Wetenschappelijk Onderzoek- en Documentatiecentrum
Ministerie van Veiligheid en Justitie

Memorandum 2017-3

Het gebruik van datagedreven analysemethoden in de (beleids)praktijk

Kansen, uitdagingen en handreikingen

S.W. van den Braak
R. Choenni

Memorandum
De reeks Memorandum omvat de rapporten van onderzoek dat door en in opdracht van het WODC is verricht.
Opname in de reeks betekent niet dat de inhoud van de rapporten het standpunt van de Minister van Veiligheid en Justitie weergeeft.

Dit memorandum is een Nederlandstalige bewerking van: Sunil Choenni, Mortaza S. Bargh, Niels Netten, Susan van den Braak, Using Data Analytics Results in Practice: Challenges and Solution Directions. In *ICT-Enabled Social Innovation for the European Social Model: a multi-disciplinary reflection and future perspectives from Internet Science, Human-Computer Interaction and Socio-Economics*, F. Davide & G. Misuraca (red.), IOS Press (nog te verschijnen).

Inhoud

1 Inleiding

De laatste jaren wordt er in de beleids- en handhavingspraktijk steeds meer *evidence-based* en informatiegestuurd gewerkt. Bij het maken en evalueren van beleid wordt bijvoorbeeld vaker gebruikgemaakt van structurele monitoring of prestatiemetingen. Dit wordt ook steeds makkelijker: door technologische ontwikkelingen komen meer, en verschillende soorten, data beschikbaar. Relevante informatie is sneller, altijd en overal beschikbaar. Daarnaast worden geavanceerde data-analysemethoden (zoals datamining en textmining) steeds toegankelijker, waardoor geprofiteerd kan worden van recente ontwikkelingen op het gebied van onder andere big data.
Bij het formuleren en het uitvoeren van beleid bieden dergelijke datagedreven analysemethoden volop kansen. Door het verzamelen, combineren en analyseren van grote hoeveelheden data kunnen nieuwe, bredere en betere inzichten verkregen worden in (opkomende) fenomenen en interventies. Met dataminingtechnieken kunnen bijvoorbeeld patronen in data gevonden worden (profielen), op basis waarvan weer voorspellingen gedaan kunnen worden. Het vinden en gestructureerd toetsen van hypotheses kan hiermee ook geautomatiseerd plaatsvinden in plaats van handmatig. Doordat grotere hoeveelheden data makkelijker en sneller beschikbaar zijn, is er ook meer directe terugkoppeling mogelijk. Een beleidsmedewerker kan daardoor sneller bijsturen als het ontwikkelde beleid niet de gewenste impact heeft. Op deze manier kunnen beter onderbouwde beleidskeuzes en -beslissingen worden genomen.
Om hiervan te kunnen profiteren, worden in de beleidspraktijk steeds meer informatiesystemen ontworpen en in gebruik genomen waarin verschillende soorten gegevens verzameld, opgeslagen en geanalyseerd worden die afkomstig zijn uit verschillende bronnen, zoals de informatiesystemen van andere (betrokken of te onderzoeken) organisaties. Daarnaast is er een groeiende behoefte om dergelijke gegevens te combineren met gegevens uit externe bronnen. Het gaat hierbij dus vaak niet om gegevens die specifiek voor analysedoeleinden zijn verzameld. Deze gegevens worden dan gebruikt voor andere doeleinden dan waarvoor ze in eerste instantie werden verzameld. Uit deze (gecombineerde) gegevens wordt vervolgens met data-analysemethoden managementinformatie gegenereerd om in de informatiebehoefte van beleidsmedewerkers (op tactisch of strategische niveau) te kunnen voorzien. Analyseresultaten die inzicht geven in knelpunten in operationele processen, kunnen daarnaast op operationeel niveau gebruikt worden om bijvoorbeeld processen te optimaliseren en beter te stroomlijnen.
De mogelijkheden van deze informatiesystemen en de datagedreven analysemethoden die erop toegepast worden, gaan echter ook gepaard met een aantal uitdagingen op het gebied van bijvoorbeeld privacy, datakwaliteit en ethiek. Om de kansen te kunnen verzilveren, dienen er dan ook antwoorden geformuleerd te worden op deze uitdagingen. Om resultaten uit analysemethoden adequaat te kunnen duiden, zal er bij gebruikers, zoals beleidsmedewerkers, meer bewustzijn moeten ontstaan en kennis moeten worden opgebouwd over de werkelijke waarde van deze resultaten. Om dit te kunnen beoordelen moeten ze inzicht hebben in de manier waarop de resultaten verkregen zijn.
Het gebruik, in de praktijk, van resultaten verkregen met behulp van datagedreven analysemethoden is niet zo vanzelfsprekend als het lijkt. We hebben te maken met enkele fundamentele problemen die het gebruik gevoelig maken voor fouten, vergissingen, vooroordelen en schade. Het onachtzaam of verkeerd gebruiken van analyseresultaten kan zelfs leiden tot schendingen van fundamentele mensen-

rechten zoals privacy, vrijheid en autonomie. Zo kan door het combineren van data uit verschillende bronnen, bijvoorbeeld (per ongeluk) de identiteit van personen of gevoelige informatie over deze personen onthuld worden. Daarnaast kunnen personen ten onrechte in een gegevensset worden opgenomen, kunnen beslissingen worden genomen die nadelig uitpakken voor deze personen, of kunnen over deze personen verkeerde conclusies worden getrokken. Zelfs als de analyses, resultaten en conclusies juist zijn, kan dit in bepaalde situaties schadelijk en/of illegaal zijn, bijvoorbeeld wanneer er sprake is van ongerechtvaardigde of oneerlijke discriminatie van personen.
Deze problemen vloeien deels voort uit een aantal fundamentele problemen die bij gebruikers van analyseresultaten niet altijd bekend zijn of onderkend worden. Ten eerste, zijn informatiesystemen gebaseerd op de aanname van een gesloten wereld (de *closed world assumption*). Dat wil zeggen dat wordt aangenomen dat de gegevens die zijn opgeslagen in een informatiesysteem waar (correct) en compleet zijn. In de praktijk houdt deze veronderstelling echter geen stand. Ten tweede, is elk informatiesysteem een model (een representatie) van een bepaald fenomeen in de werkelijkheid (de echte wereld). Dit model kan nooit een volledige beschrijving van dit fenomeen zijn en omvat sowieso geen volledige beschrijving van de relatie van dit fenomeen met andere fenomenen. Ten derde, zijn de analysemethoden die op data uit informatiesystemen worden toegepast, gebaseerd op inductie. Hierdoor zijn de resultaten met enige onzekerheid omgeven. Als gevolg van deze drie problemen zijn de resultaten van data-analysemethoden mogelijk onzeker, onvolledig en bevooroordeeld, en daarom niet altijd bruikbaar in de echte wereld.
In dit memorandum worden deze fundamentele problemen toegelicht. Daarbij wordt stilgestaan bij oorzaken van de problemen die relevant zijn bij het ontwerpen van informatiesystemen, het toepassen van data-analysemethoden, en het gebruik hiervan in de (beleids)praktijk. Hierin zijn twee belangrijke processen van belang:

1 het verzamelen van geschikte data, en
2 het selecteren van een geschikte data-analysestrategie of -algoritme.

In hoofdstuk 2 worden de uitdagingen in beide stappen toegelicht met concrete voorbeelden. Vervolgens worden richtlijnen gegeven om met deze uitdagingen om te gaan en data-analyseprojecten op een adequate manier uit te voeren. Het gaat hier specifiek om geavanceerde en grotendeels geautomatiseerde analysemethoden (zoals datamining) en niet om klassieke kwantitatieve methoden. Dergelijke nieuwe methoden maken vaak gebruik van bestaande datasets in plaats van dat de data ten behoeve van een specifiek onderzoek verzameld worden. De aanpak is daardoor anders: deze is datagestuurd en richt zich niet op het bewijzen van een vooraf opgestelde hypothese.

De auteurs bedanken Nikolaj Tollenaar, Ronald Meijer en Frans Leeuw voor het nalezen van het manuscript en hun nuttige suggesties.

2 Uitdagingen bij het verzamelen van data

Om data-analysemethoden te kunnen toepassen, moeten (geschikte) data verzameld worden. Dit gebeurt gewoonlijk in systemen zoals databases, datawarehouses en dataspaces. Hierbij spelen twee aspecten een belangrijke rol. Ten eerste, gaat het vaak om data afkomstig uit verschillende bronnen. Om deze data goed te kunnen analyseren, moeten deze eerst geïntegreerd worden. Data-integratie heeft als doel om gegevens die betrekking hebben op dezelfde entiteiten (bijvoorbeeld personen of zaken) te identificeren en zo een uniforme weergave te krijgen. Om dit te bewerkstelligen moeten overbodige (redundante) gegevens verwijderd worden en inconsistenties opgelost worden (dit doet zich voor als bepaalde gegevens in meerdere bronnen worden bijgehouden, maar de waardes afwijken), terwijl ook gekeken wordt naar de kwaliteit van de data. Ten tweede, zijn dergelijke systemen gebaseerd op de hierboven genoemde aanname van een gesloten wereld. Dit betekent dat wordt aangenomen dat de dataset compleet en foutloos is. Compleet betekent hier dat over alle relevante entiteiten informatie aanwezig is, dat voor alle entiteiten alle velden gevuld zijn én dat alle relevante velden aanwezig zijn. In veel domeinen gaat deze aanname echter niet op door registratiefouten en/of omissies.
Vaak bestaat een perfecte dataset daarom alleen in theorie. Het is daarmee ook de vraag in hoeverre een dataset de werkelijkheid (goed) representeert en iets zegt over de gehele te onderzoeken populatie (van fenomenen, objecten of personen). Dit heeft weer gevolgen voor de bruikbaarheid en validiteit van de analyseresultaten. Kort gezegd, speelt bij het verzamelen van data ten behoeve van datagedreven analyse de kwaliteit van zowel de data als het (database)model een cruciale rol. Hieronder worden beide (potentiële) problemen, en de gevolgen daarvan, nader toegelicht.

2.1 Datakwaliteit en ontbrekende data

Zoals hierboven al beschreven is de kwaliteit van databases niet altijd gegarandeerd en bevatten ze vaak ruis. Dit kan onder ander veroorzaakt worden door impliciete fouten door meetinstrumenten zoals sensoren die foute metingen doen, door willekeurige fouten in geautomatiseerde processen tijdens het verzamelen en laden van data, of door menselijke fouten bij de invoer van gegevens. Om dergelijke databases te kunnen gebruiken voor data-analysedoeleinden moeten deze fouten eerst zo veel mogelijk opgelost worden. Dit is een lastig en foutgevoelig proces, waarvoor vaak enige domeinkennis nodig is. Dit is vooral het geval als het gaat om oude of verouderde databases (ook wel *legacy*systemen genoemd). In dergelijke databases is het lastiger om fouten op te sporen en op te lossen, doordat de gegevens erin vaak slechter gedocumenteerd zijn. Als gevolg hiervan moet de betekenis van de gegevens afgeleid worden uit de inhoud van de database en de domeinkennis die daarover nog beschikbaar is in de beherende organisatie. Als de gegevens erg oud zijn, is het moeilijker om de exacte betekenis nog te achterhalen.

Een ander kwaliteitsprobleem in databases kan ontstaan doordat de gegevens erin niet volledig zijn. In dat geval ontbreken er waarden: er is voor een bepaald veld helemaal geen waarde ingevuld of het veld bevat de waarde *NULL*. In databases heeft deze waarde een speciale betekenis met verschillende interpretaties: de

Box 1 Voorbeeld verouderde database

Stel dat een verouderde database een veld “salaris” bevat, waarover de documentatie zegt dat dit het jaarsalaris van een persoon betreft. Het gemiddelde salaris kan dan berekend worden door alle waarden in het veld salaris bij elkaar op te tellen en te delen door het aantal waarden (personen).
Stel nu dat dit veld niet alleen voltijd-, maar ook deeltijdsalarissen bevat. Het berekende gemiddelde salaris kan dan erg laag uitvallen. Als dit niet expliciet in de documentatie beschreven staat en de arbeidsduur niet geregistreerd is in de database, is domeinkennis vereist om de ware betekenis van het salarisveld te achterhalen. Alleen dan kan het berekende gemiddelde goed geduid worden.

waarde is onbekend, niet van toepassing of niet gedefinieerd. Als men hiervan niet goed op de hoogte is, kunnen *NULL*-waarden tot gevolg hebben dat query’s (informatieverzoeken aan databases; zoekopdrachten die aan databases worden gegeven om gegevens terug te geven) verkeerde resultaten opleveren. Het is daarom aan te bevelen om *NULL*-waarden zo veel mogelijk te vervangen. Ook dit vereist vaak domeinkennis.

Box 2 Voorbeeld NULL-waarden

Stel dat een database informatie bevat over de leerlingen van een school waaronder een veld “examenuitslag”. Het aantal geslaagde en gezakte leerlingen kan dan berekend worden door het aantal velden met de waarde “geslaagd” en “gezakt” te tellen.
Het totale aantal leerlingen is over het algemeen gelijk aan de som van het aantal geslaagden en gezakten.
Stel nu dat dit veld meerdere NULL-waarden bevat. Het totaalaantal leerlingen is dan niet meer af te leiden uit het aantal geslaagden en gezakten. Dit totaalaantal moet berekend worden door het aantal rijen in de database te tellen. Deze query levert een ander resultaat op dan de query om het aantal geslaagden en gezakten te tellen.

Beide problemen kunnen tot gevolg hebben dat de te gebruiken dataset onvolmaakt of onvolledig is. Naast problemen met de kwaliteit van de data, is er nog een andere reden waarom datasets imperfect kunnen zijn. Dit speelt als bepaalde gegevens “verwaarloosd” worden, doordat ze niet verzameld worden (ze zijn als het ware vergeten) of niet verzameld kunnen worden (men heeft geen toegang tot de gegevens). Als gevolg hiervan kan cruciale, relevante informatie (bijvoorbeeld over een bepaalde periode of regio) in de dataset ontbreken en representeert de dataset de werkelijkheid niet (meer) goed.
Bij het selecteren en verzamelen van data kan er dus sprake zijn van een *selection bias* (bijvoorbeeld een *sampling of exclusion bias*). De steekproef is dan niet willekeurig en niet representatief voor de te onderzoeken populatie, wat problemen oplevert bij het toepassen van analysemethoden (en de validiteit van de resultaten). Belangrijker nog, de gebruiker van de dataset is er niet altijd van op de hoogte dat gegevens ontbreken (bijvoorbeeld als hij ze onbewust is vergeten). Dergelijke nalatigheid bij het verzamelen van voor data-analyse vereiste gegevens kan weer de cognitieve biases zoals bijvoorbeeld *confirmation bias* (voorkeur tot bevestiging) versterken. Dit verwijst naar de neiging van mensen om meer aandacht te hechten aan informatie die de eigen hypotheses bevestigt en tegelijkertijd minder aandacht te hebben voor informatie die deze tegenspreekt. Op basis hiervan kunnen (bewust of onbewust) verkeerde conclusies getrokken worden.

Box 3 Voorbeeld selection en confirmation bias

Een voorbeeld van een situatie waarin door *selection bias* een *confirmation bias* kan ontstaan, kan gevonden worden in een recente natuurramp waarbij via *crowdsourcing* gegevens werden verzameld. In 2012 trok orkaan Sandy over Seaside Heights en Midland Beach in de VS. Meer dan 20 miljoen berichten werden in deze periode uit deze regio verzameld en geanalyseerd. Hieruit bleek dat de meeste berichten afkomstig waren van Manhattan en niet uit de zwaarder getroffen gebieden. De reden hierachter lag in de hoge concentratie van smartphone- en Twittergebruik in Manhattan, terwijl in de getroffen gebieden sprake was van stroomuitval (waardoor telefoons leeg raakten en mobiele netwerken onbereikbaar werden). Als gevolg hiervan waren cruciale gebieden niet goed vertegenwoordigd in de gegevens die via Twitter konden worden verzameld.
Zonder rekening te houden met de context van de verzamelde gegevens en biases daarin (namelijk dat er uit de rampgebieden weinig informatie beschikbaar was), zouden verkeerde conclusies getrokken kunnen worden over welke gebieden het zwaarst getroffen waren. Stel nu dat als een extreem voorbeeld de overheid (alleen) deze informatie had gebruikt om te beslissen over te nemen reddingsmissies. En dat zij veronderstelden dat Manhattan het zwaarst getroffen was door de orkaan. Dan zouden de schaarse middelen op de verkeerde plaats zijn ingezet. De verzamelde gegevens zouden dan gebruikt zijn om de bestaande hypothese te bevestigen, ongeacht dat ze hiaten bevatten met betrekking tot de werkelijke situatie.

2.2 Modelleringsvraagstukken

Zoals hierboven al beschreven, omvatten databases slechts een gedeeltelijk model (of representatie) van de werkelijkheid. Bij het modelleren van de echte wereld in een database zijn twee principes van belang, namelijk *universe of discourse* en abstractie. Ten eerste, een database vertegenwoordigt maar een deel van de fenomenen uit de echte wereld. De in de database gemodelleerde realiteit noemen we de *universe of discourse*: dit is de set van fenomenen uit de echte wereld (ook wel: objecten) waarop een (database)model gebaseerd is. Het bepaalt daarmee over welke objecten een database gaat en welke gegevens erin verzameld worden. Ten tweede, een database bevat altijd een vereenvoudiging van de fenomenen uit de echte wereld. Dit omdat het ontwerpen van een complete database veel te complex zou zijn. Data-abstractie beoogt een vereenvoudigde, conceptuele representatie van de echte wereld te geven, door bepaalde, niet essentiële kenmerken of eigenschappen weg te laten of te verwijderen. Het bepaalt hoe gedetailleerd of specifiek

Box 4 Voorbeeld universe of discourse en abstractie

Stel dat er een database ontworpen wordt voor de benodigde onderdelen van een fabriek.
Eerst moet besloten worden over welke objecten gegevens opgeslagen dienen te worden. In dit geval kunnen dat bijvoorbeeld de leveranciers, onderdelen, en de relatie ertussen (bijvoorbeeld dat leveranciers meerdere verschillende onderdelen kunnen leveren) zijn. Tezamen vormen zij het *universe of discourse* van de onderdelendatabase.
Vervolgens moet bepaald worden welke specifieke eigenschappen over de objecten geslagen dienen te worden. Met betrekking tot de onderdelen kunnen dat bijvoorbeeld de naam, het typenummer en de prijs van het onderdeel zijn. Dit vormt dan het abstractieniveau van de onderdelendatabase.

de objecten in een database beschreven worden. Door *universe of discourse* en abstractie toe te passen kunnen economische en efficiënte databasemodellen van fenomenen uit de echte wereld gemaakt worden. Als gevolg van deze principes bevat een informatiesysteem altijd een beperkte en vereenvoudigde representatie van de realiteit.
Gezien deze modelleringsprincipes, is het hierboven beschreven probleem met betrekking tot verouderde databases (*legacy*systemen) extra pregnant. Veranderingen in de omgeving kunnen er namelijk toe leiden dat het databasemodel van de (verouderde) database niet meer (volledig) overeenkomt met de (veranderde) realiteit. Als gevolg hiervan kan het zo zijn dat, zelfs als de gegevens op het moment van opslaan juist waren (en de kwaliteit in orde), de datakwaliteit naar verloop van tijd verslechtert, doordat sommige waarden niet meer geldig zijn of geen betekenis meer hebben in de nieuwe realiteit. In veel domeinen verandert de betekenis van de opgeslagen gegevens nog wel eens, bijvoorbeeld als gevolg van wet- en regelgeving. Het is niet altijd duidelijk hoe deze wijzigingen in een databasemodel moeten en kunnen worden verwerkt.
Zowel het ongewijzigd laten als het achteraf wijzigen van waarden heeft voor- en nadelen. Als er niets gewijzigd wordt, kan dit problemen met de datakwaliteit opleveren en kunnen query's verkeerde resultaten teruggeven (er staan dan waarden in de database die in de nieuwe realiteit geen of een onduidelijke betekenis hebben). Het (oude) databasemodel representeert de nieuwe realiteit dan niet meer goed. Wijzigen kan daarentegen leiden tot (ongefundeerde) trendomkeringen. Het (nieuwe) databasemodel past dan niet meer bij de oude realiteit. Om dit te voorkomen, moeten procedures worden gedefinieerd om met betekeniswijzigingen om

Box 5 Voorbeeld verouderd datamodel

Stel dat een database informatie bevat over de geboortelanden van personen en dat in deze database personen staan die volgens de registratie in de "Sovjetunie" zijn geboren. Sinds 1991 bestaat de Sovjetunie echter niet meer. Deze database bevat daardoor (deels) verouderde informatie waarvan de betekenis veranderd is.
Als het geboorteland niet achteraf wordt aangepast, kan dit bijvoorbeeld problemen opleveren bij het tellen van het aantal personen dat in Rusland is geboren. Het uit de database verkregen aantal zal dan lager zijn dan het werkelijk aantal, aangezien personen die volgens de database in de Sovjetunie zijn geboren, (door de query) niet als geboren in Rusland worden beschouwd. Dit kan opgelost worden, door voor alle personen met als geregistreerd geboorteland Sovjetunie, het geboorteland achteraf aan te passen naar de (deel)republiek waarin zij geboren zijn (bijvoorbeeld Rusland of Oekraïne).
Als het geboorteland wel wordt aangepast, kan dit echter vreemde trendbreuken tot gevolg hebben. Op tijd t-1 (voor het uiteenvallen) was het aantal personen geboren in Rusland in de database nog heel laag (het gaat dan alleen personen die voor de eenwording van de Sovjetunie in Rusland werden geboren). Op tijd t+1, en na het aanpassen van de waarden, is het aantal in Rusland geborenen ineens geëxplodeerd. Voor het geboorteland Sovjetunie is het tegenovergestelde het geval. Dit levert dus problemen op bij het volgen van trends. Op basis van eenmaal veranderde waarden, kan de oude situatie ook niet meer gereproduceerd worden.
Een alternatief is om veranderende waarden niet aan te passen, maar wel bij te houden. Op basis van de vastgelegde domeinkennis kan de query dan zodanig opgesteld worden dat de waarde "Sovjetunie" wel meegeteld wordt bij het bepalen van het aantal in Rusland geborenen. Op deze manier blijft de database bruikbaar en valide voor zowel de oude als nieuwe realiteit.

te gaan, bijvoorbeeld door ze systematisch te gaan bijhouden. In de praktijk is het echter niet altijd makkelijk om dit te doen: het bijhouden van historie, chronologie en documentatie, en het zorgen voor reproduceerbaarheid, is vaak tijdrovend en duur.

2.3 Samenvatting

Het dataverzamelingsproces heeft als doel om gegevens uit verschillende bronnen en informatiesystemen samen te brengen in één gecombineerde database of data-set, zodat ze vervolgens geanalyseerd kunnen worden. Bovengenoemde problemen en illustrerende voorbeelden tonen aan dat het opschonen en omzetten van gegevens, met het oog op de datakwaliteit, erg belangrijk is in dit proces. Fouten en omissies (bewust of onbewust) zijn echter niet altijd makkelijk te detecteren. Hiervoor is vaak domeinkennis nodig, die niet altijd beschikbaar is (bijvoorbeeld in het geval van *legacy*systemen). Men kan daardoor nooit garanderen dat de verzamelde gegevens correct, eenduidig en volledig zijn. Bij het toepassen van data-analysemethoden moet daarom in het oog gehouden worden dat de verzamelde dataset vaak imperfect is en geen volledig en juist beeld geeft van de echte wereld. De uitdagingen die spelen bij het analyseren van data worden in hoofdstuk 3 gedetailleerder beschreven.

3 Uitdagingen bij het analyseren van data

Datagedreven analysemethoden worden in beginsel gebruikt om nieuwe kennis op te doen of nieuwe inzichten te verkrijgen. Een belangrijk verschil tussen (traditionele) statistiek en nieuwe analysemethoden zoals datamining, is het uitgangspunt van de analyse. In de statistiek worden in beginsel eerst een aantal hypotheses geformuleerd, waarna data worden verzameld om deze te verifiëren. De dataset wordt dan gebruikt om vooraf gedefinieerde (nul)hypotheses te accepteren of verwerpen. Er is hiervoor een theorie nodig over de totstandkoming van de data (de onderliggende kansverdeling en relaties tussen variabelen). Hiervoor zijn verschillende statistische toetsten beschikbaar. Datamining werkt in beginsel precies andersom: vooraf is geen hypothese bekend. Op basis van data wordt juist naar interessante hypotheses gezocht. Het heeft dus als doel om te zoeken naar statistische verbanden, correlaties of patronen in een gegeven dataset. Dit is datagestuurd, en kan verschillende doelen hebben waaronder exploratie, verklaring of voorspelling. Hiervoor zijn verschillende methoden beschikbaar, zoals classificatie, clusteranalyse en associatie-analyse. Voor ieder type bestaan er weer verschillende algoritmen en technieken. Voor classificatie kunnen bijvoorbeeld beslisbomen of neurale netwerken gebruikt worden. Op basis van dergelijke geavanceerde data-analysetechnieken kunnen beschrijvende, verklarende, of voorspellende modellen van de werkelijkheid gemaakt worden (dit wordt ook wel een profiel genoemd).

Door de gevolgde werkwijze is het mogelijk dat het gebruik van datamining onjuiste conclusies oplevert. Als er maar genoeg data verzameld en geanalyseerd worden, zal ongetwijfeld een statistische correlatie tussen variabelen gevonden worden. Dit hoeft echter nog niet te betekenen dat er ook een oorzakelijk (causaal) verband tussen de twee betreffende variabelen bestaat. Er kan namelijk nog een andere, verborgen, variabele (een *confounding* variabele, buiten het model) zijn die gerelateerd is aan zowel de verklarende als de afhankelijke variabele en dus op de achtergrond van invloed is op de oorzaak-gevolg relatie. Dit wordt ook wel een schijncorrelatie genoemd.

Een bijkomend probleem is dat de resultaten verkregen uit data-analysemethoden gebaseerd zijn op inductie. Bij inductie wordt een algemene regel (een generalisatie) afgeleid uit een reeks (specifieke) waarnemingen (die bijvoorbeeld zijn vastgelegd in een database). Zolang de aanname van een gesloten wereld geldt, is deze gevolgtrekking geldig en het resultaat correct. Er is dan geen reden om te twijfelen. Deze aanname dat de gebruikte datasets compleet en foutloos zijn, gaat in de praktijk echter niet op. Het feit dat de gebruikte datasets in beginsel onvolmaakt zijn, maakt dat de resultaten van data-analysemethoden die op deze data zijn toegepast incompleet en onzeker zijn. Daarom is voorzichtigheid geboden bij het gebruik ervan.

Om deze redenen is het interpreteren van analyseresultaten, en het gebruik ervan in de (beleids)praktijk, niet eenduidig en eenvoudig. Hiertoe moet de gebruiker van de resultaten eerst beoordelen in hoeverre de achterliggende data correct zijn, voldoende belangrijke kenmerken bevatten en de echte wereld goed representeren. Daarnaast moet de gebruiker beoordelen of de gebruikte methode valide is en op de juiste manier is toegepast. Vervolgens kan worden bepaald of de resultaten geldig, valide en bruikbaar zijn. Deze aspecten worden hieronder toegelicht.

3.1 Validiteit

Doordat datamining en andere geavanceerde data-analysemethoden gebaseerd zijn op inductie, gaan ze gepaard met enige mate van onzekerheid. Aangezien datasets in de praktijk vaak niet volmaakt en compleet zijn, kan dit leiden tot onjuiste hypotheses en, zolang men hier zich niet bewust van is, ongefundeerde conclusies. Daarom is het van belang om vooraf op de hoogte te zijn van de beperkingen van de gebruikte data en goed te kijken of deze wel geschikt zijn om data-analyse op toe te passen. Daarnaast moeten de gevonden hypotheses kritisch bekeken worden, voordat op basis daarvan conclusies worden getrokken, hiervoor is (naast kennis over de beperkingen van de data) vaak kennis over het toepassingsdomein nodig.

Box 6 Voorbeeld inductie

Stel dat een database informatie bevat over waargenomen zwanen, waaronder de kleur van iedere zwaan. In deze database worden systematisch alle waarnemingen van zwanen in een bepaald gebied geregistreerd.
Als alle waargenomen en geregistreerde zwanen wit zijn, dan is hypothese die op basis van inductief redeneren verkregen kan worden: “alle zwanen zijn wit”. Deze hypothese geldt misschien voor het betreffende gebied, maar zeker niet voor de hele wereld. In de praktijk bestaan er namelijk wel degelijk zwarte zwanen, dit weten we op basis van domeinkennis, deze zijn alleen nog niet waargenomen en/of geregistreerd. De gebruikte database is daarom incompleet en geen volledige representatie van de realiteit. Als gevolg hiervan is de gevonden hypothese in zijn algemeenheid onjuist. Deze kan eenvoudig weerlegd worden door het vinden van een zwarte zwaan.

Zelfs als de aanname van een gesloten wereld wel geldt, en de verzamelde gegevens perfect en compleet zijn, kan het zijn dat data-analyse ongeldige en onbetrouwbare resultaten oplevert. De keuze voor een geschikte methode en de configuratie ervan zijn daarbij van groot belang. Dit is problematisch als de gebruiker van (commerciële) data-analysetools de onderliggende algoritmische details niet kent of als een afnemer van analyseresultaten blindelings vertrouwt op de door derden aangeboden analyses (*analysis-as-a-service*). In deze gevallen is de kans aanwezig dat fundamentele principes die cruciaal zijn voor een succesvolle en valide toepassing van de betreffende algoritmes of technieken worden geschonden.
Als analysetechnieken verkeerd worden toegepast, ligt het gevaar van bijvoorbeeld *overfitting* op de loer. Het gevonden model volgt de data dan eigenlijk te nauw en is te complex. Het model is zo gedetailleerd dat het zelfs de ruis in de dataset heeft geleerd. Als er maar een beperkt aantal waarnemingen zijn, is de kans op *overfitting* groter. Ook als er te veel variabelen worden meegenomen in het model, is de kans op *overfitting* groot. *Underfitting* is het tegenovergestelde van *overfitting*. Het model is dan nog te simpel en kan de onderliggende verbanden in de data niet goed vangen. Het model dat het beste past bij de trainingsdata (de data die gebruikt worden om het model te leren), is niet altijd het beste voor nieuwe data, en daarom ook niet altijd geschikt om voorspellingen te kunnen doen. Vaak wordt een aparte set van testdata gebruikt om het getrainde model achteraf te kunnen beoordelen en valideren door te kijken hoe accuraat het model waarschijnlijk is op nieuwe gegevens, ofwel hoe generaliseerbaar het is; deze data zijn niet gebruikt voor het leren van het model. Dit is echter lang niet in alle domeinen mogelijk (bijvoorbeeld omdat er weinig data zijn).

Om deze redenen is het van belang om vooraf goed op de hoogte te zijn van de werking, voorwaarden en beperkingen van de gebruikte analysemethode. Het is dus zaak om steeds na te gaan of de gekozen analysemethode überhaupt toegepast kan en mag worden (gegeven de beschikbare data en het doel) en welke parameters en instellingen moeten worden gebruikt. Onachtzaam gebruik is zinloos en kan leiden tot verkeerde conclusies.
Resultaten van data-analyses kunnen bewust of onbewust verkeerd worden geïnterpreteerd. Dit kan tot gevolg hebben dat de resultaten van een analyse (onterecht) als waarheid worden gezien. Op basis hiervan kunnen besluiten worden genomen die achteraf onjuist blijken te zijn. Dit wordt versterkt door de hierboven al besproken *confirmation bias*. Deze bias brengt het gevaar met zich mee dat net zolang in de data wordt gezocht tot een vooraf bedachte hypothese bewezen kan worden. De analyse wordt dan beïnvloed door vooringenomenheid en vooroordelen.

Box 7 Voorbeeld schijncorrelatie

Een voorbeeld van een gevonden correlatie op basis waarvan verkeerde conclusies zijn getrokken verscheen een aantal jaren geleden in de media. Wetenschappers publiceerden toen een onderzoek waarin ze een opmerkelijk verband vonden tussen de consumptie van chocolade en het winnen van een Nobelprijs. In landen waar meer chocolade wordt gegeten, zijn er meer Nobelprijswinnaars.
Dit onderzoek leidde in verschillende media tot de conclusie dat je slimmer wordt van het eten van chocolade. De aanbeveling was dan ook om vooral veel (donkere) chocolade te eten. In dit geval bestaat er echter hooguit een correlatie tussen beide variabelen en is er geen sprake van een causaal verband. Deze conclusie en aanbeveling was dus te voorbarig.
Op de achtergrond speelt hier namelijk een andere variabele een rol, ontdekten andere onderzoekers. Chocolade wordt vooral gegeten in rijkere landen en deze landen kennen ook meer Nobelprijswinnaars dan armere landen. Geld speelt dus een belangrijke rol. Landen die meer geld investeren in wetenschap, hebben een grotere kans dat er veel Nobelprijzen worden gewonnen. Veel chocolade eten heeft dus geen zin, investeren in wetenschap wel.

3.2 Bruikbaarheid

De uitkomst van een data-analyse noemen we een “systeemwerkelijkheid”: een nieuwe realiteit die tot stand is gekomen op basis van data uit informatiesystemen, en daarmee slechts een representatie of interpretatie is van verbanden in de echte wereld. Aangezien een systeemwerkelijkheid dikwijls gebaseerd is op onzekere en onvolledige data, komt deze niet altijd overeen met de echte werkelijkheid. Als een systeemwerkelijkheid is gebaseerd is op verouderde data (bijvoorbeeld uit *legacy*-systemen), dan is deze vaak niet (meer) van toepassing op de tegenwoordige realiteit. Dit doordat het gevonden model is gebaseerd op waarnemingen uit het verleden. Bij het interpreteren van analyseresultaten moet daarom goed in de gaten gehouden worden op welke periode de gevonden systeemwerkelijkheid betrekking heeft. Steeds moet beoordeeld worden of de verzamelde gegevens nog steeds representatief zijn voor het heden. In domeinen waarin de omstandigheden vaak en veel veranderen, is dit van zeer groot belang. Als de systeemwerkelijkheid te veel afwijkt van de realiteit, dan is de bruikbaarheid ervan in het geding. Het is dan lastig om op basis ervan uitspraken te doen over de echte wereld.

Box 8 Voorbeeld verouderde systeemwerkelijkheid

Stel dat een database informatie bevat over alle klachten van klanten, en de afhandeling ervan, in de afgelopen 50 jaar. Deze database kan dan geanalyseerd worden om te kijken welk type klanten het meest succesvol is bij het indienen van klachten.
Stel dat op basis van een dataminingalgoritme gevonden wordt dat goed opgeleide mannen uit stedelijke gebieden de grootste kans hebben om succesvol een klacht in te dienen. Dit is dan de gevonden systeemwerkelijkheid.
Hoewel dit misschien gold voor de situatie in het verleden, is het zeer goed mogelijk dat deze regel vandaag de dag niet meer opgaat. Een van de redenen kan zijn dat vrouwen 30 tot 50 jaar geleden minder vaak hoger onderwijs genoten en (daardoor) ook minder vaak (zelf) klachten indienden. Vroeger werden vooral door hoger opgeleide mannen klachten ingediend en dat vertekent de resultaten. Tegenwoordig zijn vrouwen ook goed opgeleid en dienen ze ook vaker klachten in. Het gevonden dataminingresultaat is daarom niet noodzakelijkerwijs geldig en bruikbaar voor de wereld van vandaag. Voor het verleden is de bevinding juist wel (nog steeds) bruikbaar.

Er is nog een andere reden waarom het gebruiken en interpreteren van analyse-resultaten lastig is. Het gaat namelijk om statistische waarheden die gelden voor groepen (bijvoorbeeld personen) en niet altijd toepasbaar zijn op individuen. Een statistische waarheid impliceert namelijk een verdelingsfunctie van mogelijke uit-komsten voor een zeer grote (haast oneindige) groep waarnemingen. Daar komt bij dat de gebruikte dataset maar een bepekt aantal variabelen bevat die de individuele variatie maar deels kunnen beschrijven. Op basis hiervan kunnen daarom geen conclusies worden getrokken of voorspellingen worden gedaan die betrekking hebben op individuele gevallen. In de praktijk wordt dit toch vaak gedaan om (beleids)beslissingen te kunnen nemen. Dergelijke onjuiste interpretaties kunnen leiden tot onterechte beslissingen over individuen.
Om op basis van statische waarheden toch uitspraken over individuen te kunnen doen, zijn verschillende benaderingen denkbaar. Het gaat dan om het duiden van waarschijnlijkheidswaarden. Deze waarden zijn voor niet-statistici lastig te inter-preteren en gebruiken. De praktische betekenis is niet direct helder, en is ook afhankelijk van de betrouwbaarheid van het algoritme. Voor leken is de werking van dergelijke algoritmen echter moeilijk te doorgronden en daardoor is het bijna onmogelijk om de geldigheid van de waarde te verifiëren en te duiden. Al met al betekent dit dat resultaten verkregen met datagedreven analysemethoden zoals alle modellen die gebruikmaken van beperkte informatie, niet zomaar gebruikt mogen worden om voorspellingen over individuele gevallen te doen, en daardoor zijn ze niet altijd even goed bruikbaar in de (beleids)praktijk.

Box 9 Voorbeeld waarschijnlijkheidswaarden

Stel dat op basis van data-analyse het volgende profiel is gevonden: “80% van de mannen tussen 18 en 24 jaar die in huurauto’s rijden, zijn betrokken bij auto-ongeluk-ken”. Dit betekent dat van een oneindige groep mannen tussen 18 en 24 jaar die in huurauto’s rijden, 80% wel een auto-ongeluk krijgt en 20% niet.
Stel dat nu naar één enkel geval gekeken wordt: een man van 19 jaar die een auto huurt. Het is dan onjuist om te veronderstellen dat deze man 80% kans heeft om bij een auto-ongeluk betrokken te raken. Op basis van deze statische waarheid kan dit helemaal niet voorspeld worden. Het is dan ook niet verdedigbaar om enkel en alleen op basis van

dit analyseresultaat alle jongemannen die een auto willen huren eerst aan een rijvaardigheidscursus te laten deelnemen.
Voor individuele gevallen kan de waarschijnlijkheidswaarde geduid worden op basis van een objectief of subjectief kansbegrip.
Ten eerste, een objectieve interpretatie van de waarschijnlijkheidswaarde is dat als deze 19-jarige man een oneindig aantal ritten uitvoert, hij in 80% van de gevallen een aanrijding veroorzaakt en in 20% van de gevallen niet. Dit is echter alleen geldig als hij voorgaande aanrijdingen niet onthoudt en niet van zijn fouten leert. Deze aanname is in de realiteit niet houdbaar. Als gevolg hiervan kan nog steeds niet voorspeld worden of deze man in zijn volgende rit al dan niet een ongeluk zal veroorzaken.
Ten tweede, een subjectieve interpretatie van de waarschijnlijkheidswaarde is dat deze tot stand is gekomen op basis van observaties aangevuld met domeinkennis. Het gebruikte algoritme heeft de waarschijnlijkheid dan gekwantificeerd door veel verschillende bestuurders te onderzoeken en te kijken naar bestuurders die veel op de 19-jarige lijken. Op basis van nieuwe of aanvullende informatie kan deze schatting nog aangepast worden. Dergelijke subjectieve waarschijnlijkheidswaarden zijn echter vaak bevooroordeeld.
Voor de betrokken man levert de toegewezen waarschijnlijkheidswaarde een aantal praktische vragen op zoals: is de kans op een auto-ongeluk voor elke rit 80% of wordt de kans op een ongeluk steeds groter als ik veel veilige ritten heb gehad? Beide interpretaties geven hier geen antwoord op.

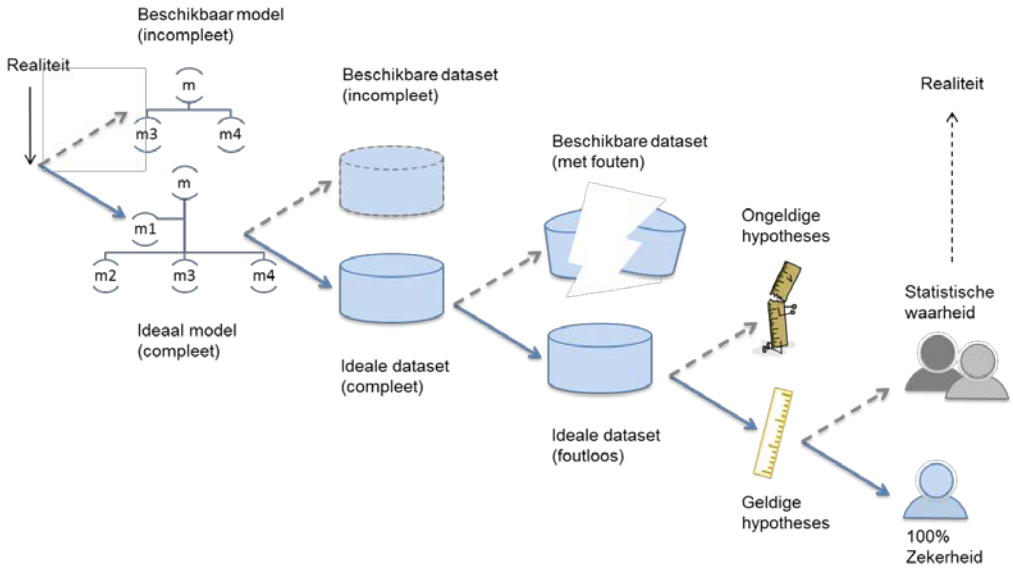
3.3 Samenvatting

Het toepassen van geavanceerde data-analysemethoden heeft als doel om op basis van verzamelde datasets tot nieuwe hypotheses te komen door te zoeken naar verbanden en patronen. Dit klinkt veelbelovend, maar bovengenoemde problemen en bijbehorende voorbeelden laten zien dat het interpreteren van de resultaten en het toepassen ervan in de (beleids)praktijk niet vanzelfsprekend is. Doordat de gebruikte data vaak gebreken hebben, zijn de analyseresultaten onzeker en mogelijk onjuist. Ook het verkeerd toepassen van analysetechnieken heeft verstrekkende gevolgen. Het is daardoor lastig om op basis van de resultaten juiste conclusies te trekken en (beleids)beslissingen te nemen. Vooral voor het doen van uitspraken over individuele gevallen zijn de resultaten lang niet altijd geschikt. In hoofdstuk 4 worden de belangrijkste schakels in het proces van data(verzameling) tot valide conclusies nog eens kort samengevat.

4 Van data naar valide en bruikbare hypotheses

Datagedreven analyse is een kwetsbaar proces waarin op verschillende punten dingen fout kunnen gaan. Als gevolg hiervan staan de verkregen resultaten soms ver af van de werkelijkheid en zijn ze lang niet altijd bruikbaar in de (beleids-) praktijk. Hierboven zijn vijf belangrijke uitdagingen beschreven, die niet helemaal los van elkaar staan, en ook invloed hebben op uitdagingen later in het proces. In figuur 1 zijn deze schematisch weergegeven. Hoe met deze uitdagingen wordt omgegaan bepaalt uiteindelijk de kwaliteit van het resultaat.

Figuur 1 Uitdagingen bij het verzamelen en analyseren van data



De eerste uitdaging heeft te maken met het modelleren van de echte wereld in een database. Door de principes van *universe of discourse* en abstractie toe te passen (om het ontwerpen van de database te vergemakkelijken), geeft dit model nooit een volledig beeld van de werkelijkheid. Er is altijd een discrepantie tussen het geconstrueerde model en het ideale model dat de echte wereld volledig en correct representeert. Door veranderingen in de wereld wordt dit verschil alleen maar groter.

Zelfs al zou de database een volledig model van de relevante fenomenen in de werkelijkheid bevatten, dan nog kunnen er problemen zijn met de kwaliteit van de dataset. De tweede uitdaging heeft daarom te maken met omissies in de data. Men kan, als gevolg van biases, nalatig zijn bij het verzamelen van benodigde gegevens en bewust of onbewust gegevens weglaten of vergeten. Als gevolg hiervan is de gebruikte dataset dus niet compleet. Een derde uitdaging is dat de kwaliteit van de gegevens in de dataset onvoldoende kan zijn. De dataset bevat dan bijvoorbeeld fouten of ruis. Vooral als het gaat om relatief oude of slecht gedocumenteerde data zijn deze tekortkomingen lastig te corrigeren.

Als gevolg van bovengenoemde problemen, is het beginpunt van een analyse vaak een imperfecte en incomplete dataset die de echte wereld niet goed, of maar gedeeltelijk, representeert. Dit heeft een negatief effect op de hypotheses die op basis van data-analysetechnieken gevonden worden. Maar ook als de achterliggende data perfect zouden zijn (en de drie voorgaande uitdagingen dus adequaat zijn geadres-

seerd), kan het verkeerd toepassen van data-analysetechnieken resulteren in foutieve hypotheses die ongefundeerd zijn. Het zorgdragen voor een juist gebruik van de verschillende algoritmen en het goed configureren ervan is daarmee de vierde uitdaging. Als dit niet goed gebeurt, komt de gevonden systeemwerkelijkheid niet (volledig) overeen met de echte werkelijkheid en kunnen verkeerde conclusies worden getrokken.

Een laatste uitdaging heeft betrekking op de interpretatie van de gevonden hypotheses. De geleerde profielen of regels bieden enkel een statische waarheid die geldt voor groepen van objecten of personen. Als ze worden toegepast op individuele gevallen, ontstaan er problemen, omdat dit niet vanzelfsprekend is. Daar komt bij dat, gegeven de vier eerdere uitdagingen, de profielen op een beperkt aantal variabelen gemaakt zijn, terwijl de individuele variatie groter is. De gevonden hypotheses zijn daardoor nooit 100% waar voor individuen. Vaak is er genoeg reden voor twijfel omdat cruciale aannames (bijvoorbeeld die van de gesloten wereld) niet gelden en de gevolgtrekkingen gebaseerd zijn op inductie. Het gebruik van datagedreven analysemethoden kent daarmee een aantal fundamentele problemen, die lastig zijn allemaal (tegelijk) op te lossen. Zelfs als het merendeel naar behoren wordt geadresseerd, dan nog kunnen de resultaten fundamenteel onjuist of onbruikbaar zijn.

Desalniettemin kunnen de analyseresultaten wel degelijk nuttig zijn voor de (beleids)praktijk. Voorzichtigheid is daarbij wel geboden. De gebruiker van de resultaten dient zich bewust te zijn van de geschetste beperkingen. Al met al geldt, hoe beter de kwaliteit van de gebruikte datasets, hoe beter de uitkomsten aansluiten bij de realiteit en hoe relevanter ze zijn voor de echte wereld. Zolang de discrepantie tussen de systeemwerkelijkheid en de echte werkelijkheid klein genoeg (of acceptabel) is, kunnen de resultaten wel degelijk helpen om betere beslissingen te nemen of processen te verbeteren. Ze kunnen dan zelfs gebruikt worden om voorspellingen te doen. Het verzamelen en analyseren van data moet daarom zorgvuldig worden gedaan. Hiertoe kunnen een aantal praktische handreikingen gegeven worden, die in hoofdstuk 5 worden besproken. Het is daarnaast de moeite waard om strategieën te ontwikkelen die helpen bij het op een goede manier interpreteren en toepassen van data-analyseresultaten. Enkele mogelijkheden worden in hoofdstuk 5 beschreven.

5 Praktische handreikingen

Gezien de problemen die spelen bij datagedreven analyses, is het zaak om weloverwogen te werk te gaan met een doordacht plan. Zo wordt de kans op biases verkleind en is de kans op een bruikbaar en betrouwbaar resultaat groter. Hoewel geavanceerde analysemethoden bij uitstek geschikt zijn om tot geheel nieuwe (en soms verrassende) inzichten te komen, is het aan te bevelen om vooraf goed na te denken over het doel van de analyse en potentiële problemen daarbij. Hieronder worden een aantal aanbevelingen voor de (beleids)praktijk gedaan om dit proces zo deugdelijk mogelijk te laten verlopen.

5.1 Datagerelateerde aanbevelingen

Het verzamelen van data kan vanuit twee verschillende manier gedaan worden: top-down en bottom-up. Bij een top-downaanpak wordt eerst de informatiebehoefte van de toekomstige gebruikers, zoals beleidsmedewerkers, vastgesteld. Er wordt dan bijvoorbeeld bekeken welke managementinformatie zij nodig hebben om beleid te evalueren en bij te sturen. Vervolgens wordt bepaald welke gegevens nodig zijn om hieraan te voldoen en hoe ze kunnen worden verkregen. In het geval van een bottom-upbenadering worden eerst gegevens verzameld, waarna onderzocht wordt wat ervan geleerd kan worden. Deze aanpak is geschikt als de informatiebehoefte vooraf niet bekend is. Men wil dan wel van de kansen die data kunnen bieden profiteren, maar weet nog niet (precies) welke inzichten ze op basis ervan kunnen verkrijgen. Deze laatste aanpak past daarom het meest bij datagedreven analysemethoden, maar brengt ook de nodige gevaren met zich mee. Daarom strekt het tot de aanbeveling om in zekere mate na te denken over het beoogde doel en de voor beleid benodigde (management)informatie.

Bij het verzamelen van data kunnen namelijk verschillende keuzes gemaakt worden. Zo kunnen data bijvoorbeeld op verschillende niveaus verzameld en opgeslagen worden. Het gaat dan om data op microniveau (bijvoorbeeld over individuen) of data op geaggregeerd niveau (statistische gegevens). Verschillende typen data vragen weer verschillende soorten systemen om ze in op te slaan.

Daarnaast speelt de kwaliteit en validiteit van de data een grote rol. Vooral als het gaat om beleidsgerelateerde (overheids)gegevens is dit van belang. Bij deze gegevens spelen namelijk verschillende problemen. Ten eerste, vindt het registreren van relevante gegevens vaak plaats op operationeel niveau (bijvoorbeeld door uitvoeringsorganisaties) voor operationele doeleinden. Dit is echter vaak niet de kernactiviteit van de betrokken organisaties en daarom komt het voor dat de data onvolledig, onnauwkeurig of inconsistent zijn. Uiteindelijk worden de gegevens dan gebruikt voor andere doeleinden dan waarvoor ze verzameld zijn. Ten tweede, komen in dit domein problemen met verouderde gegevens in *legacy*systemen relatief vaak voor. Ten derde, verandert de betekenis van de gegevens regelmatig als gevolg van wet- en regelgeving. Deze drie problemen maken het in de beleidspraktijk extra belangrijk om de kwaliteit van de gegevens te beoordelen voordat ze geanalyseerd worden. Fouten en inconsistenties moeten zo veel als mogelijk opgelost worden, maar dit is een kostbaar en tijdrovend proces waarvoor vaak domeinkennis nodig is.

In sommige gevallen moet helaas geconstateerd worden dat de vereiste gegevens niet beschikbaar zijn of van onvoldoende kwaliteit zijn (veel informatie ontbreekt of is onbetrouwbaar). Dan moet de aanbeveling zijn om niet verder te gaan

met de analyse en te kijken naar alternatieven. Er kan dan bekeken worden of op andere manieren betrouwbare managementinformatie verkregen kan worden, bijvoorbeeld door gebruik te maken van data die sterk gerelateerd zijn aan de oorspronkelijk gewenste data. Ook hierbij is domeinkennis van deskundigen van grote waarde.
Deze aanbevelingen laten zien dat het dataverzamelingsproces niet enkel voorbehouden is aan databasespecialisten en softwareontwikkelaars. Het is zaak om hierbij ook de eindgebruikers (voor het vaststellen van de informatiebehoefte) en domeinexperts (voor het interpreteren en controleren van de data) te betrekken. Samen kunnen zij dan zorgen voor een betrouwbare en geschikte dataset voor het beoogde doel.

5.2 Analysegerelateerde aanbevelingen

Ook bij het daadwerkelijk analyseren van de verzamelde data is het van belang om doordacht te werk te gaan. Hoewel de analyses voor een groot deel geautomatiseerd zijn, speelt mensenwerk een belangrijke rol, bijvoorbeeld bij het begeleiden van het proces en het beoordelen en interpreteren van de resultaten. Daarbij kunnen verschillende biases de uitkomsten verstoren. Zo kan degene die de analyse uitvoert vooringenomen zijn en een analyse (bewust of onbewust) een bepaalde kant op sturen. Keuzes in de analysefase bepalen in belangrijke mate de uitkomsten. Voor de uitvoerder van de analyse (bijvoorbeeld een data-analist) gelden daarom de volgende vuistregels.
Allereerst, moet de analist bekijken welk type analyse het meest geschikt is gegeven de beschikbare data, het doel van de analyse, en de informatiebehoefte van de gebruiker. Hierbij moet goed gelet worden op de voor- en nadelen van de verschillende technieken. Ook dient het geselecteerde algoritme op de juiste manier geconfigureerd te worden en moet op hoofdlijnen duidelijk zijn hoe het werkt. Daarnaast kunnen verschillende algoritmen gebruikt worden en de resultaten ervan vergeleken worden.
Een tweede aandachtspunt is de hoeveelheid data die gebruikt wordt in de analyse. Met het oog op *overfitting* is het niet goed om (te) weinig data te gebruiken, maar te veel data is ook niet altijd goed. Meer data betekent ook meer kans op fouten. Hoe groter de dataset, hoe groter de kans dat een significante correlatie gevonden wordt. Dit is echter lang niet altijd een betekenisvol (causaal) verband. Voor een zinvolle analyse moet dus een zekere hoeveelheid relevante data beschikbaar zijn, maar het is lastig vooraf te bepalen hoeveel “genoeg” is.
Daarnaast moet de data-analist zich bewust zijn van de kwaliteit van de gebruikte data. Om dit te kunnen beoordelen moet duidelijk zijn hoe de data tot stand zijn gekomen, hoe ze zijn verkregen, en hoe eventuele selecties zijn gemaakt. Zonder deze kennis over het verzamelproces, kan de dataset überhaupt niet zomaar als betrouwbaar worden gezien. Ook als er in het verzamelproces geen tekortkoming zijn, moet de data-analist op de hoogte zijn van de betekenis van de data en eventuele problemen met de kwaliteit. Dit vereist een verificatieproces, waarbij vaak de hulp van domeinexperts nodig is, helemaal als documentatie ontbreekt of onvolledig is. Een goede verificatie van de dataset voorkomt fouten en draagt bij aan de zekerheid van de resultaten. Voor data-analyses geldt namelijk het principe van “*Garbage in, garbage out*”. Analyses leveren alleen zinvolle informatie als zij met de juiste informatie gevoed worden.
Gerelateerd aan de hierboven genoemde aandachtspunten, moet de analist een zinvolle strategie gebruiken. Het is niet altijd het beste om willekeurig naar verbanden in data te zoeken. Die worden geheid gevonden, maar zijn lang niet altijd betekenis-

vol. Een gedegen analyse begint daarom met een set van onderzoeksvragen die passen bij de beschikbare data (en zijn gebaseerd op domeinkennis en krachtige relevante theorieën). De gevonden patronen kunnen nieuwe ideeën opleveren die vervolgens met hypotheses getoetst kunnen worden. De data-analist moet zich bij het uitvoeren van de analyse ook bewust zijn van de (cognitieve) biases die kunnen optreden. Het is zaak om open een analyse in te gaan en de gevonden hypotheses achteraf te valideren met behulp van aanvullende toetsen. Zo moet de analist bedacht zijn op schadelijke effecten zoals *overfitting* en *confounding* variabelen.

Ook voor de gebruiker van de analyseresultaten gelden een aantal aanbevelingen. Allereerst, moet de gebruiker van de data goed op de hoogte zijn van de werkwijze van de analist. Hij moet ervan op aan kunnen dat voldoende waarborgen in het verzamel- en analyseproces zijn gehanteerd (bijvoorbeeld door het gebruik van een aselecte steekproef en training- en testsets). Als er onduidelijkheden over de werkwijze (blijven) bestaan, is voorzichtigheid geboden bij het interpreteren en gebruiken van de resultaten. Dit is extra van belang als de resultaten in de praktijk, op individuele gevallen, worden toegepast. Het is daarom nodig om een goede strategie te hebben voor het toepassen van de resultaten. Voor verschillende doeleinden zijn verschillende benaderingen geschikt.

Analyseresultaten kunnen (als nieuw onderdeel van onze kennis) gebruikt worden om hypotheses over individuele zaken af te leiden. Een eerste strategie hierbij is om te zoeken naar bewijsmateriaal dat de gevonden hypothese voor het specifieke geval ondersteunt. Dit materiaal mag niet gebaseerd zijn op of afgeleid worden uit gegevens die al in de data-analyse gebruikt zijn. Dergelijke gegevens bieden namelijk geen nieuwe informatie over de hypothese. Als deze toch gebruikt worden, dan zal de hypothese op een onjuiste en onrechtmatige manier versterkt worden. Als er voldoende bewijzen uit alternatieve bronnen gevonden worden, kan de hypothese geaccepteerd worden. Een nadeel van deze strategie is echter dat het de *confirmation bias* kan versterken. Een tweede strategie is om juist te zoeken naar bewijs dat de hypothese tegenspreekt. Hierbij mogen gegevens die eerder in de

Box 10 Voorbeeld interpretatiestrategieën

Stel dat op basis van data-analyse het volgende profiel is gevonden: “jonge mannen die in postcodegebied 1234 wonen, hebben een hoger dan gemiddelde kans op het plegen van sociale zekerheidsfraude”. Stel nu dat er een individu is, een 20-jarige man, die in dit gebied woont. Op basis van het gevonden profiel kan dan de hypothese geformuleerd worden dat deze man fraude pleegt.

Conform de eerste strategie kan worden gezocht naar bewijzen die deze hypothese ondersteunen. Stel dat daarbij wordt gevonden dat deze persoon zijn afspraken met de sociale dienst steeds op korte termijn afzegt en dat uit zijn bankafschriften blijkt dat hij dagelijks meerdere keren voor kleine bedragen tankt bij steeds hetzelfde tankstation. Dit kan beschouwd worden als indicatie dat deze jongeman zwartwerkt als taxichauffeur. Op basis hiervan kan de initiële hypothese geaccepteerd worden en een aanvullend onderzoek naar deze persoon worden gestart.

Als een alternatief kan volgens de tweede strategie worden gezocht naar bewijs dat de hypothese ontkracht. Stel dat daarbij het volgende nieuwe profiel gevonden wordt: “mannen die in postcodegebied 1234 wonen en een uitkering ontvangen, plegen over het algemeen geen fraude als ze in de afgelopen 15 jaar geen (ander) misdrijf hebben begaan”. Als de 20-jarige aan dit profiel voldoet, kan dit een indicatie zijn dat hij geen fraudeur is. In dat geval zijn er twee elkaar tegensprekende profielen op basis waarvan de initiële hypothese afgewezen zou kunnen worden.

analyse gebruikt zijn, wel hergebruikt worden. Deze kunnen dan gebruikt worden om alternatieve hypotheses af te leiden. Aanvullend kunnen ook andere datasets geraadpleegd worden. Als er genoeg afzwakkend bewijs gevonden wordt, dient de initiële hypothese verworpen te worden.

Welke strategie het meest geschikt is om te gebruiken hangt af van de (aard) van de applicatie en de daarmee samenhangende impact van mogelijke *false positives* en *false negatives*. Een *false positive* is een hypothese die ten onrechte is geaccepteerd (terwijl deze eigenlijk onwaar is). Er is als het ware sprake van vals alarm. Een voorbeeld is een medische test die aangeeft dat iemand een bepaalde ziekte heeft, terwijl dat in werkelijkheid niet zo is (de hypothese van ziekte wordt op basis hiervan ten onrechte aanvaard). In de statistiek wordt dit ook wel een type I fout genoemd. Een *false negative* is een onterecht verworpen hypothese (terwijl deze eigenlijk waar is). Een voorbeeld is een medische test die aangeeft dat iemand niet ziek is, terwijl diegene dat in werkelijkheid wel is (de hypothese van ziekte wordt op basis hiervan ten onrechte afgewezen). In de statistiek wordt dit een type II fout genoemd.

De eerst genoemde strategie vermindert de kans op *false negatives*, maar verhoogt de kans op *false positives*. Voor de tweede strategie geldt het omgekeerde: er is meer kans op *false negatives*, maar tegelijk minder kans op *false positives*. Afhankelijk van de impact van verkeerde conclusies kan een van de twee strategieën (of een combinatie ervan) gekozen worden. Bij gevoelige toepassingen die een grote impact hebben op het leven van individuen, wordt de tweede strategie aanbevolen. De eerste strategie kan de voorkeur hebben voor toepassingen die minder gevoelig zijn (bijvoorbeeld voor toepassingen op het gebied van marketing). Voordat een strategie wordt toegepast, is het zinvol om een inschatting te maken van de impact van *false positives* en *negatives*, en na te denken over hoe hierop te anticiperen. Zo wordt de tweede strategie toegepast in de (Nederlandse) strafrechtspraak. Op basis van het beschikbare bewijs formuleert het Openbaar Ministerie eerst een hypothese over de (schuld van de) verdachte. Vervolgens probeert de advocaat van de verdachte deze hypothese te verwerpen door (ontlastend) tegenbewijs te presenteren. Deze strategie is gekozen om *false positives* zo veel mogelijk te voorkomen en dus onterechte veroordelingen tegen te gaan. De impact van gerechtelijke dwalingen op individuele veroordeelden is groot, en daarom is deze strategie voor deze toepassing verdedigbaar.

De eerste strategie, daarentegen, kan het beste toegepast worden in situaties waarin de impact van *false negatives* groot is. Dit is bijvoorbeeld het geval bij het inschatten van het risico op terroristische aanslagen. Het is dan van belang om zicht te krijgen op mogelijke daders (en hierop snel te anticiperen indien nodig) om grote(re) maatschappelijke schade te voorkomen. Het belang van een grote groep (potentiële) slachtoffers weegt dan zwaarder dan het belang van het individu (een onterecht als terrorist aangemerkt persoon).

6 Conclusie

In de beleidspraktijk bestaat de behoefte om te profiteren van de explosieve groei aan beschikbare gegevens en de recente ontwikkelingen op het gebied van big data. Het wordt daarom steeds gangbaarder voor overheidsorganisaties om gegevens uit verschillende bronnen en informatiesystemen te verzamelen en deze vervolgens met geavanceerde technieken (denk aan datamining) te analyseren. Dit met het praktische oogmerk om bijvoorbeeld processen efficiënter te maken en de dienstverlening aan burgers te verbeteren of uit te breiden. Het gebruik van data-analyseresultaten is in de praktijk echter om verschillende redenen niet eenvoudig.
Zo is hierboven uitgelegd dat er bij het verzamelen (en eventueel integreren) van gegevens grote uitdagingen zijn, vooral in het publieke domein waarin vaak sprake is van verouderde data en systemen. Daarnaast hebben we te maken met een systeemwerkelijkheid die hoogstens een interpretatie is van de echte werkelijkheid, en is het niet vanzelfsprekend om statistische waarheden toe te passen op individuele gevallen. Deze problemen worden veroorzaakt door een aantal fundamentelere oorzaken: we hebben te maken met een aanname een van gesloten wereld die niet geldt, een onvolledig model van de echte wereld (als gevolg van modelleringsprincipes zoals *universe of discourse* en abstractie), en afleidingen die gebaseerd zijn op inductie (en daarom nooit 100% zeker zijn). Deze aspecten belemmeren een goede interpretatie van de resultaten en de bruikbaarheid in de praktijk.
In dit memorandum hebben we een aantal praktische aanbevelingen gedaan om de hier geschetste problemen zo veel mogelijk te ondervangen. Zo hebben we twee strategieën geschetst voor (correct) gebruik van analyseresultaten in individuele zaken. Afhankelijk van het toepassingsgebied, en de impact van foute conclusies op individuen of de samenleving, kan een van de strategieën of een combinatie van beide strategieën gekozen worden. Zolang de aanbevelingen worden nageleefd, de strategieën op een juiste manier worden toegepast, en de gebruiker op de hoogte is van de beperkingen, kunnen datagedreven analysemethoden een nuttige toevoeging zijn op traditionele methodieken uit de beleidspraktijk. Ze kunnen leiden tot nieuwe inzichten die op een andere manier niet verkregen hadden kunnen worden. Over het algemeen geldt wel dat voorzichtigheid geboden is, vooral als op basis van de analyseresultaten beslissingen worden genomen over kwetsbare personen of die discriminatie of stigmatisatie tot gevolg kunnen hebben.