657)	(798)	(951)	1,617)	(2,062)		(1,496)	(776

[a] Least squares regression: RE78 on a constant, a treatment indicator, age, age^2, education, no degree, black, Hispanic, RE74, RE75.

[b] Least squares regression of RE78 on a quadratic on the estimated propensity score and a treatment indicator, for observations used under stratification; see note (g).

[c] Number of observations refers to the actual number of comparison and treatment units used for (3)–(5); namely, all treatment units and those comparison units whose estimated propensity score is greater than the minimum, and less than the maximum, estimated propensity score for the treatment group.

[d] Weighted least squares: treatment observations weighted as 1, and control observations weighted by the number of times they are matched to a treatment observation (same covariates as (a)).

Propensity scores are estimated using the logistic model, with specifications as follows:

[e] PSID-1: Prob (T_i = 1) = F(age, age^2, education, education^2, married, no degree, black, Hispanic, RE74, RE75, RE74^2, RE75^2, u74*black).

[f] PSID-2 and PSID-3: Prob (T_i = 1) = F(age, age^2, education, education^2, no degree, married, black, Hispanic, RE74, RE74^2, RE75, RE75^2, u74, u75).

[g] CPS-1, CPS-2, and CPS-3: Prob (T_i = 1) = F(age, age^2, education, education^2, no degree, married, black, Hispanic, RE74, RE75, u74, u75, education*RE74, age^3).

Tabel 4: Resultater på baggrund af matching på p(x) (Dehejia and Wahba 1999). Effekten af jobtræning kun for mændene.

Efter at have estimeret effekten træningsprogrammet på baggrund af matching ender DW med, at konkludere, at ikke eksperimentelle estimatorer kan bruges til, at efterligne et forsøg. Det skal bemærkes at alle de ikke eksperimentelle estimatorer fra Tabel 4 er mellem 587 og 2235 hvor af de fleste er omkring de 1500, som stemmer godt overens med resultatet på 1794 fra det randomiserede forsøg.

Jeg vil ikke lægge meget vægt på forskellen på resultaterne af stratificering på baggrund af propensity scoren og matching på scoren, blot gennemgå det kort her.

Efter man har estimeret P er kontrolenheder der falder udenfor mængden af estimerede propensity scores for forsøgsenhederne smidt væk. Derefter har man opdelt sandsynligheden i intervaller og samlet forsøgs og kontrolenheder i de samme intervaller. Eksempelvis er der 20 CPS kontrolenheder i intervallet fra 0.55 til 0.60, og ca. dobbelt så mange forsøgsenheder. Når et strata er defineret tjekkes det om der er signifikant forskel på baggrundsvariablene for forsøgsgruppen og de forskellige kontrolgrupper. Er der forskel estimeres P på ny med kvadrerede led og så videre.

Kritik af Dehejia og Wahba hyldest af matching på propensity scoren

I 2005 bliver studiet af DW mødt af kritik fra (Smith and Todd 2005). DW ender nemlig med, at smide 40% af LaLondes observationer væk, når de udvider forsøget til også at inkludere indkomst fra 1974, så der er to år med indkomstdata før forsøget. DW undersøger simpelthen en mindre stikprøve. En del af dem de inkluderer har ingen indkomst i 1974. Smith Todd tester DW propensity score ud på LaLondes originale eksempel og finder, at bias ikke er reduceret, sådan som DW ellers har konkluderet.

Smith og Todd De konkluderer endvidere, at når DW vil inkludere indkomst i år 1974 bruger de en variabel der hedder indkomst 13-24 måneder før forsøgets start, også selv hvis de har folk med der startede senere end en periode i 1976. Deres stikprøve er dermed mindre og med de forkerte observationer med. Smith og Todd ender med, at konkludere at denne undersøgelse med de tilgængelige kontrolgrupper reduceres bias bedst ved Dif in Dif matching estimatoren. DW har ikke som LaLonde kørt estimationer har hverken dif in dif eller dif in dif matching estimatoren.

Specifikations tests

Både Lalonde, DW og Smith og Todd tester de standardiserede differenser før estimationen. Dette er både før og efter matching. Gary King (King and Nielsen 2016) adresserer dette i sit papir. De konkluderer nemlig at lige netop på det punkt bør forskeren udnytte sin specialiserede viden omkring hvad der kan accepteres som distance. Heckman, Ichimura et al. (1997) anbefaler at bruge den del af kontrolgruppen, der valgte ikke at deltage i forsøget, men som blev tilbudt det 'No-show' gruppe, til at teste om den kontrolgruppe, man vælger at fortsætte med, er meget forskellig. Dette tyder på at bias opstår på grund af noget

uobserverbart. Ofte tester man forskellen på gennemsnittet af baggrundsvariable i de to grupper, men ikke yderligere overlap (Sekhon 2008) (Heckman, Ichimura et al. 1997).

Opsummering og diskussion

Baseret på erfaringer fra LaLonde, Dehejia og Wahba og til sidst Smith og Todd opsummeres her hvad der er vigtigt når man udvælger en kontrolgruppe og en tilhørende estimator:

1. For at matching skal have lav bias skal data inkludere mange baggrundsvariable relateret til både deltagelse i eksperiment og udfaldsvariabel (løn).
2. At kontrolgruppen skal være fra den samme overordnede population som forsøgsgruppen. I dette tilfælde skal kontrolgruppen være fra det samme marked for arbejdskraft, som forsøgsgruppen.
3. Den afhængige variabel, udfaldet, i dette tilfælde lønnen skal være sammenlignelig. Løn skal være betaling for arbejdskraft.

Antagelsen om at man kan matche på baggrund af propensity scoren hviler på, at $\boldsymbol{E(w|X, \Pr(w = 1|X)) = E(D|\Pr(w = 1|X))}$. Der er altså ikke yderligere informationer omkring tildeling af forsøgsperson status når man har betinget på $p(x)$.

Gary King anfægter de første fortalere for PSM Rubin og Rosenbaum RR. RR påviste, at $\boldsymbol{y_0 \perp w|X}$, antagelserne om overlap og SUTVA medfører $\boldsymbol{y_0 \perp w|p(x)}$. Dette er ikke det samme som, at $\boldsymbol{y_0 \perp w|p(x)}$ medfører $\boldsymbol{y_0 \perp w|X}$. Det kommer an på hvordan $\boldsymbol{p(x)}$ bliver estimeret det tager simpelthen ikke højde for variation i $\boldsymbol{X|p(x)}$. Selvom baggrundsvariable matcher i kontrol- og forsøgsgruppen $\boldsymbol{x_0 = x_1}$ dette betyder ikke at $\boldsymbol{\hat{p}(x)_0 = \hat{p}(x)_1}$.

King and Nielsen (2016) kommer ind på, at når man bruger PSM for, at sikre overlap og dermed smider observationer væk, sker dette mere eller mindre tilfældigt. Et simpelt eksempel viser dette. Udgangspunktet er en perfekt balanceret stikprøve med en mand og en kvinde i både i kontrol- og forsøgsgruppen. Der er er altså 50% sandsynlighed for, at være i forsøgsgruppen både for mænd og kvinder. I tilfælde af at man vælger at droppe observationer for, at sikre overlap vil man med 50% sandsynlighed ende med en ubalanceret stikprøve (M, M) eller (K, K), med bias og stigning i varians til følge (inefficiens).

Genetic matching

Baseret på diskussionen af matching på propensity scoren, og kritikken heraf har Sekhon (2008) udviklet en iterativ metode der optimerer balancen mellem forsøgs- og kontrolgruppe. Genetic matching lægger vægt på, at der tjekkes for balance, efter propensity scoren er estimeret. Genetic matching inkluderer i øvrigt et Kolmogorov-Smirnov test af forskel i fordeling af værdier af baggrundsvariable. Udgangspunktet er Mahalonobis metric fra før, men hvor propensity scoren er inkluderet som en variabel som man kan matche på, derfor matches der nu på Z_i . $GMD(Z_i, Z_j, W) = \sqrt{(Z_i - Z_j)^T (S^{-\frac{1}{2}})^T W S^{-\frac{1}{2}} (Z_i - Z_j)}$. W er en vægtmatrix der tildeler hver variabel en vægt afhængig af hvor meget denne variabel skal betyde når den iterative proces søger at minimere en tabsfunktion. Processen tester undervejs ved hvert trin

om gennemsnittet og hele fordelingen er ens for de to grupper. Sekhon (2008) tester Genetic Matching proceduren også på LaLondes datasæt og konkluderer, at matching på baggrund af Genetic Matching forbedres, derudover er estimaterne af effekten for hver kontrolgruppe mindre svingende.

Litteratur

Allcott, H. (2011). "Social norms and energy conservation." Journal of Public Economics 95: 1082-1095.

Angrist, J. and J.-S. Pischke (2009). Mostly harmless econometrics: an empiricist's companion.

Caliendo, M. (2006). Microeconometric evaluation of labour market policies, Springer Science & Business Media.

Dehejia, R. H. and S. Wahba (1999). "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs." Journal of the American Statistical Association 94(448): 1053-1062.

Heckman, J. J., et al. (1996). "Sources of selection bias in evaluating social programs: An interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method." Proceedings of the National Academy of Sciences 93(23): 13416-13420.

Heckman, J. J., et al. (1997). "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme." The review of economic studies 64(4): 605-654.

King, G. and R. Nielsen (2016). "Why propensity scores should not be used for matching."

LaLonde, R. J. (1986). "Evaluating the econometric evaluations of training programs with experimental data." The American Economic Review: 604-620.

Rosenbaum, P. R. and D. B. Rubin (1985). "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score." The American Statistician 39(1): 33-38.

Rubin, D. B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." Journal of educational Psychology 66(5): 688.

Sekhon, J. S. (2008). "Multivariate and propensity score matching software with automated balance optimization: the matching package for R." Journal of Statistical Software, Forthcoming.

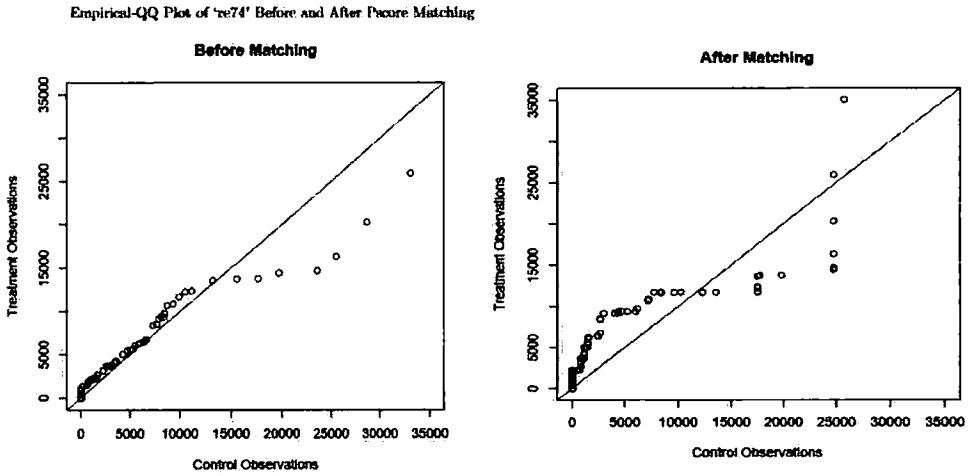
Smith, A., Jeffrey and E. Todd, Petra (2005). "Does matching overcome LaLonde's critique of nonexperimental estimators?" Journal of Econometrics 125(1–2): 305-353.

Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data, MIT press.

Appendix

Resultater fra DW propensity score (Sekhon 2008):

```
dw.pscore  <- glm(treat~age + I(age^2) + educ + I(educ^2) + black +
                  hisp + married + nodegr + re74  + I(re74^2) + re75 +
                  I(re75^2) + u74 + u75, family=binomial, data=lalonde)
```

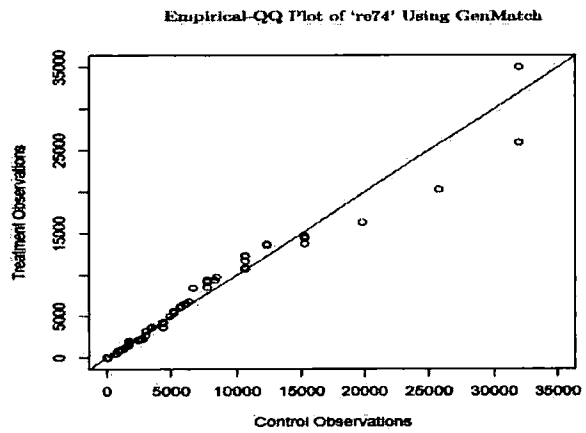


Resultater fra Sekhons Genetic Matching

```
X  <- cbind(age, educ, black, hisp, married, nodegr, re74, re75, u74, u75)

BalanceMatrix <- cbind(age, I(age^2), educ, I(educ^2), black,
                       hisp, married, nodegr, re74 , I(re74^2), re75,
                       I(re75^2), u74, u75, I(re74*re75), I(age*nodegr),
                       I(educ*re74), I(educ*re75))

gen1 <- GenMatch(Tr=Tr, X=X, BalanceMatrix=BalanceMatrix, pop.size=10000)
```



Adaptive sample size planning in repeatability studies on quantitative measurements

Oke Gerke[1,2*], Mie Holm Vilstrup[1], Ulrich Halekoh[3]

[1]Department of Nuclear Medicine, Odense University Hospital, Sdr. Boulevard 29, 5000 Odense C, Denmark.
[2]Centre of Health Economics Research, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark.
[3]Epidemiology, Biostatistics and Biodemography, University of Southern Denmark, J. B. Winsløws Vej 9b, 5000 Odense C, Denmark.

*Correspondence: oke.gerke@rsyd.dk

Abstract

Diagnostic imaging in cancer research is often done with hybrid positron emission tomography/computed tomography (PET/CT), using the glucose analogue 18F-fluorodeoxyglucose (FDG) as tracer. A frequently used measure of tumour uptake in PET imaging is the standardised uptake value (SUV). Any quantitative measurement of disease needs to be accurate and reliable to justify its clinical use. Agreement of paired measurements is well-established and analysed by Bland-Altman limits of agreement (mean difference ± 1.96 standard deviation of the differences), but sample size planning is less clear. It may focus on the width of the Bland-Altman band or on 95% confidence intervals of the limits of agreement themselves, but often a priori information is scarce. We focus on the population standard deviation of the paired differences, propose an adaptive sample size strategy using a type 1 error-spending function, and discuss alternative nonparametric test procedures for the case of non-normally distributed differences. Regarding the latter, a nonparametric quasi-range based on rank statistics is used as analogue to the population standard deviation. Our approach is exemplified with data from an ongoing study in ovarian cancer in which pre-operative scans were assessed twice by the same clinician. Interim analyses were defined post hoc after 15 and 30 patients, final analysis was done with 45 patients. We conclude that group-sequentially testing in agreement studies offers the necessary flexibility in the planning stage, but secures the experiment-wise significance level. Whilst group-sequential testing is widely used in pivotal therapeutic studies, it is underused in diagnostic research.

Background

Any quantitative measurement of disease needs to be both accurate and reliable to justify its clinical use. Reliability concerns the ability of a test to distinguish patients from each other, despite measurement error, while agreement focuses on the measurement error itself [1, 2]. Reliability is assessed with intraclass correlation coefficients, and sample size planning targets the width of respective 95% confidence intervals [3-5]. Assuming a Normal distribution for the difference of paired measurements, agreement is analysed by Bland-Altman limits of agreement [6-8], but sample size planning is less clear. It may focus on the difference of the Bland-Altman limits or on the precision of the estimated limits of agreement in terms of respective 95% confidence intervals [8, 9]. In extension of the latter, a comparison of the upper 95% confidence limit of the upper Bland-Altman limit of agreement as well as the lower 95% confidence limit of the lower Bland-Altman limit of agreement with a predefined limit for total error was proposed [10, 11]. Focusing on nonparametric prediction intervals instead, several approaches have been discussed earlier [12-15] including exact ones based on binomial distributions [16, 17].

Often, a priori information on agreement is scarce, impeding sample size planning. Instead of employing a fixed-sample design in which data on all patients is collected and first examined at the end of the study, we propose a group-sequential sample size strategy in which number and time points of interim analyses are predefined. By this means, early indications of both overly convincing agreement and clinically indefensible disagreement can be detected while the study runs, and stopping rules for fertility as well as futility can be implemented in the planning phase of the study. Sequentially testing in clinical trials with therapeutic intent is well established [18, 19], as is the handling of multiple endpoints and adaptive trial designs [20]. We are not aware of any previous application of sequential trial design methodology in agreement studies. In our setting, the experiment-wise type 1 error is ensured in the repeated hypothesis testing procedure by using type 1 error-spending functions. As starting point, we assume a Normal distribution for the difference of paired measurements and focus on its population standard deviation for testing purposes; thereby, we implicitly target the width of the Bland-Altman band. Subsequently, we move to nonparametric alternatives for which we need an analogue to the population standard deviation of the differences. We propose one half of the difference of the 84^{th} and 16^{th} percentiles as the 16-84 interquantile range (16-84 IQnR) covers, on average, 68% of all observations as does the mean ± one standard deviation in case of normally distributed data according to the 68-95-99.7 rule [21]. We derive one-sided, upper confidence intervals for ½ 16-84 IQnR and base nonparametric testing alternatives on these.

Regarding prediction intervals, we use, correspondingly, a nonparametric version of the Bland-Altman limits of agreement which are weighted functions of all data available instead of usually used simple empirical 2.5 and 97.5 percentiles. Our approach is exemplified with data from an ongoing diagnostic study in ovarian cancer in which pre-operative scans were assessed twice by the same clinician. Interim analyses were defined post hoc after 15 and 30 patients, final analysis was done with 45 patients.

Methods

Adaptive testing with α-spending functions

The spending function approach specifies a sequential design directly in terms of α_t where significance levels for interim and final analysis depend on the amount of hitherto accumulated information. We employed the α-spending function $\alpha_t = \alpha t$ [22] where α denotes the nominal experiment-wise significance level, t reflects the time that has passed by, and α_t represents the significance level to which the realized p-value is to be compared with at the respective analysis time point. We set the nominal experiment-wise level of the significance to 5%.

For instance, if one interim analysis is planned after inclusion of the first half of all patients to be recruited, t becomes 0.5 and $\alpha_{0.5}=0.025$ for $\alpha=0.05$; for the final analysis, $t=1$ and $\alpha_1=0.05$ hold. Generally, $\alpha_t \leq \alpha$ holds for all time points t in order to preserve the nominal significance level for the whole experiment. The sequential sample size strategy was exemplified post hoc with data from an ongoing clinical study (see below), including a total sample size of 45 and interim analyses with 15 and 30 patients, respectively. As a consequence, $\alpha_{0.33}=0.0167$, $\alpha_{0.67}=0.0333$, and $\alpha_1=0.05$ applied for analyses with 15, 30, and 45 patients, respectively.

Assuming normality

The sequential sample size strategy is based on a one-sided hypothesis test on the population standard deviation, σ, of the paired differences between measurements:

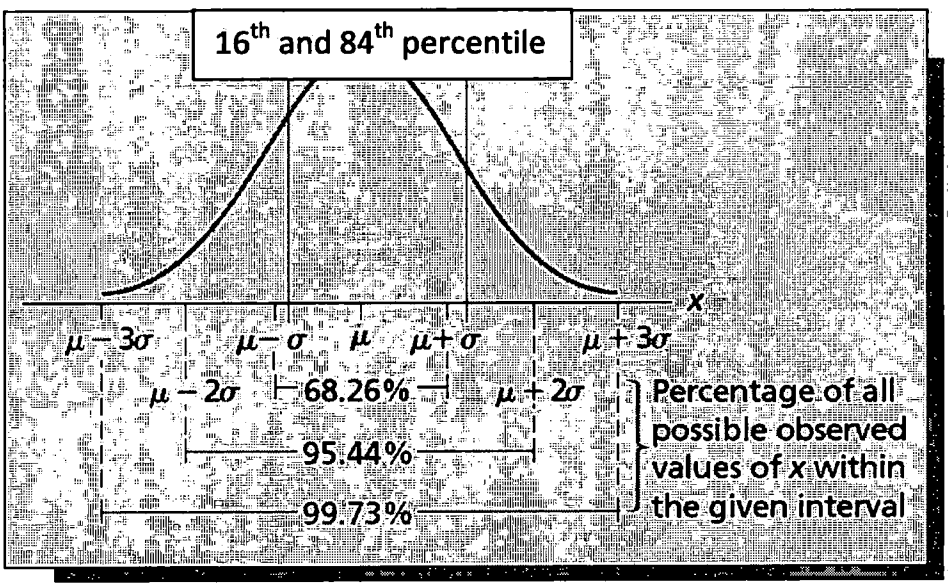
- Null hypothesis (H_0): $\sigma \geq \sigma_0$,
- Alternative hypothesis (H_a): $\sigma < \sigma_0$.

Assuming that the paired differences follow a Normal distribution, the test statistic follows a Chi-square distribution with N-1 degrees of freedom, with N denoting the number of paired measurements [21, 23]. We assessed the normality assumption visually by means of histograms for the differences including approximating Normal distributions.

A nonparametric alternative

Whenever the distribution of paired differences violates the assumption of a Normal distribution, an alternative testing approach to the above is needed; the more so as estimating second moments is more prone to uncertainty than estimating first moments. We propose a quasi-range as alternative to the population standard deviation which covers, on average, 68% of all observed differences as does the mean ± one standard deviation in case of normally distributed data according to the 68-95-99.7 rule [21]. Independently of the distributional form, the quasi-range defined by the 16^{th} and 84^{th} percentile has the abovementioned property; in case of normality the 16^{th} and 84^{th} percentile coincide with mean ± one standard deviation (Figure 1).

Figure 1: Graphical illustration of 68-95-99.7 rule [21]



Accordingly, one half of the 16-84 interquantile range (½ 16-84 IQnR) will be considered as nonparametric alternative to the population standard deviation. One-sided, upper confidence intervals for ½ 16-84 IQnR will be used in testing the set of hypotheses nonparametrically where the Greek letter ι (iota) indicates the hypothesized value:

- H_0: ½ 16-84 IQnR ≥ ι,
- H_a: ½ 16-84 IQnR < ι.

We chose four different approaches for constructing nonparametric, upper confidence limits for testing purposes:

1) An exact, distribution-free approach employing Binomial distributions [16, 17]. Given the number of observations, n, the difference of the ordered statistics $x_{(s)}$ and

$x_{(r)}$, is an upper confidence limit for ½ 16-84 IQnR, if and only if $B_n(s-1,0.84)-B_n(r-1,0.16)\geq 1-\alpha$ where $B_n(k,p)$ denotes the cumulated Binomial probability of observing up to k successes in n trials with a success probability of p.

2) The extended Harrell-Davis (HD) method estimating percentiles (and their differences) as weighted functions of all data available using Incomplete Beta distributions [15, 24, 25]. An approximate variance estimator is used [15].
3) Bootstrapping (with 3000 iterations) using
   a) a simple method based on sample quantiles (SQ) which uses weighted averages of neighbouring observations in the estimation of percentiles [26];
   b) the extended HD method [26] as in point 2).

   Upper one-sided confidence limits for ½ 16-84 IQnR are, then, respective percentiles of the empirical distributions of 3000 bootstrapped point estimates.

Clinical example

Data from an ongoing clinical study in the preoperative assessment of ovarian cancer was used which was described elsewhere [27]. In brief, this study's primary hypothesis is that dual time FDG-PET/CT (positron emission tomography/computed tomography using the glucose analogue ^{18}F-fluorodeoxyglucose as tracer) performed at 60 and 180 min after injection of tracer will increase the diagnostic accuracy of FDG-PET/CT (routinely performed at 60 min) in the preoperative assessment of resectability, provided optimal debulking is achievable. The target population consists of patients with suspicion of ovarian cancer in which the clinical suspicion of malignancy is based on initial physical (including pelvic) examination, blood tests including CA-125, and transvaginal ultrasound. Interim data from 45 patients scanned between 8 Jul 2013 and 6 Jul 2016 were used. The assessment of the FDG-PET/CT scans performed at 60 min was done twice by author MHV in order to investigate intra-observer repeatability of the post imaging process. The maximum standardised uptake value (SUVmax (g/ml)) was measured in the primary ovarian lesion when possible to identify; otherwise, the SUVmax in peritoneal carcinosis was used.

Software implementation

All analyses were performed by using Stata/MP 14.2 (StataCorp LP, College Station, Texas 77845 USA).

Results

Assuming Normality

We varied the level of the hypothesized population standard deviation, σ_0, from 0.5 and 2. At the first interim analysis (N=15), the null hypothesis could only be rejected for hypothesized population standard deviations of at least 1.5 after adjustment of multiple testing (Table 1, left panel). The corresponding upper 98.33% confidence limit of the population standard deviation was 1.495. At the second interim analysis (N=30), the null hypothesis could only be rejected for hypothesized population standard deviations of at least 1.75, with a corresponding upper 96.67% confidence limit of the population standard deviation of 1.653 (Table 1, middle panel). At the end of the study (N=45), the null hypothesis could only be rejected for hypothesized population standard deviations of at least 1.5 (Table 1, right panel). The corresponding upper 95% confidence limit of the population standard deviation was 1.288.

A nonparametric alternative

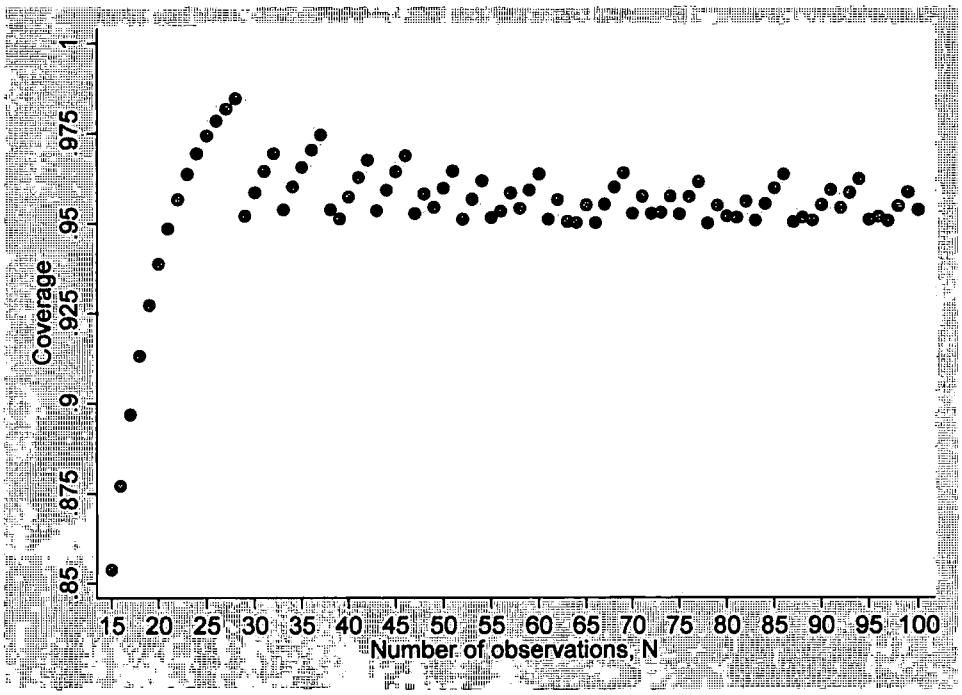
Exact upper one-sided confidence limits of ½ 16-84 IQnR exist only for samples of at least 22 paired observations when the significance level is 5% (Figure 2).

Table 1: Sequential testing on population standard deviation σ (assuming normality)

Hypothesized value σ_0	1st interim at N=15 (empirical SD: 0.909)		2nd interim at N=30 (empirical SD: 1.255)		Final analysis at N=45 (empirical SD: 1.060)	
	P-value	Upper one-sided 98.33% confidence limit of σ	P-value	Upper one-sided 96.67% confidence limit of σ	P-value	Upper one-sided 95% confidence limit of σ
0.5	1	1.495	1	1.653	1	1.288
0.75	0.89		1		1	
1	0.36		0.98		0.73	
1.25	0.082		0.55		0.08	
1.5	**0.016***		0.12		**0.0022***	
1.75	0.003		**0.014***		<0.0001	
2	0.0007		0.0014		<0.0001	

SD: standard deviation.

Figure 2: Coverage of exact, upper one-sided 95% confidence limits for 16-84 IQnR



Consequently, it did not exist at our 1st interim analysis (N=15) whereas the upper one-sided confidence limits at the 2nd interim at the final analysis exceeded those from the three other approaches by far (Table 2). At the 1st interim analysis, a bootstrapped upper-one-sided confidence limit of ½ 16-84 IQnR using the extended HD method was slightly smaller than the one derived from the extended HD method itself (1.363 vs. 1.389). At the 2nd interim and the final analysis, the extended HD method provided the smallest upper one-sided confidence limits (1.030 and 0.692, respectively) whereas bootstrapping using the SQ method led to comparably large upper one-sided confidence limits for all three analysis time points.

Table 2: Upper one-sided confidence limits of ½ 16-84 IQnR

Method for derivation of upper one-sided confidence limit	1st interim at N=15: upper one-sided 98.33% confidence limit of ½ 16-84 IQnR	2nd interim at N=30: upper one-sided 96.67% confidence limit of ½ 16-84 IQnR	Final analysis at N=45: upper one-sided 95% confidence limit of ½ 16-84 IQnR
Exact [16]	N/A[1)]	½ X_{30}-X_1=½ 6.7 = 3.35 [2)]	½ X_{43}-X_3 = ½ 2.2 = 1.1 [3)]
Extended HD method [15]	1.389	**1.030**	**0.692**
Bootstrapping: SQ method [26]	1.6	1.324	0.75
Bootstrapping: Extended HD method [26]	**1.363**	1.377	0.768

[1)] Coverage is 85.37% for r=1 and s=15. [2)] Coverage is 98.93% for r=1 and s=30. [3)] Coverage is 96.44% for r=3 and s=43.

Conclusion

Group-sequentially testing in agreement studies offers the necessary flexibility in the planning stage, but secures the experiment-wise significance level. It is adaptive in the sense that several interim analyses can (and actually should) be planned a priori in order to be able to implement early rules for stopping due to success or failure in indicating sufficient agreement. Though our example was restrospectively analysed, the potential is obvious. In the clinical development of therapeutic drugs, group-sequential testing is widely used (primarily in pivotal studies) while it has not yet reached the realm of diagnostic research in general and the assessment of agreement in particular.

References

1. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol. 2006;59(10):1033-9.
2. Barnhart HX, Haber MJ, Lin LI. An overview on assessing agreement with continuous measurements. J Biopharm Stat. 2007;17(4):529–69.
3. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420-8.
4. Dunn G. Statistical Evaluation of Measurement Errors. Design and Analysis of Reliability Studies. 2nd ed. Chichester: Wiley; 2004.
5. Shoukri MM. Measures of Interobserver Agreement and Reliability. 2nd ed. Boca Raton: Chapman & Hall; 2010.
6. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. Statistician 1983;32:307–17.
7. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307-10.
8. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res. 1999;8(2):135-60.
9. Carkeet A. Exact parametric confidence intervals for Bland-Altman limits of agreement. Optometry Vision Sci. 2015;92(3):e71-e80.
10. Stöckl D, Cabaleiro DR, van Uytfanghe K, Thienpont LM. Interpreting method comparison studies by use of the Bland–Altman plot: reflecting the importance of sample size by incorporating confidence limits and predefined error limits in the graphic. Clin Chem. 2004;50(11):2216-8.
11. Lytle FE, Julian RK, Tabert AM. Incurred sample reanalysis: enhancing the Bland-Altman approach with tolerance intervals. Bioanalysis. 2009 Jul;1(4):705-14.
12. Frey J. Data-driven nonparametric prediction intervals. J Stat Plan Infer. 2013;143:1039-48.
13. Zielinski R, Zielinski W. Best exact nonparametric confidence intervals for quantiles. Statistics. 2005;39(1):67-71.
14. Hall P, Rieck A. Improving coverage accuracy of nonparametric prediction intervals. J R Stat Soc B. 2001;63(4):717-25.
15. Steinberg SM. Confidence intervals for functions of quantiles using linear combinations of order statistics. Dissertation. University of North Carolina at Chapel Hill. Chapel Hill, North Carolina;1983.
16. Chu JT. Some uses of quasi-ranges. Ann Math Stat 1957;28(1): 173-180.

17. David HA. Order statistics. In: Wright JD (ed.). International Encyclopedia of the Social & Behavioral Sciences. 2nd ed. Vol. 17. Amsterdam: Elsevier;2015; pp.291-5.
18. Whitehead J. The Design and Analysis of Sequential Clinical Trials. 2nd ed. Chichester: Wiley;1997.
19. Moyé LA. Statistical Monitoring of Clinical Trials – Fundamentals for Investigators. New York: Springer;2006.
20. Dmitrienko A, Tamhane AC, Bretz F. Multiple Testing Problems in Pharmaceutical Statistics. Boca Raton: Chapman & Hall/CRC;2010.
21. Bowerman BL, O'Connell RT, Murphree ES. Business Statistics in Practice. 8th ed. New York: McGraw-Hill;2016.
22. Kim K, DeMets DL. Design and analysis of group sequential tests based on the type I error spending function. Biometrika 1987;74:149-54.
23. Rosner B. Fundamentals of Biostatistics. 8th ed. Boston: Cengage Learning;2015.
24. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes: The Art of Scientific Computing. 3rd ed. New York: Cambridge University Press;2007.
25. Abramowitz M, Stegun IA. Handbook of Mathematical Functions. New York: Dover;1968. Online at http://www.nr.com/aands, Chapters 6 and 26.
26. Steinberg SM, Davis CE. Comparison of nonparametric point estimators for interquantile differences in moderate sized samples. Commun Stat Theory. 1987;16(6): 1607-16.
27. Gerke O, Vilstrup MH, Segtnan EA, Halekoh U, Høilund-Carlsen PF. How to assess intra- and inter-observer agreement with quantitative PET using variance component analysis: a proposal for standardisation. BMC Med Imaging. 2016;16(1):54.

Vægtning for non-response i survey vha. egen oplyst uddannelse

Peter Linde, DST Survey, Danmarks Statistik

I en universel repræsentativ stikprøve ligger alle faktorer tæt på fordelingen i populationen. Dette kan kun sikres med en stikprøve udvalgt tilfældigt fra hele populationen, hvor der kun vil være en tilfældig stikprøve usikkerhed. Stikprøven vil for alle faktorer højst afvige fra populationen med det konfidensinterval, der er valgt, fx på 95 %. Hvor flere der vælges, hvor mindre bliver sikkerhedsintervallet. Hvis der udvælges proportional efter udvalgte faktorer, vil disse pr. definition ligge helt skarpt på fordelingen i populationen. Hvis der er stor grad af forklaret variation mht. mellem de faktorer, der vælges proportional efter, og andre faktorer, hvis de udvalgte faktorer i den proportionale udvælgelse, trække de andre faktorer med sig. Det vil blive vist med et eksempel, at de simple demografiske faktorer køn, alder og geografi, ikke trækker ret meget mht. uddannelse og andre socioøkonomiske variabler. Hvis man medtager indkomst, vil dette billede ændre sig signifikant.

I dataindsamlingen vil der være i en repræsentativ stikprøve være en non-response. Opnåelsen vil typisk ligge mellem 50 % og 60 % i en undersøgelse med mix-mode med web og telefon. Typisk vil høj social statustus bliver overrepræsenteret blandt de opnåede interviews. Det kunne her være oplagt at forsøge at veje for uddannelse. Man kan fx veje efter den faktiske uddannelse i registrene over den danske befolkning. Her skal i stikprøven have adgang til registeroplysningen for at lave en korrekt opvejning, men hvad gør man, hvis man har den egen oplyste uddannelse i spørgeskemaet, og ikke kender den faktiske fra registeret? Den egen oplyste uddannelse vil typisk overvurdere i forhold til den faktiske, hvilket giver opregningen en ny bias og ikke derfor ikke er en holdbar løsning på overrepræsentationen af høj socialt status. Hvor stor skævheden er, og hvordan man kan reducere biasen ved indirekte brug af registeroplysninger, vil blive vist.