

STAtOR

periodiek van de VVSOR jaargang 21, nummer 3, september 2020

A tale of two networks

Wetenschappelijk onderzoek klopt vaak niet

Menselijke datawetenschap

Hoe betrouwbaar zijn loterijen

Een boost voor verduurzaming; Elektrificatie van het busvervoer in de regio Rotterdam

Modelleren, een vak apart

Subtype-detectie op basis van variabelenstructuur;
Een combinatie van clustering en netwerk-methoden

Over confusiematrices, pooldiagrammen
en de orde van de dingen



STAtOR

Jaargang 21, nummer 3, september 2020

STAtOR is een uitgave van de Vereniging voor Statistiek en Operations Research (VVSOR). STAtOR wil leden, bedrijven en overige geïnteresseerden op de hoogte houden van ontwikkelingen en nieuws over toepassingen van statistiek en operations research. Verschijnt 4 keer per jaar.

Redactie

Joaquim Gromicho (hoofdredacteur), Annelieke Baller, Joep Burger, Kristiaan Glorie, Caroline Jagtenberg, Guus Luijben (eindredacteur), Kerry Malone, Richard Starmans, Gerrit Stermerdink (eindredacteur), Vanessa Torres van Grinsven en Sanne Willems. Vaste medewerkers: John Poppelaars, Gerard Sierksma en Henk Tijms.

Kopij en reacties richten aan

Prof. dr. J.A.S. Gromicho (hoofdredacteur), Faculteit der Economische Wetenschappen en Bedrijfskunde, afdeling Econometrie, Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, mobiel 06 55886747, j.a.dossantos.gromicho@vu.nl

Bestuur van de VVSOR

Voorzitter: prof. dr. Fred van Eeuwijk, db@vvsor.nl; Secretaris: Laura Zwep MSc, secretaris@vvsor.nl; Penningmeester: Judith ter Schure MSc, penningmeester@vvsor.nl; Bestuurscoördinator: Pieter Jongsma MSc, db@vvsor.nl; Webmaster: Mitra Ebrahimpoor MSc, Nikky van Buuren MSc, webmaster@vvsor.nl
Voorzitters van de secties: prof. dr. ir. Mark van de Wiel (Biometrical Section); prof. dr. Albert Wagelmans (Section for Operations Research); dr. Eduard Belitser (Section Mathematical Statistics); prof. dr. Casper Albers (Social Sciences Section); dr. Michel van de Velden (Economics Section); dr. Daniel Oberski (Section Data Science); Jonas Haslbeck MSc (Young Statisticians) dr. Sanne Willems (Section Statistics Communication).

Leden- en abonnementenadministratie van de VVSOR

VVSOR, Postbus 1058, 3860 BB Nijkerk, telefoon 033 2473408, admin@vvsor.nl. Raadpleeg onze website www.vvsor.nl over hoe u lid kunt worden van de VVSOR of een abonnement kunt nemen op STAtOR.

Voor advertenties

M. van Hootegem, hootegem@xs4all.nl
STAtOR verschijnt in maart, juni, september en december.

Ontwerp en opmaak

Pharos, Nijmegen

Uitgever

© Vereniging voor Statistiek en Operations Research
ISSN 1567-3383

TOEVAL WERKT!

Toeval werkt, zo laat ook dit nummer zien. Natuurlijk weten we dit al lang, onze tak van wetenschap berust erop. En we weten zelfs dat volgens sommige theorieën de oerknal, en dus onze existentie, het gevolg is geweest van een toevalsfluctuatie.

Maar als we dan in de redactionele praktijk ervaren dat we een afgewogen geheel van artikelen hebben die mooi op elkaar aansluiten, zonder dat daarbij enige sturing aan te pas kwam, dan kunnen we alleen maar voor de zoveelste keer onze bewondering voor het toeval uitspreken.

Zo hebben we twee verwante artikelen die beide beschrijven hoe het zoeken naar onderliggende netwerken verloopt. Wessel van Wieringen doet dit aan de hand van biochemische processen in de cel waarbij kankercellen een ander netwerk vertonen dan metastase cellen. En Vladimir Hazeleger beschrijft subtype-detectie op basis van variabelenstructuur bij psychiatrische aandoeningen. Boeiend om over twee totaal verschillende toepassingen te lezen van een in principe gelijke techniek.

Daniel Oberski levert een zeer leeswaardige kijk op menselijke datawetenschap, iets dat, in een geheel andere vorm, ook het onderwerp is van de column van John Poppelaars.

De journalisten Ans Hekkenberg en Jim Jansen interviewden Casper Albers voor de *New Scientist*. Wij zijn blij dat we dit interview mochten overnemen voor STAtOR.

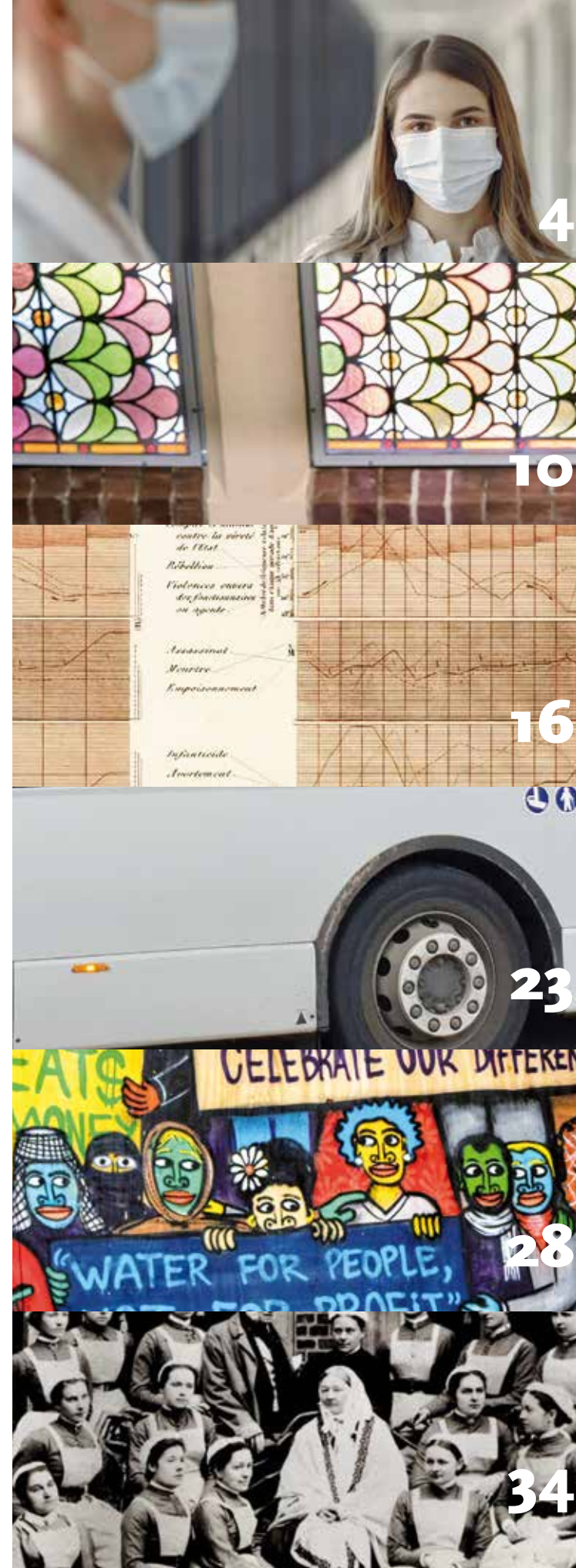
Hoe praktisch onze wetenschap kan zijn, lezen we bij Pieter van den Berg. Hij laat zien dat het omschakelen van diesel naar elektrisch bij stadbussen niet zomaar gaat, maar door OR toch lukt.

Richard Starmans schrijft over de rol van Blaise Pascal en Florence Nightingale in de ontwikkeling van de statistiek. Bijzonder aanbevolen!

In ons vorige nummer konden we op de valreep het overlijden van ons erelid Willem van Zwet melden, in dit nummer gaat Aad van der Vaart in op de grote rol die Willem wetenschappelijk en als bestuurder heeft gehad.

En geen nummer is compleet zonder columns, behalve die van John Poppelaars nog twee: over bedrog en loterijen en over zwakke magen van statistici.

Wij wensen u veel leesplezier!



INHOUD

- 2 Redactioneel – Toeval werkt!**
- 4 A tale of two networks**
| WESSEL VAN WIERINGEN
- 10 Wetenschappelijk onderzoek klopt vaak niet; Interview met Casper Albers**
| ANS HEKKENBERG & JIM JANSEN
- 16 Menselijke datawetenschap**
| DANIEL OBERSKI
- 21 Hoe betrouwbaar zijn loterijen – column**
| HENK TIJMS
- 23 Een boost voor verduurzaming; Elektrificatie van het busvervoer in de regio Rotterdam**
| PIETER VAN DEN BERG
- 26 Modelleren, een vak apart – column**
| JOHN POPPELAARS
- 28 Subtype-detectie op basis van variabelenstructuur; Een combinatie van clustering en netwerk-methoden**
| VLADIMIR HAZELEGER
- 34 Over confusiematrices, pooldiagrammen en de orde van de dingen**
| RICHARD STARMANS
- 41 Kaas en een open koe – column**
| GERRIT STERMERDINK
- 42 Willem Rutger van Zwet 1934 – 2020: vermaard wetenschapper en krachtig bestuurder – in memoriam**
| AAD VAN DER VAART
- 46 Nieuws over het World Statistics Congress 2021**
- 47 VVSOR Annual Meeting 2021: Going viral – Modeling spread**



Foto: Gustavo Fring via Pexels

A TALE OF TWO NETWORKS

De modieuze term *network medicine* refereert aan het gebruik van netwerken binnen de medische wetenschappen. Het gebruik van netwerken binnen dit domein staat nog in de kinderschoenen. Deels vanwege de wellicht terechte terughoudendheid van de medische wereld, deels omdat er weinig of geen netwerk-equivalenten bestaan van de statistische methodologie om de geëigende vragen uit dit veld te beantwoorden. Er is geen standaardtechniek, zoals de t-toets, die binnen de twee-steekproevencontext verschillen tussen netwerken identificeert. Hier volgt een tussentijdse rapportage van onze zoektocht naar zulke technieken, de voorlopige vruchten daarvan, en onze ervaringen met de toepassing van deze vruchten binnen het Amsterdam UMC.

WESSEL N. VAN WIERINGEN

De wetenschap is lang gedomineerd door de reductionistische werkwijze. Een aanpak die het geheel opdeelt en de delen apart bestudeert. Die aanpak heeft veel gebracht, maar kent beperkingen. De systeemaanpak poogt die beperkingen te omzeilen door de delen in hun totaliteit en samenhang te bestuderen. Deze aanpak heeft geleid tot nieuwe wetenschappelijke disciplines zoals systeembiologie die onder meer de complexe interacties binnen een biologische systeem als de cel poogt te doorgronden. Eerlijkheid gebiedt te vermelden dat binnen de systeembiologie omwille van de hanteerbaarheid vaak ingezoomd wordt op een *pathway*, een module binnen het regulatoire netwerk van de cel, die een specifieke functie vervult.

Een belangrijk ingrediënt van de systeemaanpak is de overdrachtelijk aantrekkelijke representatie van een systeem als een netwerk (zie figuur 1 voor een voorbeeld), dat in een oogopslag overzicht over een complex fenomeen verschaft. Een netwerk is een collectie punten die met lijnen verbonden zijn. De punten corresponderen

met de elementen van het systeem en de lijnen met hun interacties. Binnen de context van de cel representeren de element moleculen, bijvoorbeeld mRNAs, proteïnen, of metabolieten. Een interactie zou dan het aangaan van een chemische verbinding kunnen voorstellen.

Het netwerk is vaak onbekend en dient met behulp van beschikbare informatie gereconstrueerd te worden. Dat vergt een model dat netwerk en data koppelt. Dit wordt onder andere geboden door het Gaussische grafische model. Dit model veronderstelt dat de data een multivariate normale verdeling volgen, waarbij de covariantiematrix geparametriseerd wordt door zijn inverse, i.e. de precisiematrix. Toegepast op de cel beschrijft het model de stochastische activiteit in statistische toestand van een verzameling moleculen waarbij hun interacties geoperationaliseerd zijn als paarsgewijze conditionele onafhankelijkheden. Het voordeel van voornoemde parameterisatie is dat een nul in de precisiematrix onafhankelijkheid tussen het corresponderende random variabe-

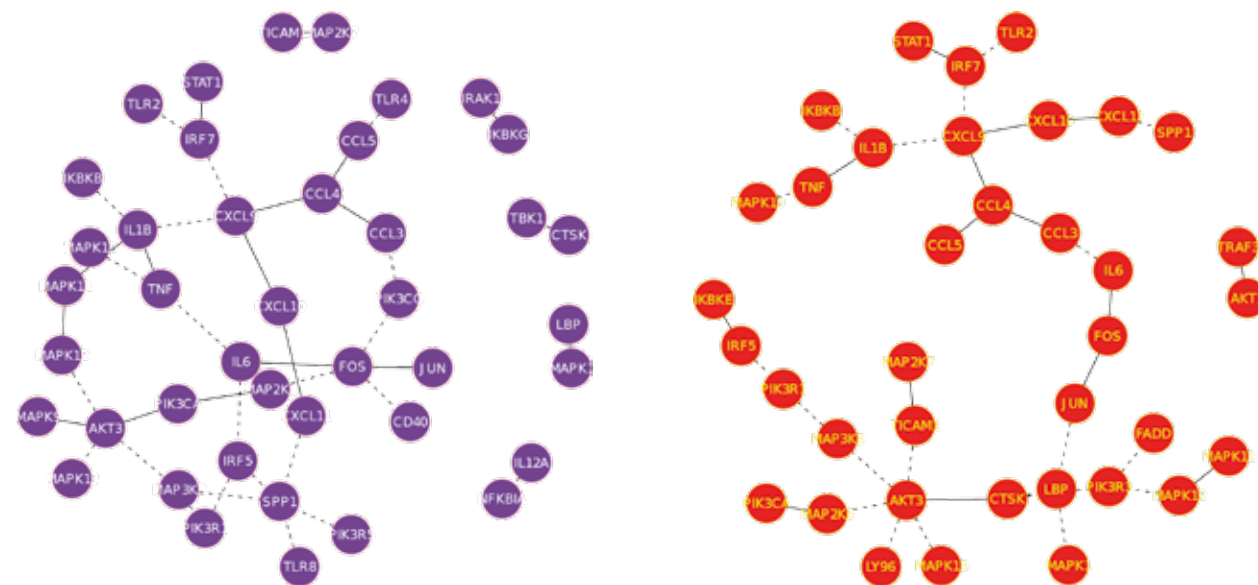
lenpaar, conditioneel op alle andere random variabelen in het model, impliceert. Dit is ook direct de link met het netwerk: twee punten, die de random variabelen geassocieerd met twee moleculen representeren, zijn niet/wel met een lijn-verbonden als zij conditioneel on/afhankelijk zijn. Met een statistisch model in de hand en data voorhanden kan het netwerk gereconstrueerd worden middels het schatten van de precisiematrix. De geschatte precisiematrix wordt gestandaardiseerd ten einde de partiële correlatiematrix, de matrix met paarsgewijze conditionele correlaties, te verkrijgen, op basis waarvan bepaald wordt welke elementen al dan niet gelijk aan nul zijn.

Medische onderzoekers raken gecharmeerd van een mooie netwerkvisualisatie, zeker als die op basis van hun data gegenereerd is. Die roze wolk drijft echter snel over als ze de praktische betekenis ervan proberen te duiden. Bijvoorbeeld omdat het gepresenteerde netwerk amper rijmt met de mentale voorstelling die de medische onderzoeker op basis van literatuur en ervaring zich ervan ge-

vormd had. Of, omdat de onderzoeker de vertaling naar de kliniek probeert te maken en één enkel netwerk geen inzicht in de verschillen tussen gezonde cel en kankercel met betrekking tot hun moleculaire interacties geeft, welke mogelijke aanknopingspunten voor een nieuwe behandeling zouden kunnen zijn. Panklare en uitgekristalliseerde methodologie voor het reconstrueren van meerdere netwerken gezamenlijk en/of hun contrast is nog niet breed gezaaid. Hieronder een schets van hoe wij daar tot nu toe een beginnend handvat aan hebben gegeven.

Regularisatie

Data zijn binnen de medische context een schaars goed. Vaak zijn data van een beperkt aantal individuen beschikbaar, terwijl het aantal betrokken random variabelen van dezelfde orde – zo niet groter – is. Ter illustratie, het aantal kankerpatiënten met hetzelfde tumortype binnen het



Figuur 1. Gereconstrueerde netwerken van de 'toll-like receptor' pathway voor de kanker- (links, paarse punten) en metastasecel (rechts, rode punten). De gestreepte en doorgaande verbindingen representeren negatieve en positieve partiële correlaties

Amsterdam UMC, dat toestemming voor wetenschappelijk onderzoek met hun weefsel verleent, is in de orde van honderd, terwijl het aantal moleculen in de cel – in principe – vele malen groter is. Dit vertaalt naar netwerkreconstructie zijn aanpassingen van de traditionele methoden geboden, bijvoorbeeld door regularisatie van de maximum likelihood-schatter van de precisie-matrix. De twee dominante regularisatiesmaken zijn de l_1 - en l_2 -regularisatie, respectievelijk 'lasso'- en 'ridge'-regularisatie genaamd (zie Friedman et al., 2008; van Wieringen et al., 2016). De regularisatie behelst het toevoegen van een met een regularisatieparameter geschaalde som van de absolute waarde (lasso) of het kwadraat (ridge) van het element van de precisie-matrix aan de loglikelihood-functie. De geregulariseerde schatter is vervolgens het argument dat maximalisatie van het geheel realiseert. De lasso-regularisatie is geliefd om zijn selectievermogen. Voor voldoende grote waarden van de voornoemde regularisatieparameter bevat de corresponderende schatting van de precisie-matrix nul elementen. De facto selecteert de lassoregularisatie welke conditionele afhankelijkheden door de data onderbouwd worden. Dat is een aantrekkelijke en praktische eigenschap. Met een beperkte steekproefgrootte resulteert lasso-regularisatie echter vaak in een leeg netwerk. Ook kent de evaluatie van de lasso- schatter convergentieproblemen als de precisie-matrix een te groot aantal niet-nul elementen herbergt. Ridge-regularisatie mag dan geen ingebouwde selectie kennen, er is wel een analytische uitdrukking voor de schatter beschikbaar. Hoewel een wiskundige toevalligheid, het heeft praktische waarde: geen convergentieproblemen. Belangrijker is dat post-hoc selectie op basis van

de l_2 -geregulariseerde schatter, bijvoorbeeld met behulp van een probabilistisch gemotiveerd afkappunt, zelden onderdoet voor die van de lasso-tegenhanger. Tevens is de fit van de ridge-schatting vaak beter. In het vervolg beperken we ons derhalve tot de ridge-regularisatie, maar een lasso-variant is nooit ver weg.

Om tot een schatter te komen moet er een keuze voor de regularisatieparameter gemaakt worden. Deze wordt gekozen op basis van data, bijvoorbeeld middels een kruisvalidatieprocedure. Zo'n procedure kiest de waarde voor de regularisatieparameter die resulteert in de beste predictie op nieuwe data. Nieuwe data zijn echter niet altijd voorhanden. In dat geval wordt een deel van de beschikbare data apart gezet om de rol van 'nieuwe' data te vervullen. De schatter, voor een gegeven waarde voor de regularisatieparameter, wordt dan bepaald op basis van alle behalve de 'nieuwe data' en het met de schatter geassocieerde predictief vermogen wordt vervolgens geëvalueerd op de 'nieuwe' data. Vaak worden op deze wijze meermaals 'nieuwe' data gegenereerd om de afhankelijkheid van het toeval in de vorming van de oorspronkelijke 'nieuwe' data te reduceren. Het predictief vermogen wordt gemiddeld over alle versies van de 'nieuwe' data. Dit predictief vermogen is dan bepalend voor de optimale regularisatieparameter.

Voorkennis

Als een medische onderzoeker een netwerk gepresenteerd wordt dat is gereconstrueerd op eigen data, is zijn of haar eerste reactie wellicht dat het niet – of in ieder geval niet volledig – overeenkomt met de netwerken die eerder in

de literatuur gerapporteerd zijn. Deels komt dat omdat de eigen data niet van een studie met identieke opzet afkomstig zijn: de populatie, de meetmethode, en de laboratoriumomstandigheden kunnen verschillen. Hoog-dimensionaal speelt steekproefvariatie ook een grote rol. Tevens zijn de in de literatuur aanwezige netwerken vaak op andere wijze, bijvoorbeeld met behulp van *text-mining*, gereconstrueerd. Hoewel deze excuses gerechtvaardigd zijn, gaan ze voorbij aan het impliciete punt van de medische onderzoeker, namelijk dat het terrein van moleculaire netwerken geen tabula rasa is. Zij dienen dan ook niet de novo gereconstrueerd te worden maar gebruik te maken van de reeds beschikbare kennis – hoe vaag en onbetrouwbaar ook – omtrent deze netwerken. Beschikbare kennis is beschikbaar middels literatuur maar ook als publiek beschikbare data van studies naar hetzelfde (of een vergelijkbaar) systeem in soortgelijke omstandigheden.

Voorkennis kan behulpzaam zijn bij het leren van een Gaussisch grafische model (van Wieringen, 2019). Voorkennis kan kwalitatief en kwantitatief zijn. Kwalitatieve voorkennis betreft informatie omtrent de topologische eigenschappen van het netwerk. Bijvoorbeeld de aanwezigheid of afwezigheid van een verbinding, of een globale ketting-structuur. Kwantitatieve voorkennis komt in de vorm van een suggestie van (een deel van) de precisie-matrix. Zo'n suggestie kan een schatting van de precisie-matrix uit een in-house pilot-studie of uit publieke data van gerelateerde studies beschikbaar via online databases, zijn. De relevantie van deze voorkennis voor de parameters van het Gaussisch grafische model voor de voorhanden zijnde data is niet bekend. Derhalve incorporeren we de voorkennis niet expliciet in het model, maar indirect via regularisatie van de schatter. Daartoe is de regularisatie-term nu de som van het kwadraat van de elementen van het verschil van de precisie-matrix en de suggestie, waarbij – in de meest algemene vorm – elk uniek kwadraat met een eigen regularisatieparameter geschaald wordt. Duidelijk moge zijn dat voor grote waarden van één van de regularisatieparameters het corresponderende element van de precisie-matrixschatting een weinig verschilt van dat van de suggestie. In het bijzonder, is het corresponderende element van de suggestie gelijk aan nul, dan resulteert een hele grote waarde van de regularisatieparameter in een nul-element van de precisie-matrixschatting. Op deze wijze kan voorkennis omtrent afwezige verbindingen geïncorporeerd worden. Natuurlijk is het ondoenlijk om voor elke parameter een regularisatieparameter te selecteren. Een oplossing hiervoor is om een laag-dimensionale parametrisatie van de regularisatieparameters te kiezen. Bijvoorbeeld, als vermoed wordt dat het netwerk

een kettingstructuur heeft, zou men de elementsgewijze regularisatieparameter van de vorm λ^α met $\lambda > 0$ en α gelijk aan de afstand van het element tot de diagonaal van de precisie-matrix kunnen kiezen, teneinde een schatting met een gestreepte structuur – welk met een kettingstructuur correspondeert – te stimuleren.

Veel kankeronderzoek richt zich op vroege detectie van de ziekte, omdat dan het aantal behandelingsmogelijkheden het grootst is en de overlevingskans het hoogst. Dit heeft tot gevolg dat kanker vaak voor het uitzaait opgespoord is. Het gevolg daarvan is dat moleculaire informatie afkomstig van uitzaaiingen schaars is. Maar, omdat kanker een voorloperstadium van de uitzaaiing is, zal het moleculaire netwerk in de uitgezaaide cel niet totaal van dat van de kankercel verschillen, ook al muteert de kankercel gestaag verder. Ergo, een schatting van het/de netwerk/precisie-matrix in de laatste, die doorgaans op een grotere steekproef gebaseerd is, kan als voorkennis voor de reconstructie van dat in/van de eerste dienen. Dit principe passen we toe in de reconstructie van het netwerk van de *toll-like receptor-pathway*, welk bijdraagt aan het uitzaaien van de kankercel. Figuur 1 toont de aldus gereconstrueerde kanker- en metastase-netwerken. De netwerken kennen overeenkomstige verbindingen maar ook enkele verschillen. Reconstructie zonder suggestie resulteert in twee netwerken met schamele overeenkomsten, en wordt – hoewel de waarheid onbekend is – vanwege deze ongerijmdheid door de medische onderzoeker als minder plausibel gezien. Daarnaast is het geruststellend dat de reconstructie van figuur 1 een systeembio-logisch principe reflecteert. In tegenstelling tot het kanker-netwerk bevat het metastase-netwerk geen cycli. Dit wijst mogelijk op de eliminatie van feedback-loops, iets dat geassocieerd wordt met verminderde controle, kenmerkend voor de evolutie van de kankercel.

Fusie

Een veel voorkomende vraag in medische context is het in een bepaald opzicht vergelijken van twee (of meerdere) groepen, bijvoorbeeld de subtypes van de ziekte DLBCL, een bepaalde vorm van bloedkanker. Zo'n vergelijk geschiedt doorgaans op basis van data afkomstig van een observationele studie waarin de betrokken individuen ongerelateerd zijn. Statistisch gezien is dit een meervoudige onafhankelijk steekproefprobleem, nu voor netwerken. Niet gehinderd door enige biologische voorkennis kan men veronderstellen dat binnen elk subtype de data gegenereerd zijn door een ander Gaussisch graphisch mo-

del. De precisie-matrices worden dan per subtype apart geschat. Deze schattingen vormen vervolgens de basis voor de reconstructie van de subtype-specifieke netwerken.

Wat niet weet, dat niet deert. Maar wellicht geven de data meer prijs als enige biologische kennis meegenomen wordt. Bijvoorbeeld dat de cellen van subtypes van een bepaalde kanker dezelfde origine kennen, nl. de gezonde cel van het betreffende weefsel. Vanuit die cel zijn de cellen van de verschillende subtypes geëvolueerd. Deze gezamenlijke oorsprong suggereert dat de regulatoire netwerken van de tot subtype geëvolueerde cellen mogelijk nog enige overeenkomst kennen. Met die kennis in het achterhoofd doet het simpelweg schatten van een precisie-matrix per subtype mogelijk niet voldoende recht aan de biologie. De precieze relatie tussen de cel-specifieke netwerken is echter onbekend en kan derhalve niet expliciet gemodelleerd worden. Mogelijke overeenkomsten moeten dan op een meer heuristische wijze gevangen worden. Dit doen we door de geregulariseerde log-likelihoodfunctie van een additionele regularisatie-term te voorzien. We gebruiken hiervoor een zgn. fusie-regularisatieterm (Bilgrau et al., 2020). Deze wordt wederom gevormd door een regularisatieparameter geschaalde som van het gekwadeerde elementsgewijze verschil tussen de groepspecifieke precisie-matrices. Voor grote waarden van de regularisatieparameter forceert fusie-regularisatie de schattingen van de precisie-matrices naar elkaar toe, terwijl voor kleine waarden deze schattingen vrij (voor zover niet ingeperkt door de reguliere, i.e. subtype-specifieke regularisatie) gekozen kunnen worden. Daar de regularisatieparameters met behulp van data geleerd worden, bepalen de data in hoeverre er gefuseerd wordt en daarmee of de schattingen van de verschillende precisie-matrices op elkaar lijken. Op deze wijze hebben de data uit de ene groep dus invloed op de schatting van de precisie-matrix in de andere groep. Fusie-regularisatie biedt dus de mogelijkheid om informatie te ‘lenen’ tussen de subtypes.

Terug naar DLBCL. De ziekte kent drie subtypes, ABC, GCB en Type III, waarvan de laatste eerder tussen de twee andere inhangt dan dat het een op zichzelf staand homogeen subtype vormt. De ABC- en GCB-subtypes hebben een duidelijk verschillende klinische prognose. De vraag is of dit een moleculaire grond heeft. Zo ja, dan uit zich dat mogelijk in hun cellulaire netwerken, hetgeen een aanknopingspunt voor de ontwikkeling van effectievere behandelingen biedt. Met behulp van de fusie-regularisatie methode hebben we per subtype een Gaussisch grafische model uit gen-expressie/activiteit data afkomstig

van DLBCL patiënten geschat. Hierin kiezen we ervoor om enkel een fusie-regularisatieparameter voor het verschil tussen ABC en GCB enerzijds en Type III anderzijds te gebruiken. De schatting van de precisie-matrices geassocieerd met de ABC- en GCB-subtypes kunnen dus enkel gelijk zijn als zij ook gelijk aan die van Type III zijn. De gekruisvalideerde fusie-regularisatieparameter is klein (t.o.v. van de subtype-specifieke regularisatieparameters) maar wel degelijk ongelijk van nul, i.e. de data van het ene subtype worden relevantie voor de precisie-matrix-schatting van het andere toegekend. Door een pragmatische post-hoc selectieprocedure worden vervolgens uit de schattingen van de precisie-matrices de netwerken verkregen. De gereconstrueerde netwerken van de ABC- en GCB-subtypes laten differentiële submodules zien, welke reeds bekende moleculaire interacties bevatten, en mogelijke clous voor de verschillende prognose tussen de subtypes vormen.

Activatie

Het meervoudig steekproefprobleem resulteert in een lijst van elementen/verbindingen van de precisie-matrix/netwerk die verschillend zijn tussen de subtypes. Hoewel ze charmant zijn vanwege hun netwerkperspectief, zijn de items uit deze lijst vaak weinig tastbaar voor medische onderzoekers. Voor hen is een conditionele (on)afhankelijkheid een lastig concept. Medische onderzoekers interpreteren de lijst doorgaans vanuit hun kennis over één van de twee moleculen, kennis die meestal gebaseerd is op marginale informatie omtrent hun (paarsgewijze) verdeling. Een additionele complicatie is dat er veel meer mogelijke verbindingen dan punten in een netwerk zijn. Als de voornoemde lijst lang is, verliest men door de veelheid aan (differentiële) verbindingen al snel zicht op het netwerk. Een probleem van een andere orde met de gevonden differentiële conditionele onafhankelijkheid is haar reproduceerbaarheid. Een gevonden verschil in de locatie van de activiteit van een molecuul laat zich met regelmaat niet valideren in het laboratorium, voor een differentiële conditionele relatie blijkt dit problematischer.

Bovenstaande problemen met de resultaten van de differentiële netwerk-analyse kunnen deels omzeild worden door de verbindingen/conditionele onafhankelijkheden niet op individueel maar op een geaggregeerd niveau te bestuderen (van Wieringen et al., 2018). Men doet dan derhalve eerder een uitspraak over het ‘systeem’ in plaats van over de delen. Zo betreft de overkoepelde vraag van medische onderzoekers niet de interactie tussen twee

moleculen, maar eerder of het pathway – waar het moleculenkoppel, met vele anderen, deel van uit maakt – geactiveerd dan wel uitgeschakeld is. In extremis kan een uitgeschakeld pathway als een leeg netwerk, i.e. een netwerk zonder verbindingen, geportretteerd worden, terwijl een actief pathway juist verbindingen kent. De praktijk is eerder een grijsschaal tussen die twee uitersten, met een laag pitje, hoog vuur, en alles er tussenin. Anderzijds, vele moleculen zijn betrokken bij meerdere pathways, sommige actief en andere inactief. Het gereconstrueerde netwerk van een inactief pathway kan dus wel degelijk enige verbindingen vertonen. Ten einde een verzwakt/versterkt netwerk te detecteren hebben we het twee-steekproevenprobleem voor netwerken als volgt aangepast. We concentreren ons op de specifieke situatie waarin de partiële correlaties, dat wil zeggen de maat van de sterkte van de netwerkverbindingen, in een van de groepen in absolute zin systematisch kleiner/groter zijn. Systematisch zwakkere/sterkere partiële correlaties vertegenwoordigen biologisch een inactief/actief pathway. Om dit te modelleren wordt in elke groep wederom een Gaussisch graphisch model verondersteld, maar hun partiële correlaties zijn proportioneel en verschillen door een multiplicatieve factor, welke gemeenschappelijk is voor alle partiële correlaties. De factor reflecteert het algehele relatieve verschil in de sterkte van de conditionele afhankelijkheden tussen de groepen. De facto, delen de twee groepen dezelfde (gestandaardiseerde) precisie-matrix op een multiplicatieve factor na. De modelparameters worden (deels) weer op geregulariseerde wijze geschat. Een geschatte waarde van de factor ongelijk aan nul representeert dan een activiteitsverschil in de pathway tussen de twee groepen, wat door een permutatietoets formeel geëvalueerd kan worden. De methodologie is aangewend om een ‘cellijn’-vondst in data afkomstig van de minder gecontroleerde omgeving van het menselijk lichaam te valideren. De cellijnstudie heeft aangetoond dat de Hedgehog signaleringspathway, verantwoordelijk voor cellulaire differentiatie, in prostaatkanker geactiveerd is. Data afkomstig van zes studies, die de genactiviteit in normaal en kanker-prostaatweefsel vergelijken, zijn gebruikt om te verifiëren of dit activiteitsverschil zich ook in vivo voordoet. Vier van de zes studies bevestigen dit middels een factor die significant van nul verschilt, één studie wijst in dezelfde richting maar haalt het significantieniveau niet, en de laatste vertoont geen enkel verschil (de factor wordt zelfs als nul geschat). Voor diagnostische doeleinde zijn vervolgens de ongerelateerde precisie-matrices van beide groepen met de gepresenteerde fusie-geregulariseerde methode geschat. Op basis hiervan is de proportionali-

teitsaannname op relatie van de partiële correlaties tussen beide groepen geëvalueerd en in de meeste datasets onproblematisch gebleken. Tevens faciliteren deze schattingen het opsporen van de verbindingen met de grootste verschillen in sterkte om zo het activiteitsverschil verder te ontleden.

Conclusie

Het bovenstaande illustreert de vruchtbare wisselwerking tussen twee disciplines. De vragen en het commentaar van medische onderzoekers op gepresenteerde netwerk-resultaten inspireren nieuw methodologisch onderzoek naar netwerk-methodes. De toepassing van de nieuwe methodes op medische data genereert nieuwe resultaten, welke weer nieuwe vragen van de medische onderzoekers oproepen, die enkel beantwoord kunnen worden met nieuwe statistiek, enzovoort. Inderdaad is er nog veel werk te verzetten. Zo beperkt het bovenstaande zich tot het Gaussische domein. Anderzijds, kunnen de data van een studie met longitudinale opzet afkomstig zijn, hetgeen afhankelijkheden tussen de observaties introduceert. Het bovenstaande concentreert zich op modelleren en schatten. Een goed paradigma voor hypothese toetsen binnen de netwerkcontext, waarbij medische vragen leidend zijn voor de relevante hypothesen, is echter ook onontbeerlijk. Kortom, wordt vervolgd.

LITERATUUR

- Bilgrau, A. E., Peeters, C. F., Eriksen, P. S., Bogsted, M., & van Wieringen, W. N. (2020). Targeted fused ridge estimation of inverse covariance matrices from multiple high-dimensional data classes. *Journal of Machine Learning Research*, 21(26), 1–52.
- Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432–441.
- van Wieringen, W. N. (2019). The generalized ridge estimator of the inverse covariance matrix. *Journal of Computational and Graphical Statistics*, 28(4), 932–942.
- van Wieringen, W. N., & Peeters, C. F. W. (2016). Ridge estimation of the inverse covariance matrix from high-dimensional data. *Computational Statistics and Data Analysis*, 103, 284–303.
- van Wieringen, W. N., Peeters, C. F. W., de Menezes, R. X., & van de Wiel, M. A. (2018). Testing for pathway (in)activation by using Gaussian graphical models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67, 1419–1436.

WESSEL N. VAN WIERINGEN werkt als universitair hoofddocent bij de afdeling Epidemiologie en Data Science van het Public Health research institute van het Amsterdam UMC, en de afdeling Wiskunde van de Vrije Universiteit Amsterdam. E-mail: w.n.van.wieringen@vu.nl

‘Statistiek is zo veelzijdig. Ik snap eigenlijk niet dat er niet veel meer statistici zijn’

WETENSCHAPPELIJK ONDERZOEK KLOPT VAAK NIET

Een wirwar van datapunten, onmogelijke grafieken en lastige kansberekeningen: statistiek kan zelfs de beste wetenschappers intimideren. Velen van hen maken dan ook statistische denkfouten, zegt Casper Albers. Of erger nog: ze gebruiken de nevelen van de statistiek om collega's en het brede publiek om de tuin te leiden.

ANS HEKKENBERG & JIM JANSEN

Hoe groot is een grote kans eigenlijk? Stel dat je op de radio hoort: er is een grote kans dat vanmiddag de zon schijnt. Ga je dan vol vertrouwen naar het strand? Of blijf je thuis, omdat die voorspelling je onzeker in de oren klinkt en je een nat pak wilt vermijden? ‘Een grote kans’ betekent niet voor iedereen hetzelfde, ontdekte de Groningse statisticus Casper Albers met een aantal collega's. Waar sommige mensen uit de bovenstaande zin concluderen dat de kans op een zonnige middag 60 procent is, menen anderen dat de zon 99 procent zeker schijnt. Verschillende interpretaties van kans-woorden zijn natuurlijk niet zo belangrijk wanneer er een middagje strand op het spel staat. Een ander verhaal wordt het wanneer wetenschappers, artsen en juristen communiceren over risico's. Een ‘grote’ kans op kanker, een ‘onwaarschijnlijke’ bijwerking of een ‘vermoedelijke’ dader – bij zulke zaken kan een Babylonische spraakverwarring levensgevaarlijk zijn. En de interpretatie van kanswoorden is maar een

van de vele manieren waarop mensen dagelijks worstelen met statistiek. Ook hebben we moeite met het interpreteren van grafieken en het herkennen van statistische valkuilen. Een voorbeeld: in mei rapporteerde de Belastingdienst zich schuldig te maken aan etnisch profileren. Mensen met twee nationaliteiten werden vaker geselecteerd voor de belastingaangiftecontrole. Het probleem: met zo'n selectiecriterium vind je wat je zoekt – altijd. Als je meer linkshandigen controleert, vind je veel fraude onder linkshandigen, gewoon doordat zij vaker zijn geselecteerd. ‘Een methodologische doodzonde’, tweette Albers. Toch verbazen zulke blunders hem niet. ‘Ook wetenschappers maken fouten in hun statistische methodes.’ En dat is waar zijn werk begint. ‘Als statisticus ben ik de assistent: ik wil de statistiek van andermans werk zo solide mogelijk maken.’ Betere wetenschap én betere voorlichting, daarvoor maakt Albers zich sterk.



Foto: Bob Bronshoff

De onderzoeken waar u aan meewerkt, lopen uiteen van het oplossen van problemen tot het helpen van jongeren met psychische stoornissen. Wat is de rode draad in uw werk?

‘Het puzzelen met getallen. De toepassingen wisselen steeds, maar wat hetzelfde blijft, is dat er een onderzoek wordt gedaan met uitdagende statistiek. Aan dat deel draag ik bij. Bij elk project ga ik in gesprek met een expert die verstand heeft van de inhoud; van het verkeer of de mentale gezondheidszorg bijvoorbeeld. Ik heb zelf geen kaas gegeten van die onderwerpen. Wel kan ik helpen ervoor te zorgen dat het onderzoek goed wordt uitgevoerd. Ik adviseer bijvoorbeeld hoe je het beste meetgegevens verzamelt en hoe je die vervolgens analyseert. Eigenlijk vertel ik onderzoekers steeds hoe ze zo goed mogelijk hun eigen vragen kunnen beantwoorden. Dat maakt mijn werk leuk: statistiek is zo veelzijdig dat ik steeds weer bij een ander vakgebied kan aansluiten en nieuwe dingen

leer. Ik snap eigenlijk niet dat er niet veel meer statistici zijn. Momenteel ben ik betrokken bij onderzoeken die menselijk gedrag in kaart brengen, bijvoorbeeld van jongeren met psychische problemen. Dat is uitdagend, want menselijk gedrag en emoties zijn ongelofelijk moeilijk te meten. Hoe meet je bijvoorbeeld hoe depressief iemand is? Dat is statistisch gezien een veel grotere uitdaging dan bijvoorbeeld verkeerstoepassingen.

Hoe ziet de adviesrol van een statisticus eruit?

In de ideale situatie ben je van begin tot eind betrokken bij een onderzoeksproject; zelfs voordat de data verzameld worden. Dan kun je namelijk meedenken over welke en hoeveel meetgegevens er eigenlijk nodig zijn om een onderzoeksvraag te beantwoorden. Maar ik heb ook wel eens meegemaakt dat collega's al data verzameld hadden en toen naar mij toe kwamen met de vraag hoe ze met die meetgegevens hun onderzoeksvraag konden beantwoor-

den. En dan moest ik zeggen dat dat met die gegevens helemaal niet kon. Daarvoor had je bijvoorbeeld op een andere manier metingen moeten doen, of aanvullende informatie moeten verzamelen. De Britse statisticus Ronald Fisher zei ooit zoiets als: “Een statisticus om hulp vragen wanneer het experiment al is uitgevoerd, is alsof je de dokter binnenroep nadat de patiënt is overleden. Het beste dat je dan nog kunt doen, is achteraf vaststellen waar het is misgegaan.”

U noemde al even het onderzoek naar jongeren met psychische problemen. Wat wilt u over deze mensen weten?
‘Het project waarbij ik betrokken ben, is van hoogleraar psychologie Maaïke Nauta. Zij wil weten of je jongeren die je eerder hebt behandeld voor een psychische stoornis kunt behoeden voor een terugval. Daarvoor moet je achterhalen welke vroege signalen erop duiden dat iemand risico loopt op zo’n terugval. Een slecht slaappatroon bijvoorbeeld, of somberheid. Hoe beter je die signalen kent, hoe eerder je kunt ingrijpen en hoe beter je iemand kunt helpen. Voor dit project heb ik bedacht welke proefpersonen we moesten rekruteren en hoe we metingen moeten doen.

Waarom is het zo lastig om die risicosignalen in kaart te brengen?
‘Omdat je bij mensen veel meer geïnteresseerd bent in het proces dat een individu doorloopt dan in een gemiddelde. Stel, een slecht slaappatroon is een mogelijke risicofactor. Dan wil je niet weten hoeveel uur iemand gemiddeld slaapt. Iemand die elke nacht acht uur slaapt, heeft een heel ander slaappatroon dan iemand die één nacht drie uur en de nacht daarna dertien uur slaapt. Beide personen slapen gemiddeld acht uur, maar die tweede persoon is veel vatbaarder voor een onregelmatig ritme. Dat geldt niet alleen voor slaap, maar ook voor emoties. Als je veel stemmingswisselingen hebt, kun je de dag ’s avonds alsnog waarderen met een gemiddelde zeven. Dat ziet er goed uit, maar sommige delen van je dag waren misschien vreselijk. Aan gemiddelden heb je dus niets. Je wilt eigenlijk live, op elk moment kunnen meten. Daarom werken wij met een app op de telefoons van de jongeren. Die app laat vier keer per dag een piepje horen en stelt dan een aantal vragen. Dat levert hele rijke, nuttige data op. Bij dit project was ik vanaf het begin betrokken. Dan kun je meedenken over elke stap. Hoe



Foto: Bob Bronshoff

CASPER ALBERS (1975) is adjunct-hoogleraar toegepaste statistiek en datavisualisatie aan de Rijksuniversiteit Groningen. Zijn onderzoek richt zich op de ontwikkeling van statistische modellen voor de sociale wetenschappen. Albers maakt zich sterk voor de verbetering van de wetenschap, onder meer als ambassadeur voor open science, de beweging die pleit voor het openlijk delen van data, analyses en onderzoeksartikelen. Daarnaast schrijft hij columns voor *de Volkskrant* en voor de Groningse universiteitskrant *UKrant*.

vaak moet zo’n alarm afgaan? Kunnen we beter vier weken lang vijf keer per dag meten, of vijf weken lang vier keer per dag? In beide gevallen krijg je twintig metingen, maar waar heb je statistisch gezien het meest aan? Dat soort zaken.

Waarom kun je bij dit soort onderzoek niet gewoon de traditionele statistiekmethoden gebruiken?

‘Het feit dat je veel metingen doet gedurende een periode voegt een ingrediënt toe aan de meetgegevens waar de ‘oude’ statistiek niet mee kan omgaan. Normaal gesproken maak je een model: je voorspelt bijvoorbeeld bij studenten hoe goed ze zullen scoren bij een tentamen op basis van metingen van hun inzet en motivatie. Zo’n model werkt prima, maar is nooit perfect. Je maakt altijd wat voorspellingsfouten. Het model houdt er bijvoorbeeld geen rekening mee dat Jim niet kon slapen, of dat Ans plots hard ging leren. In dit geval is dat niet erg, want de voorspellingsfouten zijn bij Jim anders dan bij Ans; ze zijn onafhankelijk van elkaar en stapelen niet op. Maar als je veel metingen na elkaar doet, zoals wij, dan kun je een meetfout die er bij de ene meting insloep opnieuw meenemen in de volgende meting. Om daarvoor te corrigeren, heb je een ander statistisch model nodig.’

Welke verbeteringen voert u dan door in zo’n statistisch model?

‘Eigenlijk wil je altijd weten hoe groot de meetfout is die je meeneemt. Voor hoeveel procent telt die mee in je voorspellingen? De uitdaging is om dat getal zo goed mogelijk in te schatten. Het probleem van veel modellen is dat ze aannemen dat het percentage gelijk blijft. Vaak is dat ook zo. Bijvoorbeeld als je beurskoersen voorspelt. Als de koers gisteren steeg, zal dat vandaag vast ook nog zo zijn. Maar in klinische context, wanneer je patiënten aan het behandelen bent, geldt dat niet. Die patiënten veranderen constant, zeker als hun behandeling werkt. Om daar ruimte voor te maken, moet het getal dat omschrijft voor hoeveel procent je meetfout meetelt mee veranderen. In ons nieuwe model gebeurt dat. Het getal kan daarbij geleidelijk veranderen, maar er is ook ruimte voor wat we een sudden shock noemen. Soms gebeurt er plots iets. Stel, een broer van een patiënt overlijdt. Als zo’n shock plaatsvindt, moet je model dat opmerken. We hebben dat nu getoetst in een simulatie, met succes. Nu gaan we het testen in de praktijk.’

Niet elke sociale wetenschapper heeft een statisticus aan zijn of haar zijde. Gebeurt het vaak dat wetenschappers in statistische valkuilen stappen?

‘Heel vaak. Er gebeurt zoveel bij onderzoeken dat eigenlijk niet mag. Zo zie je bijvoorbeeld weleens dat weten-

schappers een groep proefpersonen onderzoeken en geen interessant verband vinden. Maar als ze een paar ongebruikelijke meetpunten weglaten, dan verschijnt zo’n verband wél. Dan is het verleidelijk om te zeggen: ach, die paar afwijkende metingen, die nemen we niet mee. Maar dat mag natuurlijk niet. Als wetenschapper wil je verder dat de zogeheten p-waarde van je onderzoek onder de 0,05 ligt. (Dat wil zeggen dat het gevonden effect zó groot is dat het erg onwaarschijnlijk is dat je het op basis van toeval gevonden zou hebben – red. *New Scientist*). Dan is je resultaat significant. Wat je nog weleens ziet, is dat een onderzoek nét niet onder die 0,05 uitkomt. Wetenschappers voegen dan soms wat extra proefpersonen toe aan hun test, om die lage p-waarde alsnog te krijgen. Ook dat klopt statistisch niet, want zo raken de resultaten vertekend. Hier ligt een taak voor statistici. Ik geef bijvoorbeeld lezingen voor collega’s om ze erop te wijzen dat je dit soort dingen niet mag doen. Dan zie je mensen weleens schrikken. Zo van: “Oh, maar dat doe ik al jaren!” Die collega’s wilden de boel niet verdraaien; ze wisten gewoon niet dat dit niet oké is.

Gebeurt zoiets ook weleens met opzet?

‘Absoluut! Er is een anoniem onderzoek gedaan onder wetenschappers, waarbij ze vragen kregen als: ‘Hoe vaak heb je data-punten aangewezen als uitschieter en weggelaten om de p-waarde van je onderzoek te verbeteren?’ Wat blijkt: meer dan de helft van de wetenschappers maakt zich weleens schuldig aan dubieuze onderzoekspraktijken. En dat geldt voor alle vakgebieden. Je kunt ook bewust valsspelen met randvoorwaarden. In medische studies wordt bijvoorbeeld vaak de leeftijd van proefpersonen gevraagd. Die proefpersonen deel je vervolgens op in bijvoorbeeld drie groepen: jong, midden en oud. Maar met de grenzen van die groepen kun je sjoemelen. Als je bijvoorbeeld geen significant verschil vindt tussen ‘jong’ en ‘midden’, kun je de grens van jonge mensen verplaatsen van 30 naar 35 jaar. Kijken wat er dan gebeurt. Dat gesjoemel is soms overduidelijk terug te zien in een onderzoeksrapportage. Ligt de grens tussen ‘jong’ en ‘midden’ op bijvoorbeeld op 32 jaar en 2 maanden? Dat is een rode vlag. Niemand kiest zomaar zo’n willekeurige grens.

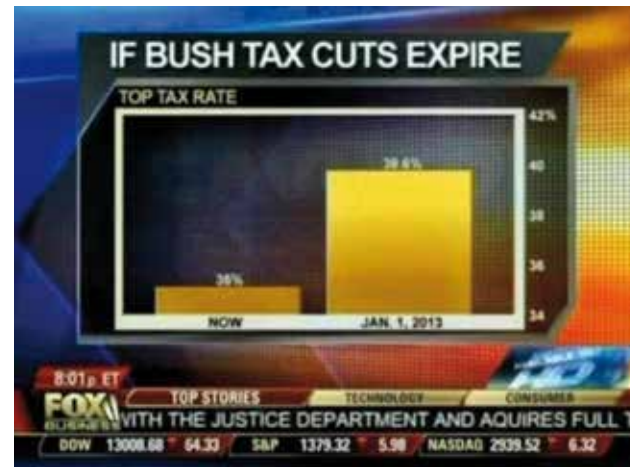
Waarom spelen zoveel wetenschappers vals?

‘Dat is voor een groot deel te wijten aan hoe de weten-

schap werkt. Als wetenschapper doe je onderzoek, schrijf je daar een artikel over en wil je dat vervolgens publiceren in een belangrijk wetenschappelijk blad. Die bladen accepteren eigenlijk alleen artikelen met interessante, tastbare resultaten. Dus móét je er eigenlijk wel voor zorgen dat er dat soort resultaten in je onderzoek zitten. Alleen dan maak je kans op een publicatie. En zo'n publicatie is cruciaal: als je niet publiceert, maak je minder kans op financiering. En als je geen financiering krijgt, wordt je contract bij de universiteit niet verlengd. Je moet eigenlijk wel valsspelen om een baan aan de universiteit te krijgen. Gelukkig zijn steeds meer mensen ervan overtuigd dat dit zo niet langer werkt. Je zou tegen zo'n wetenschappelijk blad moeten kunnen zeggen: als ik deze onderzoeksvraag beantwoord, willen jullie mijn publicatie dan hebben? Zo ja, dan zouden ze op dat moment moeten garanderen dat je werk gepubliceerd wordt – ook als de resultaten niet interessant blijken te zijn. Op die manier vallen de perverse prikkels weg om je onderzoeksdata te masseren.

Hoe groot is de impact van die perverse prikkels?
'Erg groot. Door dit soort trucjes vinden onderzoekers veel vaker significante resultaten dan wanneer het werk eerlijk zou gebeuren. Je ziet dat terug als je de p-waarden van veel onderzoeken op een rijtje zet. Dan blijkt dat er vrijwel geen studies zijn met een p-waarde van net boven de 0,05. Bij die onderzoeken masseren mensen die waarde dus naar net onder de 0,05. We zien dat dit gebeurt bij minimaal eenderde van de onderzoeksartikelen. Als je drie wetenschappelijke artikelen leest, zit er daardoor gemiddeld één tussen waarbij op een verkeerde manier met data is omgegaan. En je weet natuurlijk niet welke.

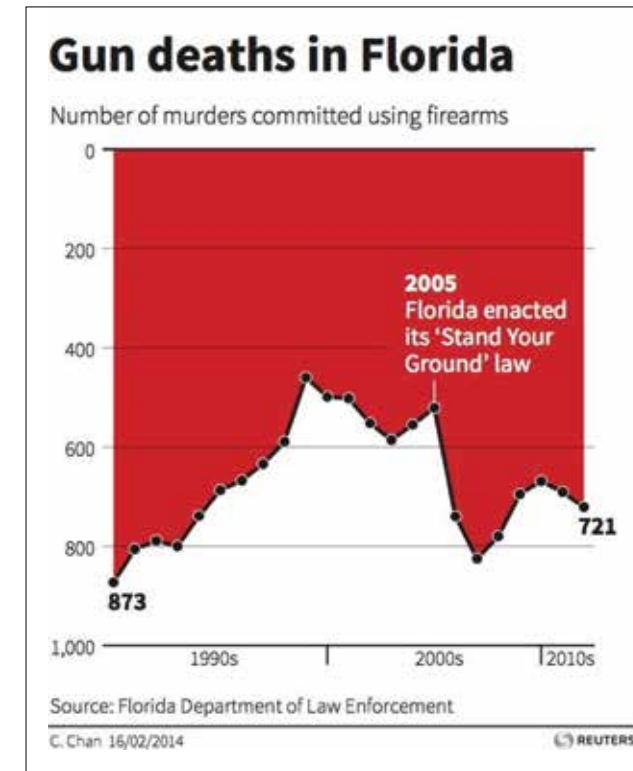
Niet alleen wetenschappers worstelen met statistiek. Welke valkuilen zijn er voor 'de gewone mens'?
'Veel mensen vinden het lastig om data goed te interpreteren. Dat wordt nog veel lastiger wanneer je geconfronteerd wordt met misleidende grafieken. Daar zijn er veel van. Zo deelde Fox News ooit een grafiek hoe de belastingtarieven zouden veranderen zodra maatregelen van Bush werden afgeschaft (figuur 1). De verandering leek enorm: na de afschaffing leek het tarief vijf keer zo hoog als ervoor. Maar Fox had de y-as van een grafiek gemanipuleerd. Die begon niet bij 0, maar bij 34 procent. Het tarief steeg in werkelijkheid maar met een paar procentpunt. Een ander voorbeeld is nog erger: het omkeren



Figuur 1. In de grafiek lijkt het verschil tussen de linker- en rechterstaaf enorm. dat komt vooral doordat de y-as niet bij 0 begint maar bij 34

van de y-as. Dat gebeurde bijvoorbeeld bij een grafiek van Amerikaanse veiligheidsautoriteiten (figuur 2). Die grafiek liet zien hoe het aantal moorden met vuurwapens in Florida veranderde met de tijd. Er was een punt aangegeven waarop er een nieuwe wet was ingevoerd. Het aantal moorden leek daarna te kelderen. Maar de y-as stond verkeerd om: hoe hoger de grafiek, hoe minder doden. Geen wonder: de geïntroduceerde wet was de *stand your ground law*. Die zegt dat als iemand op je terrein komt, je hem eerst mag neerschieten en daarna pas hoeft te vragen wat hij daar deed. Dit waren bewuste pogingen om het publiek te misleiden. Maar niet elke slechte grafiek is misleidend bedoeld. Vaak zie ik grafieken waarvan ik denk, tja, als je in Excel werkt en op de standaardgrafieknop drukt, dan is dit wat je krijgt. Dat zijn niet altijd goede grafieken.

Valt hier iets aan te doen?
Het begint met goed nadenken over hoe je je onderzoeksresultaten wil laten zien. Ik zoek altijd manieren om meetgegevens duidelijk weer te geven. Zo heb ik onlangs geholpen bij een project van een collega die de aardbevingsproblemen in Groningen had gemeten. Het rapport liet zien wanneer er bevingen waren, hoe sterk die waren en waar de schademeldingen vandaan kwamen. We hebben die data samengevat in een filmpje. Daarin zie je op een kaart bevingen verschijnen die gevolgd worden door meldingen. Je ziet soms ook geen beving, maar wel meldingen. Of juist wel een beving en geen meldingen. Dat



Figuur 2. In de grafiek lijkt het aantal doden door vuurwapens af te nemen na 2005. De y-as staat echter verkeerd om; er was dus juist sprake van een toename

filmpje maakt de informatie heel inzichtelijk.

Meer structureel: ik wil een project beginnen om uit te vinden welke mensen er beter in zijn om grafieken te snappen. Licht dat bijvoorbeeld aan het opleidingsniveau? Beeldgevoeligheid? Hoe introvert of extrovert iemand is? Kortom: met welke persoonlijkheidskenmerken hangt 'grafiekgeletterdheid' samen? Als je dat weet, kun je ook beter informatie geven, toegespitst op je doelgroep. Stel bijvoorbeeld dat vrouwen taartdiagrammen beter kunnen lezen en mannen staafdiagrammen. Dan kun je medische flyers voor vrouwen beter op een andere manier ontwerpen dan voor mannen.'

Veel mensen kijken nu meer dan gebruikelijk naar grafieken, bijvoorbeeld over het aantal coronagevallen. Wat adviseert u deze mensen?
'Om niet naar grafieken te kijken. Een grafiek laat misschien zien hoeveel zieken er nu zijn, maar dat vertelt je niets over hoe dat aantal zich gaat ontwikkelen. Je kunt die grafieken beter laten voor wat ze zijn en je laten voor-

lichten door experts. Kijk, als je met een jumbojet wilt reizen en iemand doet de motorkap open, dan kun je daar ook best naar kijken. Maar het heeft geen zin. Als er een technicus naast staat die zegt dat alles in orde is, dan ben ik geneigd om dat te geloven. Vertrouw dus de mensen die er verstand van hebben.

Gebeurt statistiekvervalsing alleen door grafieken? Of geldt het ook voor gesproken en geschreven taal?
'Mensen interpreteren kanswoorden heel verschillend. Woorden zoals 'soms', 'waarschijnlijk' en 'mogelijk'. Neem bijvoorbeeld de zin: "Ik kom soms te laat." In hoeveel gevallen kom je dan te laat? De een denkt aan 5 procent van de gevallen, de ander aan 50 procent. Bij sommige woorden loopt dat enorm uiteen. Bij andere kanswoorden is de interpretatie veel consistent. Als je over kansen spreekt, is het goed om je daarvan bewust te zijn. Ik geef graag zowel het kanswoord als de kwantitatieve informatie. Dus "soms, in 18 procent van de gevallen..." Zo neem je de onzekerheid weg.

Op Twitter bemoeit u zich vaak met maatschappelijke kwesties, van politieke stemmingen tot oneerlijke controles bij de Belastingdienst. Waar komt die drang vandaan?
'Ik denk dat ik een zendingsdrang heb, een missie. Ik hoop mensen te kunnen helpen dingen op waarde te schatten. Om fouten of onjuistheden te zien. Om ingewikkelde concepten te begrijpen. Als ik kritiek lever, is dat op wetenschappelijke gronden. Als ik dan iets zeg over klimaatverandering of Forum voor Democratie, dan gaat het op Twitter los. Sommige mensen beginnen meteen te schelden. Nou ja. Als je geen inhoudelijke argumenten hebt, lig ik daar niet wakker van; dan heb je de discussie al verloren.

Het interview verscheen eerder in *New Scientist* (2020)79 en is met dank overgenomen.

ANS HEKKENBERG is natuur- en sterrenkundige en redacteur bij de Nederlandstalige editie van het populairwetenschappelijk tijdschrift *New Scientist*. Ze spreekt regelmatig als studio-gast op radio en televisie over de laatste ontwikkelingen in de wetenschap. Je vindt Ans als @GirlforScience op Twitter en Instagram.

JIM JANSEN is hoofdredacteur bij *New Scientist* en wetenschapscoördinator bij *Het Parool*. Op zaterdag verzorgt hij de rubriek 'Eureka' voor het *Algemeen Dagblad*.

MENSELIJKE DATAWETENSCHAP

De meningen over menselijke datawetenschap zijn verdeeld in grofweg twee kampen: het ‘alles is fantastisch’-kamp en het ‘alles is waardeloos’-kamp. Als we verder willen komen met data science over mensen moeten we af van beide kampen, en menselijke datawetenschap benaderen als een echte wetenschap. Daarvoor moeten we (nog) meer samenwerken over de grenzen van de disciplines heen.

DANIËL OBERSKI

Op 2 juli 1832 presenteerde de 29-jarige datawetenschapper avant la lettre André-Michel Guerry een kort manuscript voor de prestigieuze Académie Française. Zijn presentatie, gepubliceerd als *Essai sur la statistique morale de la France* (Guerry, 1833), veranderde het leven van de spreker, die een prijs en een felbegeerde plek in de academie verdiende. Tot dat moment was Guerry een bescheiden advocaat uit een provinciestad geweest, met de ogenschijnlijk sombere taak departementale misdaadgegevens te archiveren. Door zijn tijd als ‘datamanager’ was Guerry zich echter iets gaan realiseren: data over mensen lagen voor het oprapen, en daarmee nieuwe kennis over mens en maatschappij. Een mooi voorbeeld toont figuur 1. Dit inzicht stond niet alleen aan de wieg van de sociale wetenschap, maar ook aan die van de moderne visualisatietechniek (Friendly, 2007). Je zou dus kunnen zeggen dat het praatje van Guerry niet alleen zijn eigen leven heeft veranderd.

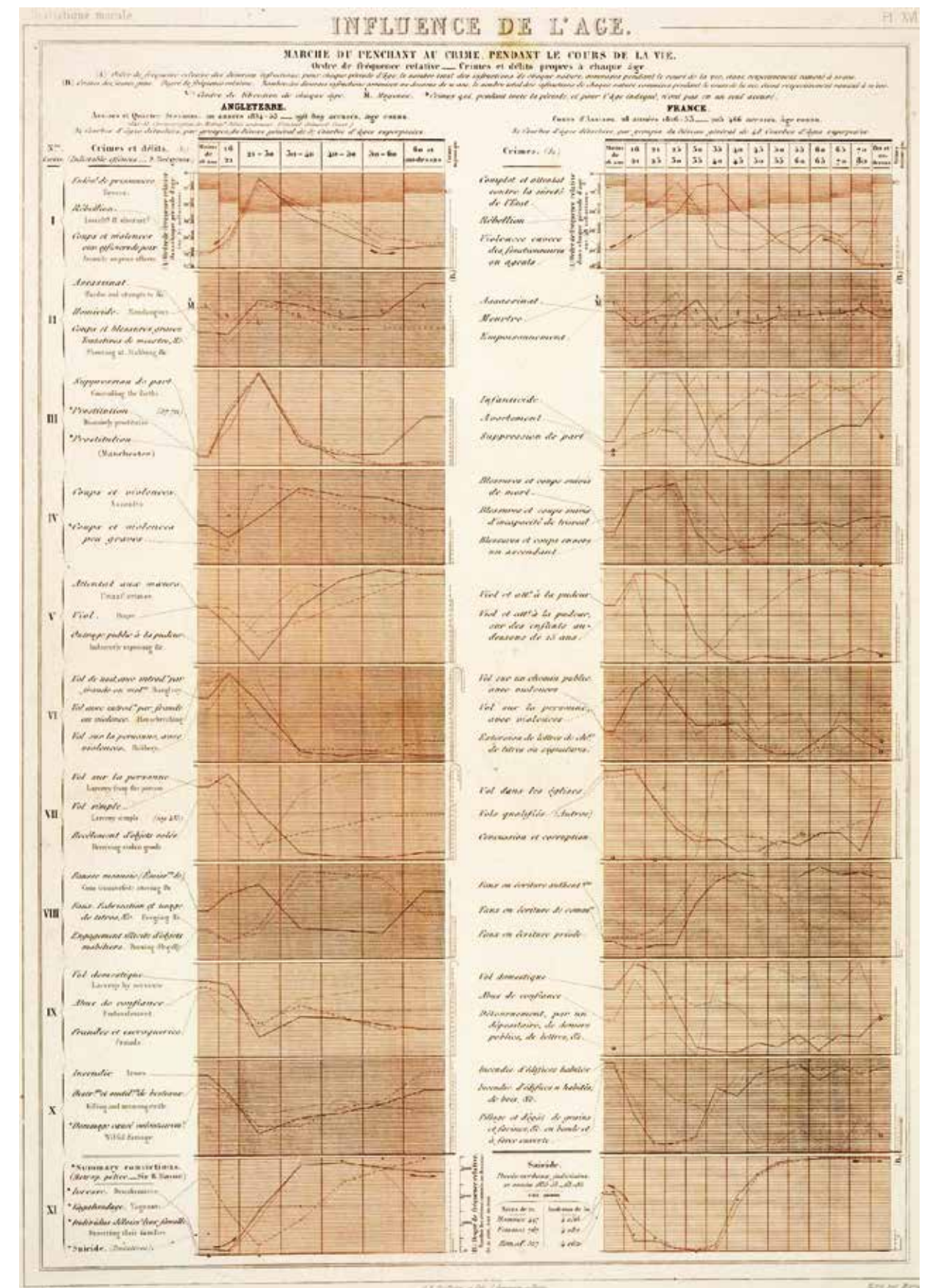
Als sinds de tijd van Guerry en zijn bekendere tijdgenoten Quetelet, Nightingale, en Playfair zijn er twee kampen: zij die datawetenschappers als de apostelen van de menselijke vooruitgang zien, en zij die vinden dat data-analyse over mensen bij voorbaat waardeloos, bedreigend of onwetenschappelijk is. Tegenwoordig staat

de krant bol van de vele gevaren waaraan kwaadaardige data scientists de mensheid onderwerpen – of juist weer van de revoluties in gezondheid en geluk, die, zo lijkt het, steeds op het punt staan bijna te gebeuren. Zelfs Lacroix, de persoon die Guerry in 1832 introduceerde bij de academie, schrijft in het voorwoord van *Essai sur la statistique morale de la France*: ‘Parmi les différen[t]s objets qui sont du ressort de la statistique, un des plus importants et des plus difficiles à traiter, consiste dans l’énumération et le classement des actions humaines.’ Hij was er duidelijk ook niet helemaal uit of menselijke datawetenschap nu dom was of briljant.

Geen van beiden weten we nu. Menselijke datawetenschap moet precies dat worden: een wetenschap. En zoals elke wetenschap zullen we die we moeten benaderen met optimisme waar gerechtvaardigd, en met zorg waar geboden.

Datawetenschap op mensen: dom of slim?

Wat zou je kunnen bereiken als data science met menselijke gegevens goed werkt? Daar zijn talloze voorbeelden van te bedenken, van automatische segmentering



Figuur 1. Guerry's visualisatie van relatieve misdaadcijfers (verticale as) voor verschillende leeftijdsgroepen (horizontale as) voor Engeland (linker kolom) en Frankrijk (rechter kolom). Vergelijkbare misdaden zijn samengevoegd. Collectie: Universitat de Barcelona, Biblioteca Patrimonial Digital. Creative Commons Public Domain Mark 1.0

van radiologische beelden, verwerking van -omics data in bio-informatica, vroegtijdige waarschuwing voor sepsis op neonatale intensive care-afdelingen, tot voorspelling van de huizenmarkt, of modellering van de verspreiding van meningen over infectieziekten én vaccinaties via sociale netwerken. De lezer zal zelf nog wel meer en betere voorbeelden kunnen bedenken.

Een opvallend kenmerk van veel datawetenschappers is een optimistische houding ten opzichte van data. Bij het maken van de prachtige grafiek weergegeven in figuur 1 bijvoorbeeld, realiseerde Guerry (1864) zich vast wel dat Engeland en Frankrijk verschillende definities van misdrijven hanteerden, dat er behoorlijk wat ontbrekende gegevens waren, en dat de Engelse gegevens minder nauwkeurig waren. Let bijvoorbeeld op de verschillen in *binning* tussen de twee kolommen (zie ook Friendly, 2007). Maar gelukkig lijkt geen van deze problemen hem op een mentaal dwaalspoor van onredelijke angsten te hebben gebracht. De oorsprong van deze optimistische houding zou kunnen liggen in het feit dat data science deels geworteld is in de ingenieurswetenschappen (Breiman, 2001). Immers, als MacGyver moest ontsnappen uit een haaienkooi verspilte hij ook geen tijd met filo-

sofische traktaten over de adhesieve eigenschappen van ducttape, en wat ducttape eigenlijk betekent, metafysisch gezien. Het was plakken of zakken. Die *can do attitude* speelt volgens mij een belangrijke rol in de huidige populariteit van menselijke data science.

Figuur 2 toont een benadering van (menselijke) datawetenschap in een geïdealiseerde cyclus van onderzoek en 'productie' (waarbij een 'product' ook een inzicht kan zijn). Denk bijvoorbeeld aan een systeem dat uit gegevens van een app en elektronisch patiëntendossier leert om te voorspellen wie moet worden doorverwezen en wie naar huis kan. Of een systeem dat voorspelt waar politie het beste ingezet kan worden. Of een systeem dat pesten op Instagram detecteert, of dat wetenschappers helpt te begrijpen hoe groepscultuur menselijk handelen beïnvloedt. Er zijn talloze systemen in omloop die deze cyclus doorlopen, met als doel de menselijke conditie te verbeteren.

Daar kunnen natuurlijk ook de ons welbekende problemen bij ontstaan. Sterker nog: bij elke keuze in figuur 2 kan een probleem ontstaan, waaronder:

- De probleemstelling kan slecht geformuleerd zijn;
- De gegevens kunnen niet beschikbaar zijn vanwege pri-

- vacyproblemen, of ze kunnen verkeerd gekozen zijn;
- De gegevens zullen meetfouten bevatten, die de analyse kunnen verstoren;
- De gegevens kunnen onrepresentatief zijn, of het model te idiosyncratisch voor de verkregen data. Beide verstoren generalisaties naar de toepassingssituatie;
- Het adviesalgoritme kan discriminatorisch zijn, of te mysterieus en ondoorzichtig om vertrouwd te worden;
- Het geven van het advies zelf kan de gegevens verstoren, denk aan algoritmische politie-inzet in een straat met hoog inbraakrisico.

Om met een twintigste-eeuwse wijsgeer te spreken: elk nadeel heb zijn voordeel. Daarom moeten we de volgende vragen beantwoorden:

- Hoe groot zijn de voordelen? Bestaan ze wel echt in de praktijk?
- Hoe groot zijn de nadelen? Zijn ze groot genoeg om de positieve effecten weg te nemen?
- Wegen de voordelen (bijv. betere gezondheidszorg) op tegen de nadelen (bijv. gebrek aan transparantie)?

Naar mijn mening is er geen manier om deze vragen in het algemeen te beantwoorden. Soms zullen ze positief uitpakken voor de menselijke datawetenschap, soms negatief. Het zal bijvoorbeeld best waar zijn dat de UK biobank selectiefouten bevat. Maar dat betekent niet automatisch dat elke analyse ervan nutteloos is. Evenzeer is het ongetwijfeld waar dat de kosten die je verzekeraar voor je maakt een *proxy* vormen voor je gezondheid, en voorspeld kunnen worden uit je patiëntendossier. Maar helaas blijkt de meetfout in die proxy het voorspellingmodel te verleiden tot onacceptabele discriminatie (Obermeyer et al., 2019; Boeschoten et al., 2020).

Met andere woorden, zowel het huidige blindstaren op de voordelen als het huidige blindstaren op de nadelen in abstracto is zinloos. Voor elk project zullen we bovenstaande vragen opnieuw moeten beantwoorden. Gelukkig hebben we daarvoor een geweldig instrumentarium, het beste zelfs dat we kennen: de wetenschap.

Een wetenschappelijke benadering

Goede menselijke datawetenschap:

- Erkent dat fouten – in gegevens, modelbeslissingen en implementatie – onvermijdelijk zijn;
- Onderzoekt de mate waarin fouten de output daadwerkelijk beïnvloeden;

- Verwijdert deze effecten waar nodig door de fouten te voorkomen of, wanneer dit niet mogelijk of kosteneffectief is, door de effecten te corrigeren.

In overeenstemming met de technische wortels van data science is goede menselijke datawetenschap ook goede engineering. Een goede ingenieur ploegt niet zomaar voort, ongeacht het terrein. In plaats daarvan ontwerpt ze een oplossing waarbij gebruik wordt gemaakt van de beschikbare materialen en waarbij de risico's worden beheerst.

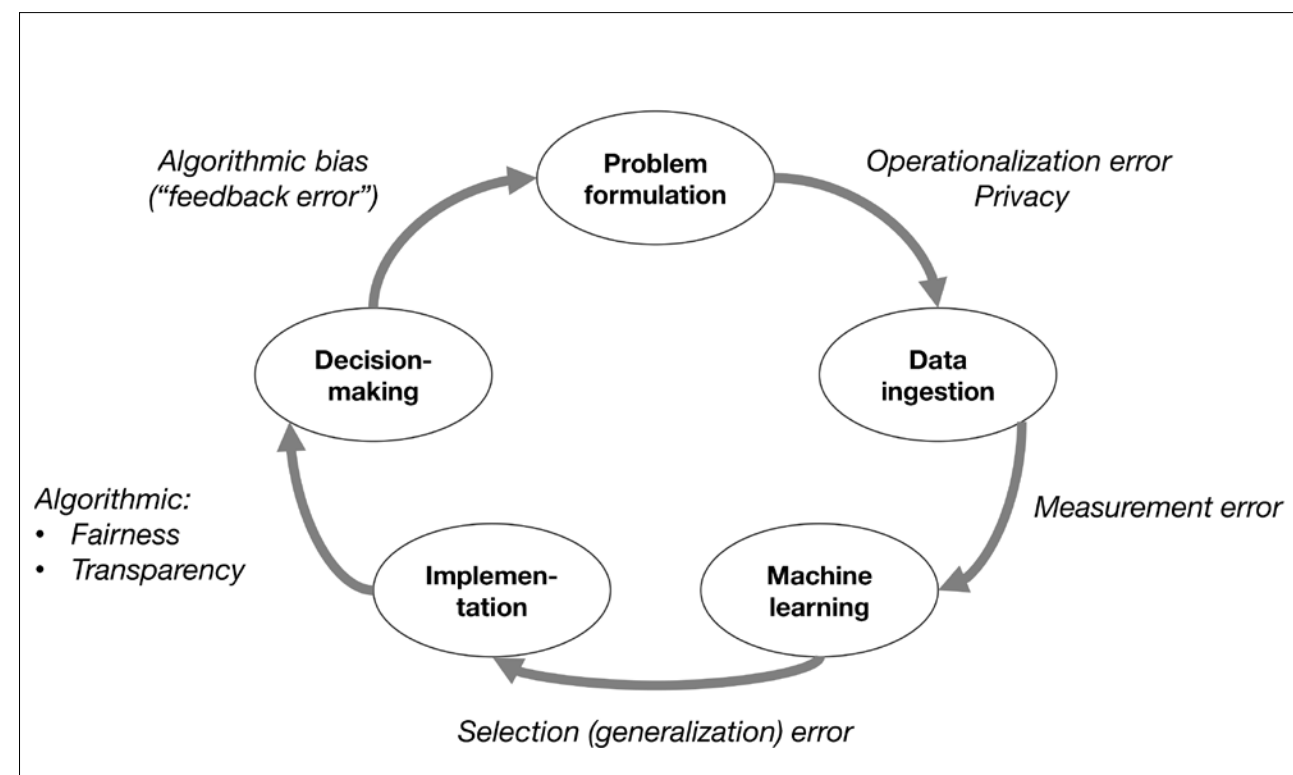
Hier zijn enkele mogelijke dingen die je zou kunnen doen om risico's te beheersen:

- Werk samen met domeinexperts om een nuttig probleem te formuleren en het op gepaste wijze te operationaliseren in een data science-project;
- Ga er niet vanuit dat het effect van meetfouten verwaarloosbaar is; gebruik bijvoorbeeld modellering om deze effecten te corrigeren;
- Gebruik causale modellen en onderzoeksontwerpen om de effecten van selectiefouten en ontbrekende waarden te corrigeren;
- Onderzoek bedreigingen voor de eerlijkheid van besluiten en gebruik bestaande technieken om modelbeoordeling zoveel mogelijk te voorkomen;
- Gebruik geschikte technieken om modelresultaten uit te leggen aan betrokkenen;
- Werk samen met cognitieve wetenschappers om te zien of dat ook echt werkt;
- Onderzoek met sociale wetenschappers en (andere) domeinexperts of en hoe je product werkt in de dagelijkse praktijk.

Hoe werken we aan het bereiken van deze verheven doelen?

Ten eerste zouden menselijke datawetenschappers routinematig een veel bredere kijk op 'modelevaluatie' moeten hebben. Een succesvol data science-project moet niet alleen goed voorspellen in out-of-sample data, maar ook: generaliseren naar andere contexten waarin het zou kunnen worden toegepast, generaliseren naar het bedoelde concept, onzekerheid juist weerspiegelen, beter zijn dan wat er al was, voordelen hebben die opwegen tegen de nadelen, en daadwerkelijk de gestelde doelen in de sociale realiteit bereiken. Om dit alles te verifiëren, zou de **uitgebreide wetenschappelijke studie van data science-projecten een routine moeten worden**.

Ten tweede kunnen **latente variabelenmodellen en causale modellen** een handig raamwerk bieden om over



Figuur 2. Bij elke stap van deze gestileerde menselijke data science-cyclus treden fouten op

een aantal van deze problemen na te denken. Latente variabele modellen zijn bijvoorbeeld handig om meetfouten te schatten en te corrigeren (Oberski et al. 2017), en causale modellering is een goede manier om na te denken over selectiefouten (Mohan and Pearl, 2014). Beide zijn belangrijk bij het onderzoeken van de eerlijkheid van modeloutput (zie Boeschoten et al., 2020).

Ten derde moeten we meer samenwerken. Veel meer. Ik pretendeer absoluut niet de eerste persoon te zijn die erop wijst dat deze kwesties belangrijk zijn. Veel van de bovenstaande punten zijn immers al hele onderzoeksgebieden, die zich uitstrekken over statistiek, informatica, ethische en juridische wetenschap, domeinwetenschappen zoals geneeskunde en biologie, en sociale wetenschappen. Maar naar mijn mening werken we niet genoeg samen en brengen we in de praktijk onvoldoende diversiteit aan perspectieven binnen (en dan heb ik het ook over mezelf!). Bij het doorlopen van de data science-cyclus, ook als statistica of -us, is het namelijk maar al te gemakkelijk om je te concentreren op wat je toevallig interesseert, ten nadele van het project als geheel. Niemand, inclusief ikzelf dus, is immuun voor dit probleem. Sommige onderzoeken naar de ethiek van gegevenswetenschap bijvoorbeeld richten zich uitsluitend op ethische, juridische, of maatschappelijke aspecten, en gaan volledig voorbij aan bestaande wiskundige tools om problemen af te wegen in een technische context. Anderzijds richt ander onderzoek zich volledig op de wiskundige tooling zonder na te gaan of het praktisch nut zal hebben in een sociale context: heb ik recht gedaan aan de intelligentie en autonomie van mensen; zullen mensen echt reageren zoals ik dat veronderstel? Natuurlijk kan niemand alles weten. Dat is dan ook precies waarom we harder zouden moeten werken om **alle velden die aan deze problemen werken te verenigen**.

Wat nu?

Van de statistische heldengalerij Guerry, Nightingale, Quetelet, en Playfair staat alleen Florence Nightingale bekend als iemand die niet alleen een uitvinding deed (het roosdiagram), maar ook daadwerkelijk bijdroeg aan het oplossen van een concreet probleem (sterfte aan het front). De anderen genieten (enige) bekendheid om hun knappe visualisaties van sterfte, criminaliteit, en opleiding, maar niet omdat we denken dat ze die dingen hebben verbeterd. Dat komt, denk ik, omdat hun tijd er

nog niet rijp voor was. Knap als Guerry was kon hij in zijn eentje niet op tegen de obstakels in figuur 2, zoals de schrijver van zijn eigen voorwoord al – zeer beleefd – aanvoelde. Onze tijd is anders. Bij zijn presentatie tegenover de opgetrokken wenkbrauwen van de academie, stond Guerry alleen. In 2020 zijn we wereldwijd letterlijk met duizenden. Het wetenschappelijke werk dat de basisvragen van de datawetenschap kan beantwoorden, kan nu beginnen.

De licentie van dit artikel is Creative Commons Attribution (CC BY 4.0). Dit artikel is een bewerking van het artikel *Human Data Science* (Oberski, 2020)

LITERATUUR

- Boeschoten, L., Kesteren, E.-J. van, Bagheri, A., & Oberski, D. L. (2020). Fair Inference on Error-Prone Outcomes. *arXiv:2003.07621 [stat.ML]*. <http://arxiv.org/abs/2003.07621>.
- Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the Author). *Statistical Science*, 16(3), 199–231. <https://doi.org/10.1214/ss/1009213726>.
- Friendly, M. (2007). A.-M. Guerry's Moral Statistics of France: Challenges for Multivariable Spatial Analysis. *Statistical Science*, 22(3): 368–99. <https://doi.org/10.1214/07-STS241>.
- Guerry, A.-M. (1833). *Essai sur la statistique morale de la France précédé d'un Rapport à L'Académie des Sciences, Par Mm. Lacroix. Crochard*.
- Guerry, A.-M. (1864). *Statistique morale de L'Angleterre comparée avec la statistique morale de la France*. Baillière.
- Mohan, K., & Pearl, J. (2014). Graphical Models for Recovering Probabilistic and Causal Queries from Missing Data. In *Advances in Neural Information Processing Systems*, 1520–8. <http://papers.nips.cc/paper/5575-graphical-models-for-recovering-probabilistic-and-causal-queries-from-missing-data>
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm Used to manage the health of populations. *Science*, 366(6464), 447–53. <https://doi.org/10.1126/science.aax2342>.
- Oberski, D. L. (2020). Human Data Science. *Patterns*, 1(4), 100069. <https://doi.org/10.1016/j.patter.2020.100069>.
- Oberski, D. L., Kirchner, A., Eckman, S., & Kreuter, F. (2017). Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models. *Journal of the American Statistical Association*, 112(520), 1477–89. <https://doi.org/10.1080/01621459.2017.1302338>.

DANIEL OBERSKI heeft een gedeelde aanstelling als universitair hoofddocent data science aan de Universiteit Utrecht, afdeling Methoden & Statistiek, en aan het Universitair Medisch Centrum Utrecht (UMCU), afdeling Biostatistiek. Zijn werk richt zich op latente variabele modellen en data science-toepassingen in de sociale, gedrags- en biomedische wetenschappen. E-mail: daniel.oberski@gmail.com

Vrouwe Justitia (Hildesheim, Duitsland). Foto: Falco via Pixabay



Hoe betrouwbaar zijn loterijen?

In een eerdere column heb ik aandacht besteed aan het loterijschandaal 'Triple Six Fix' in 1980 in Pennsylvania. In de loterij werden tijdens een tv-show in drie containers die elk 10 ballen genummerd van 0 tot en met 9 bevatten de ballen met een luchtblazer door elkaar gehusseld totdat uit elk van de containers een bal in een vacuümbuis was getrokken en in een display was uitgeworpen. De winnende combinatie van drie nummers werd bepaald door de volgorde waarin de ballen de display binnenkwamen. De groep fraudeurs stond onder leiding van Nick Perry die de presentator van de tv-show was. Perry had witte latex verf in elk van de ballen geïnjecteerd behalve in de ballen met de nummers 4 en 6. Dit betekende dat alleen ballen met één van deze twee nummers de vacuümbuis konden bereiken. Vervolgens had de groep heel veel loten gekocht met de acht combinaties 666, 664, 646, 644, 466, 464, 446 en 444. In de bewuste trekking was de winnende combinatie 666, vandaar de naam 'Triple Six Fix'. Het syndicaat van fraudeurs kon niet lang genieten van de ongeveer 1,8 miljoen dollar die ze gewonnen hadden. Ze vielen door de mand door geen voorzichtigheid te betrachten in onderlinge telefoongesprekken. Nick

Perry werd veroordeeld tot een gevangenisstraf van zeven jaar. Op het loterijschandaal van Triple Six Fix is een film gebaseerd, zoals ook binnenkort ook een film zal uitkomen over het politieke schandaal en de hypocrisie rond de Winfall-loterij in de Amerikaanse staat Massachusetts. In deze loterij werd jarenlang door de leiding van de loterij doelbewust toegelaten dat de gewone loterijspeler legaal een oor werd aangenaaid door een kleine groep van investeerders. Het script van de film is gebaseerd op de interessante longread 'Jerry and Marge go Large' van Jason Fagone in de *Huffington Post* van 1 maart 2018.

Het grootste loterijschandaal in de Verenigde Staten kwam in 2017 aan het licht nadat Eddie Tipton, de voormalige beveiligingsdirecteur van een loterij die actief was in meerdere staten, bekende dat hij de uitkomsten van meerdere trekkingen van de loterij had gemanipuleerd. De loterij gebruikte de computer om met een toevalsgenerator de winnende getallen te produceren. Dit ging als volgt: de computer leest een registratie af uit een Geigerteller die straling in de omgevingslucht meet. Deze registratie leidt tot een echt toevalsgetal. Dit getal is het startgetal (seed) voor de Mersenne Twister toevalsgenerator die

door de loterij werd gebruikt. Vervolgens werden door de toevalsgenerator de winnende getallen voor de loterijrekking gegenereerd. Tipton had zichzelf vernietigende rootkit-malware in het computersysteem van zijn werkgever geplaatst om gemanipuleerde winnende getallen te produceren voor trekkingen die op drie specifieke data van het jaar plaatsvonden. Tiptons malware was geprogrammeerd om voorspelbare resultaten te krijgen voor trekkingen op drie data van het jaar – 27 mei, 22 november en 29 december – onder de voorwaarde dat deze data een woensdag of een zaterdag waren en de trekking na 20.00 uur was. Als aan deze criteria was voldaan, werd de toevalsgenerator voor het genereren van de winnende getallen naar een ander spoor omgeleid. In plaats van een seed vanuit de Geiger-teller, gebruikte de toevalsgenerator een vooraf bepaald startgetal waardoor ongeveer tweehonderd voorspelbare combinaties van winnende getallen overbleven. Eddie, die als loterijmedewerker niet aan het loterijspel mocht deelnemen, gaf de potentieel winnende combinaties door aan zijn broer en een zakenman uit Texas, die vervolgens de loterij wonnen en de winst met hem deelde. Zo wist Tipton winnende trekkingen te regelen met in totaal meer dan \$24 miljoen aan prijzen – de grootste loterijzwendel in de Amerikaanse geschiedenis. Hoe kwam deze zwendel aan het licht? Het antwoord is simpel: zoals vele criminelen, kon Tipton het niet opbrengen te stoppen met zijn criminele activiteiten en dan is het bijna onvermijdelijk dat op een bepaald moment een fatale onvoorzichtigheid wordt begaan. Een en ander is kleurrijk beschreven door Reid Forgrave in de longread *'The man who cracked the lottery'* in het *New York Times Magazine* van 3 mei 2018. Eddie Tipton werd veroordeeld tot 25 jaar gevangenisstraf. Zowel dit loterijschandaal als het Triple Six Fix loterijschandaal werpt de vraag op hoe waterdicht loterijen in het algemeen en in Nederland in het bijzonder tegen intern misbruik beveiligd zijn.

De oprichter van de Stichting Loterijverlies pakte het beter aan dan Nick Perry en Eddie Tipton, althans in eerste instantie. Deze stichting werd in 2008 opgericht nadat bekend was geworden dat in de periode tussen 2000 en 2008 de grotere Staatsloterij-prijzen (50 duizend euro en 100 duizend euro) niet getrokken waren uit de verzameling van verkochte loten, maar uit de verzameling van alle loten die verkocht konden worden. De verzameling van alle loten was in de periode tussen 2001 en 2008 opgelopen van 7,5 miljoen loten naar 21 miljoen loten, terwijl per trekking ongeveer 3 miljoen loten werden verkocht. De Stichting Loterijverlies kwam met het voorstel een collectieve procedure tegen de Staatsloterij te voeren. Je kon tegen een eenmalige betaling van 35 euro meedoen met

deze collectieve procedure die gericht was op het verkrijgen van een forse schadevergoeding van de Staatsloterij. Dit was een uitermate slim idee want waarom zou je niet een gokje wagen voor een klein bedrag van 35 euro. De Stichting Loterijverlies kreeg de wind pas goed in de rug toen op 30 januari 2015 de Hoge Raad oordeelde dat in strijd met de wet de Staatsloterij tussen 2000 en 2008 de consument had misleid ten aanzien van de hoogte van de winkansen en het aantal grote prijzen per trekking. Het verweer van de Staatsloterij dat het minuscule kleine verschil tussen de in de reclame-uitingen voorgespiegelde winkansen (0,00000667%) en de werkelijke winkansen (0,000000953%) geen noemenswaardig effect kan hebben gehad op de beslissing van de consument om al dan niet een lot te kopen werd door de Hoge Raad verworpen. Het arrest van het hoogste rechtsorgaan had tot gevolg dat het aantal deelnemers bij Loterijverlies omhoog vloog tot ongeveer 194 duizend personen. De oprichter van Loterijverlies haalde ongeveer 7 miljoen euro binnen. Een gedeelte van dit bedrag ging op aan proceskosten, maar de schadevergoeding die Loterijverlies er voor haar deelnemers eruit sleepte was uiterst beperkt. De Staatsloterij kende een schadevergoeding toe van 40 euro per persoon en gratis deelname aan een extra trekking. De oprichter van de Stichting Loterijverlies leek aanvankelijk er financieel heel wat beter uit te springen. Begin 2016 kwam dagblad *De Telegraaf* echter met een reeks publicaties waarin ernstige twijfels werden uitgesproken of het wel zuivere koffie was wat de oprichter van Loterijverlies met de miljoenen aan inleggeden had gedaan. Dit leidde tot beslaglegging door de FIOD op goederen en bankrekeningen van de oprichter van Loterijverlies. Na nieuwe rechtszaken hierover oordeelde de Hoge Raad uiteindelijk in maart 2020 dat het strafrechtelijk beslag gehandhaafd bleef. De soap is hier nog niet mee ten einde en duurt onverminderd voort met verbitterde gevechten van de Stichting Loterijverlies tegen het dagblad *De Telegraaf* die in het nieuwsbulletin van de stichting als spreekbuis van de staat wordt gekwalificeerd. Daarnaast kent sinds april dit jaar de slepende juridische strijd om schadevergoeding te krijgen van de Staatsloterij wegens misleiding van deelnemers een tweede speler in de gedaante van de Stichting Loterij Incasso. Het wachten is op de verfilming van de oer-Hollandse soap rond de Staatsloterij.

HENK TIJMS is emeritus hoogleraar operations research aan de Vrije Universiteit en auteur van diverse leerboeken over operations research en kansrekening. Zijn meest recente boek is *Basic Probability, What Every Math Student Should Know*. World Scientific Press, 2019.
E-mail: h.c.tijms@xs4all.nl



Foto's: Pieter van den Berg

EEN BOOST VOOR VERDUURZAMING

Elektrificatie van het busvervoer in de regio Rotterdam

In lijn met de klimaatdoelen heeft de Nederlandse overheid in 2016 de ambitie uitgesproken om uiterlijk in 2030 alle dieselmotoren in Nederland te vervangen door duurzaam aangedreven bussen. Sindsdien werken het Rotterdamse vervoerbedrijf RET en de Rotterdam School of Management, Erasmus University samen om deze transitie van dieselmotoren naar elektrische bussen soepel te laten verlopen. Het gaat hierbij voornamelijk om de operationele uitdaging van het plannen van de laadmomenten van de bussen. In 2019 zijn de eerste 50 elektrische bussen in gebruik genomen en in de komende jaren zullen er snel meer volgen.

PIETER VAN DEN BERG

In de strijd tegen klimaatverandering nemen bijna alle landen in de wereld maatregelen om de uitstoot van CO₂ te beperken. Een belangrijke sector in deze energietransitie is de transportsector waarbij het met name van belang is om met fossiele brandstof aangedreven voertuigen te vervangen door duurzame voertuigen. Binnen het open-

baar vervoer gaat het hierbij vooral om het busvervoer dat nog veelal uitgevoerd wordt met dieselmotoren. In 2016 hebben de Nederlandse overheid en de vervoersbedrijven de ambitie uitgesproken om uiterlijk in 2030 alle dieselmotoren vervangen te hebben door duurzaam aangedreven bussen.

Voor de Rotterdamse aanbieder van openbaar vervoer, RET, heeft dit geleid tot een enorme operationele uitdaging. Al snel bleek dat de bestaande materieelplanning niet toegelaten zou zijn als gevolg van de beperkte actieradius en laadduur van de nieuw aan te schaffen elektrische bussen. Dit heeft als gevolg dat een bus niet langer de hele dag aaneengesloten ingezet kan worden, maar gedurende de dag gekoppeld zal moeten worden aan een laadpaal.

Daarnaast heeft de operationele onzekerheid grote invloed op het batterijniveau van de bussen. Zo kan een vertraging ervoor zorgen dat een bus een laadmoment mist waardoor de bus onvoldoende energie heeft om alle ritten uit te voeren. Ook kunnen ongunstige weersomstandigheden zorgen voor een hoger verbruik. Wanneer de energie, die voor het laden gebruikt wordt, ook nog afkomstig is van hernieuwbare bronnen als zon of wind zal dit tot extra onzekerheid leiden.

Om tot een soepele transitie te komen is er een samenwerking op gang gekomen tussen de Rotterdam School of Management, Erasmus University (RSM) en de RET. Het project bestond uit drie fasen. Allereerst was er fase 0 waarin voorbereidingen getroffen werden voor de invoering van elektrische bussen. Vervolgens werd het Noordelijke deel van het busnetwerk geëlektrificeerd in december 2019. Ten slotte is gestart met de planning van het tweede deel van het netwerk dat in de komende jaren geëlektrificeerd moet worden. In elke fase heeft RSM bijgedragen aan het mogelijk maken van de transitie.

Impact van de elektrificatie

In de vroege voorbereidingsfase van het project was RETs voornaamste doel uit te zoeken wat de impact van de elektrificatie zou zijn op de huidige planning. Hierbij werd de huidige planning geëvalueerd met verschillende parameters zoals de capaciteit van de batterij, het energieverbruik en het vermogen van de laadpalen. Ter ondersteuning van deze analyse ontwikkelde RSM een simulatietool die voor elke set met parameters evalueert hoeveel bussen een te laag batterijniveau zouden hebben aan het einde van de dag. Deze analyse liet overduidelijk

zien dat zelfs in het meest optimistische scenario met een grote batterij, laag energieverbruik en hoog laadvermogen meerdere bussen in de problemen zouden komen in de huidige planning. Deze analyse bevestigde dat de RET de materiaalplanning zou moeten aanpassen en rekening moet gaan houden met de specifieke kenmerken van elektrische bussen in de planning.

Optimaliseren van laadstrategieën

Na de voorbereidende fase lag de focus op de eerste set met bussen die geëlektrificeerd zou worden. Hierbij gaat het om 50 bussen die aan het eind van 2019 in gebruik zijn genomen. Onze werkzaamheden hebben plaatsgevonden in de voorbereiding op deze fase. De bussen voeren ritten uit op zeven verschillende lijnen aan de noordkant van de stad Rotterdam, waarbij ze gedurende de dag op zeven verschillende eindstations kunnen laden bij snelladers. Gedurende de nacht worden de bussen in de garage aangesloten op laders met een lager vermogen om weer met een volle batterij aan de volgende dag te kunnen beginnen. RET richtte zich in deze fase voornamelijk op het toewijzen van de verschillende ritten aan de bussen. Hierbij moest er voldoende ruimte tussen ritten gelaten worden om gebruik te maken van de snelladers.

RSM richtte zich op het optimaliseren van laadstrategieën die een balans zouden vinden tussen een robuuste materieelplanning en niet onnodig gebruik maken van snelladers. Het beperkte gebruik van de snelladers is noodzakelijk omdat dit grote impact heeft op de levensduur van de batterij en op het elektriciteitsnetwerk van de stad. Deze afweging werd gemodelleerd door het aantal laadmomenten gedurende de dag te minimaliseren onder de voorwaarde dat elke bus te allen tijde een minimaal batterijniveau zou hebben. Het bleek dat het aantal laadmomenten meer dan gehalveerd kon worden ten opzichte van een naïeve laadstrategie waarbij bussen alle beschikbare laadmomenten zouden gebruiken.

Al snel deed zich de vraag voor wat de impact van vertragingen zou zijn op het batterijniveau van de bussen. Vertragingen kunnen ervoor zorgen dat een bus een laadmoment mist of een lager dan gepland batterijniveau



heeft aan het einde van een rit. Als gevolg hiervan kan de bus later op de dag in de problemen komen. Om dit te evalueren hebben we de simulatie uit fase 0 uitgebreid om ook de situatie met onzekerheid te kunnen analyseren. Uit deze analyse bleek dat zelfs bij een relatief hoog minimaal batterijniveau in de planning er nog steeds meerdere bussen per dag in de problemen zouden kunnen komen als er geen real-time aanpassingen gedaan zouden worden. Als gevolg van de onzekerheid is het vooraf niet duidelijk voor welke bussen maatregelen nodig zijn. Het is daarom nodig om, op basis van de realisaties van de onzekerheid, real-time aanpassingen te doen aan het laadschema om problemen op te lossen. Hiervoor hebben we een real-time laadstrategie ontwikkeld die op basis van de optimale offline laadstrategie en de real-time informatie aangeeft welke bus op welk moment moet laden. Deze strategie blijkt een robuuste oplossing te geven terwijl het aantal laadmomenten laag gehouden kan worden.

Zonnepanelen

In december van 2019 zijn de eerste 50 elektrische bussen in gebruik genomen. Vanaf dat moment is de aandacht van het onderzoek verschoven naar de tweede set met bussen die in de komende jaren in gebruik genomen zal worden. Deze set met bussen zal vooral ingezet worden

aan de zuidkant van Rotterdam. Anders dan bij de eerste set zullen al deze bussen laden op hetzelfde busstation. Door deze clustering van de laadactiviteiten wordt het belangrijk om de interactie tussen de lijnen te beschouwen en is het benodigde aantal laadpalen minder eenvoudig te bepalen. Ook biedt deze netwerkstructuur de mogelijkheid om lokaal hernieuwbare energie op te wekken en deze direct te gebruiken voor het laden van de bussen. Hiervoor wordt overwogen om een zonnepark aan te leggen op het busstation. De onzekerheid in de hoeveelheid opgewekte energie brengt een extra laag complexiteit. De vraag was hierbij wat de impact van dit zonnepark op de totale verbruikte energie zou zijn en of het gebruik van de hernieuwbare energie verhoogd kan worden door een deel van de energie op te slaan in een batterij.

De resultaten van deze analyse laten zien dat, zelfs zonder de mogelijkheid tot opslag, tot wel 70% van de verbruikte energie in de zomer direct van de zonnepanelen kan komen. Voor de winter is dit door de lagere opwekking slechts 30%. In beide gevallen gaat er maar een klein deel van de opgewekte energie verloren. Dat zien we ook wanneer we kijken naar de meerwaarde van een batterij op het busstation waarin een deel van de ongebruikte energie opgeslagen kan worden. We zien dat dit kan leiden tot een toename van de gebruikte hernieuwbare energie van ongeveer 7%. Dit wordt voornamelijk veroorzaakt door het opladen van de batterij in de ochtend voordat de bussen moeten laden en het ontladen

van de batterij aan het eind van de dag als de zon niet meer schijnt.

Interessant genoeg zien we dat de meerwaarde van de batterij het hoogst is in de lente en de herfst en niet in de zomer. Door de hoge hoeveelheid opgewekte energie in de zomer is er maar een beperkt aantal mogelijkheden waarop de energie in de batterij nodig is voor het laden van de bussen. In de lente en de herfst komt dit, door beperktere opwekking vaker voor en heeft de batterij een grotere meerwaarde.

Lessons learned

In dit project hebben RSM en de RET nauw samengewerkt om de implementatie van de elektrische bussen zo soepel mogelijk te laten verlopen. In de verschillende fasen hebben we de nodige zaken geleerd. Zo werd al snel duidelijk dat de impact van het elektrificeren van de bussen van de RET niet onderschat moet worden en dat dit vraagt om een grote verandering aan de materieelinzet. In fase 1 hebben we de impact van vertragingen gezien, waarbij zelfs conservatieve offline laadstrategieën niet kunnen voorkomen dat bussen in de problemen kunnen komen. Het is dan ook cruciaal om real-time bijsturing toe te passen om ervoor te zorgen dat bussen met lage batterijniveaus extra laadmomenten toegewezen krijgen. Ten slotte hebben we gezien dat een netwerk met een hub voor het laden van de bussen zeer geschikt is voor het gebruik van lokaal opgewekte hernieuwbare energie en dat in zo'n situatie energieopslag maar beperkte toegevoegde waarde heeft. In december van 2019 zijn de eerste bussen volgens planning opgenomen in de dienstregeling.

LITERATUUR

- A. Abdelwahed, P.L. van den Berg, T. Brandt, W. Ketter and J. Mulder (2020). *A boost for urban sustainability: optimizing the electric transit bus network in Rotterdam*. Submitted for publication.
- A. Abdelwahed, P.L. van den Berg, T. Brandt, J. Collins, W. Ketter (2020). *Evaluating and Optimizing Opportunity Fast-Charging Schedules in Transit Battery Electric Bus Networks*. To appear in Transportation Science.

PIETER VAN DEN BERG is als universitair docent verbonden aan de Rotterdam School of Management, Erasmus University. Naast zijn onderzoek naar de elektrificatie van busvervoer richt hij zich op de logistiek van on-demand dienstverleners zoals ambulances, taxi's en wegwacht. Voor dit onderzoek ontving hij in 2019 een NWO Veni beurs. E-mail: vandenbergrsm.nl



Als muzikliefhebber kan ik enorm genieten van de composities van gitaristen als Stevie Ray Vaughan, Eric Clapton of David Gilmour maar ook van minder bekende zoals Guthrie Govan, Anne Soldaat of Jeff Beck. Hun composities zijn soms ongelofelijk complex, dan weer simpel, alles met als doel de gitaar een verhaal te laten vertellen, een gevoel over te brengen. Als amateurgitarist doe ik verwoede pogingen hun spel te benaderen, dat lukt slechts ten dele aangezien ik niet over hun vakmanschap, creativiteit en ervaring beschik. Als ik hun composities uitwerk en analyseer komt hun vakmanschap nog duidelijker naar voren. Als je dan weet dat sommige composities in een hele korte tijd geschreven zijn, zoals bijvoorbeeld *Wonderful Tonight* van Eric Clapton¹ stijgt mijn bewondering nog meer.

Modelleren is als Soleren

Naast amateurgitarist ben ik ook Operations Research-consultant. Op het eerste gezicht lijken deze twee activiteiten weinig met elkaar te maken te hebben, ze hebben echter veel met elkaar gemeen. Of je nu een goede gitarist of een goede OR-consultant wil zijn, vakmanschap, creativiteit en ervaring zijn voor beide van groot belang. Het succes van een gitaarsolo of een beslismodel hangt in hoge mate af van het vakmanschap en de creativiteit van de 'artiest'. Modelleren of op een gitaar soleren is een vaardigheid, een vak dat met veel oefenen kan worden geleerd. Leren soleren op een gitaar start met het spelen van eenvoudige patronen waarna steeds meer complexiteit kan worden toegevoegd totdat een het gewenste resultaat wordt bereikt. Het proces om een beslismodel te maken gaat analoog, het is een iteratief proces dat begint

Modelleren, een vak apart

met heel eenvoudige modellen. Het model zal geleidelijk in complexiteit groeien totdat alle relevante details en voorwaarden zijn verwerkt. De stapsgewijze benadering zorgt voor frequente feedback, wat leidt tot een beter begrip van het beslissingsprobleem (of gitaarsolo). Een dergelijke aanpak schept ook vertrouwen in het (model) leerproces en de resultaten.

Niet alles is uit data te leren

Het ontwerpen van een goed beslismodel is een cruciale factor in het succesvol aanpakken van een praktisch vraagstuk. Naar mijn gevoel wordt er in de praktijk echter steeds minder aandacht aan modelleren gegeven. In plaats van goed na te denken, gebruiken we de computer steeds meer om geautomatiseerd een model te maken dat een acceptabele oplossing oplevert. In de huidige praktijk wordt met data en een algoritme een model geschat om bijvoorbeeld het volgende onderhoudsmoment van een machine te voorspellen. Voor een individuele machine kan zo'n model prima werken, het is bovendien relatief eenvoudig en snel te maken. Echter de beslissing om onderhoud uit te voeren kan veel complexer zijn dan alleen het vaststellen van het uiterste onderhoudsmoment. Er moet bijvoorbeeld onderhoudscapaciteit worden gereserveerd, eventueel te vervangen onderdelen moeten op voorraad zijn en de gemiste productie moet worden ingehaald. Mogelijk zijn er ook beperkingen op het aantal machines dat tegelijkertijd in onderhoud mag. Deze aanvullende en machine-overstijgende beperkingen zijn echter niet uit data te leren. Met de juiste ervaring en vakmanschap kunnen deze aanvullende eisen snel duidelijk worden en in het beslismodel worden verwerkt.

Modeleren, een waardevol vak

Hoewel er veel vooruitgang wordt geboekt op het gebied van automatisch uit data leren zal kunstmatige intelligentie voorlopig niet in staat zijn modellen die alle relevante praktische details bevatten voor ons te creëren. Het blijft handwerk waarin vakmanschap, creativiteit en ervaring van groot belang zijn. Dat wil niet zeggen dat het een langdurig proces hoeft te zijn. Ik kom regelmatig in de situatie waarin in een paar uur het benodigde model al klaar is, soms tot verbazing van de klant. Is het probleem dan zo eenvoudig? Dat is het zeker niet, de snelheid is het gevolg van ervaring en kennis, het vraagstuk kan nog steeds onverminderd complex zijn. Het laat zien dat goed kunnen modelleren een krachtig instrument voor een OR-consultant is dat veel waarde heeft, wat overigens niet altijd goed over te brengen is aan een klant die alleen naar de bestede tijd kijkt. Niemand minder dat Pablo Picasso maakte dat duidelijk toen een bewonderaar aan hem vroeg een schets op een servet te maken. Picasso stemde beleefd in, voerde het werk snel uit en gaf het servet terug – maar niet voordat hij om een vrij aanzienlijk bedrag vroeg. De bewonderaar was geschokt: 'Hoe kun je zoveel vragen? Het kostte je een minuut om dit te tekenen!' 'Nee', antwoordde Picasso, 'Het kostte me 40 jaar'.

NOOT

1. Het verhaal gaat dat Eric Clapton dit nummer schreef het terwijl hij op zijn vrouw Pattie Boyd wachtte om samen uit te gaan.

JOHN POPPELAARS, Directeur LITIC B.V.
E-mail: john.poppelaars@litic.com

SUBTYPE-DETECTIE OP BASIS VAN VARIABELENSTRUCTUUR

Een combinatie van clustering en netwerk-methoden

Netwerkmethodologie biedt de unieke mogelijkheid om de innerlijke werking van complexe vraagstukken beter te begrijpen. In de klinische psychologie wordt deze methodologie steeds meer gebruikt om interacties tussen klinische symptomen in kaart te brengen. Echter, verschillende (klinische) subpopulaties lijken verschillende netwerkstructuren te hebben. Met andere woorden, netwerkmethodologie kan worden gebruikt om subtypes te onderscheiden. Voor mijn masterthese ontwikkelde ik een methode die automatisch subtypes identificeert aan de hand van netwerkstructuren. Deze methode combineert netwerkmethodologie met clustering. De methode wordt vervolgens toegepast voor het identificeren van subtypes in de publieke opinie over immigranten en vluchtelingen.

VLADIMIR HAZELEGER

Netwerkmethodologie is in een razend tempo populair aan het worden in de sociale wetenschappen. Met goed recht: netwerken hebben een solide wiskundige basis en bieden een rijkdom aan methodes om data te bestuderen. Met name *psychologische netwerken* (Epskamp, Borsboom & Fried, 2018) hebben in de afgelopen vijf jaar veel toepassingen gezien. Deze ongerichte netwerken hebben als doel klinisch psychologische stoornissen, zoals bijvoorbeeld depressie of posttraumatische stressstoornis, te helpen verklaren, door de complexe relaties tussen de individuele symptomen in kaart te brengen.

Psychologische netwerken

Een voorbeeld is weergegeven in afbeelding 1. De punten in dit netwerk staan voor individuele symptomen, en de

verbindende lijnen staan voor partiële correlaties tussen de symptomen. Deze structuur staat ook wel bekend als een voorwaardelijke onafhankelijkheidsstructuur: twee punten die niet met elkaar verbonden zijn (en dus een partiële correlatie van exact 0 delen) zijn onafhankelijk van elkaar, voorwaardelijk op de rest van de punten in het netwerk. Een dergelijke structuur kan, onder andere, worden gebruikt om symptomen te identificeren die de meeste impact hebben binnen de stoornis, en die dus als doelwit kunnen dienen voor klinische interventies (voor een uitgebreid verslag, zie Epskamp, Borsboom & Fried, 2018).

Dat deze netwerkmethodologie unieke inzichten kan bieden, wordt bevestigd door het vele gebruik van de methodologie (bijvoorbeeld Costantini, Epskamp, & Borsboom, 2015; Fried, Epskamp, Nesse, Tuerlinckx, & Borsboom, 2016; Dalege et al., 2016). De nieuwe metho-

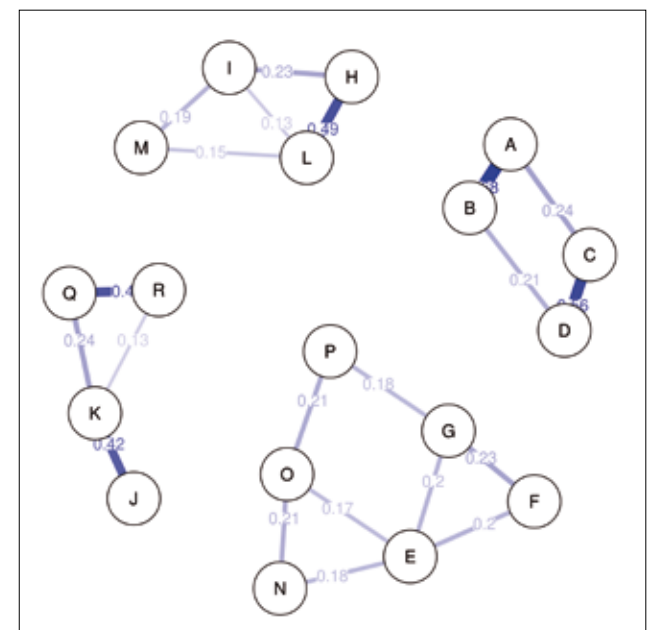
Graffiti in Gent: Foto: Matteo Paganelli via Unsplash



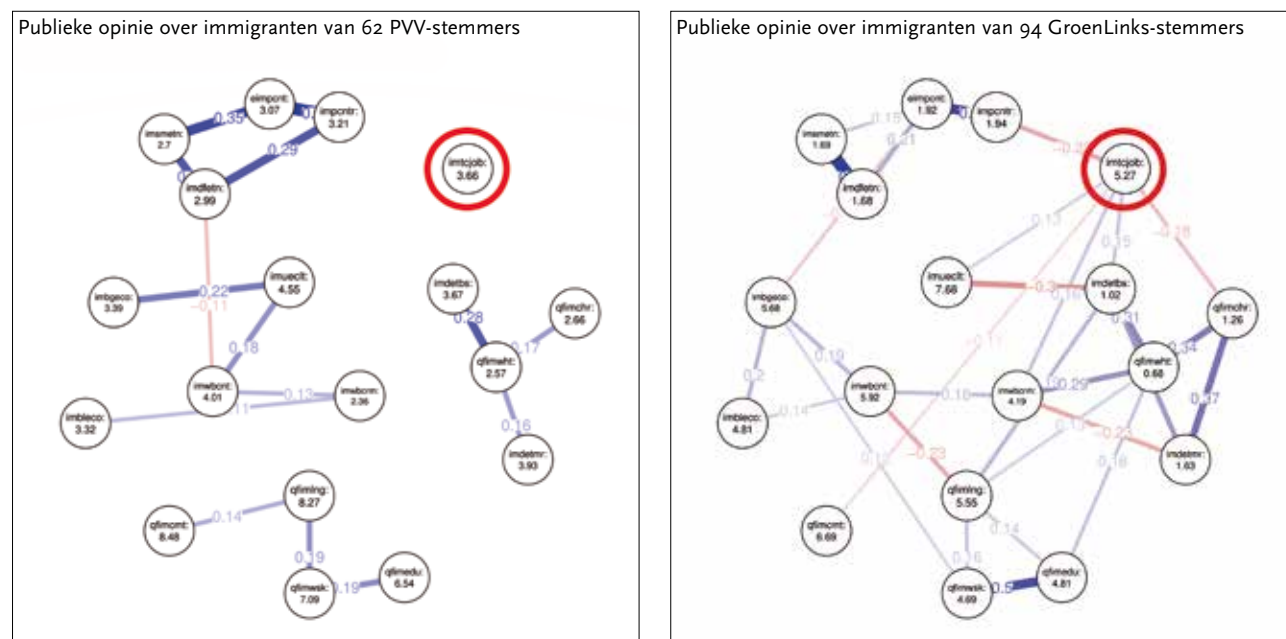
de roept echter ook een aantal vragen op, waaronder de vraag of verschillende (klinische) subpopulaties ook verschillende netwerkstructuren hebben. Deze vraag is dan ook de inspiratie geweest voor het huidige werk.

Netwerken van publieke opinie

Om een antwoord te geven op deze vraag, wendde we ons tot een publieke dataset: de European Social Survey (ESS round 7, 2014). Deze tweejaarlijkse vragenlijst wordt afgenomen in een groot aantal Europese landen, en bevat vragen over allerlei sociale, politieke en economische vraagstukken. Hieronder vallen ook vragen met betrekking tot de publieke opinie over immigranten en vluchtelingen. De publieke opinie over dit onderwerp is sterk verdeeld. Deze verdeling is onder andere te zien in standpunten van politieke partijen. Als voorbeeld gebruiken we de Nederlandse politieke partijen GroenLinks en



Afbeelding 1. Voorbeeld van een psychologisch netwerk; de punten staan voor variabelen, de blauwe (rode) verbindingen staan voor positieve (negatieve) partiële correlaties



Afbeelding 2. Netwerken van de publieke opinie van GroenLinks-stemmers (links) en PVV-stemmers (rechts); variabele namen en gemiddelde scores staan vermeld in de punten, variabele *imctjob* is omcirkeld in rood

PVV, aan de respectievelijk linkse en rechtse uiteinden van het politieke spectrum. Afbeelding 2 geeft de psychologische netwerken weer van de publieke opinie over immigranten en vluchtelingen van GroenLinks-stemmers en PVV-stemmers. De netwerken laten sterk verschillende structuren zien. In het bijzonder valt op de variabele *imctjob*. De vraag bij deze variabele luidt: ‘Bent u van mening dat immigranten banen wegnemen (1) of creëren (10)?’ De netwerkstructuur van de PVV-stemmers laat zien dat de score op deze variabele niet samenhangt met enige andere vraag voor de PVV-stemmers. Het tegendeel is waar voor de GroenLinks-stemmers: voor deze groep hangt de score op deze variabele samen met een groot aantal andere vragen.

Ontdekken van subtypes

Dit simpele voorbeeld leidt tot een tweetal conclusies. Ten eerste lijkt het waarschijnlijk dat verschillende subpopulaties verschillende netwerkstructuren hebben. Ten tweede lijkt het mogelijk om verschillen tussen de subpopulaties te interpreteren aan de hand van de verschillende netwerkstructuren. Gegeven deze aannames, is een vervolgvraag of nieuwe subpopulaties of subtypes kunnen worden ontdekt door het ontdekken van verschillende netwerkstructuren. Dit vereist een combinatie van de netwerkmethodes en bestaande methodes voor het ontdekken van groepen in data, ook wel bekend als clustering. Een dergelijke exploratieve methode biedt de moge-

lijkheid om nieuwe subtypes te ontdekken en te typeren aan de hand van hun unieke netwerkstructuur.

Publieke opinie over immigranten en vluchtelingen

Voor een demonstratie van de voorgestelde methode gebruiken we opnieuw de data van de European Social Survey, met name de items over immigratie en vluchtelingen. De vragen die worden meegenomen zijn terug te vinden in tabel 1. In tegenstelling tot het voorgaande voorbeeld passen we de methode nu toe op de volledige dataset ($N = 31.385$) en zoeken we naar subtypes van de publieke opinie over dit onderwerp.

Methode

Om de methode te introduceren worden hier kort de bestaande methodes beschreven voor het maken van netwerken en voor clustering. Vervolgens wordt beschreven hoe deze methodes gecombineerd kunnen worden om zo het gewenste resultaat te behalen. Voor een meer uitgebreide, wiskundige beschrijving van de methode en gedetailleerde beschrijving van mogelijke toepassingen, zie Hazeleger (2020).

Netwerkmethode

Voor het berekenen van psychologische netwerken wordt gebruik gemaakt van het *Gaussian Graphical Model*

KWALIFICATIE VOOR IMMIGRATIE

qfimedu	Op een schaal van 1 tot 10, hoe belangrijk vindt u het opleidingsniveau van een potentiële immigrant om te kwalificeren voor immigratie in uw land?
qfimcmt	Op een schaal van 1 tot 10, hoe belangrijk vindt u toewijding aan de lokale manier van leven van een potentiële immigrant om te kwalificeren voor immigratie in uw land?
qfimlng	Op een schaal van 1 tot 10, hoe belangrijk vindt u dat een potentiële immigrant de lokale taal beheerst om te kwalificeren voor immigratie in uw land?
qfimwsk	Op een schaal van 1 tot 10, hoe belangrijk vindt u het dat een potentiële immigrant relevante arbeidsvaardigheden heeft om te kwalificeren voor immigratie in uw land?
qfimwht	Op een schaal van 1 tot 10, hoe belangrijk vindt u het dat een potentiële immigrant blank is om te kwalificeren voor immigratie in uw land?
qfimchr	Op een schaal van 1 tot 10, hoe belangrijk vindt u het dat een potentiële immigrant Christelijk is om te kwalificeren voor immigratie in uw land?

TOEGEVOEGDE WAARDE VAN IMMIGRANTEN

imtcjob	Op een schaal van 1 tot 10, denkt u dat immigranten in uw land banen wegnemen (1) of creëren (10)?
imbgeco	Op een schaal van 1 tot 10, denkt u dat immigranten de economie in uw land verslechteren (1) of verbeteren (10)?
imueclt	Op een schaal van 1 tot 10, denkt u dat immigranten de cultuur in uw land ondermijnen (1) of verrijken (10)?
imwbcnt	Op een schaal van 1 tot 10, denkt u dat immigranten uw land slechter (1) of beter (10) maken?
imbleco	Op een schaal van 1 tot 10, denkt u dat immigranten over het algemeen meer geld uit de economie in uw land wegnemen (1) of toevoegen (10)?
imwbcrn	Op een schaal van 1 tot 10, denkt u dat immigranten criminaliteitsproblemen in uw land verergeren (1) of verbeteren (10)?

BEREIDHEID OM TOE TE LATEN

impcntr	Op een schaal van 1 tot 4, hoe bereid bent u om immigranten toe te laten uit arme landen buiten Europa?
eimpcnt	Op een schaal van 1 tot 4, hoe bereid bent u om immigranten toe te laten uit arme landen binnen Europa?
imdfetn	Op een schaal van 1 tot 4, hoe bereid bent u om immigranten toe te laten uit landen met een andere etniciteit dan in uw land?
imsmetn	Op een schaal van 1 tot 4, hoe bereid bent u om immigranten toe te laten uit landen met dezelfde etniciteit als in uw land?

IMMIGRANTEN IN HET EIGEN LEVEN

imdetbs	Op een schaal van 1 tot 10, hoe erg zou u het vinden als een immigrant met een andere etniciteit dan in uw land uw baas zou zijn?
imdetmr	Op een schaal van 1 tot 10, hoe erg zou u het vinden als een immigrant met een andere etniciteit dan in uw land met een familielid zou trouwen?

Tabel 1. Afkortingen en beschrijvingen van de 18 items uit de European Social Survey die zijn gebruikt in het huidige werk

(GGM). Het doel van de GGM is om een accurate schatting te maken van de voorwaardelijke onafhankelijkheidsstructuur van variabelen, gegeven een selecte steekproef. Voor deze netwerken wordt aangenomen dat de structuur schaars is: dat wil zeggen dat er van alle mogelijke verbindingen slechts een kleine hoeveelheid waar is. Deze structuur kan vervolgens worden weergegeven als een netwerk waarin punten staan voor variabelen en verbindingen voor partiële correlaties. De partiële correlaties ρ_{ij} kunnen direct worden berekend uit de zogenaamde precisie-matrix Ω , aan de hand van de volgende formule:

$$\rho_{ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}$$

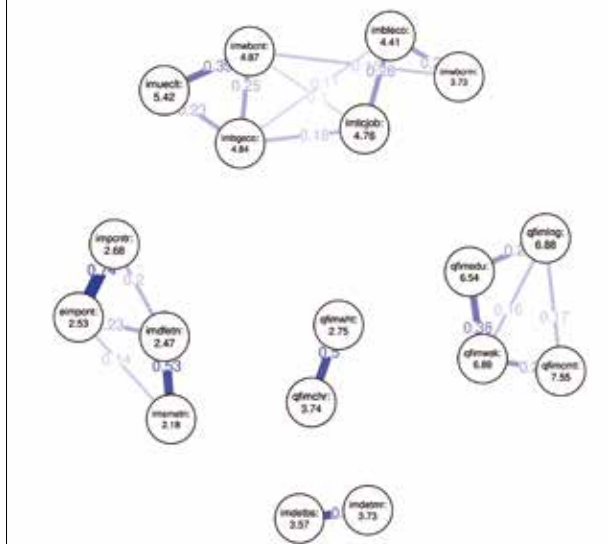
waarbij ω_{ij} staat voor waarde (i,j) van Ω . De precisie-matrix is een inversie van de ware covariantiematrix Σ van de data. Echter, de ware covariantiematrix Σ is nagenoeg nooit bekend voor een steekproef. Met de beschikbare covariantiematrix van de steekproef S kan een schatting worden gemaakt van de precisie-matrix Ω . Een directe inversie van S leidt echter tot een overschatting van Ω , met als gevolg een niet-schaarse structuur. Om overschatting te voorkomen en een schaarse structuur aan te sporen wordt gebruik gemaakt van regularisatie. Specifiek wordt er gebruik gemaakt van LASSO-regularisatie, waarbij de waardes van Ω worden gelimiteerd. Hoe groter de totale complexiteit van Ω , des te meer worden de waardes gekrompen. Dit proces wordt snel en efficiënt uitgevoerd door het *glasso* algoritme (Friedman, Hastie & Tibshirani, 2008), als volgt:

$$\Omega = \text{glasso}(S)$$

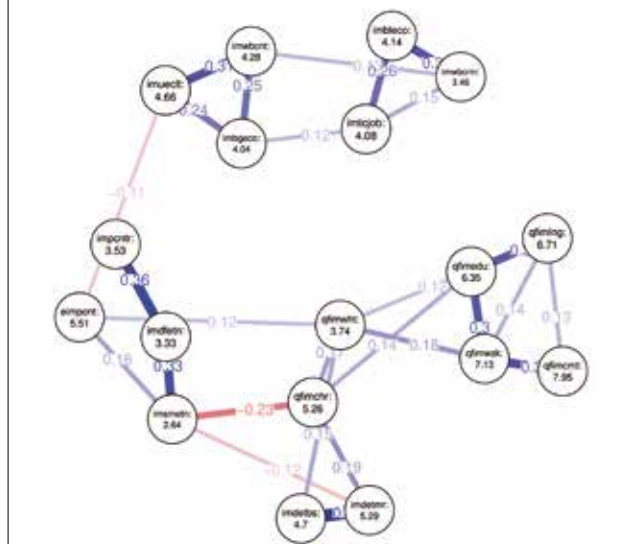
Clusteringmethode

Voor het clusteringproces wordt gebruik gemaakt van *Gaussian Mixture Models* (GMM). De GMM onderscheidt K clusters in de data door voor elk cluster een aparte Gaussiaanse (normaal)verdeling te initialiseren. Deze zogenaamde componenten bestaan uit een vector van gemiddelden μ en een covariantiematrix Σ . Vervolgens worden met behulp van *maximum likelihood estimation* (MLE) de optimale waardes van μ en Σ berekend voor elk cluster. Dit wordt gedaan door het Expectation-Maximisation (EM) algoritme. Hierbij wordt iteratief een verwachte waarde van de likelihood berekend, die vervolgens wordt gemaximaliseerd door de parameters μ en Σ te updaten. Het resultaat is een probabilistische clustering, waarbij voor elk datapunt en voor elk cluster de kans is berekend dat het datapunt bij het respectievelijke cluster hoort.

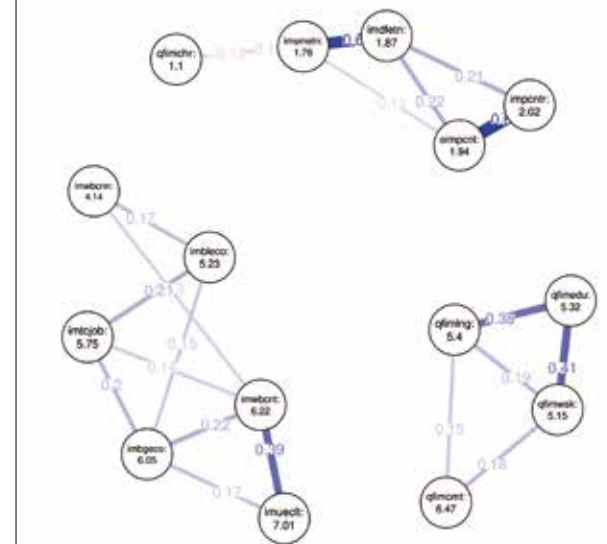
Groep 1 (Onverschillig): gamma = 0,5; threshold = 0,1; N = 19.354



Groep 2 (Tegenstanders): gamma = 0,5; threshold = 0,1; N = 4.608



Groep 3 (Voorstanders): gamma = 0,5; threshold = 0,1; N = 7.423



Abbeelding 3. Netwerken van 3 groepen geïdentificeerd door de gecombineerde methode; van links naar rechts: groep 1 (benoemd als 'Onverschillig'), groep 2 (benoemd als 'Tegenstanders') en groep 3 (benoemd als 'Voorstanders')

Combinatie

Om de methodes te combineren, gebruiken we een gemeenschappelijke parameter: de covariantiematrix Σ . In de GMM wordt een (ware) covariantiematrix geschat voor elk cluster, welke door het EM-algoritme iteratief wordt geüpdatet. De formule voor de standaard-update is de volgende:

$$\Sigma = \frac{1}{N_k} \sum_{i=1}^n \tau_{ik} S$$

Om de GGM netwerkstructuur te verwerken in de GMM, maken we gebruik van het feit dat Σ ook gezien kan worden als Ω^{-1} . Zo vervangen we de standaard-update van Σ met een *glasso* schatting van Ω , om zo de schaarse structuur te verweven in het clusteringproces:

$$\Sigma = \text{glasso}(S)^{-1}$$

Beperkingen

Een aantal beperkingen dient in acht genomen te worden bij het gebruik van deze methode. Allereerst is de methode gevoelig voor steekproefgroottes. Wanneer de grootte per groep kleiner is dan 300, is de schatting van de netwerkstructuur minder accuraat en minder stabiel, met als gevolg meer onzekerheid in de interpretatie. Ten tweede is de methode strikt exploratief. De resultaten duiden op een bepaalde structuur in de data, maar dit blijft een schatting. Vervolgonderzoek – en met name statistische toetsen – zijn nodig om de ontdekte structuur te toetsen. Ten derde mist de huidige methode een manier om het

aantal groepen in het model automatisch te bepalen. In de huidige iteratie is het aan de gebruiker om het aantal groepen in te voeren.

Resultaten

De methode wordt toegepast op European Social Survey data van de publieke opinie over immigranten en vluchtelingen. Abbeelding 3 geeft de resulterende netwerken weer voor het model waarbij 3 subtypes zijn berekend. De netwerkstructuren van de groepen verschillen noemenswaardig. Met name groep 3 heeft een unieke netwerkstructuur, omdat de variabelen *qfmwht*, *eimpnt* en *imdetbs* niet opgenomen zijn in de netwerkstructuur. Dit is een gevolg van het feit dat voor alle 7.423 leden van groep 3, de scores op deze variabelen gelijk waren aan 0, met als gevolg dat deze variabelen geen variantie hebben en dus geen correlaties kunnen hebben met andere variabelen.

Interpretatie

Een intuïtieve manier om de resultaten te interpreteren, is door elke groep een benaming te geven aan de hand van eigenschappen van het bijbehorende netwerk. Voor een uitgebreid verslag van de resultaten, zie Hazeleger (2020). Als voorbeeld bekijken we hier het netwerk van groep 2. Deze groep heeft de hoogste gemiddelden op

items die peilen hoe belangrijk men bepaalde eigenschappen (zoals opleiding, taalbeheersing, of werkvaardigheden) beschouwt bij het bepalen of een potentiële immigrant mag immigreren. Daarnaast heeft deze groep de laagste scores op items die peilen in hoeverre men gelooft dat immigranten een toegevoegde waarde hebben (bijvoorbeeld voor de economie of de arbeidsmarkt). Qua structurele eigenschappen is te zien dat de variabelen *qfmwht* (belang van blank zijn) en *qfmchr* (belang van Christelijk zijn) centraal verbonden zijn in het netwerk. Daarbij is er een sterke negatieve partiële correlatie tussen de variabelen *imsmetr* (hoeveel immigranten met dezelfde etniciteit zou u toelaten?) en *qfmchr* (belang van Christelijk zijn). Dit betekent dat respondenten die met een hoge score antwoorden op het ene item, waarschijnlijk met een lage score antwoorden op het andere item, en vice versa. Met andere woorden: mensen die het belangrijk vinden dat immigranten Christelijk zijn, maken zich minder zorgen om de etniciteit van immigranten, en vice versa. Dit lijkt een aanwijzing dat Christendom voor deze groep een proxy is van etniciteit.

Gezien het bovenstaande kan deze groep benoemd worden als de 'Tegenstanders' van immigratie. Het netwerk beschrijft een categorie Europeanen die sceptisch zijn over immigranten: ze plaatsen hoge eisen voor kwalificatie voor immigratie en zien minder toegevoegde waarde van immigranten. Daarnaast staan etniciteit en Christendom centraal in hun oordeel over immigranten.

Conclusie

De hier beschreven methode biedt een unieke manier voor het identificeren en interpreteren van subtypes in complexe datastructuren. De methode is een exploratieve manier om data te categoriseren en te visualiseren met behulp van de conditionele afhankelijkheidsstructuur van de data. Deze structuur, gevisualiseerd als netwerken, maakt het mogelijk om complexe structuren te beschrijven, te analyseren en te categoriseren via een andere invalshoek. Toegepast op de publieke opinie over immigranten en vluchtelingen, is de methode in staat om drie groepen te identificeren – namelijk Onverschilligen, Voorstanders, en Tegenstanders – die elk gekarakteriseerd worden door een unieke netwerkstructuur (voor een gedetailleerde beschrijving, zie Hazeleger (2020)).

LITERATUUR

- Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Môtus, R., Waldorp, L. J., & Cramer, A. O. J. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality*, 54, 13–29.
- Dalege, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M., & van der Maas, H. L. J. (2016). Toward a formalized account of attitudes: The Causal Attitude Network (CAN) model. *Psychological Review*, 123(1), 2–22.
- Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, 50(1), 195–212.
- ESS round 7. (2014). *European social survey round 7 data*. doi: 10.21338/NSD-ESS7-2014.
- Fried, E. I., Epskamp, S., Nesse, R. M., Tuerlinckx, F., & Borsboom, D. (2016). What are 'good' depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. *Journal of Affective Disorders*, 189, 314–320.
- Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432–441.
- Hazeleger, V. (2020). *Identification and interpretation of heterogeneous sparse conditional independence structures using Gaussian Mixture Modelling and Gaussian Graphical Modeling*. TNO-S11150. TNO

VLADIMIR HAZELEGER is afgestudeerd in cognitieve psychologie en kunstmatige intelligentie (KI). Voor zijn masterthese KI heeft hij, in samenwerking met TNO en de Universiteit Utrecht, bovenstaande methode ontwikkeld.
E-mail: Vladimir.hazeleger@gmail.com

OVER CONFUSIEMATRICES, POOLDIAGRAMMEN EN DE ORDE VAN DE DINGEN

Dit jaar is het 350 jaar geleden dat *Pensées* verscheen, het beroemdste werk van de Franse wiskundige, experimenteel fysicus, uitvinder en filosoof Blaise Pascal (1623 – 1662). De internationale statistische gemeenschap gedenkt in 2020 tevens het feit dat *datascientist* avant la lettre Florence Nightingale (1820 – 1910) 200 jaar geleden in Italië te Florence werd geboren. De toevallige omstandigheid dat beider gedenkjaren samenvallen is uiteraard onvoldoende reden hen gezamenlijk voor het voetlicht te brengen. Ook het feit dat Pascal en Nightingale beide – weliswaar op zeer verschillende wijze – een onwrikbare plaats innemen in de geschiedenis van kansrekening en statistiek, lijkt hiertoe ontoereikend. Toch beoogt dit korte essay een aantal aspecten van beider heterodoxe persoonlijkheden, overtuigingen en werkzaamheden de revue te laten passeren, die zulk een vergelijking billijken.

RICHARD STARMANS

In 1670, precies 350 jaar geleden, verscheen in Parijs de *Pensées*, zonder twijfel het beroemdste werk van de Franse wiskundige, experimenteel fysicus, uitvinder en filosoof Blaise Pascal (1623 – 1662). Zijn befaamde en ontroerende correspondentie met Pierre de Fermat – naar aanleiding van een gokprobleem dat Antoine Gombaud, Chevalier de Meré hem had voorgelegd – is vooral onder kanstheoretici genoegzaam bekend. De briefwisseling wordt door vele historici van kansrekening en statistiek als archimedisch beginpunt van het probabilistische denken in de wetenschap beschouwd. Opmerkelijk genoeg zou ook de postuum gepubliceerde *Pensées*, dat het oogmerk had een apologetisch, theologisch traktaat te vormen, in de annalen van de wetenschapsgeschiedenis worden bijgeschreven. Daarnaast herdenkt de internationale statistische gemeenschap dit jaar het feit dat Florence Nightingale (1820 – 1910) 200 jaar geleden in Italië te Florence werd geboren. Terecht heeft de International Statistical Society in deze het voortouw genomen. De le-

gendarische *Lady with the Lamp* was niet louter grondlegger en boegbeeld van de verpleegkunde, maar ook amateur-wiskundige, *datascientist* avant la lettre, politiek agitator, oorlogscorrespondent en mystica.

De Waagschaal

De *Pensées* werd al spoedig omarmd en bejubeld, maar ook scherp gekritiseerd. Neem bijvoorbeeld de meest klassieke passage uit het boek, de zogenoemde 'Waagschaal' of 'Weddenschap' van Pascal, die in geen enkel filosofisch overzichtswerk ontbreekt. Ruwweg stelt Pascal dat het veel verstandiger is in God te geloven dan niet te geloven. Indien God bestaat is de beloning of winst eindeloos, mocht hij niet bestaan dan is er weinig of zelfs helemaal niets verloren. Ja, misschien wel helemaal niets, omdat men zich levenslang geborgen wist, orde en harmonie in de kosmos zag en de zin van de contingenties



Blaise Pascal, 1691; een kopie naar een schilderij van François II Quesnel (anoniem). Collectie: Paleis van Versailles

van het bestaan onderkende. Geloofte men daarentegen niet en blijkt God toch te bestaan, dan is het verlies of de straf eindeloos; mocht hij alsnog niet bestaan, dan heeft de ongelovige hoogstens een klein voordeel, omdat hij zich geen inspanningen heeft hoeven te getroosten. Gelovigen en ongelovigen van Voltaire, David Hume en Friedrich Nietzsche tot Bertrand Russell en Richard Dawkins hebben de authenticiteit en reikwijdte van het argument gekritiseerd als een opmaat tot kil en calculerend geloven of getypeerd als een gemankeerd utilitarisme en een bedenkelijk crypto-godsbewijs.

Dat laat evenwel onverlet dat de Waagschaal alleen kan en mag worden geduid in de context van de postume publicatie van de *Pensées*, die overvloedig is gedocumenteerd in de geschiedenis van de filosofie. De volledige titel van het boek maakt al veel duidelijk: *Pensées de M. Pascal sur la religion et sur quelques autres sujets, qui ont été trouvées après sa mort parmy ses papiers*. Pascal had zich de laatste jaren van zijn leven steeds meer opgeworpen

als apologet van het Christendom en het plan opgevat een verdediging te schrijven in het licht van de moderne tijd en godsdiensttwisten in de 17e eeuw. Hij richtte zich vooral ook op collega's bij wie hij een opkomend materialisme en atheïsme bespeurde, hetgeen enigszins opmerkelijk is aangezien nagenoeg alle boegbeelden en pioniers van de wetenschappelijke revolutie in de 17e eeuw, waaronder Kepler, Descartes, Galilei, Leibnitz en Newton vrome Christenen waren, alle conflicten en scepsis ten spijt (Cohen, 2011).

Hoe dan ook, de beoogde apologie is er nooit gekomen. Na de dood van Pascal vonden zijn familie en collega's slechts een forse stapel losse aantekeningen; fragmenten van artikelen, flarden van een betoog, aforismen, doorhalingen en onuitgewerkte gedachtenspinsels. Deels waren deze geordend, maar een structuur of betoogtrant was lastig te onderkennen. Nabestaanden, collega's en uitgever legden zichzelf de ambitieuze taak op om van de documenten een min of meer samenhangend geheel te maken, dat recht zou doen aan de oogmerken van Pascal en geschikt was voor publicatie. Deze poging tot een rationele reconstructie was zonder twijfel een tour de force die zijn weerga niet kende. Het zou ongeveer acht jaar duren voordat de redacteuren een eerste proeve van hun werkzaamheden onder de naam *Pensées* deden verschijnen. De vraag waarom Pascal zijn plan niet voltooid blijft hier onbeantwoord (Starmans, 2020a). Het boek vormt door dit alles evenwel een uitnodiging tot comfortabel eclecticisme; men kan er naar hartenlust in grasduinen, diepzinnige citaten vinden, inspiratie opdoen, maar ook in de fragmenten de bronnen vinden voor een sterke narratief vóór of tegen Pascal. Voor- en tegenstanders kunnen eenvoudig hun argumenten in stelling brengen en dit is in de geschiedenis dan ook volop gebeurd. Dat geldt met name voor de Waagschaal, die evenzeer slechts als een losse aantekening in de documenten werd aangetroffen en door iedereen 'geframed' kon worden. Paradoxaal genoeg zou ook dit fragment uit de *Pensées* in de geschiedenis van de statistiek een plaats krijgen.

Confusiematrices

De waardering voor de Waagschaal kwam deels uit onverwachte hoek. Zo zag de wetenschapsfilosoof en historicus van kansrekening en statistiek Ian Hacking in Pascals Waagstuk een vroege heldere proeve van besliskunde of rationale keuzetheorie (Hacking, 1975). Met enige goede wil kan a fortiori worden beweerd dat Pascal als eerste impliciet een zogenaamde *confusiematrix* hanteerde. Ge-

makshalve verstaan we hieronder in deze bijdrage een speciaal soort 2-bij-2 kruistabel of contingentietabel, waarbij de ene dichotome variabele de (representatie van de) werkelijkheid of ‘waarheid’ vormt, terwijl de andere dichotome variabele een *oordeel* of uitspraak over voornoemde werkelijkheid uitdrukt; een *toets*, *meting* of *test*, een *waarneming* of een daarmee verbonden *actie* in diezelfde werkelijkheid. Hoe dan ook, deze ogenschijnlijk triviaal eenvoudige tabel wordt veelvuldig gebruikt in epidemiologie, bij hypothesetesten en foutenanalyse in de statistiek, in *information retrieval*, *pattern recognition* en *machine learning*. Noties als sensitiviteit en specificiteit, *recall* en *precision*, positief en negatief voorspellende waarden, likelihoodratio’s en ROC/AUC-analyses zijn er op gebaseerd en in tijden van data science worden complexe, niet zelden opake en onbegrepen algoritmen, die op basis van veel data classificaties verrichten, met behulp van een eenvoudige confusiematrix geëvalueerd. De matrix en de daarop gebaseerde associatiematen vormen daarmee in zekere zin een probabilistische weergave van de door logici gekoesterde complementaire idealen van *soundness* (‘wat afleidbaar is moet waar zijn’) en *completeness* (‘wat waar is moet afleidbaar zijn’). Ook in het alledaagse leven wordt de matrix echter – impliciet of verhuuld – gebruikt, vooral omdat ook hier stevast twee soorten fouten kunnen worden gemaakt (fout-positief en fout-negatief), die in de praktijk bovendien nooit gelijkwaardig zijn! Ook Pascal onderkende dit in zijn weddenschap. Weliswaar specificeerde Pascal nadrukkelijk geen exacte kans – al wordt daarover soms anders gedacht (Hajek, 2018) – maar gaf hij wel een gewogen nutsfunctie aan: fout-negatief is veel ernstiger dan fout-positief, correct positief is veel belangrijker dan correct-negatief.

Het beantwoorden van de vraag of Pascal daadwerkelijk de eerste was die de matrix hanteerde is een enigszins heikele onderneming, zeker in het licht van Stiglers Wet van de Eponymie, die stelt dat in de exacte wetenschappen elke bijdrage (theorie, wet, stelling, deeltje) vernoemd is naar c.q. toegeschreven wordt aan de verkeerde persoon (Stigler, 1980). Zo was de bekende driehoek van Pascal ook reeds eeuwen eerder bij de Chinezen bekend. Feit is dat Hacking de besliskundige betekenis sterk benadrukt en dat ook menig filosoof de confusiematrix hanteert alsof deze in de nagelaten papieren van Pascal is aangetroffen (Hajek, 2018). Vast staat verder dat Karl Pearson in 1904 als een van de eersten het ruimere begrip *contingency-table* hanteerde in zijn paper *On the Theory of Contingency and Its Relation to Association and Normal Correlation*. Een gezaghebbende en vroege bron waarin expliciet de confusiematrix wordt onderscheiden en – belangrijker nog – in

een algemene theorie over matrices en datarepresentatie wordt ingepast is de onvolprezen studie *Theory of Data* (1964) van de wiskundig psycholoog Clyde Coombs (1912 – 1988). In de geschiedenis van statistiek en data-analyse, waar het primaat dikwijls ligt bij de *inference experts* (Gigerenzer, 1989), hun uiteenlopende paradigma’s en bij succesvolle technieken en algoritmen, blijven de bijdragen van mathematisch psychologen nogal eens onderbelicht. Dat is betreurenswaardig, al was het alleen maar omdat zij dikwijls zoeken naar een intuïtieve correspondentie tussen formele representaties (*scaling*, *testing*, etc) en de wijze waarop de mens informatie verwerkt, uitgaande van de kantiaanse categorieën van het denken en eigenschappen van het menselijk kenvermogen. Dat geldt zeker in een tijd waarin cognitive science en AI steeds meer naar het zenit van de wetenschappelijke belangstelling lijken op te schuiven. Hoe dan ook, in dit boek onderscheidt Coombs onder meer drie soorten matrices op basis van de logische relatie tussen rij- en kolomvariabele:

- symmetrische gelijkenismatrices;
- dominantiematrices;
- conditionele gelijkenismatrices.

Daarnaast onderscheidt hij twee soorten matrices op basis van de inhoudelijke relatie tussen rij- en kolomvariabele:

- intacte matrices;
- buitendiagonale matrices.

Bekende voorbeelden van symmetrische gelijkenismatrices zijn uiteraard correlatiematrices en covariantiematrices. Bij een dominantiematrix duidt elk element een rangorde aan tussen de variabele van de rij en van de kolom. Deze kan bij voorbeeld de data van een halve dan wel een hele competitie weergeven. In het eerste geval is zij per definitie asymmetrisch, in het tweede geval is dat zeer waarschijnlijk. Bij de conditionele gelijkenisrelatie kan men rij-of kolomsgewijs de verhoudingen weergeven. Bij een intacte matrix hebben rijen en kolommen betrekking op dezelfde variabelen, bij een buitendiagonale matrix juist niet. Een kruistabel van twee verschillende dichotome variabelen is dus buitendiagonaal. Duidelijk is dat de confusiematrix een intacte conditionele gelijkenismatrix vormt. Voornoemde maten als sensitiviteit, positief voorspellende waarden, et cetera vooronderstellen zo’n intacte matrix. Bij buitendiagonale matrices zijn andere associatiematen meer geschikt, zoals oddsratio’s (in case-control studies) en relatieve risico’s (in cohortstudies) afhankelijk van het proces of mechanisme volgens welke de data zijn gegenereerd. Ofschoon veel onderscheidingen die Coombs maakt subtiel zijn en daadwerkelijk een ‘theorie van data’ constitueren, is de realiteit



Florence Nightingale te midden van de leerlingen van de Florence Nightingale School of Nursing and Midwifery van het St. Thomas' Hospital in Londen (1886). Foto: Wellcome Library, London

dat sommige zelden meer worden gebruikt. Dat geldt nadrukkelijk niet voor de confusiematrix, die in al zijn bedrieglijke eenvoud de laatste jaren een enorme opmars heeft gemaakt. Niet alleen duikt deze op in nagenoeg elk handboek op het gebied van machine learning, data science of programmeren met R en Python, maar ook in de analytische filosofie als een ogenschijnlijk bescheiden hulpmiddel bij een conceptuele analyse om werkelijkheid/ model enerzijds en data/empirie anderzijds met elkaar te verbinden.

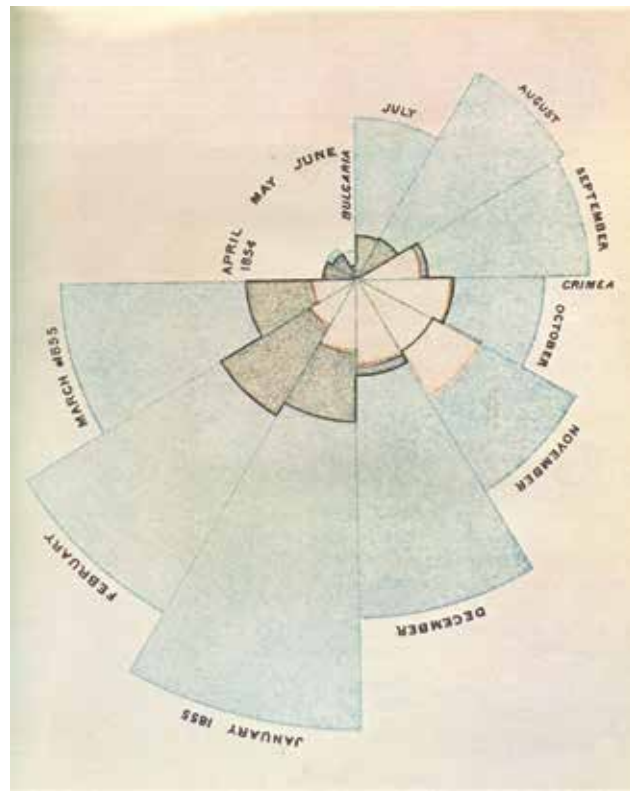
Pool diagrammen en meer

Ook bij het veelzijdige werk van Florence Nightingale moeten we ons beperken. Zij kwam uit een gegoede, enigszins vrijzinnige familie, die niet tot de Anglicaanse kerk behoorde, waardoor Florence in 1820 in het doopregister onder de *dissenters* werd ingeschreven. Hiertoe behoorden onder meer methodisten, puriteinen, unitaristen en anderen die zich niet wensten te conformeren aan bepaalde dogma’s van de staatskerk. Haar grootvader was parlementariër en pleitte vurig voor afschaffing van de slavernij, haar ouders zetten zich in voor sociale hervorming en armoedebestrijding. Desondanks was het ook voor Nightingale als hoogopgeleide vrouw in die tijd moeilijk maatschappelijk op de voorgrond te treden en in haar *Cassandra* uit 1852 richtte zij haar vizier op de verstik-

kende victoriaanse cultuur. Via privélessen en zelfstudie verdiepte zij zich in de wiskunde, maar kreeg al spoedig een roeping de noodlijdenden en zieken in oorlogsgebieden te helpen. Tijdens de oorlog op de Krim (1854 – 1864) deed zij dienst en merkte dat meer soldaten stierven door infecties en medische handelingen en dan door vijandelijke kogels. In haar pogingen de situatie te verbeteren ging zij zeer methodisch te werk. Met veel geduld en niet zonder persoonlijk risico verzamelde zij data. Waar bekende voorlopers als John Graunt en William Petty bij het analyseren van geboorte- en sterftcijfers in de 17e eeuw doopregisters en ambtelijke statistieken konden raadplegen, moest Nightingale haar data min of meer aan het front vergaren. Daarnaast onderhield zij nauwe contacten met vooraanstaande statistici als Adolphe Quetelet (1796 – 1874) en William Farr (1807 – 1883). Enerzijds had Nightingale de tijdgeest mee. De reeds genoemde Ian Hacking sprak in *The Taming of Chance* van een ‘*avalanche of numbers*’ in de 19e eeuw, overal werden statistische bureaus opgericht en trachtte men greep te krijgen op de problemen in de samenleving middels een nieuwe ‘science of data’ (Hacking, 1989). Farr hield zich in Engeland bezig met bevolkingsstatistieken, John Snow zou legendarisch worden met zijn epidemiologische studies naar de cholera-epidemie, Quetelet pleitte voor een sociale fysica, waarbij op basis van veel data naar analogie van de natuurwetenschappen algemene wetten geformuleerd moesten worden.

Anderzijds was het een realiteit dat het vaak lang duurde voordat data en de daarop gebaseerde analyses beschikbaar kwamen en nog langer voordat politici en beleidsmakers er iets mee wilden doen. Nightingale begreep dat een communicatieve inspanning van formaat nodig was. Zij vervulde zowel de rol van oorlogscorrespondent als van statisticus die nieuwe inzichtelijke technieken moest verzinnen om het vaderland te overtuigen, met name de politici. Paradoxaal genoeg bleek dat laatste vaak nog lastiger voor een vrouw dan bijval te krijgen in de door mannen gedomineerde militaire wereld, waar haar status al spoedig aanzienlijk was. In 1858 verscheen haar *Notes affecting the Health, Efficiency and Hospital Administration of the British Army. Founded chiefly on the experience of the late war*. Het boek telde vele diagrammen en illustraties, waaronder de befaamde *polar area diagrams* of *coxcomb diagrams*, waarin bepaalde punten uit de taart ‘uitgeschoven’ leken om het belang van een bepaald patroon te kunnen benadrukken. Zelf sprak zij van roosdiagrammen. Haar rapportages en visualisaties leidden ertoe dat in Engeland ook de politiek niet langer om haar heen kon. In zijn uitvoerige *The Golden Age of Statistical Graphics* uit 2009 schetst Michael Friendly welke vooruitgang op het gebied van datavisualisatie in de 19e eeuw werd gemaakt en hoe de bijdragen van Nightingale hierin moeten worden geduid. De pooldiagrammen werden weliswaar niet door haar uitgevonden, maar de wijze waarop Nightingale deze uitwerkte en toepaste heeft de datavisualisatie en de beschrijvende statistiek daadwerkelijk voortgestuwd.

Ondanks de weerstand die Nightingale ondervond was haar lot toch beduidend minder tragisch dan dat van de door haar beschreven Trojaanse prinses, die haar stadgenoten tevergeefs waarschuwde en uiteindelijk door de Grieken als slavin gevangen werd genomen. Nightingale ontving al vroeg veel waardering. Zij kwam als een held terug van de Krimoorlog, werd op haar achtendertigste Fellow van de Royal Statistical Society, later erelid van de American Statistical Association en kon samenwerken met vooraanstaande tijdgenoten. Zij effende de weg voor datavisualisatie en ook voor exploratieve data analysis, juist in tijden van het zware ‘geweld’ van de opkomende inferentiële statistiek. Tot op de dag van vandaag worden de pooldiagrammen met haar geassocieerd en zij geldt terecht als een van de *forerunners* van exploratieve data-analyse en met name datavisualisatie. Na de Krimoorlog plaveide zij mede de weg voor de epidemiologie en voor het latere *evidence-based medicine*; zij voerde als een van de eersten case-control studies uit, stuurde voortdurend surveys naar de ziekenhuizen, rationaliseerde de



Een voorbeeld van een pooldiagram (*coxcomb diagram*) waarmee Florence Nightingale de verschillende oorzaken voor sterfgevallen tijdens de Krimoorlog (1857) in kaart bracht

administratie van patiëntgegevens, waardoor prevalentie en incidentie beter zichtbaar werden, voerde nieuwe hygiënemaatregelen in, et cetera.

Twee heterodoxen?

Zowel Pascal als Nightingale waren in menig opzicht heterodoxe persoonlijkheden, in de zin dat zij niet tot de mainstream behoorden, veeleer einzelgänger vormden, die wars waren van orthodoxie en de polemiek niet schuwden. Dit verklaarde niet alleen de weerstand die zij bij leven ondervonden, maar ook de soms grillige historische receptie en waardering, die voortduurt tot op de dag van vandaag. We beperken ons tot enkele facetten. Typerend is allereerst hun beider wilskracht en onverzettelijkheid. Anders dan Pascal bereikte Nightingale een zeer hoge leeftijd, maar beide waren fysiek zeer kwetsbaar en veel historische bronnen getuigen van hun zwakke gestel, fysieke ongemakken en chronische aandoeningen. Desondanks was Pascal allerm minst een *armchair philosopher*. Hij trad naar buiten en richtte zijn pijlen op vele tegenstanders. De Jezuïeten verweet hij opportunisme in religieuze zaken, de vele speculatieve deeltjesdenkers een overmaat aan fantasie, de empiristen een onvolkomen filosofie, de metafysici een rationalistische dwaalleer. Van een over-

maat aan ontzag voor kerkelijke of wereldlijke autoriteiten was nooit sprake. Sommige van zijn polemische geschriften worden nu tot de toppen van de wereldliteratuur gerekend, zoals de *Lettres Provinciales* uit 1657.

Nightingale's onverzettelijkheid en kordaatheid is genoegzaam bekend en zij manifesteerde zich regelmatig als politiek agitator als dat nodig was om iets gedaan te krijgen. De statisticus David Spiegelhalter typeerde haar nog onlangs als '*a one-woman pressure group and think tank*' en speculeerde in een BBC-documentaire hoe zij in tijden van big data en pandemieën zou optreden. Pascal en Nightingale wisten beide al op jonge leeftijd bekendheid en status te verwerven, hadden invloedrijke vrienden, waardoor het ook mogelijk werd hun stem met kracht te laten horen.

Ook over het geloof van Pascal en Nightingale is veel geschreven. Het vormde voor beiden een drijvende kracht, maar ook in religieus opzicht waren beiden weinig orthodox. Pascal was altijd al religieus, maar geraakte steeds meer onder invloed van het Jansenisme, dat door de Jezuïeten als een ketterse stroming werd beschouwd. Cruciaal was de zogenaamde tweede bekering van Pascal, die plaatsvond op 23 November 1654, toen hij naar eigen zeggen een diepe religieuze ervaring beleefde, een visioen, een bijkans mystieke sensatie. De belangrijkste elementen ervan noteerde hij summier op een stuk perkament, dat hij in de zoom van zijn jas naaide en altijd bij zich zou dragen. Dit zogenaamde *Memoriaal* bevat de beroemde passage waarin Pascal zich richt tot de God van Abraham, Isaac en Jacob en niet (langer) tot de God van de filosofen en geleerden, een tegenstelling die ook in de *Pensées* een rol zal spelen. Uiteraard is dit slechts enkele regels tellende *Memoriaal* voer voor exegeten en psychologen gebleken. Sommigen zagen hierin een afscheid van de wetenschap, die als ijdel en onnodig divertissement terzijde werd geschoven ten faveure van een alomvattende toewijding aan God.

Ook Nightingale was zoals reeds aangegeven van huis uit weinig orthodox en had evenals Pascal op jonge leeftijd een mystieke religieuze ervaring. Wandelend in een tuin te Embley op 7 February 1837 hoorde zij naar eigen zeggen de stem van God en vanaf dat moment wist zij welk pad zij diende te volgen. Veel meer dan Pascal verbond zij haar werk met haar geloof en beschouwde de statistiek als een instrument om God te dienen. Een van haar bekendste uitspraken wordt geciteerd in het inmiddels klassieke *Games, Gods & Gambling: A History of Probability and Statistical Ideas* uit 1962, geschreven door Florence Nightingale David (1909–1993), zelf een vooraanstaande statisticus, die zowel met vader (Karl) als met zoon (Egon) Pearson zou

samenwerken en door haar ouders naar Nightingale werd vernoemd. '*The true foundation of theology is to ascertain the character of God. It is by the aid of Statistics that law in the social sphere can be ascertained and codified, and certain aspects of the character of God thereby revealed. The study of statistics is therefore a religious service.*'

Tot slot moet worden opgemerkt dat zowel Pascal als Nightingale volop werd en wordt geframed, waardoor overdreven bewondering en onterechte en eenzijdige kritiek dicht bij elkaar liggen. Zoals reeds betoogd lag bij Pascal aan deze framing de onvoltooide en in twee betekenissen 'nagelaten' belijdenis van de *Pensées* ten grondslag. Bij het curieuze *Memoriaal*, dat evenmin was gepubliceerd, is dit nog nadrukkelijker het geval. De manier waarop en de stelligheid waarmee ook serieuze historici en vooraanstaande denkers als Bertrand Russell hem op basis van een fragment van slechts enkele regels frameerden is ronduit stuitend.

Nightingale had met een andere framing van doen. Al spoedig ontstond een mythe over de Lady with the Lamp. Hagiografieën en epische verbeelding leiden evenwel – soms van de weeromstuit – tot venijnige antithesen. Opmerkelijk genoeg is juist de afgelopen twintig jaar in diverse studies getracht haar troon aan het wankelen te brengen. Nightingale werd neergezet als een narcistische en harde vrouw, wier belang voor de statistiek schromelijk wordt overschat en die door haar wijze van optreden meer schade heeft aangericht dan levens gered. Een van de bekendste pogingen de oude mythe te doorbreken is wellicht Hugh Small's *Avenging Angel* uit 1998. De laatste jaren is door een soort Hegeliaanse drieslag ook hierop weer een correctie gekomen, niet in de laatste plaats door Hugh Small zelf met zijn *The passion of Florence Nightingale* uit 2010. Wie een meer genuanceerde kijk op Nightingale prefereert kan terecht bij diverse recente studies van Lynn McDonald, die ook verantwoordelijk is voor de publicatie van de verzamelde werken van Nightingale, welke tussen 2001 en 2012 in zestien delen verschenen (McDonald, 2001).

De orde van de dingen

De ideeëngeschiedenis kent veel voorbeelden van denkers met een sterk idealisme, een hooggestemd moreel utopisme, religieuze of metafysische bevologenheid met een daarbij passend verheven kennisideaal. De queeste naar een diepere, achterliggende werkelijkheid, het wezen van de dingen vormt dan een opmaat tot speculatieve filosofie en grootse metafysische systemen; '*Was die Welt im Innersten zusammenhält*' (Goethe, *Faust*), *The Great Chain*

of *Being* (Lovejoy) en beter nog, het romantische ideaal van holistische alomvattende kennis. Of bescheidener en nuchterder, een streven naar unificatie, eenheidswetenschap of een theorie van (bijna) alles. Bij Pascal was hiervan allemaal geen sprake. Waar bij voorbeeld nagenoeg alle voornoemde pioniers van de wetenschappelijke revolutie in een Pythagoreïsch universum leefden, waarbij de wiskunde een metafysische en sacrale betekenis kreeg, zag Pascal een veel bescheidener rol voor de wiskunde, die hij zeker niet van toepassing achtte op de mens, zijn geest en begrip van zijn plaats in de kosmos. Anders dan roemruchte tijdgenoten verloor de auteur zich evenmin niet in metafysische abstracties (Kepler, Leibniz), speculatie en leerstellige deeltjestheorieën (Descartes, Hobbes), rationele godsbewijzen (Arbuthnot, Süssmilch) of gepretendeerde zekerheden voortkomend uit een axiomatisch-deductief kennisideaal. Als probabilist en uitvinder was hij probleemoplosser en in dit opzicht is er duidelijk verwantschap met Christiaan Huygens. Als experimenteel fysicus deed hij experimenten met belangwekkende theoretische resultaten, maar zonder hieraan verstrekkender implicaties te verbinden. Zijn sceptische houding ten aanzien van de kenbaarheid van de natuur en de betekenis van natuurwetenschap voor de mens leidde niet tot relativisme, maar vormde een vruchtbaar methodologisch principe, dat pas in de 19e eeuw met de probabilistische revolutie de wetenschappelijke praktijk zou veranderen. Evenmin was hij een aanhanger van rationele theologie, waarbij men tracht met rationele middelen het geloof te funderen, godsbewijzen te leveren of in ieder geval de hand van een Schepper in natuur en kosmos te onderkennen. De opkomende statistiek zou overigens al spoedig volop hiervoor worden ingezet (Starmans, 2011). John Arbuthnot, Patrick Süssmilch en Richard Price zijn hiervan slechts enkele voorbeelden. Pascal wilde geen godsbewijzen en alle pogingen de Waagschaal als zodanig te bestempelen zijn verkrampd en tot mislukken gedoemd. Dat de *Pensées* nog steeds wordt gelezen ligt aan literaire betekenis en aan het feit dat het boek voor sommigen een bron van inspiratie vormt en soms zelfs een persoonlijke leidraad op het levenspad (Starmans, 2020a). Zo schreef de Amerikaanse biostatisticus Peter Gilbert enkele jaren geleden een persoonlijk geïnspireerde leeswijzer bij de *Pensées* (Gilbert, 2012).

Gelet op de hierboven geciteerde verheven uitspraken lijkt Florence Nightingale beter te passen bij de rationele theologie. Dat is echter slechts ten dele het geval. Ook zij kenmerkte zich door een principieel empirische opstelling en had een afkeer van ijdele kennisclaims. Haar zoektocht naar een hogere orde van de dingen stond een

down-to-earth-benadering niet in de weg, zij was allereerst pragmatist. Het verzamelen, ordenen, analyseren en presenteren van data om een klein stukje van de wereld te verbeteren was Nightingale in dit opzicht voldoende. Hedendaagse wetenschapsfilosofen benadrukken dat vooruitgang in de wetenschap niet primair revolutionair, maar evolutionair is, dat het niet zozeer gaat om grootse nieuwe paradigma's en geruchtmakende theorieën, maar om stapsgewijze verfijning, niet zelden door beter tellen, meten en wegen, visualiseren en presenteren. Confusie-matrices, pooldiagrammen en talloze andere kleine bijdragen moeten dan ook in dit licht worden gezien. Het is één van de charmes van de geschiedenis van kansrekening en statistiek, dat deze ook aan de bedenkers van dergelijke bijdragen, soms als 'stille sterren' maar toch zeker op prominente wijze en in weerwil van Stiglers Wet aandacht besteedt.

LITERATUUR

- Cohen, F. (2010). *How Modern Science Came Into the World. Four Civilizations, One 17th Century Breakthrough*. Amsterdam University Press.
- Coombs, C. H. (1964). *Theory of data*. Wiley.
- McDonald, L. (2001). *The collected works of Florence Nightingale*. Wilfrid Laurier University Press.
- Gigerenzer, G. (1989). *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge University Press.
- Gilbert, P. B. (2011). *Pascal's God-Shaped Vacuum; a guided tour of the Pensées*.
- Hacking, I. (1975). *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability Induction and Statistical Inference*. Cambridge University Press.
- Hacking, I. (1975). *The Taming of Chance*. Cambridge: Cambridge University Press.
- Hájek, A. (2018). *Pascal's Wager* In: The Stanford Encyclopedia of Philosophy (Summer 2018 Edition), Edward N. Zalta (ed.).
- Starmans, R. J. C. M. (2011). *Pascal, Hume en de reikwijdte van probabilistische argumenten*. *Filosofie*, 21 (4).
- Starmans, R. J. C. M. (2020a). *De tragiek van een heterodoxe denker; Pascals nagelaten belijdenis*, In *Filosofie-Tijdschrift*, 30 (4).
- Starmans, R.J.C.M. (2020b) *Prometheus unbound or Paradise regained - the concept of causality in the contemporary AI-data science debate*. In: *Journal of the French Statistical Society, Special Issue on Causality*, Antoine Chambaz (ed), 2020.
- Stigler, S. (1980) *Stigler's law of eponymy*. *Transactions of the New York Academy of Sciences*. 39: 147–58 (Merton Festschrift Volume, edited by T. Gieryn)

RICHARD STARMANS is verbonden aan de Faculteit Bèta-wetenschappen (Department of Information and Computing Sciences) van de Universiteit Utrecht en aan Tilburg University. Hij doet onderzoek op het snijvlak van filosofie, statistiek en informatica.
E-mail: starmans@cs.uu.nl

Kaas en een open koe



In de hoogtijdagen van de Sectie Statistische Programmatuur van de VVSOR waren binnen die sectie diverse werkgroepen actief. Zo was er de groep 'Opzet van statistische pakketten in het algemeen' die enkele malen per jaar bij elkaar kwam. De bijeenkomsten werden gehouden bij het bedrijf of instituut van een der leden. Ze begonnen met een algemene introductie over die werkgroep en over hoe daar de statistiek werd toegepast.

Ik was een actief lid van die werkgroep en heb deelgenomen aan veel van die bijeenkomsten. Ik werkte bij het Rekencentrum van de toenmalige Landbouwuniversiteit en was ook een keer gastheer. De bijeenkomst was een mooie gelegenheid om de leden van de werkgroep te vertellen over het enorm brede veld van toepassingen dat aan die universiteit te vinden was. Zo waren er wiskundige modellen ontwikkeld en getoetst voor het optimaal poffen van granen, gebaseerd op fysische eigenschappen van de stoffen in een korrel. Maar ook variantie-analytische modellen voor de overerving van eigenschappen bij het kruisen van gewassen of dieren, of schema's voor optimale bemonstering van bodems in omstandigheden waar je niet precies een keurig raster overheen kunt leggen, zoals in moeilijk toegankelijke gebieden.

Een nogal bijzonder onderzoek was dat naar de mogelijkheid om van kleinere hoeveelheden geiten- en schapenmelk kaas te maken onder tropische omstandigheden. Melk is een goede bron voor eiwitten, maar bederft snel zonder koeling en kan daarom slecht verkocht en gebruikt worden buiten het eigen gezin. De oplossing ligt in kaas, die vormt een beter houdbare bron van eiwitten. Maar de westerse methoden om kaas te maken zijn in de tropen niet bruikbaar: je moet een zekere minimum hoeveelheid melk hebben om dat

te kunnen doen en de melk van meerdere dagen bewaren tot je genoeg hebt kan niet. Aan de LUW had men methoden ontwikkeld om onder tropische en vaak primitieve omstandigheden ook van kleine hoeveelheden melk kaas te kunnen maken. Ik vond dat een geweldig onderzoek dat ik zeker bij mijn presentatie wilde betrekken. Om het geheel een beetje te illustreren had ik een paar schapen- en geitenkaasjes gekregen om blokjes uit te kunnen delen.

De afdeling voorlichting van de universiteit had me een mooie video gegeven die werd getoond terwijl de schoteltjes met kaas rondgingen. Op die video werd ook een onderzoek naar de spijsvertering van runderen gepresenteerd. Daarbij werd gebruik gemaakt van een koe waarbij een fistel was aangebracht. Dat is een kunstmatige opening in de buikwand waardoor je monsters kunt nemen van de maaginhoud. Die opening is afgesloten met een grote kunststof Schroefdop. De video liet een rustig herkauwende koe zien, een laborant kwam aanlopen, schroefde de koe open, stak zijn hand met een pollepel tot aan zijn elleboog in de koe en goot een groene drabbige vloeistof in een glazen maatbeker. Alles werd goed zichtbaar en zeer plastisch in beeld gebracht. En de koe ging intussen rustig door met herkauwen.

Ik was druk bezig met het bedienen van de video-recorder en had niet direct door hoe mijn gehoor reageerde. Maar ik heb het nog jaren moeten horen dat ik hen deze beelden tijdens het genieten van kaas presenteerde. Wist ik veel dat sommige van mijn collega's zo'n zwakke maag hadden!

GERRIT STEMERDINK (eindredacteur van *STATOR*)
E-mail: gjstemerding@hotmail.com

Willem Rutger van Zwet 1934 – 2020 vermaard wetenschapper en krachtig bestuurder

AAD VAN DER VAART

Willem Rutger van Zwet, emeritus hoogleraar aan de Universiteit Leiden en erelid van de VVSOR, is op 2 juli 2020 overleden in de leeftijd van 86 jaar. Willem was wetenschappelijk kleinzoon van David van Dantzig, de eerste hoogleraar in de mathematische statistiek in Nederland. Hij heeft in de periode 1960 – 2000 een belangrijke rol gespeeld in de opbouw van onze discipline in Nederland, maar heeft vooral ook internationaal veel invloed gehad. De door de vereniging in 2012 ingestelde Willem van Zwet Prijs voor het beste proefschrift van het jaar is naar hem vernoemd.

Student en promovendus

Willem is geboren in 1934 in Leiden. Na zijn eindexamen aan het gymnasium in Den Haag studeerde hij natuurkunde en wiskunde in Leiden gedurende de jaren 1950. Natuurkunde vond hij vreselijk en 'met wiskunde kon je alleen leraar worden'.¹ Als uitweg koos Willem ervoor zich te specialiseren in de statistiek, een vak dat echter in Leiden niet gedoceerd werd. Willem ontving toestemming tot het volgen van de colleges van David van Dantzig in Amsterdam, die daar na de oorlog was benoemd tot hoogleraar in de 'collectieve verschijnselen' (dat wil zeggen statistiek), maar van origine topoloog was. In een interview uit 2000 noemde Willem de colleges van Van Dantzig 'random', niet zelden eindigend met de opdracht het zelf maar uit te zoeken, maar 'spectaculair'. Het beviel ook dat Van Dantzig een uitgesproken mening had over de wiskundige waarde

van grondbeginselen van de statistiek, zoals uiteengezet door Fisher en Savage, en die niet onder stoelen of banken stak. Het motiveerde Willem de studie voort te zetten met een promotieonderzoek, waarvoor hij aanvankelijk naar Berkeley wilde. Omdat hij lang over zijn studie had gedaan, aansluitend nog twee jaar in militaire dienst moest, en inmiddels in het huwelijk was getreden – Lucie is gedurende zijn leven aan zijn zijde gebleven, met een voor beiden bevredigende, duidelijke verdeling van taken – besloot hij in Nederland te blijven. Van Dantzig, die in 1959 overleed, werd in Amsterdam opgevolgd door Jan Hemelrijk, bij wie Willem in 1964 promoveerde.

In het genoemde interview uit 2000 noemde Willem Van Dantzig en Hemelrijk als zijn grootste professionele invloeden, naast Peter Bickel die hij later zou ontmoeten in Berkeley. Van Dantzig en Hemelrijk hadden een heel andere visie op statistiek (hoewel de tweede bij de eerste was gepromoveerd). Waar bij Van Dantzig zijn zuiver wiskundige achtergrond altijd dominant bleef en vooral de grondslagen van de statistiek voorop stonden, was Hemelrijk een toegepast statisticus, een heel goede volgens Willem, die praktische problemen van diverse aard oppakte.

Willem was tijdens zijn promotietijd verbonden aan het Mathematisch Centrum in Amsterdam (nu CWI), na de oorlog opgezet met het doel toegepaste wiskunde te bevorderen, ook met het oog op de opkomende computer. In zijn eerste jaren kwam Willem niet aan promotiewerk toe, maar had al zijn aandacht nodig voor consultatie van diverse aard. Op mathematisch



Willem van Zwet op zijn 75ste verjaardag. Foto: Chris Klaassen

statistiek gebied waren op dat moment rangtoetsen in zwang, maar Willem vond de Amsterdamse bijdrage ondermaats en besloot zijn proefschrift te wijden aan 'convexe transformaties in de kansrekening en statistiek'. Zijn proefschrift, verschenen in de reeks Mathematisch Centrum Tracts, wordt ook nu nog geciteerd. Overigens was het Mathematisch Centrum op dat moment een broedplaats voor de statistiek. Zo waren ook VVSOR ereleden Constance van Eeden en Fred Steutel in die tijd aan het centrum verbonden.

Lector en hoogleraar

Hoewel Willem hoog opgeeft over de toegepaste statistiek van Hemelrijk, is het Van Dantzig die hij in zijn latere carrière volgt. In het interview uit 2000 spreekt hij van een door hem in de jaren 1960 gesteld doel om de 'statistiek in Nederland op een behoorlijk mathematisch niveau te brengen'. Terugkijkend ziet hij dat hij daarin geslaagd is, maar betreurt dat de maatschappelijke kant van de statistiek is achter gebleven. (Het interview was in een periode van publieke discussie of het RIVM wel genoeg echte data verzamelde en zich niet te veel op theoretische modellen baseerde. Na zijn emeritaat was Willem ook zeer geïnteresseerd in de juridische aspecten van de statistiek, mede naar aanleiding van de Lucia de B. zaak.)

Na een verblijf in Oregon keerde Willem in 1965 terug naar Leiden als lector, naar zijn zeggen het type positie voor een 'vakgebied dat niet echt van belang werd geacht', maar hij werd al na enkele jaren bevorderd tot hoogleraar in de Statistiek, het vak dat 400 jaar in Leiden had ontbroken. In Leiden had hij de keuze toe te treden tot het Mathematisch Instituut, of het door Zoutendijk net opgerichte Centraal Rekeninstituut. Het werd 'natuurlijk' het tweede, wat uiteindelijk tot een toegepaste-wiskunde-instituut leidde. Wel verbaasde het Willem dat in Leiden kennelijk weinig praktisch statistische vragen leefden, anders dan in Amsterdam, waar men het Mathematisch Centrum wist te vinden. Het gaf hem gelegenheid zijn onderzoek te ontwikkelen. Willem bleef aan Leiden verbonden tot zijn pensioen in 1999, 'with no

complaints so far, volgens zijn afscheidsrede, alhoewel hij in die periode veel reisde, vooral naar Berkeley. Van 1990 – 1998 was hij ook William Newman Professor aan de Universiteit van North Carolina te Chapel Hill waar hij een deel van het jaar verbleef.

Onderzoeker en wetenschapper

Willem had een prachtige carrière als onderzoeker. Zijn werk aan asymptotische ontwikkelingen, hogere orde efficiëntie, en bootstrap methoden neemt daarin een centrale plaats in. Dat was deels samen met de al eerdergenoemde Peter Bickel, die hij in 1972 leerde kennen. Wetenschappelijk was hij zeer verbonden met de Amerikaanse school in de statistiek. De keuze voor een postdoc in Oregon was ingegeven door Fred Andrews, die het CWI bezocht had. Vanuit Oregon reed Willem met de auto naar Californië om deel te nemen aan het Second Berkeley Symposium (1967), waarvoor Andrews hem had uitgenodigd. Toen Willem in 1970 een uitnodiging voor het Third Berkeley Symposium ontving was dat voor hem ‘een sensatie’. Berkeley was een centrum voor de statistiek, met onder andere Neyman, Scott, Lehmann, Le Cam, en Peter Bickel. Peter, zes jaar jonger dan Willem, was op dat moment ook geïnteresseerd in Edgeworth ontwikkelingen. Hun ontmoeting leidde tot een levenslange samenwerking en vriendschap, veel bezoeken over en weer, en bijna tien artikelen. Daaronder enkele extreem lange (bijna 70 pagina’s in de *Annals of Statistics*) over rangtoetsen, het onderwerp dat Willem in zijn Amsterdamse tijd nog verguisde, maar waar hij op dat moment een nieuwe draai aan had gegeven. De belangstelling ging nu uit naar het gedrag van die toetsen onder de alternatieve hypothese, op de verfijnde manier van hogere ontwikkelingen voor de power functie en het door Lehmann geïntroduceerde begrip van deficiëntie.

Willem schreef daarnaast over uiteenlopende andere onderwerpen, waaronder de groei van plantencellen en spatiële stochastische processen. Hij was daarin enigszins eclectisch, zoals geïllustreerd door zijn werk aan het Kakutani probleem dat hem veel voldoening schonk. Welbekend is dat de cumulatieve dichtheid van onafhankelijk en willekeurig in een interval geplaatste punten afwijkt van de homogene verdelingsfunctie volgens een brownse brug proces. Als nu ieder nieuw punt in plaats van willekeurig in het hele interval, willekeurig in het langste interval tussen twee voorgaande punten

wordt geplaatst, zal die afwijking kleiner zijn. Willem wist dit (*Annals of Probability*, 1978) precies te maken met een benadering door twee brownse bruggen. Overigens, op een vraag in een interview uit 2006² naar zijn meest geliefde artikelen, noemde hij heel vriendelijk gezamenlijke artikelen met de meeste promovendi, en trouwens ook één met zijn zoon Erik (Van Zwet en Van Zwet, 1999).

Willem werd breed gewaardeerd om zijn wetenschappelijke werk. Hij ontving talrijke eerbewijzen, waaronder een ere-doctoraat in Praag, was lid van de Koninklijke Academie van Wetenschappen, en de Academia Europaea. Hij was de eerste ontvanger van de Van Dantzig prijs van de VVSOR (1970). Hij was een excellent en veel gevraagd spreker, en gaf onder meer de Hotelling (1988), Wald (1992), en Bahadur (2005) voordrachten.

Bestuurder en redacteur

Misschien nog meer zal Willem worden herinnerd als bestuurder en organisator. Hij was decan in Leiden, op het moment dat daar moeilijke beslissingen moesten worden genomen, maar ook lid van het bestuur van het Rijksherbarium en het Leids Universiteitsfonds. Hij was erepromotor voor Erich Lehmann (1985). Het is niet overdreven om Willem de koning van de (academische) statistiek in Nederland te noemen. Weinig, inclusief benoemingen, gebeurde zonder zijn medeweten. De jaarlijkse Lunteren-bijeenkomsten (geldend als een activiteit van de Sectie Mathematische Statistiek van de VVSOR) waren zijn initiatief; hij nodigde de sprekers persoonlijk uit in de lange periode 1972 – 1999. De oprichting van EURANDOM was het resultaat van zijn lobbywerk bij het ministerie en de collega wiskundigen, en hij was de eerste directeur. EURANDOM, nu meer een conferentiecentrum voor stochastiek, was gedurende de eerste periode een onderzoeksinstituut met rond de 25 postdocs, en een duidelijke aanjager van de stochastiek in Nederland.

En zoveel meer. Vooral ook in het internationale. Willem was de president van alle drie wereldorganisaties in de statistiek, IMS, ISI, Bernoulli Society. Hij was voorzitter van het organisatiecomité van de ISI Centenary Session in Amsterdam (1985). Hij was nauw betrokken bij het ontstaan van de tweejaarlijkse European Meetings of Statisticians. Hij was *board and corporation member* van het National Institute for Statistical Sciences in North Caroli-



Kjell Doksum, Peter Bickel en Willem van Zwet, 1982

na, en lid van de *board of directors* van de ASA. Hij zorgde voor contact tussen oost (de ‘goede’, niet partijgebonden onderzoekers daar vond hij gemakkelijk te herkennen) en west tijdens de koude oorlog. Hij was hoofdredacteur van de twee grootste tijdschriften in de statistiek. Etcetera. Willem kende iedereen. Iedereen kende Willem. Als leerling was je daar zeer van bewust, al van de verhalen in Leiden en later zelf in het internationale circuit. Als ik aan het wiskundig conferentiecentrum Oberwolfach denk zie ik Willem daar te zien zitten, achterovergeleund met een borrel of een glas wijn, omringd door collega’s en vrienden, of soms een tegenstrever.

In het interview uit 2000 zegt hij over besturen dat hij het niet onplezierig vond, ‘want het geeft inzicht in de menselijke geest, hoewel dat niet tot optimisme stemt’. Willem schuwde daar de controverse niet. Zo meende hij het recordaantal boze brieven te hebben ontvangen tijdens zijn werk als editor, soms met kwalificaties die er niet om logen. Hij ging er dan ook prat op bepaalde onderwerpen te hebben ‘uitgeroeid’.

Promotor en inspirator

Willem had 16 promovendi, die allemaal de wetenschap zijn ingegaan. Op enig moment waren bijna alle hoogleraarsplaatsen mathematische statistiek in Nederland bezet door promovendi van Willem, of door promovendi van promovendi. In mijn praatje na Willems afscheidsrede in 1999 heb ik gesuggereerd dat dit gevaarlijk klonk. Maar daaraan onmiddellijk toegevoegd, dat het dat niet was, omdat Willem niet een *school maker* was. Zelfstandigheid waardeerde hij zeer, ook als dat soms tot irritatie leidde door een eigenwijze andere visie. Willem opende Nederland voor de inhoud en het niveau van de internationale wereld van de statistiek, eerder dan dat hij die statistiek inhoudelijk leidde en beperkte. In het interview vertelt Willem overigens dat hij eigenlijk graag alleen werkte, met uitzondering dan met promovendi en Peter Bickel.

Willem was sinds 1996 Ridder in de Orde van de Nederlandse Leeuw, een onderscheiding die hem na aan het hart lag. Hij heeft zich in de afgelopen 20 jaar ingezet om een soortgelijke onderscheiding voor zijn wetenschappe-

lijke vrienden te bewerken, waaronder Peter Bickel, Jon Wellner, Friedrich Götze, Marie Hušková en Sara van de Geer. Hoezeer hij deze onderscheiding waardeerde komt tot uiting in zijn e-mail in november 2018 aan Stéphanie van der Pas, in reactie op een verzoek om advies: ‘Nog een algemene opmerking. Veel Nederlanders doen vaak een beetje lacherig over dingen zoals lintjes, maar anderen nemen lintjes heel serieus. Als je schrijft dat een lintje voor Aad “een aardige toevoeging aan het congres zou zijn” zullen die anderen denken dat “het congres misschien een aardige toevoeging aan het lintje zou zijn”. Dat is beter te vermijden.’

Willems carrière valt samen met een periode van grote veranderingen in de statistiek en in de wereld van de wetenschap in het algemeen. Het eerder genoemde interview uit 2006 met Rudy Beran en Nick Fisher belicht Willems visie, en andere aspecten van zijn leven, en is de moeite van het lezen waard. Eén van Willems hobbies komt naar voren in antwoord op de vraag: ‘*If you were stuck on a desert island with a limited choice of reading materials, which of your papers, among those available, would you take with you?*’ Willem: ‘*That is a strange question. Why would I take any of my papers? A crate of Dutch genever or bourbon if you like would make a lot more sense.*’

Wij herinneren dat het ‘oude genever’ was. Oude je-never voor oude mensen, dachten we, als jonge mensen. Maar jonge mensen worden oud, en we zullen de richtinggevende aanwezigheid van Willem van Zwet missen.

NOTEN

1. Een aantal citaten is overgenomen uit: Mathisca de Gunst, Chris Klaassen & Aad van der Vaart, Interview met Willem van Zwet, 2000, ongepubliceerd.
2. Rudy Beran & Nick Fisher. An Evening Spent with Bill van Zwet. Verschenen in *Statistical Science*, 2009. Zie <https://projecteuclid.org/euclid.ss/1255009013>.

AAD VAN DER VAART is hoogleraar Stochastiek aan het Mathematisch Instituut van de Universiteit Leiden. Van der Vaart studeerde wiskunde, filosofie en psychologie aan de Universiteit Leiden en promoveerde bij Willem Rutger van Zwet in 1987 te Leiden op *Statistical estimation in large parameter spaces* (statische schattingen in hoogdimensionale ruimtes).
E-mail: avdvaart@math.leidenuniv.nl

Nieuws over het World Statistics Congress 2021



Corona heeft ook de voorbereiding getroffen van het World Statistics Congress dat gepland staat voor 11 – 15 juli 2021 in Den Haag. Het is op dit moment geheel onduidelijk of er tegen die tijd überhaupt toestemming zal zijn een dergelijke grote fysieke bijeenkomst te houden. Afgezien daarvan zijn er overal ter wereld lockdowns en reisbeperkingen en ook daarvan is onduidelijk hoe de toestand volgend jaar zal zijn. Het bestuur van het ISI en het National Organizing Committee hebben daarom vier mogelijkheden uitgebreid onderzocht: uitstellen, afgelasten, geheel virtueel en gedeeltelijk fysiek en gedeeltelijk virtueel. Daarbij is gekeken naar aspecten als kosten, repercussies voor sponsoring, mogelijkheid voor deelnemers om fysiek aanwezig te kunnen zijn etc.

Online

Het Executive Committee van het ISI heeft, op grond van al deze onderzoeken, besloten het congres geheel virtueel te laten plaatsvinden. Op de virtuele Algemene Ledenvergadering van 27 juli is dit aan de leden meegedeeld. De reacties waren zeer positief: het is namelijk op deze manier mogelijk een veel groter aantal statistici te laten deelnemen omdat de reis- en verblijfkosten vervallen. Vooral voor hen die aan het begin van hun carrière staan vormen die kosten dikwijls een belemmering voor deelname aan dit soort grote internationale congressen. Om een bekende Nederlander te citeren: 'elk nadeel heft z'n voordeel'.

Er wordt de komende tijd druk nagedacht over de verdere uitwerking van dit nieuwe concept. Daarbij komen heel veel zaken aan de orde, zo moet er bijvoorbeeld worden gekeken over hoe om te gaan met de enorme verschillen in tijdzones van de deelnemers. Gelukkig is het

ISI niet de enige organisatie die de vorm van haar activiteiten plotseling zeer ingrijpend moet veranderen. Er zal bijvoorbeeld worden gekeken naar de ervaringen van de American Statistical Association met haar Joint Statistical Meetings die begin augustus 2020 plaatsvinden, ook dat zal een virtueel congres zijn. Onderstaand het bericht dat op 27 juli aan de ISI-leden is bekendgemaakt.

ISI World Statistics Congress 2021 goes VIRTUAL!

The International Statistical Institute is pleased to announce that its flagship biennial World Statistics Congress 2021 – The Hague, will be held virtually in July 2021.

This is an opportunity for the ISI to host a more inclusive conference than ever before. It will allow us to have the greatest ever participation, reaching out to a wide diversity of members who would not otherwise be able to afford to attend an international conference, at lower costs, and with a lower carbon footprint. We hope to see a large number of first-time participants, including many early career statisticians and students along with colleagues from developing countries.

Holding a WSC in 2021 as planned will maintain the continuity of WSCs. There will still be a strong link with The Hague, the original venue, in the virtual conference. The ISI will take the opportunity to develop new formats and harness new technology to modernise our meeting.

Further details, including a call for papers, will be announced in due course.



Annual Meeting 2020. Foto: Michiel Ton

**Going
Viral**

Modeling spread

VVSOR ANNUAL MEETING 2021

thema

Going viral – Modeling spread

**donderdag 18 maart 2021 in
In de Driehoek Utrecht | Online**

De Annual Meeting (AM) van 2021 zal in het teken staan van verspreiding. Denk daarbij niet alleen aan de verspreiding van virussen, maar ook aan informatie, meningen, goederen of mensen. De sprekers zullen vanuit verschillende perspectieven uit de Statistiek en OR hun licht laten schijnen op het modelleren van verspreiding.

Onder de voorwaarde dat de verspreiding van Covid-19 het toelaat, zal de AM, net als in 2020, plaatsvinden bij In de Driehoek in Utrecht. De dag zal worden afgesloten met een diner en een pubquiz. De AM van 2021 zal een hybride evenement worden: voor het eerst ook zal de AM ook als online streaming event plaatsvinden, zodat ook belangstellenden die niet naar Utrecht kunnen komen de meeting bij kunnen wonen.

De AM 2021 zal Engelstalig zijn. Verdere bijzonderheden kunt u vinden in de volgende editie van *STATOR* en op de website van de VVSOR.

E-mail: annualmeeting@vvsor.nl

