

STAtOR

periodiek van de VVSOR jaargang 21, nummer 2, juni 2020

Minder voedselverspilling door een dynamische
houdbaarheid en korting

De nulhypothese-significantie toets; Tijd voor
een paradigmaverschuiving

Rekenen aan maatregelen om drukke treinen te
ontlasten

Is de startprocedure bij schaatsen oneerlijk?

From killer nurses to quantum entanglement,
and back (Part 2)

Statcheck: een spellingchecker voor statistiek



STATOR

Jaargang 21, nummer 2, juni 2020

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

Redactie

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

Kopij en reacties richten aan

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

Bestuur van de VVSOR

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

Leden- en abonnementenadministratie van de VVSOR

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

Voor advertenties

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

Ontwerp en opmaak

Pharos, Nijmegen

Uitgever

© Vereniging voor Statistiek en Operations Research
ISSN 1567-3383

ACHTERHAALDE WETENSCHAP?

Wij waren nét op tijd: op donderdag 12 maart vond de zeer geslaagde Annual Meeting plaats en de dag daarop werd de lockdown een feit. Het scheelde nog geen dag of we hadden de bijeenkomst moeten afbreken en de zaal per direct moeten ontruimen. Sindsdien werd duidelijk hoe groot de rol van statistiek soms kan zijn, en ook de Operations Research is uitgebreid ingezet bij het verdelen van schaarse middelen, of dat nu IC-bedden of mondkapjes waren. De redactie overweegt een speciaal nummer samen te stellen over al die aspecten dat begin volgend jaar zal verschijnen. Natuurlijk zijn alle spontane bijdragen en suggesties van onze leden van harte welkom!

Maar eerst terug naar dit nummer: soms is de werkelijkheid sneller dan de wetenschap. En soms loopt de wetenschap voorop. Deze dooddoener lijkt in dit nummer toch enigszins geïllustreerd te worden. De artikelen die u zult aantreffen waren grotendeels al in de pijp lijn toen de corona toesloeg. Zo hadden we een artikel over manieren om drukke treinen te ontlasten ontvangen, vóór dat de bezetting van treinen een zeer drastische verandering onderging door de lockdown. Is het daarmee achterhaald? Wij denken van niet, de achterliggende methodiek kan zó worden vertaald naar andere toepassingen en waarschijnlijk zal het ooit ook in de treinen weer net zo druk worden als we het tot voor kort gewend waren. Aan de andere kant schrijven Marjolein Buisman en haar collega's over maatregelen om voedselverspilling tegen te gaan: dat onderwerp is natuurlijk al lange tijd actueel, maar het heeft in de huidige omstandigheden wel een extra lading gekregen.

De verdere artikelen zijn weer zeer gevarieerd. Er is het tweede deel gebaseerd op Richard Gills afscheidscollege, Miriam Loois analyseert de startprocedure bij het schaatsen en Michèle Nuijten presenteert Statcheck: een spellingchecker voor statistiek. Al geruime tijd is er grote aandacht voor het gebruik van de nulhypothese-significantie toets, onze veel grotere zusterorganisatie ASA heeft in haar tijdschrift *The American Statistician* (Vol. 70, 2016, Issue 2) een formeel standpunt gepubliceerd over p-Values: 'Context, Process, and Purpose'. Jonáh Stunt en Michiel de Boer behandelen dit onderwerp in een zeer boeiend artikel.

STAtOR zou niet compleet zijn zonder enkele columns, in dit nummer zelfs vier! Kennelijk hebben onze columnisten geprofiteerd van het gedwongen thuiszitten. Dat onze columns goed gelezen worden blijkt uit een reactie van Merith Pelger en Dennis Huisman op de uitlatingen van Gerrit Stermerdink in ons laatste nummer. Daar zijn we blij mee!

Wij wensen u veel leesplezier.

Tijdens het persklaar maken van deze STAtOR ontvingen we het trieste bericht dat Jacqueline Bloemhof-Ruwaard, een van de auteurs van ons openingsartikel, overleed op 5 juni 2020. Een herinnering aan haar is direct na dat artikel geplaatst.



4



10



14



19



26



34

INHOUD

- 2 Redactioneel – Achterhaalde wetenschap?
- 4 Minder voedselverspilling door een dynamische houdbaarheid en korting | MARJOLEIN BUISMAN, RENE HAIJEMA & JACQUELINE BLOEMHOF-RUWAARD
- 9 IM Jacqueline Bloemhof-Ruwaard | JOAQUIM GROMICHO & CAROLINE JAGTENBERG
- 10 De nulhypothese-significantie toets; Tijd voor een paradigmaverschuiving | JONÁH STUNT & MICHEL DE BOER
- 14 Rekenen aan maatregelen om drukke treinen te ontlasten | MENNO DE BRUYN, SIMONE GRIFFIOEN, DENNIS HUISMAN, ANDRIKE MASTEBROEK & PLEUNI NAUS
- 19 Is de startprocedure bij schaatsen oneerlijk? | MIRIAM LOOIS
- 24 Smart lock down – column | JOHN POPPELAARS
- 26 From killer nurses to quantum entanglement, and back (Part 2) | RICHARD D. GILL
- 32 Russisch Ei Roulette – column | HENK TIJMS
- 33 VVSOR-vacatures | Young Statisticians
- 34 Statcheck: een spellingchecker voor statistiek | MICHÈLE B. NUIJTEN
- 37 Balgesnak in Coronatijd; Betaald voetbal kan spannender en sneller – column | GERARD SIERKSMA
- 39 Bekroning van uitzonderlijke afstudeerprestaties: Jan Hemelrijk Award 2019 en de Willem R. Van Zwet Award 2019
- 40 NS meet wel degelijk de reizigerspunctualiteit; Reactie op de column 'Weet wat je meet' | MERITH PELGER & DENNIS HUISMAN
- 41 Daniel Dadusch en Marloes Maathuis ontvingen de Van Dantzigprijs 2020
- 42 Gekken en dwazen – column | GERRIT STERMERDINK

MINDER VOEDSELVERSPLLING DOOR EEN DYNAMISCHE HOUDBAARHEIDSDATUM EN KORTING

De hoeveelheid bederfelijk voedsel die onverkocht door supermarkten wordt weggegooid kan verminderd worden door de houdbaarheidsdatum en de prijs van een product dynamisch aan te passen. In dit artikel bespreken we de mogelijkheden daartoe en bovenal welke besparingen er mogelijk lijken.

MARJOLEIN E. BUISMAN, RENE HAIJEMA & JACQUELINE M. BLOEMHOF-RUWAARD

De traditionele houdbaarheidsdatum die op producten geprint is, is veelal conservatief ingeschat door een producent, waardoor veel kwalitatief goede producten moeten worden weggegooid omdat ze niet meer verkocht mogen worden zodra de houdbaarheidsdatum is verstreken. Door op de laatste of voorlaatste dag korting te geven, proberen supermarkten ze toch nog te verkopen. Een dynamische houdbaarheidsdatum, afhankelijk van de actuele productkwaliteit zorgt voor minder afval bij de supermarkt en leidt vaak tot minder winstverlies door middel van kortingsstickers.

Voedselverspilling in de supermarkt

In de EU wordt jaarlijks bijna 5 miljoen ton voedsel ver-

spild bij supermarkten en distributiecentra (Stenmark et al. 2016). Idealiter zou het zo zijn dat er geen enkel product wordt verspild, maar dat is haast onmogelijk gegeven dat supermarkten veel producten willen aanbieden. Klanten willen een groot assortiment (van soortgelijke producten) om uit te kiezen, waardoor de vraag per product relatief klein is, en de vraagonzekerheid relatief groot. Daarnaast is de versheid van producten van belang voor klanten, waardoor supermarkten vrijwel dagelijks de schappen bijvullen ondanks dat oude producten nog niet op zijn. Veel consumenten laten producten die dicht bij de houdbaarheidsdatum komen links liggen.

Supermarkten passen wel methodes toe om verspilling zo laag mogelijk te houden, zoals het vooraan in het schap plaatsen van de oudste producten en het plakken van de welbekende kortingssticker op producten waarvan

de houdbaarheidsdatum bijna verloopt.

Intelligente verpakkingen bevatten sensoren die de daadwerkelijke kwaliteit van een product kunnen meten en op deze manier kunnen weergeven hoe lang een product nog daadwerkelijk houdbaar is. Dit is een meer accurate manier om de houdbaarheidsdatum weer te geven dan een geprinte datum op een verpakking. Onder andere Time-Temperature-Indicators (TTI) of RFID-tags zijn voorbeelden van zulke sensoren (Heising 2014). Deze verpakkingen en sensoren zijn uiteraard duurder dan traditionele verpakkingen en daarom is het van belang te weten hoe effectief het is om zo'n sensor te gebruiken in plaats van een vooraf vastgestelde houdbaarheidsdatum. En hoe effectief is het in vergelijking met het geven van korting op producten? Om die vragen te beantwoorden hebben we een simulatiestudie uitgevoerd voor een vleesproduct.

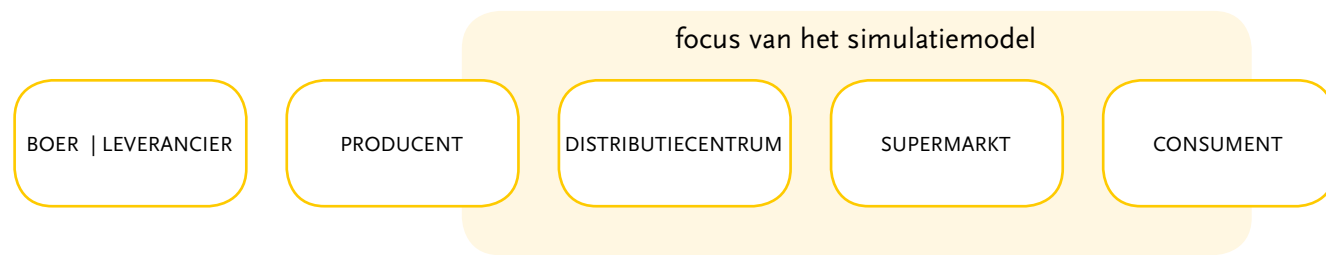
Simulatiemodel

Momenteel is een geprinte datum op een versproduct verplicht door de Europese Unie. Het is dus niet mogelijk de effectiviteit van zo'n sensor daadwerkelijk in de praktijk te testen. Om toch uit te kunnen zoeken hoeveel afval er te besparen valt, hebben we een simulatiemodel gemaakt. Figuur 1 laat zien welk deel van de vleesketen wordt gesimuleerd door het model. De simulatie begint op het moment dat het product de producent verlaat en naar het distributiecentrum (DC) van een supermarkt wordt vervoerd. In het DC wordt het product een korte tijd opgeslagen voordat het verder gaat naar de winkel. Het laatste deel van het model simuleert de verkoop van het product in de supermarkt.

Het centrale punt van het model is de supermarkt. Aan het begin van de dag, voordat de winkel opengaat,



Foto: Alexas



Figuur 1. De vleesketen en de focus van het simulatiemodel

plaatst de supermarkt een bestelling bij het DC gebaseerd op hoeveel producten de supermarkt nog in voorraad heeft. Deze bestelling wordt dan uiteindelijk geleverd als de supermarkt weer dicht is, maar voordat de volgende bestelling wordt gedaan. Gedurende de dag worden er producten verkocht aan de consument.

Effect van korting

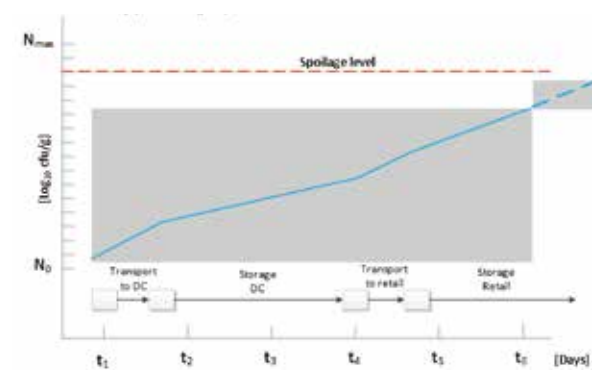
Consumenten kunnen een voorkeur hebben voor het meest verse/nieuwste product (LIFO) of ze kunnen het eerste product pakken dat ze zien, hetgeen veelal door actief schapbeheer het oudste product is met de kortste houdbaarheid (FIFO). Producten die over de datum zijn, worden aan het einde van de dag uit de schappen gehaald. Het maakt daarom nogal uit voor de supermarkt of er veel FIFO of veel LIFO-consumenten zijn. Hoe meer mensen de meest verse producten kopen, hoe meer producten er uiteindelijk over de datum gaan voor ze verkocht worden. Dit is precies de reden waarom de kortingsstickers gebruikt worden. Als de producten die kort voor hun houdbaarheidsdatum zitten goedkoper worden, dan zijn meer mensen geneigd deze producten te kopen in plaats van de duurdere producten, ook al zijn die nog langer houdbaar (Tsiros & Heilman, 2005).

Effect van dynamische houdbaarheidsdatum

Naast de korting op producten willen we ook analyseren wat het effect zou zijn van een dynamische houdbaarheidsdatum, waarbij je dus kan denken aan het vervangen van de houdbaarheidsdatum door een kwaliteits-sensor. Om dit in het model te kunnen doen, wordt er gebruik gemaakt van een microbiel groeimodel, het Gompertz-model (Bruckner, 2010). Dit model voorspelt de groei van de bacteriën die zorgen voor het bederf op

het product, op basis van het temperatuursverloop in de keten. Als het kiemgetal boven een grenswaarde komt, is het product bedorven (boven de rode lijn in figuur 2). Door in elke stap van het simulatiemodel de temperatuur mee te nemen, kan er dus precies gemodelleerd worden wanneer een product bedorven is en wanneer het nog goed genoeg is om te verkopen. Een voorbeeld hiervan is te zien in figuur 2, waar het kiemgetal wordt weergegeven per stap in de keten. In elke stap van de keten is er een specifieke temperatuur, die bepaalt hoe hard de bacteriën groeien. Hoe hoger de temperatuur, hoe sneller de groei.

Het verschil met een vaste houdbaarheidsdatum is dat er van tevoren dus niet bepaald is hoe lang het product houdbaar is. Bij een vaste houdbaarheidsdatum wordt er van tevoren, op basis van de verwachte temperatuurschommelingen in de keten, berekend hoelang het product houdbaar zal zijn, met inbegrip van een veiligheidsmarge. Bij een dynamische houdbaarheidsdatum wordt daarentegen gekeken naar de daadwerkelijke kwaliteit van het product.



Figuur 2. Kwaliteitsverloop (microbiële groei) in de keten

Analyse

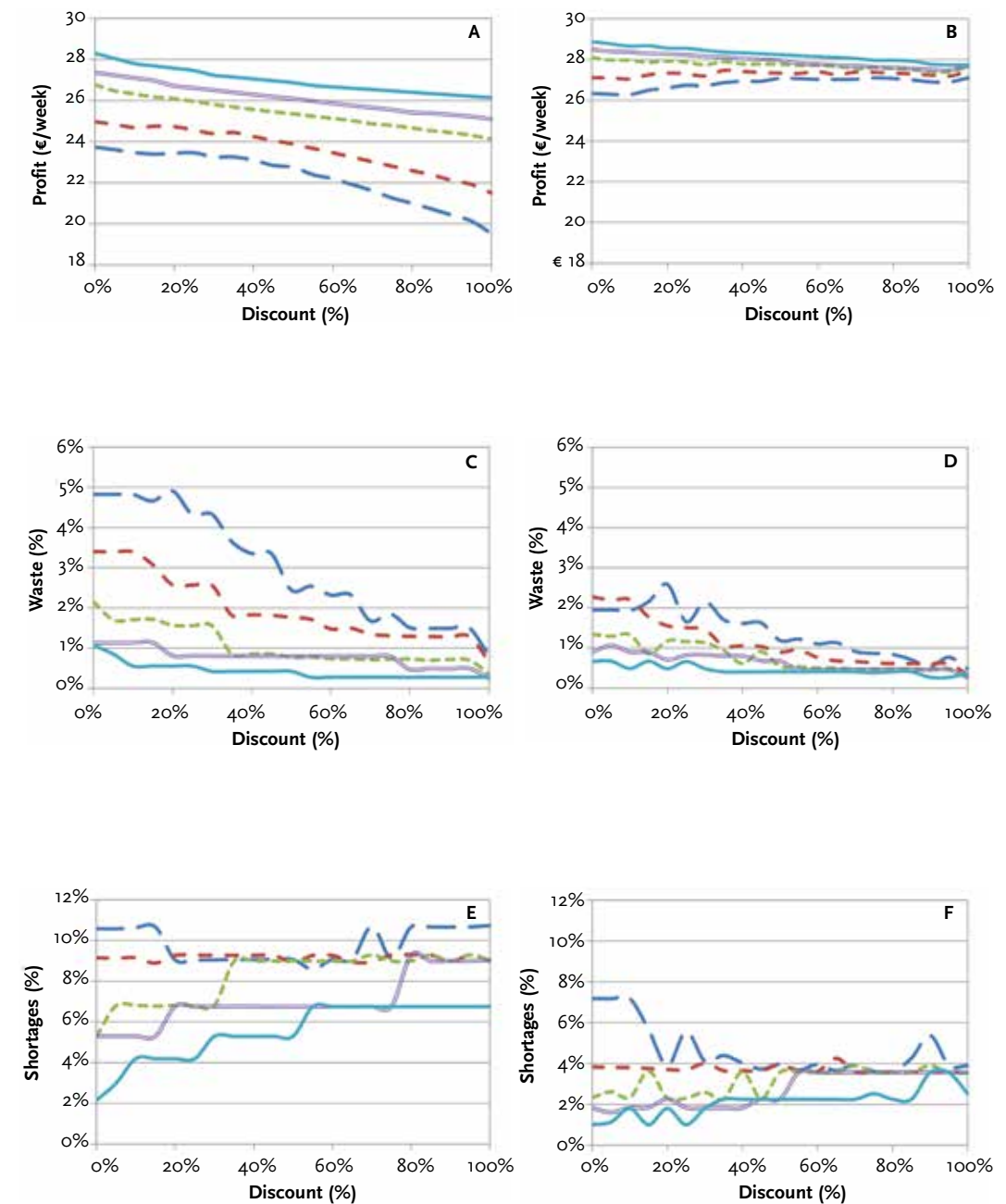
Naast dat we proberen uit te zoeken hoeveel producten er worden weggegooid bij diverse scenario's, zijn er ook andere aspecten belangrijk voor een supermarktmanager. Daarom analyseren we met het simulatiemodel niet alleen de hoeveelheid afval, maar kijken we ook naar de winst en de tekorten die bij de desbetreffende scenario's horen.

Resultaten

In figuur 3 zijn de verschillende resultaten van het model weergegeven, voor de producten met een vaste houdbaarheidsdatum (FSL) en voor de producten met een variabele houdbaarheidsdatum (DSL). Voor beide scenario's zijn de winst, de hoeveelheid afval en de tekorten berekend bij verschillende kortingspercentages. Wat duidelijk te zien is, is het verschil in afval bij vaste of variabele houdbaarheidsdatum. Ondanks dat in beide grafieken (C en D) een neergaande trend te vinden is ten opzichte van het kortingspercentage, is de initiële hoeveelheid producten die wordt weggegooid veel minder bij de variabele datum. Wanneer er geen korting wordt gegeven, ligt het weggooi-percentage voor de producten met een vaste houdbaarheidsdatum tussen de 1 en 5% (bij, respectievelijk, uitsluitend FIFO of uitsluitend LIFO-kanten). Daarentegen ligt het percentage afval bij het gebruik van een variabele datum lager, namelijk tussen de 2,2 en 0,7%.

Het effect van kieskeurige klanten

Een trend die te zien is voor beide houdbaarheidsdata is het verschil in weggooien bij verschillend koopgedrag van de consument. In grafiek C is duidelijk te zien dat er meer producten worden weggegooid wanneer de consument selectief is en altijd het meest verse product wil hebben. Wanneer er meer consumenten zijn die het voorste product in het schap kopen, daalt het afvalpercentage enorm.



Figuur 3. A. Winst FSL; B. Winst DSL; C. Waste levels FSL; D. Waste levels DSL; E. Tekorten FSL; F. Tekorten DSL

Hoe meer korting er wordt gegeven, hoe meer consumenten switchen naar het goedkopere en dus oudere product, en hoe minder producten er verspild worden. Deze trend is ook te zien in grafiek D, al is het minder extreem, aangezien de initiële afvalpercentages al lager zijn.

Winst en tekorten

Over het algemeen leidt het geven van korting, ook op producten die bijna over de datum zijn, tot minder winst. Zeker bij een vaste houdbaarheidsdatum is dit het geval, aangezien er een aanzienlijk deel van de producten wordt afgeprijsd. Bij een variabele houdbaarheidsdatum is het verlies een stuk minder groot, dit komt vooral omdat er in verhouding minder producten worden afgeprijsd. In sommige gevallen loopt de winst zelfs op wanneer er korting gegeven wordt, namelijk wanneer de meeste klanten in eerste instantie het meest verse product kozen.

Wat verder nog te zien is in figuur 3 (grafiek E en F), is dat de tekorten oplopen bij een vaste houdbaarheidsdatum wanneer er meer korting wordt gegeven. Het aantal tekorten en het afvalpercentage hangt vaak samen: hoe meer afval, hoe minder tekorten en vice versa.

Conclusie

Wat de resultaten ons leren, is dat een variabele houdbaarheidsdatum ervoor kan zorgen dat er minder producten worden weggegooid bij de supermarkt. Het grootste voordeel ligt in een iets langere verkooptijd doordat er minder veiligheidsmarge ingecalculeerd hoeft te worden. Deze veiligheidsmarge op producten met een vaste houdbaarheidsdatum is nodig om te zorgen dat producten die verkocht worden niet bedorven zijn binnen de aangegeven datum (Soethoudt et al., 2012). Maar vaak zijn producten zelfs nog na de houdbaarheidsdatum van goede kwaliteit. Als de productkwaliteit gemeten wordt, is er geen (of een veel kleinere) veiligheidsmarge nodig. Daardoor kunnen deze producten vaak een dag langer in de winkel liggen, wat de kans vergroot dat ze verkocht worden in plaats van weggegooid.

Verder wordt overduidelijk dat de klant een grote invloed heeft op de prestatie van de retailer. Misschien wel meer dan het gebruik van een vaste of dynamische houdbaarheidsdatum. Als klanten de oudste producten kopen in plaats van altijd het langst houdbare product te willen hebben, worden er een stuk minder producten weggegooid bij de supermarkt.

Voor meer informatie verwijzen we naar het artikel: ‘Discounting and dynamic shelf life to reduce fresh food waste at retailers’ (Buisman et al. 2019).

LITERATUUR

- Bruckner, S., Albrecht, A., Petersen, B., & Kreyenschmidt, J. (2013). A predictive shelf life model as a tool for the improvement of quality management in pork and poultry chains. *Food control*, 29(2), 451–460.
- Buisman, M. E., Haijema, R., & Bloemhof-Ruwaard, J. M., (2019). Discounting and dynamic shelf life to reduce fresh food waste at retailers. *International Journal of Production Economics*, 209. doi: 10.1016/j.ijpe.2017.07.016
- Heising, J. K., Dekker, M., Bartels, P. V., & Van Boekel, M. A. J. S. (2014). Monitoring the quality of perishable foods: opportunities for intelligent packaging. *Critical reviews in food science and nutrition*, 54(5), 645–654.
- Soethoudt, J. M., Van der Sluis, A. A., Waarts, Y. R., & Tromp, S. O. (2012). *Houdbaarheidsdatum, verspilde moeite?* (No. 1353). Wageningen UR-Food & Biobased Research.
- Stenmarck, A., Jensen, C., Quested, T., & Moates, G. (2016). FUSIONS Food waste data set for EU-28. European Commission, Stockholm, Sweden. Retrieved from: <https://www.eu-fusions.org/phocadownload/Publications/Estimates%20of%20European%20food%20waste%20levels.pdf>
- Tsiros, M., & Heilman, C. M. (2005). The effect of expiration dates and perceived risk on purchasing behavior in grocery store perishable categories. *Journal of marketing*, 69(2), 114–129.

MARJOLEIN BUISMAN studeerde Levensmiddelenmiddelen-technologie en Management, Economics and Consumer Studies aan de Wageningen Universiteit. Zij promoveerde in 2019 bij de leerstoelgroep Operations Research en Logistiek (ORL) op een onderzoek naar hergebruik en vermindering van afval met name bij voedselverspilling in supermarkten. Zij werkt sinds februari 2020 als universitair docent Retail Analytics aan de WHU Otto Beisheim School of Management. E-mail: marjolein.buisman@whu.edu

RENE HAIJEMA studeerde Operations Research & Management aan de Universiteit van Amsterdam. Hij promoveerde daar in 2009 bij Jan van der Wal en Nico van Dijk op onderzoek naar grote gestructureerde Markov Beslisvraagstukken, met toepassing in voorraadbeheer bij bloedbanken en het regelen van verkeerslichten. Sinds 2008 is hij universitair hoofddocent bij de leerstoelgroep ORL aan de Wageningen Universiteit. Hij specialiseert zich op voorraadbeheer van bederfelijke producten en andere veelal logistieke toepassingen van op Simulatie gebaseerde Optimalisatie, Markov beslistheorie, en Data Science technieken. E-mail: Rene.Haijema@wur.nl

JACQUELINE BLOEMHOF studeerde Operations Research en Econometrie aan de Erasmus Universiteit Rotterdam. In 1996 promoveerde zij aan de Wageningen Universiteit op een onderzoek naar de integratie tussen Operations Research en Milieumanagement. Ze werkte als universitair docent aan de Rotterdam School of Management en bij de leerstoelgroep ORL aan de Wageningen Universiteit als universitair hoofddocent, hoogleraar en leerstoelhouder. Sinds 2019 is ze met emeritaat.

IN MEMORIAM

Jacqueline Bloemhof-Ruwaard

30 mei 1969 – 5 juni 2020

Tot onze spijt geven wij kennis van het overlijden van prof. dr. Jacqueline Bloemhof-Ruwaard. Het onderzoek van Jacqueline stond in het teken van *sustainable supply chain management* en daarmee maakte ze internationaal naam. We blikken terug op haar leven en carrière.

Toen Jacqueline ging promoveren was haar domein dusdanig nieuw dat ze maar liefst drie promotoren kreeg. Als een van de eersten wilde ze haar OR-kennis toepassen om onze slechte impact op het milieu te verminderen. Een slimme dame met een nederig ideaal: helpen de aarde te redden. In de jaren daarna bleef Jacqueline verrassen. Ze koos er voor haar rol als echtgenote en moeder op de voorgrond te zetten en zette haar academische carrière tijdelijk op een lager pitje. Toch is dat aan haar impact niet te zien: meer dan tienduizend citaties, waarvan uit één paper bijna drieduizend, resulterend in een h-index van 40 en een i10-index van 73. Het is wel

duidelijk dat we het hier over een talentvolle vrouw hebben die het ver ging schoppen. Niet voor niets werd ze in 2014 hoogleraar, en leidde ze vanaf 2016 haar eigen vakgroep aan de Wageningen University & Research. Jacqueline was zeer betrokken bij de grote kring mensen om haar heen. Ter illustratie delen twee STAtOR-redactieleiden hieronder hun herinneringen.

In onze herinneringen blijft Jacqueline leven, en haar werk leeft voort in de strijd om moeder aarde te redden. We vinden het heel bijzonder in deze STAtOR een van Jacqueline’s laatste artikelen te mogen publiceren. En voor iedereen die OR wil inzetten voor een duurzamere wereld, raden wij aan Jacqueline’s advies ter harte te nemen: ‘Keep dreaming. Think big, start small, act now.’.*

* Bloemhof-Ruwaard, J. (2019). *Closing the loop: A never ending story*. Wageningen: Wageningen University & Research. doi.org/10.18174/497936

Herinneringen van Joaquim Gromicho

Jacqueline is een van de eerste Nederlanders die ik heb leren kennen. In januari 1991 kwam ik naar Nederland om promotieonderzoek te doen aan de Erasmus Universiteit Rotterdam waar ook Jacqueline studeerde. Ik zal nooit de dag vergeten waarop Jacqueline mij vroeg ‘jij mist vast de zee, kom ik laat je de Nederlandse zee zien’. We namen de trein naar Hoek van Holland. En ze had gelijk: het was voor mij een openbaring dat Nederland ook mooie zandstranden heeft! Dankzij Jacqueline werd mijn heimwee naar Portugal een stuk dragelijker. Duidelijk een mensenkenner. Maar ook een briljante geest! Al snel begon ze in Wageningen aan haar promotieonderzoek naar de integratie tussen Operations Research en Milieumanagement, dat ze in 1996 succesvol afrondde. Ze bleef wel de Erasmus Universiteit trouw bezoeken, in het bijzonder de vakgroep van Jo van Nunen.

Het bijwonen van haar afscheidsrede als hoogleraar aan de WUR was voor mij een bijzonder moment. Met een groot trots gevoel luisterde ik naar haar rustige verslag van het gedane werk; duidelijk was dat ze er graag langer mee bezig zou willen zijn. Ik kon haar nog even spreken en we hadden het erover hoe fijn het voor haar was dat ze bij Katwijk aan Zee was gaan wonen. De zee... altijd zo belangrijk voor haar.

Herinneringen van Caroline Jagtenberg

In 2017 had ik het geluk bij Jacqueline aan tafel te schuiven, op de avond dat de Jo van Nunenprijs (uitgereikt door de VLM voor het beste logistieke proefschrift) uitgereikt zou worden. Ik was een van de genomineerden, Jacqueline zat in de jury. Ik gooide mijn eigen ruiten in – althans dat dacht ik – door te zeggen dat de toepassing van logistiek van een andere deelnemer mooier was dan de mijne, diegene deed namelijk onderzoek naar sustainability. ‘Dat is mijn onderwerp!’ zei ze, en ik geloof niet dat ik ooit iemand zó heb zien stralen. Jacqueline was een voorbeeld en ze inspireerde me, niet alleen omdat zij als vrouw een hoogleraarspositie had weten te bemachtigen, maar ook vanwege de mooie onderwerpen die ze voor haar onderzoek koos. Toen al was duidelijk dat haar gezondheid niet mee zat; ze hield het niet de hele avond vol. Wat jammer, dacht ik, met deze vrouw had ik nog wel langer willen praten.

In september 2019 trad ze af bij de WUR. Helaas kon ik haar afscheidscollege niet bijwonen omdat ik aan de andere kant van de wereld verbleef. Ik was blij haar afscheidsrede online te kunnen lezen, en ik zal eerlijk zeggen: bij het dankwoord hield ik het niet droog! Ik zat toen op mijn werkplek ... en zodoende weten nu ook de collega’s van de University of Auckland wie Jacqueline is.



Illustratie: Mohamed Hassan

DE NULHYPOTHESE-SIGNIFICANTIE TOETS tijd voor een paradigmaverschuiving

De nulhypothese-significantie toets (NHST) is binnen de empirische wetenschappen de standaard procedure om conclusies te trekken over bevindingen. Ondanks het feit dat deze methode vele beperkingen kent, is het gebruik ervan alomtegenwoordig. Om te begrijpen waarom dit zo is, hebben wij onderzocht wat de visies zijn van stakeholders binnen het wetenschappelijke systeem op het gebruik van NHST, en ook alternatieve methoden voor NHST. Verschillende, met elkaar samenhangende factoren blijken het gebruik van NHST in stand te houden. Samen met eerdergenoemde stakeholders hebben wij op basis van deze factoren strategieën ontwikkeld om het gebruik van NHST te verminderen en het gebruik van alternatieve methoden te motiveren.

JONÁH STUNT & MICHEL DE BOER

De nulhypothese toets en de p-waarde

Het is de vraag waar iedere empirische wetenschapper mee worstelt: hoe bepalen we op een objectieve manier of de gegevens die we hebben verzameld berusten op

toeval of niet? In de jaren twintig van de vorige eeuw boden enkele vooraanstaande wetenschappers soelaas: Jerzy Neyman en Egon Pearson formuleerden de hypothesetoets; Ronald Fisher ontwikkelde de theorie achter de p-waarde. Neyman en Pearson poneerden een nul-

hypothese waarbij er sprake is van geen verschil tussen twee groepen of geen associatie tussen een determinant en een uitkomst (de nulhypothese; H_0). Ze waren van mening dat je niet uit kunt gaan van een nulhypothese als daar niet een plausibele alternatieve hypothese (H_1) tegenover staat die stelt dat er wél sprake is van een verschil, effect of associatie¹. Fisher nam aan dat de kans om een effect te verkrijgen gelijk aan of extremer dan het geobserveerde effect, *gegeven de nulhypothese*, bepaald kan worden met een p-waarde. Wanneer die kans kleiner is dan 5% beschouwde Fisher de kans dermate klein dat de nulhypothese buitengewoon onwaarschijnlijk wordt. Verdere onderzoeken en experimenten achtte hij dan wenselijk². De twee theorieën boden onderzoekers handvatten om hun hypothese te accepteren dan wel te verwerpen. In de wetenschap werd het de regel uit te gaan van een *cut-off* p-waarde van 0,05, waarbij de nulhypothese wordt verworpen en de alternatieve hypothese wordt aangenomen als de p-waarde kleiner is dan 0,05: de nulhypothese-significantie toets (NHST).

Het gebruik van de nulhypothese toets op basis van de p-waarde is het heersende paradigma in de empirische wetenschap geworden. Maar meer dan dat: het bereiken van de 0,05-grens voor statistische significantie is voor veel onderzoekers een doel op zich geworden. Het lijkt in de empirische wetenschappen dan ook te draaien om statistische significantie. De afgelopen decennia is de methode – alhoewel populair – op veel kritiek gestuit. NHST kent vele beperkingen³⁻¹¹, die tot gevolg hebben dat conclusies over onderzoeksgegevens vaak onjuist zijn¹⁰. Het roept de vraag op of NHST nog wel de standaard statistische procedure zou moeten zijn. Tot voor kort beperkte het debat rondom het gebruik van NHST en alternatieve methoden voor NHST zich voornamelijk tot statistici en methodologen. Een groeiend aantal wetenschappers is zich echter gaan mengen in de discussie, en is van mening dat een paradigmaverschuiving van NHST af wenselijk of zelfs noodzakelijk is.

De vraag die hier logischerwijs uit voortvloeit is waarom zoveel onderzoekers ondanks alle bezwaren NHST blijven gebruiken. Om de desgewenste paradigmaverschuiving binnen de wetenschappelijke gemeenschap te bewerkstelligen, is het noodzakelijk om hier een antwoord op te vinden. Vanuit ons onderzoek hebben wij

daarom gesproken met verschillende *stakeholders* van het wetenschappelijk systeem (onderzoekers, statistiekdocenten, onderzoeksfinanciers en redacteurs van wetenschappelijke tijdschriften), en hen gevraagd naar hun visie op het gebruik van NHST of van alternatieven voor deze methode. Op basis hiervan hebben we in kaart gebracht welke factoren van invloed zijn op het alomtegenwoordige gebruik van NHST.

Samenspel van factoren

De uitkomsten van ons onderzoek laten zien dat verschillende, met elkaar samenhangende factoren het gebruik van NHST in stand houden. Een van de belangrijkste daarvan bleek het huidige wetenschappelijke klimaat, waarbinnen conventies, gedragsnormen en overtuigingen gelden waarvan het gebruik van NHST en p-waardes nu eenmaal onderdeel is. Of anders gezegd: men weet niet beter.

Daarbij geldt ook dat er gestreefd wordt naar een significant resultaat, omdat dat volgens de respondenten een bewijs voor succesvol onderzoek zou zijn. Mogelijke consequenties voor publicatiekansen en het gevoel afhankelijk te zijn van leidinggevend en/of subsidieverstrekkingen zorgen ervoor dat onderzoekers niet geneigd zijn af te wijken van wat de normen voorschrijven. In het verlengde hiervan staan publicatiedruk, een te hoge werkdruk en de manier waarop het wetenschappelijke beloningssysteem is ingericht in de weg van gedragsverandering en het nemen van initiatieven.

Een andere factor die het gebruik van NHST in stand houdt, is het gevoel niet te kunnen voldoen aan de wetenschappelijke plicht. Men wil zowel inhoudelijk als moreel het juiste doen, maar heeft het gevoel hiervoor kennis en inzicht, vaardigheden, tijd, capaciteit en *empowerment* te missen. Velen ervaren een intern conflict: een discrepantie tussen wat men *vindt* of *wil* en wat men *doet*.

De beschreven factoren hebben tot gevolg dat het gros van de mensen die deel uitmaken van het wetenschappelijke systeem een reactieve houding heeft ingenomen. Men is weinig geneigd zelf verantwoordelijkheid te nemen en wacht actie en initiatief van anderen af. Dit verklaart waarom NHST voor bijna alle betrokkenen nog altijd als standaard geldt en de enige procedure is die gebruikt wordt.

Strategieën voor een paradigmaverschuiving

Een paradigmaverschuiving vraagt om het doorbreken van reactiviteit. Hiervoor zijn concrete maatregelen nodig. In een bijeenkomst met vertegenwoordigers van eerdergenoemde stakeholdergroepen hebben we ons daarom gericht op het ontwikkelen van strategieën om het standaardgebruik van NHST te verminderen en gebruik van alternatieve methoden te motiveren. In dit kader is het lastig dat er, alhoewel men overeenstemming vindt met betrekking tot de beperkingen van NHST, geen consensus bestaat over de juiste of beste alternatieven (zie kader). In de bijeenkomst werd echter door de meerderheid van de deelnemers het belang onderschreven dat men zich ervan bewust is dat het probleem breder is dan een gouden alternatief voor NHST. Meer dan het leren toepassen van een alternatief is het noodzakelijk kennis over en inzicht in statistische methoden te vergaren. Te weinig kennis brengt in alle gevallen risico tot verkeerd gebruik of zelfs misbruik met zich mee. Het is vooral essentieel te stoppen met het toepassen van statistische ‘trucjes’. Voorgestelde strategieën betreffen daarom onder andere het tackelen van barrières voor publicatie zonder NHST, het publiceren van tutorials over het rapporteren zonder NHST en de p-waarde, het gezamenlijk opstellen van *best practices* voor analyseplannen (zoals het belonen van transparantie, *open science*, preregistratie, etc.) en het aanmoedigen van het verbeteren van onderwijsprogramma’s Methodologie en Statistiek voor onderzoekers, docenten en studenten.

Uit ons onderzoek is verder naar voren gekomen dat er voor een succesvolle implementatie van de strategieën rekening moet worden gehouden met de volgende aspecten:

- het is essentieel de gehele wetenschappelijke gemeenschap te betrekken, niet alleen de onderzoekers;
- verandering binnen een systeem en initiatieven daar toe moeten verschillende startpunten hebben, dat wil zeggen: zowel bottom-up als top-down, zowel vanuit onderwijs als onderzoek, vanuit wetenschappelijke tijdschriften, maar ook vanuit beleid van fondsen en geldschietters;
- verandering wordt het best geaccepteerd als het plaatsvindt in kleine stappen;
- verandering moet op een juiste manier gecommuniceerd worden;
- het moet geaccepteerd worden dat onzekerheid besloten ligt in de statistiek in engere zin, en in de wetenschap in het algemeen.

Alternatieven voor NHST en de p-waarde:

In de afgelopen decennia zijn verschillende alternatieven voor NHST ontwikkeld. De meest bekende hiervan is het gebruik van Bayesiaanse statistiek²⁰ of een van de elementen van Bayesiaanse statistiek; de (minimum) Bayes factor^{7,21}. Bayesiaanse statistiek focust op de a priori waarschijnlijkheid van een verondersteld effect, waarmee de methode het probleem van indirecte inferentie ondervangt. Andere alternatieven die zijn aangedragen zijn het gebruik van informatieve hypothesetoetsing^{22, 23} en a priori inferentiële statistiek die gebruik maakt van zogenoemde *coefficient of confidence*²⁴. In het in 2019 verschenen nummer ‘Statistical Inference in the 21st Century: A World Beyond $p < 0,05$ ’ van *The American Statistician* worden veel suggesties voor alternatieven gedaan, zoals het gebruik van s-waardes en False Positive Risks, het rapporteren van de uitkomst als *effect sizes* met *confidence* intervallen en meer holistische benaderingen voor p-waardes en uitkomstmaten¹⁴. Hoewel er tal van alternatieve opties geopperd en beschreven zijn is het gebruik van alternatieven is tot nu toe zeer beperkt.

A world beyond $p < 0,05$

Bovengenoemde strategieën staan in nauwe relatie van een ontwikkeling die pas sinds kort post is gaan vatten: in 2015 besloten de redacteurs van het tijdschrift *Basic and Applied Social Psychology* (BASP) het gebruik van NHST (en p-waardes en betrouwbaarheidsintervallen) te verbieden¹². Vervolgens publiceerde de *American Statistical Association* in 2016 de zogenoemde ‘Statement on p-values’ in *The American Statistician*¹³. Dit statement bestaat

uit kritische standpunten ten aanzien van het gebruik van NHST en p-waardes en uit waarschuwingen voor het misbruik van de procedure. In 2019 wijdde *The American Statistician* een geheel nummer aan het verder doorvoeren van de hervormingen rondom het gebruik van NHST: in meer dan veertig artikelen spraken wetenschappers zich uit tegen statistische significantie en voor het omarmen van onzekerheid¹⁴. Daarnaast verschenen er in 2019 meerdere artikelen waarin – gesteund door vele ondertekenaars – een oproep werd gedaan om afstand te nemen van statistische significantie^{15,16}. Hierop is ook een aantal tegenreacties gepubliceerd^{17, 18, 19}.

Bovengenoemde publicaties hebben een constructieve bijdrage aan het NHST-debat geleverd: het aandragen van concrete maatregelen en alternatieven is bevorderlijk voor het veranderproces. Met ons project en onze concrete voorstellen tot verandering willen ook wij een betekenisvolle rol spelen. Een nieuw paradigma bewerkstelligen is voor de korte termijn geen haalbare doelstelling, maar wij ambiëren met onze geformuleerde strategieën de recente ontwikkeling te stimuleren en het veelvuldig gebruik en de sterke afhankelijkheid van NHST en de p-waarde binnen de wetenschappelijke gemeenschap te verminderen.

LITERATUUR

1. Neyman, J. & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A*, 231, 289–337.
2. Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. *Seminars in Hematology*, 45, 135–140.
3. Nickerson, R. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, 5(2), 241–301.
4. Szucs, D. & Ioannidis, J. (2017). When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. *Frontiers in Human Neuroscience*, 11(3), 390.
5. Trafimow, D., & Rice, S. (2009). A Test of the Null Hypothesis Significance Testing Procedure Correlation Argument. *The Journal of General Psychology*, 136(3), 261–269.
6. Greenland, S, Senn, S.J., Rothman, K.J., et al. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31, 337–350.
7. Goodman, S.N., & Hopkins, J. (1999). Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. *Annals of Internal Medicine*, 130, 995–1004.
8. Nuzzo R. (2014). Statistical errors: P values, the “gold standard” of statistical validity, are not as reliable as many scientists assume. *Nature*, 506(7487), 150–152.
9. Van de Schoot, R, Verhoeven, M, & Hoijtink, H. (2013). Bayesian evaluation of informative hypotheses in SEM using Mplus: A black bear story. *European Journal of Developmental Psychology*, 10(1), 81–98.

10. Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLOS Medicine*, 8(2), e124.
11. Goodman, S.N., & Hopkins, J. (1999). Toward Evidence-Based Medical Statistics. 2: The Bayes Factor. *Annals of Internal Medicine*, 130, 1005–1013.
12. Trafimow D, & Marks M. (2015). Editorial. *Basic and Applied Social Psychology*, 37(1), 1–2.
13. <https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf>.
14. Statistical Inference in the 21st Century: A world beyond $p < 0.05$. (2019). *American Statistical Association*, 73, S1.
15. Wasserstein, R.L., Schirm, A. & Lazar, N.A. (2019). Editorial. Moving to a World Beyond “ $p < 0.05$ ”. *American Statistical Association*, 73, S1, 1–19.
16. Amrhein, V., Greenland, S., & McShane, B. (2019). Retire statistical significance. *Nature*, 567(7748), 305–307.
17. Ioannidis, J.P.A. (2019). The Importance of Predefined Rules and Prespecified Statistical Analyses. Do Not Abandon Significance. *Journal of the American Medical Association*, 321(21). 2067–2068.
18. Ulrich, D. (2019). The p-value wars (again). *EJNMMI*, 46, 2421–2423.
19. Haig, B. (2019). Stats: retain significance testing. *Nature*, 569(7757), 487.
20. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). *Bayesian Data Analysis*. 3rd ed. Boca Raton: Taylor and Francis Group.
21. Kass, R.E., & Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90, 773–795.
22. Hoijtink, H., Klugkist, I., & Boelen, P., eds. (2008). *Bayesian Evaluation of Informative Hypotheses*. Springer Science & Business Media.
23. Van de Schoot, R., Hoijtink, H., Romeijn, J.W. (2011). Moving beyond traditional null hypothesis testing: Evaluating expectations directly. *Frontiers in Psychology*, 2, 1–5.
24. Trafimow, D. (2017). Using the Coefficient of Confidence to Make the Philosophical Switch From A Posteriori to A priori Statistics. *Educational and Psychological Measurement*, 77(4), 831–854.

MICHEL DE BOER is epidemioloog en universitair hoofddocent en opleidingsdirecteur van de Masteropleiding Health Sciences van de Vrije Universiteit Amsterdam. Hij geeft ruim 15 jaar onderwijs in zowel bachelor- als master- en postacademische programma’s en richt zich zowel in zijn onderzoek als onderwijs op het gebruik en misbruik van onderzoeksmethoden en statistiek. m.r.de.boer@umcg.nl
E-mail: m.r.de.boer@vu.nl

JONAH STUNT is epidemioloog en als postdoctoraal onderzoeker en docent werkzaam bij de sectie Methodologie en Toegepaste Biostatistiek, afdeling Gezondheidswetenschappen, VU, Amsterdam. Ze houdt zich hoofdzakelijk bezig met het gebruik en misbruik van nulhypothese-significantie toetsen en de p-waarde. Daarnaast werkt ze aan het herontwikkelen en innoveren van de leerlijn Methodologie en Statistiek voor de opleiding Gezondheidswetenschappen.
E-mail: j.j.stunt@vu.nl



REKENEN AAN MAATREGELN OM DRUKKE TREINEN TE ONTLASTEN

De frequente treinreiziger zal herkennen dat in de spits een zitplaats niet altijd gegarandeerd is. In 2019 kon in totaal ongeveer 95% van de spitsreizigers gedurende hun hele reis zitten. Naar verwachting kiezen de komende jaren steeds meer Nederlanders voor de trein, ook in de nu al drukke spitsuren. Al deze reizigers willen het liefste een zitplaats. Voor NS vormt dit dan ook één van de grote uitdagingen voor de komende jaren. Naast het inzetten van meer materieel (aanbod), kan de NS de vraagkant beïnvloeden. In dit artikel beschrijven we een methodiek die NS ontwikkeld heeft om verschillende maatregelen van vraagbeïnvloeding door te kunnen rekenen. We kijken daarbij naar het effect op de zitplaatskans van de reiziger en de materieelkosten van NS (en daarmee indirect de prijs van het treinkaartje). Dit artikel is geschreven voor de uitbraak van het coronavirus. De ontwikkelde methodiek en verkregen inzichten zijn echter ook in de huidige situatie nuttig en toepasbaar, bijvoorbeeld als het gaat over maximale treinbezetting.

MENNO DE BRUYN, SIMONE GRIFFIOEN, DENNIS HUISMAN, ANDRIKE MASTEBROEK & PLEUNI NAUS

Op een doordeweekse dag zijn er anno 2020 ongeveer 1,3 miljoen treinreizen. Als we bijvoorbeeld kijken naar de ochtendspits (6:30-9:00) dan fluctueert ook daarin de reizigersvraag nog flink. Rond 8 uur piekt de reizigersvraag sterk. Dit wordt ook wel de hyperspits genoemd, zie ook figuur 1, waarin de treinreizen en autoreizen op een doordeweekse ochtend getoond worden.

De benodigde hoeveelheid ingezet materieel wordt bepaald door de piek van de reizigersvraag. Materieelkosten zijn de grootste kostenpost voor een spoorbedrijf als NS. Hierin wordt een afweging gemaakt tussen de materieelkosten en de zitplaatskans voor de reiziger. Door de hyperspits af te vlakken en een klein deel van de reizigers een trein eerder of later te laten nemen, kan NS een hogere zitplaatskans bieden bij dezelfde hoeveelheid materieel of kan NS dezelfde zitplaatskans bieden met minder materieel.

In dit artikel bespreken we een aantal maatregelen

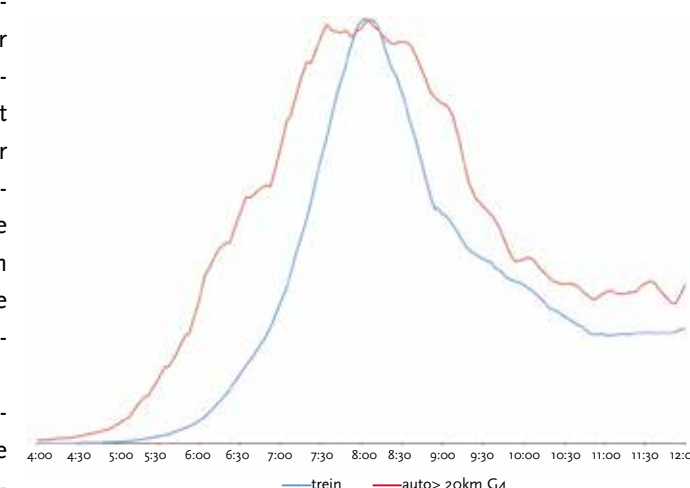
van vraagbeïnvloeding, en beschrijven we een methodiek om de effecten van deze maatregelen door te rekenen. We eindigen dit artikel met enkele conclusies die uit ons onderzoek kunnen worden getrokken.

Maatregelen om de vraag te beïnvloeden

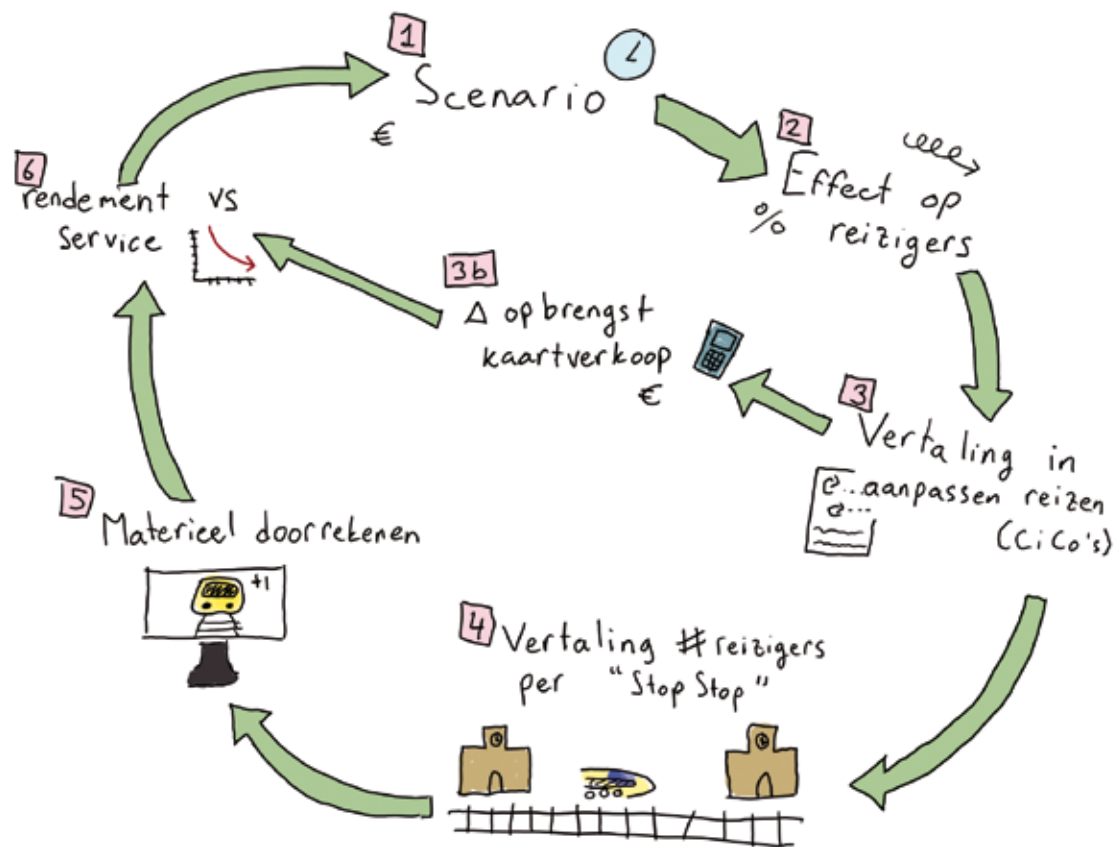
NS past verschillende vormen van vraagbeïnvloeding toe. Deze maatregelen kunnen in vier categorieën worden onderverdeeld. Dat is ten eerste informatie over drukte (traject- en treinspecifiek) ontsluiten voor de reizigers. Zo kan een reiziger zijn reiskeuze afstemmen op de verwachte drukte. Voorbeelden zijn de drukte-informatie die in de reisplanner getoond wordt, of een waarschuwing indien de trein korter is dan gebruikelijk en drukte verwacht wordt. De tweede categorie maatregelen betreft het toepassen van (niet financiële) *incentives* waarbij de reiziger beloond

wordt voor het reizen op rustigere tijdstippen. Voorbeeld hiervan is het voormalige MyOV programma waarbij punten gespaard konden worden voor het reizen in minder drukke (spits)treinen. Ten derde zijn er financiële incentives in de vorm van beprijzen. In de tarieven van NS wordt nu al het reizen op rustigere tijdstippen gestimuleerd door een korting in de daluren. Daarnaast zijn goedkope actiekaarten enkel geldig in de daluren. De vierde categorie ten slotte betreft de samenwerking met lokale partijen en overheden. Zo is in Nijmegen besloten de colleges van de Radboud Universiteit en van de Hogeschool Arnhem Nijmegen niet meer gelijktijdig te laten starten.

De maatregelen verschillen in de mate waarin ze reizigers daadwerkelijk bewegen op een ander tijdstip te reizen en ook in de mogelijkheid om de beweging te kunnen meten. Zo is het verschuifeffect van informeren over drukte in de praktijk zeer moeilijk meetbaar, maar drukte-informatie wordt door reizigers wel zeer gewaardeerd.



Figuur 1. De piek in de ochtendspits voor zowel trein als auto (bij reizen van meer dan 20 km naar de 4 grote steden). Bron: NS/Strategy Development Partners, intern rapport NS, 2016



Figuur 2. Schematische weergave van onze aanpak

Doorrekenen van maatregelen

De methodiek die we hebben gevolgd om de kwantitatieve effecten van verschillende maatregelen te bepalen is schematisch weergegeven in figuur 2. We volgen hierbij de volgende stappen.

1. We definiëren de maatregel.
2. We bepalen het effect op het reisgedrag van verschillende groepen reizigers.
3. We vertalen dit reisgedrag naar aantallen reizigers van ieder station naar ieder station op verschillende momenten van de dag (intervallen van 15 minuten). Tevens berekenen we het verschil in de opbrengsten van kaartverkoop.
4. We bepalen de verwachte reizigersaantallen per trein en traject, op basis van het reisgedrag.
5. Met deze reizigersaantallen berekenen we het effect op het materieelplan en daarmee op kosten en zitplaatskansen.
6. We combineren het verschil in opbrengsten en het effect op materieelkosten en zitplaatskansen om het totale effect van de maatregel te bepalen.

We gaan gedetailleerd in op de stappen 2 t/m 5.

Stap 2 t/m 4: reizigersgedrag modelleren en doorrekenen, uitkomst: reizigersaantallen per trein en traject, opbrengsten van de maatregel

De vraag die we willen beantwoorden is: gaan reizigers door een maatregel andere keuzes maken dan de huidige prognosemodellen voorspellen: niet meer met de trein reizen of op een ander tijdstip? En wat betekent voor het aantal reizigers in de trein? We kennen het reisgedrag bij de huidige prijsstructuren en de prijselasticiteiten bij "normale" prijswijzigingen. Om het effect te kunnen inschatten van het reisgedrag bij prijsniveaus en prijsstructuren die wezenlijk anders zijn dan we tot nu toe in de praktijk ervaren hebben, kunnen we maar deels varen op realisatie data uit het verleden.

Voor de inschatting van de verschuivingsbereidheid hebben we daarom gebruik gemaakt van een stated choice onderzoek uitgevoerd in 2017 (Van Daalen et al., 2018), waarbij ochtendspitsreizigers keuzes kregen voorgesteld voor verschillende vertrektijdstippen en tarieven. Met een discreet keuze model (Ben Akiva & Lerman, 1985) is vervolgens voor alle respondenten het nut van eerder of later vertrekken bij een gewijzigde tariefstelling

berekend. De respondent kiest dan steeds het alternatief dat het hoogste nut heeft.

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}$$

P_{ni} = kans dat respondent n vertrektijd i kiest

V_{ni} = kans nut voor respondent n van alternatief i ;

De alternatieven bestaan uit reizen met een andere vervoerwijze (bijv. auto of bus) of blijven reizen met de trein maar op een ander tijdstip met bijbehorend tarief.

Uit de chipkaartdata weet NS in grote mate van detail welke reizen er worden gemaakt: vertrekstation, vertrektijdstip, aankomststation en aankomsttijdstip worden voor reizen met een ov-chipkaart geregistreerd, overigens zonder daarbij gebruik te maken van persoonsgegevens. Omdat reizigers naast de ov-chipkaart ook met andere vervoersbewijzen kunnen reizen, moeten we deze data aanvullen om tot de totale reizigerspopulatie te komen. Dit doen we door gebruik te maken van andere databronnen zoals losse kaartverkoop en conducteursscans in de trein.

Op basis van deze informatie kunnen we de meest waarschijnlijke treinen waar de reiziger in heeft gezeten bepalen. Dit doen we met behulp van een model dat voor elke combinatie van vertrek- en aankomststation en vertrekmoment bepaalt wat de verschillende reismogelijkheden met bijbehorende kansen zijn (bijv. 30% kans de Sprinter die snel vertrekt en met 70% kans de Intercity die later vertrekt, maar er wel korter over doet).

Hoe wordt het effect van een vraagbeïnvloedingsmaatregel geschat? Voor tariefwijzigingen hebben we op basis van het voornoemde *stated choice* onderzoek de effecten op reizigersgedrag ingeschat. Voor bijvoorbeeld het verschuiven van college tijden hebben we een inschatting gemaakt op basis van inzichten uit de proef bij Nijmegen Heyendaal (Van der Aa & Hodde, 2018). Deze effecten passen we willekeurig toe op de totale reizigerspopulatie, dit door de data willekeurig te sorteren. Dit doen we om ervoor te zorgen dat niet het reisgedrag van alleen bepaalde reizigers (bijv. op specifieke trajecten) wordt aangepast. We maken daarnaast de aanname dat enkel spitsreizigers hun gedrag aanpassen bij het invoeren van maatregelen om de spits te beïnvloeden, terwijl het reisgedrag van dalreizigers gelijk blijft.

Het effect op de opbrengsten van kaartverkoop wordt berekend door het nieuwe aantal reizigers per tijdstip en kaartsoort te vermenigvuldigen met de bijbehorende

prijs. We nemen aan dat reizigers niet van kaartsoort zullen wisselen vanwege de maatregel.

Stap 5: materieelplan modelleren en doorrekenen, uitkomst kosten en zitplaatskansen

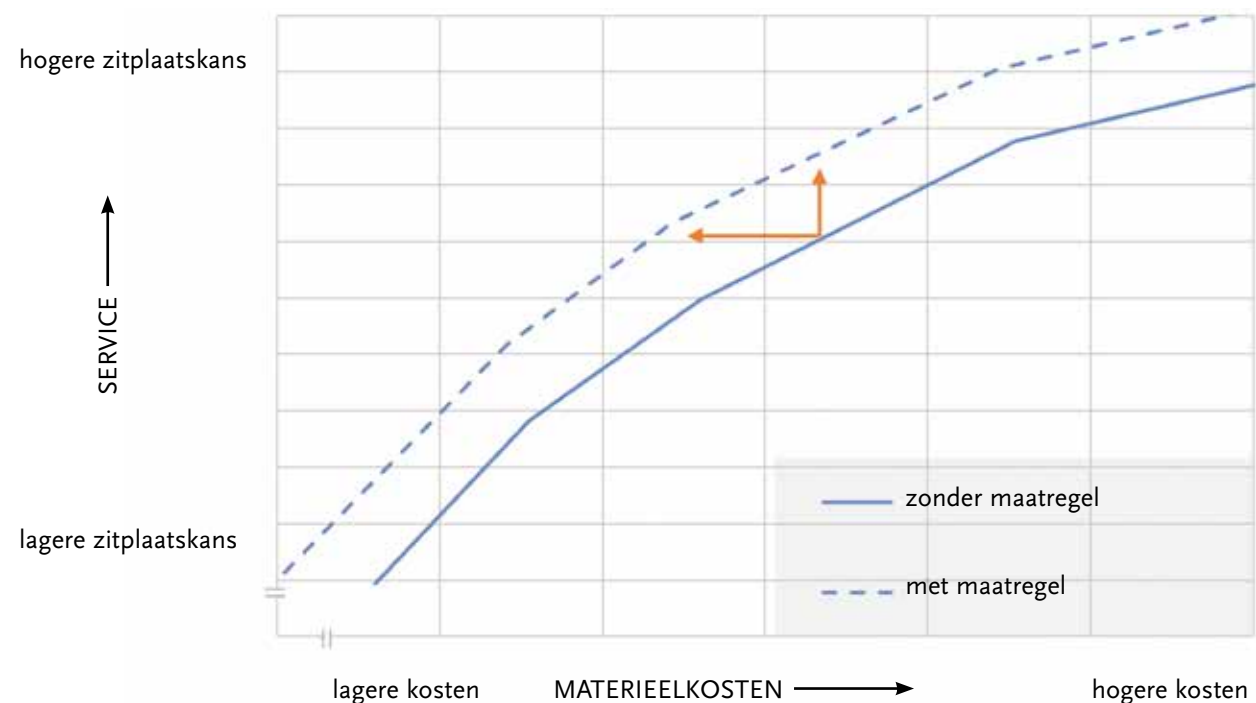
NS gebruikt de optimalisatie-software TAM om een operationeel materieelplan te maken. TAM bevat een getalig lineair programmeringsmodel, waarin voor elke trein op elk traject bepaald wordt met welke materieelsamenstelling gereden moet worden. Dit model bevat een aantal restricties. Een voorbeeld van zo'n restrictie is de treinlengte, die niet langer mag zijn dan het kortste peron waar deze stopt. De planner kan met gewichten aangeven hoe zwaar verschillende onderdelen zoals kosten (kilometerkosten) en service (zitplaatskansen) meewegen in de doelfunctie. Voor een gedetailleerde beschrijving van de achterliggende optimalisatiemodellen verwijzen we naar Fioole et al. (2006) en Nielsen (2011).

Om het effect van een maatregel door te rekenen hebben wij het model enigszins aangepast, om niet alleen de kosten op de korte, maar ook op de lange termijn mee te nemen. De oplossingen van het model voor verschillende gewichten geven ons de paretocurve van dit probleem. Een maatregel beoogt om deze curve in zijn geheel zo te verplaatsen dat een betere service geleverd kan worden tegen gelijkblijvende materieelkosten, of lagere kosten gemaakt kunnen worden voor een gelijkblijvende service, zie ook figuur 3. Uit de vorm van deze curve blijkt dat het absolute effect van een maatregel afhangt van de initiële plek op de curve: het verschil tussen twee curves is niet op alle plekken en in alle richtingen even groot.

Resultaten en conclusie

Met de bovenbeschreven aanpak heeft NS diverse manieren van vraagbeïnvloeding doorgerekend, onder meer het anders beprijzen van het treinkaartje. Hierin is bijvoorbeeld gevarieerd met de hoogte van een toeslag of korting, en de duur van de periode. Ook is er gekeken naar het beïnvloeden van specifieke doelgroepen en/of specifieke trajecten.

Vooraf maatregelen waarbij de tarieven van de treinkaartjes meer worden gedifferentieerd in de tijd hebben impact. Het aantal reizigers op het drukste moment in de ochtendspits kan hiermee tot 9% dalen, waarmee een aantal jaar reizigersgroei kan worden opgevangen bij de bestaande dienstregeling. Bij gelijke materieelkosten



Figuur 3. Materiaalkeuzes: Paretocurve van optimale materieelplannen met en zonder vraagbeïnvloeding

leidt een dergelijke kwaliteitsverbetering tot maximaal 16% minder reizigers zonder zitplaats. Vanzelfsprekend kan er ook gekozen worden om bij gelijke service de materieelkosten te verlagen.

De uitkomsten van ons onderzoek moeten echter voorzichtig worden geïnterpreteerd. De verschuivingsbereidheid van het voorkeursvertrektijdstip is immers gebaseerd op een stated choice onderzoek. Gedragsveranderingen worden in het algemeen overschat in dit soort onderzoeken. We bevelen daarom aan om ook enkele praktijkproeven te doen met verschillende maatregelen om de werkelijke verschuivingsbereidheid te bepalen en de aannames met betrekking tot verschuiving van dalreizigers en wisselen van productsoort te toetsen.

LITERATUUR

- Aa, M. van der, & Hodde, A. (2018). Slim Roosteren. Een onderzoek naar de (on)mogelijkheden van anders roosteren om de drukte in de spits te reduceren. *Tijdschrift Vervoerswetenschap*, 54(2), 18–21.
- Ben-Akiva, M., & Lerman, S. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*. Transportation Studies. Massachusetts: MIT Press.
- Daalen, T. van, Janssen, N., & Mastebroek, A. (2018). De hyperspits biedt kansen voor een betere spreiding binnen de spits. *Tijdschrift Vervoerswetenschap*, 54(2), 27–31.

Fioole, P.-J., Kroon, L.G., Maróti, G., & Schrijver, A. (2006). A rolling stock circulation model for combining and splitting of passenger trains. *European Journal of Operational Research*, 174, 1281–1297.

Nielsen, L.K. (2011). *Rolling Stock Rescheduling in Passenger Railways: Applications in Short-term Planning and in Disruption Management*, Ph.D. thesis, Erasmus Universiteit Rotterdam.

MENNO DE BRUYN is Senior Consultant Vervoer bij de afdeling Klant- & Marktdvies van NS.
E-mail: menno.debruyen@ns.nl

SIMONE GRIFFIOEN is Business Consultant Advies en Implementatie bij de afdeling Prestatieregie & Innovatie van NS.
E-mail: simone.griffioen@ns.nl

DENNIS HUISMAN is Expertise Manager Logistieke Processen bij de afdeling Prestatieregie & Innovatie van NS en hoogleraar Openbaar Vervoer Optimalisatie bij de Erasmus Universiteit Rotterdam.
E-mail: huisman@ese.eur.nl

ANDRIKE MASTEBROEK is Senior Pricing Marketeer bij de afdeling ComIT van NS en PhD student aan de TU Delft.
E-mail: andrike.mastebroek@ns.nl

PLEUNI NAUS is Data Science Consultant bij CQM.
E-mail: pleuni.naus@cqm.nl



IS DE STARTPROCEDURE BIJ SCHAATSEN ONEERLIJK?

Beorn Nijenhuis, oud-schaatser, zat eind 2015 aan tafel bij *De Wereld Draait Door*. Hij had onderzoek gedaan naar het effect van de startprocedure op de 500 meter. Zijn conclusie: Als de starter een seconde langer wacht met schieten heeft dit een nadelig effect van 0,67 seconden bij de vrouwen en 0,3 seconden bij de mannen op de eindtijd. Een enorm effect. Het leverde een storm aan kritiek op. Er volgden alternatieve berekeningen en een commentaar bij het oorspronkelijke artikel. Tijd voor een uitgebreide analyse.

MIRIAM LOOIS

Aan tafel bij Matthijs van Nieuwkerk, samen met Erben Wennemars en Mark Tuitert, pleitte Nijenhuis voor een geautomatiseerde startprocedure, zodat de tijd tussen 'ready' en het startschot altijd precies hetzelfde is. Hij

heeft de 500-meter races op de Olympische Spelen van Vancouver in 2010 onderzocht. Hij zette in zijn onderzoek de duur van het ready-start interval uit tegen de eindtijd, en bepaalde met lineaire regressie de helling

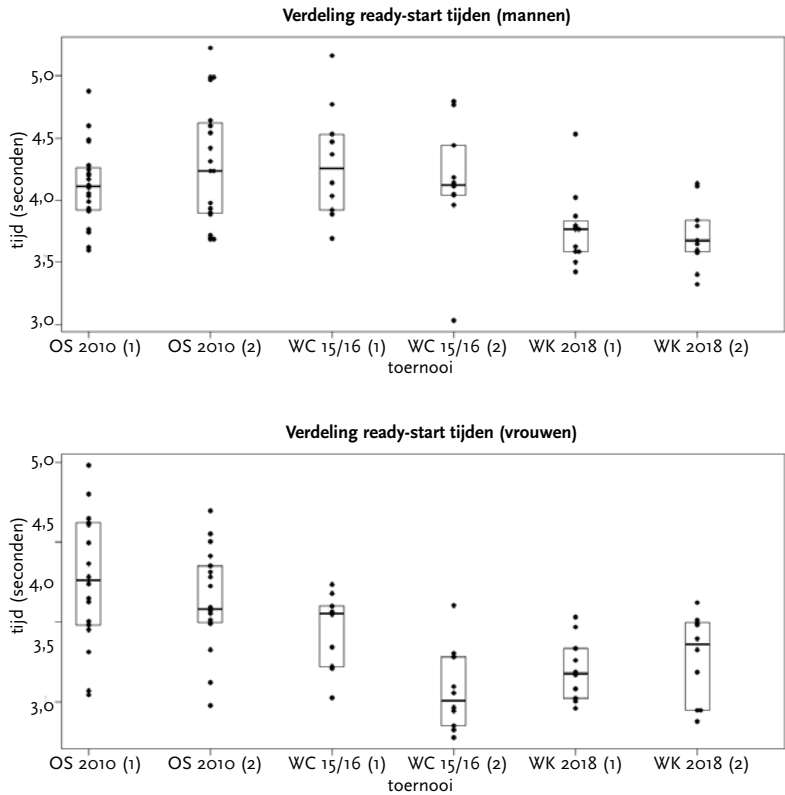
van de lijn door de punten. Hij vond een groot en significant effect: Bij de vrouwen maar liefst 0,672 seconden als de starter een seconde langer wacht, bij de mannen 0,299 seconden. In de uitzending werd vooral gediscussieerd over de romantiek van de startprocedure. Maar, die analyse klopte volgens Nijenhuis als een bus, daar hoefde verder niet over gediscussieerd te worden. In het gepubliceerde artikel zijn de auteurs overigens wel wat voorzichtiger.

Er kwam al snel kritiek, want 0,67 seconden is het verschil tussen goud en een plek in de middenmoot. Jac Orie en Michel Mulder noemden het effect wel erg groot. En er volgde ook kritiek op de opzet van de analyse, onder andere door Lex Borghans in *Economisch Statistische Berichten*, een vakblad voor economen. Wellicht duurde de startprocedure bij de betere schaatsers korter? Borghans kwam in zijn analyse uit op een niet significant effect van ongeveer twee tienden van een seconde bij zowel mannen als vrouwen. Via een collega kreeg ik een artikel onder ogen van studenten uit Groningen, gepubliceerd in het blad van hun studievereniging. Zij pleitten ervoor om handig gebruik te maken van het feit dat de schaatsers twee keer een 500 meter hebben gereden (heat 1 en heat 2). Als je het verschil in starttijd uitzet tegen het verschil in eindtijd elimineer je effecten zoals het mogelijke effect dat de starter sneller schiet bij de snellere schaatsers. Ze analyseerden meerdere toernooien en vonden ook een effect van ruim twee tienden. Een stuk minder dan Nijenhuis, maar nog steeds meer dan genoeg om invloed te hebben op de verdeling van de prijzen.

Nijenhuis en zijn medeauteurs publiceerden een commentaar bij hun oorspronkelijke artikel. Ze verwijzen daarin naar de analyse van Borghans, en ook naar een vergelijkbare analyse als die van de Groningse studenten. Wat als je kijkt naar het effect van het verschil in de starttijd op het verschil in de eindtijd? Of zelfs naar het verschil in de 100-metertijd? Het ligt immers voor de hand dat de start met name invloed heeft op de eerste honderd meter, en niet of minder

op het rondje dat volgt. In het eerste geval vinden ze een effect van 0,174 seconden op de eindtijd (gecombineerd voor mannen en vrouwen). Op de 100-meter tijd daalt de invloed van de startprocedure nog verder, naar 0,057 seconden.

Het ligt voor de hand om, net als de Groningse studenten, meerdere toernooien te analyseren. Wellicht was Vancouver net een vreemde eend in de bijt, en geven andere wedstrijden andere conclusies. Dus dat is wat we in dit artikel zullen doen. Het verzamelen en analyseren van de data is een hele klus. Ten eerste moeten er videobeelden van races beschikbaar zijn. Op Youtube is flink wat te vinden, maar vaak ontbreken er ritten of is er alleen een samenvatting te vinden. Aangezien we willen kijken naar het verschil tussen heat één en twee zoeken we ook wedstrijden waarin er twee keer gereden wordt. En laat de internationale schaatsunie nu net besloten hebben om de 500 meter nog maar één keer op het programma te zetten op Olympische Spelen en Wereldbikers. Heb je eenmaal een video gevonden, dan moet die geanalyseerd worden in een programma zoals Audacity (ook gebruikt door Nijenhuis). Op de milliseconde nauwkeurig moet



Figuur 1. Verdeling van de ready-start-intervallen van mannen en vrouwen op de drie toernooien, (heat 1, heat 2). Elk punt correspondeert met een rit. De bijbehorende boxplot geeft de mediaan, kwartiel 1 en 3 van de data weer

	MANNEN		VROUWEN	
	geschat effect δ	standaarddeviatie	geschat effect δ	standaarddeviatie
Olympische Spelen Vancouver 2010	0,066	0,023	0,089	0,025
World Cup Heerenveen 2015/2016	0,058	0,027	-0,023	0,075
Wereldkampioenschappen 2018	-0,014	0,042	-0,017	0,062
Totaal	0,051	0,016	0,066	0,022

Tabel 1. Geschatte effecten van een seconde langer ready-start interval op de 100-metertijd, en het gecombineerde effect per sekse. Het gecombineerde effect is berekend door de afzonderlijke effecten te wegen met de inverse van de variantie

geprobeerd worden om de start van het geluidssignaal van 'ready' en het schot te vinden. Urenlang schalt er 'ready...' door de huiskamer, al dan niet op dubbele afspelsnelheid. En dan blijkt dat de commentatoren soms precies door de start heen praten. Ook die video's vallen dus af. Uiteindelijk vinden we drie wedstrijden die voldoen aan alle criteria: De Olympische Spelen van Vancouver in 2010, de World Cup in Heerenveen in het seizoen 2015/2016 en de Wereldkampioenschappen Sprint in 2018.

Laten we eerst eens kijken hoe de ready-start intervallen zijn verdeeld op de drie toernooien. We maken daarbij onderscheid tussen heat 1 en 2, en tussen mannen en vrouwen. In figuur 1 is te zien dat het startinterval meestal tussen de 3,5 en 4,5 seconden ligt. In het meest extreme geval zit er 1,8 seconde tussen de snelste en de langzaamste start (heat 2 van de Worldcup in Heerenveen bij de mannen). Bij de WK bij de mannen werd er erg consistent geschoten.

Om het effect van het ready-startinterval op de 100-meter tijd te bepalen gaan we uit van het volgende lineaire regressiemodel:

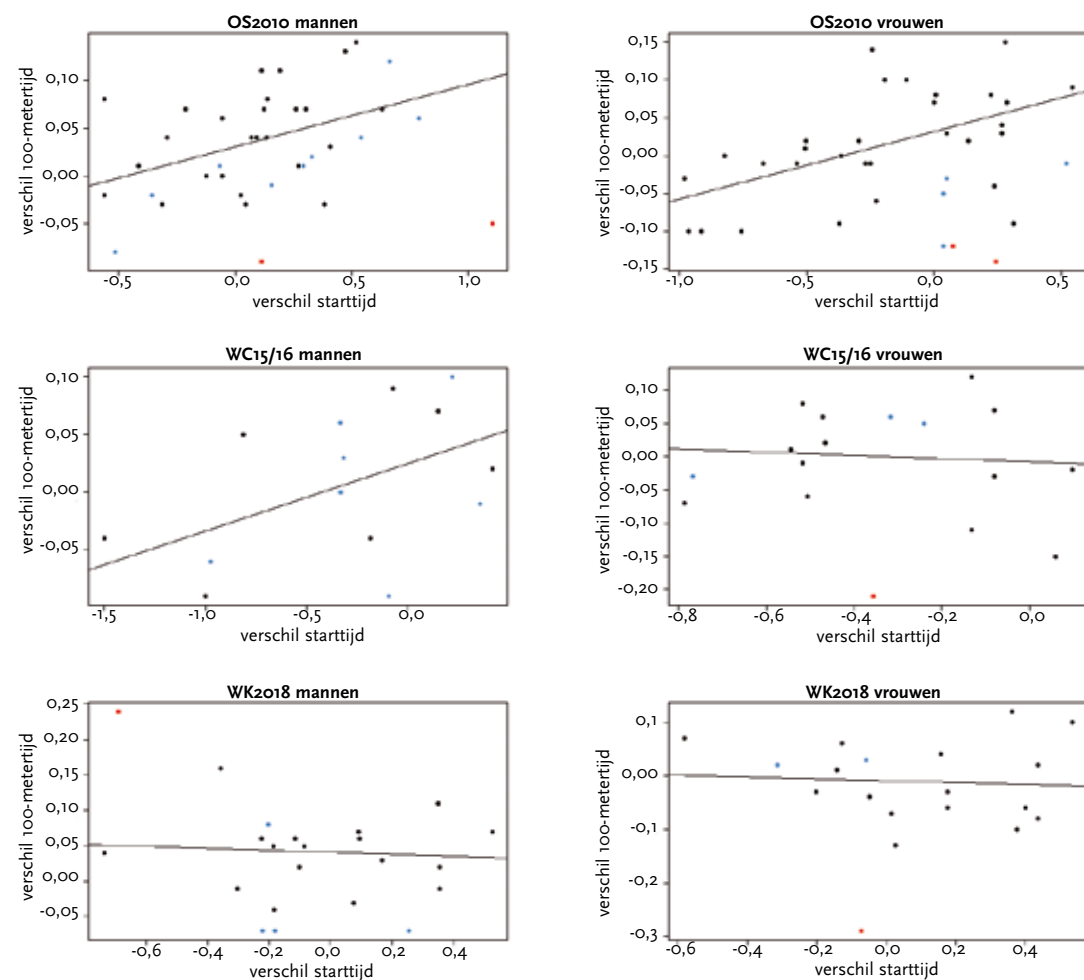
$$\Delta T_{100,i} = \alpha + \delta \cdot \Delta Tstart_i + \theta \cdot \Delta I_{false,i} + \epsilon_i$$

$\Delta T_{100,i}$ is hier het verschil in de tijd op de 100 meter (heat 2 min heat 1) van schaatser i . $\Delta Tstart_i$ is het verschil in het ready-start interval, en $\Delta I_{false,i}$ geeft aan of schaatser i een verschillend aantal valse starten had. $\Delta I_{false,i} = 1$ indien er wel een valse start was en 0 bij een goede start, dus $\Delta I_{false,i}$ is -1, 0 of 1. Dit model lijkt erg op het model van de Groningse studenten, alleen kijken we hier naar de 100-metertijd in plaats van de eindtijd, en modelleren we het effect van een valse start expliciet.

We schatten dit model eerst per toernooi en per sekse. In tabel 1 zijn de resultaten te zien. Vervolgens verwijderen we punten waarvan het residu meer dan twee standaarddeviaties afwijkt van 0, om te voorkomen dat races worden meegenomen waarbij het verschil tussen heat 1 en 2 groot is door bijvoorbeeld een forse misslag. Daarna schatten we het model opnieuw. Het gecombineerde effect van de drie toernooien wordt bepaald door de schatters te wegen met de inverse van de variantie. Hoe nauwkeuriger de schatter, hoe zwaarder deze meet. In figuur 2 zijn de data gevisualiseerd. Bij de mannen vinden we een effect van 0,051 seconde, bij de vrouwen een effect van 0,066 seconde. Beide effecten verschillen significant van 0. Wat opvalt is dat bij drie van de zes analyses juist een negatief effect wordt gevonden. In al

	MANNEN		VROUWEN	
	geschat effect δ	standaarddeviatie	geschat effect δ	standaarddeviatie
Olympische Spelen Vancouver 2010	-0,071	0,096	0,137	0,072
World Cup Heerenveen 2015/2016	-0,017	0,126	-0,375	0,188
Wereldkampioenschappen 2018	0,063	0,074	-0,212	0,183
Totaal	0,008	0,053	0,122	0,063

Tabel 2. Geschatte effecten van een seconde langer ready-start interval op de eindtijd, en het gecombineerde effect per sekse



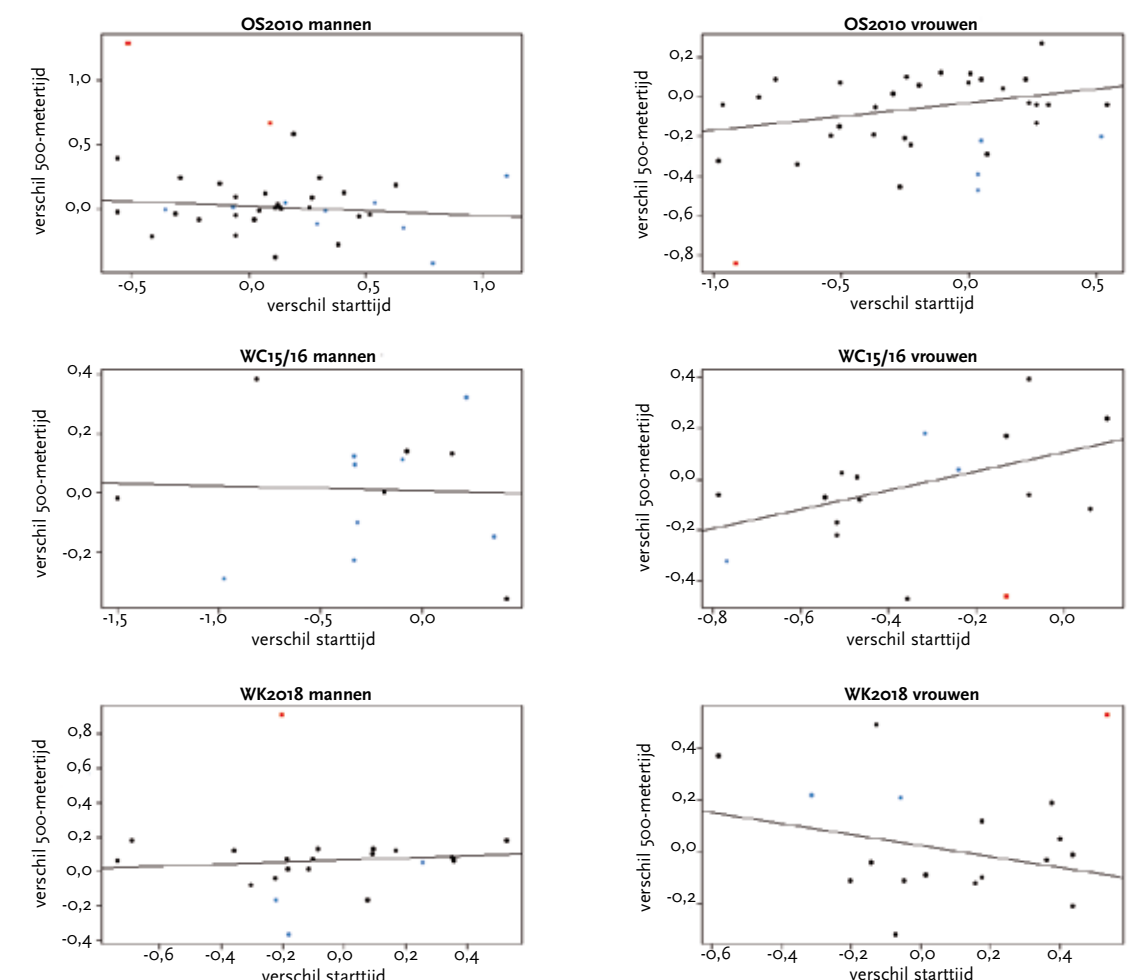
Figuur 2. Verschil in starttijd versus het verschil in 100-metertijd (rit 2 min rit 1). Punten geven individuele schaatsers weer. De rode punten geven uitschieters aan die niet mee zijn genomen in de schatting. De blauwe punten geven punten aan met een valse start in één van de twee ritten. De zwarte lijn heeft helling δ en intercept α .

die gevallen is de standaarddeviatie relatief hoog, waardoor de invloed op het gecombineerde effect beperkt is. Dit geeft wel aan dat het wenselijk zou zijn om nog meer toernooien te analyseren.

Voeren we dezelfde analyse uit voor het verschil in de eindtijd, dan vinden we bij de mannen een klein en niet-significant effect (zie tabel 2). Bij de vrouwen vinden we een relatief groot effect van 0,122 seconden dat ongeveer twee standaarddeviaties van 0 verschilt. Wat opvalt is dat de standaarddeviatie van het geschatte effect hier een stuk groter is dan bij de analyse van de 100-metertijden. Dat is ook logisch. Het rondje na de start wordt ook meegenomen, en de spreiding in het verschil in eindtijden is groter dan de spreiding in het verschil in 100-metertijden. Ook valt op dat de resultaten wat afwijken van de analyse van de Groningse studenten en de her-analyse van Nijenhuis. Dit kan komen door de manier waarop uitschieters worden geïdentificeerd en uitgesloten, door de manier waarop een valse start wordt meegenomen, en in het geval van de Groningse studenten door het gebruik van andere data.

Conclusie

Op basis van deze analyses kunnen we voorzichtig concluderen dat een langer ready-startinterval een negatief effect heeft op 100-metertijd, en daarmee waarschijnlijk ook op de eindtijd. Het gevonden effect is echter een stuk kleiner dan in de eerste analyse van Nijenhuis. Het ligt in de orde grootte van honderdsten van secondes per seconde langer ready-startinterval. Nijenhuis vond in eerste instantie 0,3 en 0,67 seconden. We moeten voorzichtig blijven, want er zijn ook kanttekeningen te zetten bij deze analyse. Zo is het nauwkeurig schatten van het ready-startinterval ingewikkeld. De onzekerheid in die tijdsintervallen is niet meegenomen in dit onderzoek. Daarbij kijkt het ready-startinterval alleen naar de tijd tussen het begin en het einde van de startprocedure. Wellicht zakt een schaatser de ene keer langzamer in waardoor hij de facto korter stilstaat, of zijn er andere verborgen variabelen. Ook wordt het totale geschatte effect voor een groot deel bepaald door één toernooi, de Olympische Spelen



Figuur 3. Verschil in starttijd versus het verschil in eindtijd

in 2010. Het zou goed zijn om meer data te verzamelen, maar dat is een tijdrovende klus.

Maar stel nu dat er inderdaad een effect is van rond de vijf honderdsten van een seconde. Hoe erg is dat dan? Het komt voor dat er binnen een wedstrijd meer dan een seconde verschil zit in de duur van het startinterval. Maar meestal ligt 50% van de intervallen (ruim) binnen een marge van een seconde. Daarmee blijft het effect beperkt tot enkele honderdsten van een seconde. Er zijn een hoop andere effecten waar je als schaatser ook geen invloed op hebt. Maakt je tegenstander wel of geen valse start? Start je in de binnen- of buitenbocht? Is de conditie van het ijs of de luchtcirculatie wel constant? Kun je bij de wissel profiteren van je tegenstander? Je kunt beargumenteren dat je alle oneerlijkheid die je weg kunt halen ook weg moet halen. Maar als de starter wordt vervangen door een robot verlies je ook een deel van de charme van de startprocedure. En het is ook aan de schaatser om te kunnen omgaan met deze wisselende omstandigheden. Wellicht een mooi compromis: Controleer of er niet een

te groot verschil zit tussen hoe lang starters wachten met schieten binnen een wedstrijd, en train ze daarin indien nodig. Dan blijft een eventueel nadelig effect in ieder geval beperkt.

LITERATUUR

- Dalmaijer, E.S., Nijenhuis, B.G., & Van der Stigchel, S. (2015). Life is unfair, and so are racing sports: some athletes can randomly benefit from alerting effects due to inconsistent starting procedures. *Frontiers in psychology*, 6, 1618.
- Dalmaijer, E.S., Nijenhuis, B.G., & Van der Stigchel, S. (2016). Commentary: Life is unfair, and so are racing sports: some athletes can randomly benefit from alerting effects due to inconsistent starting procedures. *Frontiers in psychology*, 7, 119.
- <http://deeconometrist.nl/de-schaatssport-onder-de-loep/>
<https://esb.nu/blog/20013001/schaatsers-moeten-niet-te-snel-zijn>

MIRIAM LOOIS is docent Toegepaste Wiskunde aan de Hogeschool van Amsterdam.
 E-mail: miriamloois@gmail.com



Illustratie: Thedarknut

Ik schrijf deze column als we vijf weken intelligente lockdown achter de rug hebben. Eind deze week zou ik naar Denver zijn afgereisd voor de INFORMS Business Analytics Conference, met als een van de hoogtepunten de Franz Edelman-competitie. Het is helaas anders gelopen, en de vraag is of evenementen als de INFORMS-conferentie in de 1,5-meter-economie nog wel kunnen plaatsvinden. In plaats van succesverhalen over de impact van Operations Research op de prestaties van organisaties hoor ik elke dag opnieuw analisten en experts in actualiteitenrubrieken het aantal besmettingen, sterfgevallen en de IC-bedbezetting bespreken. Zoals je kunt verwachten, leidt deze dagelijkse bijwerking van de boekhouding tot een toenemende ongerustheid in de samenleving en tot overhaaste en onverstandige beslissingen van politici. Een vergelijking met de Mexicaanse griep in 2009 dringt zich op. Toen leidde die onrust tot een massale inkoop van Tamiflu¹, nu tot de ontwikkeling (of niet) van een corona-app. Feit is dat nog veel onbekend is over covid-19, een onzekerheid waarmee we moeten leren omgaan. Goede communicatie over deze onzekerheid leidt tot een beter begrip, minder ongerustheid en naar ik hoop ook tot een *smart release*.

Onzuivere statistieken

De covid-19-crisis is de eerste wereldwijde crisis waarin we vanaf het eerste moment kunnen beschikken over grote hoeveelheden data. Voor bijna elk land ter wereld kan, bijna in real time, de ontwikkeling van het aantal besmet-

tingsgevallen en andere covid-19-statistieken worden gevolgd². Met zoveel beschikbare data lijkt het kinderspel om een inschatting te maken van belangrijke gegevens als besmettingsgraad en overlevingskans. Niets is echter minder waar. Ondanks dat er grote hoeveelheden data beschikbaar zijn kunnen we er weinig mee. John Ioannidis³, schrijver van het baanbrekende artikel 'Why most published research findings are False'⁴, noemt covid-19 een datafiasco. De achterliggende reden is een chronisch gebrek aan betrouwbare data op basis waarvan uitspraken gedaan kunnen worden over de feitelijke besmettingsgraad van de bevolking en de overlevingskans. Geen enkel land (behalve wellicht IJsland⁵) heeft tot nu toe een voldoende grote steekproef gedaan om objectief de covid-19-besmettingsgraad van zijn bevolking te schatten. De statistieken die we wel hebben zijn onzuiver en onvergelykbaar vanwege verschillende testregimes. Ook worden er verschillende definities gehanteerd, bijvoorbeeld voor sterfgevallen. Daarmee ontbreekt een datagedreven fundament voor maatregelen die voor de bestrijding van het virus moeten worden getroffen.

De beperkingen van technologie

Intussen kom ik bij het boodschappen doen steeds meer mensen tegen die zich tegen besmetting wapenen met latex handschoenen, mondkapjes en desinfecteergel. De sfeer in de supermarkt is ongemakkelijk. Winkelbezoekers proberen continu in te schatten of ze wel op 1,5

meter afstand van hun medewinkelbezoekers blijven. In een poging de restricties van de smart lockdown gecontroleerd te verruimen onderzoekt minister De Jonge met een appathon of het gebruik van technologie⁶ kan helpen om bij een verruiming van onze bewegingsvrijheid de verspreiding van het virus onder controle te houden. Van bluetooth, dat centraal staat in de te realiseren corona-app, wordt verwacht dat het de GGD'en in staat stelt om snel en effectief contacten van besmette personen te achterhalen om zo de verspreiding van het virus te kunnen beteugelen. Een droombeeld⁷ volgens Jaap Haartsen, de Nederlandse uitvinder van bluetooth. De afstandsbepaling met bluetooth wordt door vele omgevingsfactoren beïnvloed, wat zal leiden tot zowel foutnegatieve als foutpositieve meldingen. Voer voor statistici. Het laat bovendien zien dat er te veel vertrouwen in technologie wordt gesteld om onze onzekerheid over covid-19 weg te nemen. Aan de andere kant toont het aan dat onze maatschappij steeds datagedrevener wordt. Technologie – en dus ook de vakgebieden waar de VVSOR voor staat – zal een steeds centralere rol gaan spelen in het oplossen van maatschappelijke vraagstukken. Het zal echter nooit de onzekerheid kunnen wegnemen die gepaard gaat met *unknown unknowns* als het covid-19-virus.

Omgaan met onzekerheid

We zullen pas na de covid-19-crisis weten of we nu wel of niet overreageren. Geavanceerde technologie gaat ons op dit moment niet meer zekerheid verschaffen. Het zal ook niet de laatste keer zijn dat we ons in zo'n onzekere situatie bevinden. Covid-19 zal niet de laatste wereldwijde pandemie zijn; nieuwe risico's en onzekerheden zullen zich blijven aandienen. Zoals Gerd Gigerenzer in *Risk Savvy* bepleit is het belangrijk dat we leren omgaan met deze onzekerheid. We moeten leren om onzekerheid in onze beslissingen expliciet mee te wegen, in plaats van ons er door te laten gijzelen of te vluchten in complexe nieuwe technologie als AI of bluetooth. Juist als het risico complex is is een simpele oplossing effectiever, aldus Gigerenzer. In geval van covid-19 zal heldere communicatie over wat we wel en niet weten, en de mate waarin we daar zeker over zijn, bijdragen aan beter begrip en beter doordachte besluitvorming. Recent onderzoek⁸ van onder andere David Spiegelhalter en Anne Marthe van der Bles toont aan dat het een misvatting is dat 'het publiek' onzekerheidsmarges van wetenschappers niet begrijpen. Door de onzekerheid te verzwijgen zet je mensen juist op het verkeerde been. Rutte maakte de onzekerheid in zijn

besluitvorming expliciet met zijn uitspraak⁹ dat met 50% van de kennis 100% van de beslissingen moeten worden genomen in deze covid-19-crisis. Daarmee omarmt Rutte de onzekerheid in zijn besluiten en komt hij in actie. Leaders als Trump en Bolsonaro zijn kennelijk nog niet zover als Rutte en komen pas laat in actie. We kennen inmiddels de gevolgen.

Slim te werk gaan

Natuurlijk kan ook ik niet wachten tot de smart lockdown al dan niet geleidelijk wordt opgeheven. Ik sta te popelen om weer met klanten in levende lijve te kunnen spreken, met vrienden en kennissen af te spreken, rockconcerten te bezoeken en met mijn gezin er op uit te trekken. Maar laten we dat dan vooral slim doen. Niet overhaast of met de schijnzekerheid van complexe technologie, maar met simpele regels en helderheid over wat we wel en niet weten. Als risico's goed begrepen en kwantificeerbaar zijn helpt het om logica, beslismodellen en technologie te gebruiken. Als de risico's in de categorie *unknown unknowns* vallen, helpen complexe modellen niet. Ze wekken dan juist schijnzekerheid en leiden tot wantrouwen. In dat geval zijn slimme vuistregels een krachtigere oplossing. Social distancing, regelmatig je handen goed wassen en eventueel in zelfquarantaine gaan zijn goed uitlegbaar en effectief, daar heb je geen app voor nodig. Voeg daar heldere communicatie over de onzekerheid aan toe en het zal ons in staat stellen om naast een smart lockdown ook een smart release te realiseren.

LITERATUUR

1. <https://www.medischcontact.nl/nieuws/laatste-nieuws/artikel/inkoop-griepremmers-was-onbezonnen.htm>
2. https://slides.ourworldindata.org/2020_pandemic/2020_pandemic#/title-slide
3. <https://www.statnews.com/2020/03/17/a-fiasco-in-the-making-as-the-coronavirus-pandemic-takes-hold-we-are-making-decisions-without-reliable-data/>
4. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327/>
5. <https://www.government.is/news/article/2020/03/15/Large-scale-testing-of-general-population-in-Iceland-underway/>
6. <https://www.rijksoverheid.nl/onderwerpen/coronavirus-app>
7. <https://www.computable.nl/artikel/nieuws/security/6919575/250449/uitvinder-vindt-bluetooth-ongeschikt-voor-corona-app.html>
8. <https://royalsocietypublishing.org/doi/10.1098/rsos.181870>
9. <https://nos.nl/video/2326873-rutte-we-hebben-iedereen-nodig-17-miljoen-mensen.html>

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

From killer nurses to quantum entanglement, and back Part 2

In Part 1 of this paper I wrote about the Lucia case and about the Delft Bell experiment. It is now time to collect loose ends together, and also present a new data-analysis which I hope will lead to new developments in the Ben Geen case.



Foto: Stefan Schwehofer via Pixabay

RICHARD D. GILL

What I showed you in Part 1 about the Ben Geen case is horribly like the Lucia case. We now know that the Lucia numbers which I showed you are actually *wrong*. Events have been misclassified, have been shifted from one shift to an adjacent shift. Quite a few have been forgotten – they weren’t “unexplained”, so they weren’t included as “incidents”. Several real incidents (i.e., “incidents” according to the law) have been suppressed. They should have been reported to the health inspectorate, but this wasn’t done. For instance, one incident was a euthanasia which, illegally, was not reported. The data from JKZ was compiled in a great hurry by someone who was already convinced they were dealing with a serial killer and embarrassed that they didn’t catch her earlier. Lucia was in fact set up, in the sense that the hospital’s clinical director (the head paediatrician) was waiting for a final event to clinch the case she had been building up; she had already compiled the dossiers of selected earlier events.

In hospitals, accidents do happen, but they *must not happen*. The legal and financial consequences are too great. In the Lucia case, I dare to say that at least three

top medical specialists have lied (under oath) to the police and to the courts.

There are two other young male nurses in Britain (well, they were young when they got thrown into jail), sitting out life sentences, supposedly for having killed their patients. New medical insight into hypoglycaemia shows anyone with any intelligence that they are innocent. Yet the criminal justice system is not keen to let convicted people free. There is next to no financial aid for people who need to recruit very expensive lawyers to help them fight an unjust conviction. The national institution (the CCRC) which was set up to deal with possible miscarriages of justice, following a number of scandalous miscarriages, is severely underfunded. The strategy of “the system” is to do nothing and wait for the problem to go away.

The reopening of the Lucia de B. Case

Why, and how, did we succeed in getting Lucia a fair re-trial in the Netherlands? Answer: Lucia was very, very lucky! That’s all. There was an almost-inside whistleblower. The

sister-in-law of JKZ’s chief paediatrician was a medical doctor, knew her sister-in-law very well, and moreover, her son and daughter, studying medicine in Leiden, lived with their uncle and aunt in the Hague. Even so, it took a couple of years for the penny to drop.

The “true story” of the Lucia case is still taboo. While trying to raise public interest in this case and the, in my opinion doubtful, role of some particular individuals in it, I was threatened by the most famous “intellectual property” law firm in the Netherlands – a well-known law firm, often hired to (successfully) defend the reputation of very rich persons. My University was so kind as to pay for a good, expensive, Leiden educated lawyer to help me for a few months, but even he finally advised me to give way and remove some pages from my web-site, since it was going to drag on for a long, long time, cost the university his hefty monthly fee every month, and probably not succeed. It was a civil case, and I was harming the reputation, hence the earning power, of powerful and wealthy persons. You can only do this if you are very, very wealthy yourself. Journalists today who know the whole inside story do not write about it. Too many too powerful

people are still alive and kicking. I hope to outlive them.

As I mentioned, I believe that quantum mechanics did help save Lucia de Berk from lifelong imprisonment for crimes that no-one had committed. At some point a movement among scientists and journalists and decent right-minded people was gathering strength, that Lucia’s trial had been unfair; incompetent experts had given evidence; the statistical evidence wrongly interpreted. But the wheels of justice turn so, so slowly. Rumour even had it that of three possible miscarriages of justice currently being investigated by a special review committee, at most one would be allowed a retrial, since any more would too much shake the public’s faith in our justice system. It’s like, when someone runs through a burning building shouting “fire”, accusing them of undermining people’s faith in the installed fire prevention systems.

I decided to set up an internet petition and recruited many friends and colleagues to sign it. In particular, one Friday afternoon, I walked over to the physics department and knocked on the door of Professor Gerard ’t Hooft, Nobel prize-winning physicist, with whom I had discussed Bell’s theorem and quantum foundations a number of

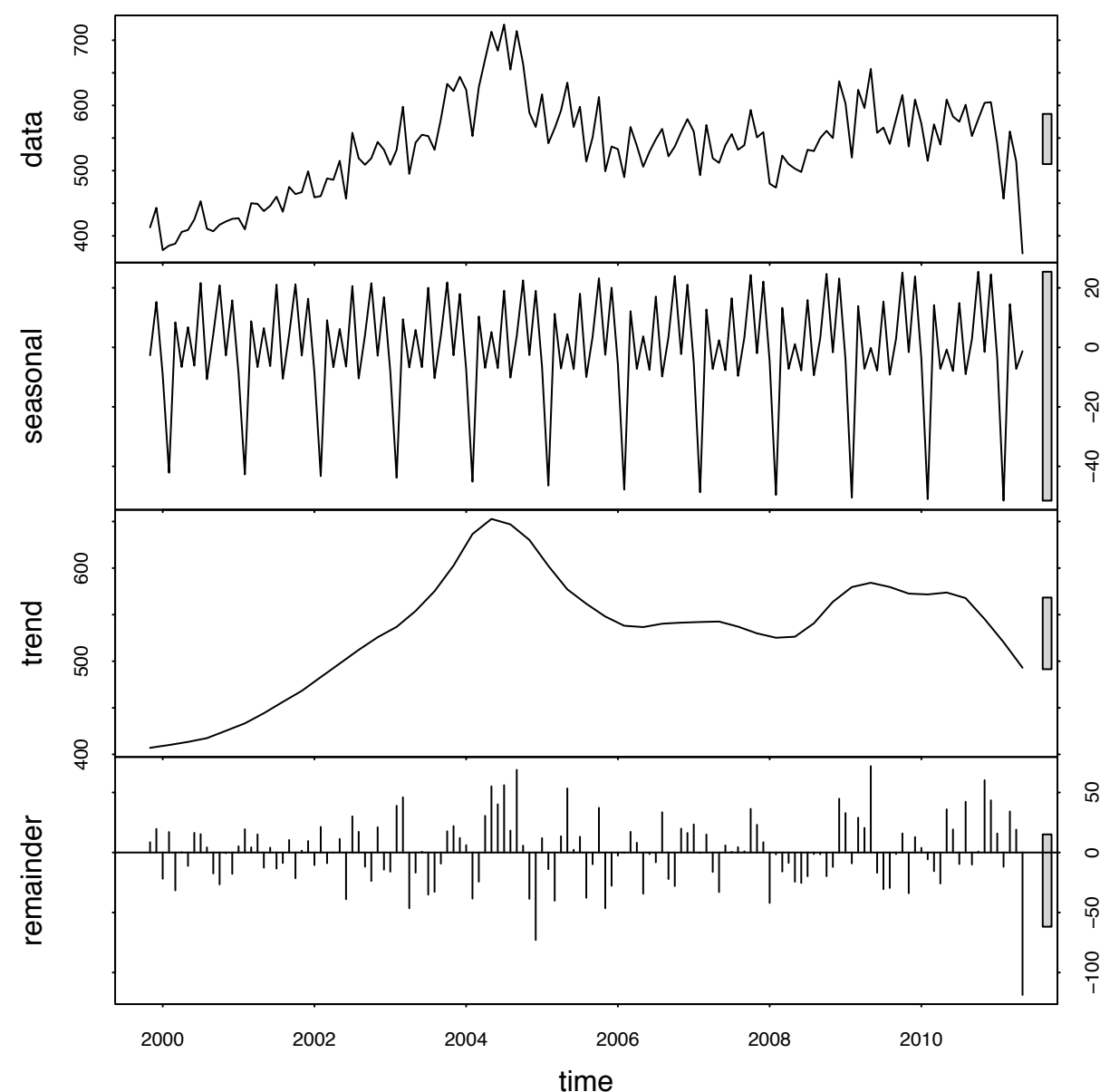


Figure 1. Monthly admissions to emergency; Decomposition of data into trend, seasonal, remainder

times. I told him why I had come – just to get his name on a petition. Which did not have anything at all to do with physics. He did already know something about the case, and I told him more. He told me he would go home and think about it, and indeed, that weekend he talked with his wife and his daughters, all of whom are medical doctors. On the Monday morning he signed the petition, together with a couple of germane sentences about the fact that in places where people often die, innocent nurses can easily be often present at deaths which at the time appear surprising. The signature was noticed by journalists and in no time picked up by journalists on *The Guardian* and *The New York Times*. And this again was noticed by

journalists in the Netherlands. And I'm sure that *this* was noticed in the corridors of power in The Hague. Where the legal establishment and powers-that-be made sure that the very best, most intelligent and independently thinking, careful and thorough legal minds were involved in the various steps which have to precede the re-opening of a definitively closed case. The advocate-general to the supreme court commissioned new research and also pointed out that a former crucial medical expert agreed with the new expert's findings, pointing out that during the earlier trial, crucial evidence had been withheld from him. He wrote that if the supreme court was petty enough to require a legal "novum" (the legal concept of a new

fact does not include new interpretations of old "facts") then this should be adequate for them.

I like to think that this helped. Though I have also heard it suggested that the petition delayed the re-trial because the establishment certainly could never give the impression that they had been influenced by some kind of social movement in the then new "social media", unchained by a bunch of unruly scientists.

Ben Geen revisited

I now want to present some completely new data from Ben Geen's hospital, never seen before. To start with, let's have a look at the monthly total number of admissions in the emergency room of his hospital over a thirteen year period roughly centred on the critical end of 2003 – beginning of 2004. I have given this time series to a standard algorithm which, using a moving window of length 21 months, draws us a slowly varying seasonal effect, a fairly smooth trend, and what is left over. This is pure data-analysis, no modelling assumptions are being made, I am just applying a fairly well known algorithm ("STL LOESS") to the data in order to let the data speak for itself. See figure 1.

What do we see? Up to summer 2004 it is getting continuously busier and busier. The number of patients being treated in Emergency almost doubles from about 400 per month to about 800 per month. I wonder if the number of nurses also doubled during this period. I very much doubt it. Then, the number collapses. Probably due to the situation which arose after Ben's arrest and trial, potential patients tended to go elsewhere, if they had the choice; and probably hospital policy changed strongly, too. The number subsides to 500 – 600 per month. There is a sudden dip in the very last month. The last month was not quite over when the data was submitted.

The seasonal effect shows a strong annual spike downwards. It's the Januaries! It is well known that in the Northern hemisphere, everywhere where there really is a "winter", people simply stay home, and in particular, don't go to hospital if they can help it, in January. There are, for instance, much less car accidents than in any other month, because much less people go onto the roads. Old people avoid slippery paths by staying at home.

Apart from that I don't see any patterns. Accidents happen, and medical emergencies happen, at pretty constant rate during the year. Correcting months for their varying numbers of days (think February) makes no discernible difference.

Anyway, we see that the number of people coming to that hospital was steadily increasing in the years before Ben's troubles. Probably we should correct the rates of various kinds of particular events to the scale of the operation, i.e., the number of admissions. One should also look at numbers of nurses; numbers of beds; and so on. We plainly see the amplitude of the "remainder" increasing with the overall scale. The bigger the numbers, the bigger the random variation. This means in particular that at the time which interests us most (winter 2003 – 2004) the random variation is largest!

Let's look at the total numbers of transfers to Critical Care (i.e., the patient is no longer waiting in the corridors for someone to make a decision, but is actually put in a hospital bed in an intensive care ward) from Emergency. We have the numbers with the interesting diagnoses of cardio-respiratory, respiratory, and hypoglycaemic arrest, but we don't know if the "arrest" had been diagnosed before the patient was brought to the hospital, or if it only occurred while the patient was waiting at Emergency. Very sick patients who have to wait a long time in Emergency before anyone can do anything with them are likely to suddenly get a lot worse while they are waiting. See figure 2.

What do we see? Just what we would expect, given the total numbers of admissions which we just studied. A slow increase, then a collapse to a stable, lower number. Take a look at what happens if we normalise the numbers by looking at monthly totals per 100 admissions to ER. Statistically, it looks to me very much as though nothing is going on here at all. December 2003 and January 2004 are low but February 2004 is high. There is a spike upwards in February 2004, and another spike upwards in November 2008. Both of those spikes coincide with new high levels of the monthly numbers of admissions to Emergency.

Ben Geen did not work at the hospital after 5 February, 2004. Half of the big peak in February 2004 has absolutely nothing to do with him. January 2004 is strikingly high but not more so, than was February 2004. The *only thing* that

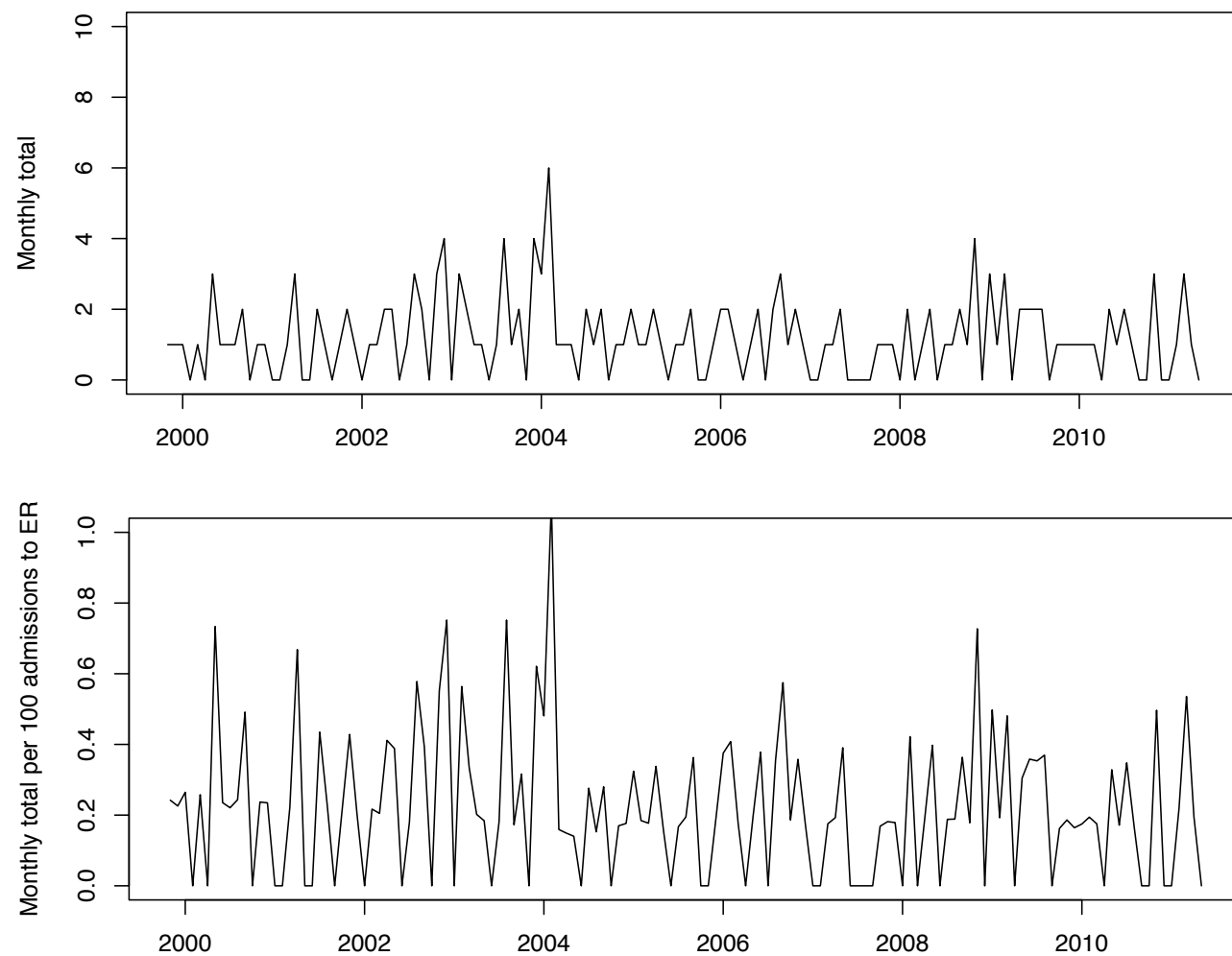


Figure 2. From Emergency to Critical Care with cardio-respiratory or hypo arrest

is unusual in those months December 2003 – February 2004 is that events are being classified as *respiratory arrest* instead of as *cardio-respiratory* or *hypoglycaemic*. Why does the number of cardio-respiratory arrests **fall so suddenly** in those three months? **Do we really believe** that there were **no** cardio-respiratory arrests in December 2004?

When were those numbers “fixed” in the official records: **day by day** as patients were admitted? Or **retrospectively** after Ben’s case started, 5 February 2004?

Of course, there is also medical evidence. There are the extensive dossiers concerning the 18 selected patients. The jury decided that Ben had deliberately harmed 17 of them, and in two cases, that this led to their early deaths (earlier than would have been expected). So 17 families received large compensations, no doubt ultimately paid for by the British tax-payer. Ben is stuck in jail, still. Somehow, he does not have the media appeal

which some other victims of miscarriage of justice do have. That damned syringe grabs the imagination.

Frankly speaking, his “legal team” has made mistakes, but they too are feeling their way in the dark. The English system means that you are helped by a solicitor who is just a family lawyer who usually has little experience in serious criminal cases. The solicitor needs to recruit a barrister (a lawyer who has been called to the bar). Very good barristers are very expensive and don’t actually have a lot of time, leaving a lot of work to paralegal assistants. In this case, the case was used as a start-up in a “freedom project”. Some law students got things moving. And then over the years, the students evaporated, their initial work going to waste. Without actually consulting with a statistician they decided to gather a lot of statistical data through freedom of information requests to numerous hospitals. A lot of data came in, typically in the form of pdf files of printouts of Excel spreadsheets prepared by

reluctant and busy hospital administrators. Often there is no indication what the meaning of a space is, often small numbers were replaced by “less than 5” in order to protect the anonymity of the patients concerned. The actual meaning of the request was definitely ambiguous. And anyway, these diagnoses are vague and multi-interpretable.

But I must admit, the statistician who Ben’s supporters recruited, yours truly, just couldn’t cope with the mess he had got into. He tried to answer the question which was put to him, but in hindsight it was a wrong question (as the CCRC was swift to point out). He did what he could in a short time (making many more hours than he had budgeted for), with the help of a succession of two students, both of whom got severely depressed by the work and moved on to something else, though fortunately not before making magnificent progress in “unlocking” the data for me. Though British, I was also not familiar at all with the English legal system. I went public with my findings and my further opinions about the case, which definitely is “not done”. It gave the lawyers of the CCRC the perfect excuse to disqualify any contribution by me or by people influenced by me.

And finally, this case, like Lucia’s, depends on lawyers and on medics. As long as Ben does not have a prestigious English *medical* expert publicly on his side, I feel that there will be little impetus for anything to happen.

How to get that to happen? One solution might be through the intermediary of a top English statistical expert, one with connections in the corridors of power ... one who rubs shoulders with influential people in the medical establishment. (I am not a prestigious English statistician. I crossed the channel to the nasty continent just across the sea long ago, and never came back). I did already gain the support of suitable persons in the British statistical world who wrote some powerful letters to the CCRC, but to no effect. The CCRC says that the statistics are a sideline. There is the syringe, and there is the medical evidence. I must ask my friends to try a different tack. We must get a famous and very senior medic on board.

But maybe this is not necessary. Maybe the law is enough. Here is an opinion of an English lawyer:

The hospital’s illegal and unqualified investigation team was only looking for evidence to secure a

conviction (Confirmation Bias) while discarding or ignoring evidence that proved Ben’s innocence. The hospital’s Serious Untoward Investigation Team initiated by Chief Nurse Brock in her capacity of Executive Lead for Governance consisted of several medical, nursing and medical records staff who were all untrained in forensic investigative techniques, crime scene preservation and the taking of witness statements. The team carried out an unlawful and flawed investigation, the material from which was later presented to medical experts appointed by the prosecution as legitimate. The opinion of these medical experts was based on flawed evidence, which had been given to those experts without their knowledge of how that evidence had been obtained. Expert opinion given on the basis of ignorance of improperly obtained evidence invalidates that medical expert evidence. The judge and jury were not aware at the trial that the evidence had been unlawfully obtained, nor of the risks to justice associated with it.

I would mention alongside of *confirmation bias* also the *Baader-Meinhof Syndrome* (or *Frequency Illusion*), partly in reverse time (retrospective reclassification of incidents). And of course, I must mention that there is no doubt that the team was acting with the best of possible intentions.

The CCRC should be able to recommend that the case is re-opened, right now. So far, this was indeed the Lucia case, all over again.

Postscript

It doesn’t end here; cases like this just keep on happening, all over the world. If any Italian statistician is interested in working on a similar *Italian* case with me, I have one for you (Daniela Poggiali). I have a lot of very interesting and challenging data which I’d like to share with you, and some ideas about what might be going on there.

RICHARD GILL is professor emeritus of Statistics at Leiden University. This article is mainly based on his ‘exauguration’ speech. Apart from other distinctions he is an Honorary Member of the VVSOR.
E-mail: gill@math.leidenuniv.nl

RUSSISCH EI ROULETTE

Pareltjes over kansrekening liggen soms voor het oprapen. *Egg Russian Roulette* is zo'n pareltje. Enkele jaren terug werd dit spel dagelijks gespeeld in de populaire *The Tonight Show* van NBC met Jimmy Fallon. Het spel werd door de gastheer Jimmy gespeeld met een gast van zijn show. De gast was steeds een beroemdheid uit de filmwereld of uit de sport. Onder de gasten waren Magic Johnson, David Beckham, Tom Cruise en Jodie Foster om enkele te noemen. De gast en Jimmy nemen om de beurt een ei uit een doos en slaan het ei stuk boven op hun hoofd. De doos bevat 12 eieren waarvan acht gekookt en vier rauw zijn, waarbij Jimmy en zijn gast niet weten welke eieren gekookt zijn en welke rauw. Degene die als eerste twee rauwe eieren stukslaat is de verliezer. Een onnozel spel dat echter uiterst populair was bij het grote publiek. Geen beter vermaak dan leedvermaak over een beroemdheid bij wie het eierstruif van het voorhoofd druipet. De gast is de eerste die een ei uit de doos pakt en stukslaat tegen het voorhoofd. Is in dit spel de gast in het nadeel? Op deze vraag is het antwoord 'ja': degene die als eerste een ei pakt zal de verliezer van het spel zijn met kans $5/9$. Op de berekening van deze kans zal hieronder nader worden ingegaan. De berekeningen laten ook het enigszins verrassende resultaat zien dat bij negen ongekoekte en drie rauwe eieren het spel eerlijk is en het dus niet uitmaakt wie als eerste een ei pakt.

Egg Russian Roulette is niet zomaar een spelletje uit een Amerikaanse tv-show. Het heeft een rijke historie die teruggaat tot de middeleeuwen. In het Engelse boerengehucht Swaton (momenteel 184 inwoners) begon het gooien van eieren rond 1322 toen de nieuwe abt van Swaton, die alle pluimvee uit de streek in bezit had, eieren als aalmoes gaf aan de trouwe bezoekers van de kerk. Als de

kerk niet bereikbaar was wanneer de door het gehucht stromende rivier Eau niet overgestoken kon worden vanwege hoog water, werden de eieren over de gezwollen rivier naar de wachtende gelovigen aan de overkant gooid. In iets andere vorm is deze traditie sinds een aantal jaren weer hersteld. Om geld in te zamelen voor liefdadigheidsdoelen wordt sinds 2006 elk jaar in het dorp een wereldkampioenschap Russian Egg Roulette gehouden die vele deelnemers vanuit de hele wereld trekt. In man-tegen-man wedstrijden wordt het spel gespeeld met zes eieren waarvan vijf gekookt en één rauw. Jammer dat zo weinig aandacht op de Nederlandse televisie besteed wordt aan dit spetterende evenement, waar overigens geen kip kwaad gedaan wordt. Het 2020 wereldkampioenschap Russian Egg Roulette in Swaton vindt plaats op zondag 28 juni. Als dit nummer van *STATOR* op tijd in de bus valt, kunnen lezers van dit blad zich mogelijk nog aanmelden als deelnemer om als een nieuwe Raymond van Barneveld ons land ook in deze tak van sport op de kaart te zetten.

Egg Russian Roulette is als kansprobleem voor elke docent van een inleidend college toegepaste kansrekening een geschenk uit de hemel. Alvorens in te gaan op het kansprobleem uit de Jimmy Fallon show, beschouw het simpeler probleem met zes eieren en twee spelers die om de beurt een ei stukslaan, waarbij de verliezer de speler is die als eerste een rauw ei stukslaat. Als onder de zes eieren slechts één rauw ei is, dan maakt het niet uit welke speler begint: de kans dat bij de k de poging het rauwe ei gepakt wordt is $5!/6! = 1/6$ voor elke k zodat de spelers ieder een verlieskans van $3 \times 1/6 = 1/2$ hebben. Zijn er twee rauwe eieren onder de zes eieren, dan zal de speler die begint de verliezer zijn met kans $3/5$. Immers

de kans op het eerste rauwe ei is $2/6 = 1/3$ bij de eerste poging, $4/6 \times 3/5 \times 2/4 = 1/5$ bij de derde poging en $4/6 \times 3/5 \times 2/4 \times 1/3 \times 1 = 1/15$ bij de vijfde poging. Hoe pak je de kansberekening aan voor het geval van 12 eieren met acht gekookte en vier rauwe eieren? Een elegante aanpak is de Markov keten aanpak met absorberende toestanden. Definieer (i, a, b) als de toestand dat i eieren uit de doos gepakt zijn waaronder a rauwe eieren door de gast van de show en b rauwe eieren door de gastheer, waarbij $0 \leq i \leq 11$ en $a+b \leq 3$. De begintoestand van de Markov-keten is $(0, 0, 0)$. De toestanden $(i, 2, 0)$, $(i, 0, 2)$, $(i, 2, 1)$ en $(i, 1, 2)$ worden absorberend genomen, oftewel het proces stopt zodra één van deze toestanden bereikt wordt. De 1-staps overgangskansen vanuit de niet-absorberende toestanden zijn simpel. Vanuit een niet-absorberende toestand (i, a, b) met i even gaat het proces naar toestand $(i+1, a+1, b)$ met kans $(4-a-b)/(12-i)$ of naar toestand $(i+1, a, b)$ met kans $1-(4-a-b)/(12-i)$, terwijl vanuit een niet-absorberende toestand (i, a, b) met i oneven het proces gaat naar toestand $(i+1, b+1, a)$ met kans $(4-a-b)/(12-i)$ of naar toestand $(i+1, a, b)$ met kans $1-(4-a-b)/(12-i)$. De gast opent het spel en de verlieskans van de gast is de som van de kansen om uiteindelijk in één van de toestanden $(i, 2, 0)$ of $(i, 2, 1)$ met $i=3, 5, 7, 9$ of 11 geabsorbeerd te worden. Deze absorptiekansen vind je door de matrix van 1-staps overgangskansen 11 keer met zichzelf te vermenigvuldigen en dan de kansen te nemen uit de bij begintoestand $(0, 0)$ behorende rij van het 11-voudige matrixproduct. Zo vind je de waarde $5/9$ voor de kans dat de gast de verliezer van het spel zal zijn (het 11-voudige matrixproduct geeft ook de verlieskans vanuit elke tussentoestand die tijdens het spel kan optreden). Op dezelfde wijze vind je de verlieskansen $1/2$ en $223/396$ voor drie respectievelijk vijf rauwe eieren in een doos van 12 eieren. Een andere aanpak om de verlieskans te vinden is computersimulatie (een Pythoncode beslaat slechts een paar regels). Een alternatieve simulatie-aanpak wordt gegeven door de real-life uitvoeringen van het kansexperiment. Op internet zijn video's te vinden van de vele afleveringen van Egg Russian Roulette in *The Tonight Show* van Jimmy Fallon met Higgins als onovertroffen sidekick die met zijn snerpande stemgeluid doet denken aan de door Marty Feldman gespeelde assistent Igor in de komische parodiefilm *Young Frankenstein* uit 1974.

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



VVSOR is op zoek naar een nieuwe voorzitter!

In het najaar van dit jaar zouden we graag een nieuwe voorzitter leren kennen, zodat we een feestelijke wissel kunnen organiseren tijdens de ledenvergadering van maart 2021. Nomineer je kandidaat door een mail te sturen naar db@vvsor.nl.

VVSOR is looking for members for the daily board!

VVSOR board membership is a great way to get to know the statistics and OR community in the Netherlands. The governing language of VVSOR is transitioning to English, so we explicitly welcome statisticians that are new to the Netherlands to help us expand the community.

Nominate a secretary or webmaster for the VVSOR daily board.

Nominate yourself or discuss your nomination with the board member-to-be and let us know at db@vvsor.nl.

Secretary takes care of member's administration, note-taking and communications.

Webmaster takes care of the online representation of the society and keeping the website up-to-date.



Dear statistics enthusiasts,

We had planned some events for the first half of 2020 and were busy organizing these events in the previous months. However, we stopped the organization of these events due to the pandemic of the coronavirus. Therefore, there were no events hosted by us in the previous months. For information about the [Young Statisticians](https://www.vvsor.nl/young-statisticians/) visit the website <https://www.vvsor.nl/young-statisticians/> where you can also subscribe for our newsletter.



Foto: via Pixabay

STATCHECK

een spellingchecker voor statistiek

We zijn gewend om een spellingchecker te gebruiken om spelfouten uit onze teksten te halen. Helaas houdt zo'n spellingchecker geen rekening met statistische 'spelfouten'. Bij het opschrijven van statistische resultaten is een typfout zo gemaakt. Een verkeerd 'gespeld' getal kan grote gevolgen hebben voor hoe een resultaat geïnterpreteerd wordt. We zouden dus eigenlijk ook een automatische check moeten hebben voor verkeerd gerapporteerde statistiek. Het gratis programma Statcheck biedt uitkomst.

MICHÈLE B. NUIJTEN

Iedereen gebruikt wel eens een spellingchecker. Logisch: het is een snelle en makkelijke manier om spelfouten uit je tekst te halen. In principe zouden we het waarschijnlijk ook wel zonder kunnen, maar in de praktijk voorkomt zo'n automatische spellingcheck toch veel onnodige fouten. Dit is eigenlijk precies het principe achter het programma Statcheck: een soort spellingchecker, maar dan voor statistiek.

Statistische spelfouten

Er zijn veel dingen die fout kunnen gaan bij het opschrijven van statistische resultaten. Ga maar eens na of het

volgende scenario bekend klinkt.

Je hebt een analyse gedaan in een statistisch programma (SPSS, R, MATLAB, etc.). De getallen die je moet rapporteren staan verspreid over de output, dus *copy-paste* je sommige stukken, en andere delen typ je met de hand over. Het kost best wat tijd om de resultaten netjes op te maken (sommige letters moeten schuingedrukt, er zitten subscripten bij, Griekse letters, wiskundige symbolen, etc.), dus als je eenmaal een resultaat op papier hebt, gebruik je deze als *template* voor de volgende resultaten: je kopieert dit ene resultaat, en vervangt alleen de getallen.

Als je bovenstaand scenario in je hoofd hebt, is het een kleine stap om je het volgende voor te stellen: In plaats van 2,18 schrijf je per ongeluk 2,81 op. Bij het re-

sultaat dat je gekopieerd had voor de opmaak vergeet je één van de getallen aan te passen. Wanneer je de resultaten van een test overneemt uit de output, selecteer je per ongeluk de vrijheidsgraden die bij een andere test horen.

Sommige statistische rapportagefouten zijn niet meer te ontdekken zonder de ruwe data. Andere, daarentegen, veroorzaken 'interne inconsistenties' die in de tekst te spotten zijn. Een interne inconsistentie ontstaat als getallen die bij elkaar horen, niet meer bij elkaar kloppen. Een simpel voorbeeld van een interne inconsistentie is $2 + 2 = 5$.

Bij het rapporteren van statistische analyses kunnen interne inconsistenties bijvoorbeeld ontstaan in de resultaten van nulhypothese-significantie toetsing (NHST). NHST is een type statistiek dat veel gebruikt wordt bij experimenteel onderzoek. Een NHST-resultaat bestaat uit drie onderdelen: een testwaarde, vrijheidsgraden, en een p -waarde. Bijvoorbeeld: $t(28) = 1,94$, $p = 0,036$. Als ik twee van de drie getallen heb, kan ik de derde uitrekenen. Zo kan ik bijvoorbeeld nagaan dat bij een t -waarde van 1,94 met 28 vrijheidsgraden een p -waarde van 0,063 hoort, en niet van 0,036.

Waarom we statistische spelfouten moeten vermijden

Een spelfout heeft meestal niet direct invloed op hoe we een tekst interpreteren, maar een inconsistentie in een statistisch resultaat kan wel degelijk invloed hebben op de conclusie. Laten we bijvoorbeeld nog eens kijken naar het bovenstaande resultaat: $t(28) = 1,94$, $p = 0,036$. In mijn vakgebied (psychologie) wordt een resultaat vaak als statistisch significant beschouwd als de p -waarde kleiner is dan 0,05. Dat betekent dat het resultaat gerapporteerd staat als statistisch significant ($0,036 < 0,05$), terwijl dat het in werkelijkheid misschien niet is ($0,063 > 0,05$).

Niet alleen inconsistenties rond die significantiegrens zijn problematisch. Ook wanneer de herberekende p -waarde aan dezelfde kant van de 0,05 valt als de gerapporteerde, kan een inconsistentie gevolgen hebben. Ten eerste weten we niet waar de fout zit: bij een inconsistentie kan het probleem in elk van de drie elementen van het resultaat zitten. Net zoals dat $2 + 2 = 5$ eigenlijk $2 + 2 = 4$ had kunnen zijn, maar ook $2 + 3 = 5$. Als we niet weten waar het mis is gegaan, kunnen we geen van de drie getallen meer vertrouwen. Dit is bijvoorbeeld problematisch in meta-analyses, waarbij testwaardes gebruikt kunnen worden om effectgroottes te berekenen.

Helaas blijkt dat inconsistenties in NHST-resultaten

veel voorkomen. Zo onderzochten we bijvoorbeeld meer dan 16 000 psychologie-artikelen met meer dan 250 000 NHST-resultaten (Nuijten et al., 2016). We vonden dat grofweg 50% van alle artikelen ten minste één statistische inconsistentie bevatte. In 12,5% van de artikelen zat ten minste één inconsistentie rond $p = 0,05$ die invloed gehad zou kunnen hebben op de conclusie.

De hoge prevalentie van statistische inconsistenties in de wetenschappelijke literatuur geeft aan dat *peer review* blijkbaar niet voldoende in staat is om dit soort fouten eruit te halen. We hebben dus een andere oplossing nodig.

Statcheck als oplossing

Om statistische inconsistenties snel en makkelijk op te sporen, heb ik samen met mijn collega Sacha Epskamp (universitair docent aan de Universiteit van Amsterdam) Statcheck ontwikkeld. Statcheck is een gratis en *open-source R package* met een bijbehorende web app op <http://statcheck.io>. Het werkt heel eenvoudig: je geeft Statcheck een artikel (of, als je het R package gebruikt, een hele map met artikelen), en Statcheck geeft je een tabelletje terug met alle NHST-resultaten en of ze consistent zijn of niet (zie figuur 1).

Achter de schermen werkt het grofweg als volgt. Statcheck zet eerst je bestand om in een kaal tekstbestand. Vervolgens zoekt het in de tekst naar specifieke combinaties van letters en cijfers die duiden op een NHST-resultaat (met zogenaamde *regular expressions*). Met de gerapporteerde testwaarde en vrijheidsgraden herberekent Statcheck de p -waarde en checkt of deze overeenkomt met de gerapporteerde p -waarde. Zo niet, dan wordt een resultaat aangemerkt als een inconsistentie.

Statcheck herkent ongeveer 60% van alle statistieken in psychologiepapers. Verder kan Statcheck corrigeren voor een aantal factoren, zoals inconsistenties die komen door afronden, of door eenzijdige toetsen. Bij het classificeren van inconsistenties behaalt Statcheck een *accuracy* van 96,2% tot 99,9%, afhankelijk van de instellingen (Nuijten et al., 2017).

Voorkomen is beter dan genezen

Wat betreft statistische spelfouten geloof ik sterk dat voorkómen beter is dan genezen. Het is tenslotte makkelijker om je getallen aan te passen als je nog bezig bent met schrijven. Als je stuk eenmaal gepubliceerd is, is het



Figuur 1. Met één druk op de knop upload je een artikel in de statcheck-app (<http://statcheck.io>), die je vervolgens een tabelletje met alle gevonden NHST-resultaten geeft, en of ze consistent zijn of niet

vaak moeilijk om fouten nog te herstellen. Er moet dan in veel gevallen een officiële correctie gepubliceerd worden; iets wat de meeste mensen liever zouden vermijden als het niet nodig is.

Ik denk dat Statcheck een handig hulpmiddel is om fouten te voorkomen. Met name de web app is uitermate geschikt voor tussentijdse zelf-checks. Verder wordt Statcheck gebruikt tijdens het *peer review* proces van een aantal wetenschappelijke tijdschriften. Mochten onderzoekers bij het indienen van hun artikel toch nog wat statistische inconsistenties over het hoofd hebben gezien, dan kan het alsnog snel opgelost worden vóór publicatie.

Statcheck kan ook een belangrijke rol spelen bij wetenschappelijk onderzoek. Met name onderzoek naar het wetenschappelijke systeem zelf ('meta-onderzoek') kan profiteren van Statcheck: het biedt een manier om snel en betrouwbaar informatie over statistische fouten te verzamelen in grote hoeveelheden literatuur. Zo kostte het maar een uurtje of twee om de kwart miljoen p-waarden te verzamelen uit onze steekproef van 16.000 psychologische artikelen.

Vooruitblik

Statcheck heeft in de afgelopen vijf jaar veel ontwikkelingen doorgemaakt. Wat begon als een handig hulpmiddelje voor mijn eigen onderzoek, is uitgegroeid tot een volledig *R package* met een web app die al meer dan 60 000 keer bezocht is en een standaard onderdeel vormt van het *peer review* proces van meerdere tijdschriften. En de ontwikkelingen staan niet stil.

Een van de geplande updates is om Statcheck uit te breiden zodat het meer typen resultaten kan herkennen.

Op het moment vindt Statcheck alleen NHST-resultaten gerapporteerd in de stijl van de American Psychological Association. Bijna alle psychologietijdschriften hanteren deze stijl, maar in andere vakgebieden worden statistische resultaten vaak anders gerapporteerd. Ook zijn NHST-resultaten niet de enige soort statistiek die in theorie gecheckt zou kunnen worden op inconsistenties. Door de zoek-algoritmes uit te breiden, zou Statcheck in meer situaties toepasbaar kunnen zijn.

Verder bekijk ik momenteel of Statcheck misschien wel meer kan doen dan alleen de inconsistenties aanwijzen. Misschien herinner je je Clippy nog? Clippy was een pratende paperclip in Microsoft Office programma's, die suggesties gaf als er iets fout dreigde te gaan: 'Het lijkt erop dat je Caps Lock aan hebt staan. Wil je dit uitschakelen?'. In principe zou Statcheck ook dit soort adviezen kunnen geven: 'Het lijkt erop dat je de decimalen van de p-waarde hebt omgedraaid. Wil je $p = 0,036$ veranderen naar $p = 0,063$ '.

Uiteindelijk zitten er grenzen aan wat Statcheck kan. Zelfs een spellingchecker die ook je grammatica controleert en verbeteringsuggesties geeft, kan er niet voor zorgen dat de argumentatie in je stuk klopt. Dit geldt ook voor Statcheck: Statcheck kan statistische spelfouten opsporen, maar niet beoordelen of je wel de juiste analyse hebt uitgevoerd.

Dat gezegd hebbende: net zoals de meeste mensen hun stukken liever zonder spelfouten opleveren, zullen de meeste mensen ook liever geen statistische inconsistenties in hun stuk willen. En bij dat laatste kan Statcheck zeker een handje helpen.

REFERENTIES

- Nuijten, M.B., Hartgerink, C.H.J., van Assen, M.A.L.M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods*, 48 (4), 1205-1226. doi: 10.3758/s13428-015-0664-2
- Nuijten, M.B., Van Assen, M.A.L.M., Hartgerink, C.H.J., Epskamp, S., & Wicherts, J.M. (2017). The validity of the tool "statcheck" in discovering statistical reporting inconsistencies. *Preprint retrieved from <https://psyarxiv.com/tcxaj/>*.

MICHÈLE NUIJTEN is in 2018 gepromoveerd op problemen rondom replicatie, statistische fouten, en bias in psychologisch onderzoek. Ze ontving voor haar promotieonderzoek de dissertatieprijs van Tilburg University. Momenteel werkt Michèle als universitair docent bij de afdeling Methodologie & Statistiek aan Tilburg University. In haar huidige onderzoek werkt ze aan efficiënte manieren om de repliceerbaarheid van resultaten in te schatten en te verbeteren. E-mail: m.b.nuijten@uvt.nl.



Foto: wgbieba via Pixabay

'BALGESNAK' IN CORONATIJD

Betaald voetbal kan spannender en sneller

De bal ligt stil. Staafdiagrammen en grafieken markeren het dodelijk spoor dat het coronaspook trekt door steden en dorpen. Menselijk leed wordt vertaald in kille statistieken. Daarbij worden ons dagelijks landenklassementen voorgeschoteld van aantallen geïnfecteerden en aantallen slachtoffers, die in de media worden begeleid door toekomstastend commentaar van neo-BNers. *Lockdowns*, mondkapjes, *intensive care's*, anderhalve-meters en corona-apps zijn de sleutelwoorden in de discussies over onzekere toekomsttrends. Nooit eerder is een ramp zo nadrukkelijk *real time* vertaald in statistieken en onzekerheidsmarges.

Natuurlijk verdringen in deze wereldomvattende crises de coronastatistieken alle sportstatistieken: sporten in competitieverband is te riskant en tot nader order verboden. Ook alle voetbalcompetities liggen stil. Elf spelers op een voetbalveld met minimaal anderhalve meter afstand tussen elk tweetal zal bij de aftrap nog gaan, maar zodra het elftal gaat rennen en vechten om de bal is deze beperking praktisch gezien *infeasible*.

Het is medio april ten tijde van het schrijven van deze column. De curve door de toppen van de staafdiagrammen vertoont een dalende tendens. De stemming achter de schermen slaat om. Plots duikt in de media het woord 'snakken' op. De sportBN-ers hebben bij brede lagen van de bevolking een 'snakken naar' de bal ontwaard. Balgesnak, kortom. De aanpalende discussies hebben na-

tuurlijk wel een probleem nodig en dat is er. Wat is het geval? De huidige voetbalcompetitie was nog lang niet klaar toen de *corona-lockdown* begon. Er moesten op dat moment nog negen speelrondes worden afgewerkt in het betaald voetbal. Ergo, er was en is nog niemand kampioen. Terwijl de *lockdown* al bijna vijf weken duurt, is het nog volstrekt onzeker of vóór augustus het virus uitgewerkt is. Het spookt nog heftig rond en wel zodanig dat de onzekerheidsmarges over de datum van de neutralisatie ervan zo groot zijn dat iedere oplossingskeuze morgen onbruikbaar kan zijn.

Je zou dit onaffe voetbalseizoen simpelweg kunnen 'vergeten', de eindstanden van het seizoen 2018/2019 als uitgangssituatie nemen voor het seizoen 2020/2021 en dit verloren seizoen een plek geven in de sportgeschiedenis: een monument als aandenken aan deze gemankeerde periode. Mijn Tilburgse collega vindt dit een slecht idee: Willem II heeft fantastisch gepresteerd dit seizoen en het is maar de vraag of

Er zijn vast veel betere ideeën, hoewel elk goed idee zijn eigen zuurpruim kent. 'Zelfs een oorlog breekt fossiele zuilen niet af', hoorde ik zo'n pruim zeggen. Dit tijdsgewricht leent zich uitstekend om na te denken over de toekomst, ook die van de sport en zeker die van het voetbal, voorheen mijn favoriete kijksport. De voetbalverslagen zijn veelal interessanter dan de wedstrijden zelf. Vooral die van Willem Visser en dan vooral zijn columns

in *de Volkskrant*. Heerlijk vind ik zijn neo-literair sporttaalgebruik, zoals op 13 april jongstleden: 'Terwijl je hoopt op liefde en licht, op licht vooral, zetten voetbalpraters en columnisten het voetbal in een surrealistisch licht-laaie' en 'Sport geeft het leven de aangename lichtheid die het nodig heeft, zeker nu zwaarte drukt op het gemoed.' Heerlijk.

Dat de tribunes elk weekend grote lege plekken vertonen, daar hoor je nauwelijks over. Vernieuwingsvoorstellen om het spel aantrekkelijk te maken voor de nieuwe generaties worden weggehoond. En dan de voetbalcompetitieopzet: een fossiel uit prehistorische tijden. Die duurt echt superlang voor hedendaagse begrippen: begint in augustus en eindigt op z'n vroegst in mei van het dan volgende jaar. Voor zover in dezelfde divisie, speelt ook nog iedere club tegen iedere andere club. Zelfs twee keer: uit en thuis. 34 speelrondes lang, zonder dat er onderweg in welke vorm dan ook een balans wordt opge maakt. Het hele seizoen zien kleintjes als VVV en Emmen uit naar de komst in hun eigen stadionnetje van het grote Ajax en naar, het misschien wel grotere, Feyenoord. De andere wedstrijden worden alleen echt trouw bezocht door de, te vaak racistisch razende, harde-kerners. Maar zelfs bij zulke 'topwedstrijden' zitten hun stadionnetjes niet vol. (VVV: van de 13 thuiswedstrijden dit seizoen slechts één uitverkocht.)

Tour de France-etappes van minimaal 200 kilometer waren geen uitzonderingen. Het jaar 1903 kende de langste etappe met 471 kilometer. Volleybalwedstrijden en schaakwedstrijden moesten worden afgebroken, omdat de noodzakelijke beslissende marges de nacht in zouden gaan. Het wereldrecord sprintend stilstaan, surplace genaamd, werd in 1968 gevestigd met 63 minuten. De UCI heeft onlangs de maximale surplace-tijd verkort tot een halve minuut. Ik schat dat mijn kleinzoon van 14 die halve minuut niet uitkijkt; bij Champions Leaguewedstrijden is

hij vooral in de weer met zijn FIFA-voetbalspel. Het kon toen allemaal niet lang genoeg duren. Hoe langer, hoe spannender. Verlengingen van voetbalwedstrijden, omdat na 90 minuten – waarvan meestal bijna 40 minuten surplace-voetbal waren en zijn – de beslissing niet was gevallen, heerlijk vonden we dat vroeger. Kom daar maar eens om in deze tijd.

We lachen vaak wat om onze zuiderburen. Maar ze hebben wel hun voetbalcompetitie ingericht met meer oog voor het noodzakelijke hedendaagse spektakel. De Belgische eredivisie is opgeknipt in twee periodes. De eerste periode kent een volledige competitie tussen alle clubs. Daarna is er een periode met een volledige competitie tussen de zes hoogst gerangschikten in de eindstand van de eerste periode. Bij de start van deze *play-offs* wordt de helft van de puntenaantallen van elk van die zes clubs, behaald in de eerste periode, meegenomen naar de tweede periode. Goed idee. Op deze wijze hebben die zes in de tweede periode een voordeel van hun score uit de eerste periode. Omdat dan met 'halven' zou moeten worden gewerkt, worden die puntenaantallen naar boven afgerond. Deze opzet en de bijbehorende puntentruc is bedacht door drie *operations research*-collega's; zie de referentie. De Belgen snappen het.

Ik waag een poging om de Nederlandse betaaldvoetbalcompetitie spannender en ook korter te maken. Net als bij de Tour de France, waar die verschrikkelijk saaie lange etappes zijn opgeknipt in kortere – die vaak ook nog te lang en te saai zijn – zou de huidige competitie van zowel de Eredivisie als de Eerste Divisie 'samengevoegd' moeten worden en het geheel opgesplitst in zes deeldivisies van elk zes clubs. (De Eerste Divisie zou twee clubs minder moeten bevatten). De gehele betaaldvoetbalcompetitie bestaat dan uit drie speelperiodes, waarin telkens iedere deeldivisie een volledige competitie afwerkt van dus 10 wedstrijden. De indeling in deeldivisies is op

basis van de eindstand van het vorige seizoen. Na elke periode promoveren twee clubs en degraderen er twee tussen opeenvolgende divisies. Elke periode beginnen de clubs met nul punten. Als je lang nadenkt, ontdek je vast wel een aantal minpunten in vergelijking met de huidige competitieopzet. Ik vertel die minpunten even niet. Maar één ervan kun je voorkomen door bij de overgang van de tweede naar de derde periode niet twee maar één club tussen de beide opeenvolgende divisies te laten promoveren en degraderen. En de echte toppers, die zich alle drie periodes hebben gehandhaafd in de Elitedivisie, die moeten tijdens de eerste twee periodes ook een worst ruiken. Zij nemen, net als bij de Belgen, de helft van de punten uit de eerste twee periodes mee naar de derde (wat mij betreft: naar boven afgerond).

Ik noem wel een aantal voordelen. Ten eerste doen de meeste wedstrijden 'ertoe'. Binnen de zes divisies zijn de clubs zeer competitief en ertussen heerst veel dynamiek: in de 'middendivisies' is in elke periode de samenstelling op vier plekken veranderd. Bovendien duurt in deze vorm de totale competitie geen 34, maar 30 speelrondes. Ook is er meer flexibiliteit. Mocht het voetbalseizoen, om wat voor reden dan ook, worden afgebroken, dan is er altijd de eindstand van een volledige (sic!) periodecompetitie. Het seizoen 2020/21 zou prima kunnen starten met deze nieuwe opzet, waarbij de indeling in de zes deeldivisies zou kunnen worden gedaan op basis van de huidige standen. Dan heeft Willem II niet voor niets zo goed gepresteerd in dit corona-seizoen.

POSTSCRIPTUM

Een bevriende collega, die een vorige versie van deze column kritisch heeft doorgelezen, wees mij op het feit dat we nu toch ook al twee periodes hebben: één voor en één na de winterstop. Dat klopt. Alleen zijn dat allebei 'halve' competities. Daarom stelt die zogenaamde winterkampioen ook niks voor. In een voetnoot meldde hij dat zijn huidig balgesnak zich niet beperkt tot zijn favoriete gehaktbal in de pauze van de thuiswedstrijden van zijn FC Emmen (van de 13 thuiswedstrijden nog maar twee uitverkocht). Balgesnak en gehaktbal in coronatijd.

LITERATUUR

- Mossou, S. (2020). Spannender, korter en flexibeler. *Algemeen Dagblad*, 25 mei 2020.
- Goossens, D.R., Beliën, J., & Spieksma, F.C.R. (2012). Comparing League Formats with respect to Match Importance in Belgian Football. *Annals of Operations Research*, 194, 223–240.

GERARD SIERKSMA is emeritus hoogleraar Kwantitatieve Logistiek en Sportstatistiek aan de Rijksuniversiteit Groningen. E-mail: g.sierksma@rug.nl



Bekroning van uitzonderlijke afstudeerprestaties

Tijdens de Annual Meeting op 12 maart 2020 zijn de Jan Hemelrijk Award 2019 en de Willem R. van Zwet Award 2019 uitgereikt.

JAN HEMELRIJK AWARD 2019

Ter bekroning van een uitzonderlijke afstudeerprestatie aan een Nederlandse instelling voor wetenschappelijk onderwijs/hoger beroepsonderwijs looft de VVSOR een scriptieprijs uit, de Jan Hemelrijk Award.

Er waren vijf inzendingen waaruit de jury de volgende winnaar heeft gekozen:

Willem Feijen voor de Master-scriptie *Computing Optimal Classification Trees*.

De jury voor de Hemelrijk Award 2019 bestond uit:

- Prof. dr. Eric Cator (voorzitter, Radboud Universiteit Nijmegen)
- Dr. Frank Phillipson (TNO)
- Dr. Herman Monsuur (NLDA)
- Dr. David Tax (TU Delft)
- Dr. Ad Ridder (VU Amsterdam)

WILLEM R. VAN ZWET AWARD 2019

Sinds 2012 reikt de VVSOR ook een prijs uit voor dissertaties: de Willem R. van Zwet Award.

Er waren vier inzendingen; de jury koos als winnaar:

Jesse Hemerik voor het proefschrift *Permutation-based Inference for High-dimensional Data*.

De jury voor de Van Zwet Award 2019 bestond uit:

- Prof. dr. Mark de Rooij (voorzitter, Universiteit Leiden)
- Prof. dr. Bernd Heidergott (VU Amsterdam)
- Prof. dr. Hendrik Boshuizen (Wageningen UR)
- Prof. dr. Fred van Eeuwijk (Wageningen UR)



Tweemaal surplace: nu en toen



Toine Mazairac (r) en Joep Meyer op het WK sprint van 1925

NS meet wel degelijk de reizigerspunctualiteit

Reactie op de column 'Weet wat je meet'



MERITH PELGER EN DENNIS HUISMAN

Onterecht denkt Gerrit Stemerding, de auteur van de column 'Weet wat je meet' in de vorige editie van *STATOR*, dat NS niet op reizigerspunctualiteit meet. Gelukkig krijgen we van de redactie van *STATOR* de kans hierop te reageren.

Als NS publiceren we niet alleen vol trots onze punctualiteitscijfers, maar is de afgelopen jaren de methodiek om de punctualiteit te meten flink verbeterd. Zo suggereert de columnist dat de kernopdracht van NS is om reizigers punctueel te vervoeren. Daar heeft hij natuurlijk helemaal gelijk in. Daarom rekent het ministerie van I&W, NS (en overigens ook ProRail) in de huidige concessie af op reizigerspunctualiteit. Deze concessie is ingegaan op 1 januari 2015. We ondersteunen het pleidooi van de columnist dan ook van harte. Middels deze reactie willen we echter zowel de columnist als u als lezer van het mooie blad *STATOR* er op wijzen dat het pleidooi van de columnist al jaren geleden in de praktijk is gebracht.

Reizigerspunctualiteit geeft een indicatie van het percentage weer van de reizen dat met minder dan χ minuten vertraging is verlopen. Een reiziger heeft dus bij aankomst op zijn uitcheckstation minder dan χ minuten vertraging gehad ten opzichte van de reis die de reiziger had kunnen maken vanaf het moment van inchecken volgens de NS Reisplanner. Als een reiziger een overstap maakt gedurende zijn of haar reis of als een trein uitvalt, dan wordt dit dus meegenomen in de berekende vertraging, er wordt immers gekeken naar de aankomst op het eindstation.

In de concessie is afgesproken om over zowel over de 5 minuten punctualiteit als over de 15 minuten punctualiteit cijfers te rapporteren. Deze gegevens kunt u dus lezen op de website van NS (<https://dashboard.nsjaarverslag.nl/hrn>). Zo kunt u terugvinden dat over het jaar 2019, 92,6% van de reizigers minder dan 5 minuten vertraging had en 97,7% minder dan 15 minuten. Deze berekening wordt gemaakt op basis van geanonimiseerde check-in en check-out gegevens, gerealiseerde uitvoeringsdata, en uiteraard de belofte die NS doet middels de NS Reisplanner.

Is dan alles perfect bij de punctualiteitsmetingen van NS? Nee, natuurlijk niet. We blijven ons ook de komende jaren inzetten om zowel de reizigerspunctualiteit beter te meten, als ook om beter te sturen op diezelfde punctuali-

teit, met als uiteindelijke doel om reizigers een nog beter product te bieden. Hiervoor werkt NS volop samen met wetenschappelijke instellingen en willen we het gebruik van technieken uit onder andere de Statistiek en Operations Research verder bevorderen middels ons initiatief binnen het programma "Kickstart AI". Hierin werken we met vier andere, grote Nederlandse bedrijven samen om het wetenschappelijk onderzoek in Nederland te versterken en de toepassing van dat onderzoek te bevorderen. We hopen daarover in de toekomst meer in dit blad te kunnen schrijven, want als deze column ons één ding heeft geleerd, dan is het wel dat niet iedereen even goed op de hoogte is van de meest recente ontwikkelingen bij NS.

Merith Pelger is senior business consultant bij de afdeling Prestatieregie & Innovatie van NS, en verantwoordelijk voor de KPI Reizigerspunctualiteit binnen NS.

Dennis Huisman is expertise manager logistieke processen bij de afdeling Prestatieregie & Innovatie van NS. Tevens is hij als bijzonder hoogleraar verbonden aan het Econometrisch Instituut van de Erasmus Universiteit Rotterdam.

Reactie Gerrit Stemerding

Ik ben erg blij met de reactie hierboven op mijn recente column. En wel om twee redenen.

Ten eerste blijkt hieruit dat ons blad aandachtig gelezen wordt, zo iets is voor een redacteur altijd prettig om te vernemen.

En ten tweede blijkt mijn hartenwens van tientallen jaren al enkele jaren geleden gerealiseerd te zijn. Al tientallen jaren? Ja, zo'n 25 jaar geleden had het toenmalige VVSOR-blad Kwantitatieve Methoden een onregelmatig verschijnende rubriek 'wat ik nu toch weer zag in de media'. Naar aanleiding van een krantenbericht over het percentage treinen dat op tijd reed heb ik toen een soortgelijke reactie geschreven.

*Als fervent treinreiziger kan ik alleen maar zeggen: Merith en Dennis, bedankt! En jullie aankondiging om in de toekomst meer voor *STATOR* te schrijven is stipt op tijd, om in stijl te blijven. Reeds dit nummer bevat zo'n artikel over 'Rekenen aan maatregelen om drukke treinen te ontlasten'.*



Daniel Dadusch en Marloes Maathuis ontvingen de Van Dantzigprijs 2020

Op 12 maart 2020 is voor de elfde keer de Van Dantzigprijs uitgereikt. De Van Dantzigprijs is de hoogste Nederlandse onderscheiding op het gebied van statistiek en besliskunde. De VVSOR reikt eenmaal in de vijf jaar de prijs uit aan een persoon niet ouder dan 40 jaar die in Nederland werkt of de Nederlandse nationaliteit heeft en die bijzondere theoretische of praktische bijdragen heeft geleverd aan statistiek of besliskunde. De leeftijdsclausule dat de winnaar niet ouder mag zijn dan 40 jaar is dit jaar toegepast met een extensie-clausule zoals ook gebruikt door het NWO. De jury bestaande uit Hein Putter, Marieke Timmerman, Aad van der Vaart (voorzitter), Harry van Zanten en Bert Zwart is op grond van nominaties en eigen onderzoek unaniem tot de conclusie gekomen dit keer de prijs uit te reiken aan twee onderzoekers: Daniel Dadusch en Marloes Maathuis



Daniel Dadusch, geboren in 1982, bezit de Franse en de Amerikaanse nationaliteit, en is thans werkzaam bij het Centrum voor Wiskunde en Informatica in Amsterdam, op het gebied van de theoretische besliskunde, optimalisatie, en algoritmische meetkunde. Hij ontvangt

de prijs in het bijzonder voor zijn recente werk aan het begrijpen van de efficiëntie van de simplex methode voor lineaire programmering, en het afleiden van grenzen voor optimalisatie algoritmen voor het vinden van geheeltallige oplossingen en in roosters. Daniel ontving eerder de A.W. Tuckerprijs voor zijn proefschrift geschreven aan Georgia Tech (2012) bij Vempala, en behaalde een Simons fellowship, een Veni subsidie en een ERC starting grant. De nominatoren schrijven: 'Several of the problems that Daniel resolved were foundational and notoriously hard

questions, that had inspired intense research activity over the last 30 years within many different areas including operations research, computer science, geometry and cryptography. A key aspect of Daniel's work is to look at all these classic problems using highly novel and creative insights and ideas, and combining them with several deep and unexpected techniques from [various areas of mathematics].'



Marloes Maathuis, geboren in 1978, heeft de Nederlandse nationaliteit, maar is werkzaam aan ETH in Zürich op het gebied van de hoog-dimensionale statistiek, causale inferentie en toepassingen van de statistiek in andere gebieden, waaronder genetica. Zij ont-

vangt de prijs in het bijzonder voor haar werk aan grafische modellen voor het trekken van causale conclusies in het geval sommige variabelen niet zijn gemeten, en haar onderzoek naar de computationele en statistische consistentie van hybride methoden voor het ontdekken van causale structuur tussen variabelen. Marloes promoveerde in 2006 aan de University of Washington (Seattle) bij Groeneboom en Wellner op een onderwerp in de survival analyse, en werd daarna docent aan ETH, waar zij een wereldexpert werd op het gebied van causaliteitsanalyse, en inmiddels tot hoogleraar is bevorderd. Een referent merkt op: '[Marloes is] sharp minded, [has] great communication skills and [is] a great scholar who aims for quality. The impact of her research – whatever that is – is big and on a very steep gradient upward.' en over haar werk aan latente variabelen 'The problem is notoriously difficult... the authors make a very substantial contribution.'



GERRIT STEMERDINK



column

Grafsteen van Max Born op het Stadtfriedhof Göttingen. Foto: Julian Herzog CC

GEKKEN EN DWAZEN

Een Nederlands spreekwoord luidt ‘Gekken en dwazen schrijven hun namen op deuren en glazen’. Men zegt ook dikwijls dat genialiteit grenst aan gekte en exacte wetenschappers zijn soms geniaal. Zouden zij ook her en der hun naam achterlaten? Ik denk het niet, al zag ik ooit in het oude Natuurkundig Laboratorium van de UvA een statief met daarop een plaatje geschroefd ‘eigendom P. Zeeman, afblijven’. Wel vinden we hen na hun dood soms terug in namen van scholen, instituten, prijzen, straten en pleinen.

De vraag kwam bij me op toen ik onlangs de autobiografie van de fysicus Max Born herlas. Ik zag weer de foto van zijn grafsteen met daarop de beroemde formule $p \cdot q - q \cdot p = \hbar / 2\pi i$. Ik ben eens verder gaan zoeken en het blijkt dat vrij veel wiskundigen en fysici een formule op hun grafsteen hebben. En als het niet op hun grafsteen is dan is het wel op een gedenkplaat voor hen. Om er maar een paar te noemen: Jakob Bernoulli, Stephen Hawking, Ludwig Boltzmann, Julian Schwinger, Georg Cantor, Ernst Abbe, Paul Dirac, Erwin Schrödinger. Het dichtst bij ons eigen vakgebied komt een prachtige tweetalige plaquette waarop wordt vermeld dat William Gosset, de Ierse statisticus én bierbrouwer, 22 jaar op Hollyville Park in Dublin heeft gewoond. Hij publiceerde onder de naam Student en ontwikkelde de t-toets. De kansverdeling daarvan is dan ook op de plaquette afgebeeld. Overigens is het niet aannemelijk dat al deze geniale wetenschappers zelf hebben besloten

‘op deuren en glazen’ te schrijven, het zullen veelal hun nabestaanden zijn geweest die hiertoe het initiatief namen.

Weliswaar geen grafsteen, maar wel verwant, is het verhaal van William Rowan Hamilton. Die maakte dagelijks een wandeling door Dublin en omstreken. Op een dag viel hem de formule voor de quaternionen in en om te voorkomen dat hij hem zou vergeten sneed hij die, volgens het verhaal, met zijn pennenmesje in de leuning van de brug. Nu is er een gedenksteen met die formule in die brug aangebracht en is dat een min of meer verplicht bedevaartsoord geworden voor wiskundigen die Dublin bezoeken.

Maar ook in Nederland hebben wij zo iets moois! In de Pieterskerk in Leiden ligt Ludolph van Ceulen begraven. Hij was de eerste die het getal π tot in 35 decimalen berekende. Enkele jaren geleden is de verloren gegane grafsteen, waarop zijn berekening van π was afgebeeld, vervangen door een plaquette ter herinnering aan hem en op de internationale π -dag (14 maart: 3.14) onthuld.

Hoe is het met u, beste lezer? Zou u voor een formule kiezen, of volgt u liever het voorbeeld van de dichter Gerrit Komrij die ooit voorstelde de volgende tekst op zijn graf te plaatsen: ‘Hier ligt Gerrit Komrij. Ik denk dat ik omrij’.

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



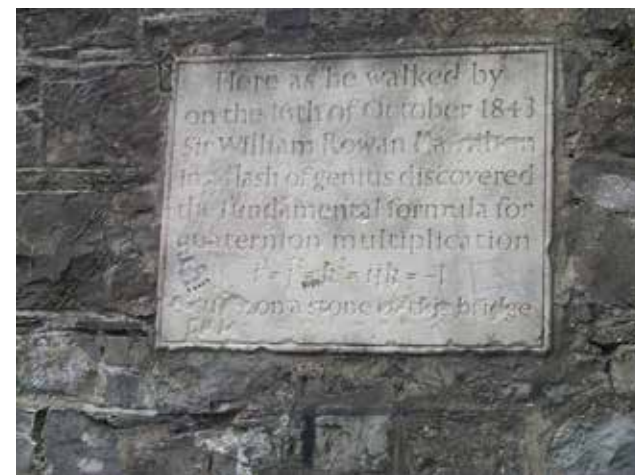
Monument voor Ernst Abbe in Jena. Foto: Daniel Mietchen CC



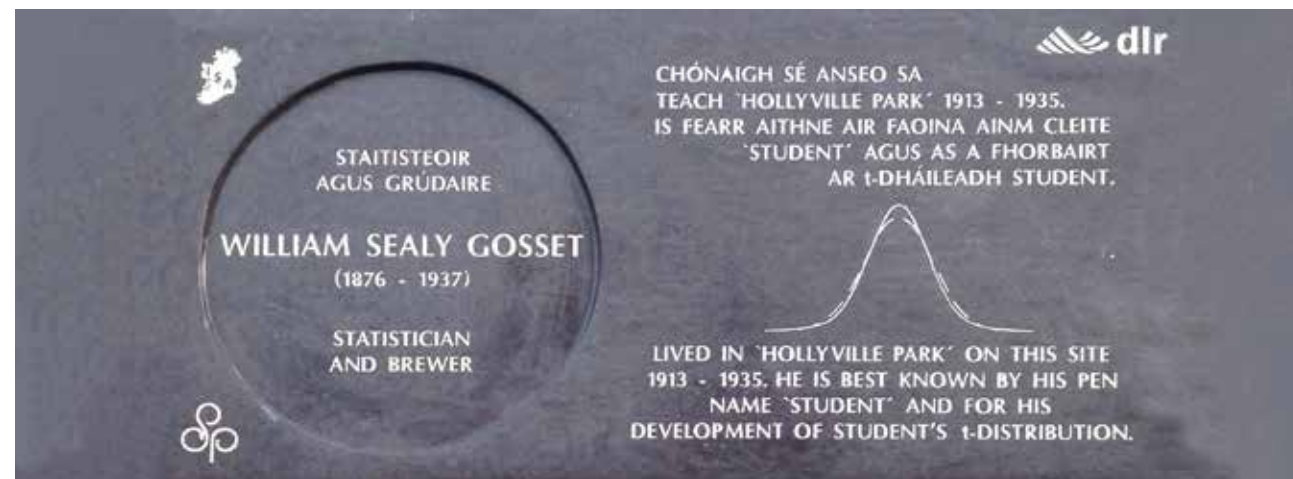
Grafmonument voor Ludwig Boltzmann in Wenen



Grafsteen voor Julian Schwinger in Cambridge (Massachusetts)



Plaquette voor William Rowan Hamilton (Dublin). Foto: Wisher CC



Plaquette voor William Gosset op Hollyvillepark in Dublin



Graf van Erwin Schrödinger in Alpbach (Oostenrijk)



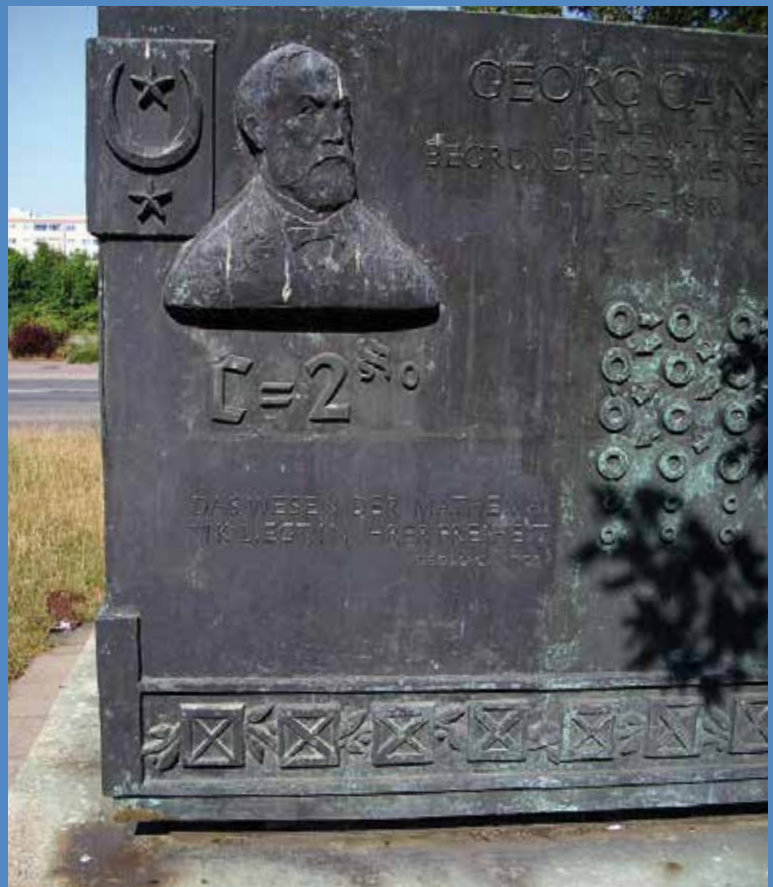
Grafschrift voor Jakob Bernoulli (met zijn beroemde spiraal) in de Munster van Bazel. Foto: Wladyslaw Sojka CC



Graf van Paul Dirac in Westminster Abbey. Foto: Stanislav Kozlovskiy CC



Graf van Stephen Hawking in Westminster Abbey, het graf ligt vlak tegenover dat van Isaac Newton



Monument in de vorm van een bronzen kubus in Halle waarmee onder anderen Georg Cantor wordt geëerd