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.

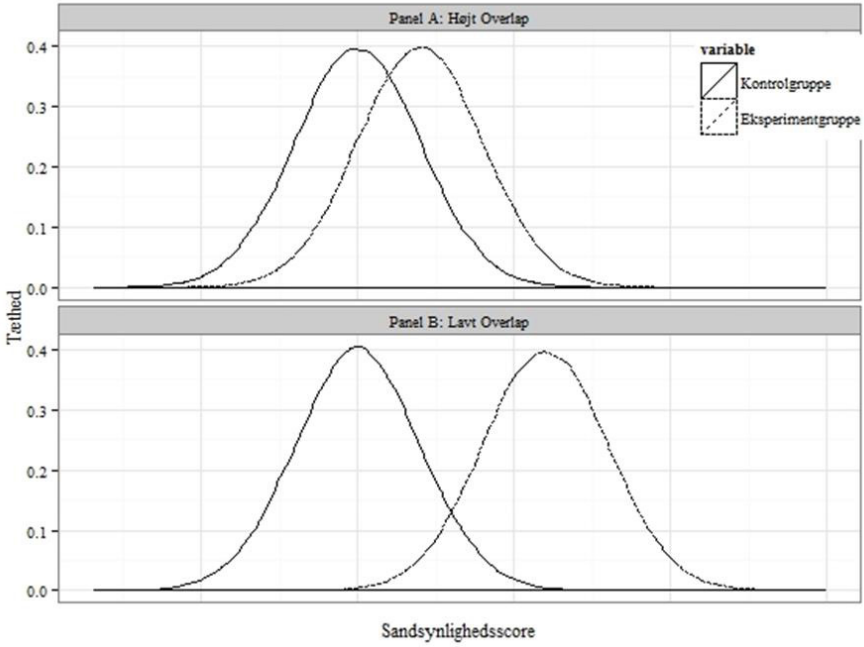
Overlap og balance

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.