

Het nut en de beperkingen van ‘onechte’ data

Als empirisch socioloog heb ik altijd wantrouwig aangekeken tegen sociologen die gebruikmaken van ‘onechte’ ofwel gesimuleerde data. Dit ondanks het feit dat slimme collega’s uit mijn eigen capaciteitsgroep zich hiermee bezighouden. Onlangs kon ik hier echter voor het eerst zelf praktijkervaring mee opdoen. Aangezien het gebruik van simulatietechnieken nauwelijks voorkomt in artikelen in *Mens & Maatschappij*, leek het me een goed idee om mijn ervaringen met u te delen.

Allereerst de vraag waar gesimuleerde data eigenlijk goed voor zijn. Het simuleren van data kan nuttig zijn als de implicaties van beleid of van een theoretisch model niet op het eerste gezicht duidelijk zijn. In het geval van beleid bestaat de simulatie dan uit het doorrekenen van de consequenties van één of meer beleidsvarianten uitgaande van aannames over hoe de bevolking zich gedraagt. Soms zijn die aannames gestoeld op empirisch onderzoek naar de gedragsintenties van mensen (‘hypothetical question’-onderzoek: ‘hoe zou u voortaan naar uw werk gaan als het openbaar vervoer gratis zou zijn?’). Bij mijn kennismaking met simulatie ging het om de implicaties van een theoretisch model. In dat geval bestaat de simulatie uit het doorrekenen van (een groot) aantal verschillende versies van het model. De resultaten daarvan worden opgeslagen (de ‘onechte’ data) en kunnen worden geanalyseerd. De resultaten van die analyse geven inzicht in de uitkomsten van het theoretisch model, met andere woorden de hypothesen die uit het model kunnen worden afgeleid.

Nu kan men zich afvragen of een theoretisch model waarvan de implicaties niet duidelijk zijn, wel gewenst is. Is zo’n model niet per definitie te ingewikkeld? We streven immers naar eenvoudige modellen om de complexe werkelijkheid af te beelden. In de praktijk blijken echter eenvoudige modellen al aanleiding te kunnen geven tot strijd over de mogelijke implicaties. Een voorbeeld is Boudons sociale-stratificatiemodel uit zijn boek *Education, opportunity and social inequality* dat in 1974 in het Engels verscheen (zie Corten, 2004). Het model bestaat slechts uit een gering aantal assumpties (over hoe succesvol leerlingen met een bepaalde sociale achtergrond zijn op school en over hoe succesvol werknemers met een bepaald opleidingsniveau zijn op de arbeidsmarkt). Over de implicaties van dit model ontstond direct na de publicatie een discussie. Is bijvoorbeeld Boudons conclusie uit het model dat afnemende ongelijkheid in het onderwijs niet leidt tot toenemende intergenerationele mobiliteit, wel waar? Boudon geeft zelf één rekenvoorbeeld waaruit dat blijkt, maar hoe zit dat als de parameters uit zijn model een andere waarde aannemen?

Simulatie heeft hier uitkomst geboden. Müller-Benedict (1999) en Corten (2004) hebben doorgerekend hoe de samenhang tussen onderwijsongelijkheid en intergenerationele mobiliteit er uitziet voor allerlei verschillende invullingen van de parameters uit Boudons model. Zij heb-

ben op deze manier laten zien dat een afnemende onderwijsongelijkheid volgens Boudons eigen model over het algemeen wel degelijk leidt tot een toename van de intergenerationele mobiliteit, dit in tegenstelling tot de overtuiging van Boudon zelf. Dat is natuurlijk een leuk resultaat.

Welke beperkingen zitten er nu aan het werken met gesimuleerde data? Een eerste beperking is dat gesimuleerde data nooit het eindstation kunnen zijn. Onderzoekers die gesimuleerde data gebruiken, doen dat om hun hypothesen op te stellen. Die hypothesen dienen in een volgende stap met 'echte' data op hun houdbaarheid getoetst te worden. Dus de bevinding dat volgens het model van Boudon er een negatieve samenhang is tussen onderwijsongelijkheid en intergenerationele mobiliteit moet nog wel degelijk empirisch worden getoetst. Zo zou Boudon – die de implicaties van zijn eigen model niet helemaal kon overzien – uiteindelijk nog gelijk kunnen krijgen als die samenhang er in de werkelijkheid niet zou zijn.

Een tweede beperking is het uitvoeren van de simulatie zelf. En dit is in mijn ogen de belangrijkste. Gesimuleerde data worden over het algemeen gemaakt door een computerprogramma te schrijven dat de mechanismen uit een model zo goed mogelijk weergeeft. Hoe vager en hoe gecompliceerder het theoretisch model is, des te moeilijker dit is. In het geval van het model van Boudon zijn de selectiemechanismen op school en op de arbeidsmarkt heel expliciet en ongecompliceerd. Desondanks valt het niet mee om deze mechanismen 'twijfelloos' om te zetten in een simulatie. Een gevolg is dat men zich bij elk resultaat van de simulatie (dus bij elke hypothese die uit het model wordt afgeleid) dient af te vragen of die nu het gevolg is van het theoretisch model of dat men het resultaat zelf gecreëerd heeft door het model niet op de juiste wijze om te zetten in het simulatieprogramma. Dit is soms voor de programmeurs van het simulatieprogramma al moeilijk te doorgronden, laat staan voor buitenstaanders (bijvoorbeeld de reviewers van een artikel).

Tot slot kan het empirisch toetsen van implicaties uit een simulatiemodel lastig zijn. Zo maakt Boudon bijvoorbeeld in zijn model gebruik van de zogenaamde 'meritocratieparameter', dat is het belang dat werknemers hechten aan het opleidingsniveau van werknemers. In het echt bestaat zo'n parameter natuurlijk niet. Men kan die parameter proberen te benaderen, bijvoorbeeld door de regressiecoëfficiënt van beroepsniveau op opleidingsniveau te schatten. Fraai is dat echter niet omdat deze regressiecoëfficiënt al een component is van de afhankelijke variabele 'intergenerationele mobiliteit'. Hierdoor worden de resultaten al gauw tautologisch. Eigenlijk dient de meritocratieparameter dus onafhankelijk gemeten te worden. Hoe verder die meting echter af ligt van het theoretische begrip, des te moeilijker het is om conclusies te trekken over de houdbaarheid van de hypothesen. Een falsificatie kan dan niet alleen liggen aan onjuistheid van het model, maar ook aan programmeerfouten of aan minder geschikte operationaliseringen.

Mijn conclusies uit deze ervaring met het werken met 'onechte' data zijn de volgende. Allereerst: het omzetten van een theoretisch model in een goede simulatie en het empirisch toetsen van de implicaties daarvan zijn moeilijke klussen. Ten tweede: gesimuleerde data zijn nuttig bij het formuleren van theoretische implicaties die niet op het eerste gezicht uit een model kunnen worden afgeleid. Ten derde: onechte data kunnen niet zonder echte data. En ten

vierde: onderzoeksartikelen waarin onechte én echte data worden gebruikt zonder al te veel in boven geschetste valkuilen te vallen, zijn welkom in *Mens & Maatschappij*.

Ineke Maas

Literatuur

- Boudon, R. (1974). *Education, opportunity, and social inequality: changing prospects in Western society*. New York: Wiley.
- Corten, R. (2004). *Education, Stratification & Simulation*. Doctoraalscriptie Sociologie, Universiteit Utrecht.
- Müller-Benedict, V. (1999). Strukturelle Grenzen Sozialer Mobilität. *Kölner Zeitung für Soziologie und Sozialpsychologie*, 51, 313-338.