STAtOR

periodiek van de VvS+OR jaargang 17, nummer 1, maart 2016

The annual meeting of the Netherlands Society for Statistics and Operations Research is the place to meet other people with an interest in statistics and operations research and people working in this field.

www.vvs-or.nl

REGISTRATION (BEFORE MARCH 10th)

Members VvS+OR
The lectures and annual meeting are open to members of the Society, speakers and invited guests, but registration through the society's website is required: <www.vvs-or.nl>. Register before March 17th and bring the confirmation e-mail to show it at the registration desk at the entrance of the lecture room.

Non-members
Interested non-members can access the day when they become a member of the Society in advance or pay the amount of € 50 (bank account: NL42 INGB 0000 202091, BIC/ SWIFT: INGBNL2, VvS+OR 'registration annual meeting 2015') before March 10th and bring proof of payment (copy of bank statement) to show at the registration desk at the entrance of the lecture room. Alternatively, you can become an ordinary member for € 70 (go to <www.vvs-or.nl> and click 'Become a Member' in the top menu; young applicants will obtain a special rate), and have free access immediately after registration.

GENERAL INFORMATION

Language
The talks at the annual meeting will be in English, the Annual General Meeting (ALV) will be in Dutch.

Annual General Meeting (Algemene Ledenvergadering)
The Annual General meeting (ALV) is scheduled at the end of the morning. The relevant documents will be provided on the website two weeks before the meeting. You can also get them by e-mail if you send a request to <info@vvs-or.nl>.

Coffee, tea and drinks after the meeting
Coffee/tea during the breaks and drinks afterwards are offered by the Society.

Lunch
Lunch is at your own cost. The restaurant of the Jaarbeurs is usually quite busy, but within walking distance of just a few minutes (e.g. in the direction of Central Station) you will find many alternatives.

ORGANIZING COMMITTEE

The annual meeting is organized by the board of the VvS+OR. For questions, contact Fetsje Bijma (secretary) by e-mail at <info@vvs-or.nl>.

THE ROLE OF STATISTICS IN DATA SCIENCE

Thursday, March 17th 2016

Annual meeting of the Netherlands Society for Statistics and Operations Research (VvS+OR)

LOCATION
Jaarbeurs Utrecht, Room 215
Jaarbeursplein 6, 3521 AL Utrecht (adjacent to Central Station)

KEYNOTE SPEAKERS
David Hand (Imperial College London)
Aad van der Vaart (Leiden University)
Tim Harford (BBC, Financial Times)

PROGRAM

9.30 - 10.00	Registration and coffee | tea
10.00 - 10.15	Opening plenary and introduction first speaker
10.15 - 11.15	DAVID HAND (Imperial College London)
11.15 - 11.30	Coffee | tea
11.30 - 12.30	Annual General Meeting (ALV, in Dutch)
12.30 - 13.30	Lunch break
13.30 - 14.40	AAD VAN DER VAART (Leiden University)
14.40 - 15.20	Ceremony of the VAN ZWET AWARD
	Presentation Van Zwet Award winner(s)
15.20 - 15.35	Coffee | tea
15.35 - 16.00	Ceremony of the HEMELRIJK AWARD
	Presentation Hemelrijk Award winner(s)
16.00 - 17.00	TIM HARFORD (BBC, Financial Times)
17.00	Drinks

The theme of the 2016 Annual Meeting of the Duch Society for Statistics and Operational Research is 'The role of Statistics in Data Science'.

Now that Big Data are emerging literally everywhere Data Science is becoming an increasingly important field of research. Statistics can play a crucial role in this development. Each from their specific background, all three speakers will focus on this topic.

1

About David Hand

David Hand is Senior Research Investigator and Emeritus Professor of Mathematics at Imperial College, London, and Chief Scientific Advisor to Winton Capital Management. He is a Fellow of the British Academy, and a recipient of the Guy Medal of the Royal Statistical Society. He has served (twice) as President of the Royal Statistical Society, and is on the Board of the UK Statistics Authority. He has published 300 scientific papers and 26 books. He has broad research interests in areas including classification, data mining, anomaly detection, and the foundations of statistics. His applications interests include psychology, physics, and the retail credit industry - he and his research group won the 2012 Credit Collections and Risk Award for Contributions to the Credit Industry. He was made OBE for services to research and innovation in 2013.

2

About Aad van der Vaart

Aad van der Vaart studied mathematics, philosophy and psychology at the University of Leiden, and received a PhD in mathematics from this university in 1987. He held positions in College Station, Texas and Paris (not in Texas), and was visiting professor in Berkeley, Harvard and Seattle. Following a long connection to the Vrije Universiteit Amsterdam he is currently *Professor of Stochastics* at Leiden University, and member of the Royal Netherlands Academy of Arts and Sciences.

His research is in statistics and probability, as mathematical disciplines, and in their applications to other sciences, with an emphasis on statistical models with large parameter spaces. Among several honours, he is recipient of the VvS+OR Van Dantzig Award (2000) and the NWO Spinoza Prize (2015). He wrote books and lecture notes (on topics such as empirical processes, time series, stochastic integration, option pricing, statistical genetics, statistical learning, Bayesian nonparametrics), as well as research papers.

3

About Tim Harford

Tim Harford is economics leader writer for the *Financial Times* and writes the 'Undercover Economist' columns on Saturdays. He first joined the FT as Peter Martin Fellow in 2003 and after a spell at the World Bank in Washington DC he rejoined the FT's leader writing team in 2006. Tim's book, *The Undercover Economist*, is a Business Week bestseller and a Sunday Times bestseller, and was number one on Amazon.co.uk. It has been translated into sixteen languages. He is now working on a sequel.

Tim is also the presenter of the BBC2 series, *Trust Me, I'm an Economist*.

STAtOR
Jaargang 17, nummer 1, maart 2016

STAtOR is een uitgave van de Vereniging voor Statistiek en Operationele Research (VvS+OR). *STAtOR* wil leden, bedrijven en overige geïnteresseerden op de hoogte houden van ontwikkelingen en nieuws over toepassingen van statistiek en operationele research. Verschijnt 3 keer per jaar.

Redactie
Joaquim Gromicho (hoofdredacteur), Annelieke Baller, Ana Isabel Barros, Kristiaan Glorie, Johan van Leeuwaarden, Guus Luijben (eindredacteur), Richard Starmans, Gerrit Stemerdink (eindredacteur), Hilde Tobi en Vanessa Torres van Grinsven. Vaste medewerkers: Fred Steutel 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, telefoon 020-5986010, mobiel 06-55886747, <j.a.dossantos.gromicho@vu.nl>.

Bestuur van de VvS+OR
Voorzitter: prof. dr. Jacqueline Meulman <president@vvs-or.nl>
Secretaris: dr. Fetsje Bijma <bestuur@vvs-or.nl>
Penningmeester: dr. Ad Ridder <penningmeester@vvs-or.nl>
Overige bestuursleden: prof. dr. Eric Cator (SMS), prof. dr. Jeanine Houwing-Duistermaat (BMS), Maarten Kampert MSc., prof. dr. Albert Wagelmans (NGB), dr. Michel van de Velden (ECS), dr. Jelte Wicherts (SWS), Nynke Krol (Young Statisticians)

Leden- en abonnementenadministratie van de VvS+OR
VvS+OR, Postbus 244, 6700 AE Wageningen, telefoon 0317 - 419572, e-mail <admin@vvs-or.nl>.
Raadpleeg onze website over hoe u lid kunt worden van de VvS+OR of een abonnement kunt nemen op *STAtOR*.

VvS+OR-website
www.vvs-or.nl

Sociale media
Wilt u uw vakgenoten ontmoeten en wilt u discussiëren over actuele thema's, volg dan de VvS+OR en de Young Statisticians via LinkedIn, Facebook, Twitter en Flickr.
Sluit u aan bij de LinkedIn-groep van VvS+OR of Young Statisticians; bekijk foto's op <www.flickr.com/photos/vvs-or/sets>; like onze Facebook-pagina; volg de President van VvS+OR op <https://twitter.com/#!/dutchstat>.

Advertentieacquisitie
M. van Hootegem <hootegem@xs4all.nl>
STAtOR verschijnt in maart, juli en december.

Ontwerp en opmaak
Pharos, Nijmegen

Uitgever
© Vereniging voor Statistiek en Operationele Research
ISSN 1567-3383

INHOUD

Dag voor Statistiek en OR 2016

Over tellen, vertellen en nog meer

Misschien is het al eens eerder opgemerkt in dit blad: eigenlijk is statistiek de oudste wetenschap! Sterker nog, strikt genomen is statistiek nóg ouder dan de ontwikkeling van het schrift.

Toen de eersten van onze voorouders zich ontwikkelden van jagers en verzamelaars tot gevestigde landbouwers kwam er behoefte aan een inventarisatie. De allereerste uitingen daarvan bestaan uit kerfjes op stokken, beenderen en potscherven. Voilà: de statistiek in oervorm was geboren. Vanuit die eerste simpele kerfjes hebben zich allerlei schriftvormen ontwikkeld.

Later, toen men zich organiseerde in grotere samenlevingen zoals steden en landen werd die kennis over bevolking en voorraden steeds belangrijker. Zo kennen we voorbeelden van administraties uit Mesopotamië en Egypte. En wie kent niet het meer dan tweeduizend jaar oude verhaal over keizer Augustus die een volkstelling organiseerde, waardoor Jozef en Maria naar Bethlehem moesten reizen.

Aan die behoefte van een samenleving om te weten wat er om gaat is zelfs de naam van ons vakgebied afgeleid, statistiek betekent niets meer of minder dan 'staat van het land'.

Tellen. Dit nummer van *STAtOR* doet volledig recht aan onze oeroude wortels. Eric Schulte Nordholt vertelt over volkstellingen en de manier waarop dat tegenwoordig kan zonder alle bewoners een formulier te laten invullen. En Jan Kardaun beschrijft hoe statistieken van doodsoorzaken tot stand komen. Ook Johan van Leeuwaarden telt, hij blijkt – via publicaties – slechts vier stappen verwijderd te zijn van Einstein.

Vertellen doen Fred Steutel en Henk Tijms in hun columns, zowel over prikkeldraad als het getal van Euler. Fred Steutel vervolgt ook onze nieuwe reeks 'Klassiekers' met zijn herinneringen aan een boek van Eugene Lukacs. Ook is er nieuws van de Young Statisticians die vertellen over hun plannen en een verslag van de LNMB-bijeenkomst in Lunteren.

Nog meer is te vinden in de rest van dit nummer. Bert Zwart, die vorig jaar de vijfjaarlijkse van Dantzig prijs mocht ontvangen gaat in op zijn wetenschappelijke plannen voor de komende jaren: statistiek en OR om betrouwbare energienetwerken te kunnen waarborgen. Ronald Meester en Timber Kerkvliet sluiten met hun artikel over kwantificeren van onzekerheid goed aan bij het artikel Peter Grünwald in ons vorige nummer. Koen Bel ten slotte schrijft over multivariate uitbreidingen van discrete keuzemodellen.

Uiteraard vindt u in dit eerste nummer van het jaar veel informatie over de Dag voor Statistiek en OR op 17 maart. Daar zal een drietal gerenommeerde sprekers ingaan op Big Data, elk vanuit hun eigen specialisme. Wij hopen veel van onze lezers daar te ontmoeten. Tot het zover is: veel leesplezier.

De STAtOR-redactie

Pieter Bruegel de Jonge, *Volkstelling te Bethlehem* (ca 1605-1610). Collectie Bonnefantenmuseum Maastricht

VOLKSTELLINGEN IN NEDERLAND EN ELDERS

In Nederland werden tot en met 1971 traditionele volkstellingen gehouden waarbij alle inwoners verplicht waren vele vragen te beantwoorden. Tegenwoordig vinden volkstellingen vrijwel ongemerkt en tegen veel lagere kosten plaats op basis van bij het Centraal Bureau voor de Statistiek (CBS) beschikbare gegevens. De basis van de huidige tellingen in Nederland wordt gevormd door de bevolkingsadministratie en het door het CBS ontwikkelde Stelsel van Sociaal-statistische Bestanden (SSB). Hoewel de meeste landen nog steeds traditionele volkstellingen houden, worden dit soort registertellingen steeds populairder. De resultaten van de verschillende recente volkstellingen in Europa zijn onderling goed vergelijkbaar als gevolg van internationale afspraken over te hanteren definities en classificaties.

Eric Schulte Nordholt[1]

Volkstellingen zijn de oudste statistieken die we kennen. In China en Egypte werd ongeveer 3000 jaar geleden de bevolking in kaart gebracht. Belangrijke redenen waren om de militaire kracht en de fiscale mogelijkheden te inventariseren. Ook in het Romeinse Rijk vonden dergelijke inventarisaties plaats. Reeds lang geleden werd in kaart gebracht hoeveel zielen de bevolking in Bethlehem telde (Willems, 2011). De schilder Pieter Bruegel de Jonge heeft daar een schilderij van gemaakt (hiernaast afgebeeld).

In de achttiende eeuw werden in Nederland verschillende regionale tellingen gehouden. De oudste integrale telling vond in 1795 plaats voor de verkiezingen in de toenmalige Bataafse Republiek (Cramer, 2005). In 1795 werden slechts 2,1 miljoen zielen geteld, terwijl Nederland tegenwoordig ongeveer 17 miljoen inwoners heeft. Sinds Nederland een koninkrijk is geworden zijn 18 volkstellingen gehouden. De eerste zeven tellingen werden door het Ministerie van Binnenlandse Zaken georganiseerd. Sinds de oprichting in 1899 is het CBS verantwoordelijk voor de Nederlandse volkstellingen. Aanvankelijk werden aparte volks-, woning- en beroepentellingen gehouden, later zijn al die tellingen geïntegreerd. Sinds 1829 zijn ruwweg elke tien jaar volkstellingen gehouden, maar in 1940 kon de volkstelling geen doorgang vinden door het uitbreken van de Tweede Wereldoorlog. In de laatste decennia is telkens een volkstelling gehouden in jaren eindigend op een 1. Dit is in alle EU-landen gedaan om de cijfers tussen deze Europese landen gemakkelijker te kunnen vergelijken.

In tabel 1 zijn enkele kerngegevens over Nederland te vinden die op basis van deze 18 tellingen zijn bepaald. Enkele zaken vallen op. In de eerste plaats natuurlijk de enorme groei van de Nederlandse bevolking. In de tweede plaats zien we in de periode na de Tweede Wereldoorlog het percentage jongeren afnemen (ontgroening) en het percentage ouderen toenemen (vergrijzing).

Traditionele volkstellingen in Nederland

Aanvankelijk werd bij volkstellingen de bevolking van Nederland op traditionele wijze in kaart gebracht. Hiertoe werden vragenlijsten ontworpen en voor alle inwoners moesten gegevens worden genoteerd. Ook werden

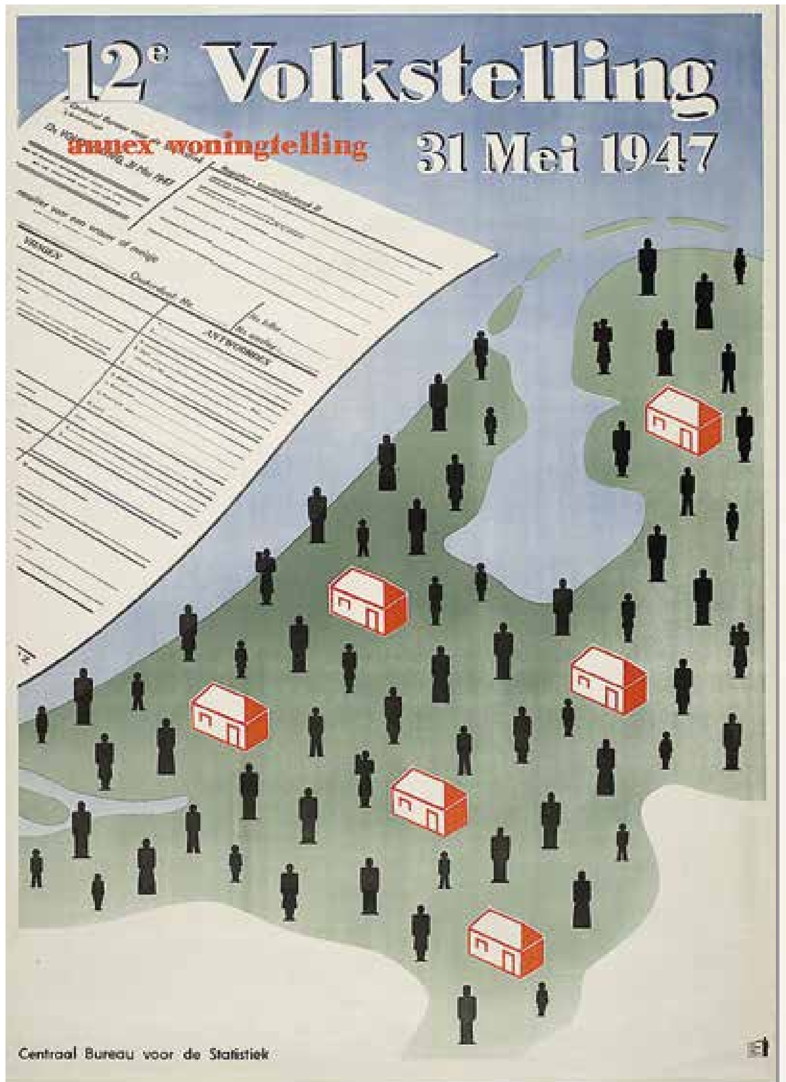
JAAR	LEEFTIJD			
	alle leeftijden	0–19 jaar	20–64 jaar	65 +
	× 1000	% van de totale bevolking		
1829	2613,3	44	50	5
1839	2861,6	45	50	5
1849	3056,9	43	53	5
1859	3309,1	42	53	5
1869	3579,5	43	52	6
1879	4012,7	44	50	5
1889	4511,4	45	49	6
1899	5104,1	44	50	6
1909	5858,2	44	50	6
1920	6865,3	42	52	6
1930	7935,6	40	54	6
1947	9625,5	38	55	7
1960	11462,0	39	53	9
1971	13060,1	36	54	10
1981	14216,9	31	57	12
1991	15070,0	25	62	13
2001	15985,5	24	62	14
2011	16655,8	23	61	16

Tabel 1. Bevolking naar leeftijdsgroep en volkstellingsjaar.
Bron: Centraal Bureau voor de Statistiek

vragen gesteld over de woningen zoals uit figuur 1 blijkt.

Deelname aan volks- en woningtellingen was verplicht. In de meeste landen ter wereld is dat nog steeds zo. Bij de Nederlandse Volkstelling van 1971 was de automatisering voortgeschreden. Er werd gebruik gemaakt van ponskaarten en daardoor kon de verwerking van de ingevulde vragenlijsten machinaal ter hand worden genomen. Helemaal probleemloos liep de verwerking bepaald niet, maar nog jarenlang werd gewerkt aan analyses van de resultaten. Deze analyses mondden onder meer uit in zogenaamde censusmonografieën.

De Volkstelling van 1971 veroorzaakte echter veel protest. De vrees voor aantasting van de privacy leidde onder meer tot de oprichting van het Comité Waakzaamheid Volkstelling dat affiches als in figuur 2 maakte. Bedacht moet worden dat de Tweede Wereldoorlog nog vers in het geheugen zat en dat Nederland in de jaren zeventig van de vorige eeuw nog geen privacywetgeving kende. In 1960 waren er maar twee weigeraars, in 1971 weigerden 35 duizend mensen mee te werken en nog veel meer mensen gaven niet thuis. Hoewel het aantal weigeraars een gering percentage van de bevolking was, besloot de Nederlandse regering enkele jaren later uiteindelijk niet tot vervolging van de weigeraars over te gaan. Het was duidelijk dat er in Nederland iets moest veranderen aan de opzet van de volkstelling. Er bestond wel vanuit (lokaal) beleid en de onderzoekswereld veel belangstelling voor de resultaten van de volkstelling, maar een nieuwe traditionele telling bleek niet meer haalbaar. In Holvast (2013) is een uitgebreide terugblik te vinden op de beroering die ontstond rond de Volkstelling van 1971.



Figuur 1. Poster van de '12e Volkstelling annex woningtelling 31 Mei 1947'

Figuur 2. Affiche tegen de Volkstelling 1971

Virtuele volkstellingen in Nederland

Het aantal studies op basis van de Nederlandse Volkstelling 1971 was groot. Het CBS heeft na de Volkstelling van 1971 gezocht naar alternatieve methoden om volkstellingsgegevens samen te stellen. Hoewel Nederland beschikt over een goede bevolkingsadministratie, zijn veel resultaten van volkstellingen niet uit die bron te halen. Denk hierbij aan de variabelen arbeidsmarktpositie, opleidingsniveau en beroep. Om toch Nederlandse gegevens te kunnen vergelijken met die van andere Europese landen zijn in het kader van de Europese Volkstellingen van 1981 en 1991 nationale tabellen samengesteld op basis van tellingen uit verschillende registraties en schattingen uit bestaande enquêtes. Hieraan is door het CBS weinig ruchtbaarheid gegeven en de oude Volkstellingswet is uiteindelijk in 1991 ingetrokken (Corbey, 1994).

In de jaren negentig van de vorige eeuw kwamen steeds meer registraties beschikbaar voor het samenstellen van de officiële statistiek. In de Wet op het Centraal Bureau voor de Statistiek en de Centrale Commissie voor de Statistiek van 1996 werd die toegang - en de mogelijkheid voor het CBS registers onderling te koppelen - voor het eerst expliciet vermeld. Er is toen een begin gemaakt met het opzetten van het SSB (Stelsel van Sociaal-statistische Bestanden). Het SSB is een uitgebreid stelsel van koppelbare registers en enquêtes. Met gegevens uit het SSB worden over uiteenlopende onderwerpen statistieken gemaakt en wordt sociaalwetenschappelijk onderzoek uitgevoerd. Het SSB is vanaf de Volkstelling 2001 een belangrijke bron geworden voor het samenstellen van volkstellingstabellen in Nederland. Meer informatie over het SSB is te vinden in Bakker, Van Rooijen en Van Toor (2014).

De huidige Wet bescherming persoonsgegevens trad in 2001 in werking ter vervanging van de Wet persoonsregistratie uit 1989. Door alle veranderingen in de opzet van de volkstellingen in Nederland en het kader waarbinnen die activiteiten plaatsvinden zijn problemen, zoals die bij de Volkstelling van 1971 speelden, nu niet meer te verwachten.

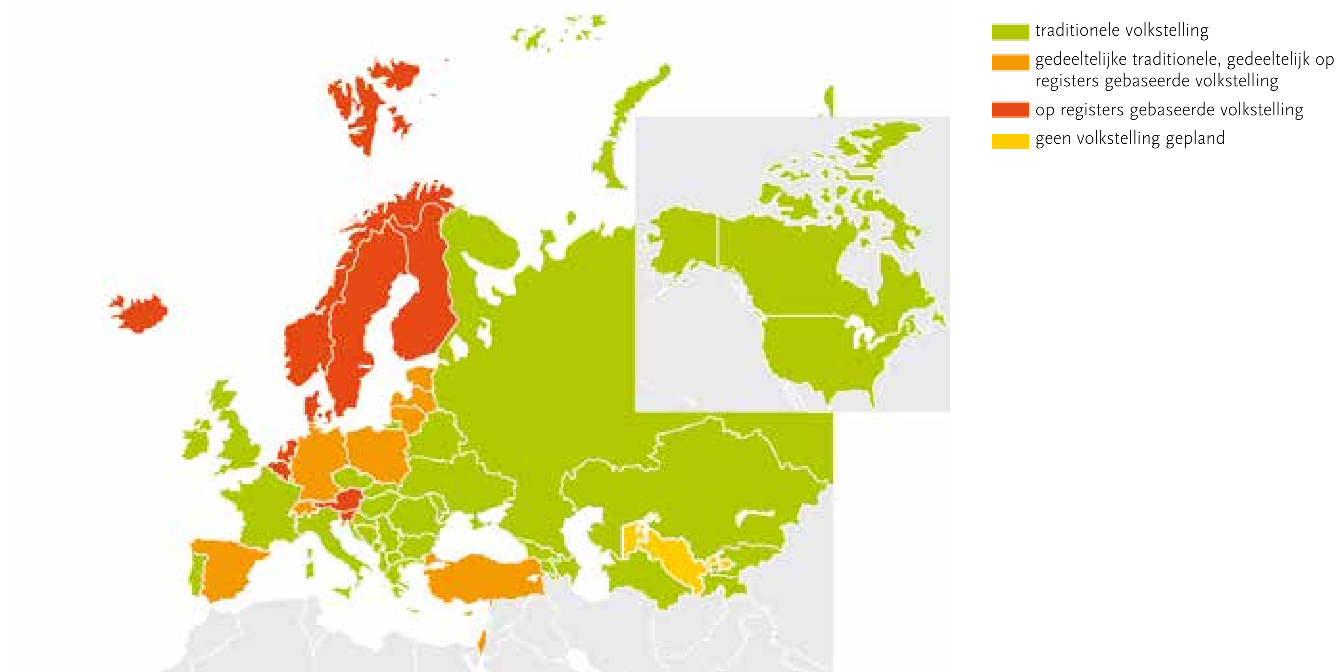
In de Europese Volkstellingsrondes van 2001 en 2011 was Nederland telkens een van de eerste landen die de onderling consistente tabellen gereed had. Meer informatie over het samenstellen van al deze tabellen en enkele hoofduitkomsten van deze tellingen kunnen worden gevonden in Schulte Nordholt (2004, 2014a en 2014b). Tegelijkertijd had haast niemand iets van deze activiteiten gemerkt, want er vond voor deze volkstellingen geen enkele enquête plaats (Van Eijk, 2004). Er wordt daarom wel gesproken van virtuele volkstellingen.

We kunnen ons afvragen hoe goed de kwaliteit van de moderne Nederlandse volkstellingstabellen is. Het is bekend dat er sprake is van enige vervuiling in het bevolkingsregister. In Bakker (2009) wordt de onder- en overdekking van de Nederlandse populatie in het bevolkingsregister in kaart gebracht. Er zijn natuurlijk vele duizenden mensen die in Nederland verblijven maar niet staan ingeschreven in het bevolkingsregister (bijvoorbeeld illegale vreemdelingen) en er zijn ook enkele duizenden mensen die ten onrechte (nog) staan ingeschreven (bijvoorbeeld recent geëmigreerden die zich niet hebben afgemeld bij de laatste gemeente waar ze woonachtig waren). Deze groepen zijn als percentage van de totale Nederlandse bevolking echter beperkt. Het valt sterk te betwijfelen of tegenwoordig betere resultaten zouden worden verkregen als in Nederland traditioneel zou worden geteld. De bereidheid mee te doen aan enquêtes is veel lager dan voorheen. Voor selectieve deelname aan enquêtes valt door middel van weging te corrigeren, maar het idee dat met een (verplichte) enquête voor iedereen in Nederland een compleet bestand van alle personen kan worden opgebouwd, is achterhaald. Terwijl in Nederland ooit volkstellingen werden gehouden om registers op te bouwen, is de situatie nu dus andersom: de meeste volkstellingsgegevens worden samengesteld op basis van informatie uit registers.

De volkstellingssituatie in andere landen

In Nederland wordt beleid gemaakt op basis van de geregistreerde bevolking. Er zijn heel veel prikkels om correct geregistreerd te staan. In de Noordelijke landen (IJsland, Noorwegen, Zweden, Finland en Denemarken) is dat ook het geval. In vele andere Europese landen is dat niet zo en zeker buiten Europa is de traditionele volkstelling nog steeds de basis voor de sociale statistieken. Wat voor volkstelling de landen die behoren tot de United Nations Economic Commission for Europe (UNECE) hebben gehouden, is geïnventariseerd in de UNECE Task Force on Census Methodology (2013). De meeste UNECE-landen hebben de afgelopen jaren een volkstelling gehouden. Drie soorten kunnen worden onderscheiden:

- traditionele volkstellingen waarbij de informatie wordt verzameld via vragenlijsten;
- gecombineerde volkstellingen waarbij een deel van de informatie via vragenlijsten wordt verzameld en een deel uit registers komt;



Figuur 3. Volkstellingsmethoden in UNECE-landen, 2013. Bron: UNECE Task Force on Census Methodology

- registertellingen waarbij geen vragenlijsten zijn gebruikt en (vrijwel) alle informatie uit registers komt en in registers ontbrekende informatie eventueel wordt geschat uit reeds bestaande enquêtes.

In figuur 3 is de situatie op een kaart weergegeven. Deze kaart toont interessante regionale verschillen in volkstellingsmethoden. Registertellingen zijn populair in Noord-Europa, terwijl gecombineerde volkstellingen vaker in Centraal-Europa voorkomen. Traditionele volkstellingen zijn de gangbare praktijk gebleven in de Engelstalige landen en in de landen van het Gemenebest van Onafhankelijke Staten (GOS)[2]. Ook buiten de UNECE-regio is de traditionele volkstelling de norm gebleven.

Er bestaan enorme kostenverschillen tussen de verschillende soorten volkstellingen. Traditionele tellingen zijn duur. Bij het houden van een gecombineerde telling kunnen kosten worden bespaard, doordat niet alle informatie aan alle inwoners hoeft te worden gevraagd. Registertellingen zijn veel goedkoper: het is niet uitzonderlijk dat bij registertellingen de kosten nog niet één procent van de kosten van een traditionele telling bedragen.

Ondanks de verschillen in opzet en kosten zijn de resultaten van de volkstellingen steeds beter vergelijkbaar. De EU- en EVA-landen[3] hebben allemaal een Volkstelling 2011 gehouden en de resultaten zijn in te zien via de zogenaamde Census Hub (https://ec.europa.eu/CensusHub2/).

De toekomst van de volkstellingen

Het belang van internationaal vergelijkbare cijfers over de bevolking wordt steeds groter, zeker in het licht van de sterk groeiende internationale migratie. Hieraan wordt binnen EU-verband onder meer invulling gegeven door een verdere harmonisatie van de volkstellingsoutput. Daarnaast valt te verwachten dat steeds meer landen de stap zetten van traditionele volkstellingen naar gecombineerde volkstellingen of registertellingen. Ten slotte is een kritisch punt dat de Europese volkstellingsresultaten zo laat beschikbaar komen. Om dat te verbeteren worden de komende tijd in Europees verband verschillende opties besproken. Een van de opties is om na de Volkstelling 2021 jaarlijks een actualisatie te gaan publiceren van een deel van de uitkomsten.

Noten

1. De in dit artikel weergegeven opvattingen zijn die van de auteur en komen niet noodzakelijk overeen met het beleid van het Centraal Bureau voor de Statistiek (CBS). De auteur dankt Paul van der Laan van het CBS voor verbeteringen in de tekst.
2. Het GOS is in 1991 opgericht en is een samenwerkingsverband van 12 van de 15 (nl. alle behalve de Baltische) landen die vroeger de Sovjet-Unie vormden.
3. De Europese Vrijhandelsorganisatie (EVA) is in 1960 opgericht en bestaat op dit moment uit IJsland, Liechtenstein, Noorwegen en Zwitserland.

LITERATUUR

Bakker, B. F. M. (2009). *Trek alle registers open!* Rede uitgesproken bij de aanvaarding van het ambt van bijzonder hoogleraar Methodologie van registerdata voor sociaalwetenschappelijk onderzoek vanwege het Centraal Bureau voor de Statistiek bij de Faculteit der Sociale wetenschappen van de Vrije Universiteit Amsterdam op 26 november 2009.

Bakker, B. F. M., Rooijen, J. van, & Toor, L. van (2014). The System of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics. *Statistical Journal of the IAOS*, (30), 411–424.

Corbey, P. (1994). Exit the population Census. *Netherlands Official Statistics*, (9), summer, 41–44.

Cramer, M. (2005). Volkstelling: nieuwe en interessante bestanden bij het CBS. *STAtOR*, (16)3, 25–29.

Eijk, D. van (2004). Het volk tellen zonder aan te bellen. *NRC*, 28 augustus 2004.

Holvast, J. (2013). *De volkstelling van 1971, Verslag van de eerste brede maatschappelijke discussie over aantasting van privacy*. Zutphen: Uitgeverij Paris,

Schulte Nordholt, E. (2004). Introduction to the Dutch Virtual Census of 2001. In E. Schulte Nordholt, M. Hartgers & R. Gircour (Eds.)*The Dutch Virtual Census of 2001, Analysis and Methodology, Statistics Netherlands* (pp. 9–22). Voorburg/Heerlen: Statistics Netherlands, <http://www.cbs.nl/nl-NL/menu/themas/dossiers/historische-reeksen/publicaties/volkstelling-2001/2001-b57-pub.htm>

Schulte Nordholt, E., 2014a. Introduction to the Dutch Census 2011. In *Dutch Census 2011, Analysis and Methodology* (pp. 7–18) The Hague/Heerlen: Statistics Netherlands, <http://www.cbs.nl/nl-NL/menu/themas/dossiers/historische-reeksen/publicaties/publicaties/archief/2014/2014-dutch-census-2011-pub.htm>

Schulte Nordholt, E., 2014b. Key results of the Dutch Census 2011 and comparisons with earlier Dutch censuses. In *Dutch Census 2011, Analysis and Methodology* (pp. 20–32) The Hague/Heerlen: Statistics Netherlands, < http://www.cbs.nl/nl-NL/menu/themas/dossiers/historische-reeksen/publicaties/publicaties/archief/2014/2014-dutch-census-2011-pub.htm>

Willems, H. (2011). Publiek bezit. *Leeuwarder Courant*, zaterdag 24 december 2011, Kerstbijlage privé, 2-3.

UNECE Task Force on Census Methodology (2013). *Census methodology: Key results of the UNECE Survey on National Census Practices, and first proposals about the CES Recommendations for the 2020 census round.* Paper gepresenteerd op de Fifteenth Meeting of the Group of Experts on Population and Housing Censuses (30 September – 3 October 2013, Geneva) door Eric Schulte Nordholt. <http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2013/census_meeting/3_E_x_15_Aug_WEB_revised_map.pdf>

Eric Schulte Nordholt studeerde wiskunde in Utrecht en econometrie in Rotterdam. Sinds 1992 is hij werkzaam bij het Centraal Bureau voor de Statistiek. Hij was projectleider van de Volkstellingen 2001 en 2011 en is thans bezig met de voorbereidingen voor de Volkstelling 2021.
E-mail: <e.schultenordholt@cbs.nl>

Fred Steutel column

Blonde Mientje

Dick Wittenberg heeft een boek geschreven over prikkeldraad. Ik heb het niet gelezen en kan dus niet van plagiaat worden beschuldigd; of kan dat dan toch? Prikkeldraad, interessant, maar is daar ook statistisch iets aan te beleven? Jazeker, of althans iets numerieks.

Er wordt per jaar ongeveer twee miljoen kilometer prikkeldraad geproduceerd. Het spul gaat ongeveer tien jaar mee. Iedere aardbewoner heeft dus bijna 3 meter tot zijn beschikking. Niet genoeg om zijn tuintje af te schermen, maar royaal voldoende voor een meervoudige halsketting. Er bestaan inderdaad ‘prikkeldraadkettingen’, van plastic en niet zo lang.

En wat kost dat spul nou? Een paar dubbeltjes tot een euro per meter. Als je 300 meter koopt via internet is het een stuk goedkoper dan wanneer je tien meter koopt in een ijzerwinkel.

Eind achttiende eeuw werd het prikkeldraad ontwikkeld, uit gewoon draad. Het was toen bedoeld om beesten buiten te houden. Op de Amerikaanse prairies liep het vee vrij rond. Stukjes bouwland werden omheind om de beesten tegen te houden. Later werd het gebruikt om vee en mensen binnen te houden; daar weten we alles van. Wittenberg zal daar ongetwijfeld uitvoerig over berichten.

Een ander numeriek aspect is het aantal talen waarin het woord prikkeldraad heel anders klinkt dan bij ons. Een selectie van tien:

barbed wire
fil barbelé
Stacheldraht
piggtråd (Noors)
piikkilanka (Fins
filo spinato (Italiaans)
jern, stikeldriet (Fries)
stekkedraod (Twents)
hakiesdraad (Afrikaans)
pikildrato (Esperanto)

Een heel andere toepassing klonk in een lied uit de mobilisatie in 1939: ‘Blonde Mientje heeft een hart met (van) prikkeldraad’. *Prikkelpoppie*, het Afrikaanse woord voor pin-up girl, heeft daar maar heel zijdelings mee te maken.

Fred Steutel is emeritus hoogleraar kansrekening aan de TU Eindhoven.
E-mail: <fsteutel@xs4all.nl>

Hoeveel stappen zijn wereldburgers van elkaar verwijderd dankzij Facebook?*

Johan van Leeuwaarden

Ik ben hoogleraar wiskunde aan de TU Eindhoven en bestudeer grote netwerken zoals het internet. Ook geef ik met veel plezier lezingen, om te laten zien hoe mooi en krachtig wiskunde is, maar ook om nieuwe contacten te leggen met mensen die mijn liefde voor wiskunde delen. Mensen die naar mijn lezing komen, vinden hopelijk die wiskunde of wat ik doe ook interessant. Vandaar dat ik tot voor kort, voorafgaand aan een lezing, onder iedere stoel een visitekaartje legde. En na afloop vroeg ik iedereen om hun eigen visitatiekaartje. Totdat iemand me deze zomer zei dat dit wel heel ouderwets is. De meeste mensen hebben geen visitekaartjes meer op zak. Eerder maken ze contact via sociale netwerken als LinkedIn of Facebook. Daarop besloot ik om een LinkedIn-profiel aan te maken. Heel eenvoudig: je zet je foto, je baan en wat gegevens op LinkedIn en vanaf dat moment gaat het netwerken vanzelf. LinkedIn weet je familieleden, vrienden en collega's in rap tempo te vinden.

Na drie maanden actief netwerken bestond mijn LinkedIn-netwerk uit 589 connecties. Een fraai object voor een wiskundige die netwerken bestudeert. Toen ik wat rekende aan dat netwerk, bleek dat er tussen die 589 mensen maar liefst 4318 verbindingen waren. En keek je iets beter, dan zag je ook dat er duidelijke groepen te herkennen waren, mensen die onderling meer verbindingen hebben dan met de overige mensen in mijn netwerk. Zo herkende ik bijvoorbeeld mijn familieleden, oud-studiegenoten, collega's aan de TU Eindhoven, bekenden in het bedrijfsleven, wiskundigen in Nederland en wiskundigen in het buitenland.

Maar genoeg over mij en mijn haast dwangmatige behoefte om te netwerken. Ieder mens wil contacten leggen. Zo schreef de Hongaarse schrijver Frigyes Karinthy in 1929 een kort verhaal, *Schakels*, waarin de hoofdpersoon een experiment wilde doen om aan te tonen dat in 1929 de wereld sterker dan ooit daarvoor verbonden was: 'We zouden alle mensen van de 1,5 miljard mensen op aarde, wie dan ook, waar dan ook, moeten selecteren. Wedden dat er voor elk individu niet meer dan vijf mensen nodig zijn, waarvan de eerste een persoonlijke kennis is, om via kennissen elk ander individu ter wereld te bereiken!' Dus neem twee mensen. In hoeveel stappen zijn deze mensen dan, door gemeenschappelijke kennissen, verbonden? Als ze elkaar kennen is dat één stap. Als ze elkaar niet kennen, maar wel een gezamenlijke kennis hebben, dan zijn het twee stappen, en Karinthy dacht dus dat niet meer dan zes stappen nodig zouden zijn. De Amerikaanse psycholoog Stanley Milgram probeerde dit in 1967 met een experiment te onderbouwen. Enkele honderden mensen in de VS kregen de opdracht een postpakket te bezorgen bij een willekeurige persoon in Boston. Mochten ze die bewuste persoon kennen (maar dat zou wel heel toevallig zijn!) dan konden ze het postpakket persoonlijk overhandigen, maar anders moesten ze het doorsturen naar een kennis, van wie ze verwachtten dat die de persoon in Boston mogelijk wel zou kennen. En inderdaad, de pakketjes die arriveerden waren meestal minder dan zes keer verstuurd.

Dat ieder mens op deze wereld via een stuk of vijf tussenpersonen ieder ander mens kent, dus dat het een 'kleine wereld' is, werd zo een populaire volkswijsheid. En toch, het bleef lastig om dit met 'echt handen schudden' of sociale relaties te testen: men had deze gegevens niet op grote schaal beschikbaar. Maar dit veranderde door de digitalisering.

Zo is er een digitaal netwerk dat beschrijft wanneer acteurs in dezelfde film hebben gespeeld en dus elkaars handen hebben geschud. Vervolgens kun je bepalen hoe groot de afstand is tussen twee acteurs. En die afstand is, tussen bijna alle acteurs, zes stappen of minder. Voor wiskundigen is een soortgelijk netwerk gemaakt. Twee wiskundigen schudden elkaars hand als ze samen een artikel hebben geschreven. Dus toen dacht ik: 'Ik zou wel eens willen weten hoeveel stappen ik als wiskundige nou verwijderd ben van de grote Albert Einstein.' Nu kun je zeggen: 'Einstein is toch een natuurkundige en geen wiskundige?' Daarover een andere keer meer (in mijn ogen was Einstein ook een wiskundige...). Maar vast staat dat Einstein wiskunde-artikelen heeft geschreven en dus, net als ik, in dat digitale netwerk zit. En wat denk je? Ik ben slechts vier stappen verwijderd van Einstein:

1. Einstein schreef een artikel met Ernst Gabor Straus in 1945;
2. Straus schreef een artikel met László Lovász in 1971;
3. László Lovász schreef een artikel met Jan Draisma in 2012;
4. Met Jan Draisma schreef ik een artikel in 2013.

Vier stappen maar! Dat lijkt uitzonderlijk, maar het geldt voor veel paren van wiskundigen. Bekend of onbekend, we zijn doorgaans in zes stappen of minder met elkaar verbonden, zelfs als de geboortejaren een eeuw verschillen. En in mijn geval is de betrekkelijkheid van mijn relatie tot Einstein eenvoudig in te zien: Jan Draisma, mijn collega, scoort zelfs een drie en is een volle stap dichterbij.

Maar dat wat betreft wiskundigen en acteurs. Hoe zit het nu met alle mensen op de wereld? Houdt het paradigma van maximaal zes keer handen schudden ook stand als we de totale wereldbevolking in een digitaal netwerk weten te beschrijven? Wat moet er gelden, wil de hypothese überhaupt waar zijn? Nou, minstens twee dingen:

- Iedereen moet, op een of andere manier, met elkaar verbonden zijn;
- De afstanden tussen mensen moeten klein blijven.

Wiskundig kun je die twee dingen proberen te begrijpen, door een model van de wereldbevolking te maken. Het eerste model werd gemaakt door de wiskundigen Paul Erdős and Alfréd Rényi. Denk aan een tafel met knopen. Die knopen dat zijn wij, en maak je met naald en draad twee knopen aan elkaar vast, dan zijn dat vrienden. Stel je doet dit volstrekt willekeurig: met de ogen dicht pak je steeds twee knopen en je verbindt deze met een draadje. Dan blijkt dat met heel weinig draadjes, ongeveer half zo veel draadjes als knopen, vrijwel alle knopen aan elkaar vast komen te zitten. En het maakt niet veel uit of je dit experiment met 70, 7000 of 7 miljard knopen doet. Dus, we zien dat door een gering aantal, vrij willekeurige verbindingen, alles verbonden raakt. Bovendien wisten wiskundigen ook uit te rekenen wat de gemiddelde afstand voor het model was en die bleek heel klein te zijn, slechts een paar stappen, zelfs voor een netwerk met 7 miljard knopen.

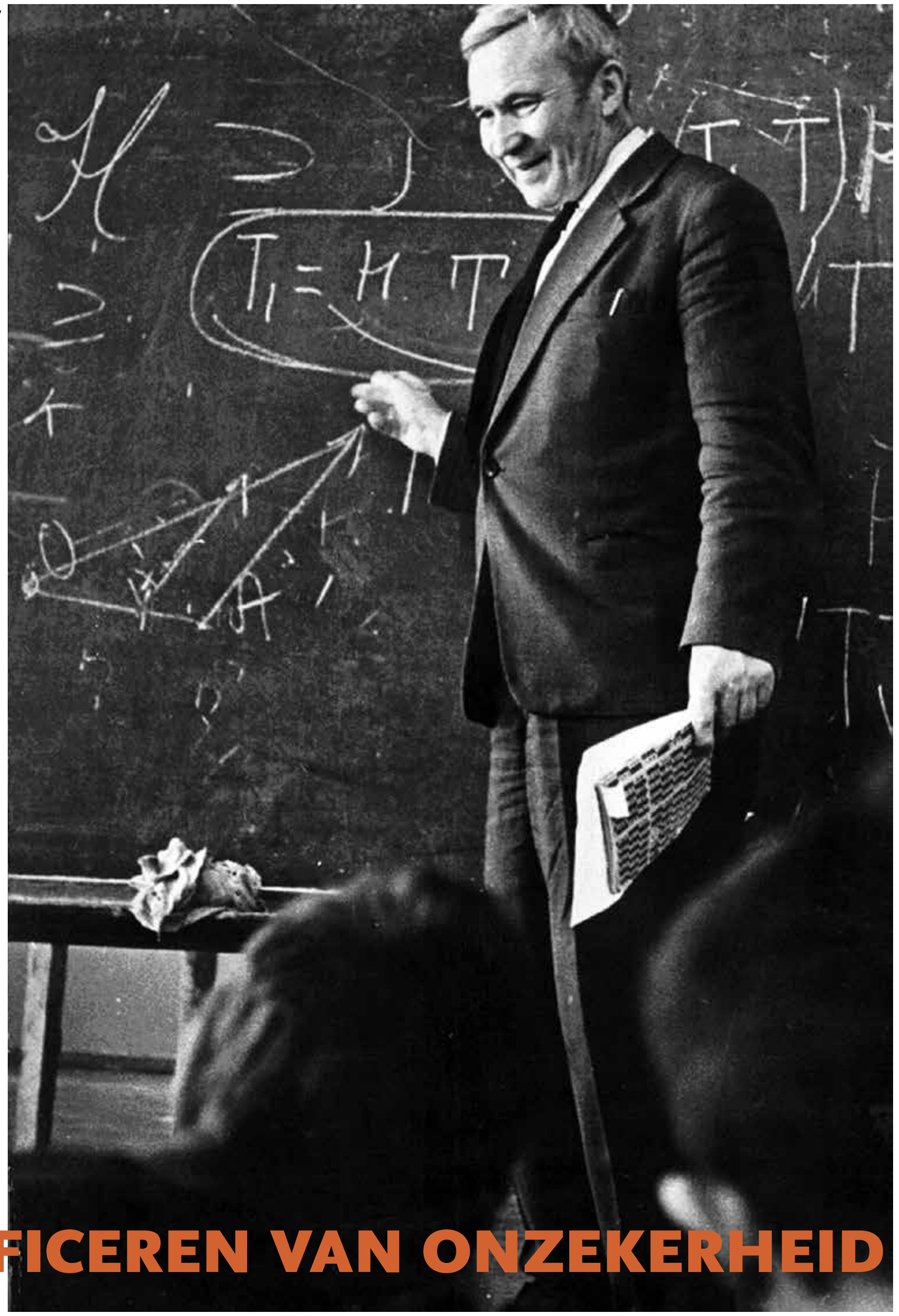
Dus we hebben de hypothese van Karinthy, het (kleine) experiment van Milgram, proeven op de som met acteurs en wiskundigen, en een wiskundig model, alle wijzend in de richting van een wereld met grote verbondenheid en kleine afstanden. En toch, zou de hypothese ook stand houden voor de echte populatie, de wereldbevolking? Waar je het vroeger moest doen met een gedachtenoefening of een klein experiment, is er nu veel data beschikbaar en hebben we enorme netwerken in beeld via het internet; netwerken als LinkedIn en Facebook. Ik heb al uitgelegd hoe mijn LinkedIn-netwerk eruit ziet, maar hoe zit het dan als je al die netwerken, van iedereen, bij elkaar neemt? We kunnen aan deze enorme netwerken rekenen. En Facebook is het grootst met 1,5 miljard gebruikers, de totale wereldbevolking in de tijd van Karinthy.

We nemen dus Facebook en bekijken alle mogelijke paren van mensen. Een hele klus, want dat aantal paren is ongeveer het aantal mensen in het kwadraat gedeeld door twee. Na lange berekeningen gedaan door Facebook in 2011 bleek dat minder dan 1% van de mensen op Facebook niet verbonden is met de overige mensen. Met andere woorden: neem twee willekeurige mensen op Facebook en de kans dat ze elkaar via een pad kunnen vinden over Facebook is vrijwel 100%. Na nog langere berekeningen bleek dat vrijwel alle gebruikers verbonden zijn in 5 stappen en 92% van alle gebruikers in slechts 4 stappen. We zullen nooit weten of het al waar was in 1929, maar door netwerken als Facebook en bruut rekenwerk kunnen we de mythe van maximaal zes stappen in de huidige tijd bevestigen. Netwerken verbinden iedereen en de al kleine afstanden worden steeds kleiner. De wereld krimpt!

* Deze tekst is gebaseerd op een van de vijf colleges die Johan van Leeuwaarden heeft gegeven bij de Universiteit van Nederland. Deze colleges van 15 minuten elk zijn te zien op <www.universiteitvannederland.nl>.

Johan van Leeuwaarden is hoogleraar wiskunde aan de TU Eindhoven en lid van De Jonge Akademie van de KNAW.
E-mail: <j.s.h.v.leeuwaarden@TUE.nl>

Andrey Kolmogorov
(Tambov, 1903 – Moskou, 1987)



KWANTIFICEREN VAN ONZEKERHEID MET BELIEF-FUNCTIES

Ronald Meester & Timber Kerkvliet

Zodra we onzekerheid willen kwantificeren of onder onzekerheid willen redeneren, grijpen we naar de kansrekening, de subdiscipline van de wiskunde die hiervoor is ontwikkeld. De moderne kansrekening is vrijwel uitsluitend gebaseerd op de axioma's van Kolmogorov.

De belangrijkste van die axioma's is ongetwijfeld de regel dat de kans op de vereniging van twee disjuncte gebeurtenissen gelijk is aan de som van de kansen op die gebeurtenissen. Deze regel is goed te motiveren met bijvoorbeeld een frequentistische kijk op de kansrekening, waarbij de kans op een gebeurtenis gelijkgesteld wordt aan de frequentie waarmee deze optreedt bij vele herhalingen van het experiment.

Het is echter een feit dat veel situaties waarin onzekerheid een rol speelt, een dergelijke herhaalbaarheid helemaal geen betekenis heeft. Bij een rechtszaak, waarbij de rechter iets moet zeggen over de vraag of verdachte wel of niet schuldig is, hebben we geen mogelijkheid om een hele serie identieke situaties te scheppen, en is een andere interpretatie nodig wanneer we over de kans op schuld zouden willen spreken. Ook zijn er situaties waarin herhaalbaarheid wel betekenis heeft, maar het helemaal niet duidelijk is of er voor alle gebeurtenissen een eenduidige frequentie bestaat waarmee ze optreden bij vele herhalingen.

In dit soort situaties wordt vaak de Bayesiaanse interpretatie van kansen gehanteerd. In deze interpretatie is de kans op een gebeurtenis, nu een hypothese genoemd, een kwantificering van iemands geloof in de waarheid van de betreffende gebeurtenis/hypothese. In deze interpretatie kan men nog steeds de axioma's van Kolmogorov motiveren, bijvoorbeeld via een Dutch Book Argument.

Het probleem is dat we in veel gevallen niet geïnteresseerd zijn in een kwantificering van *waarheid*, maar in een kwantificering van onze *kennis*. Dit is bijvoorbeeld het geval in de rechtszaal, waar de vraag niet is of een verdachte schuldig of onschuldig is, maar de vraag is of er (wettig en overtuigend) bewijs is dat een verdachte schuldig is. We geven twee voorbeelden waaruit blijkt dat de axioma's van Kolmogorov te rigide zijn indien we een kans willen interpreteren als de mate van bewijsbaarheid.

Als een getuige zich heeft vergist en per ongeluk ten onrechte iemand een alibi heeft verschaft, valt dat bewijs voor onschuld van de dader weg. Het niet meer hebben van een alibi betekent echter niet dat er meer bewijs is voor schuld van de dader. Echter, als we kansen hebben die de axioma's van Kolmogorov volgen, en de kans op onschuld wordt kleiner, dan wordt de kans op schuld groter.

Een ander probleem met de axioma's van Kolmogorov is dat we er geen totaal gebrek aan informatie mee kunnen uitdrukken. In een recente rechtszaak bijvoorbeeld was er geen onzekerheid dat roekeloos rijgedrag van een bepaalde auto tot de dood van een kind had geleid, alleen was er wel volledige onzekerheid welke van de twee inzittenden de auto had bestuurd. We zouden dus willen dat de bewijsbaarheid van beide opties 0 is. De klassieke kansrekening verhindert dit, omdat kansen moeten optellen tot 1. Bij gebrek aan een beter alternatief, wordt vaak aan beide opties kans 1/2 toegekend, maar feitelijk impliceert dit meer bewijsmateriaal dan we hebben. In de literatuur wordt de abstractie van dit voorbeeld het *eilandprobleem* genoemd (zie Slooten & Meester, 2011). We zullen de nieuwe theorie straks op dit eilandprobleem toepassen.

Al deze problemen kunnen geadresseerd worden door de axioma's van Kolmogorov flexibeler te maken wanneer we het begrip kans willen interpreteren als een kwantificering van kennis. Om verwarring te voorkomen noemen we het dan geen kans meer, maar een *belief-functie*. Deze term is geïntroduceerd door Glenn Shafer (1976, 1981). Lange tijd is zijn voorstel in de mainstream kansrekening niet serieus genomen, maar dat kwam ons inziens vooral door de gebrekkige calculus die Shafer had ontwikkeld. Onze nieuwe calculus ondervangt dit probleem en maakt belief-functies heel werkbaar en toepasbaar.

Een beetje theorie

Definitie 1

Een functie $m : 2^\Omega \to [0,1]$ heet een *basic belief assignment* als

$$m(\emptyset) = 0 \quad \text{en} \quad \sum_{C \subseteq \Omega} m(C) = 1. \tag{1}$$

Het getal $m(C)$ representeert de kans of ons vertrouwen in een uitkomst die in C ligt, zonder extra informatie over welke uitkomst het precies zal zijn. Als we bijvoorbeeld geen enkele informatie hebben behalve dat de uitkomst in een verzameling Ω ligt, dan kunnen we $m(\Omega) = 1$ en $m(A) = 0$ voor elke strikte deelverzameling van Ω nemen.

De bij m horende *belief-functie* wordt nu als volgt – definitie 2 – gedefinieerd.

Definitie 2

Gegeven een basic belief assignment m definiëren we de corresponderende *belief-functie* Bel als

$$\mathrm{Bel}(A) := \sum_{C \subseteq A} m(C). \tag{2}$$

Ons belief in A is het totale gewicht van de basic belief assignment van die verzamelingen waarvan zeker is dat ze A impliceren. Een andere manier om belief-functies in te voeren volgt uit de volgende stelling

Stelling 1

Een functie Bel is een belief-functie dan en slechts dan als

$$\mathrm{Bel}\left(\bigcup_i A_i\right) \geq \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ I \neq \emptyset}} (-1)^{|I|+1} \mathrm{Bel}\left(\bigcap_{i \in I} A_i\right). \tag{3}$$

Deze stelling zegt dat belief-functies gekarakteriseerd worden door een inclusie-exclusie formule met een ongelijkheid in plaats van de klassieke gelijkheid. Hieruit zien we onmiddellijk dat elke klassieke kansverdeling natuurlijk een belief-functie is. Dat hadden we ook eerder kunnen inzien: zodra een basic belief assignment m alleen maar positieve massa toekent aan singletons, is de bijbehorende Bel een kansverdeling.

We kunnen (3) misschien beter begrijpen door naar het geval $n = 2$ te kijken:

$$P(A \cup B) \geq P(A) + P(B) - P(A \cap B).$$

Bewijs dat naar $A \cup B$ wijst hoeft geen bewijs te zijn dat naar A of B individueel wijst, zie het eerste voorbeeld hierboven. Bewijs dat naar A wijst, wijst ook naar $A \cup B$, en evenzo voor B, maar aan de rechterkant moeten we dan natuurlijk bewijs dat naar zowel A als B wijst niet dubbel tellen. Deze overwegingen maken de termen en de ongelijkheid heel natuurlijk.

Om met belief-functies te kunnen werken moeten we weten hoe belief zich gedraagt wanneer we conditioneren op een gebeurtenis. Stel we hebben een basic belief assignment m met corresponderende belief-functie Bel, en stel dat we willen conditioneren op gebeurtenis H. Het gewicht $m(A)$ wordt nu toegekend aan $A \cap H$, als $A \cap H \neq \emptyset$. Daarna herschalen we weer zodat de nieuwe conditionele belief-functie, genoteerd met Bel_H, weer totale basic belief massa 1 heeft. Deze regel is eenvoudig toe te passen, maar in dit artikel gaan we niet echt rekenen.

Een voorbeeld

We geven ter illustratie een voorbeeld (voor meer voorbeelden zie Kerkvliet & Meester, 2016a, 2016b). Stel we weten dat een van twee personen, A en B, een misdrijf heeft gepleegd. We schrijven E voor de gebeurtenis dat de dader en A allebei een bepaald DNA profiel hebben dat met frequentie p voorkomt in de populatie. We schrijven G voor de gebeurtenis dat A de dader is.

In de reguliere Bayesiaanse aanpak, kunnen we de regel van Bayes

$$P(G|E) = \frac{P(E|G)P(G)}{P(E|G)P(G) + P(E|G^c)P(G^c)} \tag{4}$$

gebruiken om de conditionele kans op G gegeven E uit te rekenen, waarbij men dan bij gebrek aan beter $P(G) = P(G^c) = 1/2$ kiest. Dit geeft

$$P(G|E) = \frac{\frac{1}{2}p}{\frac{1}{2}p^2 + \frac{1}{2}p} = \frac{1}{1+p}. \tag{5}$$

Omdat onze theorie de klassieke generaliseert, kunnen wij dit antwoord ook verkrijgen met behulp van belief-functies. We nemen dan de volgende basic belief assigment:

$$\begin{aligned}
m(E \cap G) &= m\left(\begin{array}{|c|c|c|}\hline & G^c & G \\ \hline E^c & & \\ \hline E & & * \\ \hline\end{array}\right) = \frac{1}{2}p, \\
m(E^c \cap G) &= m\left(\begin{array}{|c|c|c|}\hline & G^c & G \\ \hline E^c & & * \\ \hline E & & \\ \hline\end{array}\right) = \frac{1}{2}(1-p), \\
m(E \cap G^c) &= m\left(\begin{array}{|c|c|c|}\hline & G^c & G \\ \hline E^c & & \\ \hline E & * & \\ \hline\end{array}\right) = \frac{1}{2}p^2, \\
m(E^c \cap G^c) &= m\left(\begin{array}{|c|c|c|}\hline & G^c & G \\ \hline E^c & * & \\ \hline E & & \\ \hline\end{array}\right) = \frac{1}{2}(1-p^2),
\end{aligned} \tag{6}$$

en vinden we dat

$$\mathrm{Bel}_E(G) = \frac{\frac{1}{2}p}{\frac{1}{2}p^2 + \frac{1}{2}p} = \frac{1}{1+p}. \tag{7}$$

Maar in onze nieuwe theorie kunnen we precies gebruiken wat we wel en niet weten. Als A het profiel heeft, hetgeen met kans $1 - p$ gebeurt, dan weten we E^c maar we weten nog niets over de (on)schuld van A. Als A en B allebei het profiel hebben, hetgeen met kans p^2 gebeurt, dan weten we E maar we weten opnieuw niets over de (on)schuld van A. Als A het profiel heeft maar B niet, met bijbehorende kans $p(1 - p)$, dan weten we dat E^c of G waar is. Dit leidt dan tot de volgende basic belief assignment:

$$\begin{aligned}
m(E^c) &= m\left(\begin{array}{|c|c|c|}\hline & G^c & G \\ \hline E^c & * & * \\ \hline E & & \\ \hline\end{array}\right) = 1-p, \\
m(E) &= m\left(\begin{array}{|c|c|c|}\hline & G^c & G \\ \hline E^c & & \\ \hline E & * & * \\ \hline\end{array}\right) = p^2, \\
m(E^c \cup G) &= m\left(\begin{array}{|c|c|c|}\hline & G^c & G \\ \hline E^c & * & * \\ \hline E & & * \\ \hline\end{array}\right) = p(1-p).
\end{aligned} \tag{8}$$

Merk op dat $\mathrm{Bel}(G) = \mathrm{Bel}(G^c) = 0$, dus deze belief-functie geeft geen enkel belief in schuld of onschuld van A, zoals het ook hoort, want daar weten we niets van. Als we nu conditioneren op E, dan leert een korte berekening dat

$$\mathrm{Bel}_E(G) = \frac{p(1-p)}{p(1-p)+p^2} = 1-p, \tag{9}$$

en deze uitkomst is kleiner dan de klassieke uitkomst, hetgeen we mogen verwachten.

Hopelijk hebben we de lezer met dit voorbeeld nieuwsgierig gemaakt naar verdere voorbeelden en toepassingen. Hiervoor verwijzen we naar de al eerder genoemde referenties.

Literatuur

Kerkvliet, T., & Meester, R. (2016a). *Assessing forensic evidence by computing belief functions – theory and applications.* Submitted, http://arxiv.org/abs/1512.01250.

Kerkvliet, T., & Meester, R. (2016b). *Quantifying knowledge with a new calculus for belief functions – a generalisation of probability theory.* Submitted, http://arxiv.org/abs/1512.01249.

Kerkvliet, T., & Meester, R. (2016c). *New axioms for epistemic uncertainty*. To appear.

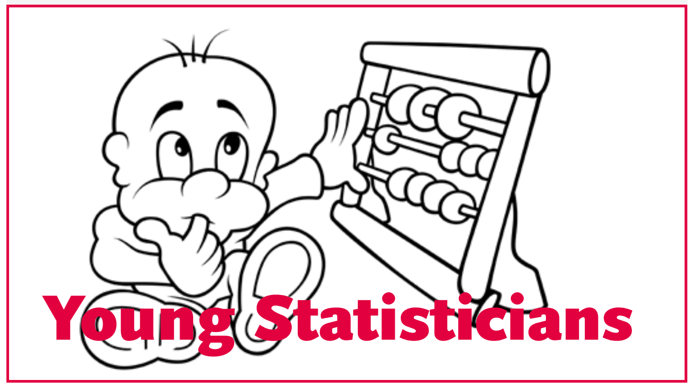
Shafer, G. (1976). *A mathematical theory of evidence*. Princeton: University Press Princeton.

Shafer, G. (1981). Constructive probability. *Synthese*, 48(1), 1–60.

Slooten, K., & Meester, R. (2011) Forensic identification: the island problem and its generalisations. *Statistica Neerlandica*, 65, 202.

Ronald Meester is hoogleraar aan de faculteit der exacte wetenschappen van de VU Amsterdam.
E-mail: <r.w.j.meester@vu.nl>

Timber Kerkvliet is PhD student aan de faculteit der exacte wetenschappen van de VU Amsterdam.
E-mail: <t.kerkvliet@vu.nl>

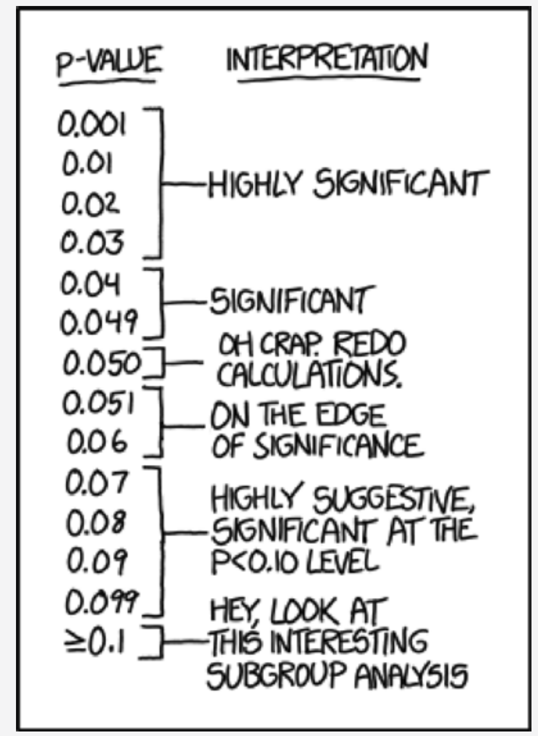


UPCOMING ACTIVITIES

It has been a while since the last activity of the Young Statisticians. Our apologies for that. To bring Young Statisticians back to life we have organized the following activities.

➔April 20th: Science Cafe: The (mis)use of p-values
➔May 30th: Workshop @ Rabobank, Utrecht

The Science Cafe's subject is the use and misuse of p-values in research. The speakers are prof. Peter Grünwald and prof. Eric-Jan Wagenmakers.



More information about the activities will follow via the mailing list and on our Facebook page. If you have any suggestions for activities, or if you want to organize something yourself, please let us know! We hope to see many of you at our activities!

Best, Nynke, Tim and Sanne
Young Statisticians Board
info@youngstatisticians.nl

RK Kerkhof in Reek (rijksmonument). Foto: Havang(nl)

DOODSOORZAKENSTATISTIEK
een oude boom met groene twijgen

Jan Kardaun

De doodsoorzakenstatistiek is een van de oudste van overheidswege bijgehouden statistieken in Nederland, en gaat terug tot de tweede helft van de negentiende eeuw. De opkomst van de staatsbemoeienis met de gezondheidszorg, onder invloed van Thorbecke, resulteerde ondermeer in de Wet Uitoefening der Geneeskunst, 1865 en de Begrafeniswet, 1868. Dit viel samen met een stroming om allerlei natuurlijke en maatschappelijke verschijnselen te bestuderen op kwantitatieve grondslag. Deze belangstelling strekte zich ook uit tot doodsoorzaken: op het Eerste en Tweede Internationale Statistische Congres (Brussel 1853, Parijs 1855) werd een lijst van doodsoorzaken opgesteld, maar niet algemeen aanvaard.

Pril begin, in het gaslichttijdperk

Doodsoorzaakgegevens, in eenvoudige en beschrijvende vorm, werden in Nederland aanvankelijk lokaal verwerkt, met het oog op snelle actie en beleidsmaatregelen in gevallen van epidemieën en problemen met drinkwatervoorziening. In enkele reorganisatiegolven (toen ook al) werd de verwerking landelijk gecentraliseerd, en vanaf 1901 aan het (kersverse) CBS toevertrouwd. Het CBS was in 1899 opgericht, en bestond aanvankelijk uit een Directeur en twaalf 'aan hem toevertrouwde ambtenaren', zodat ongetwijfeld in de begintijd het héle CBS goed op de hoogte was van het reilen en zeilen van de doodsoorzakenstatistiek.

Inmiddels was door het Internationaal Statistisch Instituut (ISI), opvolger van het Internationale Statistische Congres, een latere variant (de Bertillon-lijst) van genoemde doodsoorzakenlijst voorgedragen. Deze lijst werd in 1900 aangenomen door 26 landen, waaronder Nederland: de eerste versie van de ICD. Het ISI heeft de ICD (International Classification of Causes of Death) bijgehouden tot 1948, het jaar waarin de WHO (World Health Organization) werd opgericht. Momenteel is de tiende editie van de ICD in gebruik, inmiddels omgedoopt tot International Statistical Classification of Diseases and Health Related Problems. De WHO verzamelt en onderhoudt de WHO Mortality Database, waaraan 156 'landen' doodsoorzaakgegevens hebben geleverd (sinds 1950). 'Landen' staat hier tussen aanhalingstekens omdat ook er sommige administratieve gebieden zijn die geen land zijn, en sommige staten uit meer landen bestaan. Sinds 2000 hebben 138 'landen' gegevens geleverd, want niet alleen mensen, maar ook doodsoorzakenstatistieken komen en gaan, en zelfs staten komen en gaan.

De redenen waarom de doodsoorzakenstatistiek toen, een eeuw geleden, zo aantrekkelijk was voor 26 landen, gaan nog steeds op voor een groot aantal landen die in de huidige tijd in dezelfde situatie verkeren als de 26 landen van toen, namelijk beperkte beschikbaarheid van gegevens en infrastructuur. De doodsoorzakenregistratie is dan vaak de enige gezondheidsregistratie die landelijk dekkend is en over een welgedefinieerde gebeurtenis en over ernstige gezondheidsproblemen gaat. In landen zoals Nederland, waar een veelheid aan gezondheidsregistraties tot ontwikkeling is gekomen, is de rol van de doodsoorzaakregistratie meer die van een onderdeel van een gezondheidsstatistisch geheel geworden, en worden de doodsoorzaakfrequenties meestal in verband met andere maten gebruikt.

De doodsoorzakenclassificatie

De verzameling doodsoorzaken (kijk maar eens in de ICD, of op StatLine) is altijd een mélange geweest (sommigen beweren: een ratjetoe) van infectieziekten, orgaanaandoeningen, kankers, complicaties en 'incidenten' zoals geweld, ongeval, misdrijf. In de laatste decennia is er een verschuiving geweest van – overdreven geformuleerd – acute doodsoorzaken naar chronische doodsoorzaken. De eerste slaat op situaties waarin je min of meer gezond bent, maar een paar weken later overleden, aan bijvoorbeeld een hartinfarct, beroerte of infectieziekte. Doordat de gezondheidszorg er steeds beter in slaagt mensen met zo'n acute aandoening in leven te houden, gaan steeds meer mensen (op latere leeftijd) dood terwijl ze een aantal chronische aandoeningen hebben, die allemaal statistisch gezien het leven bekorten, maar waaraan je niet onmiddellijk doodgaat. Het is dan ook minder duidelijk waaraan je doodgaat, want vaak maakt de combinatie van aandoeningen dat je in een slechte conditie komt, kwetsbaar voor allerlei bijkomende ziekten. Het begrip afzonderlijke, onderliggende doodsoorzaak is in die situaties minder van toepassing. De huidige doodsoorzakenstatistiek (versie 10 van de ICD is ontworpen eind jaren tachtig) is echter nog steeds sterk georiënteerd op het begrip afzonderlijke, onderliggende doodsoorzaak, dat rechtstreeks afstamt van de eerste ICD uit 1900.

Korte en lange termijneffecten, clusters

De toepassingen in de beginperiode van de doodsoorzakenstatistiek lagen vooral in de snelle verwerking van de gegevens en onmiddellijke maatregelen. In de huidige geautomatiseerde tijd zou je verwachten dat het nog sneller kan en gaat, maar de voornaamste toepassing ligt in de wat langere termijntrends. Een belangrijke reden hiervoor is dat voor de snelle reactie andere mechanismen en organisatievormen (GGD-en, microbiologische laboratoria) zijn ontwikkeld om te waarschuwen voor bijvoorbeeld infectieziekten, en wel in het stadium dat de mensen ziek worden, niet pas nadat ze zijn overleden. Doodsoorzaakgegevens spelen daarbij slechts een aanvullende rol, en verlopen in dreigende gevallen (denk bijvoorbeeld aan de multi-resistente E-coli-epidemie in mei 2011) via een ander, sneller kanaal dan de doodsoorzaakopgaven voor de statistiek.

De sterke punten van de doodsoorzakenstatistiek (gegevens over *iedereen*, lange observatieperiode, in veel landen op dezelfde wijze waargenomen) komen tegenwoordig het best tot hun recht bij de bestudering van lange termijnontwikkelingen. Hiermee wordt bedoeld: perioden van enkele jaren tot enkele decennia. Dat stelt overigens extra eisen aan de stabiliteit van het meetinstrument (zie onder). Eerst echter enkele woorden over toepassingen van de doodsoorzakenstatistiek, ook in de huidige tijd, voor opflakkerende doodsoorzaken (clusters), en andere niet-periodieke verschijnselen. De dagelijkse, wekelijkse, maandelijkse en de kwartaalsterfte (nog zonder onderscheid naar doodsoorzaak) vertonen altijd schommelingen waarvoor maar in beperkte mate een verklaring te vinden is. Deze ruwe sterftegegevens zijn immers de (theoretische) som van lange-termijn ontwikkelingen, 'echte' korte-termijneffecten, stochastische ruis, meetfouten en andere artefacten. Het lukt niet altijd deze theoretische componenten a posteriori uit elkaar te rafelen, vooral niet wanneer het vermoede korte termijneffect eenmalig is. Herhaling van metingen van doodsoorzaakgegevens in dezelfde periode en regio zijn ofwel onmogelijk, ofwel zeer bewerkelijk en begrotelijk, zodat meestal met wat aannamen gegevens gebruikt worden van eerdere jaren of van andere landen of van regio's binnen Nederland om de opsplitsing in bovengenoemde componenten te onderbouwen.

Vaak moeten gegevens van enkele jaren worden samengevoegd en gemiddeld om het aandeel van de stochatische ruis te verminderen (onder de aanname dat deze jaren vergelijkbaar waren met betrekking tot de onderzochte doodsoorzaken). Veel sterk tot de verbeelding sprekende doodsoorzaken (moord en doodslag, verkeersongevallen, jeugdkanker) komen (gelukkig) niet zo heel veel voor. Bij kleine aantallen waarnemingen kun je als vanzelf percentueel grote veranderingen van jaar tot jaar aantreffen. Dan is het relativerend om bij de jaarlijkse verschillen de Poissonverdeling te gebruiken als maatstaf, om de nul-hypothese te toetsen dat de veranderingen heel goed toevallig kunnen zijn. Dat is niets nieuws. Al in 1898 liet Von Bortkiewicz zien dat jaarreeksen dodelijke ongevallen door een trap van een paard in het Pruisische leger (en enkele andere voorbeelden) geheel binnen de Poissonverdeling lagen. Die aanpak is nog steeds erg nuttig, al zouden we ons nu afvragen of zo'n trap van een paard in de categorie arbeidsongeval (bedrijfsongeval) of verkeersongeval zou thuishoren.

Stabiliteit

Als je een tijdreeks van enkele jaren of decennia bestudeert, zou je het liefst willen dat de resultaten niet beinvloed zijn door veranderingen in het meetinstrument, het geheel van de wijze van gegevensverzameling en verwerking. Dat is in absolute zin niet mogelijk en zelfs niet wenselijk. Het zou betekenen dat de doodsoorzaken nog hetzelfde ingedeeld worden als voor de wereldoorlog (u mag zelf kiezen of het de eerste of de tweede zal zijn). Er zijn op den duur veranderingen in medisch inzicht en terminologie, en hoewel er weinig echt nieuwe ziekten bijkomen, worden wel andere eenheden als ziekte herkend en benoemd. En met de verbeterde diagnostische mogelijkheden kunnen ook verfijningen in diagnostiek een plaats krijgen in de doodsoorzaakclassificatie. In het algemeen echter gaat het om diagnostische vernieuwingen die voor de behandeling van ziekten wel van belang zijn, maar die niet zo op de voorgrond treden als doodsoorzaak, zodat de doodsoorzakenclassificatie nogal conservatief is (en terecht kán zijn).

De eerste tien edities van de ICD zijn met tussenpozen van ongeveer tien jaar verschenen. De huidige editie is in Nederland al sinds 1996 in gebruik. De introductie van een nieuwe ICD brengt natuurlijk discontinuïteiten met zich mee in de statistiek, die als het een beetje meezit op globaler niveau weer wegvallen. Maar er zijn ook andere momenten – bijvoorbeeld bij wisseling van personeel, ander verwerkingsproces, andere software – waarop (onbedoelde) verschuivingen in uitkomsten kunnen binnensluipen. In het algemeen zal men bij de doodsoorzakenstatistiek een conservatief beleid voeren: streven naar de grootst haalbare continuïteit. De keerzijde daarvan is dat je niet even een paar leuke ideeën bij de vervaardiging van deze statistiek kunt uitproberen. Dat is overigens toch al geen goed idee. Het is geen snel proces om een verandering in het gehele land, bij duizenden artsen, in te voeren, en zoiets wil je zeker niet tijdelijk doen.

Reproduceerbaarheid

Naast stabiliteit wordt natuurlijk van het meetinstrument verwacht dat het reproduceerbare uitkomsten geeft. Hoewel – zoals reeds aangestipt – echte herhaling van metingen (zoals bij laboratoriumexperimenten) niet mogelijk is, kunnen sommige onderdelen van het verwerkingsproces wel op reproduceerbaarheid onderzocht worden. Een van de kritieke punten is de reproduceerbaarheid van het *coderen* (ook wel classificeren genoemd) van een doodsoorzaakomschrijving in een ICD-code, en het daarbij *selecteren* van de onderliggende doodsoorzaak (in geval er meer dan één ziekte is aangegeven op de doodsoorzaakopgave). De eerstgenoemde stap (coderen in engere zin) is daarbij makkelijker, eenduidiger en dus reproduceerbaarder dan het selecteren van de onderliggende doodsoorzaak. Bij de laatste stap komt ook interpretatie te pas. Omdat er steeds meer mensen overlijden aan een keten of een combinatie van ziekten, en om die interpretatie zo goed mogelijk te objectiveren, kent de ICD sinds de tiende editie een speciale handleiding, deel 2 van de driedelige uitgave. Dit is geen lichte kost, maar wie doodsoorzaakgegevens wil analyseren moet er toch wel degelijk kennis van nemen. En wie een doodsoorzakenstatistiek wil produceren moet het op het nachtkastje hebben liggen.

De mate waarin individuele doodsoorzaakopgaven reproduceerbaar gecodeerd worden varieert met het soort aandoening en de leeftijd van de overledene, maar veel van deze variabiliteit middelt uit op groepsniveau (dus in de statistische tabellen op StatLine). Nederland zat in 2009 op gelijk kwaliteitsniveau als andere landen die de codering geheel handmatig uitvoerden. Sindsdien is het CBS overgestapt – voor het eerst voor het jaar 2013 – op een werkwijze waarin het overgrote deel van de gevallen de codering via software (algoritmen en tabellen, zie IRIS en MMDS bij de referenties) geschiedt, in navolging van de USA en een handvol Europese landen. Dit komt de reproduceerbaarheid van de gegevens en de internationale vergelijkbaarheid zeer ten goede.

Tot slot

In het bovenstaande is in vogelvlucht weergegeven hoe de doodsoorzakenstatistiek was en is, en ook in belangrijke mate zal zijn. Onderwijl is het nodig om, onder behoud van het verworvene, te werken aan een modernisering van het begrip onderliggende doodsoorzaak. Niet alleen omdat, zoals reeds aangegeven, steeds meer mensen aan een combinatie van aandoeningen lijden, die samen tot de dood leiden, maar ook omdat ons oorzaakbegrip in de vele verstreken decennia geleidelijk is veranderd. We weten veel meer van onstaansmechanismen van ziekten en zijn ook meer gaan denken in 'statistische oorzaken'. Dan is het niet nodig dat op oorzaak A telkens binnen korte tijd gevolg B optreedt, maar het volstaat dat geruime tijd na A op groepsniveau vaker B gezien wordt. De grens tussen (harde) oorzaken, waarover je in het individuele geval vaak wel een uitspraak kunt doen en 'statistische' oorzaken, die alleen op groepsniveau aan te tonen zij, is al vervaagd, en deze vervaging gaat nog verder als risicofactoren, die bijvoorbeeld een kleine sterftekans 20 procent hoger maken ook als oorzaak aangeduid worden. Een oorzaak op groepsniveau heeft een heel andere interpretatie dan een op individueel niveau. Bij het (achteraf) invullen van een doodoorzakenopgave doet de arts een uitspraak over een oorzaak in een individuele geval en bij het maken van de statistiek moeten we ernaar streven – indien er 'concurrentie' is van meerdere oorzaken – dat in ieder geval die oorzaken goed in kaart worden gebracht die de beste mogelijkheden tot interventie of preventie bieden.

Literatuur

World Health Organization (1992–1994). *ICD-10: International statistical classification of diseases and related health problems.* (Volume I: tabular lists; Volume II: description, guidelines and rules; Volume III: alphabetical index) Geneva: WHO, <http://www.who.int/classifications/icd/en/>, <http://www.rivm.nl/who-fic/ICD.htm>.

CBS. *Doodsoorzaken in StatLine.* <http://statline.cbs.nl/Statweb/selection/?DM=SLNL&PA=7233&VW=T>.

Stabiliteit: R.H. van der Stegen, L.S. Koren, P.P. Harteloh, J.W. Kardaun, F. Janssen. A Novel Time Series Approach to Bridge Coding Changes with a Consistent Solution Across Causes of Death. *Eur J Popul.* 2014;**30**:317-335.

Reproduceerbaarheid: P. Harteloh, K. de Bruin, J. Kardaun. The reliability of cause-of-death coding in The Netherlands. *Eur J Epidemiol.* 2010;**25**:531-538.

Clusters: L. von Bortkewitsch: *Das Gesetz der kleinen Zahlen.* Leipzig: B.G. Teubner, 1898.

WHO Mortality database <http://www.who.int/healthinfo/mortality_data/en/>.

IRIS <http://www.dimdi.de/static/en/klassi/irisinstitute/index.htm>.

MMDS <http://www.cdc.gov/nchs/nvss/mmds/about_mmds.htm>.

Jan Kardaun schreef deze bijdrage op persoonlijke titel. Hij is medisch ambtenaar van het CBS en tevens bijzonder hoogleraar registratie en statistiek van doodsoorzaken bij het AMC.
E-mail: <jwpf.kardaun@cbs.nl>

Foto: Delta

Betrouwbaarheid van energienetwerken bij onzekerheid in aanbod en vraag

Ben Zwart, winnaar Van Dantzigprijs: 'De komende jaren wil ik me richten op de wiskundige analyse van zeldzame gebeurtenissen en de toepassingen hiervan op energienetwerken. Ik ben erg enthousiast over dit onderwerp: het komt niet zo vaak voor dat een onderzoeksgebied zich presenteert met interessante wiskundige vraagstukken waarvan de oplossing een belangrijke maatschappelijke impact kan hebben.'

Bert Zwart

Ondanks de komst van big data blijven zeldzame gebeurtenissen zeldzaam, en blijven er geavanceerde wiskundige methoden en technieken nodig om inzicht te verkrijgen. Binnen de stochastische operations research zijn dergelijke technieken al met succes toegepast in de financiële en verzekeringswiskunde, en computer-communicatienetwerken. Hoewel die onderwerpen relevante inspiratiebronnen blijven, is er een nieuw toepassingsgebied in opkomst met nieuwe en urgente onderzoeksvragen: dat van energienetwerken, in het bijzonder elektriciteitsnetten.

Uitdagingen

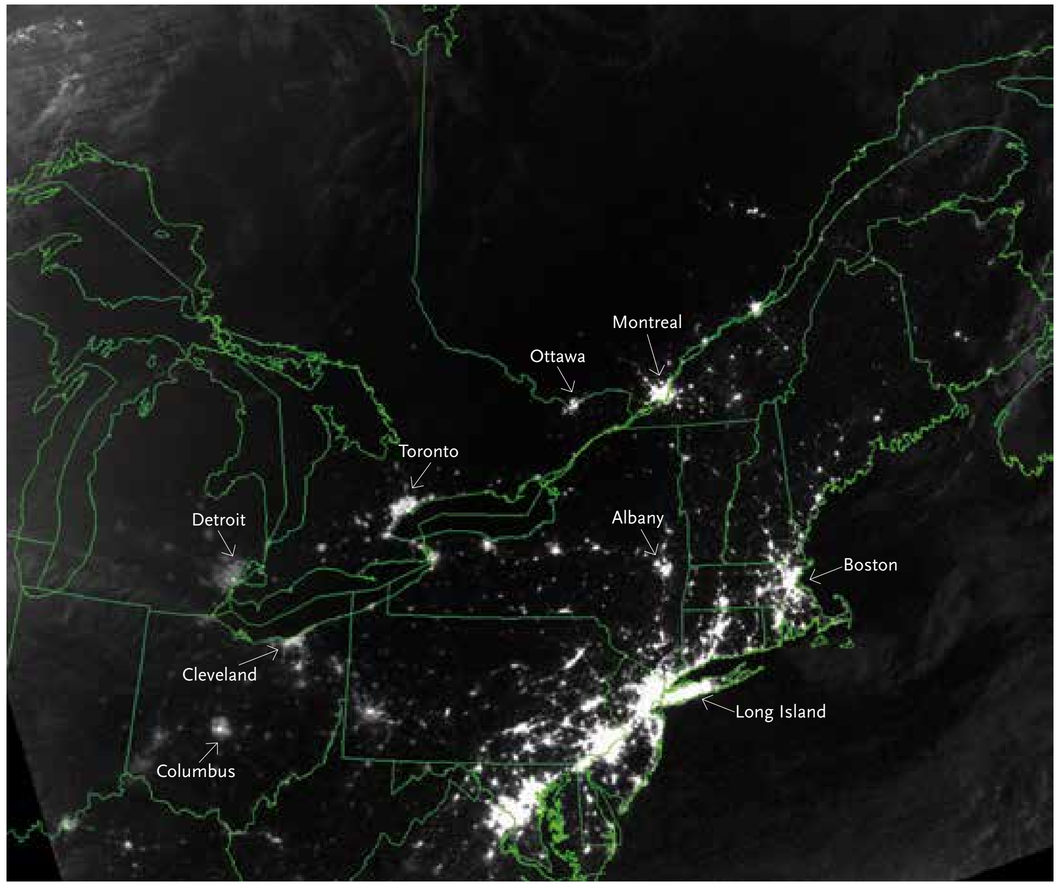
Van elektriciteitsnetten wordt verwacht dat ze ten allen tijde functioneren. De opkomst van hernieuwbare energiebronnen en hervormde elektriciteitsmarkten (waarin spotprijzen nu mogen variëren in real-time, in plaats van bijvoorbeeld om de 15 minuten) heeft een en ander onder druk gezet. Al 80% van de knelpunten in het Europese net houdt verband met de integratie van hernieuwbare energiebronnen (Yang, 2015). Transport van energie, van waar het wordt geproduceerd tot daar waar het nodig is, kost tijd en is niet 100 procent efficiënt. Natuurkundewetten beperken de mogelijkheid om elektriciteitsnetten op te delen in afzonderlijke fysieke en gebruiker lagen – iets dat cruciaal is geweest in het ontwerp van computer- en communicatienetwerken (Low, 2013).

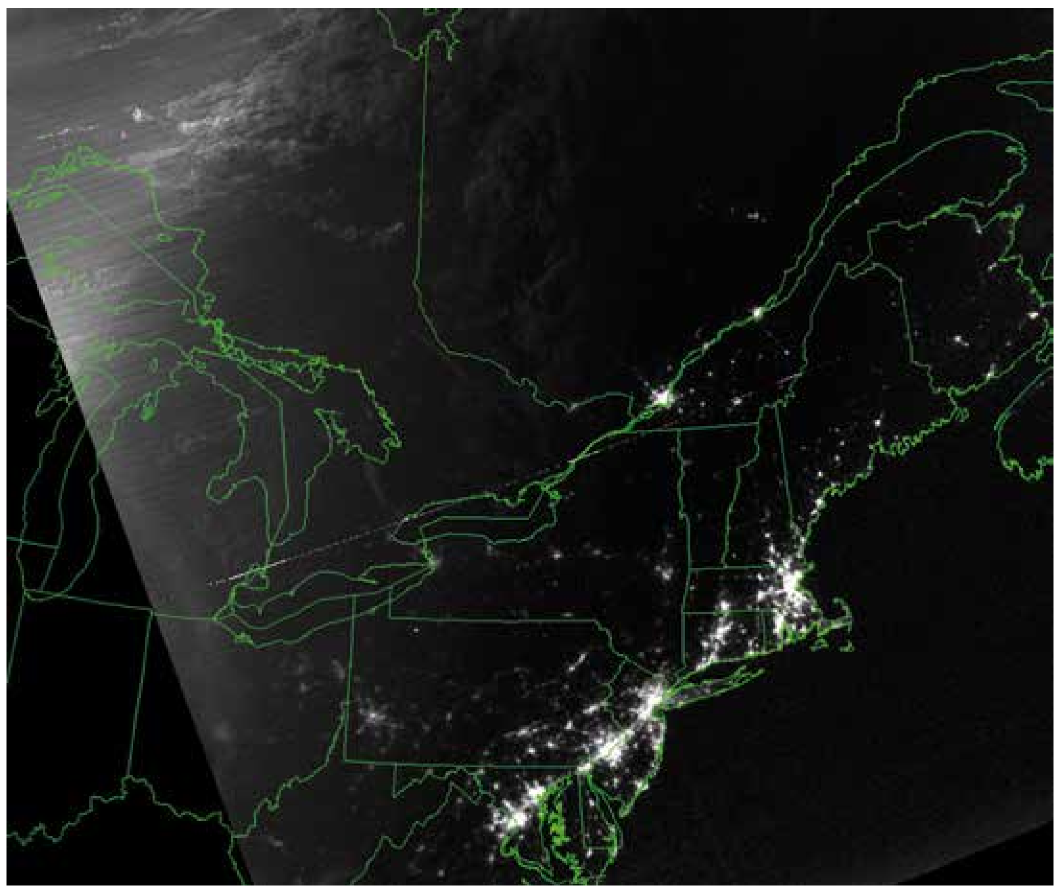
In de komende decennia moeten veel delen van de bestaande netwerkinfrastructuur worden vervangen. Het is daarom belangrijk om de betrouwbaarheid van de huidige én de toekomstige netwerkontwerpen vast te stellen. Op kortere tijdschalen, in de grootte van een paar tot circa 60 minuten, is het belangrijk om het effect van fluctuaties in zonne- en windenergie te begrijpen. Momenteel worden prognoses gezien als vaste waarden en wordt deze onvolledige informatie gebruikt om kortetermijn planning problemen zoals Optimal Power Flow (OPF) (Bienstock et al., 2012) op te lossen. Het lijkt zinvol om onzekerheid mee te nemen in deze optimalisatie procedures. Een betrouwbare netwerkplanning houdt dan rekening met bepaalde randvoorwaarden, hetgeen een netwerkstoring tot een zeldzame gebeurtenis maakt. Een voorbeeld van zo'n randvoorwaarde is dat een transportlijn in een elektriciteitsnet niet langer dan twee minuten per jaar oververhit mag zijn (Wadman et al., 2012). Op een nog kortere tijdschaal, bijvoorbeeld seconden, kunnen innovatieve oplossingen worden gebruikt die hun bruikbaarheid hebben bewezen in communicatienetwerken, zoals toegangscontrole en prijsmechanismes gebaseerd op effectieve bandbreedtes (Bosman et al., 2014). Er zijn oplossingen nodig waarmee de negatieve invloed van prijsspeculaties op fluctuaties in het elektriciteitsnet beperkt kan worden.

Een concrete onderzoeksvraag

We richten ons nu op een model geïnspireerd door een aantal van de bovenstaande ontwikkelingen en het formuleren van een aantal onderzoeksvragen. Deze vragen, alsook hun gedeeltelijke antwoorden, komen voort uit een vruchtbare samenwerking tussen de *Stochastics* en *Scientific Computing* groepen op het CWI.

We modelleren het elektriciteitsnet als een graaf $G = (V, E)$ gedurende een tijdsperiode T. De vector y van vermogens die aan de knooppunten van het netwerk op tijdstip 0 worden aangeboden heeft een aantal elementen die kunnen worden geoptimaliseerd. Het gaat er om de verzameling van deze vermogensinjecties F_p zodanig vast te stellen dat de kans dat een bepaalde randvoorwaarde qua betrouwbaarheid wordt overschreden ten hoogste p is, waarbij p in de orde 10^{-6} zal zijn. Een voorbeeld van een dergelijke beperking is dat alle stromen in het netwerk moeten worden begrensd door een voorgeschreven maximum C_{max} gedurende de planningsperiode. De reden voor een dergelijke beperking is dat een te hoge stroom kan leiden tot oververhitting van lijnen, waardoor deze uitzakken. In augustus 2003 raakte zo'n slappe lijn in het Noord-Amerikaanse elektriciteitsnet een boom, wat een gigantische storing tot gevolg had: 55 miljoen mensen zaten tot vier dagen lang zonder stroom.





Op 14 augustus 2003 viel in grote delen van het noordoosten van de Verenigde Staten en in Canada de elektriciteit uit. (boven) Satellietbeeld van het gebied op 14 augustus 2003, 20 uur voor de stroomuitval; (onder) 15 augustus, 7 uur na de stroomuitval. Bron: http://earthobservatory.nasa.gov/IOTD/view.php?id=3719

We nemen aan dat de netto stromen welke op de knooppunten aan het net wordt aangeboden beschreven kunnen worden door een multidimensionaal diffusie proces $Y(t)$, $t \geq 0$. Dit is een stochastische differentiaalvergelijking waarvan de drijvende Brownse beweging een kleine (co)variantie heeft, dat wil zeggen de covariantiematrix is geschaald (vermenigvuldigd) met een parameter ε welke klein zal zijn in onze analyse; dit is een standaard aanpak in de theorie van grote afwijkingen. Laat $P(y,\varepsilon)$ de kans zijn dat de sup-norm van de stroomvectors $C(t)$ op tijdstip t C_{max} overschrijdt in het interval $[0,T]$. Voor een vaste ε is de verzameling van toegelaten stroominjecties F_p gegeven door $F_p=\{y\colon P(y,\varepsilon)\leq p\}$.

Wanneer deze verzameling expliciet beschreven kan worden door deterministische ongelijkheden, kunnen deze meegenomen worden in een planningsprocedure zoals OPF zodat de planning voldoet aan het betrouwbaarheidscriterium. Helaas lijkt zo'n expliciete beschrijving te hoog gegrepen, de beschrijving vereist de oplossing van een zogenaamd *exit problem* voor een diffusie, waarbij de rand soms impliciet wordt beschreven door de AC power flow vergelijkingen. Om dit probleem toch aan te pakken kijken we op het CWI naar snelle simulatietechnieken en naar benaderingen die gebaseerd zijn op de theorie van grote afwijkingen (Bosman et al 2014, Wadman et al 2016). Een dergelijke benadering resulteert in een functie $I(y)$ zo dat $P(y,\varepsilon)\sim e^{-I(y)/\varepsilon}$.

Het idee is om de verzameling F_p te benaderen door de exacte kans $P(y,\varepsilon)$ te vervangen door deze benadering. De functie $I(y)$ wordt gegeven door de oplossing van een variationeel probleem, dat kan worden bestudeerd om eigenschappen van de gewijzigde versie van F_p af te leiden.

Literatuur

Bienstock, D., Chertkov, M., & Harnett, S. (2014). Chance-constrained optimal power flow: Risk-aware network control under uncertainty. *SIAM Review*, 56(3), 461–495.

Boer, A. den (2015). Dynamic pricing and learning: Historical origins, current research, and new directions. *Surveys in Operations Research and Management Science* 20(1), 1–18.

Bosman, J., Nair, J., & Zwart, B. (2014). On the probability of current and temperature overloading in power grids: A large deviations approach. *ACM SIGMETRICS Performance Evaluation Review* 42, 33–35.

Low, S. (2013). *Communication and Power Networks: architecture*. <https://rigorandrelevance.wordpress.com/2013/12/02/communication-and-power-networks-architecture-part-ii/>.

Wadman, W., Bloemhof, G., Crommelin, D., & Frank, J. (2012). Probabilistic Power Flow Simulation allowing Temporary Current Overloading. *Proceedings of Probabilistic Methods Applied to Power Systems 2012 (PMAPS)*, Istanbul.

Wadman, W., Crommelin, D., & Zwart, B. (2016). A Large Deviation based Splitting Estimation of Power Flow Reliability. *ACM Transactions on Modeling and Computer Simulation*, to appear.

Yang, Y. (2015). Hybrid Grids. *DNV GL Strategic research & innovation position paper 2–2015*.

Zwart, B. (2014). De macroscoop. *Nieuw Archief voor Wiskunde* 15(3), 176–179.

Bert Zwart studeerde in 1997 af als econometrist op de VU, en promoveerde in 2001 in de wiskunde op de TU/e. Momenteel is hij werkzaam als groepsleider op het CWI en is hij deeltijdhoogleraar aan de TU/e. Eerdere aanstellingen waren bij INRIA Rocquencourt, de VU en Georgia Tech in Atlanta. Aan het laatste instituut is hij momenteel nog steeds verbonden als adjunct-professor. In maart 2015 werd hem voor zijn werk (zoals dynamisch prijzen van goederen en macroscopische analyse van stochastische netwerken) de vijfjaarlijkse Van Dantzigprijs toegekend door de VvS+OR .
E-mail: <Bert.Zwart@cwi.nl>

VOORUITBLIK

Het hierboven beschreven onderwerp vormt de basis van het promotieonderzoek van Tommaso Nesti, die onlangs bij het CWI is begonnen in het kader van mijn VICI project *Rare Events*, en dat wordt uitgevoerd in samenwerking met DNV-GL. Dit project maakt deel uit van mijn onderzoeksambitie een uitgebreid wiskundig inzicht te verkrijgen in het vóórkomen van zeldzame gebeurtenissen in elektriciteitsnetten. Daarvoor is het ook noodzakelijk een inzicht in tweede orde effecten te verkrijgen: als een bepaalde lijn of een set lijnen breekt kan een kettingreactie van storingen optreden, hetgeen leidt tot een grote stroomstoring (een voorbeeld in de VS werd eerder genoemd – het meest recente voorbeeld in Nederland is 27 maart 2015 toen een miljoen huishoudens tot 5 uur lang verstoken waren van elektriciteit). Dergelijke *cascading failures* worden momenteel bestudeerd door Fiona Sloothaak, promovendus aan de Technische Universiteit Eindhoven, in het programma *Networks*. Sem Borst is haar tweede promotor.
Om te waarborgen dat onze modellen ook praktische relevantie hebben onderhouden we banden met Alliander, DNV-GL, Eneco en TNO. Zo gebruiken wij – Debarati Bhaumik, promovendus bij het CWI en mijn collega's Stella Kapodistria (TU/e) en Daan Crommelin (CWI) – data van bestaande windparken bij het ontwikkelen van statistische modellen voor het vermogen van windturbines toevoegingen. Debarati onderzoekt ook de voordelen van opslag (in batterijen) in elektriciteitsnetten, iets dat belangrijke connecties met wachtrij-problemen heeft. De laatste decennia is er een uiterst waardevolle connectie geweest tussen wachtrijen en communicatienetwerken. Ik verwacht dat zo'n connectie de komende decennia ook zal bestaan tussen wachtrijen en energienetwerken.

Van Dantzigprijs 2015 voor Bert Zwart

Onderzoeker Bert Zwart van het CWI heeft de Van Dantzigprijs 2015 gewonnen. Deze prijs geldt als de hoogste Nederlandse onderscheiding in de statistiek en besliskunde, en wordt eens in de vijf jaar uitgereikt. Zwart nam de prijs op 26 maart 2015 in ontvangst tijdens de Dag voor Statistiek en Besliskunde in de Jaarbeurs Utrecht.

Uit de zes genomineerden wees de jury unaniem Bert Zwart als winnaar aan. Volgens de jury '...heeft [Bert Zwart] zich ontwikkeld tot een toonaangevend onderzoeker in het brede gebied van de mathematische besliskunde. Het gaat daarbij niet om citaties of redactielidmaatschappen, maar om invloed, inspiratie en innovatie. Bert Zwart is een levend bewijs van de oude stelling dat doorbraken plaatsvinden op grensvlakken.'

Daarvan getuigt zijn werk aan *dynamic pricing*, op het grensvlak van operations research, statistiek en speltheorie, en aan de integratie van methoden uit de stochastische en deterministische scheduling. Baanbrekend is ook zijn analyse van netwerkmodellen voor verkeersopstoppingen op internet en op de weg, gebaseerd op een combinatie van resultaten uit de empirische procestheorie en de continue optimalisering. Onderzoek ongehinderd door disciplinaire grenzen en vooruitgeholpen door een diepe beheersing van een breed scala van gerelateerde gebieden.

Het werk van Bert vond eerder erkenning. Hij ontving de Gijs de Leveprijs, de Erlangprijs, een Veni, een Vidi en een Vici. En vandaag de Van Dantzigprijs. Zijn naam is een waardige toevoeging aan de lijst van laureaten van deze prijs, die de naam van David van Dantzig in ere houden en het werk aan de statistiek en besliskunde dat hij in de vijftiger jaren is begonnen hebben voortgezet en tot grote hoogte hebben gevoerd.

Bert Zwart ontving uit handen van Jacqueline Meulman, voorzitter van de VvS+OR, de Van Dantzigprijs 2015

LNMB-conferentie Lunteren, 14 januari 2016

We leven in een wereld waarin allerlei netwerken steeds belangrijker zijn geworden. Vaak zijn deze netwerken gigantisch groot en dynamisch van aard, waardoor het ontwerp en de analyse ervan bijzonder uitdagend kunnen zijn. De recente editie van het jaarlijkse Nederlands Genootschap Besliskunde (NGB) / Landelijk Netwerk Mathematische Besliskunde (LNMB) Seminar in Lunteren stond daarom geheel in het teken van dit onderwerp.

Het programma was zeer gevarieerd. Michel Mandjes (UvA, CWI) gaf een inleiding over het door NWO-gefinancierde NETWORKS project en schetste een aantal resultaten voor stochastische netwerken van afhankelijk opererende resources, die behalve in verkeers- en communicatienetwerken ook toepassingen hebben binnen de economie en biologie. Ana Barros (TNO/NLDA) sprak over de kracht van de analyse van sociale netwerken, die bijvoorbeeld gebruikt kunnen worden bij het in kaart brengen van terroristische organisaties. Pieter Cornelisse (KLM) vertelde in zijn presentatie over het belang van het wereldwijde netwerk van onze nationale luchtvaartmaatschappij en de bijbehorende kritische succesfactoren. Het ontwerp en de operationele uitdagingen van communicatienetwerken werd besproken door Richa Malholtra, die werkzaam is bij SURFnet, het Nederlandse netwerk voor onderzoek en onderwijs. Vervolgens werd door Jan van Doremalen (CQM) een overzicht gegeven van besliskundige technieken die toegepast kunnen worden bij het ontwerp van supply chains. Rommert Dekker (EUR) gaf een lezing over het ontwerp en de analyse van netwerken voor lijndiensten van containerschepen en illustreerde de gebruikte technieken aan de hand van een casus in Indonesië. Tot slot vertelde Maaike Snelder (TNO, TUD) over haar onderzoek naar data-gedreven modellen van onder meer het Nederlandse wegennetwerk en de Amsterdamse grachten.

Uit de lezingen werd duidelijk dat dankzij de besliskunde al heel veel uitdagingen zijn aangepakt, maar ook dat netwerken nog lang een vruchtbaar terrein voor onderzoek en nieuwe toepassingen zullen blijven. En natuurlijk was het NGB/LNMB Seminar weer een uitstekende gelegenheid om te netwerken!

Albert Wagelmans <wagelmans@ese.eur.nl>

Klassiekers

Eugene Lukacs over karakteristieke functies

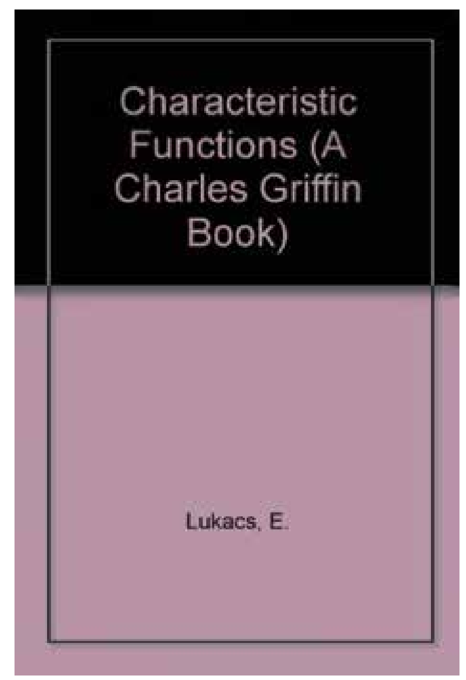
Fred W. Steutel

Van 1956 tot 1964 werkte ik op het Mathematisch Centrum (nu CWI) in Amsterdam. In 1960 kwam het boek *Characteristic Functions* van Eugene Lukacs uit bij uitgeverij Griffin. Toen ik voorstelde om me daarmee te gaan bezighouden, raadde een oudere collega mij dat af: het ging immers over Fouriergetransformeerden en daarvan was toch alles al bekend. Dat bleek toch iets anders te liggen. In mijn eigen proefschrift en in de proefschriften van twee van mijn promovendi speelde het boek een belangrijke rol.

Karakteristieke functies werden niet door Lukacs bedacht; ze komen al uitvoerig aan de orde in het boek *Fourierische Integrale* van Samuel Bochner (1955) en in die andere klassieker, *Mathematical Methods of Statistics* (1945) van Harald Cramér. Het boek van Lukacs was wel het eerste en is nog steeds het enige dat helemaal aan karakteristieke functies is gewijd.

Lukacs werd op 14 augustus 1906 in Szombathely, Hongarije geboren. Hij groeide op in Wenen en studeerde daar ook, eerst gymnasium, daarna universiteit In 1930 promoveerde hij op een meetkundig onderwerp bij Walter Meyer. Hij haalde een actuarisdiploma, waarbij hij ook wat statistiek geleerd zal hebben. In 1935 trouwde hij met studente wis- en natuurkunde Elizabeth (Lisl) Weisz Een korte tijd was hij leraar op middelbare scholen; daarna werkte hij voor een verzekeringsmaatschappij.

Toen nazi-Duitsland in 1939 Oostenrijk annexeerde emigreerde het echtpaar (zij hadden geen kinderen) naar de VS, waar hij aan verschillende universiteiten verbonden was. Hij was lange tijd directeur van het Statistical Laboratory van de Catholic University of America in Washington. Hij bleef daar tot 1972 als Professor Emeritus. Van 1973 tot 1975 was hij University Professor in Bowling Green, Ohio, waar hij bleef wonen na zijn pensionering. In 1978 keerde hij toch terug naar zijn huis in Washington.



In 1979 werkte ik een half jaar aan de Johns Hopkins Universiteit in Baltimore, niet ver van Washington. Van daaruit heb ik met vrouw en kinderen de Lukacses opgezocht; hun huis had een Weense ambiance.

Lukacs was een veelzijdig wiskundige: niet alleen karakteristieke functies, maar ook limietstellingen, karakteriseringsproblemen en onderwerpen uit de getaltheorie. Hij toonde belangstelling voor mijn werk over oneindig deelbare verdelingen. Hij nodigde mij uit voor een door hem georganiseerde bijeenkomst in het Mathematisches Forschungsinstitut in Oberwolfach in het Zwarte Woud, eldorado voor wiskundigen. Lukacs kwam daar graag en vaak. We hebben elkaar daarna nog vele malen ontmoet op wiskundige bijeenkomsten. We raakten bevriend.

Van zijn boek verscheen in 1970 een tweede druk. Hij had graag een heel nieuwe versie willen uitbrengen; maar daar wilde de uitgever niet aan: een te specialistisch onderwerp. Ten slotte kwam er in 1983 een aanvulling: *Developments in characteristic function theory*. De literatuurlijst bevat tien van mijn artikelen.

Eugene stierf op 21 december 1987. In maart van het jaar daarop was er een herdenkingsbijeenkomst in Bowling Green, waar ik voor uitgenodigd was. Zijn vrouw Elisabeth was niet meer in staat om bij de herdenking aanwezig zijn. Ik kreeg geen spreektijd, maar mij werd gevraagd een van de sprekers, Vijay Rohatgi, een naaste medewerker van Lukacs, in te leiden. Omdat het in die tijd kostbaar was om binnen korte tijd heen en weer te vliegen naar de VS, bleef ik een kleine week in Bowling Green. Het was er bitter koud was. Tot mijn verbazing zag ik een boom vol in het blad, maar even later vlogen al die 'bladeren' weg. Iets verderop trof ik een begraafplaats met honderden graven van soldaten uit de burgeroorlog; ze waren daar ver vandaan in het Zuiden gesneuveld. Herinneringen . . . .

Henk Tijms

column

EULERS GETAL *e* IS OVERAL

Slechts weinig getallen zijn zo beroemd dat ze een eigen letter hebben gekregen en wel voor de eeuwigheid. Het magische Euler getal e = 2,718281... is zo'n getal. Het getal e heeft zijn wortels in exponentiële groei zoals iedere leerling op de middelbare school tegenwoordig leert. Op uitermate originele wijze speelde Google hier op in toen het bedrijf in 2004 besloot aandelen uit te geven. Dit ging niet langs de geijkte weg van investeringsbanken, maar Google kondigde aan dat ze via een online veiling voor het grote publiek 2.718.281.828 dollars wilde ophalen. Velen in beleggingsland en daarbuiten begrepen niets van de ogenschijnlijk absurde precisie die werd nagestreefd voor het beoogde doel, maar menig wiskundige zal een glimlach niet hebben kunnen onderdrukken. Terecht was Google gefascineerd door het getal e. Dit getal is overal in de natuurwetenschappen. Het getal e speelt een hoofdrol in onder meer de beschrijving van de groei van populaties en de beschrijving van radioactief verval.

Het getal e verscheen voor het eerst in 1608 aan de oppervlakte in het werk van de Engelse wiskundige John Napier, de ontdekker van de logaritme. In 1683 herontdekte de Zwitserse wiskundige Jacob Bernouilli het getal e bij de bestudering van de groei van een bankrekening als jaar op jaar rente wordt toegevoegd. Maar het was het werk van het Zwitserse genie Leonhard Euler, één van de grootste wiskundigen aller tijden, waardoor het transcendente getal e definitief op de wiskundige kaart werd gezet. Een uitgebreid historisch overzicht over het getal e is te vinden in het prachtige artikel van Brian J. McCartin '*e*: The Master of Al' dat op internet te downloaden is.

Het getal e duikt ook overal op allerlei plaatsen in de kansrekening op en laat ik daar een aantal voorbeelden van geven. Daartoe ga ik eerst terug naar het jaar 1713 toen de Franse wiskundige Pierre Rémond de Montmort zijn boek *Essay d'analyse sur les jeux de hazard* publiceerde. In dit boek stelde De Montmort het volgende kaartspel voor dat tegenwoordig ook wel bekend staat als het Las Vegas kaartspel. De dealer draait de kaarten van een grondig geschud pak van 52 speelkaarten één voor één om en daarbij gelijktijdig roepend 'aas, twee, drie, ..., vrouw, heer, aas, twee, drie, ..., vrouw, heer,' en zo doorgaand totdat elk van de 13 rangen van de kaarten vier keer omgeroepen is bij het omdraaien van de 52 kaarten. Een 'treffer' treedt op als de omgeroepen rang overeenkomt met de rang van de omgedraaide kaart. De vraag waarvoor De Montmort gesteld stond, was de berekening van de kans op geen treffer. Dit probleem kon De Montmort niet oplossen, maar wel het simpeler probleem waarin hij zich beperkte tot 13 kaarten bestaande uit aas, twee, drie, ..., vrouw en heer. De lezer zal dit herkennen als een speciaal geval van het Sinterklaas-lootjes probleem waarin een aantal kinderen elk een cadeautje inbrengt, waarna de kinderen blindelings lootjes trekken om te bepalen welk cadeau een ieder krijgt. De vraag is dan wat de kans is dat geen enkel kind zijn eigen cadeau trekt. Deze kans kan simpel exact worden berekend en de exacte waarde is vrijwel gelijk aan de benaderingswaarde $e^{-1} \approx$ 0,3679 wanneer de groep uit zeven of meer kinderen bestaat (overeenstemming in vier of meer decimalen). Het oorspronkelijke Las Vegas kaartspel van De Montmort is exact heel wat lastiger op te lossen en de exacte oplossing vereist diepgaande combinatoriek. Echter met de Poisson heuristiek die ik mijn boek Understanding Probability uitgebreid behandel met vele toepassingen, is eenvoudig de benadering $e^{-4} \approx$ 0,0183 te vinden. Deze benadering komt goed overeen met de exacte waarde 0,0162. De exacte waarde kan worden berekend uit een integraalrepresentatie met een Laguerre polynoom van de graad vier voor de gezochte kans.

De Poisson heuristiek geeft ook direct een goede benadering in termen van het getal e voor het volgende probleem dat van hetzelfde type is als het Las Vegas kaartprobleem. Stel je hebt een familie van vijf broers en zussen met hun partners en die komen elk jaar met Kerstmis tezamen en wisselen dan onderling cadeaus uit. Elk van de 10 leden van de familie koopt een cadeau. De cadeaus worden genummerd als 1 tot en met 10, waarbij ieder familielid het nummer weet van zowel het door hem of haar gekochte cadeau als van het cadeau dat door de partner gekocht is. In een hoed gaan 10 kaartjes met de nummers 1 tot en met 10 en een ieder trekt blindelings een kaartje uit de hoed. Als iemand een kaartje trekt met het nummer van zijn of haar cadeau of met nummer van het cadeau van de partner, dan gaan alle kaartjes weer terug in de hoed en wordt de trekking opnieuw uitgevoerd. Wat is de kans dat niemand het nummer van het eigen cadeau of van het cadeau van de partner trekt zodat de trekking niet opnieuw gedaan moet worden? Een goede benadering van deze kans is $e^{-2} \approx$ 0,1353, oftewel een kans van ongeveer 13,53%. Dit ligt redelijk dicht bij de exacte waarde 12,31% van de gezochte kans. Had de familie uit zeven broers en zusters met hun partners bestaan, dan had de approximatieve kans van 13,53% dichter bij de exacte kans van 12,54% gelegen. Analyseren we nader de uitkomsten van verschillende andere gevallen, dan vinden we dat de benadering e^{-2} aangepast kan worden tot de meer accurate benadering $e^{-2}(1 - \frac{1}{2n})$.

Bij dit optreden van het getal e in de kansrekening blijft het niet. Iedereen kent waarschijnlijk wel het probleem dat bekend staat als het secretaresseprobleem of het bruidsschatprobleem van de Sultan. Laat ik dit probleem in een meer sekse-neutraal vat gieten. Een roversbende houdt haar bijeenkomst in een geheime schuilplaats. Buiten ligt een agent op de loer die bij toeval de schuilplaats ontdekt heeft en wiens enige doel is de leider van de bende te arresteren. De agent weet dat de schurken uit veiligheidsoverwegingen één voor één naar buiten zullen komen in een volledig willekeurige volgorde en dat de andere schurken gewaarschuwd zijn en ontsnappen zodra hij één van hen in de kraag grijpt. De agent zal dus alleen tot arrestatie van een bendelid overgaan als hij sterke aanwijzingen heeft dat het de roverhoofdman betreft. De agent weet uit andere bronnen dat de roverhoofdman de langste van het stel is en verder weet de agent uit hoeveel leden de bende bestaat. Hoe te handelen zodat de agent met maximale kans de roverhoofdman te pakken krijgt? Stel dat de bende uit n bendeleden bestaat, dan is het bij benadering optimaal de eerste n/e bendeleden te laten lopen en dan het eerstvolgende bendelid te arresteren die langer is dan alle voorgaande. De kans dat op deze wijze de roverhoofdman gearresteerd wordt is ongeveer gelijk aan e^{-1}. Dit is een prima benadering als het aantal bendeleden 10 of meer is. Voor n = 10 is het optimaal de eerste vier bendeleden te laten passeren en is de maximale pakkans gelijk aan 0,3987.

Tot slot laten we de rol van het getal e zien in een kansprobleem dat bekend staat als het eettafelprobleem of het Oberwolfach-probleem. Daartoe keren we terug naar bovenstaande familie van m personen, waarbij m = 10. Stel dat de familie voor zowel de avond van eerste kerstdag als de avond van tweede kerstdag een tafel heeft gereserveerd in een restaurant. De restauranteigenaar is gevraagd een ronde tafel te nemen en bij elk van de twee diners de familieleden in een *random* volgorde aan tafel te zetten. Wat is de kans dat geen twee of meer familieleden op beide avonden naast elkaar gezeten zijn? Een uitstekende benadering voor deze kans is $e^{-2}\left(1 - \frac{4}{m} + \frac{20}{3m^3}\right)$. Voor m = 10 is de benaderingskans gelijk aan 0,0821, terwijl de exacte waarde van de kans gegeven wordt door 0,0825. De correctieterm op e^{-2} kan praktisch bezien weggelaten worden als m tenminste 80 is. Echter ronde tafels met plaats voor 80 personen zijn niet op de hoek van elke straat te vinden. Dit brengt me op een uitbreiding van het probleem naar de situatie van meerdere tafels. De vraag is dan wat de kans is dat geen twee of meer personen bij twee opeenvolgende diners naast elkaar gezeten zijn wanneer bij beide diners rs personen op een volledig willekeurige wijze aan r ronde tafels gezet worden (s per tafel). Een uitdagend probleem dat voor zover ik weet onopgelost is!

Henk Tijms is emeritus hoogleraar operations research aan de Vrije Universiteit en auteur van diverse leerboeken over operations research en kansrekening.
E-mail: <tijms@quicknet.nl>

MULTIVARIATE EXTENSIES VAN DISCRETE KEUZEMODELLEN

Hoe econometrie snel en accuraat om kan gaan met veel gerelateerde keuzes*

Discrete keuzes zijn vaak aan elkaar gerelateerd. Denk daarbij aan productkeuzes in een supermarkt of antwoorden op enquêtevragen. Econometrisch onderzoek op dit gebied is schaars, omdat univariate keuzemodellen niet gemakkelijk uit te breiden zijn naar een multivariate setting. Dit artikel stelt een combinatie van model specificatie (Logit specificatie voor de conditionele verdelingen) en schattingsmethode (Composite Likelihood) voor die: (a) de relaties tussen discrete keuzes analyseert; (b) schaalbaar is naar grote dimensies; (c) parameters vindt met slechts een kleine onzuiverheid in kleine steekproeven; (d) weinig efficiëntieverlies ondervindt ten opzichte van Maximum Likelihood; en (e) parameterschattingen vindt in een acceptabel tijdsbestek. Dit alles maakt dat de praktische toepasbaarheid van de methodiek van grote meerwaarde is voor toegepast dataonderzoek.

Koen Bel

Iedereen maakt iedere dag veel discrete keuzes. De reis naar werk kan bijvoorbeeld worden afgelegd per fiets, auto of openbaar vervoer. De keuze voor producten in een supermarkt is ook een voorbeeld van zo'n discreet keuzeproces. Keuzemodellen zijn bedoeld om deze keuzes te beschrijven en te zien welke aspecten de keuzes beïnvloeden.

Discrete keuzes van een individu zijn vaak met elkaar gecorreleerd. Wanneer bijvoorbeeld een goedkoop merk brood wordt gekocht is het aannemelijk dat dit individu ook kiest voor het goedkope merk boter, of een specifiek merk hagelslag, kaas, bier, chips et cetera. Het analyseren van dit multivariate keuzeproces levert inzichten op in de relaties tussen keuze en verklarende variabele, maar ook tussen verschillende keuzes. In het supermarktvoorbeeld is deze informatie relevant voor het kiezen van advertenties en het indelen van schapruimte. Maar ook bij enquêtes is het van belang om te weten welke relaties bestaan tussen de gegeven antwoorden op verschillende vragen. Dat individuen negatief antwoorden op zowel een vraag over de regering als op een vraag over het sociale vangnet zal niet ontdekt kunnen worden door univariate analyses.

Omdat discrete keuzemodellen niet gemakkelijk uit te breiden zijn naar meer dimensies, is de literatuur over multivariate keuzes schaars. In dit artikel introduceren wij een multivariate uitbreiding op bestaande univariate keuzemodellen. Het artikel introduceert daarnaast bruikbare schattingsmethodes om de relaties tussen de discrete keuzes te kunnen duiden.

Specificatie voor het multivariate keuzemodel

Er zijn verschillende mogelijkheden om een modelspecificatie voor het multivariate keuzeprobleem op te zetten. Er kan bijvoorbeeld gedacht worden aan een groot multinomiaal keuzemodel over alle mogelijke combinaties van keuzes. Het aantal keuzecombinaties wordt echter snel erg groot, waardoor interpretatie en het schatten van de parameters erg lastig wordt. Met 10 producten in een supermarkt die een individu wel of niet kan kopen zijn er al $2^{10} = 1024$ combinaties van keuzes. Bij nog grotere aantallen worden kansen numeriek te klein om accuraat mee te kunnen rekenen.

In de literatuur is een multivariate extensie van het alom bekende multinomiale probit model voorgesteld. Helaas heeft men hier te maken met het uitwerken van integralen en dit wordt al computationeel intensief bij univariate analyses. In onze multivariate probleemstelling zal dit alleen nog maar lastiger worden.

In dit artikel gebruiken we daarom een andere aanpak. Omdat men geïnteresseerd is in relaties tussen de keuzes, is het praktisch om een overzichtelijke specificatie te hebben voor de conditionele kansen: de kans dat een individu a voor item A kiest gegeven de keuzes die hij/zij maakt voor items B, C, et cetera (en gegeven de verklarende variabelen). Deze conditionele kansen geven we dan weer door middel van een logit specificatie en bevatten alle informatie waarin we geïnteresseerd zijn; relaties tussen keuzes en effecten van exogene variabelen, zie (1).

$$\Pr[Y_{i,k} = j | Y_{-i,k}, X_i] \propto \exp(\alpha_{k,j} + X_i \beta_{k,j} + \sum_{l \neq k} \psi_{kl,jy_{i,l}}) \qquad (1)$$

Voor het schatten van de parameters zou men gebruik willen maken van de gezamenlijke verdeling. Om daar te komen maken we gebruik van de regel van Besag (1974) die zegt dat er een een-op-een relatie is tussen de volledige set van conditionele kansen en de gezamenlijke verdeling. Dit resulteert in een elegante (maar grote) multinomiale logit-specificatie (2) over alle mogelijke keuzecombinaties, welke direct gebruikt zou kunnen worden in de schattingsmethode Maximum Likelihood.

$$\Pr[Y_i = j | X_i] \propto \exp(\sum_{k=1}^{K} (\alpha_{k,y_{i,k}} + X_i \beta_{k,y_{i,k}} + \sum_{l > k} \psi_{kl,y_{i,k}y_{i,l}})) \qquad (2)$$

Probleem bij groot aantal mogelijke combinaties van keuzes

Met opzet gebruiken we het omslachtige 'zou kunnen worden' in de voorgaande zin, want de bruikbaarheid van de specificatie is binnen de standaard schattingsmethodes nog gelimiteerd. Dit heeft te maken met het volgende. Stel, er is een enquête afgegeven met 10 vragen met allemaal 7 keuzes opties, zeg van 'Helemaal niet mee' eens tot 'Helemaal mee eens'. Dan zijn er ongeveer ($7^{10} \approx 282$ miljoen mogelijke combinaties van antwoorden. De grote multinomiale logit specificatie kan hier niet mee omgaan: kansen worden numeriek te klein om zonder problemen te behandelen en computertijd wordt extreem lang. Figuur 1 laat deze computertijden zien voor verschillende aantallen keuzes en verschillende aantallen keuzeopties. Duidelijk is dat de computertijd explodeert naarmate meer keuzecombinaties mogelijk zijn. Direct Maximum Likelihood gebruiken in deze grote setting lijkt dus niet de juiste weg om te bewandelen.

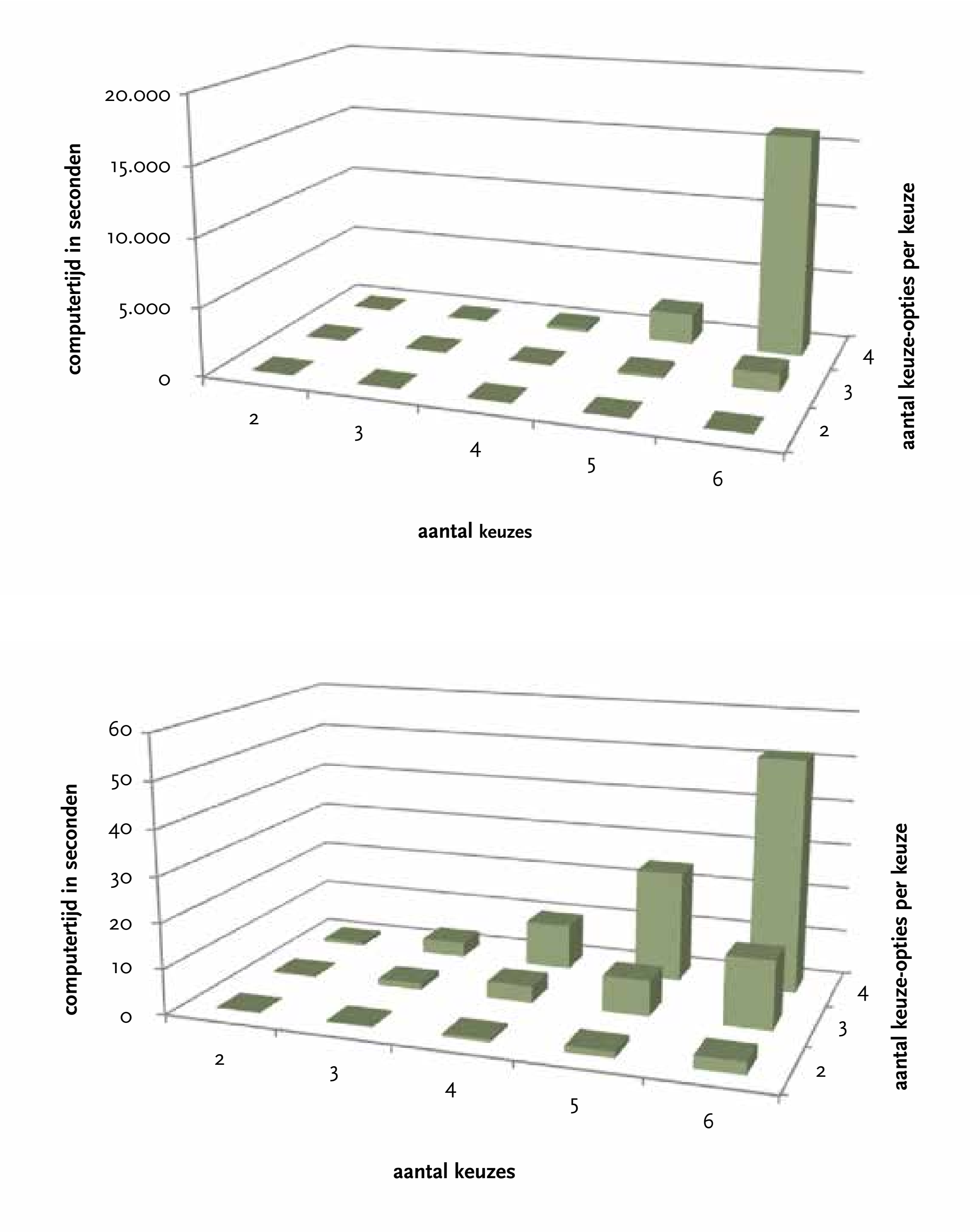
Voorgestelde schattingsmethode

We doen daarom een voorstel voor een alternatieve schattingsmethode. Eerder lieten we zien dat we het model specificeren vanuit de conditionele kansverdeling. Alle informatie waarin we geïnteresseerd zijn (relaties met variabelen en tussen keuzes) zit in de volledige set aan conditionele verdelingen. We vervangen de volledige likelihood functie door een misgespecificeerde samenstelling van conditionele verdelingen, zoals in (3). In de literatuur wordt voor deze methode de term Composite Conditional Likelihood gebruikt. Een overzichtsartikel van Varin et al. (2011) laat zien dat de Composite Likelihood methode consistente schatters oplevert. Desalniettemin hebben we te maken met een misgespecificeerde methode en zal efficiëntieverlies plaatsvinden.

$$\ell_{ccl}(\theta; y) = \sum_{i=1}^{N} \sum_{k=1}^{K} \Pr[Y_{i,k} = j | Y_{-i,k}, X_i] \qquad (3)$$

De artikelen waar dit stuk op gebaseerd is (Bel et al., 2014; Bel & Paap, 2014; Bel & Schoonees, 2015) laten zien dat de Composite Conditional Likelihood methode accurate schatters oplevert voor het multivariate keuzemodel. Simulatie studies en applicaties laten zien dat de onzuiverheid in een kleine steekproef vergelijkbaar is met de standaard methodes en dat het verlies van efficiëntie ten opzichte van deze standaard methodes verwaarloosbaar is. Daarnaast wordt de rekentijd aanzienlijk gereduceerd (zie figuur 1) aangezien we in de log likelihood functie nu te maken hebben met een sommatie van redelijke kansen in plaats van een vermenigvuldiging die leidt tot hele kleine kansen.

De Composite Conditional Likelihood methode is dan ook een goed alternatief voor Maximum Likelihood als het aantal keuzes of aantal keuze-opties groot is. Waar de Maximum Likelihood methode in het beste geval lang doet over het vinden van parameterschattingen en in het slechtste geval vastloopt omdat kansen numeriek te klein zijn, vindt Composite Likelihood een aannemelijke benadering in een acceptabel tijdsbestek.



Figuur 1. Computertijd in seconden voor de multivariate modelspecificatie gebruikmakend van standaard Maximum Likelihood (bovenste panel) en van Composite Conditional Likelihood (onderste panel) voor verschillende aantallen keuzes en keuze-opties.

Toepassingen

De nieuwe modelspecificatie voor multivariate discrete keuzes in combinatie met de Composite Likelihood schattingsmethode is wijd toepasbaar in empirisch onderzoek. Eerder in dit artikel zijn de voorbeelden over productkeuzes en gerelateerde antwoorden op enquêtevragen genoemd. Daarnaast kan gedacht worden aan keuzes voor verzekeringen, verschillende gerelateerde effecten van medicijnbehandeling, keuzes bij het bouwen van een huis en nog vele andere voorbeelden. Wanneer de volledige modelspecificatie klein blijft, is het geen probleem om Maximum Likelihood te gebruiken. Mocht de specificatie groter worden, en ook vaker worden gebruikt voor het invoegen en verwijderen van variabelen, is Composite Conditional Likelihood een bruikbaar alternatief.

Conclusie

Alternatieve schattingsmethoden worden steeds meer bruikbaar in econometrie. De huidige beschikbaarheid van grote gegevensbestanden legt veel (non-lineaire) relaties bloot. Standaardmethoden zullen veel tijd nodig hebben deze extensieve relaties te omvatten en snelle en accurate benaderingen zijn dan een goede overweging. Een benadering van de optimale oplossing is meer waardevol dan een *infeasible* en *intractable* exacte oplossing. Het huidige artikel laat zien dat een accurate benadering zeer toepasbaar is in de praktijk van geassocieerde discrete keuzeprocessen.

Literatuur

Bel, K. (2015). Multivariate extensions to discrete choice modeling. PhD thesis, Erasmus University Rotterdam. 12 november 2015.

Bel, K., Fok, D., & Paap, R. (2014). Parameter estimation in multivariate logit models with many binary choices. Rotterdam: Econometric Institute (Intern rapport, EI report serie, no EI 2014-25).

Bel, K., & Paap, R. (2014). A multivariate model for multinomial choices. Rotterdam: Econometric Institute (Intern rapport, EI report serie, no EI 2014-26).

Bel, K,. & Schoonees, P. (2015). *Extending the Dale model to multivariate ordered responses using composite likelihood estimation*. Unpublished manuscript, Erasmus University Rotterdam.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B (Methodological)*, 36(2), 192–236.

Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21, 5–42.

* Dit artikel is gebaseerd op het proefschrift *Multivariate Extensions to Discrete Choice Modeling* verdedigd op 12 november 2015 aan de Erasmus Universiteit. Het onderzoek is verricht in samenwerking met prof. dr. Richard Paap (Econometrisch Instituut, Erasmus Universiteit Rotterdam), prof. dr. Dennis Fok (Econometrisch Instituut, Erasmus Universiteit Rotterdam) en dr. Pieter Schoonees (Rotterdam School of Management, Erasmus Universiteit Rotterdam).

Koen Bel is gepromoveerd bij het Econometrisch Instituut van de Erasmus School of Economics en werkt nu bij Erasmus Q-Intelligence. Erasmus Q-Intelligence voert contractonderzoek uit op het gebied van econometrie, data science en business analytics.
E-mail: <bel@ese.eur.nl>