politica
Tidsskrift for politisk videnskab

Tidsskriftet Politica udgives med støtte fra
Forskningsrådet for Samfund og Erhverv under Det Frie Forskningsråd
Omslagsdesign: Kasper Lægård, Avail Design
Grafisk tilrettelæggelse: One Hundred Proof
Tryk: Grafisk Produktion Odense
ISSN 0105-0710
ISBN 978-87-7335-179-6

Redaktionen sluttet den 7. januar 2014

Bestilling af tidsskriftet:
Syddansk Universitetsforlag
E-mail: press@forlag.sdu.dk
Telefon: 6615 7999
Girokonto: 5 04 51 93

Politica publicerer alene artikler bedømt ved peer review, enten i form af enkeltstående artikler bedømt ved mindst to anonyme reviewers eller som del af et temanummer bedømt ved mindst én anonym reviewer. Der accepteres manuskripter på dansk, norsk og svensk.
Se www.politica.dk for kontaktoplysninger og skrivevejledning.

Politica er indekseret i *International Political Science Abstracts,* som udgives af IPSA.

Tidsskriftet Politica, c/o Institut for Statskundskab
Bartholins Allé 7, 8000 Aarhus C

Indhold

Temaartikler

Anmeldelser

politica, 46. årg. nr. 1 2014, 5-23

Jens Blom-Hansen og Søren Serritzlew

Endogenitet og eksperimenter – forskningsdesignet som løsning

Statskundskaben er fyldt med teorier om virkningen af uafhængige på afhængige variabler. Desværre er det ofte svært at fastslå, om en empirisk sammenhæng er udtryk for en kausal effekt. Uanset om den er fundet i et kvantitativt eller kvalitativt studie, kan en empirisk sammenhæng dække over mere, end at *x* påvirker *y*. Sammenhængen kan også være udtryk for, at *y* påvirker *x*. Dermed opstår et endogenitetsproblem, som indebærer, at man ikke uden videre kan udlede kausal inferens fra datamaterialet. I den situation er det relevant at overveje eksperimentet som løsning. I artiklen diskuterer vi disse forhold og giver eksempler på de væsentligste eksperimentelle designs, nemlig laboratorieeksperimentet, det naturlige eksperiment, felteksperimentet, surveyeksperimentet og kvasieksperimentet.

Hvor er videnskaben i samfundsvidenskaben?

Statskundskaben er fyldt med teorier om virkningen af uafhængige på afhængige variabler. Ifølge King, Keohane og Verba (1994: 7-9) er det et definerende træk ved *videnskabelig* statskundskabsforskning, uanset om tilgangen er kvantitativ eller kvalitativ, at målet er deskriptiv eller kausal inferens. Desværre er det ofte vanskeligt at fastslå, om en empirisk sammenhæng faktisk er udtryk for en kausal effekt. Eksempelvis har en række studier undersøgt Ostroms (1996) teori om, at samproduktion, hvor både offentligt ansatte og borgere bidrager til produktionen af offentlige ydelser, kan føre til højere effektivitet. Men det er svært at afgøre, om sammenhængen mellem samproduktion og effektivitet skyldes, at samproduktion faktisk virker – eller blot er udtryk for, at offentligt ansatte inddrager borgere, når tingene i forvejen går godt (Jakobsen og Andersen, 2013). Et andet eksempel er spørgsmålet om sammenhængen mellem størrelse og demokrati. Finifter (1970) argumenterer for, at størrelsen på en politisk enhed har en negativ effekt på borgernes politiske effektivitetsfølelse. Problemet er, at det er svært at vide, om en negativ empirisk sammenhæng er udtryk for, at Finifter har ret – eller blot skyldes, at borgere med en høj politisk effektivitetsfølelse har bosat sig i mindre enheder (Lassen og Serritzlew, 2011: 241).

Fælles for de to eksempler er, at en empirisk sammenhæng, uanset om den er fundet i et kvantitativt eller kvalitativt studie, kan dække over mere, end at *x* påvirker *y*. Sammenhængen kan også skyldes, at *y* påvirker *x*. Eller begge

dele samtidig. Med andre ord kan vi ikke regne med, at den uafhængige variabel faktisk er helt uafhængig; den påvirkes af andre faktorer i modellen og er dermed endogen. Endogenitetsproblemet kan have andre kilder, men det indebærer under alle omstændigheder, at vi ikke uden videre kan udlede kausal inferens fra datamaterialet.

Her er det relevant at overveje eksperimentet som løsning. Det har, ligesom både statistiske og kvalitative tilgange, inferens som mål. Eksperimentets styrke er, at det kan løse endogenitetsproblemet. Et eksperiment sikrer nemlig, at værdierne på den uafhængige variabel er bestemt eksogent, dvs. af forhold, der intet har at gøre med den teoretiske sammenhæng, man er interesseret i. Eksperimenter kan dog være svære at gennemføre eller finde, og kan – når det lader sig gøre – komme til at handle om nogle ret specielle empiriske cases, hvorfor resultaterne ikke uden videre kan generaliseres. Eksperimentet har altså både fordele og ulemper: Den interne validitet er høj, men den eksterne validitet kan være lav. Med andre metoder vil det ofte forholde sig lige omvendt. Et veltilrettelagt kvantitativt studie har en høj ekstern validitet, men den interne validitet er – hvis der er endogenitetsproblemer – lav. Derfor supplerer eksperimentet de traditionelle metoder godt, og derfor er det en skam, at eksperimentet er en relativt overset metode.

I artiklen fokuserer vi på endogenitetsproblemet og diskuterer, hvordan eksperimentet kan løse det. Vi giver konkrete eksempler på de væsentligste eksperimentelle designs, nemlig laboratorieeksperimentet, det naturlige eksperiment, felteksperimentet, surveyeksperimentet og kvasieksperimentet.

Traditionelle metoders akilleshæl: endogenitet

Empiriske studier i statskundskaben er fyldt med påstande om kausalitet. Det er ikke så underligt. Statskundskabens teorier handler ofte om at forklare et fænomen (*y*) ved at henvise til, hvordan det bestemmes af et andet fænomen (*x*). Formålet med et empirisk studie vil typisk være at afklare, om det faktisk forholder sig sådan, at *x* forårsager *y*. Empiriske studier har imidlertid en akilleshæl, som alt for ofte overses. For at en empirisk sammenhæng kan siges at være udtryk for en *effekt* af *x* på *y*, må *x* være eksogen. Men det er jo ikke sikkert, at denne antagelse holder, og det viser sig desværre, at konsekvenserne er alvorlige, uanset om man forsøger at undersøge sammenhængen med kvalitative eller kvantitative teknikker (King, Keohane og Verba, 1994: 185ff.).

Endogenitet kan måske nemmest forklares med udgangspunkt i den klassiske regressionsmodel:

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad (1)$$

Endogenitetsproblemet opstår, hvis x er endogen, hvilket er tilfældet, hvis x korrelerer med e. Problemet kan opstå på flere måder, men først er det værd at minde om, at den klassiske regressionsanalyse simpelthen antager, at problemet ikke eksisterer (se eksempelvis Gujarati, 2003: 71). Antagelsen er ikke desto mindre afgørende. Endogenitetsproblemet kan have sin rod flere steder. Antonakis et al. (2010: 1090; se også Meyer, 1995) peger blandt andet på udeladte variabler, simultanitet, målefejl og common source bias. Pointen er, at mange forhold kan betyde, at x korrelerer med e, og at det derfor i ethvert empirisk studie kræver en del begrundelse, hvis man vil opretholde antagelsen om eksogenitet.

Lad os se nærmere på Antonakis et al.s (2010) første problem for at illustrere, hvad endogenitetsproblemet indebærer.[1] Antag, at vi estimerer y alene med x, men at y i virkeligheden er bestemt af både x og z, der korrelerer med hinanden, således at $z_i=\gamma_1 x_i+u_i$:

$$y_i=\beta_o+\beta_1 x_i+\beta_2 z_i+e_i \tag{2}$$

I stedet for at estimere (2) udelader vi en relevant variabel (nemlig z) og estimerer i stedet denne model:

$$y_i=\varphi_o+\varphi_1 x_i+v_i \tag{3}$$

Konsekvenserne er alvorlige! De kan ses ved at indsætte, at $z_i=\gamma_1 x_i+u_i$ i (2):

$$y_i=\beta_o+\beta_1 x_i+\beta_2(\gamma_1 x_i+u_i)+e_i=\beta_o+(\beta_1+\beta_2\gamma_1)x_i+\beta_2 u_i+e_i \tag{4}$$

Her er x åbenbart korreleret med fejlleddet (som er $\beta_2(\gamma_1 x_i+u_i)+e_i$). Hvis vi tror, at vi i vores regressionsanalyse med regressionskoefficienten φ_1 har fundet effekten af x på y, tager vi fejl. φ_1 er jo *ikke* lig det, vi er interesseret i, nemlig β_1. I stedet for at opnå et estimat af β_1, rammer vi helt skævt med et estimat på $\beta_1+\beta_2\gamma_1$. Vores estimat kan være 0, samtidig med at den sande effekt, β_1, er positiv eller negativ. Estimatet kan være positivt, mens den sande værdi er negativ. Det afhænger alene af, hvordan x hænger sammen med den udeladte variabel, z, og hvordan z påvirker y. Med andre ord: Bias fra en udeladt variabel skaber endogenitet, hvilket betyder, at vi systematisk fejlestimerer effekten af x. Heldigvis kan dette problem håndteres ved statistisk kontrol for z, forudsat naturligvis at z er observerbar. Det er derfor, at det er så centralt at kontrollere for relevante tredjevariabler.

En anden kilde til endogenitet er simultanitet, nemlig at x og y påvirker hinanden gensidigt. Simultanitet er et slemt problem i dobbelt forstand. For det første er det – som vi skal vende tilbage til i næste afsnit – et meget udbredt problem, særligt indenfor samfundsvidenskaberne. For det andet er det, som vi ser på i et senere afsnit, meget vanskeligt og ofte umuligt at løse med statistiske redskaber. Her skal vi blot se, at simultanitet også leder til bias i estimatet af effekten af x. Vi tager igen udgangspunkt i (1). Hvis x er eksogen, kan vi uden videre estimere β_1. Men hvis x og y påvirker hinanden gensidigt, er det mere kompliceret. I så fald afhænger y ikke blot af x, men x afhænger også af y:

$$x_i=\gamma_1 y_i+u_i \tag{5}$$

Estimeres β_1 på basis af (1), altså under antagelse af, at $\gamma_1=0$, når man, ligesom i eksemplet ovenfor, frem til et biased estimat af effekten af x (se King, Keohane og Verba, 1994: 195-196).

Diskussionen viser, at antagelsen om eksogenitet, der ligger bag traditionelle empiriske analyser, langtfra er uskyldig. Hvis antagelsen ikke er rigtig, kan vi simpelt hen ikke stole på de empiriske analyser. Problemet kan håndteres på to måder. For det første kan man forsøge at kompensere statistisk for problemet. Det kan, hvis der er tale om bias fra udeladte variabler, ske ved statistisk kontrol. Skyldes problemet simultanitet, er det vanskeligere men ikke umuligt. For det andet kan man forsøge at fjerne problemets årsag. Det kan ske ved, at man ved hjælp af eksperimentel metode sikrer, at x faktisk er eksogen.

Endogenitet som særligt problem for samfundsforskningen

Samfundsforskere er i høj grad ramt af problemet. Der findes nemlig aktører i samfundet, som så at sige har til opgave at skabe endogenitetsproblemer. Det gælder fx embedsmændene i den offentlige forvaltning. De er ansat til at løse problemer. Det giver en særlig aktualitet til et bestemt aspekt af endogenitetsproblemet, nemlig gensidig kausalitet (eller simultanitet). Embedsmændene reagerer på problemer og prøver at løse dem med midler, de tror virker. Så problemerne skaber løsninger, som igen påvirker problemerne. Et konkret eksempel kan være en embedsmand i en kommune med økonomiske problemer. Embedsmanden ved, at udlicitering muligvis kan være en løsning. Derfor prøver han at overtale politikere til at forbedre den økonomiske situation ved at udlicitere. Hvis han har ret, vil kommunens økonomi forbedre sig.

Mere generelt er problemer ofte anledningen til at igangsætte nye initiativer. Eksempelvis fik det udvalg, som regeringen i 2012 bad give kommunalreformen fra 2007 et serviceeftersyn, følgende mandat: ”Udvalget skal i dets

vurderinger lægge vægt på at understøtte en høj kvalitet i de offentlige ydelser, effektivitet, styring og tværgående prioritering, en klar ansvarsfordeling mellem myndigheder og nærhed for borgerne" (Økonomi- og Indenrigsministeriet, 2013: 5). Udvalget blev altså bedt om at pege på tiltag med bestemte effekter. Dermed skabes problemer med gensidig kausalitet for forskere, der efterfølgende vil efterprøve, om tiltagene virker.

En så tæt sammenhæng mellem problemer og tiltag er langtfra noget særsyn. Reformer er sjældent tilfældige, men er som regel en reaktion på organisatorisk performance. I den forstand er reformer endogene. Derfor er det problematisk at måle effekten af reformerne ved at se på korrelationen mellem dem og forskellige performanceindikatorer. For denne korrelation er også et produkt af den effekt, der er fra performance til reformerne. At finde eksogene reformer og organisatoriske tiltag udgør derfor en særlig udfordring.

Løsninger på endogenitetsproblemer

Vi kan altså ikke regne med, at den empiriske korrelation mellem x og y uden videre kan fortolkes som et estimat af den kausale effekt af x på y. I samfundsforskning er det endda sådan, at vi ofte kan være sikre på, at relationen mellem x og y er gensidig. Så snart der bare er en rimelig mistanke om, at x ikke blot påvirker y, men at x kunne være indført for at imødegå et problem med y, har vi at gøre med simultanitet. Man kan tænke sig to løsninger på problemet. Den ene er at tage højde for problemet ved hjælp af statistisk metode. Den anden er at tilrettelægge designet på en måde, der sikrer, at x er eksogen. Vi ser her nærmere på den mest kendte statistiske løsning på problemet. Derefter diskuterer vi designløsningen. I næste afsnit introducerer vi eksperimentet.

Hvis vi ikke kan udelukke, at y påvirker x, hjælper statistisk kontrol os ikke. Det lader sig jo ikke gøre at kontrollere for effekten af y. Vi har brug for at få greb om den del af den uafhængige variabel, som påvirker y, men som ikke skyldes y. Det kan lade sig gøre ved hjælp af instrumentvariabler (IV). Metoden er principielt simpel: Det gælder om at finde en variabel, z, som hænger (tæt) sammen med x, men som vi kan vide med sikkerhed ikke er påvirket af y – eller mere præcist ikke korrelerer med fejlleddet (se fx Angrist og Krueger, 2001).

Hariris (2012) undersøgelse af statsdannelsesprocessers betydning for udvikling af demokrati er et godt eksempel. Hariri argumenterer for, at tidlig statsdannelse hos ikke-europæiske lande har en negativ indflydelse på demokratiet, fordi tidligt udviklede stater i ringere grad er blevet påvirket af de europæiske kolonimagters parlamentariske system. Hypotesen undersøges i en almindelig regressionsanalyse af effekten på demokratiet af, hvor tidligt staten

i 111 ikke-europæiske lande er dannet. Som forventet finder Hariri en negativ sammenhæng. Jo tidligere statsdannelse, jo mindre demokrati i dag. Problemet er, at sammenhængen også kan være omvendt: Et lands styreform kan også tænkes at påvirke dets levedygtighed (Hariri, 2012: 480). Hariri løser problemet ved at identificere et fænomen, nemlig tidspunktet for landets overgang fra jæger- til jordbrugssamfundet, som korrelerer med tidspunktet for statsdannelsen (altså med den uafhængige variabel), men som næppe kan være påvirket af graden af demokrati i landet i dag (dvs. den afhængige variabel). Det er smart. Hvis det faktisk er rigtigt, at tidspunktet for overgangen hænger stærkt sammen med tidspunktet for statsdannelsen, kan variablen bruges som IV til at forudsige tidspunktet for statsdannelsen. Disse forudsagte værdier kan umuligt være påvirket af nutidens demokrati. Hvis den forudsagte værdi for statsdannelsen påvirker demokratiet i dag, er det et stærkt indicium på, at tidspunktet for statsdannelsen faktisk har den forventede negative kausale effekt.

Den statistiske tilgang til håndteringen af endogenitetsproblemet har dog sine begrænsninger. For det første er det ret sjældent, at det er muligt at identificere en god IV. For det andet findes der ikke nogen statistisk test, der kan sikre, at en given instrumentvariabel (z) ikke påvirkes, direkte eller indirekte, af den afhængige variabel.[2] Blot en beskeden sammenhæng med den afhængige variabel vil kunne lede til betydelig bias, særligt hvis korrelationen mellem IV og den uafhængige variabel er lav (Bound, Jaeger og Baker, 1995). Med andre ord: IV-tilgangen er kun lejlighedsvist tilgængelig. Selv når man kan identificere en velegnet IV, findes der ikke nogen metode, der kan afgøre, om problemet er løst fuldstændigt.

Den designbaserede løsnings logik er derimod at eliminere selve kilden til endogenitetsproblemet. Hvis man i dataindsamlingen sikrer sig, at den uafhængige variabel er eksogent bestemt, kan man helt udelukke, at en empirisk sammenhæng skyldes endogenitet (Dunning, 2012: 24-25). Vi ser nu nærmere på eksperimentet som en designbaseret løsning på endogenitetsproblemet.

Eksperimenter som løsning

Et eksperiment er en særlig form for test. Baggrunden er normalt, at man ønsker at undersøge effekten af et bestemt tiltag, fx en reform eller organisatorisk ændring. I det følgende betegnes dette med begrebet intervention. Videre indebærer et eksperiment normalt, at en gruppe (eksperimentgruppen) udsættes for interventionen, mens en anden gruppe (kontrolgruppen) ikke gør. Effekten af interventionen fastlægges herefter som udviklingen i eksperimentgruppen sammenlignet med kontrolgruppen. Begreber som intervention, eksperiment-

og kontrolgrupper samt sammenligning er altså centrale begreber i eksperimentel forskning.

Skal et eksperiment defineres mere præcist, bliver det vanskeligt, for der er ikke enighed i litteraturen. Morton og Williams (2010: 42) definerer et eksperiment ved det forhold, at forskeren styrer interventionen og aktivt manipulerer den. Dunning (2012: 15-16; se også McDermott, 2013: 608) tilføjer, at inddelingen i eksperiment- og kontrolgrupper skal være randomiseret. Sekhon og Titiunik (2012: 35) lægger også vægt på randomisering, hvorimod forskerens kontrol over interventionen ikke er central. Problemet med at opstille en præcis definition er, at der findes mange typer eksperimenter, som definitionen helst skal omfatte. Men ingen af de nævnte bud dækker alle former for eksperimenter. Vi foretrækker derfor at arbejde med en mindre præcis overordnet afgrænsning af eksperimentbegrebet og så mere præcist afgrænse de forskellige typer eksperimenter.

Som overordnet afgrænsning vender vi tilbage til Cook og Campbells (1979: 5) klassiske indføring i eksperimentel analyse. De definerer et eksperiment således: "All experiments involve at least a treatment, an outcome measure, units of assignment, and some comparison from which change can be inferred and hopefully attributed to the treatment". I definitionen indgår dermed en ekstern intervention, men ikke forskerens grad af kontrol herover. Inddeling i eksperiment- og kontrolgrupper indgår også i definitionen, men ikke hvorvidt denne er randomiseret. De nævnte forhold, der ikke indgår i definitionen, kan anvendes til at sondre mellem forskellige eksperimenter. Det gøres i tabel 1.

Eksperimenterne er kategoriseret efter fire kriterier. Det første er, om der indgår eksperiment- og kontrolgrupper. Som det fremgår, gælder det alle typer eksperimenter. Der er altså tale om et generisk træk ved eksperimenter. Til gengæld kan andre undersøgelsesdesign også arbejde med eksperiment- og kontrolgrupper som fx traditionelle store-N-undersøgelser af observationsdata. Kriteriet kan altså ikke diskriminere mellem eksperimentel forskning og andre former for forskning.

Det næste kriterium er, om den eksperimentelle intervention er eksogen. Det gælder også alle former for eksperimenter. Og det er efter dette kriterium, at eksperimenter adskiller sig fra andre undersøgelsesdesigns. I ikke-eksperimentelle designs er der ingen garanti for eksogenitet, hverken i traditionelle store-N-undersøgelser eller singlecasestudier. Endogenitet og eksogenitet er sjældent absolutte størrelser, men som regel tilstede i et vist omfang. Hvis forskeren kan argumentere for, at interventionen ikke er fuldstændig eksogen, men omvendt ikke påvirket af undersøgelsesobjekterne i en sådan grad, at det påvirker resultatet, kan man tale om as-if eksogenitet.

Det tredje kriterium er, om der er sket en randomiseret inddeling i eksperiment- og kontrolgrupper. Randomisering behøver ikke være foretaget af forskeren. Nogle gange kan naturlige begivenheder eller politiske beslutninger påvirke forskellige dele af befolkningen eller dele af landet på tilfældig måde. I så tilfælde taler man om as-if randomisering (Dunning, 2012: 9-10). Hvis der ikke er sket randomisering, kan forskelle i resultatet skyldes initiale forskelle i eksperiment- og kontrolgruppen, der ikke har med interventionen at gøre. I så tilfælde må forskeren tage højde for disse forskelle ved fortolkningen af resultatet. Ved randomiseret fordeling kan den kausale effekt af interventionen derimod udledes direkte af forskellen mellem eksperiment- og kontrolgruppen. Som det fremgår af tabel 1, er randomisering ikke et generisk træk ved eksperimenter, idet kvasieksperimenter ikke opererer med denne måde at opstille eksperiment- og kontrolgrupper.

Endelig er det fjerde kriterium, om forskeren styrer interventionen. Hvis det er tilfældet, kan den eksperimentelle variation justeres mere præcist efter forskningsspørgsmålet end i det modsatte tilfælde, hvor forskeren er nødsaget til at analysere effekten af den variation, andre har skabt. Igen er der ikke tale om et generisk træk ved eksperimenter, men om et forhold der karakteriserer bestemte eksperimenter, jf. tabel 1.

Alt i alt gælder, at den centrale forskel på eksperimentel og ikke-eksperimentel forskning ikke er et spørgsmål om, hvorvidt der anvendes eksperiment- og kontrolgrupper, hvorvidt inddelingen heri er randomiseret, eller hvorvidt forskeren kan manipulere den eksperimentelle intervention. Den afgørende skillelinje er spørgsmålet om interventionens eksogenitet. Alle eksperimenter har eksogene interventioner. Det er der ingen garanti for ved traditionelle forskningsdesigns. Derfor er alle eksperimenter velegnede til at løse endogenitetsproblemer. Dertil kommer, at flere eksperimentelle designs også kan håndtere en anden væsentlig kilde til endogenitet, nemlig udeladte variabler. I de fleste eksperimenter holdes tredjevariabler konstante. Derved kan man udelukke endogenitetsbias fra udeladte variabler.

Eksperimentelle designs

I det følgende diskuterer vi mere udførligt de fem typer eksperimenter, der er skitseret i tabel 1. For hver type giver vi en mere præcis definition, illustrerer med et konkret eksempel fra dansk forskning og diskuterer fordele og ulemper.

Laboratorieeksperimentet

Som det fremgår af tabel 1, er laboratorieeksperimentet[3] et rendyrket eksperimentelt design. Det indeholder en sammenligning af eksperiment- og kontrol-

Tabel 1: Forskellige typer eksperimenter

	Sammenligning af eksperiment- og kontrolgrupper?	Interventionen er eksogen eller as-if eksogen	Inddeling i eksperiment- og kontrolgrupper er randomiseret eller as-if randomiseret?	Forskeren manipulerer interventionen?
Laboratorieeksperiment	Ja	Ja	Ja	Ja
Felteksperiment	Ja	Ja	Ja	Ja
Surveyeksperiment	Ja	Ja	Ja	Ja
Naturligt eksperiment	Ja	Ja	Ja	Nej
Kvasieksperiment	Ja	Ja	Nej	Nej
Traditionelt stort-N observationsstudie	Ja	Nej	Nej	Nej
Traditionelt singlecasestudie	Nej	Nej	Nej	Nej

grupper, interventionen er ikke blot eksogen, den er også randomiseret, og det er forskeren, der manipulerer interventionen. Dertil kommer, at eksperimentet finder sted i kontrollerede omgivelser. Derved sikres det, at ingen uvedkommende faktorer påvirker eksperimentet. Naturvidenskabelige eksperimenter kan kræve ganske avancerede laboratorier for at sikre dette; i statskundskaben er det typisk tilstrækkeligt med en dør, der kan lukkes.

Et eksempel på et dansk laboratorieeksperiment er Serritzlew (2003; se også 2005). Forskningsspørgsmålet er her, hvordan rammebudgettering påvirker væksten i udgifterne. Hypotesen er, at rammebudgettering har en effekt på væksten, men at effekten er betinget af beslutningstagernes præferencer, således at væksten bestemmes af disse to uafhængige variabler og deres interaktion. Det er klart, at en traditionel empirisk undersøgelse her står overfor et betydeligt simultanitetsproblem. Det er meget muligt, at brugen af rammebudgettering påvirker udgiftsvæksten, men det er også nærliggende, at rammebudgettering bringes i anvendelse, hvis væksten i udgifterne er høj. En eventuel sammenhæng mellem anvendelse af rammebudgettering og udgiftsvækst kan derfor ikke uden videre fortolkes som en kausal effekt af budgetlægningsmetoden.

Laboratorieeksperimentet blev designet som et budgetspil med fem deltagere, der med simpelt flertal skulle beslutte sig for et budget. Fire interventioner sikrede variation i de to uafhængige variabler og mulighed for at teste interaktionseffekten. Eftersom intet andet varierede mellem de fire interventioner, kunne forskelle i udgifterne fortolkes som kausale effekter af de uafhængige variabler. Eksperimentet viste, at rammebudgettering faktisk har den forventede betingede effekt på udgifterne.

Laboratorieeksperimentet har sine særlige fordele. For det første tillader det forskeren at styre, hvilke faktorer der påvirker deltagerne i eksperimentet. Dermed er det muligt at holde tredjevariabler konstant. For det andet sikrer randomiseringen, at enhver forskel mellem eksperiment- og kontrolgrupper kan tilskrives en effekt af interventionen. For det tredje kan data (deltagernes adfærd, beslutninger, holdning mv.) registreres med meget stor præcision. For det fjerde er det muligt i laboratorieeksperimenter at *inducere* præferencer. Ved at knytte belønninger (typisk mindre pengebeløb) til deltagernes adfærd, kan forskeren meget præcist styre deltagernes præferencer. Det er yderst vanskeligt i andre designs, eksperimentelle eller ej, og derfor anvendes laboratorieeksperimenter i særlig høj grad i undersøgelser af teorier om koordination og forhandling (se fx Roth og Kagel, 1995). Den væsentligste ulempe er, at laboratorieeksperimentet typisk har lav ekstern validitet. Det skyldes, at selve undersøgelsen foregår i et laboratorium, der jo typisk ikke ligner virkeligheden særlig godt. Dertil kommer, at deltagerne sjældent er repræsentative. Begge dele gør, at resultaterne er

bedre egnede til at teste kausalhypoteser end til at generalisere til en bestemt population.

Felteksperimentet

Felteksperimentet[4] ligner i sin logik laboratorieeksperimentet. Der foregår en sammenligning af eksperiment- og kontrolgrupper, inddelingen heri er randomiseret, interventionen er eksogen, og forskeren manipulerer interventionen. Forskellen er, at felteksperimentet, som navnet antyder, foregår i felten, dvs. hos de individer, organisationer eller steder, som den teori, der undersøges, handler om (Davenport, Gerber og Green, 2010: 69-71). Hvis hypotesen fx handler om vælgerregistrerings betydning for valgdeltagelsen, skal et felteksperiment lave en intervention i vælgerregistreringen (som i Gosnell (1927), der ifølge Davenport, Gerber og Green (2010) gennemførte det første politologiske felteksperiment).

Et dansk eksempel er Jakobsen og Andersen (2013), der undersøger effekten af samproduktion med et felteksperiment. Formålet var at undersøge effekten af at inddrage forældre i sprogindlæringen blandt tosprogede børnehavebørn. Det er nærliggende at forvente, at forældre med mange ressourcer gerne vil inddrages i et sådant eksperiment. Det udgør et simultanitetsproblem, som kan løses med randomiseret intervention. Forskerne opdelte derfor børnehavebørnene i en eksperiment- og kontrolgruppe, hvor børnene i eksperimentgruppen fik en sprogkuffert med hjem. Sprogkufferten indeholdt materialer, hvormed forældrene kunne læse med børnene på modersmålet. Kontrolgruppen bestod af børn, der ikke modtog en sprogkuffert. Randomiseringen betyder, at forskelle i sprogkundskaber mellem kontrol- og eksperimentgruppen kan fortolkes som en effekt af kuffertindsatsen. Studiet viser, at denne form for samproduktion har de forventede positive effekter.

Felteksperimentet sikrer som andre eksperimentelle designs en høj intern validitet. Samtidig er den eksterne validitet ganske høj. Det skyldes, at felteksperimentet jo netop gennemføres i den virkelige verden. Dertil kommer, at felteksperimentet kan have en betydelig policy-relevans. Politologiske felteksperimenter muliggøres ofte af, at en offentlig myndighed overvejer en ændring i praksis. Hvis ændringen designes som et felteksperiment, vil studiet næppe kunne undgå at have praktisk relevans. Resultaterne vil netop afklare, hvilke resultater den gennemførte ændring medførte. Felteksperimentet har også en række ulemper. For det første kan det være ganske ressourcekrævende. For det andet kan mange væsentlige politologiske forskningsspørgsmål næppe belyses med felteksperimenter. Det gælder eksempelvis forhold, der er tæt reguleret (hvilket kan forhindre randomiseret intervention), eller som har en tæt politisk

bevågenhed (hvor et felteksperiment hurtigt kan blive kontroversielt). For det tredje kræver felteksperimenter ofte betydelig koordination med aktører udenfor forskningsverdenen. Sådanne aktører kan have hensyn at varetage, som gør det vanskeligt at opstille et godt forskningsdesign. For det fjerde medfører det forhold, at eksperimentet gennemføres i felten, et vist kontroltab. Det betyder, at det kan være vanskeligt at sikre, at eksperimentet gennemføres helt som designet. Fx er der fare for, at kommunikation mellem deltagere på tværs af kontrol- og eksperimentgruppe fører til, at de to grupper ikke klart kan adskilles (Sinclair, McConnell og Green, 2012). Endelig kan felteksperimenter, som påvirker borgere direkte, være etisk problematiske, særligt hvis de stiller nogle borgere dårligere end andre.

Surveyeksperimentet

Også surveyeksperimentet[5] er et rent eksperimentelt design i den forstand, at der findes eksperiment- og kontrolgrupper, eksogen intervention, randomiseret inddeling og forskermanipulation med interventionen. Surveyeksperimentet er karakteriseret ved, at eksperimentet foregår i surveyform, dvs. ved at gruppen af respondenter tilfældigt opdeles i grupper, der udsættes for forskellige versioner af et surveyspørgsmål. Interventionen ligger altså i forskellene i spørgsmålsformuleringerne. Takket være randomiseringen kan forskellene i gruppernes svar fortolkes som effekten af interventionen.

Petersen et al. (2011) har undersøgt, hvordan holdningsdannelsen påvirkes af *cues*, altså stikord om politikkens indhold og karakter. Hypotesen er, at borgernes holdninger til velfærdspolitik afhænger af cues om modtagerens "fortjenstfuldhed", således at borgerens holdning spontant påvirkes af, om modtageren er en type, der fortjener offentlig hjælp. Petersen et al. (2011) undersøger spørgsmålet ved tilfældigt at fordele respondenter i fire grupper, der hver modtager en særlig beskrivelse af en modtager af velfærdsydelser: En ung mand, en kvinde i 50'erne, en kvinde i 50'erne med en arbejdsskade og en ældre mand, der har været på arbejdsmarkedet hele sit liv. Respondenterne i de fire grupper stilles derefter det samme spørgsmål, nemlig om aktiveringskravene burde strammes. Det viser sig, at støtten til stramningen afhænger af, hvilken type modtager respondenten er blevet præsenteret for.

Surveyeksperimentet adskiller sig ved, at det er relativt billigt og nemt at gennemføre. Randomiseringen af surveyspørgsmål kan ofte ske automatisk og en enkelt survey kan uden problemer indeholde adskillige surveyeksperimenter. En anden fordel ved surveyeksperimentet er, at det tillader et stort N. Det muliggør undersøgelser af betingede effekter, som ellers sjældent er realistiske i eksperimentelle designs. Til gengæld er det en ulempe, at interventionen i

surveyeksperimenter er lavintensiv. Hvor påvirkningen af deltagerne kan være meget betydelig i både laboratorie- og felteksperimenter, er der grænser for, hvor markant deltagerne kan påvirkes alene med formuleringsforskelle. Det betyder, at man på baggrund af surveyeksperimenter, der ikke finder en effekt af en intervention, sjældent med sikkerhed kan udelukke, at der er en effekt. Det kunne jo være, at den manglende effekt blot skyldes en svag intervention.

Naturlige eksperimenter

Naturlige eksperimenter[6] har fået deres navn, fordi data stammer fra "naturligt" forekommende fænomener. I samfundsvidenskaberne er der dog som oftest tale om produkter af sociale eller politiske processer. Det definitorisk væsentlige er, at dataene kommer udefra, ikke fra manipulation fra forskerens side. Herved adskiller naturlige eksperimenter sig fx fra laboratorieeksperimenter. Men i lighed hermed indeholder naturlige eksperimenter randomiserede (eller as-if randomiserede) eksperiment- og kontrolgrupper. Endvidere deler naturlige eksperimenter også den egenskab med laboratorieeksperimenter, at den eksperimentelle intervention er eksogen (Dunning, 2012: 41-63).

Et eksempel på et studie baseret på et naturligt eksperiment er Lassens (2005) undersøgelse af Københavns Kommunes bydelsforsøg. Kommunen var inddelt i 15 bydele, og fire bydele blev udvalgt til forsøget. I perioden 1997-2001 administrerede demokratisk valgte bydelsråd herefter en række kommunale opgaver. I de øvrige 11 bydele stod kommunen på sædvanlig vis for de kommunale opgaver. Lassen bruger forsøget til at undersøge, om information påvirker stemmeadfærd. Han udnytter her, at kommunen i 2000 afholdt en lokal folkeafstemning i hele kommunen om, hvorvidt bydelsforsøget skulle udvides til alle bydele eller helt opgives. Efter folkeafstemningen blev et repræsentativt udsnit af vælgerne spurgt, om de deltog i afstemningen. Spørgsmålet er så, om mere informerede vælgere i højere grad deltog i afstemningen. Lassens pointe er, at bydelsforsøget udgør en eksogen påvirkning af nogle vælgeres informationsniveau, nemlig beboerne i de fire udvalgte bydele, som ved mere om bydelsforsøget end de øvrige indbyggere i kommunen. Endvidere udnytter han, at informationen er nogenlunde tilfældigt fordelt blandt vælgerne, idet de fire bydele blev udvalgt, fordi de var repræsentative for hele kommunen. Alt i alt undersøger Lassen på denne måde et eksogent og as-if randomiseret stød til visse vælgeres informationsniveau. Studiet, som viser, at information højner valgdeltagelse, er et naturligt eksperiment, fordi den eksperimentelle intervention ikke er kontrolleret fra Lassens side men skabt udefra af det politiske system.

Sammenlignet med traditionelle observationsstudier, store-N-undersøgelser og singlecasestudier har naturlige eksperimenter en bedre håndtering af endogenitetsproblemet, idet den eksperimentelle variation er eksternt betinget. De har også en bedre håndtering af kontrolproblemet, idet inddelingen i eksperiment- og kontrolgrupper er randomiseret eller as-if randomiseret. Sammenlignet med andre eksperimentelle designs har naturlige eksperimenter den fordel, at de kan belyse effekten af forhold, der er svære at manipulere af forskeren selv. Hvad betyder fx sandsynligheden for militærtjeneste for mænds politiske holdninger? Politiets patruljering for kriminalitet? Valgovervågning for valgsnyd? Det er eksempler på spørgsmål, der er belyst med naturlige eksperimenter (Erikson og Stoker, 2011; Di Tella og Shargrodsky, 2004; Hyde, 2007), og som er svære at belyse i laboratorieeksperimenter.

Ulempen ved naturlige eksperimenter er, at forskeren er afhængig af udefra kommende forhold for få adgang til data. Det indebærer et vist tilfældighedselement i, hvilke spørgsmål der kan belyses. Det betyder også, at forskeren er nødsaget til at undersøge betydningen af den variation, som det naturlige eksperiment tilbyder. Det er sjældent, at den eksperimentelle variation kan gradbøjes præcist efter forskerens teoretiske spørgsmål.

Kvasieksperimenter

Kvasieksperimenter[7] ligner andre eksperimenter på den måde, at der findes eksperiment- og kontrolgrupper og en eksogen eksperimentel intervention. Endvidere har kvasieksperimenter det til fælles med naturlige eksperimenter, at den eksperimentelle intervention kommer udefra. Den er ikke manipuleret af forskeren, men tilvejebragt af naturen eller det politiske system. Men til forskel fra andre eksperimenter er inddelingen i eksperiment- og kontrolgrupper ikke randomiseret. Grupperne kan derfor være forskellige på mange andre parametre end deres udsættelse for den eksperimentelle intervention. Udfordringen ved kvasieksperimenter er derfor at fortolke, i hvilket omfang forskellen i udfaldet mellem grupperne skyldes initiale forskelle eller den eksperimentelle intervention (Cook og Campbell 1979: 6).

Et eksempel på et kvasieksperiment er Blom-Hansen, Houlberg og Serritzlews (under udgivelse) studie af administrative stordriftsfordele i kommunerne. Udgangspunktet er den danske kommunalreform i 2007, som indebar, at nogle kommuner blev lagt sammen, mens andre fortsatte uændret. De argumenterer for, at ændringen i sammenlægningskommunernes størrelse primært er eksternt bestemt, idet den ikke var et resultat af lokale forhold, men af regeringens reform. De har dermed en as-if eksogen intervention, en kontrol- og en eksperimentgruppe samt et mål for effektivitet både før og efter sammenlæg-

ningerne. Studiet er et kvasieksperiment, fordi kommunernes inddeling i eksperiment- og kontrolgrupper ikke er randomiseret. Hvorvidt en kommune blev lagt sammen eller fortsatte uændret var ikke tilfældigt. Fx havde regeringen ingen intention om at lægge de største kommuner sammen med andre. Derfor kan det ikke udelukkes, at forskelle i udfaldet i et vist omfang skyldes initiale forskelle snarere end interventionen. Blom-Hansen, Houlberg og Serritzlew kontrollerer derfor for en række strukturelle og økonomiske forskelle mellem kommunerne, inden de konkluderer, om sammenlægningerne har haft nogen effekt. Studiet viser, at store kommuner er betydeligt billigere i drift end små.

Sammenlignet med traditionelle observationsstudier har kvasieksperimenter en bedre håndtering af endogenitetsproblemet, idet den eksperimentelle variation er eksternt betinget. Men inden for gruppen af eksperimentelle designs er kvasieksperimentet videnskabeligt set det svageste, fordi der ikke sker en randomiseret inddeling i eksperiment- og kontrolgrupper. I modsætning til andre eksperimenter er kvasieksperimenter derfor nødt til at tage højde for initiale forskelle mellem eksperiment- og kontrolgruppen. Det er en betydelig udfordring, for man kan sjældent udelukke, at der selv efter omhyggelig indsamling af kontrolvariabler resterer ikke-observerede forskelle mellem grupperne.

Alligevel har kvasieksperimenter et stort potentiale. Videnskabeligt set er de stærkere end traditionelle observationsstudier, og uden for laboratoriet er kvasieksperimentets akilleshæl – randomiseringen – ofte et ideal, der er svært at opnå i praksis. Det er der mange grunde til (Cook og Campbell, 1979: 347-371; Green, 2010; Sinclair, McConnell og Green, 2012). Det kan eksempelvis være svært at fastholde, at kontrolgruppen ikke modtager interventionen, randomiseringsprocedurer kan være fejlbehæftede, deltagere i eksperimentgruppen kan nægte at lade sig udsætte for interventionen, eller der kan ske kommunikation mellem kontrol- og eksperimentgruppen. Samtidig gælder, at kvasieksperimenter ligesom naturlige eksperimenter kan belyse effekten af forhold, der er svære at manipulere af forskeren selv.

Laboratorieeksperimentet med randomiseret inddeling i eksperiment- og kontrolgruppe og med fuld forskerkontrol over den eksperimentelle intervention betragtes ofte som "guldstandarden" inden for videnskaben. Andre designs – selv andre eksperimentelle designs – kan højst udgøre en "sølvstandard". Vores diskussion af fire andre eksperimentelle designs (felteksperimenter, surveyeksperimenter, naturlige eksperimenter og kvasieksperimenter) giver et mere nuanceret billede. Alle disse eksperimentelle designs har særskilte styrker og svagheder og dermed særskilte eksistensberettigelser. Den vigtigste er, at de giver mulighed for at belyse forskningsspørgsmål, som kun meget vanskeligt lader sig underkaste laboratorieforsøg. Af samme grund hilser vi det velkom-

ment, at der ikke kun sker en stigende anvendelse af eksperimenter i samfundsforskningen, men også indhøstes erfaringer med eksperimentelle blandingsdesigns såsom *lab-in-the-field*-eksperimenter, hvor forskeren flytter laboratoriet ud i eksperimentgruppens naturlige miljø og dermed prøver at indhøste fordelene ved både laboratorie- og felteksperimentet (Morton og Williams, 2010: 296-301), og internetbaserede eksperimenter, hvor forskeren prøver at kombinere laboratorie- og surveyeksperimenter (Eckel og Wilson, 2006).

Konklusion

Traditionelt fremhæves, at eksperimenters styrke ligger i muligheden for at påvise kausalsammenhænge. Selvom to variabler samvarierer, behøver sammenhængen ikke være kausal. Med deres bedre greb om kontrol- og endogenitetsproblemet er eksperimenter bedre i stand til at belyse dette spørgsmål. Til gengæld er der ofte noget kunstigt over eksperimenter, hvorfor man ikke altid uden videre kan overføre resultaterne fra den stiliserede eksperimentsituation til virkelighedens mangefacetterede verden.

Sagen kan også fremstilles som et spørgsmål om forholdet mellem intern og ekstern validitet (Cook og Campbell, 1979: 37-94; Serritzlew, 2007; Morton og Williams, 2010: 253-277). Intern validitet handler om, i hvilket omfang forskeren kan fastslå, om sammenhængen mellem to faktorer er kausal. Den eksterne validitet handler om, i hvilket omfang resultaterne har gyldighed ud over den kontekst, hvori undersøgelsen er foretaget. Efter denne målestok ligger eksperimenters styrke i den interne validitet, mens ikke-eksperimentelle metoder kan have højere ekstern validitet. Hermed mener vi også at have sagt noget om eksperimenters nødvendighed i samfundsvidenskabelig forskning. Givet eksperimenters begrænsede udbredelse kan forskningen siges i praksis at have privilegeret den eksterne validitet på bekostning af den interne. Efter vores opfattelse er dette valg svært at retfærdiggøre. En mere balanceret hensyntagen til intern og ekstern validitet kræver mere plads til eksperimenter. Dette gælder ikke mindst, fordi kritikken af manglende eksterne validitet i eksperimentel forskning mest er rettet mod laboratorieforsøg. Andre eksperimentelle design, måske især felteksperimenter og naturlige eksperimenter, har ret høj ekstern validitet. Det er alt i alt svært at finde på gode argumenter for at bremse udbredelsen af eksperimentel statskundskabsforskning.

Noter

1. Den følgende formelle fremstilling bygger på Antonakis et al. (2010: 1091).
2. Kriteriet er, om z er korreleret med fejlleddet, og da fejlleddet er uobserveret, kan vi ikke kende korrelationen (Wooldridge, 2013: 492).

3. Se introduktion i Iyengar (2011) og på dansk Serritzlew (2007).
4. Davenport, Gerber og Green (2010) og Gerber (2011) er begge gode introduktioner.
5. Sniderman (2011) introducerer surveyeksperimentet. Politica 39 (1), 2007 indeholder en kort introduktion til surveyeksperimentet og gode eksempler på konkrete studier.
6. Dunning (2012) er en god introduktion.
7. Cook og Campbell (1979) giver en klassisk introduktion.

Litteratur

Angrist, Joshua og Alan B. Krueger (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives* 15 (4): 69-85.

Antonakis, John, Samuel Bendahan, Philippe Jacquart og Rafael Lalive (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly* 21 (6): 1086-1120.

Blom-Hansen, Jens, Kurt Houlberg og Søren Serritzlew (under udgivelse). Size, democracy, and the economic costs of running the political system. *American Journal of Political Science*.

Bound, John, David A. Jaeger og Regina M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90 (430): 443-450.

Cook, Thomas D. og Donald T. Campbell (1979). *Quasi-Experimentation. Design & Analysis Issues for Field Settings*. Boston: Houghton Mifflin.

Davenport, Tiffany C., Alan S. Gerber og Donald P. Green (2010). Field experiments and the study of political behavior, pp. 69-88 i Jan E. Leighley (red.), T*he Oxford Handbook of American Elections and Political Behavior*. Oxford: Oxford University Press.

Di Tella, Rafael og Ernesto Schargrodsky (2004). Do police reduce crime? Estimates using the allocation of police forces after a terrorist attack. *American Economic Review* 94: 115-133.

Dunning, Thad (2012). *Natural Experiments in the Social Sciences. A Design-Based Approach*. Cambridge: Cambridge University Press.

Eckel, Cathrine C. og Rick K. Wilson (2006). Internet cautions: Experimental games with internet partners. *Experimental Economics* 9: 53-66.

Erikson, Robert og Laura Stoker (2011). Caught in the draft: The effects of Vietnam draft lottery status on political attitudes. *American Political Science Review* 105: 221-237.

Finifter, Ada W. (1970). Dimensions of political alienation. *American Political Science Review* 64 (2): 389-410.

Gerber, Alan S. (2011). Field experiments in political science, kapitel 9 i James N. Druckman, Donald P. Green, James H. Kuklinski og Arthur Lupia (red.), *Cambridge Handbook of Experimental Political Science*. Cambridge: Cambridge University Press.

Gosnell, Harold F. (1927). *Getting-Out-the-Vote: An Experiment in the Simulation of Voting*. Chicago: Chicago University Press.

Green, Jane (2010). Points of intersection between randomized experiments and quasi-experiments. *Annals of the American Academy of Political and Social Sciences* 628: 97-111.

Gujarati, Damodar N. (2003). *Basic Econometrics*. Boston: McGraw Hill.

Hariri, Jacob Gerner (2012). The autocratic legacy of early statehood. *American Political Science Review* 106 (3): 471-494.

Hyde, Susan D. (2010). The observer effect in international politics: Evidence from a natural experiment. *World Politics* 60: 37-63.

Iyengar, Shanto (2011). Laboratory experiments in political science, kapitel 6 i James N. Druckman, Donald P. Green, James H. Kuklinski og Arthur Lupia (red.), *Cambridge Handbook of Experimental Political Science*. Cambridge: Cambridge University Press.

Jakobsen, Morten og Simon Calmar Andersen (2013). Coproduction and equity in public service delivery. *Public Administration Review* 73 (5): 704-713.

Kagel, John H. og Alvin E. Roth (1995). *Handbook of Experimental Economics*. Princeton: Princeton University Press.

King, Gary, Robert O. Keohane og Sidney Verba (1994). *Designing Social Inquiry*. Princeton: Princeton University Press.

Lassen, David Dreyer (2005). The effect of information on voter turnout: Evidence from a natural experiment. *American Journal of Political Science* 49: 103-118.

Lassen, David Dreyer og Søren Serritzlew (2011). Jurisdiction size and local democracy: Evidence on internal political efficacy from large-scale municipal reform. *American Political Science Review* 105 (2): 238-259.

McDermott, Rose (2013). The ten commandments of experiments. *PS: Political Science and Politics* 46 (3): 605-611.

Meyer, Bruce D. (1995). Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics* 13 (2): 151-161.

Morton, Rebecca og Kenneth C. Williams (2010). *Experimental Political Science and the Study of Causality. From Nature to the Lab*. Cambridge: Cambridge University Press.

Ostrom, Elinor (1996). Crossing the great divide: Coproduction, synergy, and development. *World Development* 24 (6): 1073-1087.
Petersen, Michael Bang, Rune Slothuus, Rune Stubager og Lise Togeby (2011). Deservingness versus values in public opinion on welfare: The automaticity of the deservingness heuristic. *European Journal of Political Research* 50: 24-52.
Sekhon, Jasjeet S. og Rocio Titiunik (2012). When Natural Experiments are Neither Natural nor Experiments. *American Political Science Review* 106: 35-57.
Serritzlew, Søren (2003). Kan udgiftsrammer begrænse væksten i budgetterne? *Politica* 35 (3): 255-273.
Serritzlew, Søren (2005). The perverse effect of spending caps. *Journal of Theoretical Politics* 17 (1): 75-105.
Serritzlew, Søren (2007). Det politologiske eksperiment. *Politica* 39 (3): 275-294.
Sinclair, Betsy, Margaret McConnell og Donald P. Green. (2012). Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science* 56: 1055-1069.
Sniderman, Paul M. (2011). The logic and design of the survey experiment: An autobiography of a methodological innovation, kapitel 8 i James N. Druckman, Donald P. Green, James H. Kuklinski og Arthur Lupia (red.), *Cambridge Handbook of Experimental Political Science*. Cambridge: Cambridge University Press.
Wooldridge, Jeffrey M. (2013). *Introductory Econometrics. A Modern Approach*. South-Western Cengage Learning.
Økonomi- og Indenrigsministeriet (2013). *Evaluering af kommunalreformen. Afrapportering fra udvalget om evaluering af kommunalreformen*. København.

politica, 46. årg. nr. 1 2014, 24-41

Derek Beach

Process tracing og studiet af kausale mekanismer

Denne artikel argumenterer for, at der kan være store gevinster ved at tage studiet af, hvad der sker imellem X og Y, mere seriøst. Det kan gøres ved at studere de kausalmekanismer, som binder X og Y sammen ved brug af dybdegående casestudiemetoder som process tracing (PT). Herved undersøges det empirisk, hvorvidt og hvordan X medvirker til at producere Y. Der er tre særlige fordele ved PT. For det første kan vi lave en meget stærk inferens om, at X er kausalrelateret til Y, fordi vi opnår detaljeret viden om den empiriske proces, som binder de to sammen i en mekanisme. For det andet opnår man en bedre forståelse for, hvordan X medvirker til at producere Y. En sidste overset fordel ved at studere mekanismer med dybdegående casestudier er, at man ikke behøver at vælge cases på baggrund af behovet for at isolere effekten af X i forhold til andre mulige årsager. Ulempen er særligt, at PT er utrolig tids- og pladskrævende, samt at mulighederne for at generalisere er begrænsede.

Man kan ikke drage inferens om, at X er årsag til Y, uden at man i det mindste kan udvikle en plausibel teoretisk forklaring på, hvorfor X kan være årsag til Y. I næsten alle ikke-eksperimentelle, regressionsbaserede analyser diskuteres der i et teoriafsnit, hvilke faktorer der binder X og Y sammen i en årsagsrelation. Men denne proces mellem X og Y undersøges som regel ikke, og hvis den gør, er det igennem analyse af en eller to intervenerende variabler.

Denne artikel argumenterer for, at der kan være store gevinster ved at tage studiet af, hvad der sker imellem X og Y, mere seriøst. Det kan gøres ved at studere de kausalmekanismer, som binder X og Y sammen ved brug af dybdegående casestudiemetoder som process tracing (PT).[1] Herved undersøges det empirisk, hvorvidt og hvordan X medvirker til at producere Y.[2]

Der er tre særlige fordele ved PT. For det første kan vi lave en meget stærk inferens om, at X er kausalrelateret til Y, fordi vi opnår detaljeret viden om den empiriske proces, som binder de to sammen i en mekanisme. Dvs. at vi kan udelukke spuriøsitet i forholdet mellem X og Y med større sikkerhed, end vi kan i store-n ikke-eksperimentelle design, fordi vi er tættere på processen i vores casestudier. For det andet opnår man en bedre forståelse for, hvordan X medvirker til at producere Y. Ofte finder man problemer med den teoretiserede mekanisme, når man undersøger den empirisk, hvilket fører til en teoretisk revision af mekanismen, der binder dem i lyset af de empiriske fund. Resulta-

tet er bedre teorier efter denne "frem og tilbage-proces", hvor teorien er i tæt dialog med empirien. En sidste overset fordel ved at studere mekanismer med dybdegående casestudier er, at man ikke behøver at vælge cases på baggrund af behovet for at isolere effekten af X i forhold til andre mulige årsager. Denne isolation af X's effekter foretages ved, at man operationaliserer de observerbare fingeraftryk, som en bestemt mekanisme forventes at have, på en måde så man maksimerer deres "unikhed".

Men specialisering har omkostninger. I PT kommer de fordele, metoden har for analysen af processen, med en forholdsvis høj pris. Ud over de praktiske udfordringer, mht. hvor meget tid og plads det kræver at lave ordentlige PT-casestudier, er der flere ulemper, der ikke er alment anerkendt i litteraturen. For det første: Givet at man arbejder med deterministiske teorier i "mekanisme som system-forståelse" (se næste afsnit), og det empiriske fokus på, hvad der sker ind imellem X og Y, i stedet for forholdet mellem X og Y, kan der være store udfordringer i indlejring af PT-casestudier i et bredere multimetodedesign – særligt når der bruges store-n regressionsbaserede metoder. For det andet: Selv hvis indlejring er mulig, kan man strengt taget kun lave inferens om mekanismer mellem X og Y i den case, man har undersøgt. Som påpeget af kritikere af mekanismer (fx Gerring, 2010), kan der være mange forskellige veje mellem det samme X og Y – hvad man kan kalde ækvifinalitet på mekanismeniveau. Problemet kan kun løses igennem en kombination af PT-studier af andre cases for at undersøge, om den samme mekanisme også binder X og Y sammen der, og en samtidig komparativ analyse af populationen for at undersøge, om der er så forskellige randbetingelser (*scope conditions*) mellem casene, eller om vi kan antage en vis kausalhomogenitet. Og til sidst: Hvis man kan overkomme de to første udfordringer, kan man stadig kun lave inferens til en begrænset population, hvor både X, Y og de relevante randbetingelser er til stede.

Artiklen starter med en kort introduktion til debatten om, hvad kausalmekanismer er. Hovedvægten vil dog være på, hvilke implikationer fokusset på mekanismer har for vores forskningsdesign. Tredje afsnit viser PT's analytiske styrker i forhold til studiet af mekanismer. Artiklen slutter med at diskutere tre udfordringer og mulige løsninger på dem.

Mekanismer som systemer

Begrebet kausalmekanisme er meget omdiskuteret i litteraturen. Det følgende vil kort diskutere de to mest almindelige forståelser efterfulgt af en præsentation af flere oversete implikationer, som "mekanisme som system" har.

Mest udbredt er en minimalistisk definition, hvor en mekanisme bare ses som en (eller flere) intervenerende variabler mellem X og Y, forstået som

X→M→Y (Falletti og Lynch, 2009: 1146; Gerring, 2007; King, Keohane og Verba, 1994: 87). M kan variere og undersøges derfor med metoder, som er egnede til studiet af variation, fx regressionsbaserede analyser, eller King, Keohane og Verbas (KKV's) tilpasning af denne logik til casestudier.

Problemet med denne definition er, at den bortdefinerer de metodiske gevinster, vi gerne vil opnå med studiet af mekanismer. Bunge siger ligefrem, at hvis vi forstår mekanismer som intervenerende variabler, er det, som vi gerne vil undersøge – det, som sker mellem X og Y – gemt væk i en "grå boks" (Bunge, 1997). Hvis M er en variabel, vil designet til at undersøge dette være, at vi maksimerer samvariationen mellem værdier af X og M og af M og Y samt den tidsmæssige rækkefølge – hvilket bedst gøres ved store-n undersøgelser i en slags tidsserieanalyse. Hvis vi fx vil undersøge forholdet mellem økonomisk udvikling (X) → demokratisering (Y), går en del teorier ud på, at den mekanisme (M), som forbinder X med Y, er fremvæksten af en middelklasse og de effekter, denne nye klasse har i det politiske system (fx Lipset, 1959; Huntington, 1991; Glassman, 1997). Hvis mekanismer bare forstås som intervenerende variabler, vil designet, der undersøger dette, kunne være at analysere graden af samvariation mellem X, M og Y ved at kigge på, hvordan middelklassen har stemt ved forskellige valg, eller bare om der er en tidsmæssig sammenhæng mellem fremvæksten af en middelklasse og stigning i graden af demokrati i et land (fx Shin, 1999; Chen og Lu, 2011). Men dette design undersøger ikke den kausale proces, hvormed fremvæksten af middelklassen har medvirket til at producere demokratisering. Og det er lige præcis det, som er gevinsten ved at undersøge kausale mekanismer.

Mange forskere mener, at mekanismer er meget mere end bare intervenerende variabler, men bedre forstås som *systemer*, der forbinder årsag med virkning igennem transmissionen af "kausale kræfter" (fx Bunge, 1997; Glennan, 1996; Mahoney, 2001; Mayntz, 2004; Waldner, 2012). Deres argument er, at ved kun at studere samvariation mellem X, M og Y bliver vi ikke særlig meget klogere på, *hvordan* X har produceret Y.

Hvis vi vil høste den metodiske værdi tilført af studiet af mekanismer, er det nødvendigt at tage mekanismer alvorligt. Dette gøres ved at bruge, hvad der kan betegnes som en "system"-definition,[3] hvor en mekanisme forstås som et *system* bestående af dele, hvor hver del består af *enheder*, som laver *aktiviteter*, der tilsammen *transmitterer* kausale "kræfter" fra X til Y (Glennan, 1996; Machamer, Darden og Craver, 2000; Machamer, 2004). En mekanisme kan sammenlignes med en maskine, hvor energi er transmitteret igennem tandhjul eller lignende enheder for til sidst at producere den ønskede effekt (Hernes, 1998: 78). Naturligvis har de fleste sociale mekanismer ikke en mekanisk, ma-

skinagtig karakter, men analogien hjælper os med at forstå, hvad det er, vi er interesserede i at undersøge empirisk, når vi snakker om mekanismer. Vores analytiske fokus er på *aktiviteter* og på, hvordan de transmitterer kausale kræfter, hvilket gør os klogere på, hvordan X har medvirket til at producere Y i den undersøgte case.

En mekanisme mellem X og Y behøver hverken at være nødvendig eller tilstrækkelig for at producere Y. Der er mange i litteraturen, der antager, at X + M er tilstrækkelig for at producere Y (Mahoney, 2001: 580; Mayntz, 2004: 241. 253; Anderson, 2012: 416; Waskan, 2011: 403). Men der er ingen logiske grunde til at kræve, at X + M er tilstrækkelig (Hedström og Ylikoski, 2010), og hvis vi gør det, vil vi reducere anvendelsen af mekanismer til de få teorier, hvor et enkelt X påstås at være tilstrækkeligt til at producere Y (fx i den demokratiske fredstese, hvor tilstedeværelse af demokrati mellem to lande er tilstrækkelig til at producere fred). De eneste to ting, der burde kræves, er, 1) at en mekanisme overfører visse "kausale kræfter" mellem X og Y, som burde være empirisk observerbare, og 2) at X er nødvendig, for at M kan påbegyndes.

For at en mekanisme fungerer, skal de fornødne randbetingelser (scope conditions) også være til stede (Falletti og Lynch, 2009). Disse defineres som: "relevant aspects of a setting (analytical, temporal, spatial, or institutional) in which a set of initial conditions leads … to an outcome of a defined scope and meaning via a specified causal mechanism or set of causal mechanisms" (Falletti og Lynch, 2009: 1152). En "ildmekanisme" kan fx ikke fungere uden ilt.

Studiet af mekanismer med casestudier indebærer flere ting, som har været overset i den eksisterende litteratur. Det første er, at man skal bruge deterministiske teorier. For det andet skal man bruge en sæt-logisk forståelse af forholdet mellem X og Y, hvilket indebærer, at man laver asymmetriske påstande om kausale forhold. For det tredje er det en logisk mulighed, at det samme X i forskellige kontekster (scope conditions) er bundet til Y af forskellige mekanismer – kaldet ækvifinalitet på mekanismeniveau. Endelig: Hvis man tager "mekanisme som systemer" alvorligt, burde forskellige mekanismer mellem forskellige årsager og Y efterlade forskellige empiriske fingeraftryk, som mindsker behovet for at isolere effekten af et enkelt X i vores casevalg. Det følgende vil introducere disse argumenter, mens de udfordringer, de rejser for forskningsdesign, diskuteres videre i fjerde afsnit.

Mekanismer er deterministiske påstande

"Mekanismer som systemer" forstås som regel som invariante og deterministiske processer, som vil starte, hvis årsagen X og de relevante randbetingelser er til stede (Anderson, 2011; Mayntz, 2004; Waldner, 2012). Der er visse forskere,

der argumenterer for, at mekanismer kan være probabilistiske teorier, men der er to grunde til at forkaste dette argument. For det første skyldes det, som kan minde om et probabilistisk forhold mellem X + M og Y som regel, at vi ikke har tilstrækkelig viden om de randbetingelser, der skal være til stede, for at en mekanisme starter. Ting sker ikke tilfældigt i verden – det er bare vores forståelse af verden, som altid er ufuldkommen. Og for det andet er en probabilistisk forståelse af kausale forhold uforenelig med enkelte casestudier, da vi i bestemte tilfælde ikke vil vide, hvorvidt vi observerer undtagelsen, der bekræfter reglen, eller om der ikke er den forventede kausaleffekt (Mahoney, 2008).

Sæt-logik og asymmetrisk kausalitet

Brugen af casestudier og deterministiske teorier har den konsekvens, at vi bruger en sæt-logisk forståelse af forholdet mellem X, M og Y. X skal være til stede for at starte M – uden X ingen M. Når X og de relevante randbetingelser er til stede, vil M altid ske. Hvis X ikke i sig selv er tilstrækkelig til at producere Y, vil M blot føre til en overførelse af kausale kræfter til Y, men Y vil ikke nødvendigvis ske. Hvis X (eller en kombination af forskellige X'er) er tilstrækkelig, vil X og M altid producere Y. I sæt-logik er begreber defineret ud fra, hvad deres teoretiske essens er (Goertz, 2006). Vigtigst er at definere den kvalitative tærskel (*difference-in-kind*) mellem, hvad der definerer en case som medlem af sættet af X eller Y og alt andet. Desuden kan gradsforskelle inden for sættet af et begreb eksistere, men disse er underordnet sætmedlemskabet.

Sæt-logik indebærer også, at vi laver asymmetriske påstande om kausale forhold. At sige, at X + M medvirker til at producere Y, er ikke det samme som at sige, at fraværet af X (ikke X, eller ~X) producerer ~Y. Vi laver derfor kun asymmetriske påstande om kausalitet, hvor vi siger noget om X + M og deres forhold til Y, men laver ingen påstande om betydning af ~X, eller hvilke andre årsager der er skyld i ~Y.

Forskellige veje mellem samme X og Y – ækvifinalitet på mekanismeniveau

En standardkritik af mekanismer er, at "for each [theoretical X→Y relationship] … one finds a litany of theoretically plausible causal mechanisms …" (Gerring, 2010: 1510). En bestemt årsag kan i teorien føre til samme Y igennem mange forskellige mekanismer. Hvilke mekanismer vil være bestemt af de randbetingelser, der er til stede i en case. Fx kan man forestille sig, at økonomisk udvikling (X) og demokratisering (Y) kan være bundet sammen af en urbaniseringsmekanisme i en bestemt kontekst, mens det i en anden kontekst kunne være igennem en uddannelsesmekanisme (Gerring, 2010: 1508).

Udfordringen er, at hvis vi finder en bestemt mekanisme i en case, kan vi ikke vide, om det er samme mekanisme, som forbinder X til Y i andre cases. Bare fordi vi finder uddannelsesmekanismen i den sydkoreanske case, er det ikke ensbetydende med, at det også er samme vej mellem X og Y i Taiwan.

Forskellige mekanismer, forskellige fingeraftryk

En anden ofte overset implikation af studiet af mekanismer er, at de burde efterlade forskellige empiriske fingeraftryk, som vil kunne fortælle os, hvorvidt de var til stede eller ej. Et klassisk eksempel er en person, der er fundet død i en ørken. Der er to overlappende og mulige årsager (overdetermineret udfald): et hul i vandflasken (tørst) og gift i vandflasken (gift). Begge kunne være tilstrækkelige til at slå manden ihjel. Hvis vi operationaliserer mekanismerne ordentligt, bør man forvente, at tørstmekanismen vil efterlade væsentligt anderledes spor end giftmekanismen. Hvis vi finder ud af, at han rent faktisk drak vandet (spor af giften i hals og mave), blev giften absorberet af kroppen (spor af giften i blodårerne), og giften førte til hjertestop (et pludseligt hjertestop efterlader bestemte stressspor), kan vi slutte, at han døde af gift. Sporene for tørstmekanismen vil være væsentligt anderledes, hvilket vil gøre, at vi kan opdatere vores tillid til, hvorvidt en bestemt mekanisme spillede en rolle, uanset om andre mulige årsager er til stede.

Et trin for trin-eksempel på et PT-casestudie af en mekanisme

Inden jeg diskuterer de udfordringer, der er forbundet med studiet af mekanismer med PT-casestudier, er det væsentligt først at præsentere metodens relative styrker. Dette gøres ved at vise et eksempel på, hvordan PT ser ud i praksis. Metodens særlige styrke er vores detaljerede viden om "processen" mellem X og Y, ved at vi undersøger, om vi finder de forventede fingeraftryk for hver del af mekanismen. Fordi vi er så tæt på vores cases, sker der ofte en frem og tilbage-proces mellem empiri og teori, hvor vi bruger det, vi har lært baseret på empiriske analyser, til at forbedre teorien. Naturligvis tester vi ikke en ny teori på det samme empiriske materiale, som er brugt til at udforme det; her er der behov for at udlede nye og uafhængige empiriske manifestationer, som kan undersøges efterfølgende.

Som eksempel bruges her teorien om forholdet mellem anvendelse af folkeafstemninger om EU-spørgsmål (X) og regeringer, som indtager positioner, der afspejler den offentlige opinion (Y) (folkeafstemninger → kongruens). Teorien siger, at i de lande, der skal afholde folkeafstemninger for at ratificere en EU-traktat, er der en større sammenfald mellem offentlig opinion og de positioner,

som regeringer indtager i forhandlingerne om en EU-traktat (Hug og König, 2002; Finke, 2009).

En plausibel mekanisme mellem X og Y kunne være som illustreret i den øverste del af figur 1, hvor de enkelte dele, som forbinder X med Y, er defineret. Her er hver del af en meget simpel mekanisme defineret ved en enhed (fx regering), som laver en aktivitet (fx indsamler information). Disse aktiviteter fanger, hvordan transmissionen af kausale kræfter i teorien sker igennem mekanismen fra X til Y. Hver del af mekanismen har *ikke* en *selvstændig* effekt på Y, men har *kun* effekter som en del af et system mellem X og Y. Det er særligt mekanismens fokus på aktiviteter, som adskiller denne systemforståelse fra en variabel mekanismedefinition, og som er kernen i de analytiske fordele, som en dybdegående sporing af en mekanisme i en enkelt case har.

Efter at en mekanisme er konceptualiseret, skal de empiriske fingeraftryk, det forventes at efterlade, udpensles. Inden hvert led af mekanismen operationaliseres, skal forskeren vælge den case, der skal undersøges. Casen skal vælges først, da de empiriske fingeraftryk, som hvert af mekanismens led efterlader, er casespecifikke. Fx er parlamentariske dynamikker forskellige i Danmark og Holland, men det betyder ikke, at der er forskellige teoretiske mekanismer på spil i de to lande.

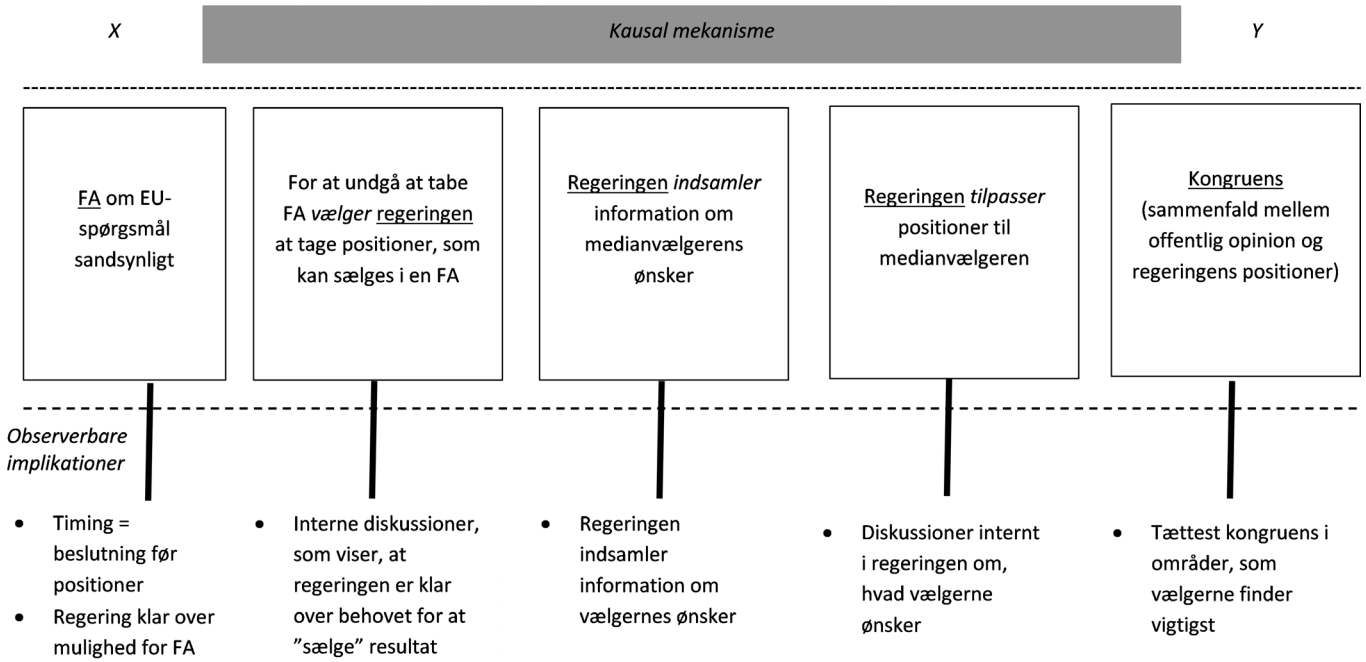
Når man vælger en case, skal både X, Y og de relevante randbetingelser være til stede ud fra devisen: "Jeg vil kun undersøge, *om* der er en mekanisme til stede i en situation, hvor mekanismen *kunne* være til stede". Hvorfor undersøge mekanismen, hvormed en folkeafstemning muligvis har produceret kongruens i et land, som ikke skal holde en afstemning? Det kunne fortælle os om andre veje til kongruens, men ikke noget om mekanismerne mellem en afstemning og kongruens.

I det konkrete eksempel har jeg valgt Irland under forhandlingen af forfatningstraktaten som case, hvor de observerbare manifestationer af hvert led er som afbildet i figur 1. Irland er medlem af både X og Y i dette tilfælde (se figur 2 nedenfor).

I praksis vil disse fingeraftryk i form af forventet bevismateriale være endnu mere udpenslede og casespecifikke, fx med en konkret beskrivelse af, hvordan jeg vil genkende "regeringen indsamler information" i den irske case, når jeg ser det.

For hver af disse empiriske fingeraftryk skal man vurdere to ting: 1) Er der andre plausible forklaringer på tilstedeværelse af et bestemt bevis? 2) Siger teorien, at jeg *skal* finde dette bevismateriale? Det første henviser til en empirisk tests "unikhed", og dermed, hvis vi finder et bevis, om det styrker vores tillid til, at et bestemt led af en mekanisme er til stede. Alternative forklaringer er

Figur 1: En mekanisme mellem folkeafstemninger og kongruens



Note: Aktiviteterne er i kursiv, enhederne understreget.

som regel ikke konkurrerende teorier, men mere casespecifikke forklaringer, såsom at politikere altid indsamler information om, hvad vælgerne ønsker, uanset om de skal bruge det til at tilpasse positionerne eller ej. Hvis det er lige så plausibelt at finde et bevis med alternative forklaringer, kan vi ikke bruge det til at styrke vores tillid til tilstedeværelsen af leddet, og modsat hvis beviset ikke kan dækkes af alternative forklaringer.

For det andet siger vores teori, at vi skal finde et bestemt fingeraftryk. Dette henviser til, hvorvidt en test er "sikker". I eksemplet med indsamling kunne man argumentere for, at "kloge" politikere vil have en veludviklet fornemmelse for folkestemningen og dermed ikke vil have behov for aktivt at indsamle information for at vide, hvad vælgerne ønsker. Derfor vil dette fingeraftryk ikke være sikkert at finde, og dermed ved vi ikke, om leddet ikke fandtes eller bare manifesterede sig på anden vis, hvis ikke vi finder det.

Derefter går man ud og undersøger, hvorvidt de forventede beviser rent faktisk findes i casen. Styrken af inferensen om tilstedeværelse af mekanismen er afhængig af, hvor stærke "tests" man har udviklet. Kan man forklare forekomsten af et bevis med alternative plausible forklaringer? Eller kan forekomsten kun forklares plausibelt, hvis den undersøgte del rent faktisk eksisterer? Hvis man ikke finder det forventede bevis, hvad kan man så sige? Var testen sikker? Her laver man inferens baseret på en bayesiansk logik i stedet for en frekventistisk logik (for en introduktion til denne logik se Beach og Pedersen, 2013: 253-256).

Når man har analyseret beviserne for hver af mekanismens dele og fx fundet klare beviser for, at de eksisterede i casen, kan man lave en stærk inferens om, at mekanismen fandtes i casen. Det betyder, at man kan lave inferensen om, at der var en kausalforbindelse mellem X og Y, og man har forklaret processen, hvormed X har medvirket til at producere Y. Som George og Bennett skriver: "process tracing provides a strong basis for causal inference only if it can establish an uninterrupted path linking the putative causes to the observed effects" (2005: 222).

Når aktiviteterne udpensles i en mekanisme, skal vores efterfølgende empiriske analyse undersøge disse aktiviteter (i observationsstudier igennem de fingeraftryk, de efterlader). Herved kan man argumentere for, at vi er tættere på at observere kausalitet mellem X og Y, end hvis vi kun undersøger korrelationer, da vi empirisk undersøger, hvad der sker mellem X og Y, og især ved at undersøge hver del af mekanismen mellem de to.[4]

Hvorfor kan vi ikke undersøge både enhederne *og* aktiviteterne på tværs af flere cases? Den primære grund til, at aktiviteterne skal droppes, illustreres i det følgende med eksemplet fra figur 1. Hvis vi vil undersøge, hvorvidt en re-

gering rent faktisk tilpasser sine positioner til vælgernes holdning, vil det kræve en dybdegående analyse af, hvad der skete internt i en regering før eller under forhandlinger af den givne EU-traktat. Man kunne undersøge dette ved blot at kigge på korrelationen mellem den offentlige opinion og regeringens positioner, men derved har man *ikke* undersøgt processen. Finke (2009) forsøger at imødekomme dette ved at lave en tidsserieanalyse, hvor han kigger på kortsigtet variation i vælgernes ønsker og udviklingen af regeringens positioner, indtil de er færdigformulerede, men selv her har han kun fanget eventuelle korrelationer og ikke undersøgt den proces, hvormed regeringen aktivt tilpasser positioner. Derved kan han ikke udelukke, at 1) regeringen enten tilfældigt har indtaget en position, som falder sammen med vælgernes ønsker, 2) at regeringen har påvirket vælgernes holdninger igennem budskaber i medierne (*cues*), eller 3) at der rent faktisk er sket en tilpasning pga. vælgerne. Problemet ved kun at undersøge korrelationen er derfor, at man ikke kan skelne mellem disse tre forklaringer – det kræver, at man i den enkelte case går ind og undersøger empirisk, hvorvidt der er beviser for, at regeringen *aktivt* har indtaget positioner, som afspejler, hvad de forventer kan "sælge" traktaten i den kommende folkeafstemning. Og dette kan kun lade sig gøre ved at lave et dybdegående casestudie af processen (mekanismen).

En anden måde at undersøge mekanismer på kunne være at følge KKV's forslag (King, Keohane og Verba, 1994: 208-230). Hvis vi forstår mekanismer som et sæt intervenerende variabler mellem X og Y, som KKV gør, bør vi undersøge den kausale effekt af hver enkelt af mekanismens dele ved at undersøge samvariationen mellem værdierne af hver del og Y – enten med store-n undersøgelser eller mindre-n design (mellem fem og 20) som beskrevet af KKV.

Men dette kræver, at vi har et design, hvor vi kan isolere effekten af hver intervenerende variabel for at undersøge størrelsen af den effekt, den har på værdien af Y, på tværs af en række cases. Men her går vi væk fra idéen om mekanismer som systemer, da de enkelte mekanismedele i en systemforståelse kun har kausale effekter som en del af hele systemet. Det er også derfor, at forskere, der tager mekanismer seriøst, mener, at hvis vi skal studere mekanismer, kræves der invariante forskningsdesign.

Samtidig, og i modsætning til ved eksperimenter, er vi ved at analysere aktiviteterne blevet klogere på, hvordan X har en kausal effekt på Y igennem en mekanisme, mens processen forbliver inden for den sorte kasse i et eksperiment, hvor man kun kan spekulere på, hvilke mekanismer der forbinder X og Y (Waldner, 2012: 76).

Konklusionen af et PT-casestudie er, at man vurderer, hvorvidt man har været i stand til i en positiv eller negativ retning at opdatere sin tillid til, at en

mekanisme mellem X og Y er til stede. Derved får man viden om processen mellem X og Y, som er svær at opnå med andre metoder, hvis ikke umulig. Men ulempen er, at da man kun har undersøgt mekanismen i en enkelt case, kan man strengt taget *kun* lave inferens inden for den undersøgte case. Hvis man har forsøgt at sige noget mere generelt om et teoretisk fænomen, hvordan kan man så lave bredere inferens på tværs af en række cases? Det er her, udfordringerne for PT som metode virkelig starter.

Udfordringer og løsninger

Hvordan indlejrer man et PT-casestudie af en mekanisme, således at man kan lave inferens om kausale forhold på tværs af en række cases? Der er en række udfordringer og begrænsninger, som har været overset i litteraturen.

Kan PT-studier indlejres i et multimetodedesign?

Den første udfordring er skabt af fokusset på mekanismer og særlig anvendelse af deterministiske teorier, hvilket gør det meget svært at få resultaterne fra sit PT-studie til at tale sammen med store-n analyser af probabilistiske teorier. I den eksisterende litteratur er disse problemer komplet overset. Ifølge Liebermans "Nested Analysis" (2005) starter man sin undersøgelse med en stor-n analyse (large-n analysis, LNA) af forholdet mellem X og Y. Hvis en robust korrelation findes, skal dette undersøges i et casestudie (small-n analysis, SNA). Man vælger en case på den fundne regressionslinje (on-lier) for at undersøge, om der er beviser for, at X og Y rent faktisk er i et kausalforhold. Hvis der findes beviser, konkluderer man, at X er en årsag til Y. Hvis ikke der findes beviser, undersøges det, om man har valgt en idiosynkratisk case, eller om teorien skal revurderes igennem en SNA, som forsøger at finde nye årsager.

Der er flere problemer med Liebermans simple model i forhold til PT og mekanismer. Det første, der skal nævnes – hvilket også er det mest alvorlige – er, at de teoretiske påstande om forholdet mellem X og Y i LNA og SNA er forskellige: Den ene er probabilistisk, og den anden deterministisk. Der er store teoretiske forskelle mellem at sige, at når værdier af X stiger, tenderer værdierne af Y at stige (probabilistisk påstand), og at X er nødvendig for, at Y finder sted (deterministisk). Desuden opstår der logiske problemer ved, at en deterministisk påstand er asymmetrisk, hvor man kun påstår noget om forholdet mellem X og Y og ikke årsager til ~Y (symbolet betyder ikke Y), mens probabilistiske påstande er symmetriske, hvor man siger noget om både Y og ~Y.

Et andet problem er, at man strengt taget undersøger to forskellige ting empirisk: X:Y-korrelation i LNA og mekanismen imellem X og Y i SNA inden for en enkelt case. Når vi kun har kigget på en enkelt case, kan vi ikke vide, om

den samme mekanisme forbinder X og Y i andre cases – et problem, som jeg vender tilbage til nedenfor.

Det tredje problem er, at hvis vi følger Liebermans råd om casevalg, kan vi ende med at vælge cases, hvor mekanismen per definition ikke kan eksistere. Dette skyldes, at bare fordi en case ligger på regressionslinjen, betyder det ikke, at mekanismen mellem X og Y er til stede. Hvis værdien af X er under den kvalitative tærskel (se afsnit 2), vil mekanismen ikke starte. Og hvorfor undersøge en mekanisme i en case, hvor vi på forhånd ved, at mekanismen ikke er til stede?

Løsningen på disse problemer er at bruge LNA-metoder, som er mere kompatible med PT, såsom QCA (kvalitative komparative analyser). QCA arbejder også med deterministiske (eller næsten deterministiske) teorier og sæt-logik, som gør dem mere forenelige med PT. Med brug af QCA kan vi på tværs af en række cases undersøge, om der er et forhold mellem X (eller en gruppe af X'er) og Y. Hvis en gruppe af årsager, som sammen er tilstrækkelige til at producere Y, er fundet i QCA, vil man vælge en case, som er "medlem" af gruppen af årsagerne og Y, for at undersøge, om der er en mekanisme (eller mekanismer), som forbinder de to i casen (Rohlfing og Schneider, 2013). Men hermed opstår de næste to udfordringer.

Ækvifinalitet på mekanismeniveau

Hvis vi finder en mekanisme mellem X og Y i en bestemt case, kan vi strengt taget ikke lave en inferens til andre cases i populationen. Dette strider imod Liebermans påstand om, at man kan nøjes med et casestudie. Han skriver: "if one or more intensive case studies can demonstrate the validity of the theoretical model – which had already passed muster in the LNA – by plausibly linking cause to effect in the expected manner, then the nested analysis provides ringing support for the model" (2005: 448). Men hvis vi tager risikoen for ækvifinalitet på mekanismeniveau alvorligt, og hvorfor skulle vi ikke det i en meget kompleks social virkelighed, betyder det, at vi ikke bare kan antage, at det er samme mekanisme, som binder X og Y sammen i andre cases. Dette er noget, vi skal undersøge empirisk.

Dette gøres ved at gentage vores PT-casestudie med en anden case, evt. lidt mere overfladisk, for at finde ud af, hvorvidt samme mekanisme er på spil. Udfordringen opstår ved, at et PT-casestudie er meget tidskrævende. Og selvom vi har tiden til det, vil det være næsten umuligt at få plads til bare et ordentligt PT-casestudie i en publikation med artikellængde, for ikke at tale om to eller flere, hvis vi gerne vil lave en inferens til resten af populationen.

Men hvis vi er i stand til det, og vi finder nogenlunde samme mekanisme i en anden case, kan vi begynde at lave en forsigtig inferens til resten af populationen baseret på logikken "Hvor sandsynligt er det, at jeg finder stærke beviser for, at mekanismen er til stede i to cases, uden at den også virker i andre cases i populationen?".

Hvis vi ikke finder samme mekanisme i den anden case, vil komparative metoder være mest velegnede til at sammenligne de to cases med hinanden og resten af populationen for at finde ud af, om casen var idiosynkratisk, eller om der er kontekstuelle faktorer, som gør, at forskellige mekanismer er på spil i forskellige cases.

Inferens til en begrænset population, hvor X, Y og randbetingelser er til stede

Efter to eller flere PT-casestudier, som finder en bestemt mekanisme, kan man begynde at lave en inferens til resten af populationen. Men den tredje udfordring er, at selv hvis man kan gøre det, så er populationen meget begrænset, nemlig til kun de cases, hvor X, Y og de relevante randbetingelser er til stede.

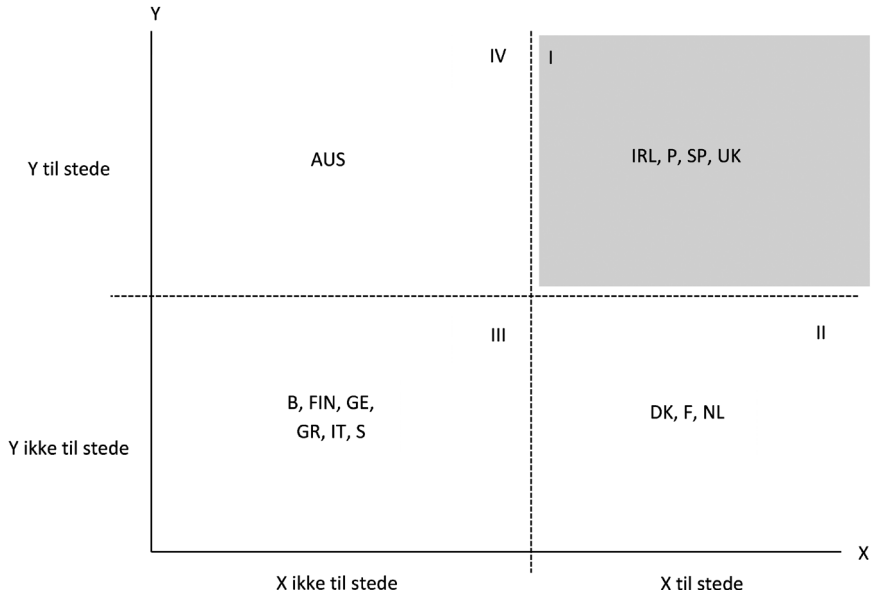
Resultatet kan se ud som figur 2, som bygger på eksemplet om forholdet mellem folkeafstemninger (X) og kongruens (Y). Her er en række cases plottet ind, alt efter om X og Y er til stede i en bestemt forhandlingsrunde (forfatningstraktaten) eller ej.

Bemærk, at grænsen mellem, om et begreb er til stede eller ikke til stede, sættes ud fra både teoretiske og empiriske grunde (Ragin, 2008; Beach og Pedersen, 2013). Her er Beachs data brugt (under udgivelse), hvor alle EU-medlemsstater i figuren er plottet ind efter, om de er medlem af X og Y.

Hvis vi er interesserede i at undersøge kausalmekanismen mellem X og Y, kan vi vælge en af casene i zone I, fx Irland. Hvis vi finder mekanismen i den irske case, kan vi dog ikke i første omgang være sikre på, at den samme mekanisme findes mellem X og Y i de andre cases i zone I (gråzone i figur 2). Men hvis vi undersøger en anden case inden for zone I (fx UK) og opdager, at samme mekanisme virker, kan vi stadig kun sige noget om fire ud af 14 cases (28 pct.).

De fleste kvalitative forskere vil ikke anse dette som et problem, da man i forvejen arbejder med idéen om, at man kun kan sige noget om begrænsede populationer (fx Goertz og Mahoney, 2009). En kvalitativ forsker vil også sige, at naturligvis er årsagerne til Y anderledes end årsagerne til ~Y. Årsagerne til fred er væsentlig anderledes end årsagerne til krig, og mens gensidigt demokrati fx kan være en årsag til fred, er manglende demokrati ikke en årsag til krig. Kausale forhold er asymmetriske.

Figur 2: Folkeafstemninger (X) og kongruens (Y) i forhandlingen af forfatningstraktaten



Note: Landenes engelske forkortelser er brugt.
Kilde: Se Beach (under udgivelse).

Cases uden for zone I kan være relevante i visse tilfælde afhængigt af vores formål med undersøgelserne. Om casene i zone III og IV ved vi på forhånd, at der ikke skulle holdes en folkeafstemning (X), og derfor kan der her rent logisk ikke være en kausal forbindelse mellem X og Y. Hvis vi har interesse i at undersøge andre årsager til Y end X, kan casene i zone IV være relevante, men her vil komparative metoder (fx et *most similar systems design*) være langt nemmere at anvende, når man forsøger at finde andre årsager end X.

Casene i zone II, hvor X, men ikke Y er til stede, er mere interessante. Disse cases minder om rygere, der har røget mange cigaretter dagligt i mange år uden at have fået lungekræft. Vi vil gerne vide, hvorfor disse cases afviger fra det, vi ellers ved fra vores forudgående undersøgelser af casene i zone I. Her kan man bruge PT til at undersøge, hvornår og hvordan denne mekanisme mellem X og Y bliver afsporet. Men det er mere komparation af den afvigende case (fx DK) med en typisk case i zone I (Irland), som gør os klogere på, hvorfor folkeafstemninger ikke fungerer i DK, mens de fører til kongruens i den irske case.

Konklusion

Denne artikel har argumenteret for, at hvis vi skal blive klogere på kausale relationer mellem årsager (X) og udfald (Y) i form af den mekanisme, som forbinder dem, er det bedste metodiske redskab dybdegående casestudier (process tracing, PT). I PT laver man en dybdegående analyse af den mekanisme, som forbinder X med Y. Mens metoder som eksperimenter kan påvise, at X har en kausal effekt på Y, ved at manipulere forekomsten af X i de undersøgte cases, bliver vi ikke klogere på den proces, hvormed X producerer Y. Det er her, PT har sin komparative styrke. PT indebærer, at man laver et dybdegående casestudie, som ser, om de forventede observerbare implikationer af hvert af mekanismens led er til stede i empirien i en enkelt case. Der er flere grunde til, at man vælger at lave enkelte casestudier. Den mest væsentlige er, at hvis vi tager mekanismer alvorligt, skal de forstås som systemer, der enten findes eller ikke findes i en bestemt case. Dvs. vi tester forekomsten af hele mekanismen i en case. Desuden er mekanismernes fingeraftryk casespecifikke og kan derfor ikke sammenlignes på tværs af casene.

Men PT er en meget specialiseret metode med mange ulemper. Blandt disse er, at mens man kan lave stærke inferenser om kausalitet inden for den enkelte undersøgte case, kan man ikke lave generaliseringer på baggrund af et PT-casestudie alene. Her er det nødvendigt, at man indlejrer PT-casestudiet i et komparativt design, hvor man kortlægger casene i populationen på X og Y. Dette er meget svært at gøre med regressionsbaserede metoder, da disse arbejder med probabilistiske teorier i forhold til PT's deterministiske teorier. Hvis man gerne vil generalisere fra den undersøgte case til andre cases, skal man undersøge mekanismen i mindst en anden case, hvor både X og Y er til stede. Og selv hvis det kan lade sig gøre at indlejre et PT-casestudie, og man er tryg ved, at mekanismen er til stede i andre cases, kan man kun generalisere til populationen, hvor X, Y og de relevante randbetingelser er til stede. Konklusionen er, at vi med PT kan få stærke påstande med belæg om kausalitet i den enkelte case, mens vores muligheder for at sige noget mere generelt ofte er meget begrænsede. Prisen er, at man lærer meget om lidt i stedet for lidt om meget.

Noter

1. Artiklen diskuterer ikke, hvorvidt mekanismer kan analyseres med andre metoder, men fokuserer udelukkende på de analytiske gevinster, PT giver, sammen med udfordringer og mangler ved metoden.
2. Læg mærke til ordet "medvirke". Det henviser til, at vi ikke per definition antager, at X er tilstrækkelig til at producere Y alene.

3. Bemærk, at denne definition ikke er en "teori", men en ontologisk påstand om, hvordan årsager og virkninger er forbundet. Derfor er det en fejllæsning at blande det sammen med systemteori fra sociologi (jf. Luhmann, 2004).
4. Der er stor debat om, hvorvidt vi direkte kan observere kausale mekanismer eller kun inferere deres eksistens ved at se på deres observerbare implikationer. Efter min mening er der ikke forskel på de metodiske implikationer af disse to standpunkter, og derved bliver det mere et filosofisk spørgsmål uden relevans for vores brug af PT-metoden. For yderligere information se Beach og Pedersen (2013).

Litteratur

Anderson, Holly (2012). The case for regularity in mechanistic causal explanation. *Synthese* 189: 415-432.

Beach, Derek (under udgivelse). Lessons from combining QCA and Process-tracing in a nested analysis of congruence. *Sociological Methods and Research.*

Beach, Derek og Rasmus Brun Pedersen (2013). *Process-tracing Methods: Foundations and Guidelines.* Ann Arbor: University of Michigan Press.

Bunge, Mario (1997). Mechanism and explanation. *Philosophy of the Social Sciences* 27 (4): 410-465.

Caren, Neal og Aaron Panofsky (2005). TQCA: A technique for adding temporality to qualitative comparative analysis. *Sociological Methods and Research* 34 (2): 147-172.

Chen, Jie og Chunlong Lu (2011). Democratization and the middle class in China: The middle class's attitudes toward democracy. *Political Research Quarterly* 64 (3): 705-719.

Falleti, Tulia G. og Julia F. Lynch (2009). Context and causal mechanisms in political analysis. *Comparative Political Studies* 42: 1143-1166.

Finke, Daniel (2009). Domestic politics and European treaty reform: Understanding the dynamics of governmental position taking. *European Union Politics* 10 (4): 482-506.

George, Alexander L. og Bennett Andrew (2005). *Case Studies and Theory Development in the Social Sciences*. Cambridge: MIT Press.

Gerring, John (2007). *Case Study Research*. Cambridge: Cambridge University Press.

Gerring, John (2010). Causal mechanisms: Yes but .... *Comparative Political Studies* 43 (11): 1499-1526.

Glassman, Ronald M. (1997). *The New Middle Class and Democracy in Global Perspective.* London: MacMillan.

Glennan, Stuart S. (1996). Mechanisms and the nature of causation. *Erkenntnis* 44 (1): 49-71.

Goertz, Gary (2006). *Social Science Concepts: A User's Guide*. Princeton: Princeton University Press.
Hedström, Peter og Petri Ylikoski (2010). Causal Mechanisms in the Social Sciences. *Annual Review of Sociology* 36: 49-67.
Hernes, Gudmund (1998). Real virtuality, pp. 74-101 i Peter Hedström og Richard Swedberg (red.), *Social Mechanisms an Analytical Approach to Social Theory*. Cambridge: Cambridge University Press.
Hug, Simon og Thomas König (2002). In view of ratification: Governmental preferences and domestic constraints at the Amsterdam Intergovernmental Conference. *International Organization* 56 (2): 447-476.
Huntington, Samuel P. (1991). *The Third Wave: Democratization in the late Twentieth Century.* Norman: University of Oklahoma Press.
King, Gary, Robert O. Keohane og Sidney Verba (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research.* Princeton: Princeton University Press.
Lieberman, Evan S. (2005). Nested analysis as a mixed-method strategy for comparative research. *American Political Science Review* 99 (3): 435-451.
Lipset, Seymour R. (1959). Some social requisites of democracy: Economic development and political legitimacy. *American Political Science Review* 53 (1): 69-105.
Luhmann, Niklas (2004). *Law as a Social System*. Oxford: Oxford University Press.
Machamer, Peter (2004). Activities and causation: The metaphysics and epistemology of mechanisms. *International Studies in the Philosophy of Science* 18 (1): 27-39.
Machamer, Peter, Lindley Darden og Carl F. Craver (2000). Thinking about mechanisms. *Philosophy of Science* 67 (1): 1-25.
Mahoney, James (2001). Beyond correlational analysis: Recent innovations in theory and method. *Sociological Forum* 16 (3): 575-593.
Mahoney, James (2008). Toward a unified theory of causality. *Comparative Political Studies* 41 (4/5): 412-436.
Mahoney, James og Gary Goertz (2006). A tale of two cultures. Contrasting quantitative and qualitative research. *Political Analysis* 14 (3): 227-249.
Mayntz, Renate (2004). Mechanisms in the analysis of social macro-phenomena. *Philosophy of the Social Sciences* 34 (2): 237-259.
Morgan, Stephen L. og Christopher Winship (2007). *Counterfactuals and Causal Inference.* Cambridge: Cambridge University Press.
Ragin, Charles C. (2008). *Redesigning Social Inquiry. Fuzzy Sets and Beyond*. Chicago: University of Chicago Press.
Shin, Eui Hang (1999). Social change, political elections, and the middle class in Korea. *East Asia: An International Quarterly* 17 (3): 28-60.

Waldner, David (2012). Process tracing and causal mechanisms, pp. 65-84 i H. Kincaid (red.), *Oxford Handbook of the Philosophy of Social Science*. Oxford: Oxford University Press.
Waskan, Jonathan (2011). Mechanistic explanation at the limit. *Synthese* 183: 389-408.

politica, 46. årg. nr. 1 2014, 42-59

Asmus Leth Olsen

Tærskelvariable og tærskelværdier: en introduktion til regressions-diskontinuitetsdesignet[1]

Regressionsdiskontinuitetsdesignet (RDD) udnytter empiriske situationer, hvor vi kan observere en variabel, der ved en given værdi opdeler de observationer, vi studerer, i en kontrol- og forsøgsgruppe. RDD har vist sig at reproducere traditionelle, eksperimentelle resultater, være intuitivt enkelt og relativt nemt at implementere i statistiske programpakker. Alligevel er det først inden for de seneste par år, at RDD er blevet anvendt i statskundskaben. Oftest har krikken af designet været, at den empiriske virkelighed sjældent tilbyder situationer, hvor RDD er anvendeligt. Denne artikel introducerer forskere og studerende til RDD i en statskundskabssammenhæng. Der vil være et særligt fokus på at vise, at RDD faktisk er brugbart til at besvare ganske mange forskelligartede kausale spørgsmål i statskundskaben.

Observationsstudier er karakteriseret ved, at vi som forskere eller studerende ikke aktivt har interveneret for at opdele de individer eller organisationer (herefter: observationer), vi studerer, i en kontrol- og forsøgsgruppe. Vi er derfor henvist til at studere observationer, der selv har selekteret sig ind i kontrol- og forsøgsgrupper eller er blevet selekteret af andre (Robinson et al., 2009). Idéen om kontrol- og forsøgsgrupper er således alene et udtryk for, at observationerne har forskellige værdier på den variabel, som vi ønsker at estimere den kausale effekt af. Hvorfor de er endt i den ene eller anden gruppe, er derfor ude af vores hænder. Denne (selv)selektion betyder, at observationerne formentlig er havnet i en særlig gruppe af gode grunde, hvorfor der som oftest vil være systematiske forskelle mellem kontrol- og forsøgsgruppe. Dette giver os et potentielt *confounding* problem, da der vil være en række faktorer, der påvirker både observationernes placering i kontrol- eller forsøgsgruppe og den afhængige variabel, vi vil undersøge. Meget sjældent har vi den fornødne viden om verden eller præcise nok data til at tage højde for confounding (Freedman, 1991). Men i en kort, simpel artikel foreslog Thistlethwaite og Campbell (1960) regressionsdiskontinuitetsdesignet (RDD), som et muligt design til at besvare kausale spørgsmål i tilfælde, hvor vi direkte kan observere selektionsprocessen i observationsstudier og dermed omgå problemet med confounding.

I korte træk udnytter RDD empiriske situationer, hvor vi kan observere en variabel, der ved en given værdi opdeler observationerne i en kontrol- og forsøgsgruppe. Der er altså et punkt på en variabel, som vi observerer, der slår en given intervention til og fra. Interventionen er som i andre observationsstudier stadig ikke vores værk eller udtryk for et formelt lodtrækningsforsøg, som i de traditionelle eksperimenter. RDD udnytter i stedet, at den præcise placering af fordelingen af observationer omkring punktet, hvor interventionen indtræder, kan være tilnærmelsesvis tilfældig (Dunning, 2012). Hermed kan vi estimere en kausal effekt af en intervention ved at sammenligne forskelle i gennemsnit på den afhængige variabel for observationer omkring interventionspunktet, hvor opdelingen i kontrol- og forsøgsgruppe sker. I RDD sker dette ved, at vi modellerer den variabel, der opdeler observationerne og/eller begrænser vores analyser til de observationer, der ligger tættest ved den værdi, som skiller kontrol- fra forsøgsgruppen (Green et al., 2009; Dunning, 2012). Den centrale antagelse er, at observationer tæt ved denne værdi ikke nøjagtigt har kunnet påvirke, om de er endt i kontrol- eller forsøgsgruppen, og at de ikke er blevet placeret intentionelt af andre.

Med afsæt i denne simple idé har RDD langsomt spredt sig på tværs af socialvidenskaberne (Cook, 2008). Og der er flere gode grund til, at RDD også bør vinde større indpas i statskundskaben: For det første kan RDD estimere kausale effekter, som kommer meget tæt på tilsvarende traditionelle, eksperimentelle kausale effekter (Aiken et al., 1998; Berk et al., 2010; Green et al., 2009; Shadish, 2011). For det andet er RDD intuitivt forståeligt, og det kræver relativt milde antagelser sammenlignet med andre naturlige eksperimentelle designs (Lee og Lemieux, 2009: 1; Dunning, 2012: 136). Sekhon (2009: 503) har argumenteret for, at observationsstudiers største udfordring i forhold til kausale spørgsmål er, at vi ikke har fundet en måde at masseproducere designs på i samme omfang som det traditionelle eksperiment lader sig masseproducere. Vi har altså ikke et simpelt *quick fix*, når vi arbejder med observationsdata. I stedet har forskellige former for regressionsanalyse, hvor vi på bedste vis kontrollerer for mulige *confounders*, været det typiske udgangspunkt for at estimere kausale effekter i observationsstudier i statskundskaben. Men som Freedman (1991) har argumenteret for, har denne praksis været en vildfarelse for statskundskaben, da vi ikke har det teoretiske fundament til at finde de rette kontrolvariable eller modellere dem med den korrekte funktionelle form. RDD er måske det tætteste, vi kommer på et sådant quick fix til kausal inferens med observationsdata: Et design der i sin idé er lige så simpelt som det traditionelle eksperiment (Morgan og Winship, 2007: 251), og som samtidig har vist sig at kunne producere lige så troværdige estimater af kausale effekter.

Alligevel har RDD allerede for længst kunnet fejre 50 års fødselsdag uden at have vundet bred udbredelse som design til at afdække kausale relationer i statskundskaben. Først inden for de seneste få år er en række RDD-studier blevet publiceret. En forklaring på den manglende udbredelse af RDD har været, at selvom designet er stærkt og intuitivt, er det i realiteten ret sjældent, at virkeligheden tilbyder muligheder for, at designet faktisk kan finde anvendelse (Cook, 2008; Lee og Lemieux, 2009). Eller rettere: RDD kræver, at forskeren afsøger empiriske områder, hvor designet kan finde anvendelse. Dvs. forskerens arbejde forskyder sig derfor fra statistisk modellering og over mod at søge empiriske kontekster, der tillader brug af RDD. I de seneste år har statskundskaben langsomt fået øjnene op for, at designet faktisk er relevant til at besvare kausale spørgsmål i alt fra valgforskning til offentlig politik (Lee, 2008; Olsen, 2013a).

I denne artikel vil jeg introducere RDD til forskere og studerende inden for statskundskaben. Hovedvægten vil være rettet mod at kommunikere idéen bag designet og kun i mere begrænset omfang selve estimationen, som er behandlet mere fyldestgørende af andre (Green et al., 2009; Dunning, 2012). Det primære formål vil være at stimulere til, at flere tænker over, hvordan RDD kan anvendes til at estimere kausale effekter inden for netop deres forskningsområde.

Artiklen er struktureret på følgende vis: Først præsenteres nogle af de grundlæggende begreber i RDD. Dernæst følger de centrale antagelser for at estimere kausale effekter med RDD og en introducerende beskrivelse af, hvordan RDD-modeller estimeres statistisk. Herefter kommer eksempler på brug af RDD inden for statskundskaben. Inden konklusionen præsenteres nogle barrierer for brugen af RDD i statskundskaben samt mulige løsninger.

Regressionsdiskonkinuitetsdesignet: de centrale begreber

Naturlige eksperimenter betegner empiriske forhold, hvor observationer tildeles værdier på en variabel af interesse på baggrund af tilnærmelsesvis tilfældige processer i "naturen" (Dunning, 2012; Robinson et al., 2009). Hermed opnår vi eksogen variation på den variabel, hvis kausale effekt vi ønsker at estimere uden aktiv intervention fra vores side. RDD skal her ses som en særlig klasse af naturlige eksperimenter, der sætter specifikke krav til de empiriske omstændigheder, herunder hvordan de undersøgte observationer har fået tildelt værdier på den variabel, som vi ønsker at estimere den kausale effekt af. Udgangspunktet i RDD er, at vi med deterministisk viden om selektionsprocessen kan tage højde for den og derved identificere en kausal effekt (Rubin, 1977; Shadish et al., 2001).

RDD består i korte træk af tre elementer: en tærskelvariabel, en tærskelværdi og en afhængig variabel, som vi ønsker at undersøge. I realiteten kan der godt

være flere tærskelvariable og tærskelværdier (Papay et al., 2011), men her ser vi alene på det simple eksempel med én tærskelvariabel og én tærskelværdi. *T*ærskelvariab*len* (*assignment*, *running* eller *forcing* variabel) er en kontinuer før-interventionsvariabel, som vi observerer for alle de enheder, vi undersøger (Shadish et al., 2001). Jo mere finkornet tærskelvariablen er desto bedre, da den skal give os så præcis som mulig information om, hvorvidt en observation modtager en intervention af interesse eller ej. Tærskelvariablen skal deterministisk bestemme, om observationer tilhører vores kontrol- eller forsøgsgruppe. Tærskelvariablen bestemmer altså observationernes værdier på den variable, som vi ønsker at estimere den kausale effekt af. At tærskelvariablen bestemmer interventionen deterministisk betyder, at en given intervention slås til eller fra ved en særlig *tærskelværdi* (*discontinuty* eller *cut-off*). Tærskelværdien er det punkt på tærskelvariablen, hvor observationerne overgår fra kontrol- til forsøgsgruppe. Den deterministiske opdeling betyder, at hvis vi kender observationernes værdier på tærskelvariablen, ved vi også med sikkerhed, om de hører til vores kontrol- eller forsøgsgruppe.

I tilfælde hvor tildelingen af forsøgs- eller kontrolstatus ikke er deterministisk, taler man om *fuzzy* RDD (Dunning, 2012). I fuzzy RDD kan observationer på samme side af tærskelværdien være havnet i både forsøgs- og kontrolgruppen. En observations placering omkring tærskelværdien giver derfor ikke længere sikker viden om observationens interventionsstatus. I disse tilfælde vil tærskelværdien alene fortælle os sandsynligheden for, at en observation tilhører enten kontrol- eller forsøgsgruppe (Angrist og Pischke, 2009: 259). Dvs. at tærskelværdien nu udgør et skarpt brud i *sandsynligheden* for at være tildelt en intervention. I disse tilfælde er der udviklet en række forskellige instrumentvariabelprocedurer til at estimere selektionsprocessen omkring tærskelværdien (se Hahn et al. (2001) og Klaauw (2002) for nærmere forklaring og eksempler). Her vil vi ikke behandle fuzzy RDD yderligere.

Hvorfra kommer tærskelvariablene?

At finde egnede tærskelvariable er kernen i RDD. En tærskelvariabel kan lyde som et abstrakt fænomen, men er i virkeligheden en ret konkret og udbredt størrelse. Faktisk er der mange af politologiens centrale variable, som bestemmes af tærskelværdier. Som Green et al. (2009: 412) har argumenteret for, er der et stort potentiale for RDD i statskundskaben pga. "de rigide forhold hvorunder institutionelle regler fordeler repræsentation og regeringsressources". Med andre ord: I politisk-administrative sammenhænge er det ofte skarpe tærskelværdier, der bestemmer værdierne på nogle af de variable, som vi er allermest interesserede i at estimere den kausale effekt af. Dunning (2012: 69)

inddeler mulige tærskelvariable i socialvidenskaberne inden for kategorierne: adgangseksamener, befolkningstærskler, størrelsesbaserede tærskler, tildelingskriterier, alderstærskler og tætte valg. Nogle af disse kategorier er imidlertid ikke klart adskilte. Et bud på nogle mere overordnede tærskelkategorier er derfor: output/outcome-information, geografi/rum, socioøkonomiske kriterier og tid/timing. Et kort kig på nogle eksempler fra litteraturen viser udmærket, hvor mange forskellige tærskelværdier der bestemmer politisk-administrative forhold inden for hver af disse kategorier:

Output/outcome-mål kan i en statskundskabssammenhæng dække over både valgregler og valgresultater, der deterministisk bestemmer, hvem der får politisk magt (Lee, 2008; Pettersson-Lidbom, 2008; Broockman, 2009; Olsen 2013b). Men det dækker også over resultatmålinger af offentlige ansatte eller organisationer, der bestemmer sanktioner og performance feedback i den offentlige sektor (Chiang, 2009; Hemelt, 2011; Olsen, 2013a). Lee (2008) var den første til at udnytte valgresultater som en tærskelvariabel. I mange valgsystemer ændrer udfaldet sig deterministisk, hvis en kandidat får over 50 pct. af stemmerne. I valg med to kandidater er der en verden til forskel på at få 49 pct. eller 51 pct. af stemmerne. Det er til gengæld ret ligegyldigt, om man får 48 pct. eller 49 pct. 50 pct. er altså den tærskelværdi, der deler sejrherrerne fra de slagne.

Geografi dækker i statskundskaben typisk over observationernes afstand til grænser mellem nationer, kommuner eller andre former for administrative og politiske enheder (Black, 1999; Keele og Titiunik 2013). Et godt et eksempel er Chen et al. (2013), der benytter individers syd/nordplacering i forhold til Huai floden i Kina, der deterministisk giver gratis adgang til kulvarme og dermed også giver variation i graden af forurening, som borgerne oplever.

Socioøkonomiske kriterier kan være mål for befolkningsstørrelse, velstand eller sociale problemer, som bestemmer, hvem der har ret til at modtage hvilke ydelser (Ludwig og Miller, 2007; Elis et al., 2009; Becker et al., 2010; Hopkins, 2011). Ludwig og Miller (2007) har set på sundheds- og uddannelseseffekter af det amerikanske Head Start Program, der gav massiv støtte til børn i fattige familier. Det særlige ved programmet var, at hjælpen blev tildelt på baggrund af tærskler i amerikanske amters fattigdomsrater i 1960. Således var det kun familier i de 300 fattigste amter, der kunne modtage hjælp. Familier i det 301 fattigste amt var udelukket fra denne hjælp.

Endelig kan tærskelvariable udspringe af tidslig variation, der regulerer borgernes rettigheder i forhold til offentlige services såsom fødselsdato, alder, i særlige tidsintervaller eller lignende. For eksempel udnytter Almond og Doyle (2011), at lovgivningen i USA giver moderen ret til to døgn på hospitalet, hvor-

for børn født efter midnat i snit vil have et længere hospitalsophold end børn født umiddelbart før midnat.

Fra tærskelvariable til estimater af kausale effekter

Hvordan kommer man fra tærskelvariablen og tærskelværdien til kausal identifikation? Den centrale antagelse er, at observationerne i nærheden af tærskelværdien ikke har haft mulighed for at selvselektere sig ind på en særlig side af tærskelværdien, og at de heller ikke intentionelt er blevet placeret af andre. Placeringen af observationer omkring tærskelværdien skal således opfylde kriteriet om tilnærmelsesvis tilfældig (*as if random*) i nærheden af tærskelværdien (Dunning, 2012). Tænker vi for eksempel på Lee (2008), så er antagelsen, at i meget tætte valg er det tilfældige forhold, der afgør, hvem der vinder. I eksemplet med Almond og Doyle (2011) er det tanken, at det er tilfældigt, om man føder lige før eller efter midnat.

Hvis antagelsen om tilnærmelsesvis tilfældighed holder, betyder det, at observationer omkring tærskelværdien ikke er systematisk forskellige på observerbare og uobserverbare før-interventionsvariable. Dvs. der er ikke andre forhold givet før interventionen, som varierer systematisk omkring tærskelværdien, hvor interventionen slås til og fra. Det eneste, der varierer, er, om observationen tilhører kontrol- eller forsøgsgruppen, dvs. om de er placeret lige over eller under tærskelværdien.

At retfærdiggøre antagelsen om tilnærmelsesvis tilfældighed omkring tærskelværdien kræver både kvalitative argumenter og empiriske test. I hvilket som helst studie er det undersøgers opgave at redegøre for, hvorvidt RDD's antagelser holder i en given empirisk kontekst (Caughey og Sekhon, 2011: 405). Kvalitativt skal det være plausibelt, at selvselektion omkring tærskelværdien ikke er mulig. Det kan for eksempel begrundes med, at tærskelværdien har været ukendt for observationerne. Det vil gøre det mere plausibelt, at observationerne ikke har kunnet påvirke deres egen værdi på tærskelvariablen og dermed sikre, at de er kommet over eller under en særlig tærskelværdi. På samme måde kan det kvalitativt godtgøres, hvorvidt dem, der har bestemt tærskelværdien, har gjort det med henblik på, at særlige individer og organisationer skulle placeres på en bestemt side af tærskelværdien. Derfor er opgaven, som med alle andre naturlige eksperimentelle designs, at vi grundigt skal sætte os ind i den empiriske kontekst og de omstændigheder, der har tildelt observationerne værdier på tærskelvariablen (Dunning, 2012). Med Friedmans (1991) ord kan vi sige, at RDD er et godt eksempel på et design, der "slider på skosålerne": Det kræver, at vi nøje undersøger og forstår den empiriske kontekst.

Dette bør også ske ved kvantitativt, empirisk at undersøge, om der er forskelle mellem observationer i et snævert vindue omkring tærskelværdien. Dette kan ske ved at se på, om observationer varierer systematisk på før-interventionsvariable tæt omkring tærskelværdien. Hvis der ikke er systematiske forskelle på tværs af en række oplagte observerbare confounders, vil det styrke antagelsen om tilnærmelsesvis tilfældighed i placeringen af observationer i nærheden af tærskelværdien. Hvis data er tilgængelig for flere tidsperioder, bør man også teste, om observationerne er systematisk forskellige på den afhængige variabel og tærskelvariablen, i perioden før observationerne blev opdelt i kontrol- eller forsøgsgruppe (Caughey og Sekhon, 2011: 405). Hvis der er for eksempel er en systematisk forskel i den afhængige variabel for observationer omkring tærskelværdien, før denne værdi var kendt og trådte i kraft, kunne det tyde på, at netop denne værdi er valgt af strategiske grunde. Derudover bør man også undersøge, om der er forskydninger i fordelingen af observationer på tærskelværdivariablen omkring tærskelværdien (Imbens og Lemieux, 2008: 621). Hvis fordelingen er ujævn eller på andre måder uregelmæssig omkring tærskelværdien, kan det være et tegn på, at observationerne har haft mulighed for at påvirke deres placering omkring tærskelværdien. For eksempel hvis det af en eller anden grund er fordelagtigt at være i forsøgsgruppen, kunne man forestille sig, at observationer ville klumpe sig sammen lige på den rigtige side af tærskelvariablen, der giver adgang hertil. Dette kan både undersøges visuelt ved at plotte frekvensen af observationer omkring tærskelværdien eller ved formelle test (McCray, 2008).

Der er oplagt, at der er mange tærskelværdier ude i verden, som deterministisk griber ind i folks liv, men hvor observationerne har gode muligheder for selvselektion eller at blive intentionelt placeret af andre. For eksempel er der massevis af tærskler i de fleste skattesystemer (eksempelvis topskattegrænser), men det er oplagt, at mennesker agerer strategisk i forhold til disser tærskler og har god mulighed for at påvirke deres egen indkomst i detaljen. Det er derfor ikke rimeligt at antage, at observationernes placering omkring tærskler af den type har været tilnærmelsesvis tilfældig. Det er altså ikke alle tærskler og tærskelvariable, der giver adgang til kausal inferens via RDD.

Hvis der er gode argumenter for at tro, at observationernes placering har været tilnærmelsesvis tilfældig, og hvis empiriske tests har kunnet underbygge dette, så kan vi sammenligne observationer omkring tærskelværdien for at få et estimat af den kausale effekt. Dvs. forskelle i gennemsnit på vores afhængige variabel omkring tærskelværdien udgør vores bedste bud på den kausale effekt af den intervention, der slås til ved tærskelværdien (Green et al., 2009). Spørgsmålet er så blot, hvordan vi kan estimere denne effekt.

Statistisk modellering af RDD

Opgaven i det følgende er at give et introducerende overblik over, hvordan RDD-modeller estimeres i praksis. Detaljerede formelle præsentationer af RDD kan findes i Hahn et al. (2001), Imbens og Lemieux (2008) og Lee (2008). Udgangspunktet er en tærskelværdidummy, som er givet ved tærskelværdivariablen:

$$\textit{tærskelværdidummy} = \begin{cases} 1, \text{ tærskelvariabel} & > \text{tærskelværdi} \\ 0, \text{ tærskelvariabel} & \leq \text{tærskelværdi} \end{cases}$$

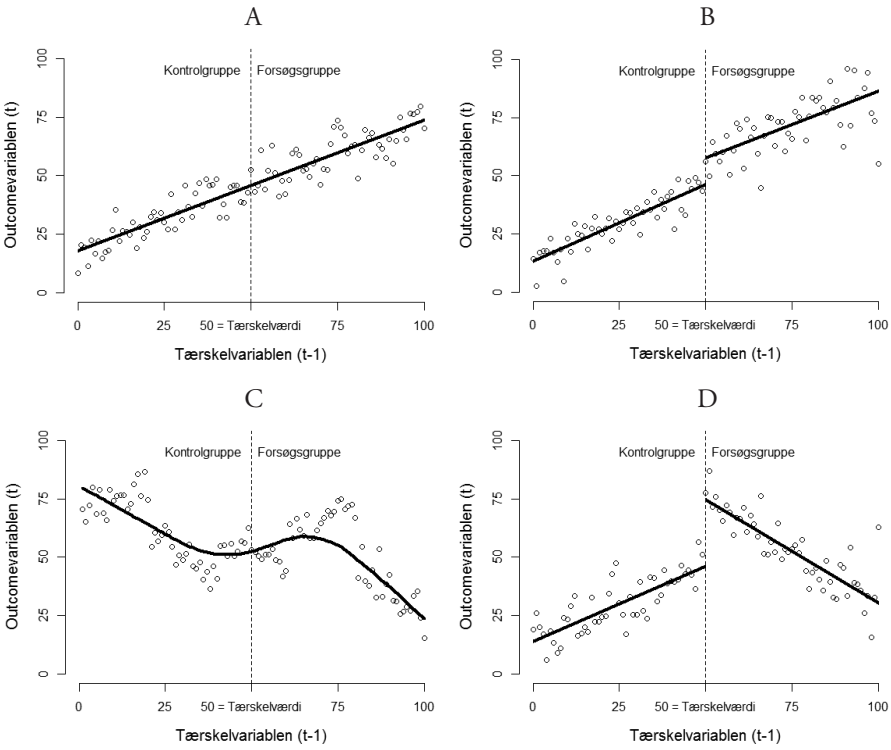
I dette tilfælde er interventionen slået til, når tærskelvariablen har værdier, der er større end tærskelværdien, mens kontrolgruppen er defineret, når tærskelvariablen antager værdier, der er mindre end eller lig med tærskelværdien. Det er vigtigt, at tærskelvariablen er målt, før den afhængige variabel bestemmes, da tærskelvariablen skal bestemme fordelingen af den intervention, som vi ønsker at estimere den kausale effekt af. Vi kan nu opskrive RDD med afsæt i denne regressionsmodel:

$$\textit{outcome} = \alpha + \beta * \textit{tærskelværdidummy} + \textit{f(tærskelvariabelen)}$$

I modellen ovenfor er den afhængige variabel en funktion af tærskelværdidummyen og tærskelvariablen. ß for tærskelværdidummien angiver RDD-estimatet af den kausale effekt. Koefficienten angiver forskydningen i værdier på den afhængige variabel ved tærskelværdien. Det er denne "diskontinuitet", vi forsøger at estimere. Visuelt er det vist i figur 1 (med inspiration fra Mellor og Mark, 1998).

I panel A ser vi udgangspunktet for RDD. Vi har tærskelvariabelen på x-aksen, som ved en given værdi tildeler observationer en kontrol- eller forsøgsstatus. I dette tilfælde er observationer med en tærskelværdi under 50 angivet som kontrolgruppe, mens observationer med en tærskelværdi med 50 eller over tilhører en forsøgsgruppe. Endelig har vi den afhængige variabel ud af y-aksen. I eksemplet forestiller vi os, at der er en positiv sammehæng mellem tærskelvariablen og den afhængige variabel. I panel A ser vi, at der ingen forskydning er i linjerne omkring tærskelvariablen, hvilket indikerer, at der ikke er nogen effekt. I panel B ser vi derimod et eksempel, hvor der er en klar forskydning efter tærskelværdien. I RDD er det i virkeligheden størrelsen på denne forskydning i linjerne, som man forsøger at estimere via β for tærskelværdidummyen. Forskellen i forskydning i linjerne på tærskelværdien udgør den kausale effekt for observationer omkring tærskelværdien (Angrist og Pischke, 2009; Dunning, 2012). I det viste eksempel vil tærskelværdidummyen være positiv svarende til

Figur 1:



en opadgående forskydning i linjen for observationerne i forsøgsgruppen over tærskelværdien.

En central udfordring i estimeringen af RDD-modeller ligger i at bestemme den funktionelle form på tærskelvariablen og det vindue af data rundt om tærskelværdien, som indgår i analysen (Dunning, 2012). Problemet består i, at vi skal separere den forskydning, vi ønsker at estimere ved tærskelværdien fra den generelle sammenhæng, der ellers måtte være mellem tærskelvariablen og den afhængige variabel. Det er ikke nogen let opgave. For det første har tærskelvariablen ikke nødvendigvis en simpel lineær sammenhæng med den afhængige variabel. At specificere den korrekte funktionelle form på tærskelværdivariablen er vigtigt for at undgå, at RDD-estimatet af den kausale effekt ikke blot afspejler ikke-linearitet omkring tærskelværdien (Angrist og Pischke, 2009: 254). En måde at gøre dette på er ved at tilføje forskellige sæt af polynomier af tærskelværdivariablen (Green et al., 2009: 405). Derved tillades forskellige ikke-linære sammenhænge mellem tærskelværdivariablen og den afhænggige

variabel. I panel C er vist et eksempel, hvor der er en ikke-lineær sammenhæng mellem tærskelvariablen og den afhængige variabel. I dette tilfælde er det tydeligt, at hvis den funktionelle form ikke blev modelleret tilstrækkeligt fleksibelt, ville man foranlediges til at tro, at der var en kausal effekt omkring tærskelværdien. Derudover bør man tillade den funktionelle form at variere på hver side af tærskelværdien. I panel D er vist et eksempel, hvor tærskelvariablens sammenhæng med den afhængige variabel varierer rundt om tærskelværdien. Det er et udtryk for, at sammenhængen mellem tærskelvariablen og den afhængige variabel ændres, idet interventionen slås til. Derfor kan det være nødvendigt at modellere en interaktion mellem tærskelvariablen og tærskelværdien.

Ud over at variere den funktionelle form for tærskelvariablen skal man også bestemme det vindue af data omkring tærskelværdien, som medtages i analysen. Dvs. man skal variere, hvor stor afstand man vil tillade, at observationer har til tærskelværdien. Herved gør man resultatet mindre overfølsomt over for, om den rigtige funktionelle form er specificeret for tærskelvariablen (Dunning, 2012). Dvs. i takt med at vi indsnævrer vinduet af data omkring tærskelværdien, bør vi kunne modellere tærskelværdivariblen stadig mere enkelt. Der er oplagt, at jo mere vi indsnævrer det vindue af data, som der analyseres på,omkring tærskelværdien, desto mere immune bliver vi over for mulige confounders og fejlspecificeret funktionelform for tærskelvariablen (Green et al., 2009: 401). Det er kun helt tæt omkring tærskelværdien, at idéen om tilnærmelsesvis tilfældighed er plausibel. Det er nemlig mere sandsynligt, at observationer, der ligger meget langt over eller under tærskelværdien, er systematisk forskellige fra hinanden (Dunning, 2012: 127). Ved at udvide vinduet kan vi derfor introducere en bias i vores estimatet af den kausale effekt, som stammer fra faktorer, der korrelerer med vores intervention. Omvendt er det ikke uden problemer at begænse vinduet af data, som der analyseres på. Ved at gøre vinduet smallere mister vi en masse data, hvilket øger usikkerheden på den estimerede effekt (Green et al., 2009: 412). En mulighed er at bruge mere avancerede ikke-parametriske estimationsmetoder, som er blevet udviklet til statistikprogrammer som R og STATA. Her benyttes oftest såkaldt lokal linear regression og procedurer til at bestemme det optimale vindue af data omkring tærskelværdien, som analyserne skal baseres på (Imbens og Lemieux, 2008; Lee og Lemieux, 2009; Imbens og Kalyanaraman, 2009). Dunning (2012: 132) har dog omvendt argumenteret for, at man altid som udgangspunkt bør estimere simple forskelle i gennemsnit for et sæt af observationer, der er placeret så tæt på tærskelværdien som muligt. Argumentet er, at hvis der er rigeligt med data omkring tærskelværdien, og hvis antagelsen om "tilnærmelsesvis tilfældig" er velbegrundet, så har vi at gøre med et lokalt eksperiment for observationer

tæt ved tærskelværdien. Af samme grund kan den kausale effekt estimeres ved at estimere en simpel forskel i gennemsnit for observationer over og under tærskelværdien. Kun i tilfælde hvor data er sparsomme tæt ved tærskelværdien, bliver vi nødsagede til at modellere den funktionelle form eller vægte observationerne efter deres nærhed til tærskelværdien for at kunne inkludere data længere væk (Dunning, 2012: 133-134). I alle tilfælde er det en god idé at estimere forskydning i linjerne omkring tærskelværdien ved så mange forskellige metoder som muligt.

Eksempler på brug af RDD i valgforskning og forvaltning

I det næste skal vi se mere specifikt på to konkrete eksempler på studier, der anvender RDD. Det ene kommer fra valgforskningen og omhandler *incumbency advantage* i den amerikanske kongres (Lee, 2008). Der er langt fra det eneste eksempel på RDD-studier af de vælgermæssige fordel ved at besidde magten (Leigh, 2008; Butler, 2009; Caughey og Sekhon, 2011; Gerber og Hopkins, 2011). Det andet eksempel stammer fra forskningen i offentlig forvaltning og politik og omhandler effekter af negative resultatmålinger på skolers performance (Chiang, 2009). Dette er blot ét blandt flere studier, der benytter RDD til at studere, effekten af resultatinformation på offentlige organisationers performance (Hemelt, 2011; Olsen, 2012, 2013a).

Eksempel 1: Incumbency advantage i den amerikanske kongres (Lee, 2008)

I amerikansk forskning har der været en stor interesse for at estimere den såkaldte incumbency advantage: dvs. er der en vælgermæssig fordel af at besidde magten? Rent deskriptivt er succesraten for genvalg i den amerikanske kongres på omkring 90 pct. Dette er dog næppe et udtryk for en ren *incumbency effect*: Politikere, der har opnået valg, er systematisk forskellige fra de kandidater, som de i sin tid vandt over. De samme kvaliteter, som gjorde det muligt for dem at vinde pladsen i første omgang, vil også påvirke deres genvalgschancer og kvaliteten af potentielle udfordrere i fremtidige valg. Nogle af disse forskelle kan måske observeres, og der kan derfor tages højde for dem i analysen, mens mange andre forhold (eksempelvis kandidatkvalitet) ikke umiddelbart kan observeres. Som tidligere nævnt foreslår Lee (2008), at tætte valg opfylder betingelserne i RDD. Ved at sammenligne vindere og tabere i tætte valg opnår vi en kontrol- og forsøgsgruppe for incumbency advantage, hvor der ikke er systematiske forskelle på hverken observerbare eller uobserverbare forhold. Ved at sammenligne valgresultater i senere valg mellem vindere og tabere opnår vi således et estimat af den kausale effekt af incumbency advantage. Lee (2008) finder på den baggrund, at der er en positiv incumbency advantage, som øger

stemmeandelen i senere valg med omkring 7-8 procentpoint. I et nyere studie har Caughey og Sekhon (2011) dog vist, at studier med RDD af tætte valg til den amerikanske kongres, herunder Lee (2008), formentlig ikke opfylder kravet om fravær af selvselektion for observationer tæt ved tærskelværdien. Således er vindere og tabere systematisk forskellige fra hinanden i selv helt tætte valg. Vindere har større politisk erfaring, flere økonomiske ressourcer, og politiske iagttagere har før valget vurderet de senere vindere som værende bedre (Caughey og Sekhon, 2011: 404). Eksemplet viser, at selv i situationer, hvor antagelserne bag RDD virker plausible, så kan nærmere undersøgelser vise, at de i realiteten ikke holder. Men det viser også, at RDD's antagelser er så relativt simple at teste, at vi hurtigt kan identificere problemer for kausal inferens via RDD. I tillæg hertil viser Eggers et al. (2013), at USA måske er en anomali, da antagelserne for RDD er opfyldt i analyser af 40.000 valg fra syv forskellige demokratier.

Eksempel 2: Effekter af negative resultatmålinger på skolers performance (Chiang, 2009)

Inden for forvaltningsvidenskaben har meget fokus været rettet mod, hvordan resultatmålinger påvirker borgere, offentlige organisationer og de offentlige ansatte. Problemet er blot, at spørgsmålet om den kausale effekt af forskellige typer performance feedback er meget kompliceret. Organisationer klarer sig på forskellig vis af gode grunde, og de selv samme grunde vil formenlig også spille ind på, hvordan organisationerne reagerer på at blive målt og vejet. Man kan for eksempel forestille sig, at en skole får at vide, at den er blandt landets dårligste målt på karakterer. De selv samme faktorer, som i første omgang fik skolen til at klare sig dårligt, kan også tænkes at pårvirke, hvordan skolen reagerer på at få at vide, at den er dårlig. Vi har hverken en teoretisk model eller de nødvendige data til at tage højde for de potentielle confounders, der kan være, når vi ønsker at slutte fra variation i resultatmålinger og til efterfølgende organisatorisk adfærd (Olsen, 2012, 2013a).

Chiang (2009) foreslår RDD som en løsning til at estimere kausale effekter af negative resultatmålinger på offentlige organisationers adfærd i en amerikansk kontekst. I staten Florida tildeles skolerne en overordnet bogstavskarakter på baggrund af tærskelværdier i et samlet kontinuiert mål for elevernes resultater. Hvis skolerne på denne tærskelvariabel opnår mindre end 280 point, så tildeles de et skamfuldt "F", som er rangslistens dårligste kategori. Hvis en skole får over 280 point tildeles den karakteren "D". Ved at sammenligne skoler lige omkring tærskelværdien på 280 point, opnår Chiang eksogen variation i de overordnede resultatmål, som den enkelte skole har opnået. Hvis en skole

får karakteren "F" får den blandt andet trusler om en lang række sanktioner, hvis ikke resultaternes forbedres. F-skoler har derfor stærk tilsyndelse til at forbedre sig. Det er ikke umiddelbar plausibelt, at skolerne kan selvselektere sig ind på en mere fordelagtig karakter i nærheden af tærsklen. Skolerne omkring tærskelværdien er derfor af samme grund ens på en lang række observerbare forhold. Således kan Chiang (2009) vise, at truslen om sanktioner sætter sig igennem via bedre resultater i matematik til børnene i F-skoler. Han viser ligeledes, at F-skoler i højere grad end D-skoler investerer i undervisningsteknologi og efteruddannelse af lærekorpset. Studiet er et godt eksempel på, hvordan RDD kan hjælpe med at besvare et ellers kompliceret kausalt spørgsmål i studiet af offentlig forvaltning og politik.

Begrænsning og udfordringer for RDD i statskundskab – og nogle løsninger

Afslutningsvis er det værd at overveje, hvilke barrierer og udfordringer der kan være for, at statskundskaben kan benytte RDD.

For det første er det oplagt, at politiske og administrative aktører med de rette incitamenter vil gøre alt for at placere sig strategisk omkring vigtige tærskelværdier. Dette er et problem i forhold til den helt centrale antagelse om, at observationerne ikke har mulighed for at selvselektere sig "til rette" omkring tærskelværdier (Cook, 2008). For politologer er der derfor en særlig pligt til både kvalitativt og kvantitativt, empirisk at vise, at tærskelværdier ikke bliver strategisk manipuleret af de undersøgte observationer eller af andre med interesser i placeringen omkring tærskelværdien.

For det andet kan der i politisk-administrative sammenhænge være problemer med at opnå eksakt viden om tærskelværdier og tærskelvariable. Typisk vil disse information ikke være offentligt tilgængelige, og det vil kræve tålmodighed at grave dem frem, hvis det overhovedet er muligt. Nogle gange kan tærskelværdier have mere uformel karakter, eller aktører kan have en interesse i at holde dem for sig selv. Et tilgrænsende problem kan være, at der i politisk-administrative sammenhænge kan være tale om uklare og derfor ikke-deterministiske tærskelværdier, hvilket betyder, at man må ty til fuzzy RDD. For politologer vil det derfor ofte kræve indgående empiriske viden og tålmodighed at identificere de nødvendige informationer, for at RDD kan implementeres korrekt.

For det tredje er RDD kun i stand til at estimere en lokal kausal effekt, der gælder for observationer omkring tærskelværdien (Imbens og Lemieux, 2008: 621). Dvs. at RDD's høje interne validitet og få antagelser har den omkostning, at estimatet af den kausale effekt er begrænset til den subpopulation af data,

som ligger omkring tærskelværdien. Udfordringen for politologer er derfor at beskrive, i hvilket omfang estimatet af den kausale effekt kan forventes at se ud for observationer længere væk fra tærskelværdien. Hvad kan vi forvente, at der vil ske med den kausale effekt i takt med, at vi bevæger os væk fra tærskelværdien? Man bør altid overveje om særlige forhold gør sig gældende for observationer omkring tærskelværdien. Dette er dog ikke væsensforskelligt fra de fleste eksperimentelle designs, hvor estimater altid vil være lokale i forhold til en særlig population, kontekst eller tid (Linden et al., 2006; Angrist og Pischke, 2009).

For det fjerde er RDD relativt inefficient og kræver derfor store mængder data for at kunne identificere signifikante effekter. Dette skyldes, at modellerne oftest kun estimeres på data med en given afstand til tærskelværdien, og at tærskelværdiindikatoren og tærskelværiablen vil være stærkt korrelerede, hvilket vil forstørre standardfejlene (Green et al., 2009). Flere har blandt andet vist, at et typisk RDD-studie kræver tre til fire gange så meget data for at vise samme effektstørrelse som i et tilsvarende traditionelt eksperiment (Goldberger, 1972; Schochet, 2009). Politologer må derfor tænke i alternative datafomer som paneldata eller multi-level data, hvilket kan give adgang til flere observationer (Chiang, 2009; Pennell et al., 2011)

Konklusion

Denne artikel har givet en introduktion till RDD for statskundskabere. Med RDD har vi et stærkt design, som med få og testbare antagelser gør det muligt at besvare kausale spørgsmål uden aktiv intervention. På mange måder er RDD lige så intuitivt og simpelt som det traditionelle eksperiment, men vi undgår de etiske og politiske problemer, som forkserintervention oftest kan medføre i statskundskaben (Campbell, 1969). Med RDD er sammenligningen af observationer omkring tærskelværdier sat i stedet for aktiv forskerintervention via lodtrækning som middel til opdeling af observationer i kontrol- og forsøgsgrupper. Sammenligninger med tilsvarende kontrollerede eksperimenter har vist, at RDD har en forbløffende evne til at fremkomme med sammenlignelige estimater (der dog er en del mindre efficiente). Endelig er RDD et af de meget få designs til kausale analyser, som udspringer af socialvidenskaberne (Thistlethwaite og Campbell, 1960; Cook, 2008).

Der er derfor god grund til at se nærmere på RDD for statskundskaben. Den virkelige udfordring ligger i at finde egnede empiriske situationer. Det er i virkeligheden her, at forskerens primære arbejdsopgave forskyder sig hen i brugen af RDD. Er der tærskelvariable og tærskelværdier, som inddeler de observationer, man ønsker at studere, i kontrol- og eksperimentgrupper? Er det

kvalitativt plausibelt, at observationernes placering i nærheden af tærskelværdien har været tilnærmelsesvis tilfældig? Og kan man vise, at observationerne i nærheden af tærskelværdien faktisk ikke er systematisk forskellige på observerbare confounders? Der er de spørgsmål, som alle, der ønsker at anvende RDD til kausale spørgsmål, først og fremmest må beskæftige sig med.

Note

1. Stor tak til redaktørerne Mogens Kamp Justesen, Robert Klemmensen og Kim Mannemar Sønderskov samt to anonyme læsere for at komme med meget konstruktive kommentarer til manus. Også tak til Ulrik Hvidman for nogle gode ændringsforslag. Eventuelt tilbageværende uklarheder og fejl er mine egne alene.

Litteratur

Aiken, Leona S., Stephen G. West, David E. Schwalm, James L. Carroll og Shenghwa Hsiung (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation. *Evaluation Review* 22 (2): 207-244.

Almond, Douglas og Joseph J. Doyle (2011). After midnight: A regression discontinuity design in length of postpartum hospital stays. *American Economic Journal: Economic Policy* 3 (3): 1-34.

Angrist, Joshua D. og Jörn-Steffen Pischke (2009). *Mostly Harmless Econometrics*. Princeton, MA: Princeton University Press.

Becker, Sascha O., Peter H. Egger og Maximilian von Ehrlich (2010). Going NUTS: The effect of EU Structural Funds on regional performance. *Journal of Public Economics* 94 (9-10): 578-590.

Berk, Richard, Geoffrey Barnes, Lindsay Ahlman og Ellen Kurtz (2010). When second best is good enough: a comparison between a true experiment and a regression discontinuity quasi-experiment. *Journal of Experimental Criminology* 6 (2): 191-208.

Black, Sandra E. (1999). Do Better Schools Matter? Parental Valuation of Elementary Education. *The Quarterly Journal of Economics* 114 (2): 577-599.

Broockman, David E. (2009). Do Congressional Candidates Have Reverse Coattails? Evidence from a Regression Discontinuity Design. *Political Analysis* 17 (4): 418-434.

Butler, Daniel M. (2009). A regression discontinuity design analysis of the incumbency advantage and tenure in the U.S. House. *Electoral Studies* 28 (1): 123-8.

Campbell, Donald T. (1969). Reforms as experiments. *American Psychologist* 24 (4): 409-429.

Caughey, Devin og Jasjeet S. Sekhon (2011). Elections and the regression discontinuity design: Lessons from close US house races, 1942-2008. *Political Analysis* 19(4): 385-408.

Chen, Yuyu, Avraham Ebenstein, Michael Greenstone og Li Hongbin (2013). Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy. *Proceeding of the National Academy of Science* 110 (32): 12936-12941.

Chiang, Hanley (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics* 93 (9-10): 1045-1057.

Cook, Thomas (2008). Waiting for Life to Arrive? A history of the regression-discontinuity design in Psychology, Statistics and Economics. *Journal of Econometrics* 142 (2): 636-654.

Dunning, Thad (2012). *Natural Experiments in the Social Sciences. A Design-Based Approach*. Cambridge, UK: Cambridge University Press.

Eggers, Andrew, Ole Folke, Anthony Fowler, Jens Hainmueller; Andrew B. Hall, James M. Snyder (2013). On the validity of the regression discontinuity design for estimating electoral effects: New evidence from over 40,000 close races. *Working paper*.

Elis, Roy, Neil Malhotra og Marc Meredith (2009). Apportionment cycles as natural experiments. *Political Analysis* 17 (4): 358-376.

Freedman, David A. (1991). Statistical models and shoe leather. *Sociological Methodology* 21: 291-313.

Gerber, Elisabeth R. og Daniel J. Hopkins (2011). When mayors matter: Estimating the impact of mayoral partisanship on city policy. *American Journal of Political Science* 55 (2): 326-339.

Goldberger, Arthur S. (1972). Selection bias in evaluating treatment effects: The case of interaction. Technical report, Institute for Research on Poverty, University of Wisconsin, Madison.

Green, Donald P., Terence Y. Leong, Holger L. Kern, Alan S. Gerber og Christopher W. Larimer (2009). Testing the Accuracy of Regression Discontinuity Analysis Using Experimental Benchmarks. *Political Analysis* 17 (4): 400-417.

Hahn, Jinyong, Petra Todd og Wilbert Van der Klaauw. (2001). Identifcation and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 69 (1): 201-209.

Hemelt, Steven W. (2011). Performance effects of failure to make Adequate Yearly Progress (AYP): Evidence from a regression discontinuity framework. *Economics of Education Review* 30 (4): 702-723.

Hopkins, Daniel J. (2011). Translating into votes: The electoral impacts of Spanish-language ballots. *American Journal of Political Science* 55 (4): 814-830.

Imbens, Guido og Karthik Kalyanaraman (2009). Optimal bandwidth choice for the regression discontinuity estimator. Unpublished manuscript, Department of Economics, Harvard University.

Imbens, Guido og Thomas Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics* 142 (2): 615-635.

Keele, Luke og Rocio Titiunik (2013). Natural experiments based on geography. *Working paper*.

Klaauw, Wilbert van der (2002). Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *International Economic Review* 43 (4): 1249-1287

Lee, David S. (2008). Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics* 142 (2): 675-697.

Lee, David S. og Thomas Lemieux (2009). Regression discontinuity designs in economics. *National Bureau of Economic Research Working Paper*.

Leigh, Andrew (2008). Estimating the impact of gubernatorial partisanship on policy settings and economic outcomes: A regression discontinuity approach. *European Journal of Political Economy* 24 (1): 256-268.

Linden, Ariel, John L. Adams og Nancy Roberts (2006). Evaluating disease management programme effectiveness: an introduction to the regression discontinuity design. *Journal of Evaluation in Clinical Practice* 12 (2): 124-131.

Ludwig, Jens og Douglas L. Miller (2007). Does head start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics* 122 (1): 159-208.

McCray, Justin (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics* 142 (2): 698-714.

Mellor, Steven og Melvin M. Mark (1998). A quasi-experimental design for studies on the impact of administrative decisions: Applications and extensions of the regression- discontinuity design. *Organizational Research Methods* 1 (3): 315-333.

Morgan, Stephen L. og Christopher Winship (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press.

Olsen, Asmus L. (2012). Regression discontinuity designs in public administration: The case of performance measurement research. Paper, ECPR Joint Sessions, Antwerpen, 11.-15. april.

Olsen, Asmus L. (2013a). Naming bad performance: Can performance disclosure drive improvements? Revise and resubmit. *Journal of Public Administration Research and Theory*.

Olsen, Asmus L. (2013b). Incumbency advantage and tax and spending effects of a unified leadership: A regression discontinuity design among Danish municipalities. Ufærdigt manus.

Papay, John P., John B. Willett og Ricard J. Murnane (2011). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics* 161 (2): 203-207.

Pennell, Micheal, Erinn M. Hade, David M. Murray og Dale A. Rhoda (2011). Cutoff designs for community-based intervention studies. *Statistics in medicine* 30 (15): 1865-1882.

Pettersson-Lidbom, Per (2008). Do parties matter for economic outcomes? A regression-discontinuity approach. *Journal of the European Economic Association* 6 (5): 1037-1056.

Robinson, Gregory, John E. McNulty og Jonathan S. Krasno (2009). Observing the counterfactual? The search for political experiments in nature. *Political Analysis* 17 (4): 341-357.

Rubin, Donald B. (1977). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Statistics* 2 (1): 1-26.

Schochet, Peter Z. (2009). Statistical power for regression discontinuity designs in education evaluations. *Journal of Educational and Behavioral Statistics* 34 (2): 238-266.

Sekhon, Jasjeet S. (2010). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science* 12: 487-508.

Shadish, William R. (2011). The truth about validity. *New Directions for Evaluation* (130): 107-117.

Shadish, William R., Thomas D. Cook og Donald T. Campbell (2001). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Stamford, Connecticut: Cengage Learning (2. udgave).

Thistlethwaite, Donald L. og Donald T. Campbell. (1960). Regression-discontinuity analysis: An alternative to the ex post factor experiment. *The Journal of Educational Psychology* 51 (6): 309-317.

politica, 46. årg. nr. 1 2014, 60-78

Mogens K. Justesen og Robert Klemmensen

Sammenligning af sammenlignelige observationer: kausalitet, matching og observationsdata[1]

Artiklen giver en introduktion til matching og diskuterer metodens styrker og svagheder i studier af kausalitet. Matching er primært en metode, der kan anvendes til at gøre observationer så sammenlignelige som muligt på observerbare forhold. Matching løser ikke i sig selv de problemer, der er forbundet med at drage kausal inferens med observationsdata, men kan bringe os et skridt i den retning, hvis den kombineres med et stærkt design, fx et naturligt eksperiment eller et kvasi-eksperiment. Artiklen giver et eksempel på, hvordan matching kan bruges til analyse af kvasi-eksperimentelle data. Til dette formål bruger vi surveydata til at analysere, hvordan en miljøkatastrofe forårsaget af en uanticiperet eksplosion på olieboreplatformen Deep Water Horizon påvirkede briternes holdninger til miljøspørgsmål.

Et af de vanskeligste spørgsmål i samfundsvidenskaben er, hvordan vi identificerer kausale effekter med data, som ikke er genereret af en eksperimentel proces, men af processer i "virkeligheden". I sin mest simple form er udgangspunktet for analyser af observationsdata, at vi sammenligner observerbare forhold for én gruppe med observerbare forhold for en anden gruppe. I modsætning til eksperimentelt genererede data, er observationsdata imidlertid karakteriseret ved, at det ikke er tilfældigt, om observationerne befinder sig i den ene eller den anden gruppe (Blom-Hansen og Serritzlew, 2014). Dette giver anledning til problemer, hvis vi ønsker at besvare kausale spørgsmål.

Eksempelvis undersøger Kam og Palmer (2008) effekten af uddannelse på politisk deltagelse. Udgangspunktet i Kam og Palmers artikel er, at tidligere resultater, der viser en tæt sammenhæng mellem uddannelse og politisk deltagelse, er problematiske pga. forhold relateret til såkaldt "selvselektion" ind i uddannelsessystemet.[2] Kam og Palmer argumenterer for, at en del af effekten af uddannelse er et resultat af social baggrund – fx forældres uddannelse – som disponerer individer til at (fra)vælge videregående uddannelser. Det betyder, at der i udgangspunktet er systematiske forskelle på dem, der tager en videregående uddannelse, og dem, der ikke gør. Derfor er det vanskelligt at sige, om en sammenhæng mellem uddannelse og politisk deltagelse er udtryk for en uddannelseseffekt – eller om den er et resultat af fx social baggrund. Kam og

Palmers analyser antyder, at uddannelse ikke har den store effekt på politisk deltagelse, når individer matches på variable som eksempelvis forældres uddannelse.

Den mest effektive løsning på sådanne selvselektionsproblemer er det eksperimentelle design, hvor randomiseringsproceduren sikrer, at de grupper, vi sammenligner, i gennemsnit er ens (Blom-Hansen og Serritzlew, 2014). Imidlertid står vi som politologer ofte i situationer, hvor vi interesserer os for kausale relationer, der vanskelligt kan – eller slet ikke bør – undersøges eksperimentelt. Eksempelvis kan (eller bør) vi ikke tildele regeringskoalitioner, borgerkrig, demokrati, handelspolitikker og finansielle kriser tilfældigt til lande, ligesom forhold som social baggrund og etnisk oprindelse vanskeligt kan (eller bør) manipuleres eksperimentelt. Har vi ikke et stærkt design eller en valid instrumentel variabel (Hariri, 2014), kan vi ofte ikke gøre andet end at forsøge at gøre de observationer, vi sammenligner, så sammenlignelige som muligt.

Matching er en metode, der forsøger at gøre dette. Matching er i stigende grad blevet et populært redskab inden for politologi (Kam og Palmer, 2008; Sekhon, 2009; Boyd, Epstein og Martin, 2010; Dinesen, 2012; Justesen, 2012), økonomi (Persson og Tabellini, 2003; Dehejia og Wahba, 1999, 2002), sociologi (Harding, 2002) og programevaluering (Imbens og Wooldridge, 2008; Khanker, Koolwal og Samad, 2009) og behandles rutinemæssigt i introducerende tekster om kausal inferens (Morgan og Winship, 2007; Imbens og Wooldridge, 2008; Angrist og Pischke, 2009).

Med matching forsøger man at simulere det eksperimentelle design ved at konstruere to (eller flere) grupper, der er så sammenlignelige som muligt på alle relevante og observerbare karakteristika – bortset fra den kausale variabel, man interesserer sig for (Ho et al., 2007; Morgan og Winship, 2007).[3] Idéen bag matching er simpel: Hver observation fra den ene gruppe matches med en eller flere ensartede observationer fra den anden gruppe, hvorefter de to gruppers forskel på den afhængige variabel beregnes. Intuitivt kan matching således opfattes som en kvantitativ version af det velkendte *most similar systems design*, som er udbredt i casestudiemetodologien (Sekhon, 2009).

Den primære styrke ved matching er, at metoden eksplicit kan bruges til at forbedre sammenligneligheden mellem observationer i data, således at empiriske resultater bygger på sammenligninger af sammenlignelige observationer. Dermed begrænses analysen til den delmængde af data, hvor der er sammenlignelige observationer, mens betydningen af observationer, der ikke har et godt sammenligningsgrundlag, ignoreres eller nedjusteres. Matching giver imidlertid ikke en magisk løsning på de vanskeligheder, der er forbundet med at isolere kausale effekter. Metodens primære svaghed er, at den sjældent *i sig*

selv kan bruges til at identificere kausale effekter. Identifikation af kausale effekter er kun mulig under antagelse af, at modellen er specificeret korrekt på baggrund af observerbare variable, og at relevante uobserverede variable ikke er udeladt. Dette understreger den generelle pointe, at kausalitetsproblemer grundlæggende kun kan imødegås ved hjælp af et stærkt design – fx et naturligt eksperiment – og ikke ved hjælp af statistik kontrol og teknik (Sekhon, 2009; Dunning, 2012).

Formålet med artiklen er at give en introduktion til matching og herunder diskutere styrker og svagheder ved metoden som redskab til at studere kausalitet med observationsdata. Vi fokuserer på den variant af matching, der kaldes *propensity score matching*. Vi begrænser artiklen til en diskussion af matching i forbindelse med tværsnitsdata.[4]

Resten af artiklen er organiseret som følger. Det næste afsnit giver en introduktion til matching, hvorefter vi diskuterer hvorvidt – og under hvilke betingelser – matching kan bidrage til at identificere kausale effekter. Herefter giver vi et empirisk eksempel på, hvordan matching kan anvendes i kombination med et kvasi-eksperimentelt design. Det sidste afsnit konkluderer.

Matching

Matching baserer undersøgelser af kausale effekter på at finde observationer, der er sammenlignelige på relevante, observerbare variable, bortset fra den kausale variabel, man interesserer sig for. Vi bruger her betegnelsen T for *kausalvariablen* – dvs. den variabel, hvis effekt på den *afhængige variabel*, Y, vi ønsker at identificere. I et matching setup er T oftest (men ikke altid) en binær variabel med to grupper, fx høj ($T = 1$) og lav ($T = 0$) uddannelse. For at relatere diskussionen til det eksperimentelle design kalder vi den første gruppe for "eksperimentgruppen" ($T = 1$) og den anden gruppe for "kontrolgruppen" ($T = 0$). X betegner sættet af observerbare *uafhængige variable*, som bruges til at matche observationer. Dette er variable, der både påvirker – eller skaber "selektion ind i" – den kausale variabel, T, og som potentielt påvirker den afhængige variabel, Y. Ved anvendelse af matching søger vi således at finde (par af) observationer, der er så ens som muligt på de uafhængige variable, X, bortset fra at den ene observation befinder sig i eksperimentgruppen ($T = 1$), mens den anden er i kontrolgruppen ($T = 0$), hvorefter forskellen i de to gruppers gennemsnit på den afhængige variabel, Y, estimeres.

Selvom matching og OLS-regression er nært relaterede teknikker (Morgan og Winship, 2007; Angrist og Pischke, 2009), kan der være fordele ved at bruge matching som udgangspunkt. For det første bliver vi med matching eksplicit konfronteret med spørgsmålet om graden af "overlap" og "balance" mellem

eksperiment- og kontrolgruppen på de variable, vi bruger til at matche. Dette er med til at sikre, at vi sammenligner observationer fra eksperimentgruppen med sammenlignelige observationer i kontrolgruppen (Gelman og Hill, 2007: 208-211; Ho et al., 2007). Matching baserer sig således på "lokale" sammenligninger af ensartede observationer i den forstand, at observationer fra eksperimentgruppen sammenlignes med deres nærmeste "tvilling(er)" i kontrolgruppen (Cameron og Trivedi, 2005: 871; Persson og Tabellini, 2003: 139).

En anden fordel ved matching er, at vi fokuserer på at modellere selektionsprocessen (Harding, 2003, 678; Angrist og Pischke, 2009: 84). Snarere end at modellere årsagerne til den afhængige variabel, modellerer vi "årsagerne til årsagen". Det betyder, at matching fokuserer på at modellere årsagerne til kausalvariablen, *T*, med udgangspunkt i et sæt af uafhængige variable, *X*. Det er fordelagtigt i tilfælde, hvor vi har bedre teori og viden om de processer og variable, der påvirker den kausale variable, *T*, end de processer og variable, der påvirker den afhængige variabel, *Y*. Imidlertid er det vigtigt at understrege, at matching ikke "løser" selvselektionsproblemer, men kan bidrage til at fokusere vores opmærksomhed på disse.[5] I det følgende fokuserer vi på den variant af matching, der kaldes *propensity score matching*. I praksis kan denne form for matching implementeres i tre trin (Khanker, Koolwal og Samad, 2009), som vi gennemgår nedenfor.

Sandsynlighedsmodellen

Hvis man som Kam og Palmer (2008) vil sammenligne forskelle i politisk deltagelse for individer med høj og lav uddannelse, er den mest intuitive måde at sammenligne observationer på at lave såkaldt "eksakt" matching, hvor observationer i eksperimentgruppen (højt uddannede) matches med observationer i kontrolgruppen (lavt uddannede), der har den eksakt samme værdi på en tredje, uafhængig variabel (fx forældres uddannelse). Denne tilgang bliver dog hurtigt problematisk pga. det såkaldte dimensionalitetsproblem. Problemet opstår, hvis man vil matche på mange uafhængige variable (Smith og Todd, 2005; Sekhon, 2009: 497). I dette tilfælde bliver det vanskelligt at finde observationer i eksperiment- og kontrolgruppen med de eksakt samme værdier på det mangedimensionelle sæt af uafhængige variable, særligt hvis disse er kontinuerte.

I en indflydelsesrig artikel viste Rosenbaum og Rubin (1983) imidlertid, at dimensionalitetsproblemet kan imødegås ved at matche på *sandsynlighedsscorer*. Sandsynlighedsscoren er defineret som $p(T = 1|X)$, dvs., sandsynligheden (p) for at en observation befinder sig i eksperimentgruppen ($T = 1$), givet de uafhængige variable, X (Cameron og Trivedi, 2005: 873; Guo og Fraser, 2010).

Sandsynlighedsscoren varierer per definition mellem 0 og 1. Idéen bag matching på sandsynlighedsscoren er simpel. I stedet for at sammenligne observationer på mange forskellige variable, sammenligner man i stedet på en endimensionel variabel – sandsynlighedsscoren – som opsummerer informationen om de uafhængige variable, *X*. Dvs., at man her forsøger at modellere "selektionen ind i" eksperiment- og kontrolgrupperne.

I praksis estimerer man en sandsynlighedsmodel – fx en logistisk regressionsmodel – hvor kausalvariablen, *T*, fungerer som den afhængige variabel, der modelleres som en funktion af de uafhængige variable, *X*. Disse uafhængige variable bør inkludere alle variable, der påvirker både kausalvariablen og den afhængige variabel, men ikke mellemkommende variable, der er en funktion af kausalvariablen (Ho et al., 2007: 216). Når sandsynlighedsmodellen er specificeret, kan man generere sandsynlighedsscoren. Dette gøres i praksis ved at estimere de *forudsagte sandsynligheder* fra den logistiske regression og gemme dem som en ny variabel. Det er denne variabel med sandsynlighedsscorer, der bruges til at matche observationer fra eksperimentgruppen med (nogenlunde) ensartede observationer fra kontrolgruppen. Matching på sandsynlighedsscoren betyder således, at man sammenligner observationer fra eksperiment- og kontrolgruppen, der har ensartede sandsynlighedsscorer.

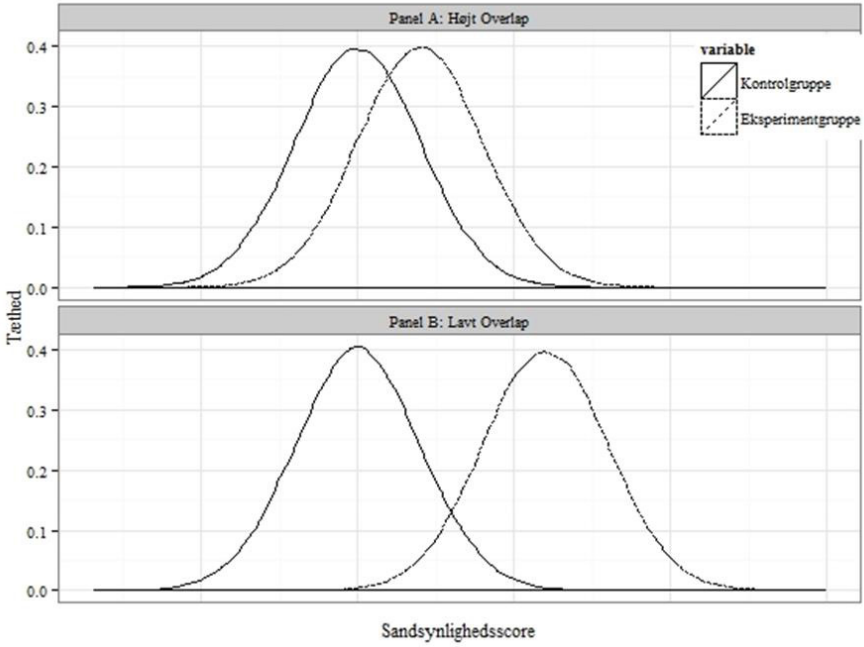
Overlap og balance

En fordel ved matching i forhold til OLS-regression er, at vi eksplicit undersøger graden af overlap (*common support*) mellem eksperiment- og kontrolgruppen på sandsynlighedsscoren (Cameron og Trivedi, 2005: 864; Khanker, Koolwal og Samad, 2009). Dette er vigtigt for at sikre, at der for observationer i eksperimentgruppen er et godt match i kontrolgruppen (og omvendt). Dette kan illustreres ved hjælp af figur 1.

Figur 1 viser fordelingerne for sandsynlighedsscoren for henholdsvis kontrol- og eksperimentgruppen i to hypotetiske scenarier. Overlapsregionen defineres ofte lidt forskelligt i litteraturen. Morgan og Winship (2007: 117) definerer overlapsregionen som det interval på sandsynlighedsscoren, der ligger imellem minimum- og maksimumværdien for kontrolgruppen, mens Persson og Tabellini (2003: 143) definerer overlap som intervallet imellem minimumsværdien for eksperimentgruppen og maksimumsværdien for kontrolgruppen.

Panel A viser en situation med en høj grad af overlap mellem observationerne i eksperiment- og kontrolgruppen. Bruger vi Persson og Tabellinis (2003) definition, er der et relativt stort interval, hvor fordelingerne er overlappende, og hvor observationer i eksperimentgruppen har sammenlignelige observationer i kontrolgruppen. I områderne uden for intervallet, er der ikke overlap.

Figur 1: Intervaller i data med højt og lavt overlap



Eksempelvis vil observationer i eksperimentgruppen med meget høje værdier på sandsynlighedsscoren – langt ude i "højre hale" af fordelingen – ikke have et godt sammenligningsgrundlag blandt observationer i kontrolgruppen. Netop fordi sådanne observationer ikke har et godt sammenligningsgrundlag, bliver de ekskluderet i den videre analyse. En vigtig pointe er her, at man ved at droppe observationer, der falder uden for overlapsregionen, gør eksperiment- og kontrolgrupperne mere homogene, hvilket potentielt reducerer bias i estimaterne (Sekhon, 2009: 495-496). Det er ligeledes vigtigt at pointere, at man *ikke* udvælger observationer baseret på den afhængige variabel, *Y*, som ikke spiller nogen rolle i denne fase af analysen, men alene på baggrund af sandsynlighedsscoren (Ho et al., 2007: 212). En vigtig del af øvelsen i matching er således at "trimme" data – i øvrigt ligesom regressions-diskontinuitetsdesignet (Olsen, 2014) – for at øge sammenligneligheden mellem observationer.

Panel B viser en situation, hvor der er en lav grad af overlap mellem fordelingerne for eksperiment- og kontrolgruppen. I dette tilfælde er der kun få observationer i eksperimentgruppen, der har sammenlignelige observationer i kontrolgruppen. En sådan mangel på gode matches er ofte tilsløret i lineær

regression (Harding, 2003). Hvis der er ringe overlap mellem eksperiment- og kontrolgruppen, vil en OLS-regression stadig estimere parametre ved at ekstrapolere lineært igennem det interval i data, der ikke indeholder observationer fra både eksperiment- og kontrolgruppen (Ho et al., 2007, 210-211). En fordel ved matching er således, at vi eksplicit bliver nødt til at forholde os til, om der er overlap mellem eksperiment- og kontrolgruppen, før vi estimerer effekten af den kausale variabel. Dermed bliver de observationer, vi sammenligner, mere sammenlignelige – i hvert fald på observerbare variable.

Eftersom pointen med matching er at skabe to grupper, der er så ens som muligt, er det vigtigt at undersøge, om eksperiment- og kontrolgrupperne er "balancerede" på de uafhængige variable, der ligger til grund for sandsynlighedsscoren. At eksperiment- og kontrolgrupperne er balancerede betyder, at deres fordelinger på de uafhængige variable, *X*, er ensartede (Morgan og Winship, 2007: 114; Ho et al., 2007: 221). Hvis det er lykkes at skabe to ensartede grupper, vil de observerede forskelle på de uafhængige variable således være små.

Balance undersøges nogle gange ved at teste, om fordelingerne for de uafhængige variable er ens for eksperiment- og kontrolgrupperne efter matching-proceduren. Andre gange inddeles observationer i intervaller (fx kvartiler) på sandsynlighedsscoren, hvorefter balance testes inden for hvert interval, således at observationer med ensartede sandsynlighedsscorer sammenlignes (Persson og Tabellini, 2003: 143-148; Khanker, Koolwal og Samad, 2009).

Det formentlig mest anvendte redskab til at undersøge balance er *t*-testen, der bruges til at teste, om gennemsnittet er ens for eksperiment- og kontrolgruppen på hver uafhængig variabel. Dette gøres ofte ved – for observationer inden for overlapsregionen – at sammenligne gennemsnittene for eksperiment- og kontrolgruppen før og efter matching. Forskellen i gennemsnit før matching er blot den rå forskel mellem eksperiment- og kontrolgruppernes gennemsnit på de uafhængige variable. Forskellen i gennemsnittet efter matching udregnes ofte som en sammenligning af gennemsnittet for eksperimentgruppen med et vægtet gennemsnit for kontrolgruppen, hvor vægten eksempelvis er givet ved antallet af gange, en observation i kontrolgruppen bruges som match for en observation i eksperimentgruppen.[6] Hvis data er balancerede, vil der ikke være forskelle i gennemsnittet for eksperiment- og kontrolgrupperne efter matching – eller i hvert fald vil forskellene blive mindre efter matching.

Problemet med *t-testen* er, at selvom gennemsnittene er ens, behøver selve fordelingerne for variablene ikke at være det. Derfor kan det være en god idé i stedet at bruge et QQ-plot (Ho et al., 2007: 221-222), som er velegnet til at undersøge, om fordelingerne (og ikke blot gennemsnittet) for en given variabel

er ens for de to grupper. Ligeledes kan den såkaldte Kolmogorov-Smirnov test bruges til at teste, om fordelinger for grupperne er ens før og efter matching.

Med observationsdata kan det dog være vanskelligt at opnå balance på alle variable. Fx kan det være vanskelligt at balancere dummyvariable, hvis der er få observationer i den ene kategori. Hvis nogle variable er ubalancerede, kan bedre balance opnås ved at specificere sandsynlighedsmodellen anderledes, fx ved at inkludere flere variable, interaktionsled eller kvadrede led på højresiden (Morgan og Winship, 2007: 115).

Sammenligning af matchede observationer

Det, vi i sidste ende er interesserede i, er at undersøge, om kausalvariablen, *T*, har en effekt på den afhængige variabel, *Y*. Derfor beregnes forskellen i de matchede gruppers gennemsnit på den afhængige variabel. Der findes flere forskellige algoritmer til dette formål. Her vil vi ikke gennemgå dem alle (se fx Cameron og Trivedi, 2005: 874-876; Morgan og Winship, 2007: 104-109; Guo og Fraser, 2010), men blot nævne to almindelige algoritmer.

En af de hyppigst anvendte algoritmer er nearest neighbor matching. Pointen er her – som navnet antyder – at hver observation i eksperimentgruppen sammenlignes med dens "nærmeste nabo" i kontrolgruppen. Afstanden mellem observationer er her givet ved sandsynlighedsscoren, således at den nærmeste nabo ideelt har en lille afstand på denne score. Et væsentligt spørgsmål er her, hvor mange observationer i kontrolgruppen man bruger som sammenligningsgrundlag for hver observation i eksperimentgruppen. Dette er vigtigt, fordi valget af antallet af matches indebærer en afvejning af bias og præcision (Dehejia og Wahba, 2002: 153; Cameron og Trivedi, 2005: 873-874; Morgan og Winship, 2007: 105-109). At matche med én nærmeste nabo (1:1 matching) har den fordel, at man sammenligner observationer med den mindst mulige afstand på sandsynlighedsscoren – dvs. de mest ensartede observationer – hvilket reducerer bias i estimaterne. Matcher man derimod med et antal, *n*, nærmeste naboer (1:*n* matching), opnår man mere præcise estimater (mindre varians), men samtidig øges afstanden på sandsynlighedsscoren mellem de observationer, man sammenligner, hvilket kan skabe større bias. Med andre ord risikerer man, at sammenligningsgrundlaget bliver mindre homogent, hvis man matcher med flere nærmeste naboer.[7]

Imidlertid kan der være situationer, hvor den nærmeste nabo er langt væk på sandsynlighedsscoren og således udgør et dårligt match. *Radius-matching* er en algoritme, der imødegår dette problem. I praksis definerer man en radius – eller maksimal afstand – på sandsynlighedsscoren, fx 0,05. Hver observation i eksperimentgruppen matches herefter med alle observationer i kontrolgrup-

pen, der falder inden for den definerede radius. Dermed bliver observationer i eksperimentgruppen kun sammenlignet med kontrolobservationer i det "nære nabolag". Dette kan være en fordel, hvis der er flere gode matches inden for en lille radius på sandsynlighedsscoren, mens ulempen er, at der kan være observationer i eksperimentgruppen, der ikke har matchende observationer i kontrolgruppen inden for den definerede radius. Størrelsen af radius involverer også et trade-off mellem bias og præcision. Eftersom sandsynlighedsscoren per definition falder mellem 0 og 1, vil en radius på 0,01 eller 0,02 være ganske lille og betyde, at man matcher med færre observationer, som til gengæld har sandsynlighedsscorer, der ligger tæt på observationerne i eksperimentgruppen. Dette giver mindre bias men også større varians (mindre præcision). Omvendt vil en større radius (eksempelvis 0,1) betyde, at man bruger flere, men mindre ensartede observationer i kontrolgruppen som matches, hvilket øger bias, men giver mindre varians (større præcision).

Det er vanskelligt generelt at sige, hvilken algoritme der er "bedst". Som fremhævet af Morgan og Winship (2007: 109), må pointen med matching dog være at mindske bias, hvorfor algoritmer, der gør det vanskelligt at generere et dårligt match, er at foretrække. Eksempelvis kan nearest neighbor matching være problematisk, hvis afstanden til den nærmeste nabo er stor, ligesom større radius også kan give større bias. Desuden er det også tydeligt, at man bliver stillet over for flere valg, når man benytter matching. Eksempelvis hvordan data vægtes med matching-algoritmen; størrelsen af radius; eller hvor mange observationer, man bruger til at matche ved nearest neighbor matching. I praksis kan det derfor være en god idé at teste, hvor sensitive ens resultater er over for disse valg.

Matching og kausalitet

Givet at man har beregnet et estimat ved hjælp af matching, er spørgsmålet, om det kan gives en kausal fortolkning. Dette er et af de forhold, der ofte er uklarhed om i empiriske anvendelser af metoden.[8] Selvom matching har intuitiv appel, er det vigtigt at pointere, at matching sjældent *i sig selv* kan identificere kausale effekter.

Helt generelt må vi for at tage springet fra korrelation til kausalitet gøre os antagelser – i øvrigt ligesom det er tilfældet for alle andre metoder, der søger at identificere kausale effekter (Angrist og Pischke, 2009; Keele og Minozzi, 2013). I tilfældet med matching kan estimaterne kun gives en kausal fortolkning under antagelse af, at de variabler, der påvirker den kausale variabel (T) – og dermed skaber "selektion ind i" eksperiment- og kontrolgruppen – er kendte og observerbare (Morgan og Winship, 2007: 91; Angrist og Pischke, 2009:

69; Keele og Minozzi, 2013).9 Dette kræver – ligesom ved OLS-regression – at alle relevante variable, der påvirker den kausale variabel (og den afhængige variabel), er observerede og inkluderet i modellen (Morgan og Winship, 2007: 91; Angrist og Pischke, 2009: 69; Keele og Minozzi, 2013). Dette er en uhyre krævende antagelse, som formentlig sjældent er opfyldt, når man arbejder med observationsdata. Eksempelvis bliver antagelsen brudt, hvis uobserverede eller uobserverbare variable påvirker både den kausale og den afhængige variabel. Grundlæggende er matching således – ligesom OLS-regression – en kontrolstrategi (Morgan og Winship, 2007; Angrist og Pischke, 2009), der kun identificerer kausale effekter, hvis selektionsprocessen er kendt, baseret på observerbare forhold og korrekt specificeret i modellen.

Det betyder, at selv i tilfælde, hvor der er perfekt overlap og balance mellem eksperiment- og kontrolgrupperne på observerbare forhold, er der ingen garantier for, at dette også er tilfældet på *u*observerede forhold. Selv i tilfælde hvor man har megen information om selektionsprocessen, kan en stor del af den interessante (selv)selektion stadig ske på uobserverbare forhold, og estimaterne er derfor følsomme over for "selektion på uobserverede" variable. I sammenligning sikrer eksperimentel randomisering, at fordelingerne for eksperiment- og kontrolgrupperne er ens på både observerede *og* uobserverede variable (Dunning, 2012). Denne betingelse er vanskellig at tilfredsstille uden et stærkt design. Som Sekhon (2009: 503) fremhæver: "Without an experiment, natural experiment, a discontinuity, or some other strong design, no amount of econometric or statistical modeling can make the move from correlation to causation persuasive". Som metode til at identificere kausale effekter fungerer matching derfor bedst i kombination med et stærkt design – fx et kvasi-eksperiment eller et naturligt eksperiment – som kan bidrage til at eliminere effekten af uobserverede variable, der kan forstyrre sammenhængen mellem den kausale og den afhængige variabel.

Empirisk eksempel

Som beskrevet ovenfor kan matching betragtes som en metode til at øge sammenligneligheden af observationer i data. Imidlertid hviler en kausal fortolkning af resultaterne på styrken af forskningsdesignet. Nedenfor illustrerer vi, hvordan matching kan anvendes i forbindelse med analyser, der hvilker på et kvasi-eksperimentelt design, men hvor mulige ubalancer i data stadig kan give udfordringer for en kausal fortolkning af estimaterne.

Data og design

Det empiriske eksempel tager udgangspunkt i et kvasi-eksperimentelt design, hvor en pludselig intervention opdeler en survey i en eksperiment- og kontrolgruppe. Den 20. april 2010 eksploderede olieboreplatformen Deep Water Horizon ud for Louisianas kyst i den Mexicanske Golf. Ved eksplosionen af den britisk-ejede British Petroleum (BP) boreplatform, døde 11 besætningsmedlemmer og det, der skulle vise sig at blive en af verdenshistoriens største oliekatastrofer, var en realitet. Oliekatastrofen tiltrak megen medieopmærksomhed overalt i verden – ikke mindst i Storbritannien, som er hjemland for BP.

Eksplosionen og det efterfølgende oliespild fandt sted samtidig med dataindsamlingen af "British Household Panel Survey", som fra 2010 er inkorporeret i forskningsprojektet Understanding Society.[10] I denne survey bliver et repræsentativt udsnit af briter interviewet om holdninger til en lang række forhold, blandt andet hvordan de opfatter miljørelaterede emner. Vi bruger data fra april måned 2010 til at undersøge, hvordan eksplosionen på Deep Water Horizon påvirker holdninger til miljørelaterede spørgsmål.

Vi refererer til analysens design som et kvasi-eksperiment – og ikke et naturligt eksperiment. Et naturligt eksperiment er karakteriseret ved, at en udefrakommende intervention "så godt som" tilfældigt inddeler observationer i en eksperiment- og kontrolgruppe (Dunning, 2012). Dette er ikke nødvendigvis tilfældet ved kvasi-eksperimenter (Blom-Hansen og Serritzlew, 2014; Hariri, Bjørnskov, og Justesen, 2013). Eksplosionen på Deep Water Horizon var uanticiperet af den britiske befolkning og var i den forstand et udefrakommende chok, der opdeler surveyen i to grupper: en kontrolgruppe interviewet før eksplosionen, og en eksperimentgruppe interviewet efter eksplosionen. Den kausale variabel (interventionen) er således en dummyvariabel, kodet som 1 for respondenter interviewet den 20. april 2010 eller senere, og som udgør "eksperimentgruppen". Respondenter interviewet før den 20. april 2010 udgør kontrolgruppen (kodet som 0). Ikke desto mindre kan der være systematiske forskelle på, hvem der blev interviewet før og efter eksplosionen. Således er selv en uventet begivenhed som denne ikke garanti for, at inddelingen i eksperiment- og kontrolgruppe er randomiseret. Det betyder, at der kan være variable, der systematisk er korrelerede med både interventionen og den afhængige variabel, således at eksperiment- og kontrolgrupperne ikke er balancerede (Angrist og Pischke, 2009: 23). Vi bruger derfor matching til at "trimme" data for at gøre observationerne i de to grupper så sammenlignelige som muligt.

Den afhængige variabel er et indeks bestående af seks spørgsmål, der vedrører forhold som miljøkatastrofer og klimaforandringer. Eksempelvis bliver respondenterne bedt om at vurdere, om "vi vil opleve større miljøkatastrofer, hvis

vi fortsætter den nuværende kurs".[11] Et andet spørgsmål beder respondenterne erklære sig enige eller uenige i, at "klimaforandringer ligger langt ude i fremtiden".[12] For alle spørgsmål er svarkategorien "Ja" eller "Nej". En faktoranalyse af de seks items viser, at én faktorløsning forklarer ca. 25 pct. af den observerede variation, og at alle items loader med mere end 0,3 på denne faktor.[13] På baggrund af de seks variable har vi lavet et sum-index, hvor den afhængige variabel varierer fra 0 til 6.

De uafhængige variable, vi matcher respondenterne i eksperiment- og kontrolgruppen på, består af en række sociale baggrundsvariable, der alle er prædeterminerede i forhold til interventionen, og som potentielt kan være kilde til ubalance mellem eksperiment- og kontrolgrupperne. Specifikt matcher vi respondenter på uddannelsesniveau, indkomst, alder, køn, britisk versus ikke-britisk nationalitet, bopæl (land-by) og et mål for respondenternes socialklasse (af pladshensyn vil en nærmere beskrivelse af variablene være tilgængelig i et webappendiks). Dels kan det ikke kan afvises, at der er systematiske forskelle mellem eksperiment- og kontrolgruppen på disse variable – fx hvis flere mænd end kvinder er interviewet før eksplosionen – dels kan de også påvirke individers holdninger til miljøspørgsmål.

Matching: procedure og estimater

Det første skridt i matching-analysen er at sikre, at kontrolgruppen og eksperimentgruppen er så sammenlignelige som muligt på de uafhængige variable. Konkret bruger vi den dummy-variabel, der opdeler data i en eksperiment- og kontrolgruppe, som afhængig variabel i en logistisk regression. Sættet af uafhængige variable i den logistiske regression er de uafhængige variable nævnt ovenfor, der potentielt er korrelerede med både interventionen og den afhængige variabel. Det skal også bemærkes, at vi ikke forsøger at estimere en kausal effekt af de uafhængige variable, men blot bruger den logistiske sandsynlighedsmodel til at undersøge, om der er systematiske forskelle på grupperne interviewet før og efter interventionen.

Givet sættet af uafhængige variable estimerer vi sandsynlighedsscoren – den forudsagte sandsynlighed for at respondenter befinder sig i eksperimentgruppen. Sandsynlighedsscoren bruges således til at matche de observationer i kontrol- og eksperimentgrupperne, der har tilnærmelsesvis ens sandsynligheder for at befinde sig i eksperimentgruppen.

Det næste skridt i analysen er at definere overlapsregionen, dvs. det interval på sandsynlighedsscoren, hvor der er respondenter i både eksperiment- og kontrolgruppen. Her definerer vi overlapsregionen som intervallet fra minimumsværdien på sandsynlighedsscoren for eksperimentgruppen til maksi-

mumsværdien på for kontrolgruppen (Persson og Tabellini, 2003: 143). Denne region udgør her intervallet imellem 0,21 og 0,56 på sandsynlighedsscoren.[14] Overlapsregionen er vist i figur A1 i appendikset. Kun otte observationer har så lave/høje værdier på sandsynlighedsscoren, at de ikke udgør et godt sammenligningsgrundlag. For de resterende observationer er der observationer fra både eksperiment- og kontrolgrupperne inden for ret snævre bånd (0,01) på sandsynlighedsscoren.

Indtil nu er matching-proceduren foregået fuldstændig uafhængigt af den afhængige variabel – indekset for miljøholdninger. Denne bliver først introduceret nu, hvor den estimerede effekt af interventionen – eksplosionen på Deep Water Horizon – beregnes. Resultaterne af matching-analyserne fremgår af tabel 1.

For sammenlignelighedens skyld viser modellerne estimatet fra en OLS-regression, som angiver, at miljøkatastrofen har en signifikant positiv effekt på briternes holdninger til miljøspørgsmål. I model 2 og 3 bruger vi en matching-algoritme, hvor observationer fra eksperimentgruppen matches med deres "nærmeste nabo" i kontrolgruppen, hvorefter forskellen i de to gruppers gennemsnit på den afhængige variabel beregnes. I model 2 inkluderes alle observationer – med og uden overlap – mens model 3 samt de efterfølgende modeller begrænser data til observationer inden for overlapsregionen. For sammenlignelighedens skyld viser vi også resultater fra matching med to og tre nærmeste naboer (model 4 og 5). Vi rapporterer også resultater fra radius-matching, hvor radius er fastsat til hhv. 0,01, 0,02 og 0,03 på sandsynlighedsscoren (model 6-8). Her er det værd at erindre, at sandsynlighedsscoren per definition varierer mellem 0 og 1. En radius på hhv. 0,01 eller 0,02 er således ganske lille og betyder konkret, at der skal findes matchende observationer med sandsynlighedsscorer, der ligger meget tæt på observationerne i eksperimentgruppen.

Det fremgår af tabel 1, at der er en signifikant effekt af miljøkatastrofen på briters holdninger til miljøspørgsmål. Eksplosionen på Deep Water Horizon havde derfor en betydning for, hvordan respondenterne opfatter miljøspørgsmål. Derudover skal det bemærkes, at der ikke er voldsomme forskelle i resultaterne på tværs af matching-algoritmerne, med undtagelse af at vi finder betydeligt større effekter af Deep Water Horizon-katestrofen på vælgernes syn på miljøet, når vi bruger nearest neighbor matching og kun matcher på én nabo. Men som nævnt giver denne metode ikke nødvendigvis retvisende estimater, hvis den nærmeste nabo er langt væk. Derudover har matching-estimaterne – bortset fra nearest neighbor estimaterne – stort set samme størrelse som OLS-estimaterne.

Tabel 1: Matching-estimater

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Metode	OLS		Nearest neighbor-matching				Radius-matching	
Koefficient	0,31** (2,59)	0,51** (2,84)	0,52* (2,85)	0,32* (2,04)	0,30* (2,04)	0,31* (2,42)	0,31* (2,40)	0,31* (2,45)
N	750	750	742	742	742	742	742	742
Antal nærmeste naboer; radius; bandwidth		1 nabo	1 nabo	2 naboer	3 naboer	Radius = 0,01	Radius = 0,02	Radius = 0,03
Overlap defineret?	Nej	Nej	Ja	Ja	Ja	Ja	Ja	Ja

Note: Alle analyser er udført i Stata 12.1 med kommandoerne pscore og psmatch2. Følgende variable er anvendt til at beregne sandsynlighedsscoren: køn, alder, klasse, indkomst, uddannelse, britisk etnicitet, bybo og ægteskabelig status. Disse variable anvendes også som kontrolvariable i OLS-regressionen (model 1). Overlapsregionen er 0,21 til 0,56. Koefficienter er average treatment effect on the treated. t-værdier i parentes. *p < 0,05; **p < 0,01.

Endelig er det vigtigt at undersøge, om eksperiment- og kontrolgrupperne er balancerede på de observerbare uafhængige variable. Hvis dette ikke er tilfældet, har vi ikke blot et potentielt problem med "selektion på uobserverbare" variable, men også med "selektion på observerbare variable", hvilket betyder, at eksperiment- og kontrolgrupperne ikke er "så godt som tilfældigt" fordelt i forhold til de uafhængige variable. Derfor tester vi, om grupperne er ensartede før og efter matching, og derved om matching-proceduren har bidraget til at øge sammenligneligheden af eksperiment- og kontrolgruppen. Til dette formål bruger vi Kolmogorov-Smirnov tests, der tester, om de uafhængige variable har samme fordeling i de matchede kontrol- og eksperimentgrupper.[15] Resultaterne fra balancetestene er tilgængelige i appendiks og viser, at kontrol- og eksperimentgrupperne generelt er balancerede på de uafhængige variable, samt at matching gør flere variable mere balancerede.

Spørgsmålet er herefter, om disse estimater kan gives en kausal fortolkning. Dvs., er estimatet udtryk for en korrelation, eller har vi identificeret en kausal effekt af Deep Water Horizon-miljøkatastrofen på holdninger til miljøspørgsmål. Som nævnt ovenfor kræver en sådan fortolkning, at der ikke er relevante, uobserverede variable, der påvirker sammenhængen mellem interventionen og den afhængige variabel. Denne antagelse er utestbar (Keele og Minozzi, 2013) og kan bedst forsvares, hvis den statistiske analyse er baseret på et stærkt design, der bidrager til at eliminere betydningen af uobserverede faktorer. Vi hverken kan eller vil afvise, at der kan være relevante variable, der påvirker sammenhængen – og det er heller ikke vores ærinde. Snarere er pointen, at i modsætning til de fleste statistiske analyser af observationsdata giver analyser, der er baseret på en kombination af et kvasi-eksperimentalt design og *ex post* statistisk justering af data, et bedre udgangspunkt for at identificere kausale effekter. Således vil designbaserede tilgange til kausal inferens formentlig reducere chancerne for, at en udeladt variabel driver resultatet.

Konklusion

I denne artikel har vi gennemgået styrker og svagheder ved matching. Matching forsøger at tilnærme sig den eksperimentelle situation ved at gøre kontrol- og eksperimentgrupperne så ens som muligt på observerbare karakteristika. Dette er i sig selv vigtigt, fordi mange af de problemstillinger, vi som politologer er interesserede i, ikke lader sig studere udelukkende ved hjælp af kontrollerede eksperimenter.

Den største udfordring i forbindelse med at isolere kausale effekter for matching – såvel som for alle andre analyser af observationsdata – er, at det ikke er tilfældigt, om observationerne i data befinder sig i kontrol- eller eks-

perimentgruppen. Den væsentligste konsekvens heraf er, at vi aldrig kan være sikre på, at vi opnår perfekt kontrol for uobserverede variable og dermed, at vi har isoleret den kausale effekt af den intervention, vi interesserer os for. Med observationsdata kan kausalitetsproblemet ikke løses alene ved at kontrollere eller justere data med statistiske metoder. Bevægelsen fra korrelation til kausalitet hviler til syvende og sidst på styrken af forskningsdesignet.

Noter

1. Vi er taknemmelige for konstruktive kommentarer fra Kim Mannemar Sønderskov samt to anonyme bedømmere.
2. Selvselektion forekommer, når individer selv vælger at deltage i et program eller en bestemt gruppe (Gelman og Hill, 2007: 168; Angrist og Pischke, 2009: 15).
3. På grund af forsøget på at simulere det eksperimentelle setup introduceres matching ofte med reference til den såkaldte *potential outcomes model* (Smith og Todd, 2005; Sekhon, 2009).
4. For en introduktion til *difference-in-differences* matching, se Smith og Todd (2005).
5. I litteraturen nævnes det ofte også, at matching ikke kræver antagelser om den funktionelle form for sammenhængen mellem de uafhængige variable og den afhængige variabel (Harding, 2003: 689; Persson og Tabellini, 2003: 139; Smith og Todd, 2005: 342). Vi anser det primære formål med matching for at være at skabe mere sammenlignelige grupper i data og diskuterer derfor ikke nærmere spørgsmålet om funktionel form.
6. Ved nearest neighbor matching kan vægten således være 1 for en kontrolobservation, der bruges som match én gang, mens vægten er 2 for en kontrolobservation, der bruges som match to gange osv.
7. En anden overvejelse er, om *den samme* observation i kontrolgruppen kan bruges som match for *flere* observationer i eksperimentgruppen (*matching with replacement*), eller om den kun kan bruges én gang (*matching without replacement*). Fordelen ved den første fremgangsmåde er, at hver observation i kontrolgruppen kan sammenlignes med flere ensartede observationer i eksperimentgruppen. Dette reducerer bias, fordi afstanden på sandsynlighedsscoren mellem de matchede observationer minimeres (Dehejia og Wahba, 2002: 153). Hvis observationer i kontrolgruppen kun bruges som match for én observation i eksperimentgruppen – og derefter ikke benyttes mere – risikerer man at matche observationer i eksperimentgruppen med uensartede observationer i kontrolgruppen (særligt hvis der er få observationer i kontrolgruppen), hvilket kan øge bias (Cameron og Trivedi, 2005: 873; Smith og Todd, 2005).
8. I to meget citerede artikler hævdede Dehejia og Wahba (1999, 2002), at de ved anvendelse af matching kunne producere estimater, der stort set var lig estimater

fra analyser af eksperimentelle data. Dette argument blev effektivt tilbagevist af Smith og Todd (2005). Politologiske artikler, der (tilsyneladende) betragter matching som en effektiv metode til at løse kausalitetsproblemer, inkluderer Kam og Palmer (2008) og Boyd, Epstein og Martin (2010).

9. Denne antagelse går under navne som *selection on observables*, *conditional independence*, *unconfoundedness*, *ignorability* eller *exogeneity*.
10. https://www.understandingsociety.ac.uk/
11. Spørgmål a_scenv_dstr.
12. Spørgsmål a_scenv_futr. De fire øvrige spørgsmål i indekset lyder: "Den miljømæssige krise menneskeheden står overfor har været overdrevet" (a_scenv_exag). "Det kan ikke betale sig for mig at gøre noget for miljøet, hvis ingen andre gør det" (a_scenv_chwo). "Vil du blive påvirket af miljøforandringer indenfor de næste 30 år?" (a_scopecl30). "Det kan ikke betale sig for Storbritannien at kæmpe mod miljøforandringer" (a_scenv_brit) .
13. Da svarkategorierne er dikotome, baserer faktoranalysen sig på en polykorisk korrelationsmatrice, som tager højde for, at Pearson-korrelationer baseret på dikotome variable vil overvurdere den reelle sammenhæng (Rigdon, 2010).
14. Af pladshensyn viser vi overlapsregionen i et webappendiks på https://sites.google.com/site/robertklemmensen/
15. Tabellen med teststørrelser for balance før og efter matching findes webappendikset.

Litteratur

Angrist, Joshua og Jörn-Steffen Pischke (2009). *Mostly Harmless Econometrics: An Empiricists' Companion*. New Jersey: Princeton University Press.

Blom-Hansen, Jens og Søren Serritzlew (2014). Endogenitet og eksperimenter. *Politica* 46 (1): 5-23.

Boyd, Christina L., Lee Epstein og Andrew D. Martin (2010). Untangling the causal effect of sex on judging. *American Journal of Political Science* 54 (2): 389-411.

Cameron, A. Colin og Pravin K. Trivedi (2005). *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.

Dehejia, Rajeev H. og Sadek Wahba (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *The Journal of the American Statistical Association* 94 (448): 1053-1062.

Dehejia, Rajeev H. og Sadek Wahba (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics* 84 (1): 151-161.

Dinesen, Peter T. (2012). Does generalized (dis)trust travel? Examining the impact of cultural heritage and destination-country environment on trust of immigrants. *Political Psychology* 33 (4): 495-511.

Dunning, Thad (2012). *Natural Experiments in the Social Sciences – A Design Based Approach.* Cambridge: Cambridge University Press.

Gelman, Andrew og Jennifer Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Guo, Shenyang og Mark W. Fraser (2010). *Propensity Score Analysis: Statistical Methods and Applications*. Thousand Oaks, CA: Sage Publications.

Ho, Daniel E., Kosuke Imai, Gary King og Elisabeth Stuart (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15 (3): 199-226.

Harding, David J. (2003). Counterfactual models of neighborhood effects: The effect of neighborhood poverty on dropping out and teenage pregnancy. *American Journal of Sociology* 109 (3): 676-719.

Hariri, Jacob Gerner (2014). Statskundskabens sammenfiltrede virkelighed og et bud på en løsning: IV-estimation. *Politica* 46 (1): 79-94.

Hariri, Jacob Gerner, Christian Bjørnskov og Mogens K. Justesen (2013). Economic shocks and subjective well-being: Evidence from a quasi-experiment. *Arbejdspapir*. Tilgængeligt på www.ssrn.com.

Imbens, Guido M. og Jeffrey M. Wooldridge (2008). Recent developments in the econometrics of program evaluation. *Working paper 14251*. National Bureau of Economic Research. Cambridge, MA.

Justesen, Mogens K. (2012). Democracy, dictatorship, and disease: Political regimes and HIV/AIDS. *European Journal of Political Economy* 28 (3): 373-389.

Kam, Cindy og Carl L. Palmer (2008). Reconsidering the effects of education on political participation. *Journal of Politics* 70 (3): 612-631.

Keele, Luke og William Minozza (2013). How much is Minnesota like Wisconsin? Assumptions and counterfactuals in causal inference with observational data. *Political Analysis* 21 (1): 193-216.

Khanker, Shahidur R., Gayatri B. Koolwal og Hussain Samad (2009). *Handbook on Impact Evaluation: Quantitative Methods and Practices*. Washington D.C.: The World Bank.

Morgan, Stephen L. og Christopher Winship (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press.

Olsen, Asmus L. (2014). Tærskelvariable og tærskelværdier – en introduction til regressionsdiskontinuitetsdesignet. *Politica* 46 (1): 42-59.

Persson, Torsten og Guido Tabellini (2003). *The Economic Effects of Constitutions*. Cambridge, MA: MIT Press.

Rigdon, Edward E. (2010). The polychoric correlation coeffiecient, pp. 789-801 i Neil J. Salkin (red.), *Encyclopedia of Research Design*. New York: Sage University Press.

Rosenbaum, Paul R. og Donald B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 (1): 41-55.
Sekhon, Jasheet S. (2009). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science* 12: 487-508.
Smith, Jeffrey og Petra E. Todd (2005). Does matching overcome LaLonde's critiques of nonexperimental estimators? *Journal of Econometrics* 125 (1-2): 305-353.

politica, 46. årg. nr. 1 2014, 79-94

Jacob Gerner Hariri

Statskundskabens sammenfiltrede virkelighed og et bud på en løsning: IV-estimation

IV-estimation er efterhånden blevet en fast og nødvendig del af den empiriske værktøjskasse i statskundskaben. Den er nødvendig, fordi vores empiriske genstandsfelt er sammenfiltret; variable hænger sammen på kryds og tværs, og det er langtfra ligetil at afgøre, hvad der er årsag til hvad. Formålet med IV-estimation er netop at skære igennem sammenfiltringen og at afklare, hvorvidt og hvor meget én faktor er årsag til en anden. Denne artikel motiverer og forklarer brugen af instrumentvariable i statistiske analyser, ligesom der gives praktiske råd til analyser med instrumentvariable.

I statskundskaben må de fleste empiriske analyser forlade sig på virkeligheden, som den udspiller sig uden for vinduet. Det er simpelthen ikke alle politologisk relevante problemstillinger, der kan undersøges eksperimentelt – hverken i laboratoriet eller i felten. Derfor studerer vi virkeligheden, og virkeligheden er kompleks: Variable hænger sammen på kryds og tværs, og derfor kan det i mange tilfælde være meget svært at afgøre, hvad der er årsag, og hvad der er virkning.

Det er denne sammenfiltrede virkelighed, hvor x påvirker y, y påvirker x, og alt muligt andet påvirker både x og y, der motiverer brugen af instrumentvariable. Idéen er simpel: Find en faktor, som kun påvirker x og derudover ingen forbindelse har til de øvrige variable i den empiriske analyse. Denne faktor kaldes en instrumentvariabel eller simpelthen et instrument. Fordi dette instrument er uafhængigt af y og alt andet i modellen, kan det isolerere variationen i x, som ikke kommer fra y (instrumentet er jo ikke forårsaget af y), og som heller ikke kommer fra de andre faktorer i modellen. Denne uafhængige eller eksogene variation i x kan derefter bruges til at identificere den kausale effekt af x på y.

Selvom idéen nok er simpel, kan IV-estimation være vanskeligt i praksis. Det er simpelthen svært at finde brugbare instrumenter, faktorer som udelukkende påvirker x og intet andet i en statistisk model. Af samme grund benytter vi i næste afsnit to illustrative eksempler til at motivere og forklare brugen af IV-estimation. Det ene undersøger effekten af politi på kriminalitet, hvor valgår bruges som instrument for antallet af politifolk i en by (Levitt, 1997).

Det andet undersøger effekten af økonomisk vækst på sandsynligheden for militær konflikt i Afrika syd for Sahara. Her bruges nedbør som instrument for økonomisk vækst. Disse eksempler gør forhåbentlig statistikken håndgribelig. Artiklen fortsætter med en grafisk og siden lidt mere teknisk fremstilling. Afslutningsvis gives et antal praktiske råd til IV-estimation.

To eksempler på IV-estimation

Eksempel 1: politi og kriminalitet

Steven Levitt forsøgte i 1997 at besvare et tilsyneladende simpelt kausalt spørgsmål: Falder kriminaliteten i en by, hvis der ansættes flere politifolk? Og hvor meget i så fald? Dette er det relevante spørgsmål for politikere, der skal prioritere offentlige midler: Hvor meget kan kriminaliteten forventes at falde for en given bevilling til politiet?

Selvom spørgsmålet er simpelt, er svaret det ikke. På den ene side gælder det, at flere politifolk givetvis begrænser kriminaliteten. På den anden side ansporer kriminalitet antageligt et byråd til at prioritere politiindsatsen; højere kriminalitet giver således mere politi.

Forventningen er derfor i udgangspunktet, at flere politifolk på den ene side er årsag til mindre kriminalitet, mens mere kriminalitet på den anden side er årsag til flere politifolk. En velkendt OLS-regression af kriminalitet på antallet af politifolk indeholder begge sammenhænge, og i Levitts data (som i mange studier før hans) var der faktisk en positiv sammenhæng mellem kriminalitet og antallet af politifolk; selv når der blev kontrolleret for socioøkonomiske og demografiske faktorer, størrelsen af sociale ydelser, uddannelsesniveau arbejdsløshed og andre relevante faktorer. Den positive effekt af kriminalitet på politistyrken dominerer altså den negative effekt af politistyrken på kriminalitet.

I dette tilfælde er det oplagt, at den observerede korrelation ikke kan udlægges som den kausale effekt af politi på kriminalitet; det lyder usandsynligt, at flere politifolk i sig selv skulle være årsag til mere kriminalitet. Lige så oplagt er det, at korrelationen er ubrugelig som baggrund for policy-anbefalinger: Den logiske konsekvens ville være at anbefale at fyre alle politifolk for derved at begrænse kriminaliteten mest muligt.

Steven Levitts IV-analyse tager afsæt i den betragtning, at antallet af politifolk i amerikanske byer ganske ofte stiger i de år, hvor der er valg til byråd og borgmesterpost: Når valget nærmer sig, sættes der flere penge af til politiet. I Levitts data stiger antallet af politifolk da også i gennemsnit med 2 pct. i valgår. Levitt benytter disse elektorale cykler til at isolere den del af variationen i antallet af politifolk, som er uafhængig af graden af kriminalitet i byerne. Da valgperioderne er faste, bliver tidspunktet for valg til byrådene ikke påvirket

af kriminaliteten, og derfor bliver den del af variationen i antallet af politifolk, som kan henføres til valgcyklerne, heller ikke påvirket af kriminaliteten. Levitt bruger altså valgcyklerne som instrumentvariabel for antallet af politifolk, hvilket umiddelbart ser ud til at fjerne problemet med omvendt kausalitet, nemlig at graden af kriminalitet også forårsager variation i politistyrken. Analysen er dog ikke skudsikker. For det kan meget vel tænkes, at der også er valgcykler i fx socialpolitikken eller på uddannelsesområdet. Hvis det er tilfældet (således at det kommunale forbrug på disse områder også vokser i valgår), og hvis disse områder også varierer systematisk med kriminaliteten, ja så vil Levitts analyse overvurdere den kausale effekt af politi på kriminalitet: Den estimerede effekt vil så indeholde effekten af en forbedret socialpolitisk indsats og øgede udgifter til uddannelse, som må formodes også at virke dæmpende på kriminaliteten. Levitt kontrollerer imidlertid for disse faktorer og mener dermed at have isoleret den rene kausale effekt af politi på kriminalitet. For fuldstændighedens skyld runder vi lige eksemplet af og nævner, at ifølge Levitts IV-analyse har antallet af politifolk en betydelig dæmpende effekt på voldskriminalitet men ingen effekt på berigelseskrimalitet.

Eksempel 2: nedbør, vækst og militær konflikt

Antag, som det andet eksempel, at vi ønsker at identificere den kausale påvirkning af økonomisk vækst på sandsynligheden for borgerkrig på tværs af lande i Afrika. Dette er en vanskelig empirisk problemstilling. For det første er der endnu engang omvendt kausalitet, idet sandsynligheden for borgerkrig påvirker væksten, lige så vel som væksten påvirker sandsynligheden for borgerkrig. Dette problem kan ikke løses ved simpelthen at betragte værdier for national indkomst på et tidligere tidspunkt end borgerkrigen: Den nationale indkomst bliver skabt af borgere, som træffer deres beslutning om økonomisk aktivitet på baggrund af deres forventninger om fremtiden – herunder naturligvis deres forventning om sandsynligheden for borgerkrig. For det andet er der næsten med sikkerhed historiske og kulturelle faktorer, som ikke kan observeres, men som hænger sammen med både væksten og sandsynligheden for borgerkrig. Alt dette betyder, at den nationale indkomst ikke er eksogen i denne sammenhæng, og størrelsen af den kausale påvirkning herfra på sandsynligheden for borgerkrig kan ikke umiddelbart identificeres.

Edward Miguel, Shanker Satyanath og Ernest Sergenti (2001) løste problemet ved i et kendt studie at benytte variation i nedbøren som kilde til eksogen variation i den nationale indkomst i afrikanske lande. Da landbrug stadig spiller en betydelig økonomisk rolle i Afrika syd for Sahara, har mængden af nedbør målbare nationaløkonomiske konsekvenser. Derfor benyttede forfat-

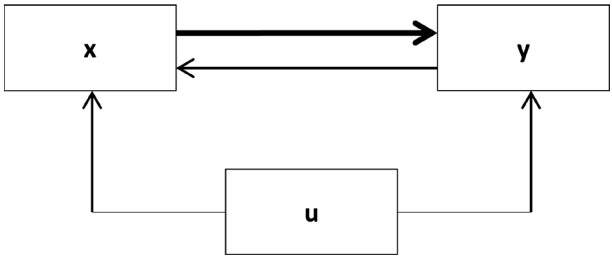
terne kun den del af variationen i væksten, som afhænger af nedbøren, til at identificere den kausale påvirkning af indkomst på sandsynligheden for borgerkrig. Dette giver kausal identifikation, fordi nedbør i denne sammenhæng er eksogen: Sandsynligheden for borgerkrig påvirker ikke mængden af nedbøren, ligesom det nok er begrænset i hvilken grad, at de faktorer som påvirker nedbøren (fx atmosfæriske trykforhold) påvirker sandsynligheden for borgerkrig andet end gennem den økonomiske vækst. Med nedbør som instrument for vækst finder Miguel et al. en meget stærk effekt af økonomisk recession på sandsynligheden for militær konflikt i Afrika syd for Sahara.

I begge disse eksempler er der betydelig omvendt kausalitet, idet den afhængige variabel også påvirker den uafhængige. Desuden er der et væld af faktorer, som påvirker både den uafhængige og den afhængige. Derfor er der intet som *a priori* tilsiger, at politi skulle påvirke kriminalitet, eller økonomisk recession skulle påvirke konflikt snarere end omvendt. Netop derfor er det en væsentlig opgave for empirisk samfundsforskning at forsøge at udrede trådene og identificere så præcist som muligt, hvor meget man kan forvente, at kriminaliteten falder, hvis politistyrken øges, og hvordan økonomien påvirker sandsynligheden for konflikt. Redskabet til at isolere envejspåvirkningen var i begge tilfælde IV-estimation, og i de følgende afsnit udfoldes tankegangen bag metoden. Først grafisk og siden sættes metoden på formel.

IV-estimation grafisk set

Figur 1 illustrerer politologiens sammenfiltrede virkelighed. Der er en sammenhæng mellem to faktorer, *x* og *y*, men vi kan ikke i udgangspunktet afgøre, hvorfra sammenhængen kommer.

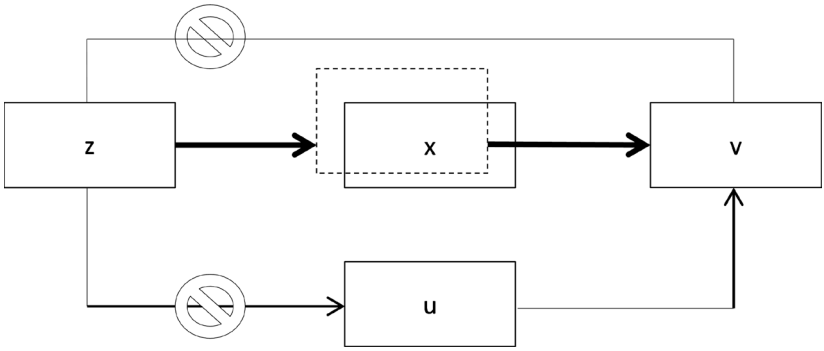
Figur 1: Korrelationsanalyse hvor effekten af *x* på *y* ikke kan identificeres



Er *x* årsag til *y*, *y* årsag til *x*, forårsages de begge af *u*, eller er det i virkeligheden lidt af alle disse sammenhænge på en gang? Hvis vi gerne vil identificere hvor meget, *x* påvirker *y* – den fede pil i figuren – er det altså ikke nok at betragte

korrelationen mellem de to faktorer. Denne afspejler jo potentielt samtlige pile i figuren. Pointen med IV-estimation er netop at eliminere de forstyrrende pile i figuren, således at forskningsspørgsmålet kan besvares, og effekten af *x* på *y* kan udsondres.

Figur 2: IV-estimation



Modsat figur 1 er der i figur 2 to pile. Det illustrerer, at sammenhængen mellem *x* og *y* ved IV-estimation estimeres i to trin. Det andet trin besvarer det egentlige forskningsspørgsmål (hvor stor er effekten af *x* på *y*), mens det første trin udelukkende tjener til at eliminere de pile, der forstyrrede den egentlige analyse. Hvordan så det?

I første trin af analysen estimeres sammenhængen mellem instrumentet og den uafhængige variabel. Instrumentet er illustreret ved kassen *z*. Den stiplede firkant, der omfatter en del af *x*, illustrerer, at instrumentet aldrig kan opfange al variationen i den uafhængige variabel. Det er dog væsentligt, at instrumentet opfanger en væsentlig del af variationen i den uafhængige, grafisk svarende til at den stiplede kasse så vidt muligt er sammenfaldende med den uafhængige variabel. Hvis instrumentet kun opfanger en begrænset del af variationen i den uafhængige variabel, bliver effekten af den uafhængige på den afhængige variabel estimeret upræcist i analysens andet trin. I det tilfælde siger vi, at instrumentet er svagt. Eller med andre ord: Jo stærkere korrelationen er mellem instrumentet og den endogene uafhængige variabel, desto stærkere instrument og desto mere præcist bliver estimatet i analysens andet (og egentlige) trin.

For at IV-estimation kan løse problemerne med omvendt kausalitet (pilen fra *y* til *x*) og udeladte variable (korrelationen mellem *u* samt *x* og *y*), skal instrumentet opfylde følgende: Det skal ikke have nogen sammenhæng med *y* andet end via det *x*, som det instrumenterer for. Dette illustreres af de "forbudte" pile

fra *z* til *y* og fra *z* til *y* via *u*. Hvis instrumentet opfylder dette, er det eksogent (uafhængigt af de andre faktorer i modellen), og analysens første skridt kan således bruges til at isolere den eksogene variation i *x*. Hvis instrumentet udelukkende er korreleret med *y* via det endogene *x*, er instrumentet validt.

Opsummerende kræves det altså, at instrumentet skal være (i) stærkt (stærkt korreleret med det endogene *x*, som det er instrument for; illustreret ved sammenfaldet mellem den stiplede kasse og *x*) og (ii) validt (ikke korrelere med *y* andet end via det endogene *x*; illustreret ved de forbudte pile). Hvis disse betingelser er opfyldt, giver IV-estimation et konsistent estimat af den kausale effekt af *x* på *y*. Det ser vi nærmere på i næste afsnit.

Før vi kommer dertil, skal det nævnes, at særligt den anden betingelse i praktisk arbejde går under forskellige betegnelser. Hvis den anden betingelse gælder, siges det fx også, at instrumentet er ortogonalt på fejlleddet, eller at eksklusionsrestriktionen er opfyldt. "Ortogonalt på fejlleddet" betyder blot ukorreleret med alt det, vi burde have men ikke har med i modellen. I den anden formulering ligger, at hvis instrumentet udelukkende påvirker *y* gennem *x*, kan instrumentet udelades fra modellen, når *x* er indeholdt. Givet *x* har instrumentet altså ingen selvstændig forklaringskraft på *y*.

IV-estimation sat på formel

I sin simpleste form foregår IV-estimation simpelthen som almindelig OLS i to trin og kaldes af samme grund *two-stage least squares* (ofte forkortet 2SLS eller TSLS). I første trin benyttes instrumentet som nævnt til at isolere eksogen variation i den uafhængige variabel, og i andet trin benyttes denne eksogene variation så til at estimere den kausale effekt af den uafhængige variabel på den afhængige variabel.

De to trin i en 2SLS-analyse kan matematisk beskrives som følger:

$$x_i = c + az_i + dX' + e_i \quad (1)$$
$$y_i = k + bx_i + fX' + u_i \quad (2)$$

Ligning (2) beskriver analysens *second stage*, som er det egentlige forskningsspørgsmål; nemlig effekten af *x* på *y*. Denne er givet ved parameteren *b*. *b* svarer grafisk til pilen længst mod højre i figur 2. OLS-estimation af ligning (2) ville give et skævt resultat (*biased* koefficienter), fordi u_i og x_i per antagelse er korreleret. Ligning (1) repræsenterer analysens *first stage* og tjener til at identificere den eksogene variation i *x* ved hjælp af instrumentet *z*. Parameteren *a* angiver korrelationen mellem instrumentet og den endogene uafhængige variabel og svarer grafisk til pilen længst mod venstre i figur 2. Grafisk tjener analysens

første trin altså til at eliminere alle de forstyrrende pile, der indgik i figur 1. X' angiver forskellige (eksogene) kontrolvariable, som både indgår i analysens første og andet trin.

IV-estimation giver et konsistent estimat af den kausale effekt af *x* på *y*, hvis to betingelser er opfyldt. Matematisk ser de ud som følger:

$$\text{Cov}(z, x) \neq 0 \quad \text{(i)}$$
$$\text{Cov}(z, u) = 0. \quad \text{(ii)}$$

(i) siger, at der skal være sammenhæng mellem instrumentet, *z*, og den endogene variabel, *x*. Hvis (i) ikke er opfyldt, er instrumentet svagt. (ii) siger, at instrumentet ikke må være korreleret med fejlleddet i analysens andet trin, hvilket er det samme som, at instrumentet ikke må være korreleret med den afhængige variabel, *y*, andet end gennem *x*. Hvis (ii) er opfyldt, er instrumentet validt. Samlet set giver parameteren *b* et konsistent estimat af den kausale effekt af *x* på y, når både (i) og (ii) er opfyldt.

Den velkendte OLS-estimator er *unbiased*. Det betyder, at uanset stikprøvestørrelsen vil koefficienterne fra en OLS-analyse i gennemsnit ramme rigtigt i forhold til koefficienterne i befolkningen. Det samme gælder desværre ikke for 2SLS-estimatoren, som ikke er unbiased men dog *konsistent*. Det betyder, at koefficienterne fra 2SLS-analyser nærmer sig befolkningskoefficienterne, jo større stikprøven bliver. 2SLS-estimatoren er ikke unbiased, fordi analysens første trin er resultat af en estimation (i ligning (1)) og dermed indeholder lidt tilfældig variation. En del af denne tilfældige variation kommer fra det endogene *x*, som jo er den afhængige variabel i første trin. Og da *x* per definition er korreleret med fejlleddet i analysens andet trin (ellers ville vi jo ikke behøve at instrumentere), siver en smule af den bias, som oprindeligt motiverede brugen af 2SLS frem for OLS, alligevel ind i estimatoren (Angrist og Pischke, 2009: 209).

I de følgende to afsnit zoomer vi ind på de to betingelser, (i) og (ii). Hvordan ved vi, om de er opfyldt, og hvilke konsekvenser har det for analysen, hvis de ikke er?

Svage instrumenter?

I figur 2 illustrerede den stiplede kasse, at instrumentet sjældent kan forklare al variationen i den endogene uafhængige variabel. Hvis instrumentet i første trin kun forklarer en mindre del af variationen i *x*, bruges kun en lille del af variationen i *x* til i analysens andet trin at forklare *y*. Og dermed bliver estimatet af effekten af *x* på *y* upræcist. Instrumentvariable, som kun forklarer en

begrænset del af variationen i det endogene *x,* kaldes *svage instrumenter*. Svage instrumenter har to konsekvenser.

For det første forværres den indbyggede bias af 2SLS-estimatoren. 2SLS-estimatoren er som nævnt ikke unbiased, fordi første trin indeholder lidt tilfældig variation i *x*, hvoraf noget uvægerligt kommer fra fejlleddet *u*. Hvis instrumetet kun forklarer en lille del af *x*, kommer en relativt større del af den tilfældige variation i første trin fra *x* selv – og dermed fra *u*, som *x* jo per definition er korreleret med. Derfor bliver 2SLS-estimatoren mere biased, jo svagere instrumentet er. I yderste instans, hvor instrumentet ikke blot er svagt men decideret irrelevant (svarende til Cov(z, x) = 0) og slet ikke bidrager til at forklare *x*, svarer 2SLS-estimatoren til OLS-estimatoren. Første trin bortfalder så at sige, så analysen udelukkende består af ligning (2), hvilket i sig selv er en almindelig OLS-regression. Generelt gælder det således, at 2SLS-estimatoren bliver biased i retning af OLS-estimatoren, jo svagere instrumentet er.

For det andet og mere teknisk betyder svage instrumenter, at 2SLS-estimatoren ikke længere nødvendigvis er normalfordelt. Ligesom ved OLS, kræver statistisk inferens i 2SLS-analyser, at estimatoren er normalfordelt i store stikprøver. Derfor kan vi ikke stole på konfidensintervaller og *p*-værdier i 2SLS-analyser, når instrumenterne er svage (se fx Stock og Watson, 2007: 440).

Hvornår er et instrument svagt? Tommelfingerreglen siger, at hvis F-teststørrelsen i analysens første trin er mindre end 10, er instrumentet svagt (Stock, Wright og Yogo, 2002).[1] Der er selvsagt ikke tale om et absolut kriterium, men det anbefales i så fald at undersøge instrumentets robusthed.

Generelt gælder det, at F-teststørrelsen bliver mindre, hvis man medtager forklarende variable, der er svagt korreleret med den afhængige variabel. Dette gælder også i IV-analysens første trin: Hvis vi lidt forsimplende antager, at der er to instrumenter og disse er ukorrelerede, så er $F = \frac{1}{2}(t_1 + t_2)$; gennemsnittet af de kvadrede t-teststørrelser (fx Stock og Watson, 2007: 228). Skrevet på denne form er det klart, at F-teststørrelsen – og dermed styrken af instrumenterne i første trin – falder, hvis man tilføjer et instrument, som er svagt korreleret med det endogene *x*. Implikationen heraf er, at man ikke skal medtage relativt svage instrumenter, hvis man allerede har ét forholdsvist stærkt instrument. Hellere ét stærkt instrument end mange svage!

Hertil kommer at bias i 2SLS alt andet lige er voksende i antallet af instrumenter (Angrist og Pischke, 2009: 209). Faktisk gælder det, at når der præcis er ét instrument per endogen uafhængig variabel, er 2SLS-estimatoren tilnærmelsesvist unbiased. Derfor kan det i tilfælde med relativt svage instrumenter være en god idé at gentage analysen med kun ét instrument per endogent *x* (her

skal man naturligvis i hvert tilfælde vælge det stærkeste). Da estimatoren er her tilnærmelsesvist unbiased, minimerer dette problemet med svage instrumenter.

Holder eksklusionsretstriktionen?

Formålet med IV-estimation er jo i udgangspunktet at isolere den del af variationen i x, som ikke er korreleret med u. Men hvis instrumentet selv er korreleret med u, er det klart, at IV-analysen ikke kan give et konsistent estimat.

Hvor det første krav til IV-estimation var let at undersøge, er det straks værre med det andet. Det andet krav involverer nemlig noget potentielt uobserverbart – fejlleddet i analysens andet trin – som ikke må være korreleret med instrumentet. Derfor kan det ikke testes direkte, og her er der i praksis ingen vej uden om simpelthen at tænke sig om. Virker det plausibelt, at instrumentet udelukkende påvirker den afhængige variabel gennem x?

Hvis IV-modellen er overdetermineret (flere instrumenter end endogene uafhængige variable), kan man undersøge om ét af instrumenterne er eksogent under antagelse af, at de andre er. Tanken bag den sådanne "overidentifikations-tests" er at sammenligne IV-koefficienterne fra ét sæt instrumenter med de tilsvarende koefficienter baseret på et andet sæt af instrumenter. Hvis instrumenterne er eksogene, vil disse koefficienter ikke adskille sig meget fra hinanden. Problemet med sådanne tests er, at z_1's eksogenitet kun kan testes under forudsætning af, at z_2 faktisk er eksogen. Hvilket vi sjældent kan vide med sikkerhed. Hertil kommer, at hvis nulhypotesen (at instrumenterne er valide) afvises, kan man ikke umiddelbart afgøre, om det er z_1, z_2 eller dem begge, som ikke opfylder eksklusionsrestriktionen.

Alternativt kan man medtage sit instrument direkte i analysens andet trin. Hvis instrumentet udelukkende påvirker y via x, skal koefficienten på z være insignifikant, når det optræder sammen med x. Problemet her er imidlertid, at siden vi i udgangspunktet laver IV, må x være endogen, hvilket betyder, at alle koefficienter i analysen bliver estimeret med bias (Sovey og Green, 2011: note 4). Derfor kræves det også i dette tilfælde, at modellen er overdetermineret: På grund af endogeniteten skal x instrumenteres med z_2, hvis z_1 inkluderes direkte i modellens andet trin.

Der kræver altså altid mere end ét instrument per endogent x, hvis det skal testes, om eksklusionsretstriktionen holder – og selv hvis det er opfyldt, bør sådanne tests aldrig stå alene. Et overidentifikationsstest kan dog være et udmærket supplement til et godt argument for, hvorfor eksklusionsrestriktionen er overholdt. Det primære bør imidlertid være det gode argument.

Praktiske råd i IV-analyser

Rapporter altid F-teststørrelsen i analysens første trin

Som nævnt fungerer IV-estimation kun, hvis instrumentet er (tilstrækkeligt) korreleret med den endogene uafhængige variabel. Da dette i praksis oftest måles ved hjælp af F-testsstørrelsen i analysens første trin, bør man altid rapportere denne i sin IV-analyse. Til gengæld er R^2, som måler modellens forklaringskraft, ikke informativ i analysens andet trin.[2] Formålet med IV-estimation er altid at opnå det mest præcise estimat af den kausale effekt af *x* på *y*; det er altså en *effects of causes*-analyse. Hvis vi var interesserede i, hvordan en statistisk model som helhed forklarer en afhængig variabel (en *causes of effects*-analyse), er IV-estimation ikke anvendelig, da R^2 ikke har samme naturlige fortolkning som i OLS (fx Wooldridge, 2003: 494).

Medtag de samme kontrolvariable i første og andet trin

De samme variable skal indgå i analysens første og andet trin; 2SLS-estimatoren bliver således inkonsistent, hvis man i analysens første trin kun medtager nogle af de (eksogene) kontrolvariable fra analysens andet trin (fx Wooldridge, 2002: 93). I praksis er dette dog ikke det store problem, da de fleste statistikprogrammer (herunder STATA) af sig selv medtager alle eksogene uafhængige variable i begge analysens trin.

Brug ikke endogene kontrolvariable i IV-analyser

Af det foregående råd følger umiddelbart, at man ikke kan bruge endogene kontrolvariable i 2SLS-analyser. Hvis de medtages i første analyses første trin, er eksklusionsrestriktionen jo ikke opfyldt. Og hvis de ikke gør, er IV-estimatoren som nævnt ikke konsistent. Derfor benyttes der i IV-analyser oftes ret generiske kontrolvariable (såsom regions- og tidsdummier) samt kontroller, hvis eksogenitet er nogenlunde plausibel. I mange sammenhænge kan det alligevel være relevant at kontrollere for endogene faktorer, simpelthen for at få afgjort om den postulerede sammenhæng mellem *x* og *y* også findes, når en given faktor holdes konstant. Hvis der ikke findes gode instrumenter for disse endogene faktorer, bør man inkludere dem i en almindelig OLS-analyse, som er at foretrække frem for en misspecificeret 2SLS-analyse (jf. fx Hariri, 2012a: 486).[3]

"Intuitiv eksogenitet" er ikke "økonometrisk eksogenitet"

En faktor er endogen i en given sammenhæng, hvis den bliver bestemt inden for den model, som beskriver pågældende sammenhæng. Hvis ikke er den eksogen. Mange geografiske, topografiske eller klimatiske variable er i de fleste samfundsvidenskabelige analyser eksogene i den intuitive forstand, at de ikke

bliver bestemt af samfundsmæssige faktorer. Deraf følger imidlertid ikke, at de nødvendigvis er eksogene i økonometrisk forstand. Således kræver IV-estimation, at instrumentet er ukorreleret med alt (andet end *x*), som potentielt kunne være korreleret med *y*. Selvom forekomsten af malaria i en befolkning fx er intuitivt eksogen, er et lands evne til at imødegå og håndtere malaria en funktion af blandt andet økonomisk udvikling og sundhedsmæssig og bureaukratisk infrastruktur. Derfor kan forekomsten af malaria i en given samfundsvidenskabelig analyse godt være økonometrisk endogen. Pointen er her, at geografiske eller andre "intuitivt eksogene" faktorer ikke per definition er brugbare instrumenter. Det kræver som sagt, at de i en given sammenhæng er ukorrelerede med fejlleddet.

Pas på med laggede instrumenter

Hvis værdien på den uafhængige variabel bestemmes på t_0, og værdien på den afhængige bestemmes på t_1, kan værdien på den afhængige ikke direkte være årsag til værdien på den uafhængige. Man kunne derfor være fristet til at bruge x_{t0} som instrument for x_{t1} i analyser af effekten af x_{t1} på y_{t1}. Det er dog sjældent en god idé. Hvis der er stiafhængighed i det system, der undersøges, vil den laggede værdi af den endogene uafhængige variabel også være endogen. Det er nogenlunde intuitivt, for hvis x_{t1} er korreleret med fejlleddet, og x_{t0} er korreleret med x_{t1} (hvilket jo er stiafhængighed), så vil x_{t0} også være korreleret med fejlleddet. I situationer med stiafhængighed vil laggede instrumentvariable derfor blot skubbe et uløst endogenitetsproblem tilbage i tid, fra t_1 til t_0 (Hariri, 2012b: 192-194).

Feedback og retning på bias

Det er en gængs misforståelse, at OLS-estimatoren bliver biased i numerisk opadgående retning, hvis feedback-effekten fra *y* på *x* har samme fortegn som effekten fra *x* på *y*. At feedback-effekten fra *y* til *x* således lægges oven i effekten fra *x* til *y*, og OLS-estimatet fanger dem begge. Antag fx at effekten af demokrati på økonomisk udvikling (BNP) er positiv, og omvendt at effekten af BNP på demokrati også er positiv. OLS-estimatet af demokrati på BNP indeholder naturligvis feedback-effekten (BNP på demokrati) – men selvom de begge er positive, er OLS estimatet ikke nødvendigvis større end IV-estimatet. Retning på bias afhænger nemlig af den relative størrelse af *x* på *y* og *y* på *x*. Hvis effekten af BNP på demokrati er mindre end effekten af demokrati på BNP, så vil OLS-estimatet undervurdere den kausale effekt af demokrati på BNP.[4] OLS-estimatoren skal således snarere ses som gennemsnittet af den kausale effekt af *x* på *y* og *y* på *x* end summen af de to. Hvis feedback-effekten af *y* på *x*

er numerisk større (end x på y) og i samme retning bliver OLS-estimatet biased i opadgående retning; hvis effekten af y på x er numerisk mindre og i samme retning bliver OLS-estimatet biased nedad.

"An instrument does not a theory make"

Instrumentvariablen er en kilde til eksogen variation i den uafhængige variabel; ikke en dybereliggende forklarende variabel (Rodrik, Subramanian og Trebbi, 2004: 153). Og det er som sådan IV-estimatoren skal fortolkes. I en berømt artikel undersøgte Joshua Angrist (1990) de privatøkonomiske konsekvenser af at have tjent i det amerikanske forsvar. Analysen er vanskelig, fordi der potentielt forekommer selvselektion, således at personer med forholdsvis begrænsede indtægtsmuligheder i det civile selvselekterer ind i militæret. Angrist forsøgte at løse dette problem ved at bruge lodtrækningsnummeret til session (mere specifikt det såkaldte Vietnam draft lottery) som instrument for tjeneste i hæren. Da lodtrækningen er tilfældig, løser det problemet med selvselektion.

Pointen er her, at Angrists artikel naturligvis ikke introducerer en teori om lodtrækning til session som årsag til individuel indkomst. Lodtrækningen bruges mere beskedent som identifikationsstrategi, og generelt gælder det som sagt, at instrumentet er et middel til at identificere eksogen variation og ikke en forklarende variabel.

Pas på med genbrug

Det er som nævnt i indledningen ikke nogen helt let sag at finde brugbare instrumenter, som korrelerer (stærkt) med x men ikke med andet i modellen. Derfor ser man jævnligt instrumenter brugt i én sammenhæng blive genbrugt i en lidt anden sammenhæng. Her i Morck og Yeungs (2011: 50) fine formulering af problemet:

> A Tragedy of the Commons has led to an overuse of instrumental variables and a depletion of the actual stock of valid instruments for all econometricians. Each time an instrumental variable is shown to work in one study, that result automatically generates a latent variable problem in every other study that has used or will use the same instrumental variable, or another correlated with it, in a similar context. We see no solution to this. Useful instrumental variables are, we fear, going the way of the Atlantic cod.

Her skal det blot understreges, at hvis et instrumentet, som i et tidligere studie var stærkt korreleret med x_1, nu skal bruges som instrument for x_2, så kræver det naturligvis, at x_1 og x_2 er fuldstændig ukorrelerede. Dette vil i mange tilfæl-

de ikke være opfyldt, hvorfor eksklusionsrestriktionen ikke er opfyldt. Derfor skal man passe på med og være på vagt over for genbrug af instrumentvariable.

Brug gerne IV-estimation som et supplement til OLS
Hvis eksklusionsrestriktionen ikke er opfyldt, eller instrumentet er svagt, er IV-estimatoren både biased og upræcis. Derfor kan det i mange sammenhænge være en god idé at betragte IV-analysen som én test blandt flere. Som altid afhænger den konkrete empiriske strategi dog af forskningsspørgsmålet. Nogle gange er OLS-analysen decideret ubrugelig (som fx i Levitts analyse af effekten af politistyrken på kriminalitet); her giver det ikke meget mening at betragte OLS- og IV-estimation som komplementære. I sammenhænge hvor man har mere tiltro til OLS, kan IV-analysen bruges som et tjek blandt flere, der undersøger, hvor troværdig den kausale udlægning af resultaterne er.

Kilder til eksogen variation (eller: Hvor er de gode instrumenter?)
De formelle betingelser for IV-estimation er beskrevet ovenfor – hvor det også blev betonet, at det kan være sin sag at finde et godt instrument. Der er ingen kogebogsfremgangsmåde til at finde et godt instrument. Det kan dog anbefales at konsultere litteraturen, der forklarer den endogene uafhængige variabel. Hvis denne fx er førkolonial statsdannelse, kan man konsultere den antropologiske eller arkæologiske litteratur om statens opståen. Lidt mere generelt og til inspiration kan det nævnes, at gængse kilder til eksogen variation er tilfældigt udtrukne tal (lotterier), "naturen" (fx vejrforhold, geografi, topografi), historiske faktorer eller lovgivningsmæssige kriterier. Et eksempel på det første kunne være at bruge sessionsnummer som instrument for militærtjeneste i analyser, hvor fx indkomst eller holdninger er den afhængige variabel. Det at arbejde i militæret kan sagtens være korreleret med ikke-observerbare personlighedstræk, som også korrelerer med holdninger eller indkomst; militærtjenesten er altså ikke den eneste forskel på soldater og andre, hvorfor en simpel sammenligning giver skæve estimater af effekten af militærtjeneste. Da sessionsnumre udtrækkes tilfældigt er de en god kilde til eksogen variation (Angrist, 1990). Et velkendt eksempel på vejrforhold som kilde til eksogenitet blev givet i eksemplet fra Miguel et al. ovenfor. Et andet kunne være Feyrer og Sacerdote (2009), der bruger vind og strømforhold i Stillehavet og Atlanterhavet som instrumenter for, hvilke øer der blev koloniseret hvornår og af hvem. Administrative eller lovgivningsmæssige diskontinuiteter (fx kommunegrænser, højdekravet for at blive garder, det at børn født lige før/efter nytår starter i skole et år forskudt osv.) kan under visse omstændigheder bruges som instrument for en uafhængig

variabel af interesse. Dette er et såkaldt *fuzzy regression discontinuity*-design, som behandles andetsteds i dette nummer.

LATE: Hvad IV-estimatoren måler

IV-estimatoren måler præcist den kausale effekt af den uafhængige variabel på den afhængige variabel for det subsample, hvor værdien på den uafhængige variabel bestemmes af instrumentet. Dette subsample er velafgrænset og identificerbart, hvis instrumentet er en binær variabel. I sprogbrugen fra den nyere *treatment*-litteratur kaldes dette en *Local Average Treatment Effect*, forkortet LATE. Antag fx at vi vil estimere afkastet af uddannelse: Hvor meget stiger lønnen i gennemsnit for forskellige uddannelser. Uddannelse og indkomst er korreleret med fx evner og interesser, som er vanskelige at inkludere som kontrolvariable. Et muligt instrument kunne da være, om folk bor i en by med universitet eller ej. IV-estimatoren giver os her effekten af uddannelse på indkomsten, for de mennesker som begynder at studere på universitetet, fordi de bor i en universitetsby. Hvis instrumentet ikke er binært (men i eksemplet fx afstand i kilometer), giver IV-estimatoren os ikke et præcist kausalt estimat for et veldefineret subsample (Blundell og Costa Dias, 2009: 612). Dog gælder det som ovenfor anført, at jo bedre instrumentet i gennemsnit forklarer den uafhængige variabel (jo større F-teststørrelse i analysens første trin), jo mindre lokal og tættere er IV-estimatoren på at gælde hele stikprøven. I grænsen hvor R^2 er 1 i analysens første trin, er instrumentet jo sammenfaldende med den uafhængige variabel, således at IV-estimation, hvis antagelserne holder, giver den gennemsnitlige kausale effekt for hele stikprøven.

Opsamlende

Estimation ved hjælp af instrumentvariable er efterhånden en uundværlig del af den kvantitative værktøjskasse i empirisk samfundsforskning, hvor det er notorisk vanskeligt at afgøre, hvorvidt og hvor meget en faktor er årsag til en anden. Derfor har artiklen forsøgt at give en ikke-teknisk introduktion til og motivation for brugen af instrumentvariable.

Selvom idéen bag IV-estimation måske er simpel, er det i praksis meget vanskeligt at finde overbevisende instrumentvariable. Lidt karikeret kan man sige, at hvis eksklusionsrestriktionen er opfyldt, er instrumentet formentlig svagt. Og hvis instrumentet er stærkt, er eksklusionsrestriktionen formentlig ikke overholdt. Eller med andre ord vil de faktorer, som udelukkende påvirker *y* via *x*, ofte være så perifære i forhold til den samfundsfaglige problemstilling, at de heller ikke er specielt stærkt korreleret med *x*. Og hvis de er stærkt korreleret med *x*, hænger de formentlig også sammen med alt muligt andet i modellen.

Derfor bør estimation med instrumentvariable heller aldrig stå alene i en empirisk undersøgelse – og gør det sjældent i praksis. IV-estimation skal snarere ses som én blandt flere analyser. Hvis instrumentet er svagt, forstærkes nemlig den indbyggede bias ved IV-estimation. Og hvis eksklusionsrestriktionen ikke er opfyldt, er IV-estimatet både skævt og upræcist sammenlignet med mere almindelige metoder. IV-estimation er dog stadig en uhyre brugbar metode i empirisk samfundsforskning, for den tvinger forskeren til at fokusere på den fundamentale betingelse for kausal identifikation; nemlig hvorvidt den uafhængige variabel kan siges at være eksogen. IV-estimation tvinger så at sige samfundsforskeren til at stille de rigtige – og kritiske – spørgsmål til sin egen analyse. Samtidig gør gode IV-analyser os klogere på effekterne af de årsager, vi i politologien interesserer os for. Og præcis identifikation af, hvilke effekter en given faktor (her årsag) har, er en forudsætning for at designe og evaluere policy-tiltag i praksis. Sine svagheder til trods er der derfor ingen tvivl om, at IV-estimation er et blivende redskab i den empiriske politiologis værktøjskasse.

Noter

1. F-teststørrelsen i første trin er nogenlunde omvendt proportional med den relative bias af 2SLS-estimatoren i forhold til OLS-estimatoren. Et F på 10 svarer altså nogenlunde til en relativ bias på 10 pct.
2. Faktisk kan R^2 blive negativ i 2SLS-analyser (se fx Wooldridge, 2003: 503).
3. Formelt set skal instrumentet blot være eksogent betinget af kontrolvariablene i analysen. Der kan dog der være grund til skepsis, hvis eksklusionsrestriktionen kun hævdes at holde, når der er medtaget flere kontrolvariable. Kontrolvariablene skal jo alle selv være økonometrisk eksogene.
4. Her ser vi bort fra udeladte variable, som naturligvis også kan skabe bias i OLS-estimatet. Hvis en udeladt variabel er positivt korreleret med både *x* og *y*, skaber dette opadgående bias i OLS-estimatet.

Litteratur

Angrist, Joshua D. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *American Economic Review* 80: 313-336.

Angrist, Joshua D. og Jörn-Steffen Piscke (2009). *Mostly Harmless Econometrics*. Princeton, MA: Princeton University Press.

Blundell, Richard og Monica Costa Dias (2009). Alternative approaches to evaluation in empirical microeconometrics. *The Journal of Human Resources* 44 (3): 565-640.

Feyrer, James og Bruce Sacerdote (2009). Colonialism and modern income: Islands as natural experiments. *Review of Economics and Statistics* 91 (2): 245-262.

Hariri, Jacob Gerner (2012a). The autocratic legacy of early statehood. *American Political Science Review* 106 (3): 471-494.

Hariri, Jacob Gerner (2012b). Kausal inferens i statskundskaben. *Politica* 44 (2): 184-201.

Levitt, Steven D. (1997). Using electoral cycles in police hiring to estimate the effect of police on crime. *American Economic Review* 87 (3): 270-290.

Miguel, Edward, Shanker Satyanath og Ernest Sergenti (2001). Economic shocks and civil conflict: An instrumental variables approach. *Journal of Political Economy* 112 (4): 725-753.

Morck, Randall og Bernard Yeung (2011). Economics, history, and causation. *Business History Review* 85: 39-63.

Rodrik, Dani, Arvind Subramanian og Francesco Trebbi (2004). Institutions rule: The primacy of institutions over geography and integration in economic development. *Journal of Economic Growth* 9: 131-165.

Sovey, Allison og Donald Green (2011). Instrumental variables estimation in political science: A readers' guide. *American Journal of Political Science* 55 (1): 188-200.

Stock, James H., Jonathan H. Wright og Motohiro Yogo (2005). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics* 20: 518-529.

Stock, James H. og Mark W. Watson (2007). *Introduction to Econometrics*. New York: Pearson.

Wooldridge, Jeffrey M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: The MIT Press.

Wooldrige, Jeffrey M. (2003). *Introductory Econometrics*. South-Western: Mason, OH.

politica, 46. årg. nr. 1 2014, 95-113

Peter Bjerre Mortensen

Granger kausalitet[1]

Denne artikel giver en grundig introduktion til begrebet Granger kausalitet, der inden for statskundskaben har haft stor betydning for forståelsen af den kausale relation mellem to variable observeret over tid. Artiklen redegør for definitionen på Granger kausalitet og gennemgår, hvordan man empirisk kan teste for Granger kausalitet. Derudover diskuteres ved inddragelse af en række empiriske eksempler nogle af de begrænsninger og udfordringer, der knytter sig til analyser af Granger kausalitet. Artiklen afrundes med en perspektivering til metoder, der kan håndtere mere end to variable samt til metoder, der udvider Granger kausalitetstesten til analyser af sammenhænge både over tid og på tværs af enheder.

Siden sin introduktion for godt og vel 40 år siden har analyser af Granger kausalitet været blandt de mest populære metoder til at undersøge kausalrelationen mellem to variable. Inden for statskundskaben har Granger kausalitetsanalyser været anvendt i studier af eksempelvis dagsordensfastsættelse, hvor et centralt spørgsmål er, hvorvidt forskellige typer dagsordener påvirker hinanden. Metoderne har også været benyttet i analyser af det klassiske spørgsmål om, hvorvidt demokrati eller økonomisk velstand kommer først, samt i analyser af sammenhængen mellem partitilhørsforhold og politisk popularitet, sammenhængen mellem politisk tillid og social kapital, sammenhængen mellem indenrigs- og udenrigspolitik samt i analyser af sammenhængen mellem bureaukrati og performance. Flere af disse studier vender jeg tilbage til senere.

Artiklen indledes med en introduktion til, hvad Granger kausalitet betyder samt en kort beskrivelse af, hvordan Granger kausalitet kan undersøges ved brug af tidsseriedata. Denne introduktion følges op af nogle eksempler på, hvordan Granger kausalitet har været anvendt i politologiske analyser. Derefter følger en diskussion af en række af de væsentlige problemer og begrænsninger, man bør være opmærksom på, når man arbejder med analyser af Granger kausalitet. Artiklen afrundes med en kort perspektivering til metoder, der udvider den simple situation med kun to variable til multivariate analyser samt til metoder, der udvider Granger testen til analyser af sammenhænge over både tid og på tværs af enheder. I lyset af emnet for dette temanummer er der i artiklen valgt at fokusere på Granger kausalitet, og artiklen kan derfor ikke læses som en generel introduktion til tidsserie-analyse.

Granger kausalitet

I udgangspunktet har tidsseriedata den store fordel, sammenlignet med tværsnitsdata, at man kan adskille den forklarende og den afhængige variabel tidsligt. Som Koop (2000: 175) formulerer det: "… tiden løber ikke baglæns. Det vil sige, hvis begivenhed A indtræffer før begivenhed B, så er det muligt, at begivenhed A forårsager begivenhed B. Det er imidlertid ikke muligt, at begivenhed B forårsager begivenhed A. Med andre ord, fortidige begivenheder kan påvirke nutidige begivenheder. Fremtidige begivenheder kan ikke" (min oversættelse). Frem for blot at blive argumenteret på grundlag af teoretiske ræsonnementer giver observationer over tid altså en mulighed for faktisk at undersøge, hvorvidt der eksisterer et asymmetrisk forhold mellem årsags- og virkningsbegivenhed.

Det er denne logik, økonomen Clive Granger bygger på, da han i artiklen "Investigating Causal Relations by Econometric Models and Cross-spectral Methods" fra 1969 formulerer det, der senere er blevet kaldt Granger kausalitet.[2] Den grundlæggende idé er, at hvis vi har to variable, X og Y, observeret over tid (t) kan X_t betegnes som årsag til Y_t, hvis vi kan forklare Y_t bedre ved at inddrage fortidige observationer af X_t (dvs. X_{t-1}, X_{t-2}, …, X_{t-k}), end hvis vi kun havde inddraget fortidige observationer af Y_t (dvs. Y_{t-1}, Y_{t-2}, …, Y_{t-k}). Eller som Granger (1969: 428) formulerer det: "We say that Y_t is causing X_t if we are better able to predict X_t using all available information than if the information apart from Y_t had been used".

Til at illustrere de mulige kausale relationer mellem to tidsserier, X_t og Y_t, tager vi udgangspunkt i følgende to ligningssystemer, der beskriver to relativt enkle VAR-modeller:[3]

$$X_t = \sum_{i=1}^{n} \alpha_i X_{t-i} + \sum_{j=1}^{n} \beta_j Y_{t-j} + u_{1t} \quad (1)$$

$$Y_t = \sum_{j=1}^{n} \lambda_j Y_{t-j} + \sum_{i=1}^{n} \delta_i X_{t-i} + u_{2t} \quad (2)$$

Beskrevet på denne måde er der fire mulige kausalrelationer mellem X og Y. For det første kan der være en *ensidet kausal relation*, enten ved at X_{t-i} kan forbedre forudsigelsen af Y_t, eller ved at Y_{t-j} kan forbedre forudsigelsen af X_t. Hvis det er tilfældet, har vi en situation, hvor enten X_{t-i} Granger forårsager Y_t eller Y_{t-j} Granger forårsager X_t. Imidlertid kan der også eksistere en *feedback relation* mellem X_t og Y_t, hvor ikke blot fortidige værdier af X_{t-i} forbedrer forudsigelsen

af Y_t, men hvor også fortidige værdier af Y_{t-j} forbedrer forudsigelsen af X_t. Hvis det er tilfældet, eksisterer altså en *tosidet kausal relation* mellem X_t og Y_t. For det tredje kan de to tidsserier være *ukorrelerede* i den forstand, at inddragelse af fortidige værdier af X_{t-i} ikke forbedrer forudsigelsen af Y_t ligesom inddragelse af fortidige værdier af Y_{t-j} ikke forbedrer forudsigelsen af X_t.

Endelig bruger Clive Granger selv begrebet *samtidig* ("instantaneous") *kausalitet* om den situation, hvor samtidige observationer af X_t og Y_t er korrelerede. At benævne denne sammenhæng en kausal relation er imidlertid omdiskuteret, og i hvert fald falder det centrale argument om tidslig adskillelse af årsags- og effektbegivenhed i dette tilfælde bort. Som påpeget af Charemza og Deadman (1992: 189) eksisterer samtidig kausalitet strengt taget ikke, da der altid vil være en tidsforskydning imellem to uafhængige begivenheder. At det imidlertid i praksis kan være svært at adskille Granger kausalitet fra samtidig kausalitet, gives eksempler på senere i denne artikel.

Granger kausalitetsbegrebets popularitet hænger formodentlig sammen med, at der allerede i begyndelsen af 1970'erne blev udviklet egentlige statistiske tests, der kan bruges til at vurdere, hvorvidt der eksisterer Granger kausalitet mellem to variable. Mest udbredt inden for politologien er tilsyneladende den såkaldte "Direct Granger test" med følgende fremgangsmåde:[4] Først gennemføres en OLS-regressionsanalyse af følgende model:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_j Y_{t-j} + \text{residual}_t \quad (3)$$

Herefter gennemføres en OLS-regressionsanalyse af en udvidet model, der inkluderer laggede værdier af X:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_j Y_{t-j} + \delta_1 X_{t-1} + \ldots + \delta_i X_{t-i} + \text{residual}_t \quad (4)$$

Næste skridt er nu at teste nulhypotesen H_0: $\delta_1 = \ldots = \delta_i = 0$, hvilket implicerer, at laggede værdier af X ikke bidrager til forklaringen af Y, ud over hvad laggede værdier af Y kan forklare. Det er altså værd at bemærke, at denne Granger kausalitetstest faktisk er en test for Granger *ikke-kausalitet*. For at teste H_0 gennemføres en simpel F-test, hvor F ratioen kan beskrives ved følgende formel:

$$F = \frac{(SSR_0 - SSR_1)/i}{SSR_1/(T - i - j - 1)}$$

hvor SSR_0 er kvadratafvigelsessummen (sum of squared residuals) for model 3, og SSR_1 er kvadratafvigelsessummen for model 4. T er antallet af observatio-

ner. Hvis F-testen er statistisk signifikant, afvises H_0, hvilket fortolkes således, at X Granger forårsager Y.[5] Samme procedure kan nu blive gentaget for en test af, hvorvidt Y Granger forårsager X.

Et konkret eksempel på anvendelsen af den direkte Granger test er vist i tabel 1, der er tilpasset fra John R. Freeman's artikel i *American Journal of Political Science* fra 1983 med titlen "Granger Causality and the Time Series Analysis of Political Relationships". Ud over en grundig diskussion af fordele og ulemper ved forskellige metoder til at teste for Granger kausalitet viser Freeman (1983), hvordan den direkte Granger kausalitetstest kan anvendes til at belyse et af datidens centrale spørgsmål om, hvorvidt det var den ene eller den anden af den kolde krigs to supermagter, der drev våbenkapløbet frem. Konkret undersøger Freeman, hvorvidt fortidige værdier af de amerikanske (sovjetiske) forsvarsudgifter kan forklare nutidige værdier af de sovjetiske (amerikanske) forsvarsudgifter. Resultatet er gengivet i tabel 1 og kan fortolkes sådan, at det er de amerikanske forsvarsudgifter, der "Granger forårsager" de sovjetiske forsvarsudgifter, og dermed amerikanerne, der i den undersøgte periode (1949-1975) drev våbenkapløbet frem.

Tabel 1: Eksempel på "Direkte Granger (ikke-) kausalitetsanalyse"

Retning på kausalitet	F-test	Konsekvens
Amerikanske forsvarsudgifter →Sovjetiske forsvarsudgifter	F = 4,24**	H_0 afvises
(H_0: Fortidige værdier af amerikanske forsvarsudgifter forudsiger *ikke* sovjetiske forsvarsudgifter)		
Sovjetiske forsvarsudgifter → Amerikanske forsvarsudgifter	F = 1,40	H_0 afvises ikke
(H0: Fortidige værdier af sovjetiske forsvarsudgifter forudsiger *ikke* amerikanske forsvarsudgifter)		

Note: Tabellen viser kun et uddrag fra tabel 4 i Freeman (1983: 352-353), og der henvises til denne artikel for information om de nærmere specifikationer af analysen. Freeman benytter i denne test 2 lags for de endogene variable og 4 lags for de eksogene variable (se senere diskussion af antallet af lags). Tidsenheden er år. **p < 0,01.

Granger kausalitetstesten har også været anvendt i analyser af et andet aspekt af de militære udgifter, nemlig i relation til spørgsmålet om interaktionen mellem et samfunds militærudgifter og samfundsøkonomien, sidstnævnte målt ved eksempelvis BNP (se Dunne og Smith, 2010 for en kritisk gennemgang af denne litteratur). Spørgsmålet er, hvorvidt udviklingen over tid i den ene variabel (militærudgifter) kan hjælpe med at forudsige udviklingen i den anden

variabel (økonomien) og omvendt. Som Dunne og Smiths (2010) grundige analyser viser, er svaret på dette spørgsmål imidlertid ikke enkelt, og hovedkonklusionen er, at resultaterne er særdeles sensitive over for, hvordan de statistiske modeller specificeres (se diskussionen nedenfor).

Et helt andet genstandsfelt, hvor brugen af Granger kausalitetstests har været udbredt, er studier af dagsordensfastsættelse. Kort fortalt kan dagsordensfastsættelse forstås som en kamp om, hvilke emner der på et givet tidspunkt skal have mediernes, befolkningens og politikernes opmærksomhed (se Dearing og Rogers, 1996: 2). Det dominerende spørgsmål inden for denne tradition har været: Hvem påvirker hvem? Er det eksempelvis medierne, der påvirker befolkningens og politikernes dagsorden; befolkningen, der påvirker politikernes og mediernes dagsorden eller i virkeligheden først og fremmest politikerne, der påvirker mediernes og befolkningens dagsorden?

I kombination med tilgængeligheden af gode tidsseriedata er det oplagt, at dette forskningsspørgsmål har ført til hyppig brug af Granger kausalitetsanalyser (se fx Gonzenbach, 1992; Wood og Peake, 1998; Soroka, 2002; Walgrave, Soroka og Nuytemans, 2008; Green-Pedersen og Stubager, 2010; Green-Pedersen og Mortensen, 2010). På tværs af studierne er det imidlertid svært at identificere et entydigt svar på det overordnede forskningsspørgsmål, og konklusionen synes at være, at svaret i høj grad er betinget af værdierne på en række institutionelle og politiske 3. variable (se Walgrave og Van Aelst, 2006; Thesen, 2012).

Andre eksempler på brug af Granger kausalitetsanalyser finder man inden for studier af sammenhængen mellem demokrati og økonomisk vækst (se Heo og Tan, 2001; Baum og Lake, 2003), sammenhængen mellem partitilhørsforhold og politisk popularitet (MacKuen, Erikson og Stimson, 1989), sammenhængen mellem politisk tillid og social kapital (Keele, 2007), sammenhængen mellem indenrigs- og udenrigspolitik (Moore og Lanoue, 2003) samt et originalt studie af Meier, Polinard og Wrinkle (2000), der benytter Granger kausalitetsanalyse til at undersøge, hvorvidt det er en organisations resultater, der påvirker bureaukratiet i organisationen, eller bureaukratiet, der påvirker organisationens resultater. Granger kausalitetsanalysen i Meier, Polinard og Wrinkle (2000) peger på, at det er dårlig performance, der giver øget bureaukrati, og det resultat genfindes i Thorgaard og Andersens (under udgivelse) analyse af sammenhængen mellem karakterer og bureaukrati i den danske folkeskole.

Stort set alle disse studier bevæger sig på forskellig vis ud over blot at studere den bivariate sammenhæng mellem to tidsserier, men de bygger ikke desto mindre alle på den kausalitetsopfattelse, der mest præcist er blevet formuleret

af Clive Granger i 1969. Næste afsnit ser nærmere på nogle af de udfordringer og begrænsninger, der kan være forbundet med at anvende Granger kausalitetsanalyser i praksis.

Potentielle problemer og begrænsninger ved analyser af Granger kausalitet

Der er en række forbehold og begrænsninger, man bør være opmærksom på, når man anvender de ovenfor beskrevne Granger kausalitetstests. Først nævnes i dette afsnit nogle af de begrænsninger, der følger af, at testen bygger på OLS-regression. Derefter diskuteres en række særlige udfordringer, der knytter sig til brug af lags i Granger kausalitetsanalyser samt betydningen af tidsfrekvens og tidsforskydning i ens observationer.

Forudsætninger for den direkte Granger test

For det første kan der ofte være risiko for, at residualerne (i model 3 og 4, se ovenfor) ikke er uafhængige, og det kan derfor være nødvendigt at benytte en variant af GLS-regression, når modellerne estimeres. Denne metode tager højde for samtidig korrelation mellem residualerne og er derved mere efficient end OLS-estimeringen. Hvis residualerne ikke er korrelerede, vil GLS-estimeringen blot producere resultater, der er identiske med den enkle direkte Granger test (se Freeman, 1983: 333-334).

På grund af de laggede afhængige variable gælder for det andet, at den ovenfor beskrevne F-test kun er valid asymptotisk, hvilket betyder, at den er sensitiv over for et lavt antal observationer. Sagt med andre ord kan tidsserien være "for kort" i den forstand, at den indeholder for lidt information til en tilfredsstillende test af H_0-hypotesen. Konsekvensen af et lavt antal observationer vil som hovedregel være, at testen producerer for lave p-værdier, hvorfor man risikerer at forkaste H_0-hypotesen på forkert grundlag (se Dunne og Smith, 2010: 431).

For det tredje bygger Granger testen på en antagelse om, at tidsserierne er stationære. Det ligger uden for emnet for denne artikel at redegøre nærmere for stationaritetsbegrebet i tidsserieanalyse, herunder kriterier for svag og stærk stationaritet, men det er vigtigt at være opmærksom på denne forudsætning.[6] Antagelsen om stationaritet er ofte brudt i politologisk relevante tidsserier såsom udviklingen i de offentlige udgifter over tid, udviklingen i mængden af love og regler over tid, antallet af offentligt ansatte mv. For alle disse tidsserier gælder som hovedregel, at de udviser en klar opadgående trend over tid, og korrelerer man to tidsserier, der begge har den samme trend, er konsekvensen som hovedregel, at p-værdierne bliver lavere end de sande p-værdier – altså en risiko for, at den korrelation, man finder mellem tidsserierne, er spuriøs (se Dunne og

Smith, 2010: 431). Budskabet er, at Granger kausalitetstesten også i denne situation kan finde anvendelse, men det er afgørende, at man inden den egentlige test får håndteret eventuelle forudsætningsbrud ved hjælp af den brede vifte af redskaber, der er udviklet inden for litteraturen om tidsserieanalyse.[7]

Endelig kan modellen naturligvis være fejlspecificeret, hvis ikke der er taget højde for relevante 3. variable, der kan forårsage både X og Y (se Freeman, 1983: 330). Denne risiko for en spuriøs sammenhæng vender vi tilbage til sidst i artiklen, hvor der kort introduceres til metoder til multivariat analyse af Granger kausalitet.

Antal lags, tidsfrekvens og tidsforskydning

En særlig udfordring ved analyser af Granger kausalitet er spørgsmålet om, hvor mange lags, der skal inkluderes i analysen. Som illustreret i tabel 2, kan det få stor betydning for analysens konklusioner, hvor mange lags der inkluderes. Tabellen er tilpasset fra Dunne og Smith (2010: 432), der undersøger sammenhængen mellem USA's militærudgifter og landets BNP. Anvendes kun et års lag, tyder analysen på, at de amerikanske militærudgifter Granger forårsager landets BNP (p = 0,0342), imens H_0-hypotesen om, at BNP *ikke* Granger forårsager landets militærudgifter *ikke* kan afvises (p = 0,4959). Konklusionen bliver imidlertid præcis den modsatte, hvis der inkluderes tre lags i analysen (altså t-1, t-2 og t-3), og inkluderes to lags, er der tegn på tosidet kausalitet, hvis signifikansniveau på p < 0,10 procent anvendes.

Tabel 2: Betydningen af antallet af lags

	p-værdier for H_0-hypotesen	
	Militærudgifter → BNP (H_0: Fortidige værdier af militærudgifter forudsiger ikke BNP)	BNP →Militærudgifter (H_0: Fortidige værdier af BNP forudsiger ikke militærudgifterne)
1 lag	0,0342	0,4959
2 lags	0,0959	0,0201
3 lags	0,1684	0,0346

Note: Tabellen er tilpasset fra Dunne og Smith (2010: tabel 1) og bygger på amerikanske data for perioden 1950-2009. Tidsenheden er år.

En sådan sensitivitet over for antallet af inkluderede lags er langt fra usædvanlig (se fx Soroka, 2002). Der findes imidlertid ikke nogle enkle svar på problemet. Ofte benyttes informationskriterier såsom Akaikes Information

Criterion (AIC) og Bayesian Information Criterion (BIC) til at hjælpe med at afgøre antallet af relevante lags, men de forskellige informationskriterier giver ikke nødvendigvis samme svar (se eksempelvis Dunne og Smith, 2010: 431). Et andet hensyn kan være at inkludere så mange lags, at man undgår autokorrelation i residualerne, hvilket kan inspiceres med standard statistiske tests for autokorrelation (Soroka, 2002: 126).

Der er altså redskaber til at hjælpe, men de erstatter ikke et vist indslag af fornuft og omtanke baseret på kendskab til og substantielle betragtninger om, hvilken tidsforskydning mellem årsag og effekt man kan forestille sig. Arbejder man eksempelvis med data på mediedagsordenen, der generelt har en flygtig karakter, giver det måske ikke meget mening at inkludere lags, der går tre og fire kvartaler tilbage i tid, hvorimod det fint kunne give mening, hvis det var data på den politiske beslutningsproces, der typisk er præget af betydelig mere inerti og inkrementalisme.

Givet at man ofte har et begrænset antal observationer til rådighed, er den grundlæggende udfordring, når det gælder antallet af lags, at hvis man tilføjer ekstra lags, reducerer man risikoen for fejlspecificering og bias i resultaterne, men samtidig øger man standardfejlene, hvorved testens styrke reduceres (jf. Dunne og Smith, 2010: 431). En lav teststyrke betyder i dette tilfælde, at man har lavere sandsynlighed for at finde statistisk signifikant Granger kausalitet mellem to variable. Omvendt, hvis man inkluderer for få lags i modellen, øges sandsynligheden for at påvise Granger kausalitet, men resultatet kan være en falsk afvisning af H_0-hypotesen.

En anden udfordring knytter sig til fortolkningen af de enkelte lagkoefficienter. Granger kausalitetstesten besvarer et specifikt men noget snævert spørgsmål, og det kunne ved mange problemstillinger være interessant at se på fortegnet og den statistiske signifikans af de enkelte lagkoefficienter, der er inkluderet i modellen. Det er imidlertid ikke usædvanligt, at både størrelse og signifikans af de enkelte laggede estimater også afhænger af antallet af inkluderede lags, ligesom fortegnet på estimaterne kan variere for forskellige lags.

Et eksempel på sidstnævnte fænomen er gengivet i tabel 3, der bygger på Green-Pedersen og Stubagers (2010) analyser af sammenhængen mellem mediernes dagsorden og antallet af § 20-spørgsmål i Folketinget. Typisk vil man ikke have tilstrækkeligt præcise teoretiske forventninger til at hjælpe med at fortolke de skiftende fortegn på forskellige lags. Dertil kommer, at der ofte vil være stærk multikollinearitet imellem de laggede observationer, hvilket kan gøre det umuligt at estimere de individuelle effekter tilstrækkeligt præcist (men bemærk, at multikollinearitet inden for en gruppe af variable ikke påvirker F-testen for den samlede gruppe af variable).

Som en konsekvens af disse problemer ser man ofte, at man blot afrapporterer de summerede lagkoefficienter (se fx Soroka, 2002). I tabel 3 er de summerede lags altså beregnet som summen af Lag 1 + Lag 2 + Lag 3 + Lag 4 + Lag 5.[8] Kigger vi i tabel 3 på den samlede signifikanstest (Granger kausalitetstesten), er konklusionen på baggrund af den specifikation, der er valgt i tabel 3, at mediedagsordenen Granger forårsager spørgsmålene i folketingssalen, imens det modsatte ikke synes at være tilfældet. Dog bør det bemærkes, at en kortsigtet effekt af § 20-spørgsmål på indslag i radioavisen ikke kan udelukkes, og i den forstand kan tabel 3 også tjene som eksempel på de mulige udfordringer ved at introducere lags med ingen eller meget svag effekt (se Lag 2 til Lag 5 i venstre side af tabellen).

Tabel 3: Sammenhængen mellem mediernes og politikernes dagsorden

§ 20-spørgsmål → indslag i radioavisen		Indslag i radioavisen → § 20-spørgsmål	
Lag 1	0,014*	Lag 1	0,043†
Lag 2	-0,003	Lag 2	0,045†
Lag 3	0,009	Lag 3	-0,006
Lag 4	-0,005	Lag 4	-0,015
Lag 5	0,000	Lag 5	0,029
Summeret lags	0,015	Summeret lags	0,097
P-værdi af samlet test	0,278	P-værdi af samlet test	0,017

Note: *p < 0,05, †p < 0,10. Den samlede test er for H_0 = Lag 1 + Lag 2 + Lag 3 + Lag 4 + Lag 5 = 0. For yderligere information om analysen, se Green-Pedersen og Stubager (2010: 672).

Vedrørende kausalfortolkninger bør man desuden være opmærksom på det, man kunne kalde "pseudo lags". Ved studier af tidsserier af politiske data såsom budget- og udgiftsdata ser man ofte, at de forklarende variable er laggede 1 år ud fra det argument, at udgifter, der er registreret i år t, typisk er bestemt i år t-1. Det betyder jo så samtidig, at beslutningen om udgifterne (Y) er truffet i samme tidsinterval (år), som de forklarende variable (X) er målt. Finder man dermed, at X_{t-1} bidrager til at forklare Y_t har man de facto kun påvist "samtidig kausalitet" og altså ikke Granger kausalitet.

Denne diskussion fører videre til en anden udfordring, som man typisk står over for, når man arbejder med Granger kausalitet: Nemlig hvad tidsfrekvensen af data egentlig betyder for kausalitetsanalysen. Clive Granger kommer

selv ind på problemstillingen i sin artikel fra 1969 (s. 427): "It might be true that when quarterly data are used, for example, a simple causal model is not sufficient to explain the relationships between the variables, while for monthly data a simple causal model would be all that is required … It has been shown elsewhere […] that a simple causal mechanism can appear to be a feedback mechanism if the sampling period for the data is so long that details of causality cannot be picked out".

Mange politiske processer har en langvarig karakter, der gør, at det – modsat eksempelvis udsving i aktiekurser – ikke giver meget mening at observere dem fra dag til dag. Klassiske variable som budgetter vedtages eksempelvis typisk én gang om året, hvorfor årlige observationer synes at være den oplagte og da også oftest benyttede tidsenhed. Ofte er intervallet i tidsseriedata et vilkår, som man ikke kan ændre ved, men der kan også være tilfælde, hvor man har et valg i forhold til, hvorvidt man vil aggregere observationer til eksempelvis uger, måneder, kvartaler eller år.

At dette valg kan få betydning for resultatet af kausalitetsanalysen giver Nannestad (1999) et illustrativt eksempel på. Den substantielle problemstilling handler om, hvorvidt udviklingen i antal artikler om etniske minoriteter Granger forårsager udviklingen i befolkningens modvilje mod etniske minoriteter eller omvendt. Resultaterne ved mediedækning opgjort for henholdsvis kvartaler og måneder er gengivet i tabel 4, der er tilpasset efter tabellerne 4.2 og 4.4. i Nannestad (1999).

Den øverste del af tabel 4 viser resultatet af direkte Granger tests ved brug af 1-4 lags, hvor tidsenheden er kvartaler. Baseret på disse analyser vil konklusionen være, at de to tidsserier ikke Granger forårsager hinanden. I den nederste del af tabel 4 er Granger analyserne imidlertid foretaget med medievariablen opgjort pr. måned – befolkningens holdning er stadig målt kvartalsvis af grunde, der er nærmere redegjort for i Nannestad (1999: 138-139). Det interessante i denne sammenhæng er, at hvis et lempeligt signifikansniveau på $p < 0{,}10$ anvendes, vil man på baggrund af analyserne af månedsdata konkludere tosidig kausalitet ved brug af en måneds tidsforskydning. Ved brug af henholdsvis to og tre måneders tidsforskydning bliver konklusionen, at det er befolkningens holdninger, der Granger forårsager udviklingen i mediedækningens omfang. Ikke blot får vi altså endnu et eksempel på, at antallet af lags kan have afgørende betydning for testens udfald, men også at det anvendte interval mellem observationerne over tid har betydning. Og også i forhold til sidstnævnte kommer man ikke uden om et vist indslag af fornuft og omtanke baseret på forudgående kendskab til og teoretiske overvejelser om tidsdynamikken i de kausale processer, som man studerer.

Tabel 4: Sammenhængen mellem mediedækningens omfang og befolkningens holdninger

	Udviklingen i mediedækningens omfang → udviklingen i modviljens omfang		Udviklingen i modviljens omfang → udviklingen i mediedækningens omfang	
	Testværdi (F)	Prob. nulhypotese sand	Testværdi (F)	Prob. nulhypotese sand
Lag (kvartaler)				
1	0,0814	0,778	1,6917	0,207
2	0,0576	0,944	0,6859	0,516
3	0,0505	0,984	1,5613	0,238
4	0,6550	0,634	2,0917	0,140
Lag (måneder)				
1	4,0895	0,055	4,4587	0,046
2	2,0511	0,154	3,5566	0,073
3	1,3172	0,296	4,8633	0,039

Note: Tabellen er en gengivelse af resultaterne i tabel 4.2 og tabel 4.4 i Nannestad (1999).

Vi har i dette afsnit set på en række potentielle problemer og begrænsninger, man bør være opmærksom på, når man arbejder med Granger kausalitetsanalyser. I det næste afsnit introduceres nogle videreudviklinger af den enkle bivariate test.

Videreudviklinger af den enkle bivariate Granger kausalitetsanalyse

En oplagt risiko ved de bivariate Granger kausalitetstests er, at de er fejlspecificerede, fordi der ikke er kontrolleret for vigtige 3. variable. Det vil sige, man kan have påvist en bestemt tidsrækkefølge mellem to variable, men ikke påvist fravær af spuriøsitet. Rent teknisk er det imidlertid relativt enkelt at udvide de enkle bivariate VAR-modeller beskrevet ovenfor med flere variable. I praksis behandler disse multivariate VAR-modeller de fleste eller alle variable i modellen som endogene, og hver variabel estimeres som en funktion af tidligere værdier af den pågældende variabel samt tidligere værdier af de andre variable i modellen.[9] Tilsvarende den direkte Granger test beskrevet ovenfor benyttes

ofte F-tests til at vurdere den statistiske signifikans af enkeltvariable og/eller blokke af variable (se Freeman, Williams og Lin, 1989).

Tabel 5 giver et eksempel på, hvordan tilføjelsen af en tredje variabel kan påvirke testen af Granger kausalitet mellem to andre variable. Eksemplet er hentet fra Dunne og Smith (2010: 438) og bygger videre på tabel 2 ovenfor, men der fokuseres her på, hvordan sammenhængen mellem arbejdsløshed og militærudgifter ændrer sig, når der kontrolleres for BNP – og altså ikke på hovedsammenhængen mellem BNP og militærudgifter.[10] Som det fremgår af tabel 5, har det stor betydning for sammenhængen mellem arbejdsløshed og militærudgifter, hvorvidt et mål for BNP inkluderes i analysen. Hvis BNP inkluderes, er konklusionen, at arbejdsløshed ikke Granger forårsager militærudgifterne. Inkluderes BNP ikke, er konklusionen derimod, at arbejdsløshed Granger forårsager militærudgifterne.

Tabel 5: Betydningen af antallet af variable

	p-værdier for H_0-hypotesen
	Arbejdsløshed → militærudgifter (H_0: fortidige værdier af arbejdsløshed forudsiger *ikke* militærudgifter)
Model 1	
VAR-model hvor kun arbejdsløshed og militærudgifter indgår	0,0083
Model 2	
VAR-model hvor der kontrolleres for BNP	0,8448

Note: Tabellen er tilpasset fra Dunne og Smith (2010: tabel 2) og bygger på amerikanske data for perioden 1950-2009. Antal lags = 2 i analyserne, der også inkluderer trend.

Dunne og Smith (2010) giver en række yderligere eksempler på, at Granger kausalitetsanalyserne kan være temmelig sensitive over for antallet af variable, der inkluderes i analysen. Konklusionen er ikke, at denne sensitivitet gør Granger analyserne ubrugelige, men det er et redskab, der bør bruges med omtanke, og hvor man omhyggeligt bør gennemføre og afrapportere analyser af, hvor robuste ens resultater er i forhold til valg af antal lags, valg af variable mv. Som Dunne og Smith (2010: 439) formulerer det: "Given the large number of possible specifications and the danger of data-mining, searching for results in accord with one's beliefs; there is an issue about how results should be reported".

Eksemplet i tabel 5 kunne invitere til, at man generelt forsøger at kontrollere for et stort antal potentielt relevante variable. Problemet ved den strategi kan imidlertid være, at antallet af parametre, der skal estimeres, hurtigt bliver meget stort. Hvis der er n variable i en VAR-model med k antal lags, vil hver ligning have n*k parametre plus eventuelle eksogene variable. Hvis ikke tidsserierne er tilstrækkeligt lange, vil meget store VAR-modeller have dårlige statistiske egenskaber, og som en følge heraf vil Granger kausalitetstestene ofte være svage.

Som svar på disse udfordringer ser man i stigende grad, at Granger kausalitetstests anvendes til at analysere sammenhænge i såkaldt "Time Series Cross Section" (TSCS) data, hvor data er målt både på tværs af enheder og på tværs af tidspunkter.[11] I artiklen "Two Sides of the Same Coin? Employing Granger Causality Tests in a Time Series Cross-Section Framework", fremhæver Hood, Kidd og Morris (2008: 326), at: "Within TSCS frameworks, Granger tests generate meaningful results with significantly shorter time spans, incorporate significantly more observations, and produce more efficient results than Granger tests in conventional contexts". Dertil kommer James Stimsons (1985: 915) pointe om, at tidsserieanalyser af eksempelvis policy output – der typisk er ordnet i årlige intervaller – ofte ender med at blive studier af fortiden, fordi man bliver nødt til at gå meget langt tilbage i tid for at få observationer nok. Politik i 1950'erne var ikke nødvendigvis kendetegnet ved de samme mekanismer som politik i 00'erne, og er man eksempelvis interesseret i at forstå samspillet mellem medier og politik i nutiden, kan det være hensigtsmæssigt at forsøge at indsamle kortere tidsserier fra et større antal enheder.

Der er altså mange gode grunde til at benytte TSCS data, hvor antallet af tværsnitsobservationer til en vis grad kan kompensere for længden af tidsserier. Som Hood, Kid og Morris (2008) påpeger, er der imidlertid to væsentlige inferensproblemer forbundet med at anvende konventionelle Granger kausalitetstests på TSCS data. Det ene spørgsmål handler om at tage højde for potentiel heterogenitet i form af forskellige konstantled på tværs af enhederne. Det problem kan som hovedregel adresseres ved brug af såkaldte fixed-effect modeller, hvor der estimeres et parameter for hver enhed (se Andersen, 2007 for en introduktion til denne metode).

Det andet inferensproblem er ligeledes knyttet til spørgsmålet om heterogenitet på tværs af enheder og vedrører risikoen for enten at konkludere, at en kausal sammenhæng gælder på tværs af alle enheder, hvis det i virkeligheden kun er en del af enhederne, den gælder for, eller omvendt afvise en kausalsammenhæng på trods af, at den gælder for nogle af de observerede enheder.

Sidstnævnte problem, der vedrører en central antagelse om kausal homogenitet på tværs af enheder, er velkendt (Beck, 2007; Wilson og Butler, 2007) men ikke helt enkelt at håndtere i praksis. Med udgangspunkt i Granger kausalitetsanalyse beskriver Hood, Kid og Morris (2008: 328) imidlertid en procedure, der kan bruges til at estimere forskelle i Granger kausalitet på tværs af enheder i TSCS data. Konkret giver proceduren mulighed for at undersøge følgende kausale scenarier, som sagtens kunne være relevante for en række teoretisk vigtige problemstillinger, når man eksempelvis arbejder med typiske TSCS data såsom skole-, kommune- eller landedata:

1. Der eksisterer en identisk kausal relation mellem X og Y på tværs af alle enheder
2. Der er ikke en kausal relation mellem X og Y i nogle af enhederne
3. Der eksisterer en kausal relation mellem X og Y inden for nogle af enhederne, men karakteren af den kausale relation er ikke konstant på tværs af enhederne.

Substantielt illustrerer Hood, Kid og Morris (2008) deres metode i et Granger kausalitetsstudie af sammenhængen mellem opbakningen til det republikanske parti og afroamerikaneres politiske mobilisering over tid og på tværs af 11 amerikanske sydstater. En nærmere introduktion til studiet samt teknikken i deres metode ligger uden for denne artikel, men tilsvarende den relativt enkle direkte Granger test er testproceduren opbygget omkring en række F-tests og kan relativt enkelt implementeres i et standard statistikprogram som STATA. Substantielt tyder Hood, Kid og Morris's (2008) analyse på, at på tværs af alle sydstaterne fører politisk mobilisering blandt afroamerikanere til øget opbakning til det republikanske parti, men kun i det område, der benævnes "deep south" finder de tosidet kausalitet i den forstand, at øget opbakning til det republikanske parti også Granger forårsager øget politisk mobilisering blandt den afroamerikanske befolkning.

Konklusion

Denne artikel har givet en række eksempler på, hvordan Granger kausalitet kan testes empirisk, herunder introduceret til en række af de potentielle problemer og begrænsninger, der knytter sig til denne type analyser. Det er imidlertid værd at bemærke, at selv med "perfekte data" er der grænser for, hvad Granger kausalitet egentlig fortæller om kausalrelationen mellem to variable. Granger kausalitet handler først og fremmest om inkrementel, forbedret forudsigelse. Vejrudsigter Granger forudsiger eksempelvis vejret, men de færreste vil betragte vejrudsigterne som årsager til vejret. Der ligger altså i tilgangen en

risiko for det, James Tobin i en berømt artikel fra 1970 har beskrevet som "post hoc, ergo propter hoc", hvilket kan oversættes med *efter begivenheden, altså som følge af den*. Og denne sidestilling af kausalitet og timing, som Tobin kritiserer, kan være særlig problematisk i analyser af samfundsmæssige forhold, hvor individer såvel som kollektive aktører kan træffe beslutninger under anticipering af fremtidige forhold og begivenheder (se også Nørgaard, 2007).

I en diskussion af kausal inferens er det derfor også i denne sammenhæng relevant at understrege de tre klassiske kriterier, som en korrelation skal opfylde, før den kan betragtes som kausal inden for statskundskaben: 1) tidsrækkefølge, 2) teori og 3) fravær af spuriøsitet (se Andersen, 2010).[12] Granger kausalitet er især stærk til empirisk at undersøge tidsrækkefølgen imellem årsags- og effektvariable. Som ovenfor beskrevet kan Granger analyser også adressere spuriøsitetskriteriet gennem kontrol for relevante 3. variable, men Granger analyse bidrager i sagens natur ikke med teoretiske forklaringer. I det klassiske eksempel, hvor en dansk undersøgelse tilsyneladende påviste en sammenhæng mellem antallet af storke i et område med, hvor mange børn der blev født i samme område, kunne man sagtens forestille sig, at en Granger kausalitetsanalyse ville vise, at antallet af storke Granger forårsager børnefødsler. Da vi imidlertid ikke teoretisk kan begrunde en sådan kausal sammenhæng, kan vi måske nok påvise Granger kausalitet, men vi kan ikke bruge dette til at drage kausal inferens.

På trods af disse forbehold synes Granger kausalitetstests dog stadig at være et stærkt analyseredskab sammenlignet med tværsnitsstudier, hvor opdelingen i effekt og årsagsvariable ofte alene beror på et teoretisk ræsonnement.[13] Det synes klart, at sammenlignet med kontrollerede, randomiserede eksperimenter, vil den interne validitet i en Granger kausalitetsanalyse altid være lavere. Fastholder man imidlertid en vis interesse for observationsstudier, vil der som hovedregel være værdifuld viden at hente om kausalrelationen mellem ens tidsserier i en Granger kausalitetsanalyse. Det er formodentlig også derfor, at der godt og vel 40 år efter Granger skrev sin artikel, stadig publiceres Granger kausalitetsstudier i de bedste politologiske tidsskrifter.

Som forklaring på hvorfor begrebet om Granger kausalitet fik så stor gennemslagskraft på trods af dets begrænsninger, peger Clive Granger selv på, at diskussioner om kausalitet typisk har været præget af, hvad kausalitet ikke er, hvorfor anvendelsesorienterede økonomer/samfundsforskere hurtigt forstod at værdsætte det enkle og positivt formulerede Granger kausalitetsbegreb. Vi lader økonomen og pragmatikeren Clive Granger få det sidste ord med et uddrag fra et læseværdigt essay fra 1980, hvori han beskriver, hvordan hans egen tilgang til begrebet kausalitet adskiller sig fra mere filosofiske tilgange:

> The philosophers are not constrained to look for operational definitions and can end up with asking questions of the ilk: "If two people at separate pianos each strike the same key at the same time and I hear a note, which person caused the note I hear?" The answer to such questions is of course: "Who cares?" … One interesting aspect of the philosophers' contribution is that they often try to discuss what the term causality means in "common usage", although they make no attempt to use common usage terms in their discussion. Rather than trying to decide what the public thinks they mean by such a difficult concept as causality, it may be preferable to try to influence common usage towards a sounder definition.

Noter

1. Forfatteren takker *Politicas* anonyme bedømmere samt redaktørerne af temanummeret for nyttige kommentarer.
2. Man kan støde på betegnelsen Granger-Wiener kausalitet, da en lignende idé blev foreslået af Wiener allerede i 1956 – et slægtsskab Granger da også selv noterer i sin artikel (Granger, 1969: 428). I nærværende artikel benyttes imidlertid blot Granger kausalitet, der synes at være den mest udbredte betegnelse.
3. I den univariate autoregressionsmodel (AR-model) beskrives evolutionen i en variabel som en lineær funktion af variablens tidligere værdier. Vektor-autoregression (VAR) er en generalisering af den univariate AR-model til et N-variabel system, der beskriver hver variabel som en funktion af variablens tidligere værdier samt de tidligere værdier af de resterende N-1 variable. I dette og de efterfølgende afsnit fokuseres på VAR modeller, hvor N = 2.
4. En sammenligning af styrker og svagheder ved forskellige Granger tests kan blandt andet findes i Nelson og Schwert (1982) samt Geweke, Meese og Dent (1983). Christopher A. Sims (1972) Granger test har ligheder med en "Direct Granger test", men adskiller sig blandt andet ved, at den typisk kræver flere observationer, idet både fortidige og fremtidige observationer af X inkluderes i testen af, hvorvidt kausaliteten ensidigt går fra X til Y (se Sims, 1972: 545). Givet de mange forskellige Granger tests kan det være hensigtsmæssigt at afprøve robustheden af sine konklusioner ved at underkaste dem forskellige typer tests (se fx Freeman, 1983 samt Dunne og Smith, 2010).
5. I praksis er det særdeles enkelt at gennemføre denne test ved brug af standard statistikprogrammer såsom SPSS og STATA.
6. For en relativt tilgængelig og pædagogisk introduktion til tidsserieanalyse, herunder stationaritet og ikke-stationaritet, se Clarke, Norpoth og Whiteley (1998).
7. En populær tilgang er den såkaldte ARIMA-metode, der blev udviklet af Box og Tiao (1975). En relativt tilgængelig og pædagogisk gennemgang af metoden er

givet i Clarke, Norpoth og Whiteley (1998). Det skal dog understreges, at der også findes en række andre tilgange til at inspicere og håndtere egenskaberne ved tidsseriedata (se fx Soroka, 2002: Appendiks A for en kortfattet oversigt over litteraturen).

8. I nogle publikationer benævnes sådanne tests af summen af de laggede koefficienter "neutralitetstests" (se fx Zarnowitz, 1992).
9. Variable, der er åbenlyst eksogene – såsom trend-variable eller dummy-variable for sæson-effekter – kan også inkluderes i VAR-modeller.
10. Der henvises til Dunne og Smith (2010: 437-439) for de teoretiske argumenter bag analysen samt for en uddybende diskussion af resultaterne.
11. Der skelnes af og til mellem paneldata (flere tværsnitsobservationer end tidsperioder) og TSCS data (flere tidsperioder end tværsnitsobservationer), men her anvendes blot TSCS som en generel forkortelse for data, der er målt både på tværs af enheder og tid.
12. For en kritisk diskussion af disse kriterier, se Hariri (2012).
13. Her ses bort fra de muligheder, der knytter sig til naturlige eksperimenter og/eller kvasieksperimenter, diskontinuitetsdesigns og lignende.

Litteratur

Andersen, Lotte Bøgh (2010). Forskningskriterier, i Lotte Bøgh Andersen, Robert Klemmensen og Kasper Møller-Hansen (red.), *Metoder i statskundskab*. København: Hans Reitzels Forlag.

Andersen, Simon Calmar (2007). Multilevel-modeller : en introduktion og et eksempel. *Statskundskabens metoder* 39 (3): 294-316.

Baum, Matthew A. og David A. Lake (2003). The political economy of growth: Democracy and human capital. *American Journal of Political Science* 47: 333-347.

Beck, Nathaniel (2007). From Statistical Nuisances to Serious Modeling: Changing How We Think About the Analysis of Time-Series-Cross-Section Data. *Political Analysis* 15: 97-100.

Box, G. E. P og G. C. Tiao (1975). Intervention analysis with application to economic and environmental problems. *Journal of the American Statistical Association* 70: 70-92.

Charemza, Wojciech W. og Derek F. Deadman (1992). *New Directions in Econometric Practices*. Worcester: Edward Elgar Publishing Limited.

Clarke, Harold D., Helmut Norpoth og Paul Whiteley (1998). It's about time: Modelling political and social dynamics, i Scarbrough, Elinor og Eric Tanenbaum (red.), *Research Strategies in the Social Sciences. A Guide to New Approaches*. Oxford: Oxford University Press.

Dearing, James W. og Everett M. Rogers (1996). *Agenda-Setting*. London: Sage Publications.

Dunne, Paul J. og Ron P. Smith (2010). Military expenditure and granger causality: A critical review. *Defence and Peace Economics* 21: 427-441.

Freeman, John R. (1983). Granger causality and the time series analysis of political relationships. *American Journal of Political Science* 27 (2): 327-358.

Freeman, John R., John T. Williams og Tse-min Lin (1989). Vector autoregression and the study of politics. *American Journal of Political Science* 33 (4): 842-877.

Geweke, John, Richard Meese og Warren Dent (1983). Comparing alternative tests of causality in temporal systems. *Journal of Econometrics* 21: 161-194.

Gonzenbach, William J. (1992). A time-series analysis of the drug issue, 1985-1990: The press, the president and public opinion. *International Journal of Public Opinion Research* 4 (2): 126-147.

Granger, Clive (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37 (3): 424-438.

Granger, Clive (1980). Testing for causality. A personal viewpoint. *Journal of Economic Dynamics and Control* 2: 329-352.

Green-Pedersen, Christoffer og Peter B. Mortensen (2010). Who sets the agenda and who responds to it in the Danish parliament? A new model of issue competition and agendasetting. *European Journal of Political Research* 49: 257-281.

Green-Pedersen, Christoffer og Rune Stubager (2010). The political conditionality of mass media influence: When do parties follow mass media attention? *British Journal of Political Science* 40: 663-677.

Hariri, Jacob Gerner (2012). Kausal inferens i statskundskaben. *Politica* 44 (2): 184-201.

Heo, Uk og Alexander C. Tan (2001). Democracy and economic growth: A causal analysis. *Comparative Politics* 33: 463-473.

Hood III, M.V., Quentin Kidd og Irvin L. Morris (2008). Two sides of the same coin? Employing Granger causality tests in a time series cross-section framework. *Political Analysis* 16: 324-344.

Keele, Luke (2007). Social capital and the dynamics of trust in government. *American Journal of Political Science* 51 (2): 241-254.

Koop, Gary (2000). Analysis of Economic Data. New York: John Wiley & Sons.

MacKuen, Michael B., Robert S. Erikson og James A. Stimson (1989). Macropartisanship. *American Political Science Review* 83 (4): 1125-1142.

Meier, Kenneth J., L. Polinard og Robert D. Wrinkle (2000). Bureaucracy and organizational performance: Causality arguments about public schools. *American Journal of Political Science* 44: 590-602.

Moore, Will H. og David J. Lanoue (2003). Domestic politics and U.S. Foreign Policy: A study of cold war conflict behavior. *Journal of Politics* 65: 376-396.

Nannestad, Peter (1999). *Solidaritetens pris*. Aarhus: Aarhus Universitetsforlag.

Nelson, Charles R. og G. William Schwert (1982). Tests for predictive relationships between time series variables: A Monte Carlo investigation. *Journal of the American Statistical Association* 77: 1-18.

Nørgaard, Asbjørn Sonne (2007). God statskundskab: Heksekunst eller håndværk? *Politica* 39 (3): 233-255.

Sims, Christoper A. (1972). Money, income, and causality. *The American Economic Review* 62 (4): 540-552.

Soroka, Stuart N. (2002). *Agenda-Setting Dynamics in Canada*. Vancouver: UBC Press

Stimson, James A. (1985). Regression in space and time: A statistical essay. *American Journal of Political Science* 29 (4): 914-947.

Thesen, Gunnar (2012). When good news is scarce and bad news is good. *European Journal of Political Research* 52 (3): 364-389.

Thorgaard, Peter og Mikkel Munk Quist Andersen (under udgivelse). Hønen eller ægget – Bureaukrati og performance i den danske folkeskole. *Politica* (http://politica.dk/kommende-artikler/)

Tobin, James (1970). Money and income: Post hoc ergo propter hoc? *The Quarterly Journal of Economics* 84 (2): 301-317.

Walgrave, Stefaan og Peter van Aalst (2006). The contingency of the mass media's political agenda setting power. Towards a preliminary theory. *Journal of Communication* 56 (2): 88-109.

Walgrave, Stefaan, Stuart N. Soroka og Michiel Nuytemans (2008). The mass media's political agenda-setting power: A longitudinal analysis of media, parliament, and government in Belgium (1993 to 2000). *Comparative Political Studies* 41 (6): 814-836.

Wood, B. Dan og Jeffrey S. Peake (1998). The dynamics of foreign policy agenda setting. *American Political Science Review* 92 (1): 173-184.

Wilson, Sven E. og Daniel M. Butler (2007). A lot more to do: the sensitivity of time-series cross-section analyses to simple alternative specifications. *Political Analysis* 15: 101-123.

Zarnowitz, Victor (1992). *Business Cycles: Theory, History, Indicators, and Forecasting*. Chicago: University of Chicago Press.

Anmeldelser

Meredith Rolfe, *Voter Turnout. A Social Theory of Political Participation*, Cambridge University Press, 2012, 229 s.

Hvorfor stemmer nogle borgere oftere end andre? Så simpelt kan hovedspørgsmålet i Meredith Rolfes bog opstilles. På den måde beskæftiger hun sig med et både klassisk og hyppigt undersøgt spørgsmål indenfor samfundsvidenskaben. Og som titlen antyder, er hun særligt interesseret i de sociale faktorer, der spiller ind på vælgernes beslutning om at deltage i valgene eller blive hjemme på sofaen.

Rolfes hovedargument er, at det sociale skal sættes før alt andet, når vi ønsker at forklare valgdeltagelse og i øvrigt også mange andre former for politisk adfærd, og konsekvenserne af denne tankegang er ganske omfattende. Det gælder både specifikt for Rolfes analyser, men også for en række tidligere og kommende studier i politisk adfærd, hvis man altså vælger at følge alle hendes fodspor. Således argumenterer Rolfe for, at sociale forhold ikke blot skal ses som et supplement til individuelle faktorer, når vi forklarer politisk adfærd. Der er nemlig ingen individuelle faktorer, der påvirker borgernes stemmeafgivning i sig selv. De individuelle faktorer skal derfor forstås igennem sociale forhold. Således afhænger stort set alle borgeres valg af, hvad andre borgere foretager sig. Det er ifølge Rolfe omfanget af borgernes netværk (både i mængde og tæthed), der er afgørende for, om borgerne stemmer eller ej. De fleste borgere er nemlig villige til at samarbejde, hvis andre også gør det. Dermed kan få *first movers* igangsætte en kædereaktion, som spreder sig langt videre, end individuelle faktorer kan forklare.

I bogens første del udlægger Rolfe det teoretiske udgangspunkt og opstiller flere matematiske modeller for borgernes beslutningstagen, som hun tester gennem en række simulationer. Anden del, som er væsentlig mindre i sit omfang end første del, starter med en teorigennemgang, og herefter er der to kapitler med empiriske undersøgelser.

Rolfe lægger ud med at beskrive den, ifølge hende, konventionelle tilgang i forskningen. Her gælder det, at borgerne er fuldt oplyste, atomistiske individer, som derfor træffer komplet rationelle beslutninger. Det er denne tilgang, som Rolfe skriver sig op imod. Således er borgernes valg truffet under begrænset viden og præget af en række andre bevæggrunde end de snævert rationelle. Det er i den henseende forfriskende, at Rolfe skriver modigt og direkte og ikke er bange for at sætte tingene på spidsen. Man kan dog indvende, at Rolfes tan-

kegang om, at politisk adfærd generelt og stemmeadfærd specifikt er sociale handlinger, ikke i sig selv er så radikalt et nybrud, som man får indtryk af, hvis man kun læser hendes udlægning af den eksisterende forskning. Således er disse idéer allerede at finde tilbage i Campbells *The American Voter* (1960), og de seneste ti år er der kommet flere værker, der trækker på lignende tankegods (eksempelvis *The Social Logic of Politics* af Alan Zuckerman (red., 2005)). Dette betyder ikke, at bogen mister sin relevans, men det virker en anelse forstyrrende, at Rolfe skriver som om, at hun her har opdaget noget helt nyt, når der faktisk er en række andre, der behandler politisk adfærd med et lignende udgangspunkt.

Rolfe udlægger i første del en række forskellige argumenter, der tilsammen danner basis for hendes sociale teori om stemmeafgivelse. Her er et kapitel om borgeres betingede beslutninger, et kapitel om betinget samarbejde samt et kapitel om den sociale mening ved at stemme. Disse tre kapitler fører frem til et kapitel om, hvordan vælgeres valg er socialt betingede, hvilket endeligt munder ud i det samlede teorikapitel. Pointerne kan kortfattet sammenfattes til, at borgeres valg i dagligdagen er betinget af, hvad andre borgere foretager sig. En relativt lille andel (typisk 10-15 pct.) af borgerne vil være first movers i givne sociale situationer, som stemmeafgivning altså også hører under, og deres beslutning om at "samarbejde" (fx stemme til valg) eller ej spreder sig så som ringe i vandet i borgernes sociale netværk. Derfor er borgernes sociale netværk, hvad angår mængde af interaktioner og dybde/intensitet i disse sociale sammenhænge, den mest centrale faktor for at forklare social og politisk adfærd.

Kapitlerne baserer sig primært på en række simulationer, hvor Rolfe tester matematiske modeller. Det iøjefaldende ved denne del er således, at Rolfe ikke stiller sig tilfreds med at fremlægge sit teoretiske udgangspunkt og de antagelser, som enhver teori lægger til grund. Rolfe forsøger at dykke ned i centrale antagelser og forklare, hvordan mekanismerne fungerer. Det gælder eksempelvis, når hun udlægger, at borgernes handlinger og beslutninger er præget af begrænset rationalitet. Her inddrager hun en række tidligere undersøgelser fra blandt andet behavioristisk økonomisk forskning og tager data herfra i betragtning, når hun designer modellerne. To anker ved hendes tilgang i denne del af bogen skal her fremsættes. Substantielt er det for det første et væsentligt spørgsmål, om stemmeafgivning blot er en normal, social hverdagshandling på linje med eksempelvis dilemmaet om, hvorvidt man skal give penge til en gademusikant. Rolfe argumenterer for, at de to situationer er sammenlignelige, Men er det så simpelt og direkte overførbart? Hvis ikke, så hænger hendes præmis en anelse tyndt på dette punkt. Den anden anke handler om omfang og gentagelser. Således minder flere af kapitlerne unægtelig meget om hinanden,

og den grundige teorigennemgang kunne med fordel være kortet ned og eventuelt erstattet med flere empiriske analyser.

Den interne validitet er ganske høj i første del af bogen, og modellernes evner til at opføre sig som ventet er imponerende. For at understøtte, at modellerne afspejler sig i forhold ude i virkeligheden, inddrager Rolfe til tider data fra tidligere undersøgelser. Som læser får man dog den fornemmelse, at hun hovedsageligt inddrager de data, der passer bedst med pointerne. Således gøres der i flere tilfælde brug af en række undersøgelser fra 80'erne uden nogen forklaring på, hvorfor det netop er disse undersøgelser, der bruges. Og dette selvom tilsvarende undersøgelser er gentaget flere gange senere hen. Dette er særligt problematisk i forhold til den eksterne validitet, når Rolfe argumenterer så kraftigt for, at de sociale netværk er afgørende. For hvordan har borgernes sociale netværk ændret sig siden eksempelvis 1985? Som læser tænker jeg her især på internettets fremkomst, og hvordan det kan tænkes at påvirke sociale netværks størrelser og intensitet. Men Rolfe forholder sig på intet tidspunkt til internettet, selvom der er lavet en række undersøgelser af, hvordan nettets sociale netværk påvirker borgernes sociale og politiske adfærd (mest kendte eksempel på dette er nok Christaki og Fowlers bog *Connected* fra 2009).

I anden del af bogen, som afhængigt af definitioner udgør to-tre ud af bogens ni kapitler, prøves teorien af på nogle cases. Den mest iøjefaldende konklusion i denne del af analysen er, at mobiliseringstiltag er afgørende for, at borgerne overhovedet stemmer til valg. Således er der en god portion determinisme, når Rolfe skriver, at "få borgere, hvis nogen overhovedet, ville stemme, hvis ikke politikere lavede mobiliseringsaktiviteter" (s. 107, egen oversættelse). Derfor er det for valgdeltagelsens skyld centralt, at der er politiske kampagner omkring valgene, men indholdet af kampagnerne er ligegyldigt. Kampagner virker nemlig ved, at de øger diskussionen indenfor netværkene, som på den måde øger stemmeprocenten. Man kan så spørge, om nogle kampagneformer og budskaber skaber mere opmærksomhed og diskussion end andre, men det forholder Rolfe sig ikke til.

I anden del af bogen dedikerer Rolfe et selvstændigt kapitel til forholdet mellem uddannelse og politisk adfærd. Ifølge Rolfe er den konventionelle visdom, at mere uddannelse fører til øget valgdeltagelse. Men grunden til, at forskere ofte finder en korrelation mellem uddannelse og valgdeltagelse, er, at uddannelse er med til at udvide folks sociale netværk, hvad angår både mængde og tæthed. Det er derfor en fejlslutning, at det er uddannelsens påvirkning på borgernes viden om samfundsforhold, der øger sandsynligheden for, at universitetsuddannede stemmer hyppigere end borgere med lavt uddannelsesniveau. Sandsynligheden for at stemme er ifølge Rolfe ens på tværs af alle uddannel-

ses- og indkomstgrupper, men varierer afhængigt af borgernes sociale netværk og mængden af politisk diskussion i disse grupper. Det er absolut interessant, at uddannelse på den måde er en proxy for socialt netværk, og Rolfe gør en del ud af, at hendes pointe på dette punkt er banebrydende. Kritikeren vil dog indvende, at flere tidligere undersøgelser også har peget på, at sammenhængen mellem uddannelse og politisk deltagelse i mere eller mindre grad faktisk handler om socialisering og ikke udelukkende om den individuelle påvirkning, som borgeren udsættes for gennem uddannelse.

Samlet set har Meredith Rolfe skrevet en bog, der har sine klare styrker i sin teoretiske tilgang og udbygning. Bogens udgangspunkt og konklusioner er i mine øjne ikke så banebrydende, som Rolfe gerne selv vil gøre det til, men mindre kan også gøre det. Hvis man er til nyere og grundige empiriske analyser eller har brug for en introduktion til feltet, vil man næppe få stor glæde af Rolfes bog. Men hvis man kender lidt til feltet i forvejen og trænger til en grundig og relativt matematisk-teoretisk tilgang til social adfærd og valgdeltagelse, er bogen bestemt et relevant bud at gå til.

Jonas Hedegaard Hansen
Ph.d.-studerende
Institut for Statskundskab
Københavns Universitet

Rune Stubager, Kasper Møller Hansen og Jørgen Goul Andersen, *Krisevalg. Økonomien og folketingsvalget 2011*, København: Jurist- og Økonomforbundets Forlag, 2013, 214 s.

Denne bog, skrevet af tre fremtrædende valgforskere fra hver sit universitet, er et led i rækken af publikationer fra Det Danske Valgprojekt. Dette projekt begyndte med folketingsvalget 1971 og strækker sig derved over en periode på 40 år og med 16 folketingsvalg. Det skal dog siges, at ikke alle valgene er lige grundigt undersøgt – for eksempel har der været fyldige analyser af folketingsvalgene i 2001 og 2005, men ikke af 2007-valget. Men da interviewmaterialet er tilgængeligt for alle valgene, er der rige muligheder for ph.d.-studerende og andre interesserede til at supplere den eksisterende forskning op med nye analyser.

Denne bog har imidlertid et andet fokus end de foregående. Igennem en række år har vi vænnet os til, at folketingsvalgene udfolder sig som en strid mellem venstresiden og højresiden. Striden har stået om både fordelingspoliti-

ske og værdipolitiske mål, og ikke mindst de sidstnævnte har delt vælgerne ved de senere års valg. Folketingsvalget i september 2011 kunne derimod ventes at sætte disse ideologiske holdninger på vågeblus som følge af den økonomiske krise, der satte ind i 2008. I stedet kunne en regering, der præsterede så dårlige økonomiske nøgletal, ventes at blive straffet, især af de vælgere der havde lidt de største økonomiske tab.

Som bekendt tabte regeringen da også valget, men kun med en meget lille margen, eftersom den blå blok fik 49,7 procent af stemmerne. Den blev altså ikke straffet særlig hårdt – eller også blev den straffet, men så var der faktorer, der trak i modsat retning. Det gør det på en måde ekstra spændende at følge med i forfatternes analyser, når der tegner sig en sådan komplikation forude.

Forfatterne omtaler flere mulige forklaringer: 1) at vælgerne ikke mærkede så meget til krisen, måske fordi velfærdssystemet afbødede følgerne af den; 2) at vælgerne betragtede krisen som udefra kommende og ikke gjorde regeringen ansvarlig; 3) at en rød regering ikke blev anset for at kunne klare økonomien bedre end den nuværende; eller 4) at den blå regering på andre områder blev set som mere kompetent end en rød regering.

Næsten som i en god krimi skal vi igennem en længere miljøskildring og forhistorie for at komme frem til en løsning. Denne forhistorie har form af en klassisk model, den såkaldte årsagstragt eller "kausalitetstragt" (*funnel of causality*), der ser vælgeren som styret af forskellige langtids- og korttidsfaktorer frem mod beslutningen på stemmetidspunktet.

Faktisk skal vi hen til midten af bogen, før vi kommer til det afgørende kapitel 5 omhandlende økonomisk stemmeadfærd. Forinden gennemgår Jørgen Goul Andersen i kapitel 2 den økonomiske udvikling, hvor man bider mærke i, at såvel Fogh Rasmussen-regeringen som de danske nyhedsmedier var meget uvillige til at erkende dybden af den økonomiske krise – en nedtur på 6,4 pct. i nationalindkomsten fra toppen i 2007. Måske derfor var der selv i 2011 over dobbelt så mange, der var bange for ikke at kunne blive plejet tilstrækkeligt, når de blev gamle, som der var af dem, der var bange for at miste deres arbejde.

Kapitel 3 rummer et tiltrængt gensyn med de sociale baggrundsfaktorer, som sidst blev undersøgt ved 2001-valget. Vi noterer os, at klasseposition fortsat betyder mindre og mindre, hvorimod kønsforskelle i partivalget fortsætter med at vokse – 58 pct. af kvinderne mod kun 44 pct. af mændene stemte rødt. Endnu større er dog forskellen mellem offentligt ansatte, hvor 70 pct. stemte rødt, og privatansatte, hvor det kun var 39 pct. Som noget nyt har de danske interviews også stillet spørgsmål om svarpersonernes formueaktiver i form af ejerbolig, sommerhus, opsparing, forretning, værdipapirer og udlejningsejendomme. Det viser sig at være en effektiv prediktor af partivalget, idet både

Venstre, Konservative og Liberal Alliance står betydeligt stærkere blandt dem, der havde flere typer aktiver, end blandt dem der ikke havde aktiver. Det gælder især aktiver forbundet med høj risiko, nemlig de tre sidstnævnte typer.

De samme sociale faktorer, som farver partivalget rødt, virker også ind på de politiske holdninger af den klassiske venstre-højre type. Selvom partierne her forsvinder ud af billedet, er paralleliteten temmelig slående. Som det fremgår af kapitel 4, er kvinder mere venstreorienterede end mænd, personer uden formue mere venstreorienterede end folk med risikable aktiver, og den berømte 68-generation er også ved overgangen til pensionsalderen mere venstreorienteret end andre generationer. Interessant er det at konstatere, at klassesamfundet, som i kapitel 3 blev erklæret dødt på det partipolitiske niveau, i kapitel 4 genopstår på det ideologiske niveau. I en regressionsanalyse vises det, at selvstændige og funktionærer er højreorienterede, arbejdere derimod venstreorienterede, på et indeks bestående af syv ideologiske holdninger. Hvorfor slår det ikke igennem i partivalget? Antagelig, skriver forfatterne (s. 112), fordi der skabes et krydspres mellem de klassiske holdninger og de nyere værdipolitiske holdninger. Disse sidstnævnte samles dog først op henimod slutningen af bogen. For det var jo ikke det, valget handlede om, vel?

Scenen er nu sat for en analyse af virkningerne af den økonomiske krise. Her viser det sig snart, at regeringen faktisk blev straffet. Der var nemlig 78 pct. af vælgerne, der mente, at samfundsøkonomien var blevet dårligere, mod kun 7 pct. der mente, at den var blevet bedre; og blandt de førstnævnte fik den røde blok mere end halvdelen af stemmerne. Kunne dette ikke forklare regeringens tilbagegang? Jo – måske. Forfatterne opererer med nogle "forudsagte sandsynligheder" for at stemme blåt, når andre uafhængige variable sættes til deres gennemsnit. Disse sandsynligheder må ses som et forsøg på at anskueliggøre effekterne i logistiske regressionsmodeller, et tilbagevendende problem i nyere vælgerundersøgelser. Men det er et spørgsmål, om læseren får det lettere med disse fiktive tal. For eksempel beregnes det i tabel 5.5, at 65 pct. af de vælgere, der mente, at samfundsøkonomien var blevet "meget bedre", forudsiges at stemme for den blå blok. Men det er en ringe trøst for Lars Løkke Rasmussens regering, for sådanne vælgere eksisterede ifølge tabel 5.1 simpelthen ikke.

Selvom valget endte med at være meget lige, må det ikke glemmes, at regeringspartierne V+K gik tilbage med 5 pct. af vælgerne fra 2007 til 2011. Halvdelen af denne tilbagegang kan tilskrives almindelig træthed med en regering (*cost of ruling*, som omtales i kapitel 6), men det levner dog plads for en reel effekt af den økonomiske krise. Effekten blev bare ikke nær så stor, som den tegnede til få måneder før valget. Denne dæmpning af effekten kommer let til at lægge beslag på opmærksomheden, når man forventer et jordskredsvalg.

Den økonomiske effekt viser sig endvidere at være en effekt af samfundsøkonomiens udvikling, ikke udviklingen i den enkelte vælgers private økonomi. Det er man efterhånden vant til, for det samme har været tilfældet ved de foregående valg, ligesom det gælder i de fleste andre lande. Der er dog noget, der tyder på, at man ved at stille svarpersonerne over for et valg mellem et privatøkonomisk gode (i form af en skattelettelse) og et samfundsøkonomisk gode, for eksempel en lavere pensionsalder, finder en helt forskellig reaktion hos rød og blå bloks gennemsnitsvælger. Undersøgelsen heraf ligger i forlængelse af Kasper Møller Hansens artikler (sammen med Mickael Bech) og tilhører en ny og mere eksperimenterende tilgang til vælgerforskning. Den første af disse artikler går tilbage til 2007 og beregnede blandt andet, at lavindkomstvælgere ville betale 4293 kr. om måneden for at få Fogh snarere end Lykketoft som statsminister! Denne gang virker beregningerne mere troværdige, selvom det stadig er et problem at gøre det forståeligt for den almindelige læser, præcis hvordan ræsonnementet forløber. Afsnittet burde fylde mere, end det gør. Og det er lidt synd, for det er et meget spændende nyt indslag i vælgerforskningen.

Kapitel 6 og 7 diskuterer de øvrige forklaringer på, at regeringen ikke tabte et jordskredsvalg som følge af krisen. Allerede i kapitel 5 har vi fået at vide, at næsten ingen vælgere anede, hvor stor nedgangen i nationalproduktet havde været. Nu viser der sig flere sider af krisebevidstheden. Vurderingen af Danmarks økonomi begyndte at gå nedad allerede i begyndelsen af 2008, hvorimod der gik et par år, før vælgerflertallet mente, at krisen var "alvorlig". Man kan således ikke beskylde de danske vælgere for at hidse sig op eller gå i panik. Krisen spillede da heller ikke ind i den gennemsnitlige vurdering af privatøkonomien, som ifølge figur 6.1 ikke på noget tidspunkt blev negativ. Utroligt når man tænker på faldet i huspriserne, de fallerede byggefirmaer og landmænd, den stigende arbejdsløshed m.m.; men det forklarer måske, hvorfor vælgerne ikke var i oprør. Som ved et trylleri var krisen forvandlet til et offentligt underskud, og nu måtte der spares. Det accepterede vælgerne – som noget ret usædvanligt var der i 2011-undersøgelsen et stort flertal, der mente, at der ikke var råd til lønforhøjelser, samt et endnu større flertal, der mente at der ikke var råd til skattelettelser. Med hensyn til den sociale velfærd tegnede der sig et flertal for at afskaffe efterlønnen og nedjustere forventningerne generelt. Men alt det kom ikke den to år gamle Løkke Rasmussen regering til skade, for det var jo holdninger, den selv gav udtryk for.

På denne måde slap regeringen så godt igennem krisestemningen, at den sandsynligvis havde kunnet blive siddende, hvis valget var blevet udskrevet fjorten dage senere. Det kan undre nogen, men minder det ikke om det, man husker fra tidligere krisevalg? Poul Hartling havde ved valget i 1975 og Anker

Jørgensen i 1977 og 1979 ligefrem succes blandt vælgerne med at styre landet under en krise. Denne gang var det tæt på at gå på samme måde.

Bogen slutter med et kapitel, hvor nogle vigtige faktorer samles op og hældes ned i "kausalitetstragten". Det gælder for det første værdipolitikken, der egentlig hører hjemme højere oppe i tragten, og for det andet partilederne, der hører hjemme tæt på tragtens udløb – blandt andet debuterede Lars Løkke Rasmussen jo som partileder og statsminister. Med hensyn til holdningen til indvandring og andre værdipolitiske emner er der igennem 20 år gravet grøfter, som ikke let kan sløjfes igen, krise eller ikke krise. Værdipolitikken bidrager da også stadig, side om side med fordelingspolitikken, til at placere vælgerne i de respektive partier. Den passer blot ikke så godt med den rød/blå opdeling, som er et gennemgående træk i bogen.

At partilederne har en effekt på vælgerens partivalg er indiskutabelt, men den kan måles på flere måder og falder højst forskellig ud. Interviewundersøgelserne spørger om, hvor stor sympati man har for de forskellige ledere, og det er jo næsten som at spørge om, hvilket parti man stemmer på. Men undersøgelserne spørger også om, hvor stor sympati man har for de forskellige partier. Kontrollerer man for denne faktor, som det gøres her, bliver effekten af lederen beskeden. Mon ikke der senere bliver mere at sige om Lars kontra Helle? Der er temmelig mange bolde i luften her til sidst, og det gælder i det hele taget, at de gode hensigter fra forordet om at nå ud til et "bredere publikum" nok ville tilsige en større grad af enkelhed og ensartethed bogen igennem. De kapitler, der er lettest at læse, er pudsigt nok dem, hvor partierne forsvinder ud af billedet, som kapitel 2, 4 og 6.

Der er masser af interessante data, hypoteser og analyser i bogen for dem, der vil fordybe sig i de enkelte problemstillinger. Litteraturhenvisninger er anført i slutningen af hvert kapitel, hvilket er belejligt, hvis man vil fotokopiere kapitlet (med behørig tilladelse selvfølgelig). I det hele taget viser bogen, at det danske valgprojekt, trods sit 40 års jubilæum, er mere levende end nogensinde.

Ole Borre
Professor emeritus
Institut for Statskundskab
Aarhus Universitet

Thomas Risse, Stephen C. Ropp and Kathryn Sikkink (eds.), *The Persistent Power of Human Rights: From Commitment to Compliance*, Cambridge University Press, 2013, 372 p.

Building on their 1999 book, *The Power of Human Rights*, Thomas Risse, Stephen Ropp and Kathryn Sikkink are continuing their enquiry into the diffusion of human rights norms in *The Persistent Power of Human Rights: From Commitment to Compliance*. The first book proposed a "spiral model" of human rights change, which has since been widely applied by scholars seeking to understand why countries move from a state of rights violation to that of rights protection. The model entailed a five-phase transformation of state practice: 1) state repression; 2) a denial that human rights exist or apply to its actions; 3) tactical concessions that implicitly recognise the validity of human rights; 4) prescriptive status which signals a rhetorical and formal commitment to human rights; and 5) "rule-consistent behaviour" whereby the state meets its various human rights obligations.

The book focuses on the last three phases of the spiral model, which are interpreted as a movement from commitment to compliance, from "talking the talk" to "walking the walk". To guide this investigation of changing norms and practice, the book poses the following research question: Under what conditions and by which mechanisms will actors – states, transnational corporations, other private actors – make the move from commitment to compliance? This question broadens the study to include non-state actors and enables the contributors to explore a set of social mechanisms and scope conditions. But the focus on the differences between human rights commitment and compliance also frames the problem of human rights violations as primarily a legal problem.

The book has four parts. In Part I, two chapters review recent qualitative and quantitative research that tests the spiral model. Part II discusses and develops the study's core conceptualisations: the mechanisms, one of the scope conditions, and the notions of commitment and compliance. Part III presents three case studies of states (the US, China, and Tunisia and Morocco) and a chapter that investigates whether treaty ratification and UN criticism lead to compliance. The last part moves beyond the state to analyse transnational companies, rebel groups and family practices.

To theoretically develop tools for the analysis of the commitment/compliance move, the introductory chapter (by Thomas Risse and Stephen C. Ropp) proposes four mechanisms of socialisation as well as five scope conditions which mediate their impact. The mechanisms are coercion, incentivization,

persuasion and discourse, and capacity building. Ryan Goodman and Derek Jinks (Ch. 6) offer an internal critique of the assumption that the four social mechanisms complement each other in the movement towards human rights compliance. To them, the combination of several mechanisms may lead to "crowding-out effects" whereby "the operation of one mechanism of influence might undercut the operation of another" (p. 105). For instance, coercion or material rewards for compliance may undercut a sense of self-determination and sovereignty, thereby generating a political backlash.

The ways the different mechanisms may have adverse effects is a useful contribution which could explain why anti-human rights discourses sometimes have popular appeal. For instance, in March 2013 the majority of Kenyan voters elected as president a suspected perpetrator of crimes against humanity. And in Uganda, a popular parliamentary bill proposes capital punishment for homosexuality, while the human rights perspective is branded as Western neo-imperialism. None of the book's case studies, however, consider such possible crowding-out effects, with the result that Chapter 6 is left curiously hanging. Instead, anti-human rights "counter-discourses" in China and the US (Ch. 8 and 9) seem inexplicable.

The five scope conditions are presented as deriving mainly from the shortcomings of the 1999 book. These are regime type, the nature of statehood, the control over rule implementation, the extent of material vulnerability to external and internal pressure, and the extent of social vulnerability to the same. The idea is that these characteristics pertaining to state and non-state actors condition the impact of the various mechanisms and explain why the spiralling towards compliance happens, stalls, reverses, or stops. The scope conditions are defined as a set of binary choices between democracy/authoritarianism, consolidated/limited statehood, centralised/decentralised rule implementation, material vulnerability/resistance, and social vulnerability/resistance.

In Chapter 4 on statehood, Tanja A. Börzel and Thomas Risse take their point of departure in the observation that consolidated statehood forms the exception rather than the rule in the international system. They argue that human rights promotion and research implicitly assume that states are consolidated and therefore unwilling to protect human rights. In fact, the human rights violations derive from state and non-state actors which cannot be controlled by the state because it lacks capacity. This, they argue, requires a re-conceptualization of the human rights agenda: "positive incentives, sanctions or persuasion will not do the trick, but have to be matched by institution- and capacity-building" (p. 83). This policy recommendation encapsulates the thinking about the state in the spiral model and many of the chapters: The

state is the receiver of and reactor to domestic and international pressures and inducements. By and large, its actions can be explained by the absence, presence and nature of this transnational human rights movement.

Thus, the differences in human rights compliance in Guatemala and Georgia "can be explained by the variation in exposure to the international (Western) community as well as by the fact that Western organizations have invested many more resources in capacity-building in Georgia than in Guatemala" (p. 81). In Indonesia, "[p]rogress has indeed been made" since ratification of the Torture Convention "but in the context of intense [UN] monitoring" (Ch. 7 by Ann Marie Clark, p. 134). Moving from commitment to compliance, Morocco "responded to the continued mobilization of transnational networks and an increased engagement of international actors by appropriating the human rights discourse" (Ch. 10 by Vera van Hüllen, p. 198).

Ch. 8 on the US (by Kathryn Sikkink) suggests, however, that this reactive story is not all there is to state practice in the area of human rights. Her examination of the case of officially sanctioned torture and "extraordinary renditions" during the War on Terror provides a more nuanced account of the state. As the government was not morally or materially vulnerable to domestic and international pressures, these had little impact in stopping the human rights violations. Instead, torture and kidnappings caused "fierce opposition within the [Bush] administration", i.e., among different government departments and professional groups. The US case furthermore illustrates, according to Sikkink, that "even a quite firm commitment to international law, signaled by ratification and implementation in strong domestic statutes, can be undermined by a relatively small group of powerful political operators in the context of a security threat, a compelling anti-terrorism discourse, and domestic indifference to the rights of others" (p. 162).

Although not proposed by Sikkink, her analysis of the US case suggests that the spiral model is perhaps less a path towards (or from) compliance, and more a spectrum of practices that can co-exist within a larger regime. While this would remove the spiral from the spiral model, the chapter on rebel groups in effect shows this. It also illustrates the difficulties of empirical investigation if you cannot or will not speak to those you study.

Hyeran Jo and Katherine Bryant's chapter on rebel groups (Ch. 13) has difficulties translating the concepts of commitment and compliance into the practices of outlawed insurgents. To test the move towards compliance, they operationalise commitment and compliance by the number of civilian deaths and rebel-granted visits to the civilian population by the International Committee of the Red Cross. Commitment to human rights entails the granting of

visits and killing of more than 25 civilians in a year, while "[g]roups that grant full visits and kill fewer than twenty-five civilians" comply with human rights law. Jo and Bryant, then, do find that some rebel groups comply with humanitarian norms. But as insurgent violence is most often brutal, the identified and narrowly defined pockets of compliance are perhaps better seen as practices that co-exist in a tense and contradictory relationship with intense violence (i.e., phase 1, repression). It is also difficult to know if the rebels killed less than 25 civilians because of a belief in human rights or government protection.

In the concluding chapter, Risse and Sikkink summarise the volume by asking whether "actors [are] moving to greater compliance with international human rights law, and if so, *why and how* are they doing so" (p. 275, emphasis in original). While many of the chapters conclude in the affirmative, they do not explain why this development is occurring. This may stem from the book's legalistic approach to human rights: As the law is per se legitimate, it does not anticipate an interpretation of non-compliance. Thus it is not clear why rebels sometimes comply or why some families abandon repressive practices. It is not even clear why the Bush administration legalized torture or why Indonesia reduced it. While the book suggests that the power of human rights persist, readers are not made to understand the source of this power.

Line Engbo Gissel
Department of Political Science and Government
Aarhus University

Abstracts

Jens Blom-Hansen and Søren Serritzlew

Endogeneity and experiments – research design as solution

There are many theories within political science on the effect of independent variables on dependent ones. Unfortunately, it is often difficult to establish whether an empirical relationship is causal. Whether found in a quantitative or a qualitative study, empirical relationships may be due to more than effects of x on y. It may also be due to y having effects on x. When this is the case, there is an endogeneity problem which means that causal inference cannot be made directly from the data. In this situation, it is relevant to consider an experimental solution. In the article we discuss these issues and provide examples of the most important experimental designs, namely the lab experiment, the natural experiment, the field experiment, the survey experiment, and the quasi-experiment.

Derek Beach

Process tracing and the study of causal mechanisms

This article contends that there are large methodological advantages in tracing mechanisms linking a given cause (or causes) with an outcome. Causal mechanisms are the theoretical process that link X and Y together. By tracing mechanisms, we gain knowledge about how X contributes to producing Y. There are three advantages of PT. First, PT enables us to make strong inferences about X being causally related to Y because we gain detailed within-case evidence of the process that links X and Y together. Second, we gain a better understanding of the process linking the two. Finally, when tracing mechanisms we do not need to control for other confounding causes. However, the disadvantages of PT are both on the practical-level (enormously demanding to undertake), and the limited ability to generalize findings from the single case.

Asmus Olsen

Assignment variables and assignment values: An introduction to the regression discontinuity design

The regression discontinuity design (RDD) exploits empirical settings where we observe a variable which at a given value separates the observations we study into a control or treatment group of interest. RDD has shown ability to reproduce results similar to experimental benchmarks while at the same time being

intuitively simple and relatively easy to implement in most statistical software. However, RDD has only very recently been applied in political science. A central criticism of RDD has been that actual empirical settings seldom offer the necessary conditions for the use of RDD. Here the logic of RDD is introduced to scholars and students in political science with focus on how RDD is useful in terms of answering a diverse set of causal questions in political science.

Mogens Kamp Justesen and Robert Klemmensen

Comparison of comparable observations: Causality, matching and observational data

This article provides an introduction to matching and discusses the strengths and weaknesses of matching methods in studies of causality. Matching is mainly a method that enables us to increase the comparability of observations on observed variables. Matching methods do not enable us to draw causal inferences based on observational data. If we want to estimate causal effects, matching works best in combination with a strong design – such as a natural experiment or a quasi-experiment. We also provide an example of how matching can be used to analyze quasi-experimental data. To this end we use survey data to analyze how a major environmental disaster caused by an unanticipated explosion on the British-owned oil rig Deep Water Horizon affected attitudes to environmental issues in the British population.

Jacob Gerner Hariri

A non-technical introduction to instrumental variables estimation

This article provides a non-technical introduction to and motivation of instrumental variable (IV-)estimation. IV-estimation is immediately motivated by the complex nature of what political scientists study: Variables are associated in many different ways and it is often far from obvious what is cause and what is effect. The purpose of IV-estimation is to cut through the complexity by teasing out a source of exogenous variation in the independent variable in an empirical analysis.

Peter B. Mortensen

Granger causality

This article gives an introduction to the concept of Granger causality, which has had a great influence on the understanding of the causal relationship between two variables observed over time. The article accounts for the basic definition of Granger causality and provides several examples of its uses in the discipline of political science. Furthermore, based on a range of empirical studies, the article provides a critical discussion of the challenges and problems involved in conducting Granger causality analysis. The article concludes with a discussion of how the basic Granger analysis can be extended to include more than two variables as well as to include observations in both time and space.

Om forfatterne

Jens Blom-Hansen, professor i offentlig forvaltning, Institut for Statskundskab, Aarhus Universitet. Interesserer sig for forholdet mellem embedsmænd og politikere, stat/kommune-forholdet og EU-politik. E-mail: jbh@ps.au.dk

Jacob Gerner Hariri, postdoc, Institut for Statskundskab og Økonomisk Institut, Københavns Universitet. Er ph.d. i statskundskab fra Københavns Universitet og har blandt andet publiceret i *American Political Science Review* og *British Journal of Political Science*. E-mail: jgh@ifs.ku.dk

Mogens Kamp Justesen, ph.d., lector, Department of Business and Politics, Copenhagen Business School. Hans forskningsinteresser vedrører bland andet spørgsmål om demokrati og økonomisk udvikling. Blandt nyere publikationer er bidrag i *Comparative Politics* og *Electoral Studies*. E-mail: mkj.dbp@cbs.dk

Robert Klemmensen, ph.d, professor (mso), Institut for Statskundskab, Syddansk Universitet. Har bidraget med artikler i tidsskrifter som *Comparative Political Studies*, *Comparative Politics* og *Journal of Theoretical Politics*.
E-mail: rkl@sam.sdu.dk

Peter Bjerre Mortensen, ph.d., professor (mso) i statskundskab, Institut for Statskundskab, Aarhus Universitet. Forsker blandt andet i partikonkurrence, dagsordensfastsættelse, bureaukratisering og effekter af NPM-reformer. Blandt de seneste publikationer er artiklerne "Government Responses to Fiscal Austerity: The Effect of Institutional Fragmentation and Partisanship", med Carsten Jensen (under udgivelse i *Comparative Political Studies*) og "Avoidance and Engagement Issue Competition in a Multi Party System", med Christoffer Green-Pedersen (under udgivelse i *Political Studies*). E-mail: peter@ps.au.dk.

Asmus Leth Olsen, cand.scient.pol., ph.d., adjunkt, Institut for Statskundskab, Københavns Universitet. Forsker i krydsfeltet mellem politisk psykologi og den offentlige sektor med afsæt i eksperimentelle metoder. Har senest publiceret i tidsskrifterne *Judgment and Decision Making* og *Public Choice*.
E-mail: ajlo@ifs.ku.dk

Søren Serritzlew, professor i offentlig forvaltning, Institut for Statskundskab, Aarhus Universitet. Interesserer sig blandet andet for effekten af reformer i den offentlige sektor. E-mail: soren@ps.au.dk