STAtOR

Statistics and Operation Research for the Energy Transition *Program Annual Meeting 2022 & VVSOR Conference*

Flexibele Statistiek; Van p-waarden naar e-waarden

De kortste route langs 57.912 rijksmonumenten

Cognitieve achteruitgang detecteren met behulp van spraakherkenning via sensoren

Imputeren van huishoudsamenstellingen met machine learning

Welke test 'loopt' het best?

Beter prevalenties meten door het combineren van schattingsmethoden

De reikwijdte van de counterfactual; Over causaliteit, potential outcomes en grafische modellen

Foutlokalisatie in enquêtedata met behulp van (zachte) regels en machine learning

periodiek van de VVSOR jaargang 23, nummer 1, maart 2022

STAtOR

Jaargang 23, nummer 1, maart 2022

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

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

Kopij en reacties richten aan
Prof. dr. J.A.S. Gromicho (hoofdredacteur), Universiteit van Amsterdam Faculteit Economie en Bedrijfskunde, Sectie Operations Management | Amsterdam Business School, Plantage Muidergracht 12, 1018 TV Amsterdam, j.a.s.gromicho@uva.nl

Bestuur van de VVSOR
Voorzitter: prof. dr. Casper Albers, db@vvsor.nl; Secretaris: Pieter Jongsma MSc, secretaris@vvsor.nl; Penningmeester: Judith ter Schure MSc, penningmeester@vvsor.nl; Algemeen bestuurslid: Thomas Wise MSc, db@vvsor.nl; Webmaster: Eugenio Traini MSc: webmaster@vvsor.nl.
Voorzitters van de secties: prof. dr. ir. Mark van de Wiel (Biometrical Section); prof. dr. Albert Wagelmans (Section for Operations Research); dr. Eduard Belitser (Section Mathematical Statistics); dr. Rebecca Kuiper (Social Sciences Section); dr. Michel van de Velden (Economics Section); dr. Daniel Oberski (Section Data Science); Marije Sluiskes MSc (Young Statisticians) dr. Sanne Willems (Section Statistics Communication).

Leden- en abonnementenadministratie van de VVSOR
VVSOR, Maarsbergseweg 20, 3956 KW Leersum, admin@vvsor.nl. Raadpleeg onze website www.vvsor.nl over hoe u lid kunt worden van de VVSOR of een abonnement kunt nemen op *STAtOR*.

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

Ontwerp en opmaak
Pharos, Nijmegen

Uitgever

ISSN 1567-3383

Data? Niet vanzelfsprekend!

STAtOR heeft het vorige jaar afgesloten met een geweldig nummer, geheel gewijd aan wat ons vakgebied kan betekenen bij het optreden van een pandemie. Nu ligt een nieuw jaar ligt voor ons, met de stellige verwachting dat we de covid-beperkingen achter ons kunnen laten. Toch zal onze samenleving blijvend zijn veranderd, zo is deels thuiswerken niet meer weg te denken.

Op de Annual Meeting (AM) op 17 maart zullen we elkaar eindelijk weer fysiek kunnen ontmoeten in De Balie in Amsterdam. Schrijf u snel in voor deze dag! Het wordt een hybride bijeenkomst dit jaar, zodat leden en niet-leden de AM ook online kunnen bijwonen. Vorig jaar hebben we geleerd dat we daarmee deelnemers van over de hele wereld kunnen trekken.

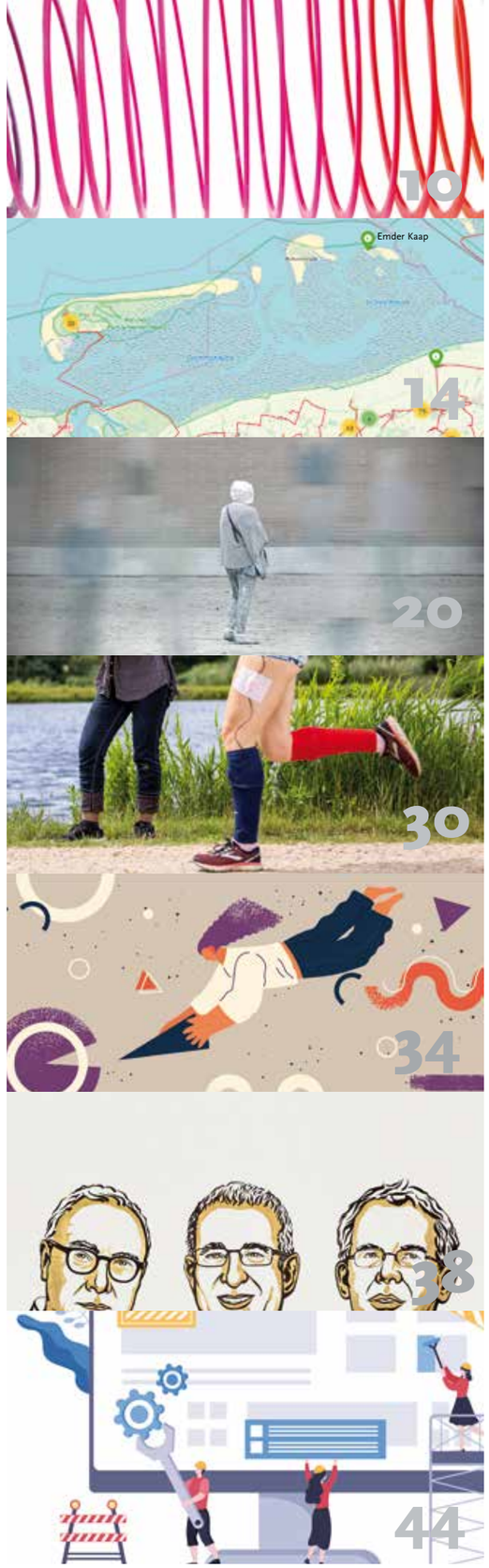
Naast de informatie over onze hybride Annual Meeting, die gaat over de mogelijkheden van ons vakgebied bij de komende Energie transitie bevat dit nummer een flink aantal artikelen. Bijzondere aandacht vragen we voor ons openingsartikel: Rosanne Turner, winnares van de Jan Hemelrijk Award van vorig jaar, heeft het over Flexibele Statistiek.

In veel van de artikelen ligt de nadruk op data. Data lijken vaak zo vanzelfsprekend, ze worden verzameld en kunnen dan worden geanalyseerd. Maar juist in het eerste deel van dat traject zijn er aspecten als volledigheid en consistentie die aandacht behoeven. Zie hiervoor de artikelen van Daalmans, Kloos en Keijzer. Ook kunnen data op een innovatieve manier worden verzameld, van Hoek et al beschrijven het detecteren van cognitieve achteruitgang met behulp van spraakherkenning via sensoren in een handschoen!

Ook de verdere inhoud is gevarieerd. Fietsland Nederland is rijk aan monumenten, liefst 57,912 hebben de status Rijksmonument. De Ruiter vertelt hoe de kortste route per fiets langs deze monumenten is gevonden. Starmans schrijft over over causaliteit, potential outcomes en grafische modellen en Boiten analyseert het meten van loopgedrag.

En natuurlijk is *STAtOR* niet compleet zonder onze columnisten: Henk Tijms over een levensgevaarlijke brug, Jelke Bethlehem over de onwenselijkheid van enquêtes via Twitter en Gerard Sierksma over een obsessief zoekende wiskundige.

Wij wensen u zoals altijd veel leesplezier!



INHOUD

2 **Data? Niet vanzelfsprekend!**

Energy Transition — **Statistics and Operations Research for the Energy Transition**

4 **Letter from the president**

6 **Program Annual Meeting & Conference on March 17, 202** | online & De Balie, Amsterdam

10 **Flexibele Statistiek. Van p-waarden naar e-waarden; Robuustere conclusies met minder data** | Rosanne Turner

14 **Record-handelsreizigersprobleem; De kortste route langs 57.912 rijksmonumenten** | Frans de Ruiter

18 **Peilingpraktijken. Een Twitter-peiling? Doe maar niet!** – column | Jelke Bethlehem

20 **Cognitieve achteruitgang detecteren met behulp van spraakherkenning via sensoren** | Rianne Drijver, Sigrid van Hoek, Jonas Klingwort & Rob Willems

24 **Imputeren van huishoudsamenstellingen met machine learning** | Jacco Daalmans

27 **Over P≠NP en een Eeuwige Student** – column | Gerard Sierksma

30 **Welke test 'loopt' het best?** | Marjolein Boiten

34 **Beter prevalenties meten door het combineren van schattingsmethoden** | Kevin Kloos

38 **De reikwijdte van de counterfactual; Over causaliteit, potential outcomes en grafische modellen** | Richard Starmans

44 **Foutlokalisatie in enquêtedata met behulp van (zachte) regels en machine learning** | Tania Keijzer

47 **De doodlopende Glazen Brug** – column | Henk Tijms

48 **Good news from Young Statisticians**

48 **Leden gezocht voor de werkgroep Open Statistica**

STATISTICS AND OPERATIONS RESEARCH FOR THE ENERGY TRANSITION

Letter from the President

With this letter, I follow up on a tradition that has been going on for over a decade, to address the members in the first *STAtOR* issue of the year. The past few years have been difficult years for everyone – so difficult that our society couldn't even celebrate it's 75th birthday – but also difficulties come with interesting statistical challenges, as was clear from the previous issue of *STAtOR*. These difficulties also imply that we cannot have the Annual Meeting in the format that we had pre-corona. However, as the prospects are now better than last year, we also don't have to have a fully online Annual Meeting either. We'll combine the best of both worlds with the first hybrid Annual Meeting of the society.

New concept, new location, same quality

The Annual Meeting will be held on 17 March of this year. We will break with the tradition of hosting the Annual Meeting in Utrecht: this time the event will take place at De Balie in Amsterdam. We've decided for this venue as this has the technical facilities to combine an in person meeting with an online live stream.

For a limited number of members, there is the possibility to register (for a fee) for a live attendance. We would have preferred to open this option to everyone, but for obvious reasons we cannot fill the venue with statisticians packed close together. All those who can't, or don't want to, register for the in person meeting, can register (for free) for online participation – including the possibility to ask the speakers questions. We've learned from last year's Annual Meeting that having the event (also) online can be a great success, with attendees from all over the globe. Registration details for the Annual Meeting are given on the following pages.

Just as in previous years, the organizing committee has managed to come up with an exciting programme, with speakers covering the different areas of Statistics and OR talking about one of the major societal challenges of our time: the energy transition.

Energy Transition

By now, there is abundant scientific consensus that man-made climate change has serious consequences and endangers our way of living. In most countries, there is also political consensus that this is problematic and needs to be acted upon – although consensus on how to act often is lacking. Many citizens are aware that we need to change our way of living, but are often confused about how to act due to contradicting information.

One of the most important steps in the energy transition is moving from e.g. coal and gas based energy to renewable sources such as wind and solar. This is easier said than done: recently there were warnings in the news that the Dutch electricity system is so saturated that no more solar panels should be connected. Thus, there is a need for smart models (read: statistics) in order to have the transition go smoothly.

Flexible pricing – making electricity more expensive when the sun doesn't shine and the wind doesn't blow – can also help to shift energy demand to moments where the energy supply is ample. Again, this is easier said than done, and sophisticated models are required to come up with good flexible prices.

With more and more people generating electricity themselves, via e.g. photovoltaic panels and wind turbines, and even storing the electricity themselves in large batteries, small communities arise that are (nearly) self-sufficient with regards to energy consumption. Such communities influence how we model the energy grid.

The first three speakers, Otto Swertz, Ronald Huisman and Miguel Anjos, will discuss the above mentioned statistical and OR challenges in modeling energy grids. However, we know that humans aren't only driven by rational arguments. Having an algorithm simply computing that the transition is a sensible step, isn't sufficient for making this step happen. In the fourth talk, Linda Steg will discuss the psychology behind climate change: how should we approach people in order to encourage them to show pro-environmental behaviour? Also here statistical techniques help answer this question.

Besides these four talks, the Annual Meeting will also host the award ceremonies for the Jan Hemelrijk Award, for the best undergraduate thesis, and the Willem van Zwet Award, for the best PhD thesis.

General Assembly

The meeting will start with the Algemene Ledenvergadering, or General Assembly, which is for members only. During the general assembly we will discuss the state of affairs of the society. Besides the annually recurring points as the budget, there will be some new points for discussion and approval of the members. First, Marianne Jonker from Radboud UMC will be nominated as new board member with the plan that she will take over Judith's role as treasurer in 2023. Furthermore, a committee on 'Open Statistica' will be established – see the announcement elsewhere in this *STAtOR*. Some minor changes to the awards procedures for the Van Zwet award and the Hemelrijk award will be proposed, as well as a simplification in the regulation concerning free student-membership. All documents for the general assembly will be distributed via e-mail two weeks before the event.

Casper Albers

VVSOR Annual Meeting

March 17, 2022

Energy Transition

The 2022 Annual Meeting will be about the energy transition. The need for transition towards using renewable energy has recently been discussed more and more in the media. Subsidies to purchase solar panels have been provided to many households in recent years, turning consumers into prosumers.

This recently even caused a problem, having a case of oversupply of solar electricity on the grid in the North of the Netherlands, where this generation could not be handled. On the other hand, energy prices have been increasing in the Netherlands, so much so that some companies are shutting down, temporarily. While some companies even went out of business. Consumers are now being confronted with increases in energy prices up to 67%, which caused the government to intervene and provide financial compensation to Dutch households.

How can Operations Research and Statistics shed some light on the challenges around the energy transition? Four speakers will discuss their research on the energy transition, all of them with a different perspective on the topic and coming from another field of study.

Annual Meeting of the Netherlands Society for Statistics and Operations Research (VVSOR)

Thursday March 17, 2022

11:00 – 17:00

online (hybrid) & at De Balie

Kleine-Gartmanplantsoen 10, 1017 RR Amsterdam

How can Operations Research and Statistics shed some light on the challenges around the energy transition? Four speakers will discuss their research on the energy transition, all of them with a different perspective on the topic and coming from another field of study.

- **Ir. O. (Otto) Swertz**
- **Dr. R. (Ronald) Huisman**
- **Prof. dr. M.F. (Miguel) Anjos**
- **Prof. dr. L. (Linda) Steg**

For the first time, this year the Annual Meeting will be a hybrid event at De Balie in Amsterdam and broadcasted live online. During the Q & A questions can be asked via the chat function. We will have a general assembly for members (in Zoom), followed by the actual event with four talks and two award presentations. The AM 2022 will be in English.

Registration is now open, please register on the vvsor-website
https://www.vvsor.nl/articles/vvsor-annual-meeting-2022.
Attending this year's annual meeting online is free of charge.
Attending the meeting at De Balie costs 75 euro (including drinks and lunch).
Reduced price for students: 25 euro.
E-mail: annualmeeting@vvsor.nl

DATE
Thursday, March 17, 2022

VENUE
Online via Vimeo
and at De Balie, Kleine-Gartmanplantsoen 10, 1017 RR Amsterdam

REGISTRATION
Registration for the conference is mandatory at www.vvsor.nl/articles/vvsor-annual-meeting-2022. Detailed information can be found on our website.

LANGUAGE
The talks at the annual meeting will be in English.

ALGEMENE LEDENVERGADERING (ALV)
The Annual General Meeting of members (ALV) takes place on March 17, 11:00 – 12:00, via Zoom. The relevant documents will be provided on the website two weeks before the meeting.

SNACKS AND DRINKS
Lunch and drinks during the breaks will be provided.

ORGANIZING COMMITTEE
The annual meeting is organized by a special committee in cooperation with the board of the VVSOR.
For questions, contact the administration by email at annualmeeting@vvsor.nl.

PLEASE REGISTER BEFORE MARCH 15

10:30 – 11:00	**Walk in**
11:00 – 12:00	**Annual General Meeting (ALV), for members only, also via Zoom**
12:00 – 13:00	Lunch break
13:00 – 13:15	**Welcome** and **Opening Talk** by Casper Albers, President of the VVSOR
13:15 – 13:45	**The history of the Dutch energy system and the current energy transition** Otto Swetz, *Statistics Netherlands*
13:45 – 14:15	**Data analytics will help energy markets to better accommodate renewable energy** Ronald Huisman, *Erasmus University*
14:15 – 14:45	Break
14:45 – 15:15	**Prosumers and the Future of Smart Electricity Grids** Miguel Anjos, *University of Edinburgh*
15:15 – 15:40	**Ceremony of the Willem R. van Zwet Award & the Jan Hemelrijk Award** Prize winners will be presented by the juries, followed by a short presentation by the laureates
15:40 – 16:00	Break
16:00 – 16:30	**The psychology of climate change** Linda Steg, *University of Groningen*
16:30 – 17:00	**Final panel discussion with speakers (Q & A) & Closure**
17:00 – 18:00	**Drinks**

13:15 – 13:45

THE HISTORY OF THE DUTCH ENERGY SYSTEM AND THE CURRENT ENERGY TRANSITION

Ir. O. (Otto) Swertz

Statistics Netherlands (CBS)

The presentation will give an overview of the history of the energy system in the Netherlands focusing on the energy flows. It will show the earlier transitions in the country like amongst others the rapid transition to natural gas in the 1960s and the impact of the 1973 oil crisis and the slow but steady rise of renewable energy in this millennium. Further shown will be relevant variations within energy consumption and the drivers for this. An example is the influence of the North-West European electricity market and/or the wholesale market prices on energy consumption in the Netherlands.
As an introduction to the next speakers, the main policy goals for energy transition and emission reduction in the light of climate change will be summarized. The relation between energy products and greenhouse gases will be shown. Policy targets are internationally agreed like the Paris Agreement from 2018 and the recent European Green Deal. Also shown will be the Dutch translation of the national goals to regional goals within the Dutch Climate Agreement (Klimaatakkoord). The regional datasets developed for communities to help them in working on the local energy transition will be presented.

OTTO SWERTZ is since 2017 heading the energy statistics team within Statistics Netherlands, better known as CBS. This is the National Statistical Institute by task responsible for the statistics publication within the country. The energy team consists of over 20 persons. From 2007 till 2017 Otto was project manager responsible for the production of the energy balance and the monthly statistics on fossil fuels and electricity.

13:45 – 14:15

DATA ANALYTICS WILL HELP ENERGY MARKETS TO BETTER ACCOMMODATE RENEWABLE ENERGY

Dr. R. (Ronald) Huisman

Erasmus University

The increase of supply from renewable energy sources challenges power markets. As supply from renewables is driven by weather conditions it is not perfectly predictable. This frequently results in a mismatch between supply and demand with outages and extreme prices as a result. My presentation shows that energy markets have to become more flexible to accommodate renewable energy. I'll argue that increased flexibility can be achieved with statistics, data analytics, and models. In the second part I'll discuss a quantile regression approach to better predict when power prices will be extremely high or low.

RONALD HUISMAN studied econometrics at the Erasmus University Rotterdam and has a PhD in financial economics from Maastricht University. Currently, he is associate professor financial economics at the Erasmus University Rotterdam. There he teaches about and research topics on financial economics applied to (renewable) energy markets, sustainability, and impact investing. Besides academia, Ronald has co-founded Energy Global, a data-driven energy company. After leaving Energy Global he co-founded Modex Analytics (a data management and analytics company), and Floyd Davis Finance (a company that helps impact entrepreneurs to become investment ready).

14:45 – 15:15

PROSUMERS AND THE FUTURE OF SMART ELECTRICITY GRIDS

Prof. dr. M.F. (Miguel) Anjos

University of Edinburgh

A smart grid is the combination of a traditional electrical power system with information and energy both flowing back and forth between suppliers and consumers. We focus on how the accessibility and reducing cost of decentralized renewable energy sources are stimulating the emergence of small-scale residential prosumers who can produce and consume electricity. Such prosumers may increase the uncertainty of consumption behaviour, reduce consumption from the grid, and potentially disconnect altogether from the grid. Alternatively, they may remain connected, and their energy potential can provide flexibility as a service to the grid. The behaviour of such prosumers depends on tariff policies, investment conditions, and environmental and operational conditions. In particular, the rapid improvement in commercial storage technologies has made it possible for prosumers to become fully electricity self-sufficient. We propose a decision-support framework accounting for these factors and combining strategic and operational planning optimization models. Using this framework, we show how changes in tariff policy may not only create a financial incentive for self-generation and self-consumption but may also push prosumers towards disconnecting from the grid. Our results motivate a thoughtful reconsideration of current schemes for the economic integration of prosumers in the energy system.

MIGUEL F. ANJOS is Chair of Operational Research at the School of Mathematics, University of Edinburgh, and Schöller Senior Fellow at the University of Erlangen-Nürnberg. His research interests are in the theory, algorithms and applications of mathematical optimization. He is particularly interested in the application of optimization to problems in power systems management and smart grids. He is the Founding Academic Director of the Trottier Institute for Energy at Polytechnique Montreal, and current President of the INFORMS Section on Energy, Natural Resources, and the Environment. He is a Fellow of the Canadian Academy of Engineering.

16:00 – 16:30

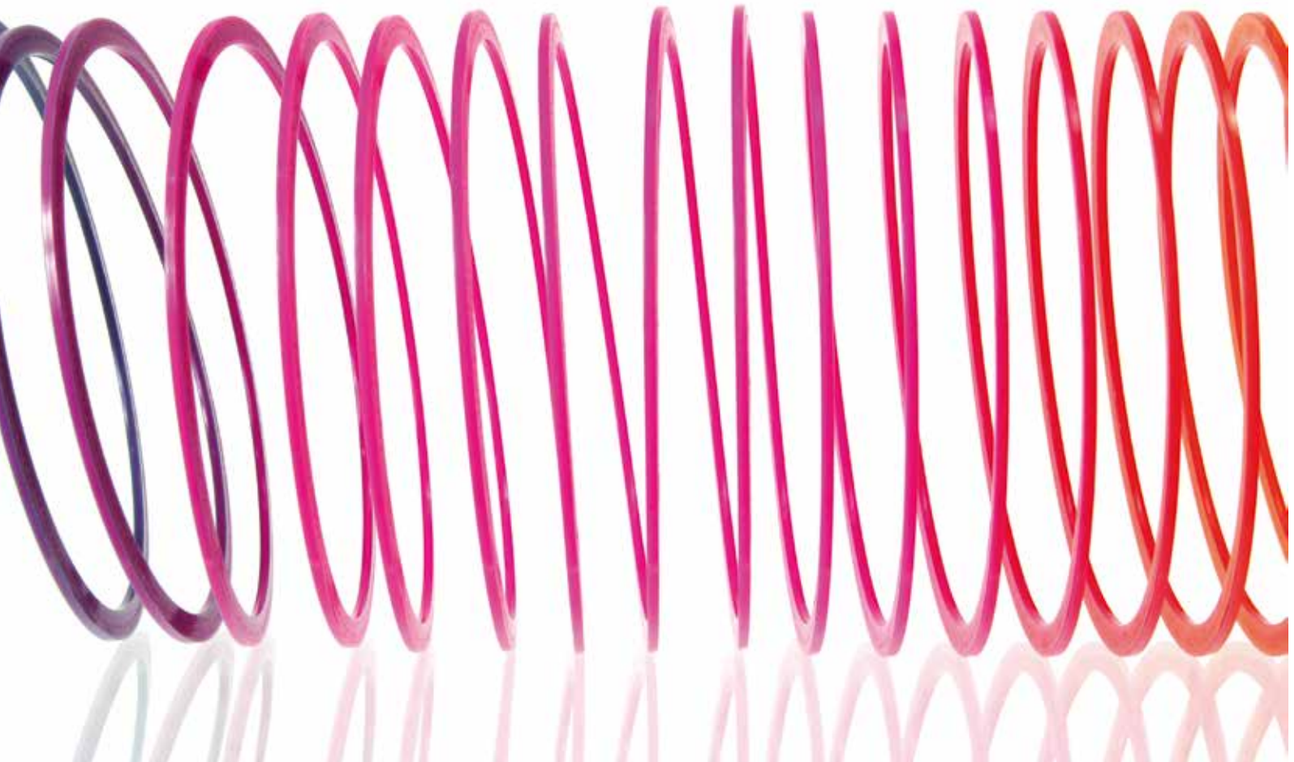
THE PSYCHOLOGY OF CLIMATE CHANGE

Prof. dr. L. (Linda) Steg

University of Groningen

Common approaches to encourage pro-environmental behaviour typically target extrinsic motivation, by offering incentives that change personal costs and benefits of behaviour. I will explain why such approaches are not always as effective as assumed. Next, I will discuss factors and strategies that can foster and secure intrinsic motivation to engage in pro-environmental behaviour. Intrinsically motivated people behave without being coerced or incentivised, even when pro-environmental behaviour is somewhat costly, as doing so is meaningful and makes them feel good.

LINDA STEG is professor of environmental psychology at the University of Groningen. She studies factors influencing sustainable behaviour, the effects and acceptability of strategies aimed at promoting sustainable behaviour, and public perceptions of technology and system changes. She is a member of the Royal Netherlands Academy of Sciences (KNAW) and the European Academy of Sciences and Arts. She is laureate of the Dutch Royal Decoration with appointment as the Knight of the Order of the Netherlands Lion, and laureate of the Stevin prize of the Dutch Research Council. She is lead author of the IPCC special report on 1.5°C and AR6, and participates in various interdisciplinary and international research programmes in which she collaborates with practitioners working in industry, governments and NGOs.



FLEXIBELE STATISTIEK

Van p-waarden naar e-waarden; Robuustere conclusies met minder data

De rigide natuur van p-waarden maakt ze steeds minder passend bij huidige mogelijkheden van dataverwerking: een groeiende dataset continu analyseren en data uitwisselen tussen verschillende instituten was nog nooit zo makkelijk. Een flexibeler alternatief zijn e-waarden, waarbij de e staat voor *evidence*. In dit artikel laten we zien hoe voor een breed scala aan scenario's van nulhypothesetoetsen in datastromen op simpele wijze e-waarden kunnen worden ontworpen, en hoe deze kunnen worden toegespitst op het zo snel mogelijk verzamelen van bewijs voor een alternatieve hypothese.

Rosanne J. Turner

De afgelopen jaren zijn er veel goede ontwikkelingen geweest op het gebied van data-infrastructuur voor wetenschappelijk onderzoek, zoals bijvoorbeeld de *Personal Health Train* en het FAIR-data initiatief (Van Soest et al., 2018; Wilkinson et al., 2016). Wetenschappers hebben nog nooit zoveel mogelijkheden gehad om veilig data uit te wisselen met collega's in andere instituten en om via dashboards data live te analyseren. Deze ontwikkelingen brengen interessante statistische uitdagingen met zich mee. Wat moet een onderzoeker bijvoorbeeld doen als hij na het behalen van zijn geplande steekproefgrootte een net niet significant resultaat behaalt, en een collega aanbiedt dat hij tien extra samples toe kan voegen? Of als de geplande steekproefgrootte nog niet behaald is, maar de onderzoeker tussendoor uit nieuwsgierigheid alvast een p-waarde uitrekent, die dan significant blijkt te zijn?

Sequentieel toetsen met klassieke p-waarden

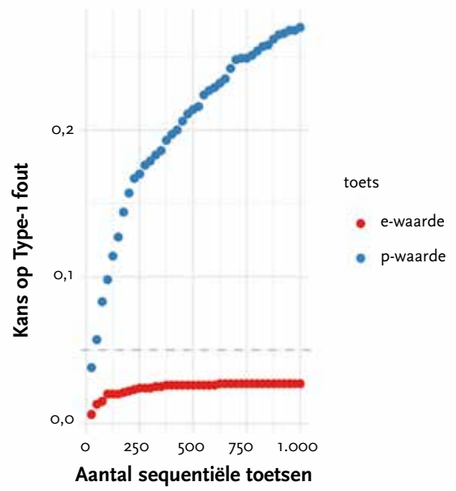
Peter Grünwald gaf eerder al in *STAtOR* een mooi overzicht van hoe het gebruik van klassieke nulhypothesetoetsen en p-waarden in (onder andere) dit soort situaties problematisch is (Grünwald, 2015). Klassieke toetsen zoals de gepaarde t-toets of Fishers exacte toets geven alleen een garantie op de kans om onterecht de nulhypothese te verwerpen, de Type-1 fout (*false positive rate*), als vooraf de steekproefgrootte exact wordt vastgesteld.

Stel dat we als statistiekconsultant een onderzoeker proberen te adviseren die de nulhypothese wil toetsen dat medicijn *a* een even grote kans op succes biedt als medicijn *b*. De onderzoeker heeft een dashboard voor hun studie opgezet en krijgt iedere keer als in beide groepen één patiënt het medicatietraject heeft afgerond een update. De data komen dus binnen in gebalanceerde blokken van groepsgrootte $n_a = 1$ in groep *a*, en $n_b = 1$ in groep *b*. Het liefst zou de onderzoeker op ieder moment dat het dashboard geüpdatet wordt een nulhypothesetoets doen, om zo snel mogelijk de studie af te kunnen ronden. In figuur 1 is geïllustreerd wat er zou gebeuren als we voor deze analyse Fishers exacte toets zouden gebruiken: onze kans om een Type-1 fout te maken blijft alsmaar stijgen naarmate we meer en meer datapunten verzamelen. Eigenlijk is onze hele statistische analyse oninterpreteerbaar geworden!

E-waarden

In figuur 1 is ook te zien dat met een analyse met zogenoemde e-waarden in plaats van p-waarden de Type-1 fout wel begrensd blijft. E-waarden zijn een alternatief voor p-waarden voor het doen van nulhypothesetoetsen, waarbij de e staat voor *evidence*. Zoals deze naam misschien al doet vermoeden zijn de e-waarden een maat voor bewijs voor de alternatieve hypothese in de data. Door de definitie van e-waarden blijven deze naar verwachting laag als data in werkelijkheid gegenereerd worden onder de nulhypothese: alle niet-negatieve *random variables* (kansvariabelen) met een verwachte waarde van hoogstens 1 onder alle verdelingen in de nulhypothese zijn e-waarden (Grünwald et al., 2019; Vovk & Wang, 2021).

Door deze definitie blijft het product van sequentieel verzamelde e-waarden zelf ook weer een e-waarde, ongeacht de methoden die de onderzoeker heeft gebruikt om te besluiten of de studie door moest lopen of stoppen (details in Grünwald et al., 2019). Dit betekent dat we onze onderzoeker zouden kunnen adviseren iedere keer als het dashboard geüpdatet is een e-waarde uit te reke-



Figuur 1. De geschatte kans op het maken van een Type-1 fout bij het toetsen van de nulhypothese dat de kans op succes in twee groepen hetzelfde is, onder sequentieel toetsen met Fishers exacte toets en e-waarden (aangepast uit Turner et al, 2021)

nen met deze nieuwe datapunten. En nu komt het mooie: uit de eigenschap dat e-waarden onder de nulhypothese een verwachte waarde van hoogstens 1 hebben volgt dat we ze kunnen gebruiken voor een nulhypothesetoets. We verwerpen de nulhypothese als de e-waarde groter is dan $1/\alpha$, wat ons een Type-1 fout garantie geeft van α, zelfs als we na iedere update opnieuw ons product van sequentiële e-waardes toetsen. Als onze onderzoeker bijvoorbeeld een significantieniveau van 0,05 aan wil houden, adviseren we de nulhypothese te verwerpen zodra het product van de sequentiële e-waarden groter is dan 20.[1]

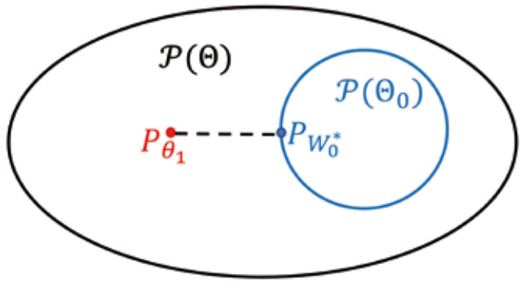
Zoveel mogelijk bewijs

E-waarden bieden dus die Type-1 fout garantie onder sequentieel toetsen, maar hoe gedragen ze zich als de alternatieve hypothese waar is? De lezer kon zich bij de brede definitie van e-waarden in de vorige paragraaf wellicht gelijk een aantal 'domme' keuzes voorstellen, bijvoorbeeld een *random variable* die altijd de waarde 1 aanneemt. In Grünwald et al. (2019) wordt voorgesteld e-waarden zo te kiezen, dat ze optimaal bewijs verzamelen voor de alternatieve hypothese in de sequentiële setting die we hierboven beschreven. De e-waarde die dit doet voor een puntalternatief $p_{\theta_1}^*$ kan gevonden worden door *reverse information projection*:

$$E^*(z^n) = \frac{p_{\theta_1^*}(z^n)}{p_{W_0^*}(z^n)}, \text{ zodat } D\left(P_{\theta_1^*} \parallel P_{W_0^*}\right) = \min_{W_0 \in W(\Theta_0)} D\left(P_{\theta_1^*} \parallel P_{W_0}\right). \quad (1)$$

Hierbij is D de Kullback-Leibler divergentie (of: relatieve entropie), een divergentiemaat uit de informatietheorie. De verzameling $\left\{P_{W_0} : W_0 \in W(\Theta_0)\right\}$, kan gezien worden als de verzameling van Bayesiaanse marginaalverdelingen voor de data z^n, ééntje voor elke prior W_0 op de nulhypothese Θ_0.[2] We kiezen dan de prior W_0^* zodat $P_{W_0^*}$ de marginaal over de nulverdelingen is die het 'minst verschilt van' of het 'dichtste bij' ons puntalternatief staat, zoals geïllustreerd (in Euclidische afstand) in figuur 2.

Het mooie aan vergelijking (1) is dat we een algemene uitdrukking hebben om voor elke combinatie van een alternatief en nulhypothese een goede e-waarde te definieren, maar het vinden van een oplossing lijkt wellicht niet triviaal. Voor een heel aantal toetsscenario's is ook nog geen analytische oplossing gevonden, en moet gebruik gemaakt worden van een numerieke benadering via optimalisatie, zoals geïllustreerd in Turner, 2019 en Lardy, 2021.



Figuur 2. De *reverse information projection* van een puntalternatief op de nulhypothese (figuur samengesteld analoog aan, als simplificatie van, Grünwald et al., 2019, figuur 1)

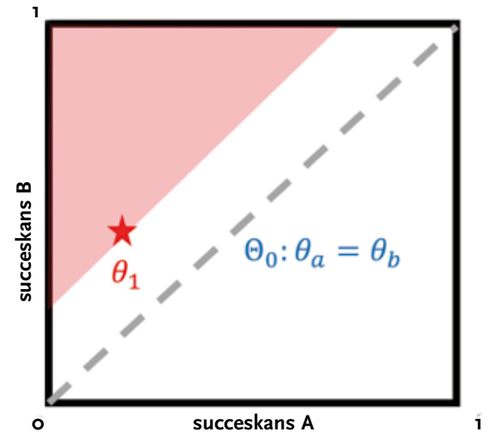
Een simpele oplossing voor data in twee groepen

Verrassend genoeg blijkt er voor een grote groep scenario's, namelijk waarbij we kansverdelingen in twee groepen vergelijken, een simpele generieke vorm van e-waarden te bestaan (Turner et al., 2021). Voor deze scenario's nemen we aan dat er een set aan kansverdelingen is, $P(\Theta) := \left\{P_\theta, \theta \in \Theta\right\}$, en dat datapunten $y_{i,a}$ en $y_{i,b}$ in groep a en b gegenereerd worden door een bepaalde verdeling uit deze set, P_{θ_a} en P_{θ_b} – dit is ons puntalternatief. Verder leggen we van tevoren vast hoeveel datapunten we in dit blok in elke groep gaan verzamelen: n_a in groep a, en n_b in groep b (tussen verschillende datablokken in mogen we deze aantallen wel veranderen). Indien we dan de nulhypothese willen toetsen dat de kansverdelingen in groep a en b gelijk zijn, versimpelt de e-waarde tot:

$$\frac{\prod_{i=1}^{n_a} p_{\theta_a}(y_{i,a})}{\prod_{i=1}^{n_a}\left(\frac{n_a}{n} p_{\theta_a}(y_{i,a}) + \frac{n_b}{n} p_{\theta_b}(y_{i,a})\right)} \frac{\prod_{i=1}^{n_b} p_{\theta_b}(y_{i,b})}{\prod_{i=1}^{n_b}\left(\frac{n_a}{n} p_{\theta_a}(y_{i,b}) + \frac{n_b}{n} p_{\theta_b}(y_{i,b})\right)}. \quad (2)$$

Deze formule kan zelfs verder versimpeld worden als de nulhypothese een bepaalde eigenschap bezit, convexiteit. Dan kunnen we de noemers in vergelijking (2) vervangen door de waarschijnlijkheid van de data onder een bepaalde kansverdeling uit onze set, $p_{\theta_0} = \frac{n_a}{n} p_{\theta_a} + \frac{n_b}{n} p_{\theta_b}$. In dit geval weten we ook dat deze e-waarde de beste e-waarde is om bewijs te verzamelen voor ons puntalternatief.

In het voorbeeld van onze onderzoeker is dit toevallig zo. Stel dat onze onderzoeker verwacht dat in groep a het medicijn succesvol is bij 20% van de patiënten, en dat het in groep b twee keer zo goed zal werken (de rode ster in figuur 3). De beste e-waarde voor dit alternatief ziet er dan als volgt uit: in de teller komt simpelweg de kans op de data in elke groep onder Bernoulli verdelingen met parameterwaardes 0,2 en 0,4, en in de noemer de kans op de gehele data onder een Bernoulli verdeling met parameterwaarde 0,3.



Figuur 3. De parameterruimte voor het nulhypothesetoets-scenario waarbij proporties in twee groepen vergeleken worden. De gestippelde lijn, Θ_0, geeft de parameterruimte van de nulhypothese aan

Leren van eerder geziene data

In de praktijk komt het niet vaak voor dat men genoeg voorkennis heeft om een de e-waarde toe te spitsen op een puntalternatief, maar bijvoorbeeld wel op een minimaal klinisch relevant verschil tussen de groepen (de roze driehoek in figuur 3), of een minimale odds ratio. In dat geval kan een prior op de parameterruimte geplaatst worden, en kan in combinatie met data gezien in eerdere blokken de beste e-waarde om bewijs voor de alternatieve hypothese van vorm (2) benaderd worden (details in Turner et al., 2021; en Turner et al., 2022). Het blijkt dat we met deze 'lerende' sequentiële methode net zo snel of soms zelfs sneller de nulhypothese kunnen verwerpen dan met klassieke hypothesetoetsen, waarbij we met de e-waarden altijd nog de mogelijkheid houden de dataset verder uit te breiden. Ook al gebruiken we hier priors, de aanpak is niet echt Bayesiaans. De Type-I foutgarantie geldt namelijk altijd, wat voor prior we ook kiezen – de keuze beinvloedt alleen hoe snel we de nulhypothese kunnen verwerpen als de nulhypothese niet waar is.

Kortom, e-waarden bieden mooie mogelijkheden data op een meer flexibele, maar nog steeds robuuste, manier te analyseren. Hopelijk kan het aanbieden van dit soort makkelijk aanpasbare statistische methoden bijdragen aan het verminderen van toekomstige *research waste* (Glasziou & Chalmers, 2018).

Deze tekst is gebaseerd op eerdere publicaties (Turner, 2019; Turner et al., 2021). Dank aan co-auteur Peter Grünwald.

Noten

1. Indien de onderzoeker heel graag een p-waarde wil rapporteren kan dat ook: 1 delen door de e-waarde geeft een conservatieve p-waarde.
2. Definitie van een Bayesiaanse marginaalverdeling: $P_{W_0}(Z^n) = \int w_0(\theta_0)\, P_{\theta_0}(Z^n) d\theta_0$.

Literatuur

Glasziou, P., & Chalmers, I. (2018). Research waste is still a scandal; An essay by Paul Glasziou and Iain Chalmers. *British Medical Journal, 363*.

Grünwald, P. D. (2015). Paranormale statistiek: Over de vele problemen met p-waarden, en een redelijk alternatief. *STAtOR, 4*.

Grünwald, P.D., De Heide, R. & Koolen, W. (2019). *Safe testing*. Preprint available on arXiv as arXiv:1906.07801.

Lardy, T. (2021). *E-values for hypothesis testing with covariates*. Master's thesis. Leiden University.

Turner, R. J. (2019). *Safe tests for 2 x 2 contingency tables and the Cochran-Mantel-Haenszel test*. Master's thesis, Leiden University.

Turner, R. J., Ly A., & Grünwald, P.D. (2021). *Two-sample tests that are safe under optional stopping, with an application to contingency tables*. Preprint available on arXiv as arXiv:2106.02693 [stat.ME].

Turner, R. J., Ly, A., Pérez, M. F., ter Schure, J. A., & Grünwald, P. D. (2022). Safestats: Safe anytime-valid inference. R package version 0.8.6, available at https://cran.r-project.org/web/packages/safestats/index.html..

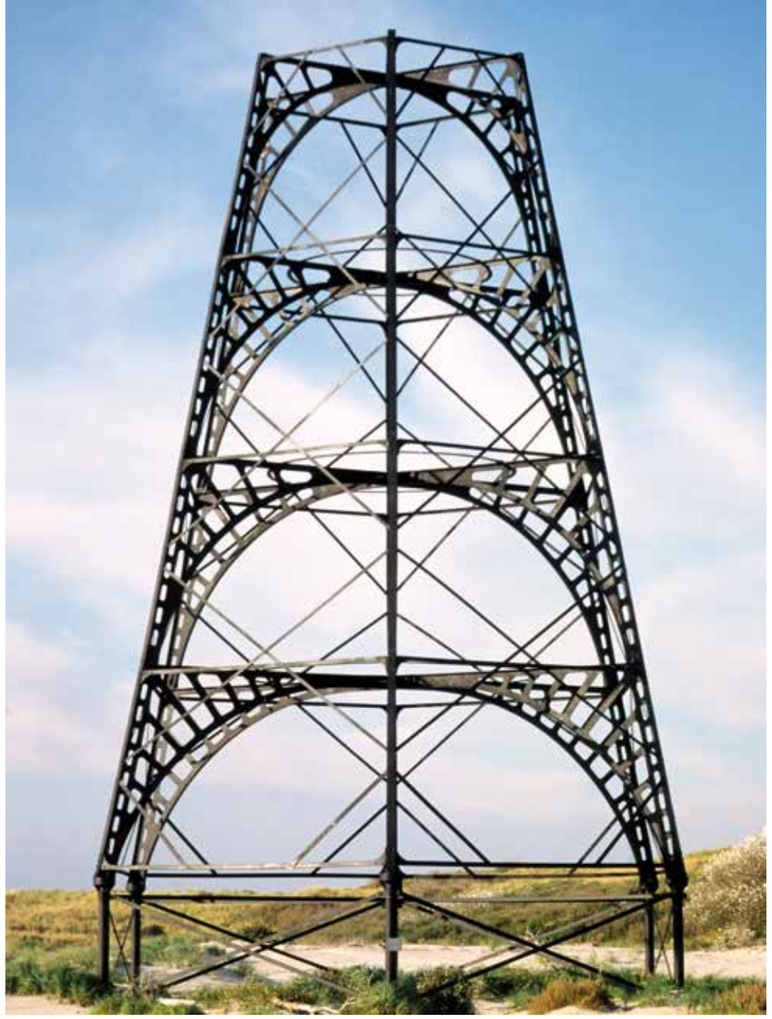
Van Soest, J., Sun C., Mussmann O. et al. (2018). Using the personal health train for automated and privacy-preserving analytics on vertically partitioned data. In *Building continents of knowledge in Oceans of Data: The future of co-created eHealth,* 581–585.

Vovk, V., & Wang, R. (2021). E-values: Calibration, combination, and applications. *The Annals of Statistics, 49*(3), 1736–1754.

Wilkinson, M.D, Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 3*, 160018.

Rosanne Turner is PhD student statistiek bij Professor Peter Grünwald en Professor Floortje Scheepers aan het Centrum Wiskunde en Informatica en het UMC Utrecht. In 2018 behaalde zij een PhD in de Geneeskunde en in 2019 sloot zij haar Master Statistical Science for the Life and Behavioural Sciences af aan de Universiteit Leiden, waarbij zij voor haar scriptie de Jan Hemelrijk Award van de VVSOR ontving.
E-mail: Rosanne.Turner@cwi.nl

Nederland heeft tienduizenden rijksmonumenten die je op een mooie dag met de fiets kunt bezoeken. Maar heb je je al eens afgevraagd wat de kortste fietsroute langs al deze monumenten is? De oplossing, met een lengte van 20.253.062 meter is, op moment van schrijven, het grootste optimaal opgeloste handelsreizigersprobleem ter wereld. In dit artikel leggen we meer uit over de achtergrond en de technieken voor het oplossen van deze uitdaging.



De Emder Kaap op Rottumeroog. Foto: Rijkswaterstaat | Rob Jungcurt CC

RECORD-HANDELSREIZIGERSPROBLEEM

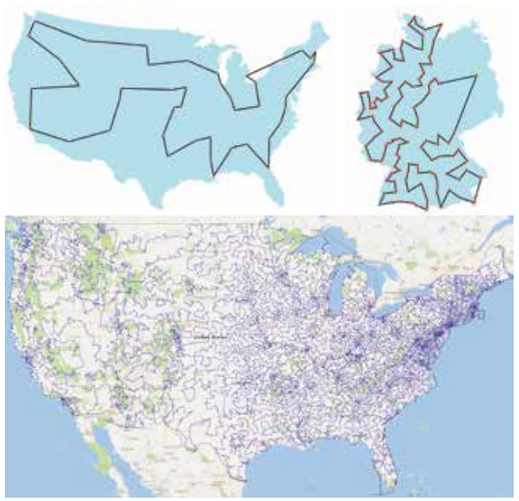
De kortste route langs 57.912 rijksmonumenten

Frans de Ruiter

In het najaar vindt de jaarlijkse Open Monumentendag plaats. In het tweede weekend van september openen duizenden monumenten gratis hun deuren voor publiek. Een drukbezocht evenement dat natuurlijk de nodige voorbereiding vergt van deze bijzondere locaties. Een lange aanloop was er ook voor het team van CQM, Bill Cook (hoogleraar aan de Universiteit van Waterloo in Canada) en Keld Helsgaun (Universiteit van Roskilde in Denemarken). Samen zijn zij de uitdaging aangegaan om de kortste fietsroute te vinden waarmee je alle 57.912 locaties kunt bezoeken. Om dit te kunnen bepalen moet je eerst alle afstanden tussen de locaties weten en vervolgens nog een van de moeilijkste wiskundige problemen oplossen: het handelsreizigersprobleem.

Historische records

Het probleem om de kortste rondrit te vinden langs een aantal locaties, gegeven de afstanden tussen alle punten, wordt het handelsreizigersprobleem genoemd (*traveling salesman problem*). Het vinden van een oplossing voor dit probleem, én bewijzen dat het de kortste is, wordt al decennia als een grote uitdaging gezien. In 1954 werd het eerste record gevestigd, toen Dantzig, Fulkerson en Johnson een methode beschreven voor de kortste rondrit langs 49 Amerikaanse steden over autowegen. In 1977 werd het record aangescherpt naar 120 steden voor een rondrit door wat toen West-Duitsland was door Groetschel. Na deze tijd werden de meeste records voor het handelsreizigersprobleem bedacht voor routes die geometrisch waren: locaties werden punten op een stuk papier en de afstanden een rechte lijn tussen de punten. De logische reden hiervoor was dat het lastig wordt om echte afstanden te bepalen op wegenkaarten. In die tijd werden afstandstabellen uit atlassen gebruikt, die alleen de afstanden tussen grote steden bevatten. Voor slechts 100 steden heb je $100 \times 99 = 9.900$ afstanden nodig om alle paren te krijgen. Zelfs als je aanneemt dat van A naar B reizen dezelfde afstand is als andersom, van B naar A reizen, moet je tabel nog steeds duizenden afstanden bevatten. Enkele jaren geleden werden door een team onder leiding van Bill Cook wel weer grote problemen op wegenkaarten opgelost. In 2016 is de optimale route gevonden voor 49.603 Amerikaanse monumenten en in 2018 een ludieke kortste kroegentocht langs alle 49.687 pubs in het Verenigd Koninkrijk. Al deze routes maakten gebruik van loopafstanden uitgerekend door Google Maps.

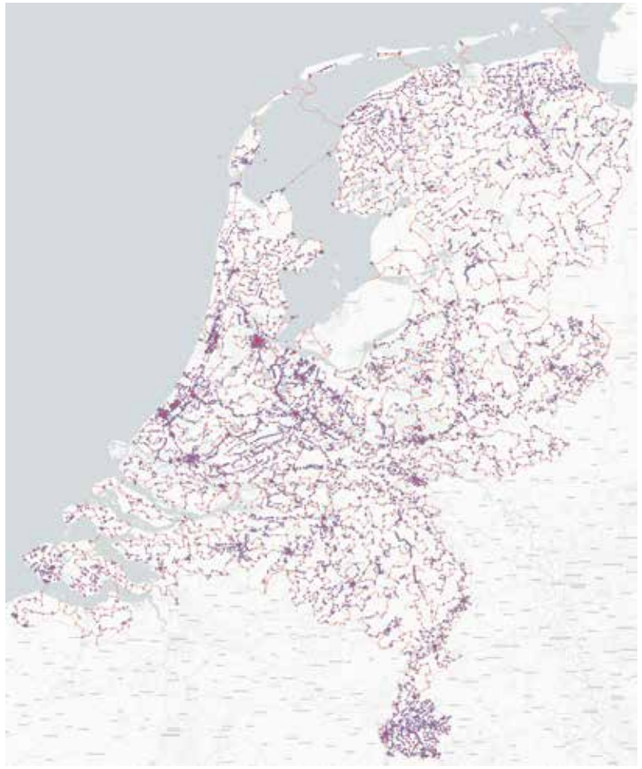


(linksboven) De eerste recordroute uit 1964 door 49 Amerikaanse steden.
(rechtsboven) Recordroute uit 1977 door 120 steden in West-Duitsland.
(onder) Recordroute uit 2016 langs 49,603 Amerikaanse monumenten.

Data over 57.912 monumenten

49.603 monumenten in Amerika is natuurlijk een erg indrukwekkend aantal. Maar we moeten niet vergeten dat we in Nederland ook een hele rijke historie hebben met veel Nederlands cultureel erfgoed. Uit het Rijksmonumentenregister hebben we een lijst van 63.287 monumenten kunnen halen. Hierop staan bekende rijksmonumenten zoals het Museumplein en de Zaanse Schans, maar ook oude boerderijen, vuurtorens en honderden voorgevels van huizen langs de Amsterdamse grachten. In de datasets komen echter maar 57.912 locaties voor. Dit heeft twee belangrijke redenen. Allereerst zijn er zo'n 5.000 locaties waarvan de exacte coördinaten overeenkomen met een ander monument in de dataset. Verder zijn er zo'n 500 monumenten waarvan de locatie helemaal niet bekend is. Voor het berekenen van de afstanden is ook een wegenkaart nodig van goede kwaliteit. Aangezien we de kortste fietsroute willen bepalen, moeten de afstanden ook wel per fiets afgelegd kunnen worden. Even snel de A2 meepakken zit er dus niet in. Hiervoor kon gelukkig gebruikt worden gemaakt van het fietsennetwerk van Geodan, specialist op het gebied van *location intelligence*, dat de onderliggende kaart heeft aangeleverd om de afstanden te bepalen. De tweede uitdaging is dat de monumenten meestal niet direct op een fietspad liggen. Alhoewel het overgrote deel prima te bereiken is per fiets, zijn er toch bijna 1.000 monumenten die meer dan 200 meter van

Locaties van rijksmonumenten in Nederland; op sommige locaties is meer dan één monument te zien



De record oplossing voor een handelsreizigersprobleem: de kortste fietsroute langs alle Nederlandse rijksmonumenten

een fietspad liggen, twintig monumenten moet je van meer dan een kilometer ver weg bekijken. En dan is er nog één monument waar echt een verrekijker voor nodig is omdat het meer dan zeven kilometer van een fietspad af ligt: de Emder Kaap op het beschermde Rottumeroog (zie de foto op pagina 14 en de kaart op pagina 16). Alle locaties van de monumenten worden daarom geprojecteerd op de fietswegenkaart zodat je legaal met de fiets op het dichtstbijzijnde punt uit komt.

1.676.870.916 afstanden berekenen

Het handelsreizigersprobleem moet de kortste route vinden *gegeven de afstanden tussen alle paren van locaties*. Als je uitgaat van symmetrische afstanden, waar van A naar B dezelfde afstand is van B naar A, dan heb je afstandstabel met 57.912 x 57.911 / 2 = 1.676.870.916 regels nodig. Voor de vorige records, de kroegentocht in de UK en de Amerikaanse monumentenroute, heeft het team van Bill Cook gebruik gemaakt van Google. Echter, dit zorgde voor grote beperkingen, omdat afstanden niet allemaal in een keer kunnen worden uitgerekend via de Google services. Zo konden ze per dag maar enkele duizenden afstanden uitrekenen. CQM heeft eigen algoritmes ontwikkeld die heel snel grote afstandstabellen kunnen uitrekenen. Om meer dan een miljard routes snel te berekenen, moet je de kaart dan wel op een slimme manier voorbewerken. Een van de manieren om dat te doen, is door het opsplitsen van de kaart in kleinere stukjes. In Nederland heb je vaak rivieren of meren die Nederland op een natuurlijke manier verdelen in deze stukjes. Op een kleiner niveau kunnen slimme graaf partitionerings-algoritmen deze structuren ook herkennen in woonwijken en grachtengordels. Voor een afstandsberekening tussen A en B kun je dan elk component waar deze twee punten niet in zitten, vervangen door een veel makkelijkere graaf met minder verbindingen. Op die manier kan de afstandsbepaling een stuk sneller gaan. Zo snel zelfs, dat de afstandsberekening met de slimme opknipping slechts twee uur duurde op een snelle computer van CQM

Het vinden van de kortste route

Nu alle ruim 1,6 miljard afstanden bekend zijn, kan een kortste route worden gevonden. De meest succesvolle heuristiek voor dit probleem is de zogeheten Lin-Kernighan-Helsgaun (LKH) heuristiek. Hierbij wordt een bepaalde tour steeds verbeterd om te kijken of een klein aantal verbindingen, bijvoorbeeld vier, op een andere manier kunnen worden gelegd. Binnen drie dagen werd hierdoor een tour gevonden met een lengte van 20.253.564 meter. Na wat verbeteringen van Keld Helsgaun door een parallelle implementatie werd op 27 september een tour gevonden die nog eens 186 meter korter was. Dit is een hele goede tour, maar op dat moment weet je nog niet of het nog beter kan. De 'Concorde' software van Bill Cook voert verschillende stappen uit om te laten zien dat er geen betere tour bestaat. Eerst worden kleinere subproblemen opgelost waardoor hij kon aantonen dat de kortste route niet langer dan 20.251.395 meter kon zijn. Dat is dus nog een verschil van 1.983 meter. Je zou kunnen zeggen dat dit niks is op 20.000 kilometer, namelijk maar 0,01%. Maar om het record te halen, moet het verschil natuurlijk teruggebracht worden tot nul meter. Via een uitgebreide brand-and-bound search, waarbij je heel veel verschillende subproblemen oplost, kun je uitsluiten dat er een beter route bestaat. Dit koste wel héél veel tijd. De berekeningen werden uitgevoerd op een netwerk van 10 servers van de Universiteit van Waterloo met ieder 32 processorkernen. Daardoor duurden de berekening 'slechts' drie maanden. Als je het op één processor had moeten uitrekenen, dan had je dit 96,9 jaar aan rekentijd gekost.

The bigger picture

Het is natuurlijk fijn dat er nu een fietsroute langs alle monumenten in Nederland is van 20.000 kilometer. Helemaal geruststellend is het dan ook om te weten dat je geen meter te veel hoeft te trappen omdat er een kortere route zou bestaan. De maatschappelijke relevantie is echter groter dan enkel voor een paar fanatieke fietsers. Deze technieken worden gebruikt door bedrijven als Amazon om de kortste routes te bepalen voor het afleveren van pakketjes. Daarnaast worden de algoritmes om snel afstanden te berekenen door CQM voor haar klanten gebruikt als onderdeel van tal van planningsvraagstukken die snel moeten worden opgelost. Zo gebruiken ze de algoritmiek bijvoorbeeld voor het dagelijks plannen van 5.000 tot wel 15.000 taxiritten voor de volgende dag.

Literatuur

Cook, W. J. (2011). *In pursuit of the traveling salesman*. Princeton University Press.

Brandhof, A. van den. (2021). *In 20.000 kilometer heb je alles gezien*. *NRC*, 9 september..

https://monumentenregister.cultureelerfgoed.nl/monumentenregister

NL monuments webpage and data: https://www.math.uwaterloo.ca/tsp/nl/

Frans de Ruiter is gepromoveerd op het gebied van robuuste optimalisatie bij Tilburg University. Sinds 2017 is hij consultant bij CQM in Eindhoven. Hier werkt hij aan grote optimalisatie en planningsprojecten. Een dag in de week is hij als research scientist werkzaam bij de Universiteit van Wageningen. E-mail: deruiter@cqm.nl

De Emder Kaap op Rottumeroog ligt meer dan zeven kilometer van een fietspad

PEILINGPRAKTIJKEN

Een Twitter-peiling? Doe maar niet!

Figuur 1. De Twitter-peiling van de provincie Zuid-Holland, 13 juni 2021

Het is mogelijk om met Twitter een kleine opiniepeiling uit te voeren. Dat lijkt een interessante mogelijkheid om goedkoop, snel en eenvoudig gegevens te verzamelen en zo onderzoek te doen. Zo'n Twitter-peiling heeft echter nogal wat methodologische tekortkomingen. Daarom is het beter om dit soort-peilingen te vermijden. Helaas kom je ze toch nog regelmatig tegen. Een voorbeeld is een peiling van de provincie Zuid-Holland die op zondag 13 juni 2021 op Twitter verscheen (figuur 1). Deze peiling had maar één vraag. Je moest reageren op de stelling 'Ik blijf vaker thuiswerken, ook als dit niet meer hoeft'. En er waren twee mogelijke antwoorden: 'Eens' en 'Oneens'. Op het moment van bekijken van deze peiling hadden 42 personen de vraag beantwoord. De peiling bleef kennelijk nog een dag zichtbaar op Twitter ('1 dag resterend').

Tekortkomingen

Een Twitter-peiling heeft een aantal ernstige beperkingen.

Slechts één vraag

In de eerste plaats kun je maar één vraag stellen. Daarmee is het wel een heel korte peiling. Je kunt wel meer vragen stellen maar dat moet je dan doen in een reeks opeenvolgende losse tweets. Je hebt dan geen controle over de volgorde van het beantwoorden van de vragen. Verder kun je vragen overslaan. En je kunt ook geen samenhang tussen de antwoorden op verschillende vragen onderzoeken. De Twitter-peiling van de Provincie-Zuid-Holland had inderdaad maar één vraag.

Te korte teksten

Een tweede tekortkoming is dat de tekst van de vraag niet langer mag zijn dan 280 tekens. Dat kan te weinig zijn voor wat langere vragen of toelichtende teksten. En de tekst van de antwoordmogelijkheden mag niet langer zijn dan 25 tekens. Dat is in veel gevallen te weinig. De beperkte beschikbare ruimte was voor de peiling van de Provincie-Zuid-Holland wel voldoende.

Alleen een gesloten vraag

Een derde tekortkoming is dat je alleen maar een gesloten vraag met maximaal vier mogelijke antwoorden kunt stellen. Gesloten vragen kunnen natuurlijk veel meer antwoordmogelijkheden hebben. Denk maar eens aan een vraag naar politieke voorkeur. De respondenten moeten dan kunnen kiezen uit een reeks politieke partijen. En dat zijn er meestal veel meer dan vier. En er zijn ook nog andere typen vragen, zoals de open vraag ('Wat voor soort werk doet u?'), de numerieke vraag ('Hoe oud bent u?') en de gesloten vraag met meer dan één antwoord ('Welke kranten leest u?'). Dat kan allemaal niet in een Twitter-peiling. De vraag van de Twitter-peiling van de provincie Zuid-Holland had gelukkig maar twee mogelijke antwoorden, al kun je je afvragen of nog niet een derde mogelijkheid had moeten zijn ('Ik weet het nog niet').

Onduidelijke doelgroep

Een vierde tekortkoming van een Twitter-peiling is dat je niet een geschikte doelgroep voor je peiling kunt kiezen. Alleen je volgers op Twitter zien je peiling. Weliswaar kunnen ook andere twitteraars bij de betreffende tweet terechtkomen, maar dan moeten ze die wel bewust opzoeken, bijvoorbeeld via een hashtag. Welke doelgroep ben je zo aan het onderzoeken? Is dat een interessante doelgroep? Waarschijnlijk niet. Wat nu als je de vraag wilt voorleggen aan alle Nederlanders, of alle leraren, of alle politieagenten, of alle mensen in de zorg? Dit is niet mogelijk. De vraag zal dus steeds weer zijn: voor welke doelgroep is de Twitter-peiling nu eigenlijk representatief?

Ook bij de Twitter-peiling van de Provincie Zuid-Holland was de doelgroep niet duidelijk aangegeven. Ging het om alle Nederlanders? Of alle Nederlanders met een baan? Of alle Nederlanders van 18 jaar en ouder? Of alle Zuid-Hollanders? Of alle Zuid-Hollanders met een baan? Of alle Zuid-Hollanders van 18 jaar en ouder? Er was helaas geen uitleg voor wie de peiling bedoeld was.

Slechts zeven dagen

Een vijfde tekortkoming is dat een Twitter-peiling niet langer dan 7 dagen open kan staan. De standaard-instelling is 1 dag, maar je kunt die periode uitbreiden tot maximaal 7 dagen. Dat kan in sommige situaties wel eens te kort zijn. Zo mis je mogelijk heel wat respondenten.

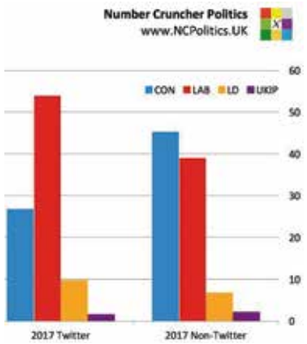
Niet representatief

Een zesde tekortkoming is het gebrek aan representativiteit. Wie doen er mee aan een Twitter-peiling? Dat zijn mensen met internet. En ze moeten ook een Twitter-account hebben. Verder moeten ze de maker van de peiling via Twitter volgen. Of ze moeten hem bewust opzoeken, bijvoorbeeld via een hashtag. En als ze de peiling zien, moeten ze ook nog spontaan besluiten mee te doen. Dit selectieproces zal zeker geen representatieve steekproef opleveren. Het is dus geen aselecte steekproef.

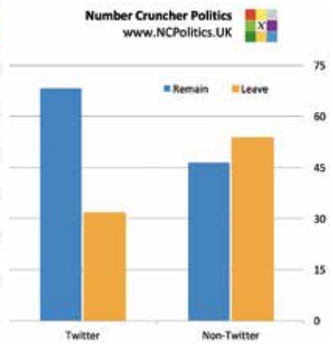
Die representativiteit kan nog verder worden aangetast als groepjes mensen samen besluiten om gericht de vraag op Twitter te beantwoorden en zo proberen de uitkomst van de peiling te manipuleren.

Wegen niet mogelijk

Een zevende tekortkoming is dat je een Twitter-peiling niet kunt wegen. Je zou kunnen proberen een Twitter-peiling te corrigeren voor het gebrek aan representativiteit. Dat zou kunnen met een weging. Daarvoor moet je in je peiling weegvariabelen bij de respondenten meten. Voor die variabelen moet je ook de verdeling in de hele doelgroep kennen. Voorbeelden zijn geslacht, leeftijd, regio (bijvoorbeeld keuze tussen stad of platteland), opleidingsniveau en stemgedrag bij de vorige verkiezingen. Dan kun zien welke groepen onder- en oververtegenwoordigd zijn. Door het toekennen van gewichten kun je vervolgens daarvoor corrigeren. Helaas, een Twitter-peiling heeft maar één vraag. Dus je kunt geen weegvariabelen meten. Daardoor is wegen onmogelijk. Het gebrek aan representativiteit blijft.



Figuur 2. De relatie tussen Twitter-gebruik en stemgedrag. Bron: Matt Singh, *British Election Study*



Figuur 3. De relatie tussen Twitter-gebruik en de mening over Brexit. Bron: Matt Singh, *British Election Study*

Een Brits voorbeeld

Het grootste probleem van een Twitter-peiling is toch wel het gebrek aan representativiteit. Matt Singh van de website Number Cruncher Politics laat dit nog eens duidelijk zien aan de hand van een Brits voorbeeld.

Op 8 juni 2017 waren er parlementsverkiezingen in het Verenigd Koninkrijk. Vlak daarna werd de British Election Study (BES) uitgevoerd. In deze peiling zaten, onder anderen, vragen over stemgedrag bij de verkiezingen en ook over het gebruik van sociale media. De BES is een goed onderzoek. De steekproef is netjes aselect geloot uit alle adressen in de VK, interviewers verzamelden de gegevens met face-to-face interviews, de respons was relatief hoog, en de onderzoekers wogen de uitkomsten. Daarom kun je spreken van een representatief onderzoek.

Matt Singh heeft gegevens uit de BES gebruikt om na te gaan of er verband is tussen stemgedrag en gebruik van sociale media. Dat verband is er zeker. De grafiek in figuur 2 vergelijkt hij het stemgedrag van Twitter-gebruikers met het stemgedrag van respondenten die geen Twitter gebruiken.

Duidelijk is te zien dat Labour behoorlijk oververtegenwoordigd is bij de twitteraars. 54% heeft Labour gestemd en slechts 27% Conservatief. En het patroon is juist andersom bij mensen die niet twitteren. Daar zijn de conservatieven in de meerderheid met 45%, terwijl maar 39% Labour stemde. Kortom, in een Twitter-peiling zit te veel Labour en te weinig Conservatieven.

Matt Singh laat aan de hand van een voorbeeld zien waartoe dit kan leiden. Zie de grafiek in figuur 3. Bij de twitteraars is een grote meerderheid (68%) tegen Brexit (*Remain*). Slechts 32% wil een Brexit (*Leave*). Bij de niet-twitteraars hebben de voorstanders van Brexit echter de meerderheid (54% tegen 46%).

Kortom, het is onverstandig om Twitter-peilingen te gebruiken, want ze zijn niet representatief. Een Twitter-peiling? Niet doen!

Jelke Bethlehem werkte bij het CBS en is emeritus hoogleraar aan de Universiteit Leiden. Hij is een expert op het gebied van steekproeven, vragenlijsten en weergave van onderzoeksresultaten. Deze onderwerpen behandelt hij regelmatig in zijn blog.
E-mail: mail@jelkebethlehem.nl

Cognitieve achteruitgang detecteren met behulp van spraakherkenning via sensoren

In de enquête SHARE meten interviewers onder andere cognitieve achteruitgang van 50-plussers in pan-Europese landen. Cognitieve achteruitgang kan bijvoorbeeld een indicatie zijn van beginnende dementie. In het kader van de Sensor Data Challenge hebben wij onderzoek gedaan naar alternatieven om kwaliteit van spraak op een objectieve manier te meten door middel van sensoren.

Rianne Drijver, Sigrid van Hoek, Jonas Klingwort & Rob Willems

Op 22 en 23 april 2021 heeft het Centraal Bureau voor de Statistiek (CBS) samen met De Haagse Hogeschool, het Rijksinstituut voor Volksgezondheid en Milieu (RIVM) en de Universiteit Utrecht voor de derde keer de Sensor Data Challenge georganiseerd. Tijdens deze 24 uur durende hackathon werkten studenten en professionals – met een achtergrond in statistiek, data science, sensor studies, engineering en IT – in acht kleine groepen van 3 tot 5 personen aan verschillende challenges met behulp van sensoren. De challenges gingen over specifieke problemen op het gebied van gezondheid, veiligheid en tijdsgebruik.

Vanwege de coronamaatregelen werd de hackathon online gehouden. De vorige twee edities van de Sensor Data Challenge waren op locatie, te weten bij het CBS en op de Haagse Hogeschool. Teams waren daardoor flexibeler in de keuze voor sensoren vanwege de mogelijkheid om hardware te lenen. Deze keer kregen de deelnemers vooraf een sensorkit van Arduino. Deze kit maakt het mogelijk om automatisch verschillende variabelen zoals lichaamstemperatuur en vochtigheid te meten en te analyseren.

Sensoren worden steeds lichter en goedkoper. Door continue innovaties in Sensor Technology worden de metingen ook van steeds betere kwaliteit. Bovendien hebben waarnemingen met behulp van sensoren door het objectieve karakter de potentie uitkomsten van een vragenlijst aan te vullen of zelfs gedeeltelijk te vervangen. Vandaar dat onderzoeksinstellingen waaronder het CBS al enige jaren onderzoek doen naar de toepasbaarheid van deze sensoren. Zo onderzoekt het CBS bijvoorbeeld de mogelijkheid om statistieken te verbeteren met behulp van Weigh-in-Motion sensoren, gegevens van verkeerslussen and Automatic Identification System (AIS) (zie [1, 2, 3]). Als deel van dit onderzoek wordt de Data Challenge georganiseerd.

Challenges

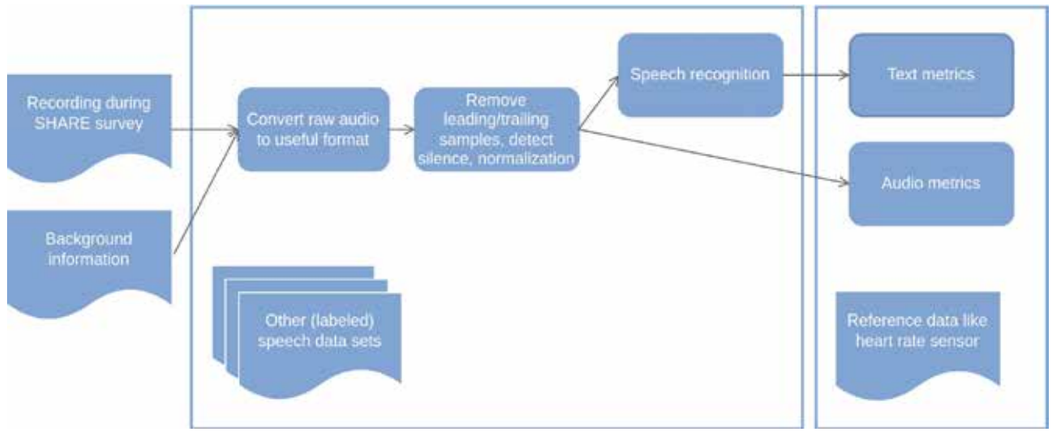
In de challenge van 2021 werden de volgende vier onderwerpen aangedragen door een vijftal Nederlandse organisaties. Ze werden vervolgens verdeeld over de acht deelnemende groepen:

- Het meten van fysieke en mentale stress (aangedragen door het Ministerie van Defensie)
- Het meten van tijdsgebruik binnenshuis (aangedragen door het CBS)
- Het meten van blootstelling aan gevaarlijke stoffen (aangedragen door het RIVM en de Arbo Unie)
- Het meten van kwaliteit van spraak en sociale interactie (aangedragen door SHARE)

Wij gingen aan de slag met het laatste onderwerp: het meten van kwaliteit van spraak en sociale interactie. De Survey of Health, Aging, and Retirement in Europe (SHARE) gebruikt op dit moment een vragenlijst om de kwaliteit van spraak te meten onder 50-plussers. Met interviews gehouden in 28 verschillende Europese landen en Israël is dit de grootste pan-Europese studie die het mogelijk maakt om microdata op het gebied van publieke gezondheid en sociaaleconomische levensomstandigheden longitudinaal en over landen heen te vergelijken. SHARE doet onderzoek naar vroege tekenen van cognitieve achteruitgang die bijvoorbeeld kunnen wijzen op voortekenen van dementie. Het doel van onze challenge was om een objectieve manier te vinden om spraakkwaliteit te schatten met behulp van een meting die geïntegreerd kan worden in het SHARE-interview. Als verdere eis werd gesteld dat de spraakkwaliteit-indicator onafhankelijk moet zijn van kennis, woordenschat en taalgebruik van de respondent.

Opzet van de spraakkwaliteit-indicator

Het ontwikkelen van een spraakkwaliteit-indicator is een ingewikkelde taak. Gelukkig zijn er veel akoestische indicatoren beschikbaar om spraakkwaliteit te meten (zie bijvoorbeeld [4, 5]). Wij hebben een programma in de veelgebruikte computertaal Python geschreven dat geluid opneemt middels de microfoon van de laptop en vervolgens de snelheid van het spreken (`speech rate') en het



Figuur 1. Schema om de spraakkwaliteit-indicatoren te berekenen

gemiddelde aantal lettergrepen per seconde berekent. Hiervoor is een Speech-to-Text Classifier nodig. Vanwege de beperkte tijd hebben we de Google Cloud Speech API gebruikt, die in veel talen (waaronder Nederlands) beschikbaar is. De stappen om de indicatoren te berekenen zijn weergegeven in figuur 1. Naast deze twee indicatoren zouden zogenaamde *filled pauses* zoals 'uhm' en correcties tijdens het spreken kunnen duiden op cognitieve achteruitgang. We hebben een begin gemaakt aan het berekenen van de lengte en het aantal *filled pauses* door de geclassificeerde tekst te vergelijken met de geluidsgolven.

Naast het analyseren van de stem kan er ook worden gekeken naar andere lichamelijke factoren. Temperatuur, transpiratie en hartslag zijn factoren die kunnen veranderen wanneer iemand stress ervaart. Wanneer mensen moeite hebben met het lezen van een tekst of het beantwoorden van vragen, kan de lichaamstemperatuur omhooggaan, kunnen ze meer gaan transpireren of kan de hartslag omhoog gaan. Daarom gebruiken we sensor-gebaseerde metingen als aanvulling op de spraakkwaliteit-indicatoren. We hebben daarom tijdens de challenge een handschoen samengesteld waarin zich sensoren bevinden die temperatuur, hartslag en galvanische huidreactie meten. De sensor die galvanische huidreactie meet detecteert transpiratie in de hand. Figuur 2 toont de handschoen die wij tijdens de challenge hebben gebruikt. De sensoren in de handschoen kunnen uitgelezen worden met een Arduino en de metingen kunnen bedraad of wireless verstuurd worden naar een computer waar de data opgeslagen en geanalyseerd worden.

Figuur 2. Handschoen met sensoren voor het meten van temperatuur, hartslag en galvanische huidreactie

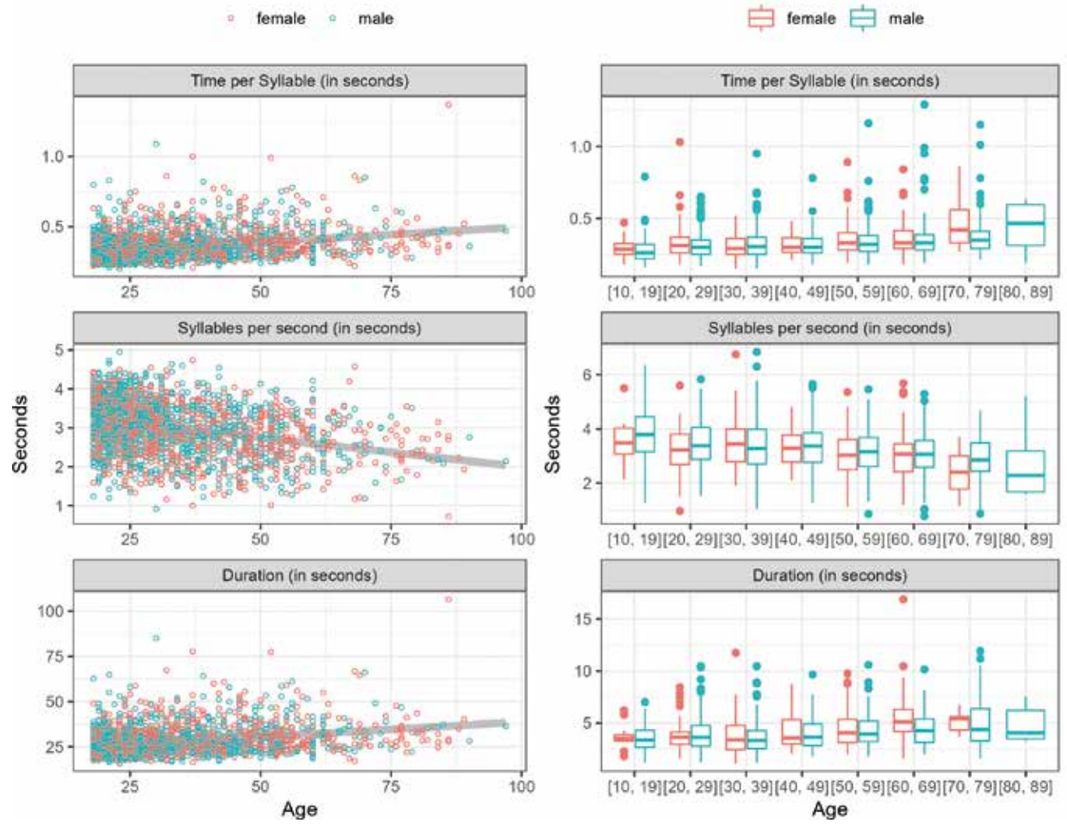
We stelden voor onze indicatoren in SHARE te integreren door deelnemers een korte passage te laten lezen en een kort verhaal te laten verzinnen (bijvoorbeeld door deelnemers te laten kiezen uit een reeks onderwerpen/afbeeldingen) zoals [6]. Dat laatste vereist het samenstellen en uitspreken van zinnen, en is daardoor cognitief gezien meer uitdagend. Op basis van de tekst en de opgenomen audio kunnen vervolgens de indicatoren berekend worden. Voor het SHARE interview begint, kan de desbetreffende persoon de handschoen aandoen, vervolgens kunnen er basiswaarden gemeten worden om later te vergelijken met de waarden die tijdens het interview gemeten zijn. Doordat de stem van de persoon en de sensoren allebei worden opgenomen kan de correlatie tussen de tekst, stem, temperatuur, hartslag en activiteit van de zweetklieren in de hand worden bepaald. De sensoren leveren op deze manier extra informatie over een eventuele cognitieve achteruitgang.

Indicatoren

In deze paragraaf illustreren we onze indicatoren met behulp van twee openbaar beschikbare datasets. De eerste dataset bestaat uit opnames van personen die allen dezelfde korte passage lezen, de tweede dataset bestaat uit opnames van personen die verschillende zinnen met een verschillend aantal woorden en verschillende inhoud lezen (de dataset zijn beschikbaar via [7, 8]). De resultaten van de indicatoren zijn weergeven in figuur 3. De drie spraakkwaliteit-indicatoren zijn afgezet tegen de leeftijd van respondenten, uitgesplitst naar geslacht. In beide datasets is te zien dat het aantal lettergrepen per seconde afneemt met de leeftijd, terwijl de tijd per lettergreep en de lengte van de spreektijd een lichte toename vertonen.

We concluderen dat onze methodologie in staat is onderscheid te maken tussen spraakkwaliteit van verschillende leeftijdsgroepen in de dataset. Hoewel het nog verder ontwikkeld en gevalideerd dient te worden, concluderen we dat onze indicatoren potentie hebben om te gebruiken voor het meten van in ieder geval dit aspect van kwaliteit van spraak. De betrokkenen partijen in SHARE bleken geïnteresseerd te zijn in onze oplossing en vroegen om verdere uitwerking.

De eerste plek van de challenge was behaald door een team van studenten dat voorstelde NFC-chips te gebruiken om tijdsgebruik binnenshuis te meten. Het elegante



Figuur 3. De drie spraakkwaliteit-indicatoren naar leeftijd en geslacht op basis van de eerste dataset (links) en de tweede dataset (rechts)

idee, samen met het werkende prototype en de presentatie maakte dit team tot winnaar. Het CBS zal dit idee verder oppakken en de mogelijkheden onderzoeken om het ook in de praktijk in te zetten.

Onze dank gaat uit naar Gerard Draadjer (Student Bedrijfskunde, Haagse Hogeschool), die deel uitmaakte van het team tijdens de Sensor Data Challange.

Referenties

[1] Klingwort, J., Buelens, B., & Schnell, R. (2019). Capture-recapture techniques for transport survey estimate adjustment using permanently installed highway sensors. *Social Science Computer Review, 39*(4). 527–542. doi: 10.1177/0894439319874684.

[2] Puts, M. J. H., Daas, P. J. H., Tennekes, M., & de Blois, C. (2019). Using huge amounts of road sensor data for official statistics. *IMS Mathematics, 4*(1), 12–25. doi: 10.3934/Math.2019.1.12.

[3] Consten, A., Puts, M., de Witt, T., Bisoti, E., Papandreou, C. P. K., Bis, M., Bliska, A., & Langsrud, Ø. (2017). *ESSnet Big Data, Work Package 4 AIS Data, Milestone 4.10 Progress and technical report of 1st internal WP-Meeting*. https://ec.europa.eu/eurostat/cros/sites/default/files/WP4_Milestone_4.10_2017_11_13.pdf.

[4] Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*(2), 145–164.

[5] Weiner, J., & Schultz, T. (2018). Selecting features for automatic screening for dementia based on speech. *International Conference on Speech and Computer*, 747–756. Springer, Cham.

[6] Rodgers, J. D., Tjaden, K., Feenaughty, L., Weinstock-Guttman, B., & Benedict, R. H. (2013). Influence of cognitive function on speech and articulation rate in multiple sclerosis. *Journal of the International Neuropsychological Society, 19*(2), 173.

[7] Tatman, R. (2017). *Speech Accent Archive: Parallel English speech samples from 177 countries*. https://www.kaggle.com/rtatman/speech-accent-archive.

[8] Mozilla. (2017). *Common Voice*. https://www.kaggle.com/mozillaorg/common-voice

Rianne Drijver (riannedrijver@gmail.com) studeert Sensor Technology – major elektrotechniek en is honours student bij de Hanzehogeschool. Sigrid van Hoek (st.vanhoek@cbs.nl), Jonas Klingwort (j.klingwort@cbs.nl) en Rob Willems (rma.willems@cbs.nl) methdologen bij het Centraal Bureau voor de Statistiek.

Het maken van officiële statistieken wordt gecompliceerd door ontbrekende gegevens. Eén van de manieren om hier mee om te gaan is door te imputeren. Dit betekent dat schattingen worden gemaakt van ontbrekende waarden. Verschillende imputatiemethoden zijn beschikbaar zoals: nearest neighbour, ratio- en regressie-imputatie. Ook binnen *machine learning* zijn er technieken voorhanden om ontbrekende gegevens te schatten. Deze worden echter (nog?) niet wijdverbreid toegepast in de officiële statistiek. Dit artikel beschrijft een proof of concept van een nieuwe imputatiemethode voor de huishoudensstatistiek. De methode is gebaseerd op Extreme Gradient Boosting, een techniek uit machine learning.

IMPUTEREN VAN HUISHOUDSAMENSTELLINGEN MET MACHINE LEARNING

Jacco Daalmans

Jaarlijks publiceert de huishoudensstatistiek over het aantal huishoudens, uitgesplitst naar samenstelling (éénpersoonshuishoudens, paren met en zonder kinderen, etc.). De statistiek wordt afgeleid uit de huishoudensbus: een longitudinaal bestand van alle huishoudens op alle adressen. De huishoudensbus wordt gevoed vanuit verschillende bronnen. Een belangrijke is de Basisregistratie Personen (BRP). Voor een deel van de adressen levert deze bron geen eenduidige resultaten op. Voor die adressen wordt geprobeerd om de huishoudens deterministisch af te leiden. Als bijvoorbeeld twee volwassenen tegelijk naar een adres verhuizen gaat het CBS ervan uit dat het gaat om één huishouden. Een relatief klein aantal adressen kan ook niet deterministisch worden afgeleid. Die overgebleven groep, ruim 900.000 huishoudens, wordt modelmatig geïmputeerd. Omdat dit voor alle onbekende adressen gebeurt spreekt men van massa-imputatie; zie het kader verderop voor de valkuilen van die techniek. Momenteel worden ontbrekende gegevens geschat met een regressiemethode. Hierin wordt de relatie gelegd tussen huishoudsamenstellingen en hulpvariabelen, zoals leeftijdsverschil, wel of niet hetzelfde geslacht van de bewoners en stedelijkheid van de gemeente. Goede resultaten van *machine learning* methoden voor vergelijkbare classificatieproblemen geven aanleiding om te onderzoeken of Extreme Gradient Boosting ook kan worden aangewend voor de imputatie van huishoudsamenstellingen. Als eerste is er een *proof of concept* gemaakt, waarin gefocust is op een vereenvoudigd probleem: de classificatie van adressen met twee volwassen bewoners in: 1. twee (eenpersoons)huishoudens en 2. één (tweepersoons) huishouden.

Gradient Boosting

Classificatie met Gradient Boosting betekent dat een onbekende doelvariabele wordt geschat op individueel niveau. De schattingen worden gebaseerd op achtergrondkenmerken, die idealiter sterk samenhangen met de doelvariabele. Eerst wordt een model geschat met data waarvoor doel- en hulpvariabelen bekend zijn, vervolgens wordt dat model toegepast op eenheden waarvoor alleen hulpvariabelen beschikbaar zijn. Voor classificatieproblemen worden de kansen geschat op iedere categorie. Hier wordt dus voor ieder adres kansen geschat op 'twee huishoudens' versus 'één huishouden'. Op basis van deze geschatte kansen wordt er vervolgens geïmputeerd.

Gradient Boosting is een zogenaamde ensemble-techniek. Dit betekent dat de methode verschillende schattingen voor één eenheid combineert. Binnen de klasse van ensemblemethoden bestaat een onderscheid tussen bagging en boosting. Bij bagging worden verschillende schattingen onafhankelijk gemaakt. De uiteindelijke schatting wordt verkregen door het combineren van de afzonderlijke schattingen; bijvoorbeeld door te middelen. Bij boosting zijn de opeenvolgende schattingen afhankelijk. Iedere schatting probeert het resultaat van de vorige schatting (verder) te verbeteren. Zoals de naam al aangeeft, behoort Gradient Boosting tot de boosting methoden.

Een formelere manier om de boosting methode te beschrijven is als volgt. Stel dat we geïnteresseerd zijn in een doelvariabele $\mathbf{y}$. In dit geval staat y_i voor de kans op twee huishoudens op adres *i*. De kans op één huishouden is gelijk aan één minus deze kans. Om $\mathbf{y}$ te schatten hebben we hulpinformatie tot onze beschikking, die wordt aangeduid met $\mathbf{X}$. In eerste instantie wordt de best mogelijke schatting voor $\mathbf{y}$ bepaald, zeg $f_1(\mathbf{X})$. Dit is de schatting die optimaal is volgens een wiskundig criterium. Deze eerste schatting noteren we met $\hat{\mathbf{y}}_1 = f_1(\mathbf{X})$. De schatting $f_1(\boldsymbol{X})$ kan op verschillende manieren worden gemaakt. Gradient Boosting gebruikt beslisbomen, maar het zou bijvoorbeeld ook met regressie kunnen. Waar veel andere methoden hier ophouden, voert Gradient Boosting vervolgstappen uit om de eerste schatting $\hat{\mathbf{y}}_1$ te verbeteren. Iedere stap is erop gericht om de schattingsfout na afloop van de voorgaande stap, het zogenaamde residu, zo goed mogelijk te voorspellen. De gedachte is dat als het lukt om de fouten te schatten dat men daarvoor ook kan corrigeren. In de tweede stap wordt dus geprobeerd om de residuen, $\mathbf{r}_1 = \mathbf{y} - \hat{\mathbf{y}}_1$ te schatten. De uitkomst van stap 2 zijn schattingen $f_2(\mathbf{X})$ voor $\mathbf{r}_1$. De schatting voor de doelvariabele $\mathbf{y}$ na stap 2, is de som van de schatting $f_1(\mathbf{X})$ na stap 1 en de geschatte correctie $f_2(\mathbf{X})$ We krijgen dus dat $\hat{\mathbf{y}}_2 = f_1(\mathbf{X}) + f_2(\mathbf{X})$. Vervolgens wordt in stap 3 een schatting $f_3(\mathbf{X})$ gemaakt van het residu na stap 2. Op deze manier volgt dat de schatting na stap **J** te schrijven is al een som $\sum_{j=1}^{J} f_j(\mathbf{X})$. In de praktijk wordt vaak een gewogen som toegepast, maar daar gaan we omwille van de eenvoud niet verder op in.

Zoals eerder opgemerkt, worden de schattingen $f_j(\mathbf{X})$ gemaakt met beslisbomen. Bomen bestaan uit knopen en uit vertakkingen (zie figuur 1). In binaire beslisbomen wordt de dataset in iedere knoop verdeeld in twee delen, die beide zo homogeen mogelijk zijn met betrekking tot de te schatten doelvariabele. Er wordt een stopcriterium gehanteerd dat bepaalt wanneer er wordt gestopt met het verder vertakken van de boom (oftewel: opsplitsen van de data). Wanneer men te ver doorgaat met vertakken loopt men het risico op overfitting. Dit betekent dat de boom goede schattingen geeft voor de data die zijn gebruikt om de boom af te leiden (de zogenaamde trainingset), maar minder goed werkt op een onafhankelijke dataset, waarop men het model wil toepassen (de testset). Om overfitting te voorkomen wordt het aantal takken van de boom beperkt. Het algoritme probeert dus om zo goed mogelijke schattingen te maken met zo eenvoudig mogelijke beslisbomen. Bij veel toepassingen van Gradient Boosting wordt een groot aantal, soms wel honderden, relatief eenvoudige beslisbomen gecombineerd. Een

Zoals eerder opgemerkt is het beschreven advies een toepassing van massa-imputatie: het grootschalig schatten van ontbrekende gegevens op microniveau. De bedoeling is om die schattingen te gebruiken voor verschillende doeleinden. Deze werkwijze is niet onomstreden.

Idealiter houdt men bij het schatten van ontbrekende gegevens rekening met het doel waarvoor men de imputaties wil gebruiken. Stel dat iemand geïnteresseerd is in het verband tussen het aantal adressen met twee eenpersoonshuishoudens en de stedelijkheid van de gemeente. Als men dan bij het imputeren geen rekening houdt met deze relatie dan kan het zo zijn dat de geïmputeerde data een onjuist beeld geven over de samenhang tussen beide variabelen. Een berucht voorbeeld is het zogenaamde hondenbrokkenprobleem. Stel dat we een data set hebben met daarin de uitgaven aan hondenvoer. Alle ontbrekende waarden worden geïmputeerd. Het gegeven of iemand wel of geen hond heeft wordt echter niet meegenomen bij het maken van de schattingen. Deze informatie is niet beschikbaar of wordt niet relevant beschouwd. Een onderzoeker koppelt vervolgens de geïmputeerde data aan een tweede data set, waarin het hebben van een hond (wel) is opgenomen. Omdat bij het imputeren van uitgaven aan hondenbrokken geen rekening is gehouden met het feit of iemand een hond heeft, kunnen er vele hondenbezitters worden gevonden die geen geld aan hondenbrokken uitgeven, omgekeerd kunnen er veel niet-hondenbezitters zijn die toch geregeld hondenbrokken kopen. Dit kan dus leiden tot een verkeerde conclusie over de relatie tussen twee variabelen.

Een beter alternatief voor een generieke vorm van massa-imputatie is dat alle gebruikers zelf imputaties afleiden, die specifiek bedoeld zijn voor het doel waarvoor zij de data nodig hebben. In de praktijk kan dit echter lastig zijn omdat de gebruikers niet altijd alle data hebben. Bovendien betekent het veel werk voor de gebruiker en kan het inconsistente uitkomsten opleveren als verschillende gebruikers verschillende imputatiemethoden toepassen. Vanwege de bovenstaande argumenten en vanwege het relatief lage aantal imputaties voor huishoudens, is ervoor gekozen om, ondanks de bezwaren toch massa imputatie toe te passen. Bovenstaande betekent echter niet dat massa imputatie in het algemeen is aan te bevelen voor iedere statistiek.

Figuur 1. Een fictief voorbeeld van een beslisboom

eindknoop van een boom, een zogenaamd blad, hoort bij een specifieke, homogene groep. Bijvoorbeeld alle duo's op één adres van hetzelfde geslacht met minder dan 15 jaar leeftijdsverschil. Voor ieder blad wordt er ofwel een waarde van de doelvariabele geschat (hier: de kans op twee huishoudens op één adres), of een correctie ten opzichte van een eerdere schatting.

Als resultaat van Gradient Boosting krijgen we voor ieder adres een geschatte kans voor één versus twee huishoudens. Deze kansen kunnen we gebruiken voor het afleiden van imputaties. Dit is gedaan door het trekken van random getallen tussen nul en één. Stel dat we voor een bepaald adres afleiden dat er 60% kans is op één huishouden en 40% op twee huishoudens. We trekken dan vervolgens uit een uniforme verdeling op het interval [0,1]. Als de uitkomst kleiner of gelijk is aan 0,6 dan imputeren we 'één huishouden', is de uitkomst groter dan imputeren we 'twee huishoudens'. Deze stochastische methode om imputaties af te leiden wijkt af van de gangbare methode bij machine learning die kansen afrondt. Als er een kans van 0,6 is geschat op één huishouden, dan wordt die afgerond naar 1 en wordt er dus 'één huishouden' geïmputeerd. Hoewel deze benadering op individueel niveau de kleinste schattingsfout oplevert, heeft deze als effect dat de verdeling van de doelvariabele sterk af kan wijken van die van de geobserveerde waarden. Stel bijvoorbeeld dat voor alle adressen 60% kans wordt geschat op 'één huishouden'. Afronden zou dan betekenen dat voor alle adressen 'één huishouden' wordt geïmputeerd. De categorie 'één huishouden' wordt dan dus geobserveerd in 60% van de adressen, maar komt voor in 100% van de imputaties. Voor veel toepassingen bij statistische bureaus is dit zeer ongewenst, aangezien een juiste verdeling op geaggregeerd niveau belangrijker is dan een nauwkeurige schatting op individueel niveau.

Resultaten

Gradient Boosting blijkt betere schattingen te geven dan de huidige regressiemethode. Door toepassing van de methode op woningen met een bekende huishoudsamenstelling (zogenaamde cross-validatie) is een inschatting te maken van de precisie van de imputaties. Gradient Boosting classificeert 77% van de adressen correct, terwijl dit percentage voor de huidige regressiemethode op 74% ligt. Daarnaast is er ook gekeken naar de zogenaamde AUC (Area Under the Receiver Operating Characteristics (ROC) Curve). Eén van de interpretaties hiervan is hoe waarschijnlijk het is dat een grotere kans voor klasse 1 wordt voorspeld voor iemand uit klasse 1 vergeleken met iemand uit klasse 0. De score ligt tussen 0,5 en 1. Een score van 0,5 betekent dat een model willekeurig gokt en 1 staat voor een perfecte discriminatie tussen de twee groepen. De AUC/ROC voor het regressiemodel en Gradient Boosting bedragen respectievelijk 0,87 en 0,90.

Gradient Boosting versus regressie

Zoals hierboven beschreven geeft Gradient Boosting nauwkeurigere schattingen dan de huidige regressiemethode. Gradient Boosting heeft echter ook andere voordelen. Zo is deze methode makkelijker toepasbaar. Bij regressie moet van te voren worden bepaald wat de relatie is tussen de doelvariabele en de verklarende variabelen. Dit kan bijvoorbeeld een lineair verband zijn, maar ook een exponentieel verband. Bij Gradient Boosting is het niet nodig om dit van te voren te vast te leggen; dit wordt door de methode bepaald. Ook kan Gradient Boosting eenvoudiger dan regressie overweg met ontbrekende waarden in verklarende variabelen. Een nadeel van Gradient Boosting is dat de uitkomsten lastiger zijn te duiden. Het is niet erg eenvoudig om achteraf te achterhalen hoe specifieke schattingen tot stand zijn gekomen. Bij regressie is dit makkelijker. Vanwege de bovenstaande voordelen is geadviseerd om Gradient Boosting te implementeren. Momenteel wordt de methode verder uitgewerkt en worden er voorbereidingen getroffen om de methode in het productieproces op te nemen.

Jacco Daalmans heeft econometrie gestudeerd aan Tilburg University. Hij werkt als methodoloog voor het Centraal Bureau voor de Statistiek en is in 2019 gepromoveerd op toepassingen van macro-integratie in de officiële statistiek.
E-mail: j.daalmans@cbs.nl

Gerard Sierksma

column

Over P≠NP en een Eeuwige Student

In het decembernummer van het tijdschrift *NewScientist* staat een mooi verhaal met de titel 'P=NP?'. Een vraag als titel dus. Die vraag betreft een van de zeven beroemde millenniumproblemen met een miljoen dollar voor elke eerste oplosser. Het antwoord laat al ruim 50 jaar op zich wachten en Tamara Florijn, de auteur, twijfelt of het er ooit van komt. Ze eindigt nogal pessimistisch met zinsneden als: '(...) dat het nog wel honderd jaar kan duren voordat (...)' en 'Misschien zullen we wel nooit weten of (...)'. Je moet kennelijk wel een beetje gek zijn om er tijd en energie in te steken. Ik heb zo'n 'gek' gekend, een eeuwige student.

Iedereen heeft wel een beeld van een eeuwige student, maar wat P≠NP of P=NP betekent is minder bekend, zeker ook doordat de uitleg ervan een behoorlijke dosis wiskundige voorkennis vereist. Dat de P en NP niets van doen hebben met parkeren of zo lijkt me duidelijk, maar met wat dan wel? Om te beginnen, de P staat voor 'polynomiaal'. Je zou denken dat NP dan voor 'niet-polynomiaal' staat, maar dat is dan weer niet het geval. Schiet niet op dus, zelfs als ik toevoeg dat NP staat voor 'niet-deterministisch polynomiaal'. Dan maar kort-door-de-bocht uitgelegd, wat Tamara ook doet.

De letter P staat voor de klasse van wiskundige problemen die (met een algoritme) zijn op te lossen in polynomiale tijd, wat kort-door-de-bocht wil zeggen 'snel op te lossen'. En wat is 'snel' dan wel? Ik kom daar zo op terug. Eerst even naar NP. Dat is de klasse van problemen waarvan 'snel kan worden gecheckt' of een gevonden 'oplossing' echt wel oplossing is. Oké, nu 'snel'. Interessant in deze context is dat er een helder onderscheid is tussen 'snel' en 'niet snel'. Een probleem heet 'snel', sommigen zeggen zelfs 'makkelijk', oplosbaar als voor dat probleem een algoritme bestaat waarmee, voor alle mogelijke input data, een oplossing wordt geproduceerd binnen een beperkte tijdspanne, ook wel *real-time* genoemd. Een voorbeeldje. Als om vier uur 's ochtends de vrachtwagens moeten beginnen met het rijden van de routes en alle bezoekadressen zijn bekend om, zeg, drie uur in de morgen, dan heeft het algoritme maximaal een

uur ter beschikking om een optimale planning te berekenen. En dat dag-in-dag-uit, met elke dag andere input data en elke dag dat ene uurtje. Als dat lukt heb je in je computer een snel algoritme ter beschikking en heet het (routerings)probleem snel oplosbaar. Daarentegen een probleem dat 'niet snel' oplosbaar is, kent deze luxe niet en moet het doen met algoritmen die in het slechtste geval uren-dagen-maanden, ja zelfs eeuwen moeten rekenen om een gewenste (vaak een optimale) oplossing te berekenen. Het 'handelsreizigersprobleem' is hiervan het prototypevoorbeeld. In de praktijk heb je aan dergelijke algoritmen dus helemaal niets en wordt gewerkt met algoritmen die slechts suboptimaal de wensen (van de planner) vervullen. Hierbij moet worden aangetekend dat het aantal kandidaatoplossingen exponentieel toeneemt met de omvang van de input data: één locatie erbij betekent onmiddellijk een verdubbeling van het aantal routes (de zoekruimte) waaruit de beste berekend moet worden. Exponentiële groei van de zoekruimte derhalve.

De vraag 'P=NP?' betekent dan: Als je snel kunt controleren of een oplossing van een probleem klopt, kun je dan het probleem zelf ook snel oplossen? Voor alle problemen, waarvoor dus een snel checking-algoritme bestaat, zou dan ook het probleem zelf snel zijn op te lossen, hoe bizar de input data er ook uit ziet. En geloof het of niet, maar als op de vraag P=NP? het antwoord 'ja' is, dan kunnen computers vervolgens veel werk van wiskundigen overnemen en zelfs pincodes van banken kraken. Geen wonder dat we hier te maken hebben met een 1-miljoen dollarprobleem. Nog korter door de bocht met een metafoor: Controleren of een aangeboden 'voorwerp' inderdaad de gezochte speld in de gigagrote hooiberg is moge simpel zijn, maar dat daarmee ook het vinden van die speld zelf simpel is ligt gelukkig niet echt voor de hand. Geen paniek over krakende codes dus, voorlopig.

De nu 82-jarige informaticus Stephen Cook stelde ruim 50 jaar geleden als eerste de vraag 'Is P gelijk aan NP?' Tegenwoordig spreken we van het 'PvsNP-probleem' – met 'versus' dus tussen P en NP –, omdat het gilde der complexiteitswiskundigen inmiddels is gepolariseerd: De paniekzaaiers (pincodes worden gekraakt) die geloven in P=NP, met daartegenover de 'ongelovigen', zij die hun ziel en zaligheid hebben verkocht aan P≠NP. De gelovigen vormen een kwijnende minderheid en in lijn met Tamara moeten we ook vrezen voor het uitsterven van het P≠NP-ras. Ik reken mezelf tot de ongelovigen: ooit wordt bewezen dat P niet gelijk is aan NP en worden geen pincodes gekraakt. En misschien is het nog beter om Tamara te zien als de voorloper van een derde denominatie, waar ik ook wel iets in zie: de PvsNP-agnosten, die beweren dat de vraag zelf eigenlijk helemaal niet heeft bestaan en ook niet zal bestaan. Zeker is dat de kwestie niet bestond voor Cooks eerste formulering ervan. En ooit zou de vraag compleet kunnen verdwijnen zonder antwoord. Misschien, heel misschien gebeurt dat zodra de kwantumcomputer zijn intrede heeft gedaan en NP-problemen, zoals het handelsreizigersprobleem, wel snel maar niet polynomiaal worden opgelost. Het zou dus weleens heel lang kunnen gaan duren. Tja, en dan is de 1-miljoen dollar niks anders geweest dan een in de eeuwigheid verdwenen bijna-grijpbare wortel voor de neus van de ezel.

Koffer is bestemd voor de wortel: de PvsNP 1-miljoen dollar

Gert Tijssen zal midden vijftig zijn geweest. Het was in de zomer van 2005. De televisie stond nog aan toen zijn zuster hem vond. Op de glazen tafel naast hem lag de as van z'n laatste sigaret. De fles leeg, het glas vol. Gert is vanuit Groningen bijgezet in het familiegraf in zijn geboorteplaats Apeldoorn. Een zoektocht op het internet levert mij niks op: geen sterfdag, geen bio, niks. *Zapped into eternity*? Een dikke tien jaar heb ik zijn obsessieve P≠NP?-gezwoeg meegemaakt. Hij was zeker niet gek, wel markant en uiterst scherpzinnig. *Never a dull moment* in de tien jaren met Gert. Maar wat heeft het opgeleverd?

In 1992 studeerde Tijssen cum laude af als de laatste 'eeuwige student'. Daarna heb ik het gewaagd om hem een promotieplaats aan te bieden als vervolg op zijn succesvolle afstudeerscriptie over *cutting stock* problemen.

Gert Tijssen (links) bij zijn afstuderen; rechts aan tafel van boven naar beneden: Caspar Schweigman, Ton Steerneman, Jaap Ponstein, Ton Wansbeek en Gerard Sierksma.)

Hij heeft die aanbieding graag geaccepteerd. In de perioden waarin zijn manische toppen overgingen in depressieve dalen was Gert een uiterst aimabele man die echter meer en meer begon te vergen van luisterende oren. Na een paar maanden als PhD-student veranderde plots zijn gedrag: geen lange monologen meer bij de koffieautomaat en er kwam een soort grauwe stilte om hem heen te hangen. Ineens was hij verdwenen met, in een laatste e-mail, de mededeling dat hij voor een half jaar naar Griekenland was vertrokken, naar zijn zuster in Athene.

Zo'n zes maanden later, op een goede maandagochtend, trof ik hem aan bij de koffieautomaat, monter en zijn oude maatpak wat strakker om de inhoud. Het oude vertrouwde tussen hem en mij was niet verdwenen en zonder omhaal meldde hij te zijn gestopt met snijproblemen. En met een grijns: 'die zitten allemaal in P en zijn dus opgelost.' Maar wat dan wel? Weer die grijns: 'Het wordt vanaf nu de simplexmethode met een nieuwe pivottruc die LP in P brengt' en voegde eraan toe 'LP gaat straks ook met simplex in P.' Ik ben akkoord gegaan, onder de voorwaarde dat er binnen een half jaar een toptijdschriftrijp *paper* van hem zou zijn. *Believe it or not*, in juli 1995 lag zo'n artikel op mijn bureau. Het duurde daarna minstens drie jaar tot de publicatie in *Mathematical Programming*. Die lange aanlooptijd kwam doordat we in eerste instantie het *paper* hebben aangeboden aan het toptijdschrift *Mathematics of Operations Research* en de toenmalige *editor-in-chief* het terugstuurde met de mededeling '*this is folklore*'. Kort gezegd, ging het onder meer over de stelling dat voor LP-modellen met eindige oplossingen geldt: *The dimension of the optimal primal face is equal to the degeneracy degree of the corresponding optimal dual face.* Door Tijssen fraai bewezen met Balinski-Tucker simplex tableaus. Dat klinkt toch niet als folklore, zou ik zeggen. Maar dit terzijde.

Tijssen is gepromoveerd op 24 januari van het jaar 2000. In datzelfde jaar was het dat het PvsNP-probleem verheven werd tot een van de zeven millenniumproblemen met voor elk probleem de hoofdprijs van 1-miljoen dollar, uit te keren door het *Clay Mathematics Institute* aan de allereerste oplosser. Slechts een van de zeven problemen is inmiddels opgelost; 6 miljoen ligt dus nog op de plank. Tien jaren heeft Gert de 1-miljoenwortel voor z'n neus zien bungelen.

In de nadagen van zijn leven veranderde Tijssen in de rijzige grijzende man, die veel ouderen onder ons zich zullen herinneren. Gerts heilig geloof in het vinden van een bewijs voor P≠NP leek zijn regelmatig terugkerende depressies te verhevigen. Zeker driemaal heeft hij me bij de koffieautomaat 'zijn bewijs' van P≠NP verteld. Euforisch. De driemaal die ik me herinner had hij de nacht ervoor niet geslapen, maar wel te diep in het glas met P≠NP gekeken. Een vierde 'bewijs' is er niet meer van gekomen: de fles was leeg, het glas nog vol. De 1-miljoen dollar zit nog in het vat en is voor eeuwig waardevast geïndexeerd. Voor eeuwig?

P.S. Een meer dan opmerkzame lezer van een eerdere versie van deze column vroeg zich af waarom ik niet voor de titel 'In Memoriam Gerhard Antonie Tijssen' heb gekozen. Die uitleg over P en NP kan iedereen toch vinden op Wikipedia? Ze heeft hier een punt. Voor een in memoriam vind ik het in het geval van Gert Tijssen te laat. Tja en dat lange verhaal waarin ik kort-door-de-bocht... En of ik NP-agnost wordt of zoiets? Ik hou me daar eerlijk gezegd niet mee bezig. Wel raad ik elke jonge wetenschapper aan, die een van de zes overgebleven 1-miljoenen wil verdienen, eerst een ontdekking te doen die minimaal in *Math of OR* verschijnt.

Literatuur

Tijssen, G. A., & Sierksma, G. (1998), Balinski-Tucker Simplex Tableaus: Dimensions, degeneracy degrees, and interior points of optimal faces. *Mathematical Programming, 81*, 349–372.

Tijssen, G. A. (2000), *Theoretical and practical aspects of linear optimization*. PhD Thesis, SOM Research School, University of Groningen.

Tijssen, G. A., & Sierksma, G. (2006), Simplex adjacency graphs in linear optimization. *Algorithmic Operations Research, 1,* 46–51.

Gerard Sierksma is emeritus hoogleraar Operations Research aan de Rijksuniversiteit Groningen.
E-mail: g.sierksma@rug.nl

Marjolein Bolten wordt gemonitord tijdens haar looptraining. Foto: Rikkert Harink

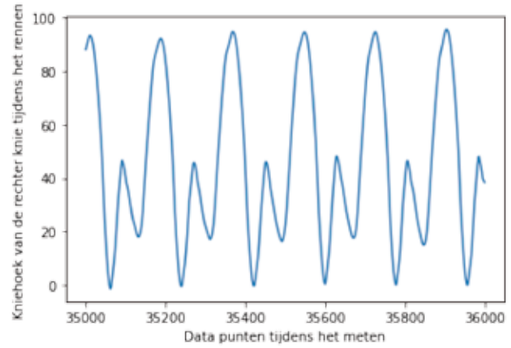
Welke test ‘loopt’ het best?

Marjolein Bolten

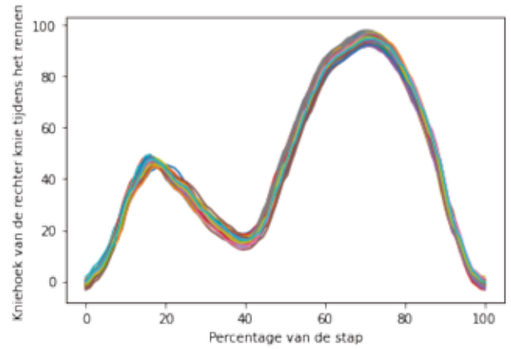
Hardlopen is een populaire en laagdrempelige sport. Een paar hardloopschoenen is al voldoende om een rondje te lopen. Het kan op jouw moment of in een gezellige groep. Bovenal, het is aantoonbaar gezond en geestverruimend. Echter, de voordelen kennen ook een nadeel, blessures. Eenderde van de blessures ontstaan tijdens het hardlopen, betreft een knieblessure. Deze ontstaan vaak door overtraining en/of te weinig rust na de training, oftewel het gaan trainen terwijl de spieren nog vermoeid zijn (Mechelen, 1992).

In het kader van deze blessures is het zinvol om looptrainingen te monitoren en bijvoorbeeld te kijken naar de hartslag of spierbewegingen. Het statistisch analyseren van de data kan helpen om de blessures te voorkomen. Elke stap is anders en daardoor zullen de bewegingspatronen van meerdere stappen niet perfect over elkaar heen lopen. Deze variatie kan gemodelleerd worden als statistische ruis, zodat de data kunnen worden beschreven als de echte gemiddelde stap plus deze statistische ruis.

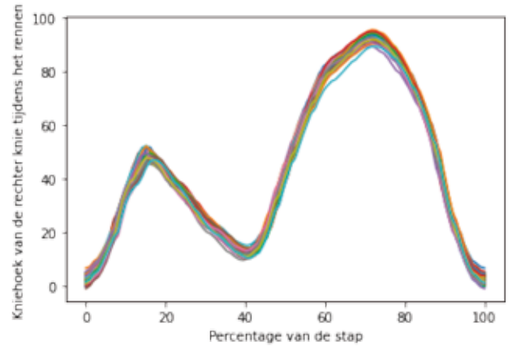
Het analyseren van spierbewegingen is een stuk moeilijker dan het analyseren van een hartslag. Dit komt doordat spierbewegingen in feite kunnen worden gezien als een tijdreeks, waarbij meerdere functies achter elkaar de data beschrijven. Waar bij de hartslag wordt gekeken naar een reeks van getallen voor het analyseren, wordt bij spierbewegingen gekeken naar een reeks van trajecten, waarbij een traject bestaat uit meerdere getallen die bij elkaar horen. Voor zulke tijdreeksen zijn er minder analyse-



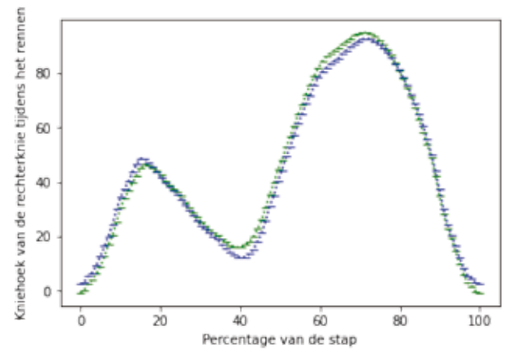
Figuur 1. Ruwe data



Figuur 2. Niet vermoeid



Figuur 3. Vermoeid



Figuur 4. Gemiddelde trajecten

mogelijkheden bekend dan voor datasets met reeksen van getallen.

Figuur 1 toont de bewegingen van de knie in het sagittale vlak, het anatomische vlak dat het lichaam verdeeld in een links en rechts vlak, tijdens het hardlopen. Deze bewegingen kunnen worden gezien als een tijdreeks. In dit artikel worden drie verschillende statistische methodes vergeleken en uitgevoerd op de data om te kijken welke methode het meest geschikt is voor het analyseren van de bewegingen van de kniehoeken tijdens het hardlopen.

Data

De data zijn verzameld tijdens een vermoeidheidsrun op een loopband waarbij de hardloper zo lang mogelijk moest hardlopen op 103% van zijn 8 km tempo. Door op dit tempo te lopen was vermoeidheid gegarandeerd aan het eind van het protocol. De eerste 60 stappen worden gebruikt als data groep 1, de niet-vermoeide groep en de laatste 60 stappen worden gebruikt als data groep 2, de vermoeide groep. De data zijn weergegeven in figuren 2 en 3, met groep 1 in figuur 2 en groep 2 in figuur 3, waarbij elke 60 stappen over elkaar heen geplot zijn om het verschil te verduidelijken.

Zoals in figuur 2 en 3 te zien is, zijn er verschillen tussen de groepen, maar de vraag is of dit verschil te verklaren is door de aanwezigheid van ruis of dat er een significant verschil is. Om deze vraag te beantwoorden zijn er drie verschillende statistische methodes geanalyseerd door te testen op verschil in de gemiddelde trajecten van de twee groepen. Hiervoor zijn dus eerst de gemiddelde trajecten berekend en deze zijn te zien in figuur 4.

In de statistische methodes wordt gebruik gemaakt van de volgende nul- en alternatieve hypothese:

H_0: De twee gemiddelde trajecten zijn identiek

H_1: De twee gemiddelde trajecten zijn niet identiek

De statistische methodes zullen met een 95% betrouwbaarheidslevel de nulhypothese accepteren of verwerpen. Door het verwerpen van de nulhypothese kan geconcludeerd worden dat de twee trajecten niet identiek zijn.

Zoals in figuur 4 te zien is, zijn deze twee gemiddelde trajecten niet identiek aan elkaar, maar welke methodes zijn in staat dit verschil ook te detecteren en zullen dus de nulhypothese verwerpen?

Methoden

De drie statistische methoden die vergeleken worden in dit artikel zijn, door middel van A. betrouwbaarheidsintervallen, B. Tweezijdige t-test en C. Bootstrap test. De eerste twee methoden vallen onder de categorie gecombineerde testen. Dat betekent dat er meerdere testen tegelijkertijd naast elkaar uitgevoerd worden met uiteindelijk één conclusie: de nulhypothese verwerpen of accepteren. Dit is een simpele combinatie van univariate methoden, die niet de specifieke structuur van het hele traject meenemen. De laatste methode is geen gecombineerde test en test wel direct op het hele traject.

Voor de eerste twee testen is elke stap in 100 individuele punten opgedeeld, waarbij op elk van die losse punten de betreffende methode is uitgevoerd. Zo zijn er 100 betrouwbaarheidsintervallen opgesteld die, na een multipliciteit correctie (zie volgende paragraaf), samen de 95%-betrouwbaarheidsband geven voor de twee groepen. Elke betrouwbaarheidsinterval is zo opgesteld dat met een betrouwbaarheid van 95% gezegd kan worden dat elke waarde van een willekeurige stap uit die groep op dat punt in het interval zal liggen.

De tweede methode die geanalyseerd is is de Student's t-test, dit is een veel voorkomende test in de statistiek en is daarom ook meegenomen in dit artikel. Voor de tweezijdige t-test is op diezelfde 100 punten de student's t-test uitgevoerd en na een multipliciteit correctie kan de conclusie getrokken worden.

Beide testen zullen de nulhypothese – de trajecten zijn identiek – verwerpen als er op tenminste een van de 100 losse testen die naast elkaar uitgevoerd worden een verschil te detecteren is. Als de trajecten op een punt verschillen kan daaruit direct afgeleid worden dat de volledige trajecten niet identiek zijn.

Multipliciteit correctie

Bij elke statistische test die uitgevoerd wordt is er een kans dat de nulhypothese onterecht verworpen wordt: α. Deze kans wordt gecontroleerd door het betrouwbaarheids level, (1 – α)100%. In alle methodes is gerekend met een alpha van 0,05 en dus een betrouwbaarheids level van 95%. Dit betekent dat als een test 100 keer uitgevoerd wordt, in maximaal 5% van de gevallen de test ten onrechte de nulhypothese verwerpt.

Voor de testen waarbij gelijktijdig 100 testen worden uitgevoerd moeten we verwachten dat er in totaal 5 valse verwerpingen zijn. Hiervoor worden de individuele alpha's aangepast met behulp van een multipliciteit correctie (Dudoit, Shaffer & Boldrick, 2003), in dit geval Bonferroni correctie. De gecorrigeerde α voor de individuele punten is dan $\alpha_i = \alpha/s$, met s het aantal individuele punten, 100, waarop de test wordt uitgevoerd. Hierdoor blijft het betrouwbaarheids level van de gecombineerde test op 95%.

Bij de laatste test wordt er direct op het hele traject getest en is er dus ook geen multipliciteit correctie nodig. Met behulp van bootstrapping wordt de kritieke waarde, de drempelwaarde voor statistische significantie, bepaald door middel van pseudo observaties. Deze worden gecreëerd door een random residufunctie, op te tellen bij het traject. De residufunctie is het verschil tussen een van de 60 trajecten en het gemiddelde traject. Dit wordt gedaan voor alle 60 trajecten en zo wordt er een nieuwe pseudo observatie gecreëerd. Bij bootstrapping wordt dit proces heel vaak herhaald, in dit geval 1000 keer, en zo wordt er een pseudo dataset gecreëerd.

Voor elke pseudo observatie is de pseudo teststatistiek S_N^+ berekend volgens

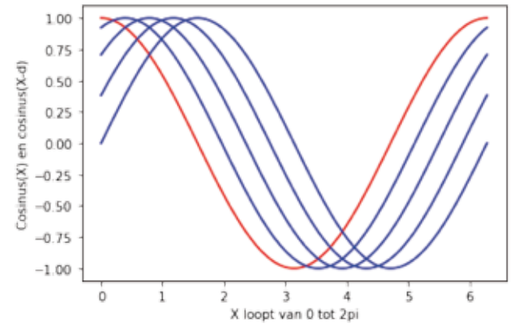
$$S_N^+ = \frac{n_1 n_2}{N} \left| \overline{X}_{1,n_1} - \overline{X}_{2,n_2} \right|^2$$

en de kritieke waarde, C_α van deze test is geschat door de 1 – α percentiel te nemen van de gesorteerde lijst met deze pseudo observaties (Paparoditis & Sapatinas, 2016).

Om uiteindelijk een conclusie te kunnen trekken wordt de teststatistiek van de originele dataset vergelijken met de kritieke waarde, en zal de nulhypothese verworpen worden als de teststatistiek groter is dan de kritieke waarde. In dit geval is het verschil tussen de groepen zo groot, dat dit niet door willekeurige ruis zal zijn.

Resultaten

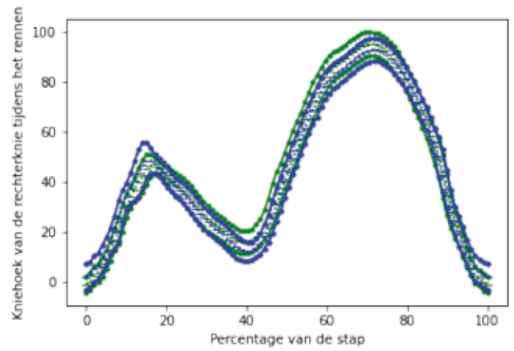
Elke methode is eerst getest op kunstmatige data zodat de validiteit van de testen kon worden geschat. Om zulke kunstmatige data te verkrijgen wordt een echt stap patroon en algemene ruis gekozen. Hiervoor is een cosinus en random normaal verdeelde ruis gebruikt. De methoden waren gevalideerd zodra de testen op kunstmatige data gecontroleerd waren door alpha en het betrouwbaarheidsniveau. Voor elke methode is de power geschat, dit is de kans dat de test correct de nulhypotheses verwerpt. De power is ingeschat door te testen op de rode groep met een blauwe groep uit figuur 5 waarbij de blauwe groepen lopen van cosinus tot sinus, waarbij het verschil tussen de groepen dus steeds duidelijker wordt. Zoals in tabel 1 gezien kan worden is de test met de bootstrap het best om kleine verschillen te detecteren



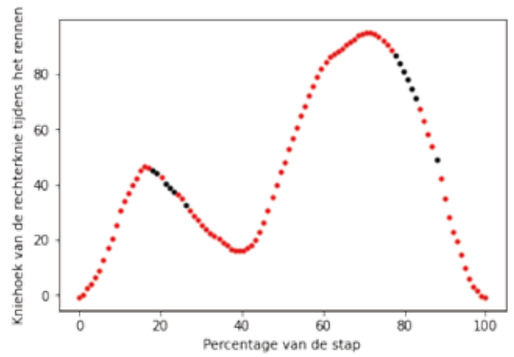
Figuur 5. Power

δ	% verwerpingen A betrouwbaarheidsintervallen	% verwerpingen B student's t-test	% verwerpingen C bootstrap test
0	0	4	4
$\frac{\pi}{200}$	0	7	8
$\frac{2\pi}{200}$	0	18	46
$\frac{3\pi}{200}$	0	34	96
$\frac{4\pi}{200}$	0	62	100
$\frac{5\pi}{200}$	0	86	100
$\frac{6\pi}{200}$	0	100	100
⋮	⋮	⋮	⋮
$\frac{3\pi}{10}$	0	100	100
$\frac{4\pi}{10}$	11	100	100
$\frac{\pi}{2}$	96	100	100

Tabel 1



Figuur 6. Convidence bands



Figuur 7. T-test

en heeft daardoor dus de hoogste power.

Tot slot zijn de methoden uitgevoerd met de twee groepen data van het vermoeidheidsprotocol. De resultaten hiervan zijn te zien in de figuren 6 en 7. In figuur 6 is methode A door middel van betrouwbaarheidsintervallen te zien. Volgens deze methode zullen twee gemiddelde trajecten verschillend zijn omdat de betrouwbaarheids banden overlap hebben op alle punten. Figuur 7 toont methode B. De tweezijdige t-test verwerpt de nulhypothese wel; er zijn in figuur 7 meerdere punten te vinden waarop de trajecten niet identiek zijn. Voor methode C. Bootstrap test S_N = 24843,21 is groter dan C_α = 598,87. Dus de nulhypothese wordt verworpen. Dit betekent dat, volgens deze testen, het erg onwaarschijnlijk is dat de verschillen komen door willekeurige ruis, en dus komen door het verschil, de vermoeidheid, in de twee groepen.

Conclusie

Over het algemeen blijkt een test door middel van betrouwbaarheidsintervallen niet geschikt om de bewegingen van de kniehoeken te analyseren, deze is namelijk niet in staat om zulke kleine verschillen te detecteren. Daarentegen zijn de tweezijdige t-test en de bootstrap test dit wel. De power van de bootstrap test bleek het grootst. Daarom is de bootstrap test het best om te gebruiken voor verder onderzoek richting blessurepreventie tijdens het hardlopen.

Literatuur

Dudoit, S., Shaffer, J. P., & Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science, 18*(1), 71–103. http://www.jstor.org/stable/3182872

Mechelen, W. van. (1992). Running injuries: A review of the epidemiological literature. *Sports Medicine, 14*(5), 320–335.

Paparoditis, E., & Sapatinas, T. (2016). Bootstrap-based K-sample testing for functional data. *arXiv* 1409.4317.

Bolten, M. M. (2021). *Detection of changes in movement patterns of runners*. http://essay.utwente.nl/86782/

Dankwoord van Marjolein Bolten

Dit artikel is een samenvatting van mijn bacheloropdracht voor de bachelor Applied Mathematics bij het project Sports, Data & Interaction, opgericht door de faculteit EEMCS van de Universiteit van Twente (een samenwerking van Applied Mathematics, Technical Computer Science and Electrical Engineering). Ik werkte samen met dr. Katharina Proksch en Rupsa Basu MSc, die ik ook wil bedanken voor hun hulp tijdens het schrijven van dit artikel. Ook wil ik prof. dr. Nico van Dijk bedanken voor het proeflezen en de suggesties bij het schrijven van dit artikel.

E-mail: m.m.bolten@student.utwente.nl

Beter prevalenties meten door het combineren van schattingsmethoden

Machine learning algoritmes worden veel ingezet voor classificatiedoeleinden. Met een hoge nauwkeurigheid kunnen algoritmes veel objecten juist classificeren. Maar wat als we niet geïnteresseerd zijn in de classificatie van ieder object, maar juist in de groepsgrootte van iedere klasse? Classificeren en optellen lijkt een logische aanpak, zeker als het algoritme een hoge nauwkeurigheid heeft. In dit artikel vinden we uit waarom dit geen goede manier is om te kwantificeren en, nog veel belangrijker, hoe we dat op kunnen lossen.

Kevin Kloos

In de officiële statistiek zijn veel statistieken gewijd aan het schatten van prevalenties. Deze lopen uiteen van het percentage 65-plussers met een internetverbinding tot het percentage werknemers dat lid is van een vakbond. De laatste jaren zijn *machine learning* algoritmes enorm in opkomst, ook binnen de officiële statistiek. Nieuwe toepassingen liggen bijvoorbeeld in het bepalen of een woning zonnepanelen heeft aan de hand van satellietbeelden, of bepalen of een retailer een webwinkel heeft aan de hand van diens webpagina. Deze algoritmes zijn echter imperfect en kunnen dus classificatiefouten maken. Als er in de ene groep meer fouten worden gemaakt dan in de andere groep ontstaat er *misclassification bias*, oftewel vertekening door misclassificatie. Deze vertekening ontstaat niet alleen bij slimme algoritmes. In de *STAtOR*-editie van juni 2021 schreef Anton Meijburg (Meijburg, 2021) welke invloed classificatiefouten van PCR-testen hebben op het schatten van de prevalentie positieve uitslagen. Om de prevalentie juist te schatten is het vanzelfsprekend om de vertekende schatting te corrigeren.

Paradox

De vertekening van de geschatte prevalentie hangt af van drie parameters: 1. de daadwerkelijke prevalentie, 2. het aantal juist geclassificeerde positieven (sensitiviteit) en 3. het aantal juist geclassificeerde negatieven (specificiteit). Het lijkt logisch om te denken dat een verbetering in de sensitiviteit en/of specificiteit leidt tot een betere schatting van de prevalentie. Aan de hand van een simpel voorbeeld is het gemakkelijk om te zien dat dat niet altijd hoeft te zijn. Algoritme A schat 20 personen fout in: 10 fout positieven en 10 fout negatieven. Algoritme B schat 15 personen fout in: 9 fout positieven en 6 fout negatieven. Algoritme B kan beter classificeren, omdat zowel de sensitiviteit als de specificiteit van algoritme B hoger ligt dan die van algoritme A. Algoritme A kan echter beter de prevalentie schatten, omdat de fout positieven en de fout negatieven tegen elkaar weggestreept worden, terwijl dat in algoritme B niet het geval is. In algoritme B zijn, in tegenstelling tot algoritme A, de fouten niet gelijk over de groepen verdeeld en ontstaat er dus een vertekening van de geschatte prevalentie.

Steekproef

Om de prevalentie beter te kunnen schatten, is er meer informatie nodig over de specificiteit en sensitiviteit. Deze twee waarden kunnen worden bepaald met een aselecte steekproef. We nemen een willekeurige groep bestaande uit *n* observaties waar naast de geschatte classificatie ook de daadwerkelijke classificatie bekend is. In het voorbeeld van de webpagina's en de zonnepanelen kan er bijvoorbeeld handmatig gecontroleerd worden wat de daadwerkelijke classificatie is. Deze informatie kan worden samengevat in een kruistabel (tabel 1). De sensitiviteit en de specificiteit kunnen beiden uit deze kruistabel worden geschat, alsmede een schatting van de daadwerkelijke prevalentie.

Corrigeren

Aan de hand van deze steekproef (tabel 2) kan de vertekende schatting van de prevalentie ((TP + FP) / n = (900 + 300) / 2000 = 0,6) gecorrigeerd worden. Er zijn drie bekende en makkelijk toepasbare manieren om de prevalentie te corrigeren. De makkelijkste manier om de prevalentie te corrigeren is simpelweg het aantal daadwerkelijke positieven tellen en te delen door de omvang van de steekproef ((TP + FN) / n = (900 + 100) / 2000 = 0,5). Een andere bekende manier om de prevalentie te corrigeren is door de vertekende schatting te corrigeren met een kansenmatrix (tabel 3). Een kansenmatrix kan makkelijk geconstrueerd worden vanuit de steekproef door simpelweg de rijen te laten optellen tot 1, ook wel het rij-normaliseren van een matrix genoemd. De gecorrigeerde schatting kan gemaakt worden door de inverse

	Geschat positief	**Geschat negatief**
Daadwerkelijk positief	TP (terecht positief)	FN (fout negatief)
Daadwerkelijk negatief	FP (fout positief)	TN (terecht negatief)

Tabel 1. Voorbeeld van een kruistabel

	Geschat positief	**Geschat negatief**
Daadwerkelijk positief	900	100
Daadwerkelijk negatief	300	700

Tabel 2. Fictief voorbeeld van een steekproef

	Geschat positief	**Geschat negatief**
Daadwerkelijk positief	TP / (TP + FN) = 0,9	FN / (TP + FN) = 0,1
Daadwerkelijk negatief	FP / (TN + FP) = 0,3	TN / (TN + FP) = 0,7

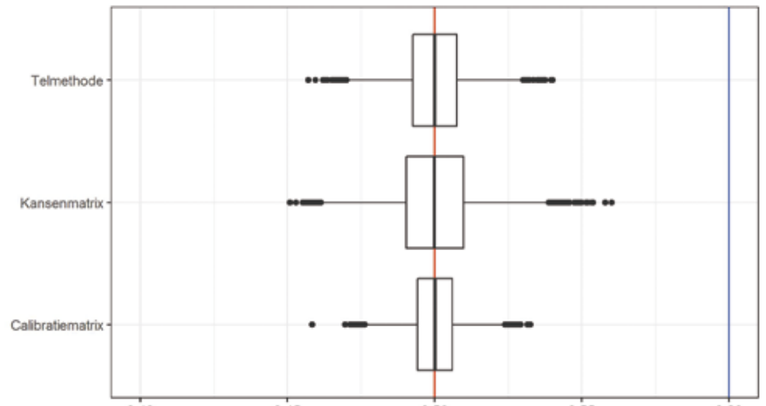
Tabel 3. Kansenmatrix van fictieve steekproef

	Geschat positief	**Geschat negatief**
Daadwerkelijk positief	TP / (TP + FP) = 0,75	FN / (TN + FN) = 0,125
Daadwerkelijk negatief	FP / (TP + FP) = 0,25	TN / (TN + FN) = 0,875

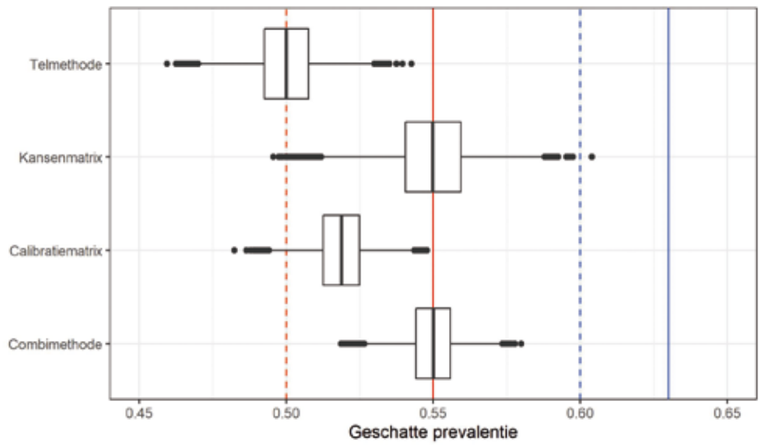
Tabel 4. Kalibratiematrix van fictieve steekproef

van de getransponeerde kansenmatrix te berekenen en te vermenigvuldigen met de vertekende schatting van de prevalentie. Een minder bekende manier om de prevalentie te corrigeren is door de vertekende schatting te corrigeren met een kalibratiematrix (tabel 4). Waar de kansenmatrix geconstrueerd wordt door het rij-normaliseren van de kruistabel van de steekproef, kan de kalibratiematrix geconstrueerd worden door het kolom-normaliseren van diezelfde kruistabel. De gecorrigeerde schatting kan gemaakt worden door de kalibratiematrix te vermenigvuldigen met de vertekende schatting van de prevalentie.

De drie correctiemethoden geven een onvertekende schatting van de prevalentie, maar de variantie is verschillend bij ieder van deze methoden. Een simulatie van 10.000 steekproeven laat zien dat alle schattingen van iedere methode dichter bij de daadwerkelijke prevalentie zitten dan de vertekende prevalentie (figuur 1). De methode met de kalibratiematrix varieert het minst en de methode met de kansenmatrix varieert het meest onder deze set van parameters. Het is bewezen dat de methode met de kalibratiematrix altijd minder varieert dan de methode met de kansenmatrix. Dit komt omdat er bij de



Figuur 1. Boxplot van 10.000 simulaties om de prevalentie te schatten. De daadwerkelijke prevalentie is 0.50 (rood), terwijl de vertekende prevalentie 0.6 (blauw) is. Voor de simulatie zijn de dezelfde gegevens als in de tabellen zijn gebruikt: sensitiviteit = 90%, specificiteit = 70%, daadwerkelijke prevalentie = 50%, steekproefomvang van 2.000 op een zeer grote populatie



Figuur 2. Boxplot van 10.000 simulaties om de prevalentie te schatten. De nieuwe daadwerkelijke prevalentie is 0,55 (rood, solide), terwijl de vertekende prevalentie 0,63 (blauw, solide) is. De oude daadwerkelijke prevalentie is 0,50 (rood, stippellijn) en de oude vertekende prevalentie is 0,60 (blauw, stippellijn). Voor de simulatie zijn de dezelfde gegevens als in de tabellen zijn gebruikt: sensitiviteit = 90%, specificiteit = 70%, daadwerkelijke prevalentie = 50%, steekproefomvang van 2.000 op een zeer grote populatie

methode met de kalibratiematrix niet geïnverteerd hoeft te worden, in tegenstelling tot de methode met de kansenmatrix. Een hogere sensitiviteit of specificiteit leidt tot minder variantie bij de `matrixmethoden`, maar heeft geen invloed op de simpele telmethode. Verder is het vanzelfsprekend dat een grotere steekproef leidt tot minder variantie in alle correctiemethoden.

Verandering over tijd

De prevalentie blijft vanzelfsprekend niet constant over tijd. Om een prevalentie te schatten over tijd, is in een ideale situatie een nieuwe aselecte steekproef nodig. Hier is vaak geen tijd en/of geld voor. Een mogelijke oplossing is om de oude steekproef ook te gebruiken op het nieuwe tijdspunt. Aan de andere kant is het wél relatief makkelijk om een nieuwe vertekende schatting te maken met het slimme algoritme. De informatie van de oude steekproef en de nieuwe vertekende schatting worden gecombineerd tot een nieuwe gecorrigeerde schatting. Het is de vraag of de drie correctiemethoden ook werken op het nieuwe tijdspunt.

De situatie in figuur 1 wordt als uitgangspositie genomen. Stel dat de prevalentie verandert van 50% in de oude situatie naar 55% in de nieuwe situatie. De vertekende prevalentie stijgt van 60% naar 63%, wat betekent dat het verschil niet gecorrigeerd kan worden met alleen de stijging van de vertekende prevalentie. Het valt op dat twee van de drie correctiemethoden een vertekende schatting maken (figuur 2). Ten eerste is het is logisch dat de telmethode een vertekende schatting maakt. De steekproef is notabene getrokken uit een populatie met een prevalentie van 50%. De schatting met de kansenmatrix blijft onvertekend, omdat de waarden van de specificiteit en sensitiviteit onveranderd blijven in de nieuwe situatie. De schatting met de kalibratiematrix is echter vertekend. De verandering van de prevalentie over tijd kan niet helemaal gecorrigeerd worden; de gemiddelde schatting ligt tussen de oude en de nieuwe prevalentie in. Dit komt omdat de kansen in de kalibratiematrix afhankelijk zijn van de daadwerkelijke prevalentie, terwijl de kansen in de kansenmatrix onafhankelijk zijn van de daadwerkelijke prevalentie. De methode met de kansenmatrix kan dus goed de verandering over tijd schatten, maar heeft veel variantie voor de initiële schatting. Daarentegen kan de methode met de kalibratiematrix een goede initiële schatting kan maken, maar kan het niet goed de verandering over tijd verwerken.

De verandering over tijd zorgt er dus voor dat er nog geen goede methode is om de prevalentie ook over tijd goed in te kunnen schatten. Een mogelijke oplossing is om de methodes met de kansenmatrix en de kalibratiematrix te combineren tot de combimethode. Als startpunt wordt de methode met de kalibratiematrix genomen, maar de verandering over tijd wordt beschreven door het verschil in de methode met de kansenmatrix. In figuur 2 valt het op dat de waarden van de combimethode ongeveer dezelfde variantie heeft als de methode met de kalibratiematrix uit figuur 1. De combimethode zorgt ervoor dat de verandering over tijd beter gemodelleerd kan worden, zonder dat er verdere ingewikkelde algoritmes gebruikt hoeven te worden. Het is bewezen dat de combimethode beter presteert dan de individuele correctiemethoden, op een paar extreme situaties na, en dus een goede toevoeging is op het bestaande palet van correctiemethoden.

Conclusie

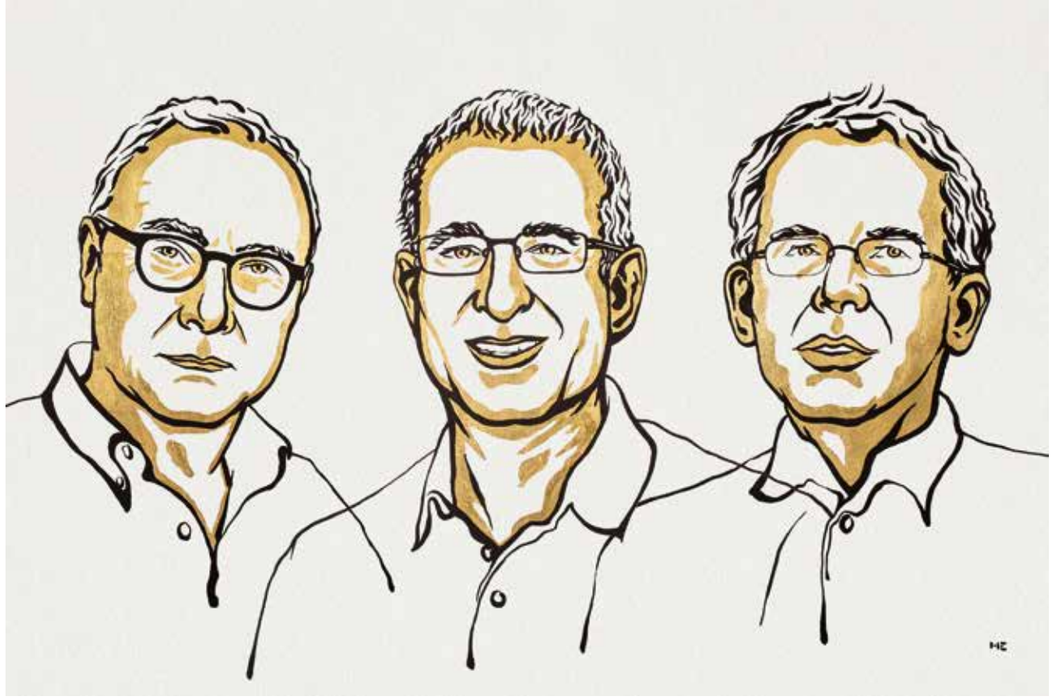
De voornaamste conclusie uit dit artikel is dat prevalenties schatten veel meer is dan alleen goed classificeren en dan optellen. Zelfs goedwerkende classificatie-algoritmes kunnen vertekende uitkomsten genereren die gecorrigeerd moeten worden. Een aantal correctiemethoden is gepresenteerd in dit artikel. Klassieke correctiemethoden werken goed als de prevalentie constant blijft, terwijl de nieuwe combimethode goed werkt als de prevalentie verandert over tijd. Mocht u geïnteresseerd zijn in de wiskundige uitwerkingen van de methodes, wijs ik u graag door naar mijn master thesis (Kloos, 2021) of neem contact met me op.

Literatuur

Meijburg, A. (2021). Sensitiviteit, specificiteit en COVID-19. *STAtOR, 22*(2), 40–44. (https://www.vvsor.nl/wp-content/uploads/2021/07/STAtOR-2021-2-40-44-Meijburg.pdf)

Kloos, K. (2021). *Comparing correction methods to reduce misclassication bias*. Leiden University. Master Thesis. (https://www.github.com/kevinkloos/MasterThesis)

Kevin Kloos doet promotieonderzoek in Quantification Learning aan de Universiteit van Leiden. Gedurende zijn masteropleiding Statistical Science volgde Kevin het deeltraject European Master of Official Statistics (EMOS), waardoor hij zijn scriptie kon schrijven bij het Centraal Bureau voor de Statistiek (CBS). Zijn wetenschappelijke artikel over de combimethode leverde de tweede prijs op in de internationale IAOS Young Statisticians Prize. Dit artikel is een samenvatting van zijn masterscriptie.
E-mail: k.kloos@fsw.leidenuniv.nl



David Card, Joshua Angrist en Guido Imbens, winnaars van de Nobel Economics Prize 2021. Illustratie: Niklas Elmehed | Nobel Prize Outreach

DE REIKWIJDTE VAN DE COUNTERFACTUAL

over causaliteit, potential outcomes en grafische modellen

Richard Starmans

In 2011 ontving de Amerikaanse informaticus Judea Pearl (1936) de door de Association for Computing Machinery (ACM) ingestelde A.M. Turing Award voor zijn fundamentele bijdragen aan de AI *'through the development of a calculus for probabilistic and causal reasoning'*. In 2021 werd aan de van oorsprong Nederlandse econometrist Guido Imbens (1963) en de Amerikaans-Israëlische econoom Joshua Angrist (1960) de Nobel Memorial Price in Economic Sciences toegekend *'for their methodological contributions to the analysis of causal relationships'*. Overigens sleepten Imbens en Angrist samen 'slechts' de helft van de prijs in de wacht, de andere helft viel ten deel aan de Canadese econoom David Card (1956) voor zijn empirische bijdragen aan de arbeidseconomie en de ontwikkeling en toepassing van *natural experiments*. Dat laatste begrip vormt een belangrijke verbindende schakel tussen de drie winnaars. Volgens de jury leverden de laureaten te onderscheiden, maar complementaire bijdragen, waarin wordt gedemonstreerd hoe oorzaak-gevolg relaties kunnen worden vastgesteld en geanalyseerd in natuurlijke experimenten, die betrekking hebben op complexe *real-life problems* en daarmee verbonden beleidsvraagstukken met een grote maatschappelijke impact.

Causaliteit

Het gegeven dat prestigieuze wetenschappelijke prijzen zoals de Nobelprijs voor Economie en de Turing Award worden toegekend aan onderzoek naar *causal inference* of *causal reasoning* is saillant, aangezien causaliteit van oudsher geldt als een obscuur en weerbarstig begrip. Het denken over oorzaak-gevolg relaties kent een lange en moeizame genealogie, die teruggaat tot Aristoteles en de Stoa en door sommigen – terecht of onterecht – louter met metafysica werd geassocieerd en niet met een wetenschappelijk wereldbeeld. Het gaf volop aanleiding tot controversen, kende geen bruikbare formaliseringen en leidde nauwelijks tot vooruitgang. De ideeëngeschiedenis telt dan ook vele prominente denkers die ex cathedra verordonneerden dat in het wetenschappelijk discours geen plaats behoort te zijn voor zulk een archaïsche notie. Men denke aan Ernst Mach, Bertrand Russell, Karl Pearson, Ludwig Wittgenstein of Paul en Patricia Churchland. Ook in tijden van big data, data science en vooral *deep learning* weerklinkt met name in de populaire literatuur de opvatting dat het begrip causaliteit toch in het gunstigste geval als obsoleet dient te worden beschouwd. Onder meer voormalig Google-onderzoeksdirecteur Peter Norvig, econoom Victor Mayer-Schönberger, wetenschapsjournalist Chris Anderson en 'The Master Algorithm'-auteur Pedro Domingos droegen bij aan een rijk geschakeerd palet aan stellingnames, dikwijls gelardeerd met anti-causalistische allusies, waarvan sommige uiteraard genuanceerder en gematigder zijn dan andere.

De toekenning van voornoemde prijzen mag dan aantonen dat vooruitgang op het gebied van causaliteit wel degelijk mogelijk is, de vernieuwde of voortgezette belangstelling ervoor leidt allerminst tot een verenigd en eensgezind veld. Veeleer is er sprake van een veelstromenland, dat een verzuilde aanblik biedt met tal van tegenstellingen en naast elkaar bestaande benaderingen, die (nog steeds) weinig interactie vertonen, zeker in de overvloedige filosofische literatuur (Illari, 2014) (Starmans, 2018; 2020). Dat laatste is wellicht niet verrassend voor diegenen die menen dat in de filosofie nu eenmaal alleen 'vooruitgang' mogelijk is door het plegen van een intellectuele vadermoord, waarbij radicaal wordt gebroken met de traditie waaruit men voorkomt. De tegenstellingen blijken evenwel ook manifest als we ons beperken tot moderne probabilistische benaderingen, waarin de geschetste vooruitgang nu juist werd geboekt en die vooral in de statistiek, de AI, econometrie en deels in de sociale wetenschappen opgang hebben gemaakt. Opmerkelijk genoeg staan ook de voornoemde laureaten Imbens en Pearl in een aantal opzichten tegenover elkaar en bekennen zich tot twee causale tradities, waarvan de verschillen nogal eens worden uitvergroot (Pearl, 2018; 2021; Imbens, 2020). Waar Pearl vooral binnen de AI furore maakte met zijn grafische modellen of *directed acyclic graphs* (DAG), zijn *do-calculus* en zijn *backdoor-* en *frontdoor* criteria voor identificatie van causale effecten, werkt Imbens vooral binnen de economie aan de integratie van de methode der instrumentele variabelen met de zogenaamde Potential Outcome benadering (PO). Enkele aspecten van de problematiek brengen we hier kort voor het voetlicht.

Counterfactuals

De tegenstelling tussen de PO- en DAG-benadering is opmerkelijk, omdat de verbondenheid verder gaat dan louter de constatering dat beide benaderingen probabilistisch zijn en claimen dat causale verbanden wel degelijk in observationele data kunnen worden vastgesteld. Allereerst zijn de methoden wiskundig grotendeels equivalent; een stelling binnen het DAG-raamwerk is dan een stelling binnen de PO-benadering en vice versa. Daarnaast komen beide voort uit twee statistische en sterk causaal-georiënteerde tradities, die teruggaan tot de vroege jaren twintig van de vorige eeuw. De DAG's vinden hun wortels in de padmodellen van Sewall Wright, een causalist van het eerste uur aan wie Pearl zich in (Pearl, 2018) schatplichtig toont en die hij roemt als herauut en pionier van de door hem geproclameerde causale revolutie (Wright, 1921). De PO-benadering is terug te voeren tot het vroege werk van Jerzy Neyman, die de methode introduceerde in het kader van inzichten van Ronald Fisher op het terrein van *experimental design* en inferentiële statistiek (Neyman, 1923). Later zou Donald Rubin de methode verder ontwikkelen in een context van observationele data. De PO-aanpak wordt dan ook dikwijls getypeerd als Neyman-Rubin modellen. Daarbij mag ook de econometrische invalshoek niet worden vergeten, met name de simultaneous equations methode van Jan Tinbergen, Philip Wright en later Trygve Haavelmo. Philip Wright, de vader van Sewall Wright, wees als een van de eersten op het concept van instrumentele variabelen bij het oplossen van het identificatieprobleem bij vraag-aanbod modellen.

De voor dit korte essay meest relevante overeenkomst tussen DAG's en PO is echter filosofisch en betreft het gebruik van *counterfactuals* als sleutel tot causaliteit. Zo beschouwd impliceert causaal redeneren nadenken over, verwijzen naar, postuleren van en uitspraken doen over tegenfeitelijke werelden, parallel aan de feitelijke of actuele wereld; uitspraken waaraan wel een waarheidswaarde in die tegenfeitelijke wereld lijkt te kunnen of zelfs moet worden toegekend vanuit de actuele wereld. Counterfactuals verwijzen naar verschillende klassen van uitspra-

ken, die we hier gemakshalve typeren als uitspraken van de vorm: als P het geval zou zijn / het geval was geweest / had plaatsgevonden, dan zou Q het geval zijn / het geval zijn geweest / hebben plaatsgevonden. Uiteraard kunnen antecedens, consequens of beide de negaties van P en Q bevatten. Essentieel is hierbij dat P kan verwijzen naar een observatie, feit, toestand, handeling of gebeurtenis en dat P in de actuele wereld verondersteld wordt niet waar te zijn, niet te hebben plaatsgevonden, maar in de tegenfeitelijke wereld wel en vice versa. Ter illustratie wordt vaak het onderscheid gemaakt tussen indicatieve conditionals (als P het geval is, dan is ook Q het geval) en subjunctive conditionals (Als P het geval zou zijn, dan zou ook Q het geval zijn), dat iets van de problematiek verheldert met de toevoeging dat elke counterfactual een subjunctieve conditional is, maar niet omgekeerd. In de algemene vorm van de subjunctieve conditional kan de spreker agnostisch zijn over P, zonder een sterke claim over een tegenfeitelijke wereld, terwijl dit bij een counterfactual wel het geval is (Starmans, 2021a).

Hoe dan ook, counterfactuals werden al vroeg verbonden met de studie van causaliteit en het was de sceptische filosoof David Hume (1711–1776) die daarbij het voortouw nam. Zowel zijn vroege en lijvige *A Treatise of Human Nature* uit 1739 als zijn latere, meer toegankelijke en populaire *An Enquiry concerning Human Understanding* uit 1748 bevat in dit opzicht cruciale passages. In Sectie VII van de *Enquiry* stelt de auteur onder meer: '*We may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second. Or, in other words, where, if the first object had not been, the second never had existed*.' Met name de frase 'in other words' is frappant en verbindt het eerste gedeelte (fragment A), dat een notie van causaliteit als regulariteit of constante conjunctie c.q. opeenvolging in de tijd suggereert met het tweede gedeelte (fragment B), dat feitelijk een counterfactual behelst. Duidelijk is dat de auteur hiermee een sterke verbondenheid van beide suggereert; twee typeringen die equivalent zijn, of – iets zwakker wellicht – middels een relatie van logisch gevolg op elkaar betrokken zijn, of die toch zeker een sterke semantische verwantschap vertonen, waarbij fragment B dient ter verduidelijking of verdere afbakening van fragment A. Waarschijnlijk beschouwde Hume zelf de tegenfeitelijke beschrijving als een voor de lezer meer vertrouwde en intuïtieve notie, een commonsense uitleg, die kon helpen de eerste beschrijving te duiden.

Causaliteit en counterfactuals

In ten minste drie opzichten is dit alles opmerkelijk. Allereerst moet worden opgemerkt dat in de daaropvolgende eeuwen de fragmenten A en B zich langs gescheiden paden zouden ontwikkelen en uitgroeien tot zelfstandige benaderingen van causaliteit. Vandaag de dag gelden de regulariteitsbenadering en de counterfactual-benadering als twee afzonderlijke invalshoeken, naast vele andere. Beide zijn noch tot elkaar te herleiden, noch fungeert de een ipso facto als verduidelijking van de ander. Daarbij komt dat vanuit Hume's strikte concept-empirisme, waarin begrippen rechtstreeks tot impressies te herleiden moeten zijn, het speculeren over niet-geobserveerde c.q. niet-observeerbare toestanden of mogelijke werelden twijfelachtig is. Ze mogen dan als vanzelfsprekend opduiken in gedachte-experimenten, die de filosofie van oudsher heeft gekend, maar een strenge empiristische epistemologie doet doorgaans geen beroep op noties die niet-waarneembaar, speculatief, metafysisch of wellicht zelfs incoherent zijn en niet tot sense data of impressies zijn te herleiden. Zij kunnen hoogstens een illustratieve of heuristische functie vervullen en dienen met de nodige scepsis te worden benaderd. In de derde plaats is het opmerkelijk dat Hume toen reeds de counterfactual te hulp riep om de causale relatie te duiden, omdat de counterfactual een notoir lastig begrip was en in zekere zin nog steeds is. Lange tijd golden counterfactuals als obscuur, een precieze interpretatie bleek problematisch en pogingen tot een waarheidsfunctionele semantiek te komen waren voor velen niet overtuigend. Het zou duren tot ver in de twintigste eeuw voordat men enige greep kreeg op counterfactuals, onder meer door de logisch-linguïstische benadering van Saul Kripke (1940) en zijn mogelijke werelden semantiek, maar vooral met het werk van David Lewis (1941–2001), wiens Counterfactuals uit 1973 inmiddels de status van een moderne klassieker heeft verworven. Maar nog steeds bestaat over de reikwijdte ervan geen volledige consensus. Een traditioneel bezwaar betreft het feit dat counteractuals niet waarheidsfunctioneel zijn: de waarheidswaarde van het geheel is geen functie van de waarheidswaarde van de delen, de deelpremissen. Daar komt bij dat counterfactuals in verschillende talen verschillend worden gerepresenteerd, vooral met betrekking tot tense, modaliteit en aspect. Dat geldt zeker voor het gebruik van de zogenaamde *fake-tense* om de tegenfeitelijke wereld te evoceren. De notie roept dan ook tal van filosofische vragen op: *ontologisch* (kan zo'n wereld bestaan en hoe dan?), *epistemologisch* (hoe kunnen we kennis verwerven van deze wereld, die we niet waarnemen en waarin we niet kunnen tellen, meten en wegen en evenmin interveniëren?), *semantisch* (hoe kunnen we betekenis toekennen aan uitspraken over die wereld vanuit de actuele wereld?) en *pragmatisch* (Is het gebruik ervan eigenlijk wel nodig of wenselijk, zeker in een wetenschappelijke context?). Vele aspecten blijven hier buiten beschouwing, maar de vraag is uiteraard in hoeverre het mogelijk is in een gedachte-experiment wijzigingen in de actuele wereld te bedenken, deze vervolgens te projecteren op een tegenfeitelijke wereld, de (logische en fysische) consequenties hiervan te kennen, vervolgens de coherentie of bestaanbaarheid ervan te postuleren om zo tot een zinvolle vergelijking tussen beide werelden te komen. Men neme het volgende voorbeeld:

> Als Julius Caesar een kat of een priemgetal was geweest, dan zou de wereldgeschiedenis nu een keizer/veldheer hebben gehad die goed kon klimmen of een persoon die bij de rivier de Rubicon de woorden veni, vidi, vici uitsprak en die enkel deelbaar is door 1 of door zichzelf.

Dit voorbeeld mag artificieel zijn en lijkt tot absurditeiten te voeren, waarbij alle klassieke taalfilosofische problemen betreffende *rigid designators*, *crossworld-identity* en consistentie, overerving van eigenschappen, presupposities, ceteris paribus clausules, et cetera opduiken. De kernvraag is uiteraard welke beperkingen dan blijkbaar moeten worden opgelegd aan de menselijke fantasie en zijn ongebreideld vermogen tot imaginatie en aan zijn taalgebruik om daarmee een vorm van hypothetisch redeneren af te dwingen, waarbij deze tegenfeitelijke wereld wel zinvol gedacht kan worden en er, al dan niet in de vorm van een counterfactual, betekenisvolle uitspraken over kunnen worden gedaan. Welke toestanden of variabelen kunnen in zo'n gedachte-experiment redelijkerwijs gewijzigd worden en welke interacties treden op? Kortom: hoe moet de counterfactual beteugeld worden? (Starmans, 2021b)

PO of DAG?

Al met al kan worden opgemerkt dat Hume in zekere zin een vooruitziende blik had, omdat de counterfactual sedert de late 20e eeuw een ware zegetocht beleeft in causale probabilistische formalismen. Daartoe moesten wel de nodige obstakels worden overwonnen en het succes schuilt deels in de wijze waarop de counterfactual aan banden wordt gelegd. In de PO-aanpak wordt de oorzaak-gevolg relatie geanalyseerd door aan elke onderzoekseenheid een tweetal *potential outcomes* Y(1) en Y(0) toe te kennen, die op een subtiele wijze worden geassocieerd met een gemeten afhankelijke variabele Y. Een causaal effect wordt geschat door de wereld waarin een proefpersoon interventie A ondergaat, $Y_i(1)$, te vergelijken met een wereld waarin diezelfde proefpersoon interventie A niet ondergaat, $Y_i(0)$. Het is uiteraard een fysische onmogelijkheid om in beide werelden tegelijkertijd te vertoeven en metingen te verrichten, hetgeen door sommigen zelfs wordt beschouwd als *The Fundamental Problem of Causal Inference*.

De PO-benadering is derhalve een uitgewerkt gedachte-experiment, waarin dit alles wel mogelijk is en men waarden gaat toekennen aan variabelen c.q. grootheden in zowel de feitelijke of 'actuele' wereld als in een tegenfeitelijke wereld. Men kan dit alles ook opvatten als een missing-data problem, waarbij elke individu twee extra kolommen in de dataset ontvangt, waarvan precies één kolom een missing value heeft, afhankelijk van interventie A. Men kan dan met imputatiemethoden alsnog een ieder twee scores toekennen, een individueel causaal effect $Y_i(1) - Y_i(0)$ berekenen, om daarna bijvoorbeeld voorspellingen voor nieuwe cases te doen, et cetera. Degene voor wie dit een brug te ver is kan ook een causale grootheid definiëren, zoals b.v. een average treatment effect ATE = E [Y(1) – Y(0)], gedefinieerd in termen van potential outcomes, die dan gereduceerd moet worden tot een statistische grootheid, zoals bijvoorbeel *associational difference* E(Y= 1 | A=1) – E(Y=1 |A=0), waarin de potential outcomes zijn geëlimineerd en op basis van geobserveerde data de causale relatie met behulp van conditionele afhankelijkheden wordt geduid. Uiteraard is zulk een reductie niet triviaal. Vanwege lineariteit van de verwachtingsoperator E geldt weliswaar E[Y(1) – Y(0)] = E[(Y1)] – E[(Y0)], maar dit is doorgaans niet gelijk aan E(Y | A=1) – E(Y |A=0). Anders zou het beroemde maxime *causation is not association* onwaar zijn. Zeker als de data niet middels een RCT zijn gegenereerd moet er rekening worden houden met confounding. De associatie tussen confounder Z en interventie A wordt 'gebroken' via gangbare methoden als stratificatie en vooral matching, of

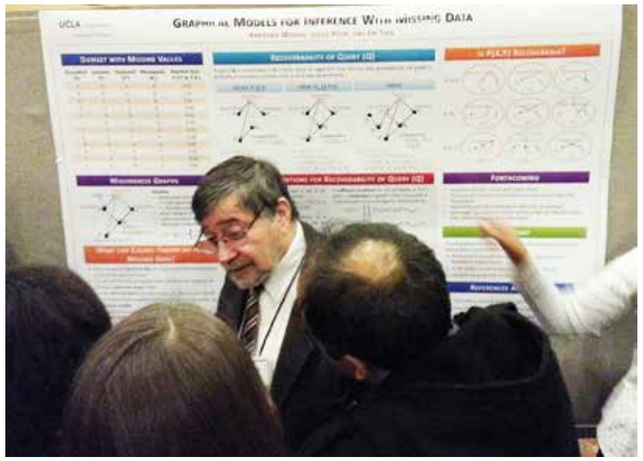
door het opbouwen van een pseudo-populatie middels (*inverse probability*) *weighting*. Een belangrijke aanname om de reductie mogelijk te maken is die van *ignorability* oftewel $(Y(1), Y(0)) \perp\!\!\!\perp A$, waardoor de potential outcomes conditioneel onafhankelijk zijn van A en de *missing values* van de *potential outcomes* genegeerd of *ignored* kunnen worden. Deze aanname wordt ook wel *exchangeability* genoemd, die vrijwel identiek is en inhoudt dat het verwisselen van de *treatment* en *non-treatment* groepen leidt tot dezelfde *potential outcomes*. De groepen zijn dan op de interventie na gelijk en uitwisselbaar. Helaas is deze aanname van *ignorability/exchangeability* in observationele data onrealistisch en men zou daarom kunnen proberen dit in subgroepen van een relevante derde variabele W te bewerkstelligen door te conditioneren. De assumptie van *conditional exchangeability*, oftewel $(Y(1), Y(0)) \perp\!\!\!\perp A \mid W$ leidt dan tot de bescheidener reductie:

$$E[Y(1) - Y(0) \mid W] = E[Y(1) \mid W] - E[Y(0) \mid W] = E(Y \mid A=1, W) - E(Y \mid A=0, W).$$

Het bedoelde marginale effect kan worden bereikt door W 'uit te marginaliseren'.

$$E[Y(1) - Y(0)] = E_W [E[Y(1) - E[Y(0) \mid W]] = E_W [E [Y \mid A=1, W] - E[Y \mid A=0, W]]$$

Deze zeer onvolledige weergave van PO negeert assumpties zoals *positivity* en *unmeasured confounders*, maar illustreert hoe een aloud gedachte-experiment formeel kan worden uitgewerkt en hoe de counterfactual beteugeld kan worden door het streven werelden identiek te houden en alleen te laten verschillen met betrekking tot de interventie. Nog prominenter duikt de counterfactual op in de DAG-benadering van Pearl, die decennialang heeft gewerkt aan een probabilistische benadering die zijns inziens meer recht doet aan de oorzaak-gevolg relatie dan de traditionele statistische benaderingen. Hij spreekt in (Pearl, 2018) zelfs van de 'ladder van causaliteit', waarbij de door hem bepleite verandering in het wetenschappelijke denken middels een drietal stappen of treden gestalte moet krijgen: die van associatie, vervolgens interventie en tot slotte de counterfactual. Hume's regulariteit en de traditionele statistiek blijven volgens de auteur steken op de eerste trede, die van associatie en correlatie, waarbij we alleen kunnen observeren. Bijbehorende vragen als 'Hoe hangen X en Y samen?' en 'Hoe verandert kennis



Judea Pearl tijdens de Conference on Neural Information Processing Systems in 2013

van Y indien X wordt waargenomen?' laten geen diepere, causale conclusies toe. De tweede trede betreft niet alleen waarnemen, maar ook handelen, interveniëren en laat vragen toe als: 'Hoe verandert Y als we actief X veranderen?' De derde trede is volgens Pearl essentieel om echt causaal te kunnen redeneren. Het gaat om *imagining, retrospection* en laat vragen toe als 'Zou Y het geval zijn geweest, als X het geval was geweest?'. Daarmee vormt de counterfactual sterker dan in de PO-aanpak het sluitstuk van de causale redenering en een wezenlijk kenmerk van de menselijke conditie; zonder deze dreigt het gehele project van de sterke AI te mislukken en wordt / blijft de mens overgeleverd aan deep learning. (Pearl, 2018) vormt in feite een lange, soms polemische aanklacht tegen de traditionele statistiek en epistemologie, die volgens hem op de eerste trede van de ladder zijn blijven steken, maar soms ook tegen de methode van Rubin en Imbens, en tegen iedereen die het belang van de grafische representaties niet erkent. Die vormen inderdaad een belangrijk hulpmiddel bij het specificeren van het causale model, identificatie van het effect en het begrijpen van confounding, mediation of effect-modification, maar dit aspect moet hier verder buiten beschouwing blijven.

Imbens lijkt overigens niet geïnteresseerd in de mentalistische claims van Pearl en onderzoekt waarom de DAG's in de economische literatuur nauwelijks voet aan de grond krijgen (Imbens, 2020). Zo is de PO-benadering zijns inziens, ook historisch beschouwd, veel geschikter voor echte, grootschalige (causale) problemen in de economie, terwijl de DAG-benadering veeleer een oplossing lijkt 'op zoek naar' een probleem of vooral tracht *toyproblems* op te lossen. Ook stelt Imbens dat (Pearl, 2018) vooral gaat over identificatie van het causale effect en niet over wat er aan vooraf gaat (model-specificatie, het tekenen van de causale graaf) en wat er na komt (inference!). Bovendien kent het domein van de economie voorwaarden zoals monotoniciteit, die in de DAG-benadering niet goed te formuleren zijn. A fortiori suggereert Imbens dat Pearls ladder met een vierde trede moet worden uitgebreid, die van *reversed causality*, zodat ook van gevolg naar oorzaak kan wordt geredeneerd. Het weerwoord (Pearl, 2021) liegt er evenmin om en het laatste woord over deze kwestie is uiteraard nog niet gezegd. Toch wordt de soep niet altijd zo heet gegeten. Veel moderne literatuur en MOOCS op het gebied van causal inference zijn hybride en eclectisch; concepten en notaties van PO en DAG's worden door elkaar gebruikt en vooral selectief, wanneer het uitkomt en dat geldt met name binnen *machine learning*.

Epiloog

De filosoof Willard V. O. Quine schreef ooit in zijn *Methods of Logic* (1950)*: '... any adequate analysis of the contrafactual conditional* (counterfactual, RS) *must go beyond truth values and consider causal connections, or kindred relationships, between matters spoken of in the antecedent of the conditional and matters spoken of in de consequent*.' Quine, nota bene zelf empirist pur sang, behaviorist en fysicalist stelt dat de counterfactual alleen te begrijpen is vanuit het 'obscure' begrip der causaliteit en niet tot een extensionele logica is te herleiden. Indien causaliteit nodig is om de counterfactual te begrijpen en vice versa dreigt een vicieuze cirkel te ontstaan, al lijkt de recente vooruitgang in het probabilistische causale onderzoek dit te logenstraffen. Toch is het gebruik van causale uitdrukkingen, waarvan ons taalgebruik is doordrenkt, dikwijls kwalitatief, deterministisch en zeker niet probabilistisch. Pogingen het begrip te axiomatiseren indachtig het adagium 'zonder counterfactual geen causaliteit' zijn bovendien problematisch omdat de counterfactual in het dagelijkse taalgebruik moeilijker te beteugelen lijkt. Iemand die in een dialoog/taalspel de uitspraak 'P veroorzaakt Q' doet, daarmee een taalhandeling verricht en bijbehorende commitments aangaat inzake de feitelijke wereld die hij kent, zou zich evenzeer moeten committeren aan uitspraken en een daarmee geassocieerde tegenfeitelijke wereld, die hij niet kent en wellicht niet kenbaar of zelfs incoherent is. Zouden minder weerbarstige 'als–dan'-beweringen niet ten minste overwogen moeten worden? De vraag welke verdedigingsplicht iemand op zich neemt is vooral relevant in tijden van Explainable AI (EAI). (Miller, 2019) beargumenteert dat het zoeken naar (causale) verklaringen zonder een sociale en linguïstische context niet zinvol is en bovendien betreft EAI niet zozeer de context van discovery en de onderliggende causale structuur van de werkelijkheid, maar veeleer de context of justification, waarin oorzaken, redenen en motieven in een gespecificeerd, retorisch gefaseerd taalspel gestalte moeten geven aan de gezochte verklaringen (Starmans, 2020).

Dit alles neemt niet weg dat met de opmars van *causal inference* de counterfactual als aloud gedachte-experiment volop in ere is hersteld.

Literatuur

Illary, P., & Russo, F. (2014). *Causality; philosophical theory meets scientific practice*. Oxford.

Imbens, G. (2020). Potential Outcome and Directed Acyclic Graph approaches to causality; Relevance for empirical practice in economics. *Journal of Economic Literature, 58*(4)

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. *Statistical Science, 5*(4), 1923.

Pearl, J and D. MacKenzie (2018). The Book of Why, the new science of cause and effect, New York.

Pearl, J. (2021); http://causality.cs.ucla.edu/blog/index.php/2020/01/29/on-imbens-comparison-of-two-approaches-to-empirical-economics/

Starmans, R. J. C. M. (2020). Prometheus unbound or Paradise regained: the concept of causality in the contemporary AI-data science debate. *Journal of the French Statistical Society, 161*(1), 4–41.

Starmans, R. J. C. M. (2021a). Over padmodellen, structurele vergelijkingen en de roep om Explainable AI. *STAtOR, 22*(2).

Starmans, R. J. C. M. (2021b). De tegenfeitelijke wereld van de counterfactual; obscuur gedachte-experiment of sleutel tot de causale redenering. *Filosofie Tijdschrift, 31*(5).

Wright, S. (1921). Correlation and Causation. *Journal of Agricultural Research, 20*, 557–585.

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

Foutlokalisatie in enquêtedata met behulp van (zachte) regels en machine learning

Het Centraal Bureau voor de Statistiek (CBS) maakt jaarlijks productiestatistieken. Om betrouwbare statistieken te kunnen maken is het van belang dat de data van goede kwaliteit zijn. Er wordt daarom onder andere gewerkt aan het verbeteren van het automatische data-opschoningsproces, zodat fouten in enquêtedata worden gelokaliseerd. Na deze opschoning zitten er nog steeds fouten in de data. Daarom worden de data aanvullend handmatig opgeschoond. Voor mijn afstudeerscriptie (onderdeel van het project EBN2.x) heb ik twee methoden onderzocht voor het lokaliseren van fouten in enquêtedata om het opschoningsproces te verbeteren.

Tania Keijzer

Project EBN2.x heeft als doel de enquêtedata zo vroeg mogelijk consistent te maken, zo veel mogelijk automatisch te corrigeren en het handmatige werk te sturen via kwaliteitsmaten. In dit omvangrijke project wordt onder andere gewerkt aan het verbeteren van het automatische opschoningsproces. Hierbij worden voor de hand liggende fouten gecorrigeerd, foutieve waardes opgespoord en vervangen door geldige waardes met behulp van harde regels. Harde regels zijn gedefinieerde regels waar de data altijd aan moeten voldoen, zoals regels over het domein van variabelen en regels die het verband tussen variabelen aangeven. Na het uitvoeren van het opschoningsproces zitten er nog steeds veel fouten in de enquêtedata die handmatig gecontroleerd moeten worden. Dit kost veel tijd en moeite. Hoe zou dit opschoningsproces verbeterd kunnen worden, zodat er minder handmatige controles uitgevoerd moeten worden?

Er zijn twee mogelijke verbeteringen onderzocht. Ten eerste worden in het huidige opschoningsproces naast de harde regels ook zachte regels opgenomen. Dit zijn regels die wijzen op opvallende waarnemingen die als fout gezien kunnen worden, maar waarbij dit niet altijd het geval hoeft te zijn. Ten tweede wordt gekeken of de fouten gelokaliseerd kunnen worden met behulp van een *machine learning* benadering. Dit onderzoek is toegepast op de enquêtedata van de productiestatistieken. Ik heb de methodes toegepast op bedrijven uit de industriesector met 20 tot 49 werkzame personen.

Mixed integer programmeerprobleem

Per bedrijf worden er meer dan honderd variabelen uitgevraagd. Het is van belang om te achterhalen welke variabelen fout zijn bij elk bedrijf. Door het grote aantal variabelen is het niet evident welke (combinatie van) variabelen ervoor zorgen dat regels geschonden worden. In het huidige opschoningsproces worden foutieve waardes opgespoord met behulp van een mixed integer programmeerprobleem (MIP). De doelfunctie, een gewogen som van het aantal foutieve variabelen, wordt geminimaliseerd waarbij aan alle harde regels moet worden voldaan:

$$D_{HARD} = \sum_{j=1}^{p} w_j y_j$$

waarbij $y_j = 1$ als de variabele x_j als foutief wordt aangewezen, anders $y_j = 0$, w_j is de bijbehorende wegingsfactor en p is het aantal variabelen dat gewijzigd kan worden in de data (De Waal, Pannekoek & Scholtus, 2011). Met dit model worden met de harde regels 89% van de fouten niet gedetecteerd (*recall* = 0,11) en 31% van de voorspelde fouten is officieel geen fout (*precision* = 0,69). Dit levert een f1-score, het harmonisch gemiddelde tussen de recall en precision, van 0,19. In het werkelijke productieproces zijn deze cijfers waarschijnlijk anders dan in dit onderzoek, omdat hier niet de data van het laatste stadium van het opschoningsproces gebruikt kon worden als input van het MIP-model. Na het lokaliseren van de fouten worden geldige waardes ingevuld zodat aan alle harde regels wordt voldaan.

In het kader is een voorbeeld van het MIP-model met de harde regels uitgewerkt. Hierin wordt maar één harde regel meegenomen in het MIP-model. In werkelijkheid zijn er honderden harde regels bij de productiestatistieken, wat het probleem zo complex maakt.

Stel de variabelen 'winst' ($x1$), 'omzet' ($x2$) en 'kosten' ($x3$) hebben respectievelijk de wegingsfactoren 2 ($w1$), 4 ($w2$) en 3 ($w3$), de harde regel 'winst is omzet minus kosten' geldt en de volgende dataset geeft per rij de ingevulde enquête van een bedrijf weer:

Bedrijf	Winst	Omzet	Kosten
1	800	1000	300
2	300	500	200

De bovenste rij schendt de harde regel. Daarom moet minstens één van deze variabelen gewijzigd worden om aan de harde regel te voldoen. De variabele 'winst' wordt als fout aangemerkt en vervangen door een missende waarde, omdat deze variabele de laagste wegingsfactor heeft. Op deze manier wordt de doelfunctie geminimaliseerd. In het vervolg worden de missende waardes vervangen door geldige waardes. Bij dit probleem zou eventueel de volgende zachte regel geformuleerd kunnen worden: 'winst mag niet groter zijn dan 50% van de omzet'.

Mixed integer programmeerprobleem met de zachte regels

Door de zachte regels toe te voegen aan het MIP-model wordt de doelfunctie gewijzigd naar het minimaliseren van de som van het aantal foutieve variabelen en het aantal zachte regels dat wordt geschonden (Scholtus, 2013). Hierbij geldt nog steeds de voorwaarde dat aan alle harde regels moet worden voldaan.

$$D = D_{HARD} + D_{ZACHT}$$

$$D_{ZACHT} = \sum_{k=1}^{K_s} S_k z_k$$

waarbij z_k=1 als de zachte regel k niet voldoet, anders z_k=0, S_k is de bijbehorende wegingsfactor en K_s is het aantal zachte regels. Bij het minimaliseren van deze doelfunctie wordt dus een afweging gemaakt tussen de wens om zo weinig mogelijk variabelen te wijzigen en de wens om aan zo veel mogelijk zachte regels te voldoen. Het aantal zachte regels dat wordt meegenomen in het MIP-model is heel flexibel, waarbij de rekentijd wel groter wordt als er meer regels worden meegenomen.

In dit onderzoek zijn met verschillende technieken empirisch zachte regels geformuleerd in de vorm van een lineaire ongelijkheid. Hierbij is enkel een selectie aan zachte regels meegenomen in het MIP-model. Uit de evaluatie blijkt dat toevoegen van de zachte regels aan het huidige MIP-model ervoor zorgt dat er meer fouten worden gelokaliseerd, maar dat er ook meer locaties onterecht als fout worden aangewezen. Met de MIP-modellen met de zachte regels worden maximaal 3% meer fouten gelokaliseerd dan het huidige MIP-model, wat een relatief kleine verbetering is.

Machine learning benadering

Als alternatief voor het toevoegen van zachte regels aan het MIP-model worden verschillende machine learning benaderingen toegepast. De fouten moeten in meerdere variabelen gelokaliseerd worden, waardoor het een multi-label classificatieprobleem is. Het machine learning model multi-label K-Nearest Neighbour (MLKNN), dat met een multi-label dataset kan omgaan, is toegepast.

MLKNN is een uitbreiding van het K-Nearest

Neighbour model (Zhang & Zhou, 2005). In de multi-label dataset heeft elk bedrijf niet één maar meerdere labels, één voor elke variabele. Elk label bestaat uit twee klassen. De klasse '1' betekent dat een variabele voor een bedrijf fout is, anders de klasse '0'. Allereerst worden voor een nieuw bedrijf de K dichtstbijzijnde buren bepaald met de Euclidische afstand. Vervolgens wordt voor elke variabele de kans bepaald dat deze fout is. Daarna wordt het maximum posteriori principe gebruikt om de labels van het nieuwe bedrijf te bepalen.

Daarnaast zijn de data eerst getransformeerd naar meerdere single-label classificatieproblemen, zodat vervolgens de machine learning modellen Naive Bayes en Extreme Gradient Boosting toegepast kunnen worden. Na het uitvoeren van alle machine learning modellen geldt dat nog niet aan alle harde regels wordt voldaan. Om deze reden wordt na een machine learning model altijd het huidige MIP-model uitgevoerd om waar nodig meer fouten aan te wijzen.

Elk machine learning model heeft zijn eigen parameters die afgesteld kunnen worden, bijvoorbeeld het aantal dichtstbijzijnde buren bij MLKNN. Na het afstellen van de juiste parameters door middel van kruisvalidatie geeft MLKNN voorafgaand aan het huidige MIP-model de hoogste recall (0,50). Daarentegen behaalt het machine learning model Extreme Gradient Boosting voorafgaand aan het huidige MIP-model de hoogste f1-score (0,48). Het machine learning model MLKNN behaalt niet de hoogste f1-score, omdat dit model veel meer locaties onterecht als fout aanwijst, waardoor de precision een stuk lager is. Met de machine learning benadering worden maximaal 39% meer fouten gelokaliseerd dan het huidige MIP-model.

Conclusie en aanbevelingen

Tijdens dit onderzoek zijn meerdere modellen onderzocht die tot een verbetering van het huidige MIP-model leiden. Het uitvoeren van de machine learning benadering leidt tot de grootste verbetering. Het machine learning model MLKNN voorafgaand aan het huidige MIP-model lokaliseert veel meer fouten dan het huidige MIP-model, maar wijst ook veel meer locaties onterecht als fout aan. Daarom wordt aanbevolen eerst te onderzoeken of de locaties die onterecht als fout zijn aangewezen nauwkeurig vervangen kunnen worden door geldige waardes voordat dit model in de praktijk wordt toegepast.

Een alternatief is het machine learning model Extreme Gradient Boosting voorafgaand aan het huidige MIP-model. Hiermee worden nog steeds veel fouten gelokaliseerd, maar er worden minder locaties onterecht als fout aangewezen ten opzichte van MLKNN. Er geldt wel dat nog steeds meer locaties onterecht als fout worden aangewezen in vergelijking met het huidige MIP-model. Daarom kan overwogen worden om in een vervolgonderzoek de methodes voor het invullen van geldige waardes te verbeteren als de locaties niet nauwkeurig vervangen kunnen worden door geldige waardes.

Discussie

In dit onderzoek is alleen een voorspelling gemaakt voor bedrijven uit de industriesector met 20 tot 49 werkzame personen. Stel in de toekomst worden de zachte regels in het huidige opschoningsproces meegenomen, dan moet gekeken worden welke zachte regels voor bedrijven in andere sectoren opgesteld kunnen worden. Er worden namelijk verschillende versies van enquêtes uitgevraagd voor de productiestatistieken. Voor het toepassen van een machine learning benadering in de praktijk moet gekeken worden op welke data de algoritmen getraind kunnen worden voor de voorspelling van bedrijven uit andere sectoren.

In dit onderzoek is aangenomen dat een model beter presteert als er meer fouten gelokaliseerd worden, een verhoging van de recall, zonder dat dit voor een extreme verlaging van de precision zorgt. Daarom moet de f1-score ook zo hoog mogelijk zijn. Echter, een model kan ook verbeteren als er minder locaties onterecht als fout worden aangewezen, een verhoging van de precision. De eindpresentatie die ik over mijn onderzoek gaf, eindigde daarom met de vraag: 'Wat is beter: een hogere *recall* of *precision*?'.

Literatuur

De Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. Wiley, Hoboken.

Zhang, M.-L., & Zhou, Z-H. (2005). A k-nearest neighbor based algorithm for multi-label classification. *IEEE International Conference on Granular Computing*, Vol. 2, 718–721, doi10.1109/grc.2005.1547385.

Scholtus, S. (2013). Automatic editing with hard and soft edits. *Survey Methodology, 39*(1), 59–89.

Tania Keijzer is afgestudeerd in de Toegepaste Wiskunde aan de Haagse Hogeschool. Voor haar afstudeerscriptie bij de afdeling Methodologie van het CBS heeft ze onderzoek gedaan naar het lokaliseren van fouten in enquêtedata. Haar scriptie is beoordeeld met een 8,5.
E-mail: taniakeijzer@gmail.com

Henk Tijms

column

Foto: Alejandro Garay via Pixabay

DE DODELIJKE GLAZEN BRUG

De Squid Game is één van de meest bekeken Netflix-series die ooit uitgezonden is. In een krankzinnig sadistische ratrace – gebaseerd op ouderwetse kinderspelletjes – krijgen een paar honderd aan lager wal geraakte mensen de kans om nog wat van hun leven te maken. Hoewel de serie een leeftijdsgrens van 16 jaar en ouder kende, was de serie onder basisschoolleerlingen razend populair en werd massaal nagespeeld op het schoolplein, tot ongenoegen van ouders en leerkrachten. In het bloedstollende spel Glass Stepping Stones uit de zevende aflevering van de serie moeten 16 deelnemers een brug met 18 treden oversteken[1]. Eén voor één proberen de deelnemers veilig de overkant te bereiken. Bij elke trede heeft een deelnemer de keuze om het linkerpaneel of het rechterpaneel te kiezen, waarbij één van de twee panelen uit normaal glas bestaat dat meteen breekt als er op wordt gestapt en de andere paneel gehard glas bevat, waar veilig op gestapt kan worden zonder dat het glas breekt. Voor elke trede is er een kans van 50% dat het linkerpaneel gehard glas heeft, en als de linkerkant gehard glas heeft, heeft de rechterkant dat niet (en omgekeerd). Het is onmogelijk om het verschil te zien tussen het normale en het geharde glas. Het slechte nieuws is dat als een deelnemer op een paneel met normaal glas springt, het glas breekt en de deelnemer naar beneden tuimelt met de dood als gevolg. Het goede nieuws is dat het offer niet voor niets is geweest, want het gebroken paneel geeft alle overgebleven deelnemers waardevolle informatie over wat de juiste weg naar veiligheid is. Verder wordt verondersteld dat elke deelnemer ook de informatie heeft wat de veilige panelen zijn die door voorgaande deelnemers gekozen zijn. In volgorde probeert elke deelnemer de brug over te steken en de deelnemer blijft in beweging totdat de deelnemer ofwel succesvol alle 18 treden op de brug heeft overgestoken, ofwel tussentijds naar beneden is getuimeld. Wat is het verwachte aantal overlevende deelnemers, wat is voor elke deelnemer de kans op overleven en wat is de kansverdeling van het aantal deelnemers dat overleeft?

Dit spel is dodelijker dan het spel van Russisch roulette. Het Markovketen concept is ideaal voor een kanstheoretische analyse van het spel. Beschouw een absorberende Markovketen met 20 toestanden i = 0, 1, ..., 18, 19. Toestand i met $1 \leq i \leq 18$ betekent dat het spel gevorderd is tot trede i waar echter echter een deelnemer op het normale glas van deze trede gesprongen is en jammerlijk het leven gelaten heeft, toestand 19 betekent dat een deelnemer veilig op trede 18 beland is en dus de overkant bereikt heeft, en toestand 0 is een hulptoestand die het begin van het spel markeert. Toestand 19 wordt genomen als een absorberende toestand van de Markovketen, dat wil zeggen als het proces toestand 19 bereikt heeft dan blijft het daar in. Voor i = 0, 1, ..., 18, worden voor de Markovketen de één-staps overgangskansen p_{ij} van toestand i naar toestand j gegeven door

$$p_{ij} = \left(\frac{1}{2}\right)^{j-i} \text{ voor } j = i+1, .., 18 \quad \text{en} \quad p_{i,19} = \left(\frac{1}{2}\right)^{18-i}.$$

Voor toestand 19 is $p_{19,19} = 1$ en de overige p_{ij} zijn 0. Laat $\mathbf{P}$ de 20×20 matrix van één-staps overgangskansen zijn. De simpele berekeningen gaan nu als volgt. De matrix producten $\mathbf{P}^k$ worden berekend voor $k = 1, ..., 16$. Noteer met a_k de kans dat deelnemer k overleeft en met d_k de kans dat precies k deelnemers overleven voor $k = 0,1..., 16$. Dan

$$a_k = p_{0,19}^{(k)} \quad \text{voor } k = 0, 1, \ldots 16,$$

dus a_k wordt gegeven door het (0, 19)de element van $\mathbf{P}^k$. De d_j's kunnen vervolgens worden berekend met

$$d_{16-k+1} = a_k - a_{k-1} \quad \text{for } k = 1, 2, \ldots, 16,$$

waarbij $a_0 = 0$. Immers a_k geeft ook de kans dat 16–k+1 of meer deelnemers overleven, omdat elke deelnemer precies weet wat de paneelkeuzes van de voorgaande deelnemer zijn geweest. Dus als deelnemer k veilig de overkant bereikt, dan bereiken ook degenen na deelnemer k veilig de overkant. Dit betekent dat de kans dat in totaal d–k+1 deelnemers overleven gevonden wordt door de kans a_{k-1} af te trekken van de kans a_k voor k ongelijk aan nul. De kans d_0 wordt uiteraard gegeven door $1 - \sum_{k=1}^{16} d_k$. De verwachte waarde van het aantal overlevende deelnemers wordt berekend als $\sum_{k=1}^{16} d_k$. De matrix berekeningen leiden tot

$$a_1 = 0{,}000,\ a_2 = 0{,}000,\ a_3 = 0{,}001,\ a_4 = 0{,}004,$$
$$a_5 = 0{,}015,\ a_6 = 0{,}048,\ a_7 = 0{,}119,\ a_8 = 0{,}240,$$
$$a_9 = 0{,}407,\ a_{10} = 0{,}593,\ a_{11} = 0{,}760,\ a_{12} = 0{,}881,$$
$$a_{13} = 0{,}952,\ a_{14} = 0{,}985,\ a_{15} = 0{,}996,\ a_{16} = 0{,}999.$$

De verwachte waarde van het aantal overlevenden is 7,000076, een waarde vrijwel gelijk aan 7 (in de Netflix-episode was het werkelijke aantal overlevenden gelijk aan 3, een opmerkelijk klein aantal in het licht van $d_0 + d_1 + d_2 + d_3 = 0{,}047$). De waarde 7 krijg je met het volgende heuristische argument: als het aantal treden groot genoeg is, dan is het aantal treden waarvan de samenstelling door een deelnemer onthuld wordt bij benadering geometrisch verdeeld met parameter 1/2 en verwachtingswaarde 2, en dit maakt het plausibel dat gemiddeld genomen ongeveer 9 deelnemers moeten worden opgeofferd opdat de resterende 16 – 9 = 7 deelnemers veilig de overkant kunnen bereiken. Een ware slachting onder de deelnemers zou het geval zijn geweest als de spelers alleen de kapotte panelen hadden kunnen zien en geen verdere informatie hadden gehad. Simulatie lijkt de enige praktische aanpak voor de probabilistische analyse van deze variant. Honderdduizend simulatie runs geven de schatting 0,24 voor het verwachte aantal overlevenden en de schatting 0,11 voor de kans dat de laatste deelnemer overleeft.

Noot

1. https://www.youtube.com/watch?v=190FNytuOZO

Henk Tijms is emeritus-hoogleraar operations research aan de Vrije Universiteit en auteur van diverse leerboeken over operations research en kansrekening. Zijn meest recente boeken zijn *Basic probability; What every math student should know* (World Scientific Press, 2021, 2e druk) en *Operations Research; An introduction to models and methods* met de co-auteurs R. Boucherie en A. Braaksma, (World Scientific Press, 2021).
Homepage: https://personal.vu.nl/h.c.tijms/
E-mail: h.c.tijms@xs4all.nl

Good news from Young Statisticians

On May 3rd of 2021, the Young Statisticians section celebrated its 10th anniversary; on November 11th, current and former board members gathered to raise a glass to this occasion. We enjoyed a delicious dinner at Scarlatti in Leiden and concluded the evening with a Young Statisticians anniversary cake. The current board members were curious to hear more about the stories and experiences of the previous boards and received many tips on the organization of future events. We received helpful advice on pub quiz topics, discussed which of all past Young Statisticians logos was the best, and listened to mushroom hunting adventures.

Due to the corona measures, we unfortunately had to cancel our traditional New Year's drinks. However, we are definitely planning to organize various offline events again in the 'measure-low' months. So keep an eye on our newsletter (and that of the VVSOR) for the dates!

LEDEN GEZOCHT VOOR DE WERKGROEP OPEN STATISTICA

Afgelopen jaar is op de ALV door een aantal leden de wens uitgesproken om van *Statistica Neerlandica* een volledig *open access* blad te maken, mogelijk ook zonder *article processing charge*. Dit vanwege de evidente voordelen die *open science*, en *open access* in het bijzonder, met zich meedragen.

Zo'n wijziging is echter niet zonder gevolgen. *Statistica Neerlandica* werkt momenteel succesvol samen met Wiley en krijgt vanuit de uitgever de nodige support. Daarnaast is *Statistica* een grote bron van inkomsten voor de vereniging: de inkomsten uit royalties van Wiley zijn van dezelfde ordegrootte als de inkomsten uit contributies.

Bij een eventuele structuurwijziging moeten we dus niet over één nacht ijs gaan. Op de Algemene Ledenvergadering van 17 maart zal daarom een werkgroep Open Statistica worden opgericht. De nieuwe werkgroep gaat zich verdiepen in verschillende mogelijkheden om meer open access bij *Statistica Neerlandica* te bewerkstelligen en de consequenties van die mogelijkheden. Op de ALV van 2023 zal de werkgroep verschillende scenario's presenteren.

Heb je interesse om deel te nemen aan deze werkgroep? Neem dan contact op met het bestuur, db@vvsor.nl.