STAtOR

periodiek van de VvS+OR jaargang 17, nummer 3, december 2016

thema **SMART SOCIETIES**

STAtOR

Jaargang 17, nummer 3, december 2016

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

Redactie
Joaquim Gromicho (hoofdredacteur), Annelieke Baller, Ana Isabel Barros, Joep Burger, Kristiaan Glorie, Guus Luijben (eindredacteur), Richard Starmans, Gerrit Stemerdink (eindredacteur) en Vanessa Torres van Grinsven. Vaste medewerkers: Johan van Leeuwaarden, Fred Steutel en Henk Tijms.

Kopij en reacties richten aan
Prof. dr. J.A.S. Gromicho (hoofdredacteur), Faculteit der Economische Wetenschappen en Bedrijfskunde, afdeling Econometrie, Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, telefoon 020-5986010, mobiel 06-55886747, <j.a.dossantos.gromicho@vu.nl>.

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

Leden- en abonnementenadministratie van de VvS+OR
VvS+OR, Postbus 1058, 3860 BB Nijkerk, telefoon 033 - 2473408, e-mail <admin@vvs-or.nl>.
Raadpleeg onze website over hoe u lid kunt worden van de VvS+OR of een abonnement kunt nemen op *STAtOR*.

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

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

Ontwerp en opmaak
Pharos, Nijmegen

Uitgever

ISSN 1567-3383

thema SMART SOCIETIES

overig

Uitmuntende Master's of Ph.D. thesis begeleid?

OPROEP om kandidaten te nomineren voor de Jan Hemelrijk Award 2016 en de Willem R. van Zwet Award 2016

Ter bekroning van een uitzonderlijke afstudeerprestatie aan een Nederlandse instelling voor wetenschappelijk onderwijs of hoger beroepsonderwijs looft de VvS+OR al sinds 1989 een scriptieprijs uit. Sinds 2014 staat deze bekend als de Jan Hemelrijk Award. Sinds 2012 is er ook een prijs voor dissertaties: de Willem R. van Zwet Award.

De VvS+OR roept op tot nominaties voor deze prijzen. Beide prijzen bestaan uit een oorkonde en een geldbedrag van 1000 euro. De prijzen zullen worden uitgereikt op de jaarlijkse VvS+OR dag, op donderdag 23 maart 2017.

Genomineerd kunnen worden personen die tussen september 2014 en september 2016 zijn afgestudeerd resp. gepromoveerd en die nog niet eerder zijn genomineerd.

Hierbij worden supervisors (begeleiders) opgeroepen om een uitmuntende afstudeerscriptie (Master) of dissertatie (Ph.D.) te nomineren voor de Jan Hemelrijk dan wel Willem R. van Zwet Award 2016.

De indiening van een nominatie dient vergezeld te gaan van een aanbevelingsbrief van de supervisor van de genomineerde. De precieze procedure voor beide prijzen, alsmede de reglementen en het nominatieformulier zijn te downloaden op de website van de VvS+OR, <www.vvs-or.nl>. De nominatie dient binnen te zijn vóór **10 januari 2017**.

Namens de VvS+OR,
Prof. dr. Eric Cator *(Juryvoorzitter Jan Hemelrijk Award)*, prof. dr. Jelle Goeman *(Juryvoorzitter Willem R. van Zwet Award)* & drs. Sander Scholtus *(secretaris der beide jury's)*

January 17 - 19, 2017, Lunteren, The Netherlands
42th CONFERENCE on
the Mathematics of Operations Research

The LNMB ((Landelijk Netwerk Mathematische Besliskunde) conferences 'The mathematics of operations research' form the annual meeting of Dutch senior and junior researchers in the Mathematics of Operations Research. The program offers presentations on high-quality research and applications and should appeal to both academic researchers and to management consultants in trade and industry.
List of international speakers: Constantinos Daskalakis (Massachusetts Institute of Technology, USA), Matteo Fischetti (University of Padova, Italy), Edward H. Kaplan (Yale School of Management, USA), Kurt Mehlhorn (Max Planck Institute for Informatics, Germany)
For information see www.lnmb.nl/conferences/2017/

Thursday January 19, 2017, 10.00-17.00
Lunteren, The Netherlands
NGB/LNMB Seminar Back to school, learn about the latest developments in Operations Research
Topic: Developments and applications in the area of Game Theory

Noted for the first time in literature in 1713, game theory blossomed really starting from the 1930s. The prisoners' dilemma is dating from the 1950s and that is about the same time John Nash did his most famous work that would deliver him the Nobel prize 44 years later. But since then the research has not stopped. The wide application area of game theory include economics, biology, philosophy, politics, and computer science. In today's world where resources, supply chains, and profit are shared, game theory plays an important role in determining what is 'fair'. But also in making sure that government policies have the desired effect, or in how market leaders can influence the market developments, game theory can be used. In another area, think about the Internet of Things, how individual sensors and components measure, make decisions, act, and interact, all based on rules and game theory principles. Game theory is used more than most of us are aware.
This year, at the 3rd day of the annual LNMB congress in Lunteren, we will further explore the developments and applications of the game theory area of expertise.
For information see www.lnmb.nl/conferences/2017/

Smart Societies

Het thema van dit dikke nummer is Smart Societies, dat zijn samenlevingen die op een slimme en innovatieve manier gebruik maken van gegevens over allerlei zaken binnen die samenleving.

Het is op deze plaats al eens eerder gesteld: Statistiek heeft een oeroude oorsprong. Tot de eerste vormen van gegevens vastleggen behoort een soort inventarisatie van goederen zoals schapen of manden graan. Toen onze verre voorouders zich van jagers-verzamelaars tot landbouwers hadden ontwikkeld, kregen zij behoefte aan dit soort informatie. Men legde dat vast met kerfjes in beenderen of stokken, of met indrukken in klei. Daarmee staat statistiek in feite aan de oorsprong van het schrift.

Toen er grotere gemeenschappen als steden en landen ontstonden bleef die behoefte groeien, we kennen vele voorbeelden uit Egyptische papyrussen en Mesopotamische kleitabletten. Het Nieuwe Testament begint zelfs met een verhaal over een volkstelling in het Romeinse Rijk. En rond 1500 ontstond de term statistiek met als betekenis: de toestand van de staat.

Tegenwoordig is statistiek niet meer weg te denken uit een moderne samenleving: hoe kunnen we bijvoorbeeld verstandig plannen en er voor zorgen dat we over zeg 15 jaar voldoende capaciteit op vervolgopleidingen hebben zonder een goed inzicht in de bevolkingsopbouw.

Maar statistiek kan in deze tijd veel meer. Er worden dagelijks vele terabytes aan gegevens gegenereerd, afkomstig van mobiele telefoons, in- en uitcheck apparaten in het openbaar vervoer, elektronische betalingen, bewakingscamera's, uitslagen van medische onderzoeken etc. Die overweldigende hoeveelheid gegevens, vaak big data genoemd, biedt veel mogelijkheden om regelend op te treden in een samenleving. Zo is bij Sail 2015 in Amsterdam gebruik gemaakt van allerlei mobiliteitsgegevens om de grote hoeveelheid bezoekers via verschillende routes te leiden en daarmee opstoppingen te vermijden. En we kennen allemaal de alternatieve routes die onze GPS-systemen suggereren bij grote files op de snelwegen.

U vindt in dit nummer artikelen over heel diverse onderwerpen. Bijvoorbeeld over het slim plannen van het legen van afvalcontainers – waardoor in Twente twee vuilniswagens bespaard worden – maar ook over een betere screening op darmkanker of over het optimaal plannen van het transport van de enorme hoeveelheden contant geld die vervoerd worden van en naar geldautomaten, banken en bedrijven.

Statistiek en OR zijn onontbeerlijk bij het gebruik van al deze gegevens, wij mogen best eens wat meer naar voren treden met alles wat we met ons vakgebied bijdragen aan de samenleving. Misschien is het een goed idee als het VvS+OR-bestuur dit nummer, en daarmee ons vakgebied en onze vereniging, onder de aandacht brengt van ministers en kamerleden.

De big data zijn er, we hoeven ze alleen maar op een goede manier te analyseren en gebruiken. Op naar de Smart Society!

Naast alle aandacht voor ons thema vindt u in dit nummer verenigingsnieuws, bijdragen van onze vaste columnisten, de rubriek Klassiekers en helaas enkele artikelen waarin de overleden leden Paul van der Laan en Leo Kroon worden herdacht.

Ten slotte kunnen we melden dat wij blij zijn met het besluit van het bestuur dat we met ingang van 2017 weer vier maal per jaar kunnen verschijnen. We hopen nog vele nummers te mogen maken, met interessante verhalen uit de volle breedte van ons vakgebied. Veel leesplezier!

De redactie

Foto: Verkeersinformatiedienst (VID)

SMART ROUTING

PRAKTIJKPROEF AMSTERDAM – METHODE EN EVALUATIE

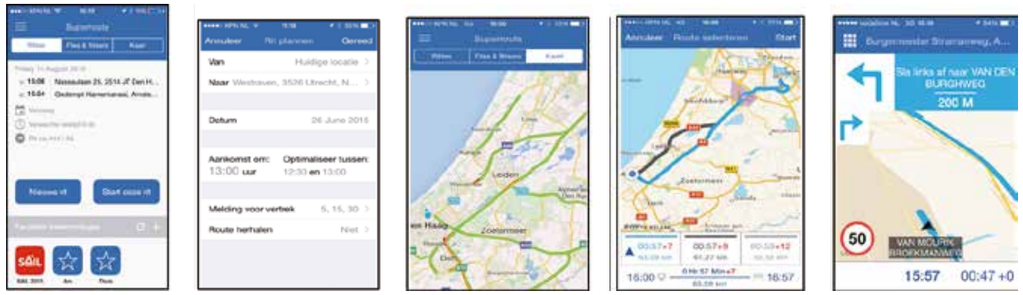
Op wegen in en tussen stedelijke gebieden wordt vaak veel verkeer afgewikkeld. Een gevolg hiervan is dat er verwacht en onverwacht reistijdverlies ontstaat. Diverse maatregelen kunnen worden ingezet om reistijdverlies tegen te gaan wat van cruciaal belang is voor steden. Dit artikel richt zich op het spreiden van het verkeer door slim te routeren (smart routing) en door het geven van vertrektijdstip-advies. In het kader van de PraktijkProef Amsterdam (PPA) is een grootschalige test uitgevoerd met deze maatregelen. Het is een mooi voorbeeld van een 'Smart City oplossing' gericht op doorstroming waarbij veel verschillende databronnen van verschillende partijen aan elkaar zijn gekoppeld.*

Maaike Snelder

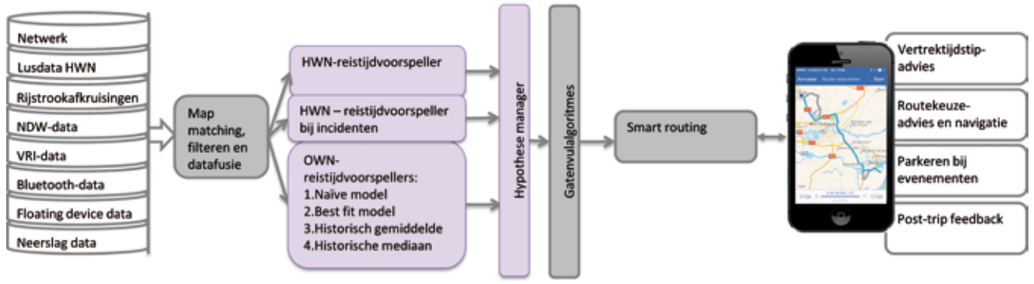
De PraktijkProef Amsterdam (PPA) is een grootschalige test waarbij innovatieve technieken gebruikt worden met als doel het verminderen van de files in de regio Amsterdam. In de eerste fase van de PraktijkProef zijn systemen langs de weg en in de auto afzonderlijk getest. Dit artikel zoomt in op een service die in de auto is aangeboden via de Superroute app en Super P-route app.

De Superroute app (voor regulier verkeer) en Super P-route app (voor evenementenverkeer) geven pre-trip vertrektijdstipadvies en routeadvies. Een extra functionaliteit van de Super P-route app is parkeeradvies en de mogelijkheid van tevoren een parkeerplaats te reserveren bij evenementen. Figuur 1 toont een aantal screenshots van de Superroute app. In de app worden meerdere routemogelijkheden gepresenteerd, waarbij aangesloten wordt bij de persoonlijke voorkeuren van de gebruiker. Tijdens een rit bepaalt het zogeheten Smart Routing-algoritme op basis van actuele en voorspelde reistijden voortdurend of er een betere (snellere) route mogelijk is en geeft de app – als dit het geval is – wederom meerdere routemogelijkheden. Bij de selectie van een route vindt een beperkte vorm van *load balancing* (spreiding over het netwerk) plaats door gebruikers niet allemaal over dezelfde route te sturen. Hierbij wordt gebruik gemaakt van persoonlijke voorkeuren van gebruikers en eerder gegeven route-adviezen aan de gebruiker zelf en aan andere gebruikers en van restcapaciteiten op alternatieve routes.

De Superroute en Super P-route app bieden *in-car* reisinformatie en reisadviezen en werken op smartphones en tablets met besturingssystemen iOS en Android, 3G- (of 4G-)verbinding en een netwerklocatie-fix. Tijdens de proefperiode in 2015 werkte de app een jaar lang, 24 uur per dag, 7 dagen per week, in heel Nederland.



Figuur 1. Screenshots Superroute app



Figuur 2. Architectuur data en modellen

Data en modellen voor smart routing

Figuur 2 geeft een overzicht van de data en modellen die zijn ingezet om tot een smart routing advies te komen. Links staan de gebruikte databronnen. Deze databronnen bevatten historische data en de meest recente data van één of enkele minuten geleden. Met behulp van een map-matching-algoritme worden deze data gekoppeld aan een kaart, gefilterd en gefuseerd om een zo goed mogelijk beeld te krijgen van de snelheden op alle wegen, intensiteiten (aantal voertuigen per uur) op alle snelwegen en rijstrookafkruisingen ter indicatie van een incident.

De data worden gebruikt voor het maken van reistijdvoorspellingen (paarse blok in de figuur). Er bestaan veel methodes waarmee reistijdvoorspellingen kunnen worden gemaakt. De methodes richten zich meestal op specifieke omstandigheden. Om deze reden draaien meerdere reistijdvoorspelmodellen gelijktijdig. Voor het hoofdwegennetwerk (HWN – alle Rijkswegen van Nederland, aangevuld met een aantal provinciale wegen) worden voorspellingen voor 15 minuten vooruit tot 3 uur vooruit op basis van een best-fit-voorspelling gemaakt. Voor de kortetermijnvoorspelling voor 0 tot 15 minuten worden de verkeerspatronen (snelheden en intensiteiten) vooruit gepropageerd met behulp van *kinematic wave theory* (Newell, 1993) rekening houdend met de voorspelde verkeerssituatie over 15 minuten. Indien er een incident (rijstrookafkruising) plaatsvindt, treedt de HWN-voorspeller bij incidenten in werking. Dit is een marginaal model dat op basis van de gemeten verkeerssituatie en de capaciteitsreductie met behulp van *kinematic wave theory* bepaalt hoe snel de file achter het incident opbouwt en later weer afbouwt en welke wegen hierdoor worden beïnvloed. De verwachte extra reistijd als gevolg van het incident volgt hier uit. Voor het onderliggend wegennetwerk (OWN) zijn vier OWN-voorspellers ingezet. De naïeve voorspeller scoorde het best voor een korte voorspelhorizon (tot 15 minuten vooruit) en voor situaties zonder congestie. De best-fit-voorspeller scoort het best bij de overgang naar spitsperiodes en de historische-mediaan-voorspeller scoort het best gedurende de overige periodes. De zogenaamde 'hypothesemanager' combineert deze voorspellingen tot één voorspelling. In (Snelder & Calvert, 2015) zijn de algoritmes voor reistijdvoorspellingen en de hypothesemanager in meer detail toegelicht.

Een vergelijking tussen de reistijdvoorspellingen en de gerealiseerde reistijden heeft aangetoond dat deze eenvoudige voorspelmethodes al goed presteren, omdat de relatieve fout over het algemeen kleiner is dan 20%. De architectuur is zodanig opgezet dat eenvoudig andere reistijdvoorspellers kunnen worden toegevoegd en de kwaliteit van de voorspellers die opgenomen zijn in het raamwerk kan worden gemonitord en verbeterd.

Bovengenoemde stappen kunnen worden uitgevoerd voor wegvakken in het wegennetwerk waarvoor data van voldoende kwaliteit beschikbaar zijn. Het is echter ook essentieel om voor de onbemeten wegen reistijdvoorspellingen te genereren omdat anders al het verkeer over de onbemeten wegen wordt gestuurd. Daar wordt immers geen vertraging gemeten. Hiertoe zijn 'gatenvulalgoritmes' ontwikkeld.

De reistijdvoorspellingen zijn de basis voor een slimme routeringsmodule (smart routing) waarmee via de Superroute app en Super P-route vertrektijdstipadvies, routekeuze-advies, advies over te kiezen parkeergarages bij evenementen en feedback aan de gebruiker wordt gegeven.

Evaluatie

Er was veel belangstelling voor de dienst die is aangeboden door Amsterdam Onderweg. Uiteindelijk hebben ruim 28.000 mensen zich via de website aangemeld voor de proef met de Superroute app. Daarvan hebben 21.428 de app op hun smartphone geïnstalleerd, en van 19.865 deelnemers is er minstens één rit geregistreerd. Van de deelnemers gebruikte 15 procent de app minstens wekelijks, en bijna 30 procent minstens tweewekelijks. De overige 55 procent gebruikte de app sporadisch.

De dienst is geëvolueerd van een alleen op on-trip routeadvies gerichte dienst naar een dienst waarin meerdere functies geïntegreerd zijn, waaronder ook pre-trip advies (vertrektijdstipadvies, en reistijd- en routeinformatie). De deelnemers maakten veel gebruik van de pre-trip adviezen; het (on-trip) routeadvies werd minder gebruikt. Hierdoor kon geen effect gemeten worden op het aantal voertuigverliesuren of de betrouwbaarheid van de reistijden. Verwacht was dat de app meer gebruikt zou worden bij bijzondere omstandigheden, zoals slecht weer, incidenten, wegwerkzaamheden en grote drukte. Dit was echter niet in de data terug te zien. De peiling van de waardering van de dienst gaf een gemengd beeld met zowel positieve als negatieve beoordelingen.

De Praktijkproef Amsterdam heeft aangetoond dat het mogelijk is om real-time binnen 1 minuut kwalitatief hoogwaardige voorspellingen voor een groot netwerk (116.000 links en 68.000 knopen) te maken op basis van grote hoeveelheden data en op basis daarvan smart routing advies te geven aan een grote groep gebruikers.

De publiek-private samenwerking in de proef was succesvol. De samenwerking in deze proef was een belangrijke stap naar verdergaande samenwerking tussen overheid en markt. Ten bate van de proef zijn gegevens ontsloten waar marktpartijen eerder nog niet over konden beschikken, en deze data kunnen nu ook (semi-) geautomatiseerd verwerkt worden. Ook zijn goede ervaringen met betrekking tot het delen van regelscenario's opgedaan. Deze scenario's werden beschikbaar gesteld door de wegbeheerders en zijn meegenomen in de routeadviezen aan de deelnemers. Dankzij de publiek-private samenwerking die in de proef opgezet is, kon binnen zes weken een uitbreiding van de Superroute app (met onder andere OV-informatie) voor het grootschalige evenement SAIL Amsterdam gerealiseerd worden.

Dankwoord

Dit artikel is tot stand gekomen in het kader van het project Praktijkproef Amsterdam In-Car, Perceel Regulier Verkeer en Perceel Evenementen Verkeer. Dank gaat uit naar alle betrokkenen en in het bijzonder naar Isabel Wilmink, Eline Jonkers, Marco Duijnisveld en Simeon Calvert voor hun bijdrage aan eerdere publicaties waar dit artikel op is gebaseerd.

Literatuur

Jonkers, E., Wilmink, I., & Duijnisveld, M. (2016). Resultaten en geleerde lessen PraktijkProef Amsterdam in-car. *Verkeerskunde*, <www.verkeerskunde.nl/internetartikelen/vakartikelen/resultaten-en-geleerde-lessen-praktijk-proef.45132.lynkx>.

Jonkers, E., & Wilmink, I. (2016). Resultaten PraktijkProef Amsterdam in-car. *Nationaal verkeerskundecongres 2016*. <http://nationaalverkeerskundecongres.nl/nieuws/paper-resultaten-praktijkproef-amsterdam-in-car.344666.lynkx>.

Newell, G.F. (1993). A simplified theory of kinematic waves in highway traffic, Part I: General theory. *Transportation Research Part B: Methodological*, 27, 281-287.

Snelder, M., & Calvert, S.C. (2015). Real-time reistijdvoorspelling voor routekeuze- en vertrektijdstipadvies; Een toepassing in de Praktijkproef Amsterdam. *Colloquium Vervoersplanologisch Speurwerk*. pp. 1–15.

Snelder, M., Calvert, S.C., Bakri, T. Heijligers, B., & Van Huis, J. (2015). Real-time reistijdvoorspellingen voor routekeuze- en vertrektijdstipadvies; Een toepassing in de Praktijkproef Amsterdam. *Tijdschrift Vervoerswetenschap, 4*, 145-161.

Snelder, M., Calvert, S.C., Bakri, T. Heijligers, B., & Van Huis, J. (2016). Raamwerk voor real-time reistijdvoorspellingen. *Verkeerskunde*, <www.verkeerskunde.nl/internetartikelen/vakartikelen/raamwerk-voor-real-time-reistijdvoorspellingen.45131.lynkx>.

Calvert, S.C., Snelder, M., Bakri, T. Heijligers, B., & Knoop, V.L. (2015). *Real-Time Travel Time Prediction Framework for Departure Time and Route Advice*. TRB Paper No. 15-3916, Transportation Research Record No. 2490, pp 56-64, http://dx.doi.org/10.3141/2490-07

Maaike Snelder heeft econometrie gestudeerd aan de Erasmus Universiteit en is in 2010 cum laude gepromoveerd op het ontwerpen van robuuste wegennetwerken. Zij is werkzaam bij TNO en de Technische Universiteit Delft en is gespecialiseerd in het ontwikkelen, toepassen en toetsen van (datagedreven) modellen voor de evaluatie en het ontwerp van wegennetwerken en realtime informatievoorziening en advies. Maaike heeft diverse prijzen gewonnen zoals de Dewis Award, de European Excellence Award en de Greenshields Prize.
E-mail: maaike.snelder@tno.nl

BIG DATA VOOR EFFECTIEVERE SCREENING OP DARMKANKER

Mark Hoogendoorn

Big data zijn te vinden en te gebruiken in een grote varieteit aan domeinen. Een domein waarin de toegevoegde waarde evident is, is het medische vakgebied. Hierin werkt de tijd ook erg mee: er vinden veel ontwikkelingen plaats in de medische wereld die ervoor zorgen dat er steeds meer data van patiënten beschikbaar komen. Denk hierbij aan scans, laboratoriumuitslagen, voorgeschreven medicatie, gestelde diagnoses, doorverwijzingen die hebben plaatsgevonden, uitgevoerde procedures, et cetera. Als deze data voor grote hoeveelheden patiënten gecombineerd worden – uiteraard geanonimiseerd – biedt dit mogelijkheden die tot voor kort ondenkbaar waren, zoals het vinden van nieuwe patronen in die zeer informatierijke data. Voorbeelden hiervan zijn het vinden van voorspellers voor bepaalde ziektes en het voorspellen van de effectiviteit van bepaalde behandelingen voor een patiënt. Voorheen was dit soort onderzoeken erg kennisgedreven: de specialisten of wetenschappers bedachten een relatief kleine set van variabelen die interessant waren, verzamelde daar data over en kwamen met een model. Nu kunnen er met behulp van *machine learning* technieken relaties tussen alle mogelijke variabelen ontdekt worden. Echter, de data uit het medische domein vereisen wel ontwikkeling van nieuwe technieken die kunnen omgaan met de karakteristieken van deze data.

Een samenwerkingsproject van de Vrije Universiteit Amsterdam, het Leids Universitair Medisch Centrum, het Utrechts Medisch Centrum en het VU Medisch Centrum en gedeeltelijk gefinancierd door zorgverzekeraar VGZ heeft een eerste stap gezet om deze technieken te ontwikkelen. Specifiek is er gekeken naar het voorspellen van darmkanker om effectiever te kunnen screenen en patiënten op een proactieve wijze te kunnen benaderen. We zullen in dit artikel laten zien hoe we te werk zijn gegaan. Eerst worden de specifieke ziekte en de bijbehorende data die we tot onze beschikking hadden in kaart gebracht, waarna de uitdagingen in de medische datasets worden omschreven. Daarna beschrijven we de oplossing die is onderzocht en de uitkomsten. De gedetailleerde resultaten zijn gepubliceerd in *Computers in Biology and Medicine* (Kop et al., 2016).

Darmkanker

Darmkanker is een kankersoort met de op twee na hoogste incidentie wereldwijd van alle kankersoorten. De grootste moeilijkheid bij darmkanker is dat de symptomen heel aspecifiek zijn: de symptomen zijn niet uniek voor darmkanker en komen ook regelmatig voor bij veel minder ernstige ziektes. Dit terwijl een vroege diagnose bij darmkanker cruciaal is. De overlevingskansen nemen namelijk dramatisch af als de ziekte in een te laat stadium wordt geconstateerd. Er zijn op dit moment in diverse landen, waaronder Nederland, screeningsprogramma's voor darmkanker. Deze werken aan de hand van een test op bloed in de ontlasting. Niet alle mensen doen echter mee aan deze test. Daarnaast blijkt bij een positieve uitslag, waarna mensen een zeer ongemakkelijk en duur onderzoek in het ziekenhuis moeten ondergaan, slechts een relatief laag percentage mensen daadwerkelijk de diagnose darmkanker te krijgen. Naar ons idee kan big data helpen om deze screening effectiever te laten verlopen.

Patiëntendata

Natuurlijk begint een big-data-project bij de data. De universitair medische centra hebben samenwerkingen met huisartsenpraktijken die geanonimiseerde data delen van patiënten die daartegen geen bezwaar hebben. Hiermee zijn wij in staat geweest om totaal van meer dan 260.000 patiënten data te verkrijgen over een periode van vijf jaar. Van deze patiënten was informatie aanwezig over gestelde diagnoses gedurende bezoeken aan de huisarts (meer dan 11 miljoen in totaal), medicatie (meer dan 23 miljoen recepten), doorverwijzingen (ongeveer 4,4 miljoen) en laboratoriumuitslagen (circa 22 miljoen uitslagen). Hiertussen bevonden zich 1.292 gevallen van darmkanker. De data zijn allemaal gecodeerd en op een gestructureerde manier opgeslagen. De dataset bevat een drietal uitdagingen die aangepakt moeten worden om machine learning technieken effectief toe te kunnen passen:

1. De data zijn erg *temporeel* van aard. Er zijn gebeurtenissen die op bepaalde momenten plaatsvinden. Juist relaties over de tijd zouden voorspellend kunnen zijn, bijvoorbeeld eerst milde buikklachten waarna steeds ernstiger diagnoses gesteld worden. Echter, veel machine learning technieken kunnen niet goed omgaan met dit soort data: ze negeren de tijdscomponent en daarmee gaat mogelijk waardevolle informatie verloren.
2. De data zijn vaak *onvolledig*. Patiënten komen lang niet voor alle klachten langs en presenteren ook niet alle klachten aan de huisarts. Daarnaast kan het huisartsinformatiesysteem een bepaalde manier van registreren in de hand werken.
3. Sommige waardes in het systeem betekenen alleen iets als ze in een bepaalde *context* gezien worden. Denk bijvoorbeeld aan het gewicht van een patiënt. Dat iemand 80 kilogram weegt zegt niet direct iets. Een combinatie met lengte of bijvoorbeeld bouw is nodig om te bepalen of het gewicht "gezond" is. Het is ook mogelijk dat juist een verandering van het gewicht interessant is.

Deze drie uitdagingen vergen een specifieke aanpak. Hiertoe hebben we een soort pijplijn ontwikkeld om de data als het ware voor te bereiden voordat er machine learning wordt toegepast om voorspellende modellen voor darmkanker te vinden.

Machine learning pijplijn

De pijplijn bestaat uit drie delen die de bovengenoemde uitdagingen aanpakken. In het eerste deel wordt de *context* van bepaalde waardes toegevoegd. Specifiek gaat het in dit geval om laboratoriumuitslagen. We onderscheiden twee situaties: (1) we weten bepaalde normen voor de waardes (denk hierbij bijvoorbeeld aan de bloeddruk), en (2) we weten dit niet en kijken naar trends over de tijd. In het eerste geval vervangen we de numerieke waarde door een categorie, te weten, *laag*, *normaal* en *hoog*. In het tweede geval kijken we naar hoe de waardes zich over de tijd tot elkaar verhouden en leiden af of de waarde *dalend*, *stijgend* of juist *stabiel* is. In feite vervangen we de waarde dus wederom door een categorie.

In het tweede deel van de pijplijn proberen we iets te doen aan de *onvolledigheid* van de data. Het medische domein beschikt over enorm veel kennis. Deze kennis is in diverse gevallen formeel gespecificeerd in een medische ontologie. Hierin worden bijvoorbeeld relaties tussen medicatie en diagnoses vastgelegd (medicijn x kan gebruikt worden als diagnose y gesteld wordt). Op het moment dat we zien dat een medicijn wordt voorgeschreven en dit wordt alleen gedaan bij een bepaalde diagnose die niet geregistreerd is kunnen we aannemen dat de betreffende diagnose ook in dit geval gemaakt is en voegen we deze diagnose toe aan de dataset.

Ten slotte kijken we naar het *temporele* aspect. Wat we in dit deel doen is het zoeken naar patronen die over de tijd plaatsvinden, bijvoorbeeld eerst obstipatie gevolgd door anemie (ijzergebrek) of juist op hetzelfde tijdstip. We zoeken dan patronen die bij voldoende patiënten voorkomen maar ook onderscheidend zijn tussen de twee groepen patiënten die we hebben, te weten de groep die wel een darmkanker diagnose krijgt, en de groep waarvoor dat niet geldt.

We passen deze pijplijn als volgt toe. Eerst selecteren we voor ieder van de 260.000 patiënten in de dataset een relevante geschiedenis aan data. Voor de patiënten met een darmkankerdiagnose zijn dit de zes maanden voor de diagnose. Voor de overige patiënten nemen we een willekeurige periode van zes maanden. We passen vervolgens op deze selectie van de data de drie stappen toe die we zojuist beschreven hebben. Tenslotte tellen we voor de patiënt hoe vaak bepaalde gebeurtenissen voorkwamen gedurende de geselecteerde periode en gebruiken dat om machine learning op uit te voeren. Bijvoorbeeld, een patiënt heeft vier keer een diagnose obstipatie gekregen, had vijf keer een normale bloeddruk, had een keer een patroon van obstipatie gevolgd door anemie en de patiënt had de diagnose darmkanker. Op deze manier vatten we dus eigenlijk de hele geschiedenis van de patiënt samen. Het laatste probleem is dat de groep van darmkanker patiënten enorm ondervertegenwoordigd is in de data, we lossen dit op door deze gevallen een hoger gewicht te geven bij het leren van het model. Dit heeft tot gevolg dat zij prominenter in de aanpak zullen worden meegenomen.

Resultaten

We passen verschillende machine learning algoritmes toe. We vinden dat we in staat zijn om significant beter darmkanker te voorspellen dan reeds bestaande modellen. Met name het temporele deel van onze pijplijn blijkt veel toegevoegde waarde te hebben. Verder bevat dit model, zonder enige sturing gegeven te hebben, de bekende voorspellers voor darmkanker (onder andere obstipatie, anemie, leeftijd, buikpijn). In aanvulling daarop is er een nieuwe voorspeller gevonden, te weten het metaboolsyndroom, een stofwisselingsziekte. Er was wel een relatie bekend tussen darmkanker en het metaboolsyndroom, maar dit is de eerste keer dat deze relatie ook echt als voorspellend naar voren komt.

Conclusie

Medische big data is een enorm rijke bron van informatie die kan worden gebruikt om de zorg van patiënten te verbeteren. Om een goed resultaat te bereiken met technieken uit het domein van machine learning moet echter wel rekening gehouden worden met de karakteristieken van de data. In dit artikel is een voorbeeld beschreven van zo'n aanpak. In de toekomst zien we voor ons dat hetzelfde proces (of beter nog, een verbetering ervan) kan worden toegepast op andere ziektebeelden.

Literatuur

Kop, R., Hoogendoorn, M., Ten Teije, A., Büchner, F. L., Slottje, P., Moons, L. M., & Numans, M. E. (2016). Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records. *Computers in Biology and Medicine, 76*, 30-38.

Mark Hoogendoorn is universitair docent kunstmatige intelligentie aan de Afdeling Informatica van de Vrije Universiteit Amsterdam. In zijn onderzoek richt hij zich met name op het grensvlak tussen machine learning en de gezondheidszorg. Hij is in 2007 gepromoveerd aan dezelfde universiteit.
E-mail: m.hoogendoorn@vu.nl

Hoe laat laad ik mijn auto?

Pia Kempker, Nico van Dijk, Werner Scheinhardt, Hans van den Berg & Johann Hurink

De verkoop van elektrische auto's heeft de afgelopen jaren een sterke groei gekend. Niet alleen in absolute zin rijden er meer elektrische auto's over onze wegen, maar ook het marktaandeel neemt toe. Naast de toenemende maatschappelijke bewustwording van milieuvriendelijk rijden, als veruit de belangrijkste factor, speelt ook een puur oer-Hollandse geaardheid hierbij een rol: zuinigheid. Ondanks de aanmerkelijk hogere aanschaf- en afschrijvingskosten, komt men over een gebruiksperiode van 5 jaar al gauw tot een besparing van enkele honderden euro's per jaar, deels (tot voor kort) dankzij een belastingvoordeel, maar vooral ook dankzij het lagere energie verbruik per kilometer. (zie bijvoorbeeld Rademaker, 2013, 2016).

Een verdere reductie van kosten zou dan ook een aardige duit in het zakje kunnen doen om huidige of toekomstige twijfelaars alsnog over de streep te trekken om elektrisch te gaan rijden. Dit lijkt haalbaar door elektrische auto's slim op te laden en dat is waar dit artikel zich op richt.

Met *smart technology* wordt in de toekomst de bijzondere mogelijkheid geboden gebruik te maken van sterke prijsfluctuaties voor elektriciteit. Deze fluctuaties vloeien voort uit fluctuaties in vraag en aanbod, waarbij aan de aanbodkant (met steeds meer zonne- en windenergie) veranderende weersomstandigheden natuurlijk een belangrijke rol spelen. De prijzen worden uiteindelijk bepaald door een gecompliceerd proces om vraag en aanbod middels biedcurves te matchen (zie kader over de PowerMatcher). De kunst voor de autobezitter is om zo goed mogelijk op prijsfluctuaties in te spelen, dus zoveel mogelijk op te laden tegen lage tarieven zonder het risico te lopen dat de accu tijdens het rijden onvoldoende is opgeladen.

Dit artikel richt zich op het afleiden van zo'n optimale laadstrategie. We doen dit in de context van de in het kader beschreven (two-timescale) PowerMatcher.

Concreet kijken wij naar een situatie waar een elektrische auto over een gegeven aantal tijdintervallen moet worden opgeladen. Hiertoe moet vóór het begin van ieder tijdsinterval beslist worden welke biedingscurve naar de PowerMatcher gestuurd wordt zonder dat men al volledige informatie heeft over de prijzen in toekomstige tijdsintervallen. Dit leidt tot de volgende doelstelling:

BIEDINGSPROBLEEM het verkrijgen van zo goed mogelijke biedingscurves voor toepassing in de PowerMatcher (zie kader) in de situatie met prijsfluctuaties en onzekerheden;

wat neerkomt op:

LAADPROBLEEM het bepalen van de hoeveelheid af te nemen energie voor iedere mogelijke waarde die de prijs kan aannemen in het volgende tijdsinterval.

We gaan inzoomen op het laadprobleem, dat we met behulp van klassieke stochastische OR-optimalisatie aanpakken. We formuleren het probleem als een stochastisch dynamisch programmeringsprobleem (SDP) en ontwikkelen op basis van deze formulering een heuristische oplossing van het laadprobleem. Ten slotte laten wij zien hoe deze strategie het doet in de praktijk.

SDP-formulering

In essentie vertoont het laadprobleem de structuur van een dynamisch beslissingsprobleem, te weten: op achtereenvolgende momenten $t = 1, 2, ..., T$ kan een beslissing genomen worden over hoeveel in die periode $t : [t,t+1)$ te laden, met de eis dat aan het einde van de totale laadperiode, dus op moment $T+1$, aan de totale behoefte L voldaan moet zijn.

Wanneer de prijzen $p_1, ..., p_T$ op de momenten $t = 1, ..., T$ vooraf bekend zouden zijn, is dit laadprobleem een standaard rugzak (*Knapsack*) probleem. Deze prijzen zijn echter vooraf onbekend dan wel onzeker als gevolg van verschillende externe en tussentijds veranderende factoren als weercondities, gebruikersprofielen, biedingen enz. (zie kader over de PowerMatcher). In dit artikel zullen wij deze onzekerheid van prijzen modelleren als stochastisch, dat wil zeggen door middel van de prijzen als stochastische grootheden: $P_1, ..., P_T$. Op basis van de realisatie voor de huidige periode t kan dan vervolgens besloten worden hoeveel in die periode moet worden opgeladen, rekening houdend met het stochastisch karakter van de prijzen in de daaropvolgende perioden. In het kader hieronder wordt voor een fictieve situatie getoond dat dit probleem te beschrijven is met behulp van een eenvoudige beslisboomstructuur.

De beslisboomstructuur leidt tot een SDP-probleem (als algemenere opzet van een Markov Decision Problem, zie bijv. Tijms, 1986 of Puterman, 2014) met de volgende beslissingsstructuur: gegeven de situatie (toestand) op een moment t, dient een beslissing genomen te worden. Dit heeft twee gevolgen:

- Directe kosten (in dit geval voor het opladen) in de periode $[t,t+1)$;
- De overgang naar een nieuwe situatie (toestand) op tijdstip $t+1$.

Omdat de toestandsbeschrijving voldoende informatie moet bevatten voor deze twee gevolgen dient in de formulering hiervan ook de prijs opgenomen te zijn. Dit leidt tot toestanden (x_t,p_t) op tijdstip t met

x_t: de resterende nog te laden hoeveelheid energie,
p_t: de huidige prijs per kWh.

Voor de éénstapskosten $r_t(x_t,p_t,u)$ van het opladen van u kWh in de toestand (x_t,p_t) nemen wij de kosten evenredig aan de laadhoeveelheid $r_t(x_t,p_t,u) = up_t$. De nu te bepalen waardes zijn

$V_t(x_t,p_t)$: minimale verwachte kosten vanaf t vanuit toestand (x_t,p_t).

Op basis van verwachtingswaarden kan het SDP vervolgens van achteraf beginnend (op moment T) teruglopend naar t=1 worden doorgerekend, waarbij de waarden $V_t(x_t,p_t)$ iteratief bepaald worden volgens de SDP vergelijkingen:

$$V_{T+1}(x,p) = \begin{cases} 0 & \text{als } x = 0 \\ \infty & \text{anders} \end{cases}$$

$$V_t(x,p) = \min_{u \le u_{max}} [up + E_{P_{t+1}} V_{t+1}(x-u, P_{t+1})]$$
voor $t = T, ..., 1$.

De uiteindelijke minimale laadkosten en de daartoe vereiste beslissingen in elke mogelijke tussentoestand (x_t,p_t) voor de momenten $t = 1, ..., T$ volgen vanuit $V_1(L,p_1)$: de minimale verwachte laadkosten startend met prijs p_1 op t=1.

Deze oplossingsstructuur blijft ook geldig bij tijdsafhankelijke en zelfs afhankelijke prijsverdelingen, maar deze generalisaties brengen computationeel gezien wel com-

POWERMATCHER – DECENTRAAL ENERGIEMANAGEMENT

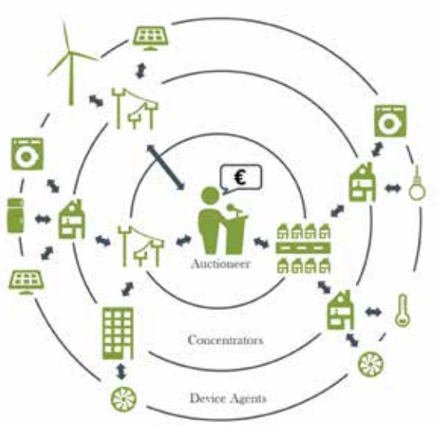
De ontwikkeling naar een maatschappij met duurzame energie betekent dat de energieproductie meer decentraal wordt: in plaats van enkele grote energieproducenten krijgen we te maken met opwekking van energie door een groot aantal verschillende partijen, zoals ook de buurman met zonnepanelen op het dak. Om het elektriciteitsnetwerk stabiel te houden moet echter wel worden voldaan aan de voorwaarde opwekking = verbruik. In de toekomstige decentrale setting is een gecoördineerde afstemming van opwek en verbruik een flinke uitdaging en vraagt om een decentraal energiemanagement.

De PowerMatcher[1] is een voorbeeld van een op de markt gebaseerde coördinatietechnologie die tijdsafhankelijke prijssignalen ondersteunt. De prijs voor energie wordt bepaald door een veiling, waarbij iedere marktpartij een biedcurve (*bidding curve*) aanlevert die aangeeft hoeveel energie de partij bij verschillende prijzen wil afnemen of leveren. Een coördinerende agent verzamelt deze biedcurves en bepaalt een prijs die ervoor zorgt dat de markt in evenwicht is; dit gebeurt in real-time (zie ook figuur 1).

Om dit marktmechanisme te realiseren, moeten er voor apparaten biedcurves opgesteld worden. Deze biedcurves verschillen echter per apparaat:

- Een zonnepaneel levert de opgewekte energie, onafhankelijk van de stroomprijs.
- Een elektrische verwarming heeft meer flexibiliteit en kan afhankelijk van de prijsverwachting voor de nabije toekomst besluiten nu aan te slaan of juist iets later.
- Voor een elektrische auto is er 'winst' te behalen door alleen op de goedkoopste momenten op te laden, zolang dit geen beperkingen oplevert voor de gewenste mobiliteit.

Voor veel apparaten zoals elektrische auto, batterij en verwarming is hierbij het hebben van informatie over toekomstige prijzen van belang om een betere bieding te maken. De zogenoemde two-time-scale PowerMatcher is een uitbreiding van de PowerMatcher die hierop inspeelt. Hierbij wordt in eerste instantie door middel van een bieding een inschatting van de gemiddelde prijs voor een iets langere periode verkregen (bijvoorbeeld voor de komende 24 uur). Hierna wordt net zoals bij de gewone PowerMatcher het verbruik en de opwekking op korte termijn middels biedingen bepaald. Apparaten zoals elektrische auto's, batterijen en verwarming kunnen de in de eerste bieding verkregen gemiddelde prijs gebruiken om hun korte-termijn biedcurve aan te passen. In dit artikel laten wij zien dat deze extra informatie tot behoorlijke kostenreducties kan leiden, vergeleken met de gewone PowerMatcher.

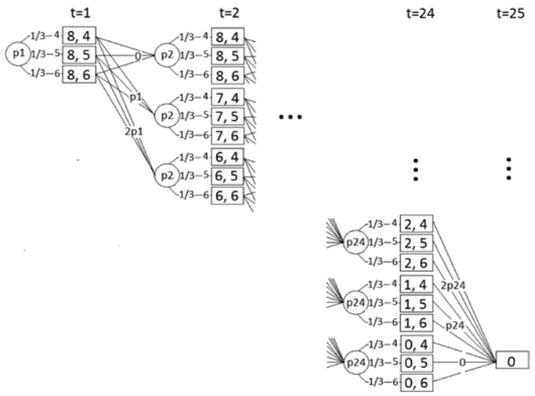


Figuur 1. Visualisatie van het PowerMatcher concept. Bron: <http://flexiblepower.github.io/technology/powermatcher/>

VOORBEELD

Ter illustratie bekijken wij een eenvoudige fictieve situatie. Gegeven is een tijdsperiode van 8 uur 's avonds tot 8 uur 's ochtends, opgedeeld in 24 intervallen van een half uur. In ieder interval hebben wij dezelfde, maar onafhankelijke, discrete verdeling voor de prijzen: de drie prijzen zijn 4, 5, 6 en de kans voor ieder van deze prijzen is 1/3 voor alle intervallen. In de totale tijdsperiode van 12 uur moet een auto geladen worden met L = 8 kWh. Het bijbehorende laadprobleem leidt tot een beslisboom zoals getoond in figuur 2. Per half uur moet worden beslist, afhankelijk van de nog te laden hoeveelheid en de huidige prijs (de rechthoeken in de figuur), hoeveel de auto in het komende half uur wordt opgeladen.

Met een dergelijke beslisboomstructuur kunnen, van achteraf terugrekenend, de minimale verwachtingswaarden worden bepaald, samen met de daarvoor benodigde beslissingen in elk van de mogelijke beslispunten (de rechthoeken).



Figuur 2. Beslisboom voor het opladen van een auto; essentieel is hierbij een 2-dimensionale toestand (zie uitleg SDP)

plicaties met zich mee. In het vervolg gaan we uit van onafhankelijke prijzen voor de achtereenvolgende periodes, met allemaal dezelfde kansverdeling $F(.)$.

Het aardige van deze specifieke SDP-formulering is dat een analytische oplossing te vinden is, met een drempelstructuur die intuïtief begrijpelijk is. Door voor de huidige periode t de huidige prijs p ($=P_t$) te vergelijken met de verwachte *order-statistics* (laagste waarden) van de toekomstige prijzen P_{t+1} tot en met P_T, wordt besloten of er wel of juist niet moet worden bijgeladen. Er wordt bijgeladen op maximale laadsnelheid u_{max} wanneer p lager ligt dan de verwachte k-na laagste prijs, waarbij k het aantal perioden is dat er nog minimaal moet worden opgeladen, en er wordt niet bijgeladen wanneer p hoger ligt dan de verwachte (k+1)-na laagste prijs. Echter is het bepalen van de verwachte order-statistics in de praktijk computationeel weinig aantrekkelijk.

Een alternatief is om deze optimale strategie te vervangen door een heuristiek (zie kader). Deze stelt losjes gezegd het volgende: als het verwachte aantal goedkopere periodes in de toekomst niet groter is dan het minimale aantal periodes dat nog moet worden bijgeladen, dan wordt maximaal bijgeladen, terwijl er in het tegenovergestelde geval doorgaans niets wordt bijgeladen.[2]

De praktijk

De gevonden strategie is getoetst op basis van echte data voor het Belgische netwerk in de eerste helft van 2015 (180 dagen). Deze data bevat voor elk kwartier de hoeveelheid elektriciteit die in totaal daadwerkelijk werd verbruikt, de hoeveelheid daarvan opgewekt uit wind, de voorspellingen voor het verwachte totale verbruik en de voorspelling van de verwachte windopwekking. Deze data is voor ons voorbeeld opgeschaald naar een toekomstige situatie met een groot aandeel windenergie. Uitgaande van (fictieve) vaste prijzen van 1 cent per kWh voor windenergie en 10 cent per kWh voor overige (fossiele) energie, zijn hieruit schattingen voor een gerealiseerde en een verwachte prijs per kWh bepaald. De laatste komt overeen met $E[P_t]$, oftewel de verwachtingswaarde behorend bij de in het model gehanteerde verdelingsfunctie F.

We gaan uit van een elektrische auto met een kleine accu van 8 kWh (waarmee een afstand van ruwweg 40 km kan worden afgelegd), die iedere nacht tussen 20.00 uur 's avonds en 8 uur 's ochtends volledig moet worden opgeladen. Op basis van de bovenbeschreven data zijn nu de volgende strategieën met elkaar vergeleken:

1. Naïef: zo snel mogelijk. Dit komt overeen met het aansluiten van de oplader om 20.00 uur waarna de accu met u_{max} wordt opgeladen tot deze vol is;
2. PowerMatcher Hierbij wordt de strategie uit het bovenstaande kader gevolgd met een normaal verdeelde F waarbij de parameters altijd gelijk worden gekozen, op basis van de statistieken over een jaar (u=5, a^2=1);
3. Two-time-scale PowerMatcher Hierbij wordt de strategie uit het bovenstande kader gevolgd met een normaal verdeelde F waarbij de parameters elk kwartier worden bijgewerkt op basis van de voorspellingen.

In figuur 3 is voor één van de nachten in 2015 te zien hoe de naïeve strategie vanaf het begin oplaadt, en de PowerMatcher de 'dure periodes' overslaat. Maar ook is te zien hoe de two-time-scale PowerMatcher er nog veel beter in slaagt om de werkelijk 'goedkope periodes' te vinden. De resultaten voor het halve jaar zijn samengevat opgenomen in tabel 1. Ten opzichte van de naïeve strategie doet de PowerMatcher het beter (2,2%), maar de two-time-scale PowerMatcher geeft een veel grotere verbetering, namelijk 8,2%. Ter vergelijking, wanneer we voorkennis van de werkelijke prijzen zouden hebben, zou de maximaal haalbare verbetering 10,8% bedragen!

STRATEGIE

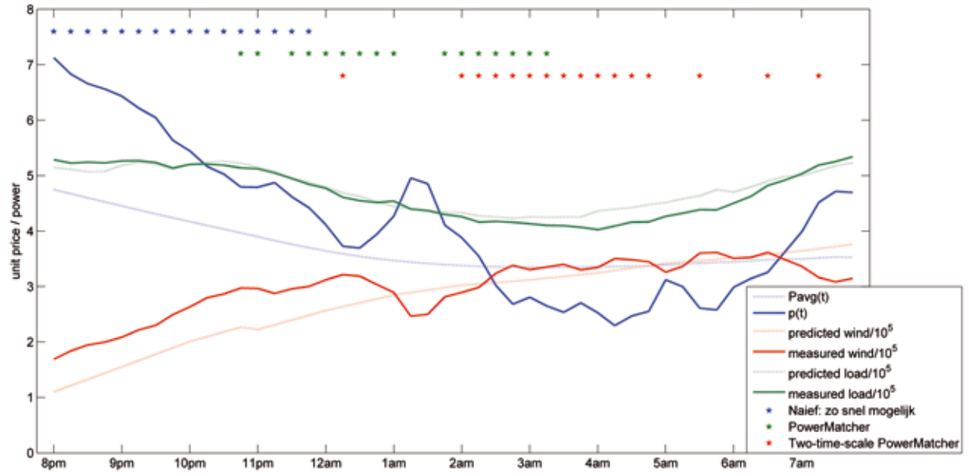
Een goede heuristiek voor de hoeveelheid op te laden energie $u_t(x, p)$ in periode t bij een totaal nog te laden hoeveelheid x en een huidige prijs p wordt gegeven door:

$$u_t(x,p) = \begin{cases} \min(u_{max}, x) & \text{als} \quad (T-t+1)F(p) \le k \\ x - ku_{max} & \text{als } k < (T-t+1)F(p) \le k+1 \\ 0 & \text{als} \quad (T-t+1)F(p) > k+1, \end{cases} \qquad (3)$$

waar $k = \left\lfloor \frac{x}{u_{max}} \right\rfloor$ het minimale aantal perioden is dat er nog moet worden opgeladen bij maximale laadsnelheid. Verder is F de verdelingsfunctie van de prijzen. In de praktijk zullen we deze normaal verdeeld veronderstellen.

STRATEGIE	KOSTEN	%
Naïef: zo snel mogelijk	108,61	-
PowerMatcher	106,26	2,2
Two-time-scale PowerMatcher	99,71	8,2
(Optimal)	96,83	10,8

Tabel 1. Vergelijking van totale (fictieve) kosten in euro's voor het opladen van een auto over de eerste 180 nachten van 2015, en de procentuele verbetering t.o.v. de 'naïeve' strategie.



Figuur 3. Data en resultaten voor één nacht in 2015, gegenereerd op basis van de gegevens voor het Belgische netwerk. De gekleurde sterretjes geven aan in welke periodes de auto wordt opgeladen bij de gebruikte strategieën, de blauwe lijnen geven de huidige prijs en de verwachte gemiddelde prijs voor het restant van planning periode. De rode lijnen geven de voorspelde en de gerealiseerde windopwekking en de groene lijnen het voorspelde en het gerealiseerde totale verbruik weer

Conclusie

In dit artikel hebben wij, in het verlengde van het onderzoek in Kempker et al., 2015, het probleem van het slim opladen van een elektrische auto bij onzekere en over de tijd variërende elektriciteitsprijzen onderzocht. Onder de aanname dat deze prijzen als onafhankelijke stochastische variabelen gemodelleerd kunnen worden, hebben we met behulp van Stochastic Dynamic Programming een analytische oplossing gevonden voor de optimale oplaadstrategie, en daarmee tevens voor de biedingsstrategie in de context van de (two-time-scale) PowerMatcher. Op basis hiervan is vervolgens een praktisch bruikbare heuristiek ontwikkeld, en deze is in een vergelijkende numerieke studie op basis van echte data getest.

Hierbij lijkt het dat de resultaten een grote mate van robuustheid hebben ten aanzien van precieze modelveronderstellingen: *hoe* de toekomstverwachting wordt meegenomen maakt niet zoveel uit, het gaat er vooral om *dat* deze wordt meegenomen. De two-time-scale PowerMatcher biedt goede mogelijkheden om hier in te voorzien, terwijl het DP-karakter van de strategie (kortetermijn feedback loop) garandeert dat suboptimale tussentijdse beslissingen goed worden opgevangen.

Met dank aan Leon Kester, Koen Kok, Pamela MacDougall (allen TNO) en Yvonne Prins.

Noten

1. Voor een gedetailleerde beschrijving van de PowerMatcher, zie bijvoorbeeld <http://flexiblepower.github.io>.
2. In een tussenliggend geval wordt bijgeladen, en wel met een zodanige hoeveelheid dat na afloop van de huidige periode de hoeveelheid nog bij te laden energie (x) een veelvoud is van u_{max}.

Literatuur

Rademaker, P., (2013, 2016). *Elektrisch rijden wint aan populariteit. Maar ... is het ook iets voor jou?* http://wqd.nl/89UK

Tijms, H.C. (1986). *Stochastic modelling and analysis; A computational approach*. John Wiley & Sons.

Puterman, M.L. (2014). *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons.

Kempker, P., Van Dijk, N.M., Scheinhardt, W.R.W., Van den Berg, J.L., & Hurink, J.L. (2015). Optimization of charging strategies for electric vehicles in PowerMatcher-driven smart energy grids. In: *Proceedings 9th EAI International Conference on Performance Evaluation Methodologies and Tools (Valuetools)*. Berlin.

Pia Kempker is gepromoveerd in de systeemtheorie aan de VU Amsterdam en wiskundig onderzoeker bij de groep Cyber Security and Robustness van TNO. E-mail pia.kempker@tno.nl

Nico van Dijk bekleed de leerstoel Stochastic Operations Research (SOR) en is verbonden aan het Centre for Healthcare Operations Improvement and Research (CHOIR) van de Faculteit EWI, Universiteit Twente. E-mail: n.m.vandijk@utwente.nl

Werner Scheinhardt is universitair docent bij de leerstoel Stochastic Operations Research (SOR) van de Faculteit EWI, Universiteit Twente. E-mail: w.r.w.scheinhardt@utwente.nl

Hans van den Berg is senior onderzoeker bij TNO en als deeltijdhoogleraar ICT tevens verbonden aan Faculteit EWI, Universiteit Twente. E-mail j.l.vandenberg@tno.nl

Johann Hurink is als hoogleraar Mathematische Besliskunde verbonden aan de Faculteit EWI van de Universiteit Twente en directeur van het Landelijk Netwerk Mathematische Besliskunde (LNMB). E-mail: j.l.hurink@utwente.nl

Foto: Twente milieu | Léonie Vaarhorst

DYNAMISCHE AFVALINZAMELING

Efficiënter en slimmer

Doorgaans worden er voor afvalinzameling vaste routes en schema's gebruikt. De containers worden dagelijks of wekelijks geleegd, of ze nu vol zitten of niet. Ondergrondse containers kunnen vrij gemakkelijk uitgerust worden met sensoren die registreren hoe vol de containers zitten. Deze informatie maakt het mogelijk om dynamisch het moment te kiezen waarop de containers moeten worden geleegd. De inzameling op basis van slim plannen zorgt voor kostenbesparing, minder geluidsoverlast, minder uitstoot en minder irritaties over volle containers.

Martijn Mes & Marco Schutten

In Nederland wordt afval op gemeentelijk niveau verzameld en verwerkt. In dit proces worden verschillende containertypes gebruikt voor de tijdelijke opslag van afval bij de mensen thuis en bij bedrijven. Zo zijn er minicontainers (ook wel bekend als kliko's) per huishouden en blokcontainers die worden gedeeld door meerdere huishoudens. De minicontainers moeten periodiek (bijvoorbeeld op dinsdag in de even weken) aan de straat gezet worden, zodat de afvalinzamelaar ze kan legen. Twente Milieu, die onder andere de afvalinzameling verzorgt voor Enschede, Hengelo en Almelo, gebruikt sinds 2009 ook ondergrondse afvalcontainers. In eerste instantie werden deze containers alleen geplaatst bij appartementencomplexen en commerciële gebouwen, zoals restaurants. Tegenwoordig zijn ze ook te vinden in woonwijken.

Ondergrondse containers hebben een aantal voordelen ten opzichte van mini- en blokcontainers. Ze zijn bijvoorbeeld minder zichtbaar dan blokcontainers en hoeven niet (zoals minicontainers) op vaste momenten geleegd te worden. Bovendien kunnen ze vrij gemakkelijk uitgerust worden met sensoren die de klepbewegingen bijhouden of de vulgraad meten. Deze informatie kan vervolgens centraal worden uitgelezen en kan gebruikt worden om dynamisch het moment te kiezen waarop een container geleegd moet worden. Dit kan leiden tot minder transportbewegingen, wat resulteert in kostenbesparingen en vermindering van verkeersopstoppingen en emissies van schadelijke stoffen.

In een plan voor het legen van ondergrondse containers moet per dag bepaald worden welke containers er die dag geleegd worden en welke route de

vrachtwagens hiervoor rijden. Merk hierbij op dat de beslissingen voor een dag gevolgen hebben voor de dagen erna: als vandaag een bijna volle container niet geleegd wordt, dan zal dat morgen zeer waarschijnlijk wel moeten gebeuren. Het probleem dat opgelost moet worden kan gezien worden als een 'omgekeerd' *Inventory Routing Problem* (IRP). In een IRP moet een leverancier zijn klanten van voorraad voorzien op basis van actuele voorraadinformatie. De belangrijkste beslissingen zijn (1) wanneer welke klant beleveren, (2) hoeveel te leveren en (3) middels welke route deze klanten te beleveren. Voor de IRP bestaan verschillende geavanceerde oplosmethoden. Deze zijn echter minder geschikt voor ons probleem van het plannen van de lediging van ondergrondse containers vanwege het grote aantal locaties in een relatief klein gebied (in een stad kunnen honderden ondergrondse containers staan).

Voor het maken van een goed plan is een aantal observaties vanuit de praktijk belangrijk. Zo is de hoeveelheid afval die in een container gestort wordt afhankelijk van de container en van de dag van de week. Daarnaast zijn er seizoensinvloeden over de weken van het jaar. Ook worden de containers niet geleegd in de weekenden; dit geeft een piek voor de containers die geleegd moeten worden op maandag (als er van tevoren niet op geanticipeerd is). Een methode voor dynamische afvalcollectie zal dus moeten proberen de werklast over de dagen van de week te verspreiden, door bijvoorbeeld op vrijdag al wel een container te legen die op maandag niet geleegd zou worden bij gelijke vulgraad.

Het doel van onze studie is om een relatief eenvoudig methode te ontwikkelen (dit vergemakkelijkt de toepassing in de praktijk) waarvan de werking via parameters aangepast kan worden om bijvoorbeeld de werklast goed te verdelen over de dagen van de week. Daarnaast willen we een methode ontwikkelen voor het automatisch bepalen van de beste parameterinstelling op basis van historische gegevens. We gebruiken hierbij Twente Milieu als case studie. Twente Milieu heeft na het uitvoeren van onze studie de dynamische afvalinzameling doorgevoerd en de daaruit resulterende besparingen gerapporteerd (Tubantia, 2013).

Heuristiek voor dynamische afvalinzameling

Het probleem van dynamische afvalinzameling bestaat grofweg uit twee delen: 1. bepalen welke afvalcontainers op een dag geleegd zullen worden en 2. bepalen welke routes de vrachtwagens hiervoor gaan rijden. Vanwege het dichte wegennetwerk in steden en het grote aantal containers in een relatief klein gebied is vooral de bepaling van de containers die geleegd moeten worden van belang. Hieronder richten we ons dus vooral op deze beslissing; in ons onderzoek worden de routes voor de vrachtwagens bepaald met een eenvoudige heuristiek (*Cheapest Insertion*) en in de praktijk zullen deze routes bepaald worden door een Transport Management Systeem (TMS).

Voor het kiezen van de te legen containers op een dag maken we gebruik van het concept waarin we de containers op een dag classificeren als '*MustGo*', '*MayGo*' of '*NoGo*'. Een MustGo-container moet op de bewuste dag geleegd worden. Een MayGo-container mag geleegd worden als dat goed uitkomt, bijvoorbeeld als deze in de buurt ligt van MustGo-containers. Een NoGo-container mag op de bewuste dag niet geleegd worden. Het classificeren van containers doen we op basis van de vulgraad van de containers aan het begin van de dag, of beter gezegd, op basis van de verwachting over hoe lang het nog duurt totdat de container vol is. Naast de informatie van de sensoren gebruiken we dus stochastische informatie over de hoeveelheid afval die in een container gestort wordt.

Op basis van twee grenswaarden voor het aantal dagen dat het nog duurt totdat een container naar verwachting vol is, wordt deze geclassificeerd als MustGo, MayGo of NoGo. Aangezien we de classificatie afhankelijk willen maken van de dag van de week (om de werklast net na het weekend niet te groot te laten worden en omdat de hoeveelheid afval die gestort wordt afhankelijk is van de dag van de week) geeft dat 10 parameters (5 werkdagen maal 2 grenswaarden). Om nog beter onderscheid te kunnen maken in de dag van de week, hebben we daarnaast een parameter per dag van de week die aangeeft hoeveel containers er die dag maximaal geleegd mogen worden. In totaal moeten we dus 15 parameters van een waarde voorzien.

Grofweg werkt onze heuristiek voor dynamische afvalinzameling als volgt. Na het classificeren van de containers worden er routes gemaakt voor de MustGo's. Daarna worden er MayGo's aan de routes toegevoegd totdat er geen MayGo's meer zijn of totdat het maximale aantal containers voor die dag bereikt is. De volgorde voor het toevoegen van MayGo's gebeurt op basis van een ratio die een afweging maakt tussen de hoeveelheid afval in een container en de afstand die er extra voor gereden moet worden om deze te legen.

Optimal Learning

Voor het bepalen van de waardes voor de 15 parameters maken we gebruik van *Optimal Learning*. Optimal Learning is een techniek waarmee met zo min mogelijk metingen informatie wordt verzameld met als doel toekomstige beslissingen te verbeteren. In ons geval betekent het doen van metingen het uitvoeren van een computersimulatie. De simulatie dient dan om te bepalen hoe goed een specifieke combinatie van parameterwaardes werkt, dat wil zeggen, wat de kosten op de lange duur gemiddeld zijn bij gebruik van de gegeven combinatie. Nu is het onmogelijk om alle combinaties van parameterwaardes te evalueren aangezien 1. er enorm veel combinaties van mogelijke parameterwaardes bestaan, 2. elk simulatie experiment een stochastische uitkomst geeft (d.w.z. bevat ruis in het antwoord waardoor we mogelijk het experiment een aantal keren moeten herhalen) en 3. elk simulatie experiment relatief veel tijd kost. Het doel van Optimal Learning is om de beste combinatie van parameterwaardes te bepalen, zonder expliciet iedere combinatie te evalueren. Voor meer details over hoe we Optimal Learning hebben toegepast voor dit probleem, verwijzen we naar Mes et al. (2014).

Resultaten

We hebben onze methode getest met behulp van computersimulatie, waarbij we zowel de case van Twente Milieu als ook virtuele probleeminstanties gebruiken. Belangrijke conclusies zijn dat de voorgestelde aanpak goed werkt, dat verschillende type netwerken (qua afstanden, afvalvolumes, etc.) andere parameterinstellingen nodig hebben, dat de parameterwaardes verschillen voor de dagen van de week en dat een half uur rekentijd voldoende is om goede parameterwaardes te leveren op een standaard computer.

Om met veranderende omstandigheden om te gaan, kunnen de parameterwaardes van onze heuristiek aangepast worden. Zelfs als het netwerk van beschikbare containers niet verandert, kunnen veranderingen van andere aspecten (zoals het aantal beschikbare voertuigen of de gemiddelde hoeveelheid gestort afval) ervoor zorgen dat de parameterwaardes aangepast moeten worden. Dit betekent bijvoorbeeld dat bij de start van de zomervakantie of het veranderen van de seizoenen de parameterwaardes aangepast zouden moeten worden. De beperkte benodigde rekentijd voor onze Optimal Learning methode maakt dit mogelijk. De simulatieresultaten laten zien dat door het aanpassen van de parameterwaardes kostenbesparingen tot 40% mogelijk zijn ten opzichte van het gebruik van standaardinstellingen. Voor meer details en verdere analyse van de resultaten verwijzen we naar Mes et al. (2014).

Uiteindelijk heeft Twente Milieu een vereenvoudigde versie van de beschreven methode geïmplementeerd. Resultaten van deze implementatie laten zien dat dynamische afvalinzameling significante besparingen oplevert (Tubantia, 2013). Door dit concept konden twee vrachtwagens voor de afvalinzameling worden afgestoten.

Literatuur

Mes, M., Schutten, M., & Pérez Rivera, A. (2014). Inventory routing for dynamic waste collection. *Waste Management, 34*(9), 1564-1576. <dx.doi.org/10.1016/j.wasman.2014.05.011>.

Forse winst voor Twente Milieu. (2013, 2 juli). *Tubantia,* <http://iturl.nl/snk3l>.

Martijn Mes is universitair hoofddocent bij de afdeling Industrial Engineering en Business Information Systems van de Universiteit Twente (UT). Hij haalde zijn mastertitel Applied Mathematics (2002), deed zijn promotieonderzoek bij de faculteit Management and Governance van de UT (2008) en zijn postdoc aan Princeton University.
E-mail: m.r.k.mes@utwente.nl

Marco Schutten is universitair hoofddocent bij de afdeling Industrial Engineering en Business Information Systems van de Universiteit Twente. Hij haalde zijn mastertitel Applied Mathematics (1992) en zijn PhD in Industrial Engineering (1996). Tot 2006 was hij senior consultant bij ORTEC bv.
E-mail: m.schutten@utwente.nl

Johan van Leeuwaarden

Illustratie: Kheng Ho Toh

Wiskundigen en hokjesdenken

De Amerikaanse site Careercast.com maakt ieder jaar een ranglijst van de beste banen, gebaseerd op vier criteria: werkomgeving, inkomen, stressfactoren en uitzicht op doorgroeien. 'Plezier & Poen', zeg maar. In 2014 voerde het beroep van wiskundige deze lijst aan. Voor velen kwam dit als donderslag bij heldere hemel. In 2013 stond de wiskundige nog op plaats 18. En het beroep van statisticus (een wiskundige gespecialiseerd in data) werd in datzelfde jaar 2014 van plaats 20 naar plaats 3 gekatapulteerd. De dubbele reuzensprong markeerde een ophanden zijnde kentering van de arbeidsmarkt.

Na verbazing komt duiding. Allereerst de vraag: wat is dan een wiskundige? Careercast geeft van alle beroepen een compacte definitie en voor wiskundige was dat in 2014: 'Past wiskundige theorieën en formules toe om te onderwijzen of om problemen op te lossen in een commerciële, educatieve of industriële omgeving.' Nu valt geen enkel beroep te vangen in een regel, maar toch dekt deze regel de lading. De beroepsmogelijkheden voor wiskundigen zijn eindeloos. Ze kunnen aan de slag bij grote bedrijven, overheidsinstellingen, de zakelijke dienstverlening en het onderwijs. Je leert modelleren (complexe vragen kwantificeerbaar maken) en gecombineerd met rekenvaardigheden, data doorgronden en patronen ontdekken, geeft dit een sterke basis voor heel wat carrières. Al die carrières bijeengeveegd noemen we dan 'wiskundige'. En ligt de nadruk op statistische methoden dan kun je het beroep ook 'statisticus' noemen. De statisticus trouwens, heeft een eigen definitie: 'Ordent, analyseert en interpreteert de numerieke resultaten van experimenten en enquêtes.' Nou, niet zo heel sexy zult u zeggen, maar toch goed voor de bronzen plak in 2014. En het wordt nog sterker, want ook een actuaris is half wiskundige/statisticus, kijk maar naar de definitie van Careercast: 'Interpreteert statistieken om kansen te bepalen op ongevallen, ziekte, overlijden en verlies van eigendommen door diefstal of natuurrampen.' In 2014 bezette het drietal samen plaats 1, 3 en 4 (de hoogleraar op plaats 2 voorkwam de triple).

U begrijpt het, sinds 2014 ben ik me gaan verdiepen in dit fenomenale verhaal. In 2015 was plaats 3 voor wiskunde, plaats 1 voor actuaris (net als in 2013), plaats 4 voor statisticus en plaats 6 voor data scientist: vier wiskundige beroepen in de top 6! Nieuwkomer data scientist bestond niet (in de lijsten) voor 2015 en kreeg als definitie: 'Combineert informatica met statistische analyse en andere disciplines om trends die voortkomen uit data te kunnen interpreteren.' Een wiskundige/statisticus dus, die ook kan programmeren. Een soort spin-off beroep. Careercast vat ieder jaar de totale arbeidsmarkt samen in een korte slogan. En in 2015 was dat 'Math matters'.

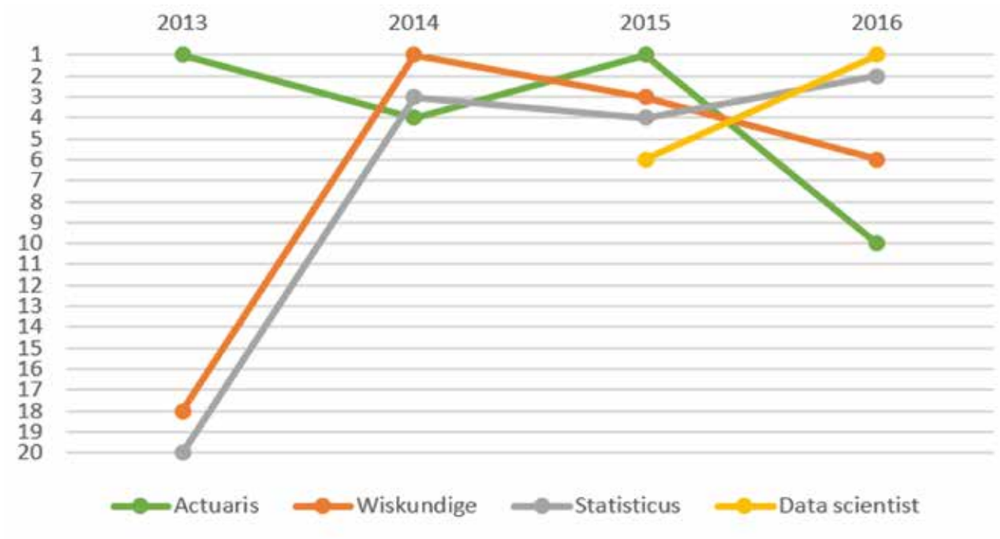
Zelf vergelijk ik een wiskundige met een zelfstandige, je hebt je wiskundige gereedschapskoffer en je kunt overal aan de slag. Je kunt jezelf, afhankelijk van de klus, even-goed actuaris, statisticus of data scientist noemen. Wat we dan niet moeten doen? Wiskunde terug in het traditionele kleine hokje stoppen. En toch vrees ik dat Careercast de eerste stap daartoe heeft genomen. Wat hebben ze gedaan namelijk: door de druk van de oprukkende data scientist (lees: wiskundig zzp'er met praktische instelling) zijn de definities van de andere drie beroepen aangepast. Ik heb dit zelf moeten ontdekken na een kleine data-analyse.

Aanleiding voor de data-analyse was de trendontwikkeling die ik in figuur 1 laat zien. We zien de opmars en dominantie van de wiskundige beroepen. Wel vroeg ik me af wat er in 2016 met de actuaris was gebeurd. Van plaats 1 naar 10 is geen vrije val, maar wel opmerkelijk. Toen ik in de rapporten van 2013–2016 dook, bleek dat in 2016 de definitie van actuaris was aangepast in 'Analyseert de financiële kosten van risico en onzekerheid'. Tsja, een fors kleiner en minder spannend hokje dan in voorgaande jaren. De val viel te begrijpen. En wat was er dan gebeurd met de statisticus? Ook die werd in 2016 opnieuw uitgevonden, maar kwam juist heel wat sterker terug: 'Gebruikt statistische methoden om data te verzamelen en te analyseren teneinde praktijkproblemen op te lossen in het bedrijfsleven, de industrie, de gezondheidszorg en andere sectoren.' Wacht even, dat lijkt verdacht veel op de definitie van wiskundige! Ik rook onraad. Met het hart in de keel baande ik me een weg in de onaangekondigde ruilverkaveling. En daar was dan de definitie van een wiskundige anno 2016: 'Doet onderzoek om wiskundige principes te ontwikkelen en te begrijpen.' Die definitie past mij, maar niet mijn talloze evenknieën buiten de universiteit. Dat doet pijn, snapt u, alsof de klok honderd jaar wordt teruggedraaid. Maar u weet beter. Er staat geen maat op wiskunde. Wiskunde doet er toe.

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

JOB	PERIOD	JOB DEFINITION
MATHEMATICIAN	≤ 2015	Applies mathematical theories and formulas to teach or solve problems in a business, educational, or industrial climate
	≥ 2016	Conducts research to develop and understand mathematical principles
STATISTICIAN	≤ 2015	Tabulates, analyzes, and interprets the numeric results of experiments and surveys
	≥ 2016	Uses statistical methods to collect and analyze data and to help solve real-world problems in business, engineering, healthcare, or other fields
ACTUARY	≤ 2015	Interprets statistics to determine probabilities of accidents, sickness, and death, and loss of property from theft and natural disasters
	≥ 2016	Analyzes the financial costs of risk and uncertainty
DATA SCIENTIST	≥ 2016	Combines information technology, statistical analysis and other disciplines to interpret trends from data.

Tabel 1. Definities van beroepen volgens Careercast.com



Figuur 1. Trend in beste beroepen volgens Careercast.com

Foto: Transport en Logistiek Nederland (TLN)

COMBINING AUTOMATIC MONITORING AND SAMPLE SURVEYS TO PREDICT ROAD USE

To predict future transportation demand, policy makers have always relied on traditional sample surveys under road users and secondary sources like company databases. Nowadays, advanced monitoring systems are installed in the road network, offering more abundant and detailed transport information than traditional methods. This dissertation demonstrates how these big data from road monitoring devices can be used to make more accurate forecasts of the demand for freight transportation.

Yinyi Ma

For policy makers, accurate forecasts of future road use are very important. It allows them to plan road capacity, avoid congestion, and reduce pollution. Traditionally, the data needed for those forecasts have always been collected with surveys that statistical agencies sent out to companies in the transport sector. For companies, filling out these surveys is not only time consuming, the results often also lack in precision because not every company reports their trips in the same detail and with the same time intervals.

The now available road monitoring techniques produce data with very specific characteristics. Loop detectors that are installed in the road surface on many spots generate data on vehicle length. Cameras above the road can be used to read the license plates and track a vehicle between several points. GPS can provide the vehicle trajectory and give insight into the trip chains of freight trucks. The novel statistical approach that I developed to model freight transportation now allows all these different types of data to be combined and seamlessly connected to the more traditional survey data.

A statistical model has been created for predicting freight truck demand in much more detail, both in time and in place. Thanks to the sheer amount of data that automatic monitoring generates, even for a small city within a short time period, there is enough data to be processed and interpreted into improved estimation accuracy, especially when combined with the existing transport surveys. And these surveys will probably not disappear. As valuable as big data from road monitoring is, surveys under road users will still be needed to find out about people's purpose for a trip and on how they combine their trips.

Disaggregating origin-destination pairs

Traditional transportation modeling only considers the number of travelers from a certain origin to a certain destination: the origin-destination (OD) pair. This traditional OD pair is homogeneous with the assumption that each trip is independent from each other. It ignores the fact that travelers plan ahead and choose attributes of each trip such as a mode of transport (car, train, etc.). In my novel statistical model, a new concept of OD tuple is defined as a sequential dependence of the OD pairs. Unlike the OD pair based on anonymous loop detector data, the new concept of OD tuple considers the trip chain of travelers, using high definition road cameras capable of identifying vehicles and providing path. For instance, traditionally we consider there are 500 travelers from in3 to out4 and 600 travelers from in6 to out7, as illustrated in Figure 1. With the help of cameras allocated in HS3 and HS6, we know there are 100 travelers from both in3 to out4 and from in6 to out7. The two OD pairs in3–out4 (500) and in6–out7 (600) can be disaggregated into three OD tuples in3–out4 (400), in6–out7 (500) and in3–out4–in6–out7 (100).

This new concept of OD tuple bridges transportation modeling, which considers only OD pairs, and activity-based modeling, which focuses on travel behavior with the trip chain concept. OD tuple is defined and tested in the freight transportation field for the first time.

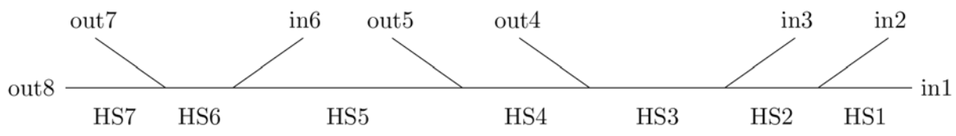


Figure 1. Part of A15 motorway from Hoogvliet to Havens

Hierarchical Bayesian Networks

The thesis demonstrates how different data sources, such as survey data and monitoring big data, can be fused using Hierarchical Bayesian Networks (HBNs), also called belief networks. HBNs belong to the field of Bayesian inference. Bayesian inference updates a prior distribution of parameters to be estimated, e.g. transport demand, with observational data, e.g. flow observations, resulting in an improved posterior distribution of par-ameters. HBNs are probabilistic models that represent a set of stochastic variables and their conditional dependencies via a directed acyclic graph. This HBN approach is originally applied in the multidisciplinary field of transportation planning and official statistics. Combining survey data and monitoring data extends the data collection methods and dramatically increases the accuracy of estimation and forecasting.

Estimating transportation demand

To estimate transportation demand, methods from information theory were applied. In my thesis I proposed to replace the widely used information minimization method by the Kullback-Leibler divergence method, which is also called the information measure or relative entropy method. The Kullback-Leibler divergence is a non-symmetric measure of the difference between two probability distributions P and Q. Typically, P represents the true distribution of data, observations, or a precisely calculated theoretical distribution, while the measure Q represents a prior, theory, or approximation of P.

Compared with the information minimization approach, the Kullback-Leibler divergence method represents better the underlying concept of information minimization for estimating transportation demand and provides accurate results without any approximation. No matter whether the flow data are small or large, the Kullback-Leibler divergence method has lower average deviations between the estimated demand and the assumed ground truth demand than the information minimization method.

From just one to several previous days

To exploit the observed demand in several previous days in forecasting transportation demand, I applied the multi-process model associated with auto-regressive dynamics. Multi-process models are a class of models consisting of a mixture of dynamic linear models. A prior probability is given to each pre-defined combination of the previous days' demand, and the model yields post-erior probabilities for each combination. Usually, one of the combinations gains a unit probability, and the other combinations have zero probability. With this multi-process model, people's experience is taken into account in the pre-defined combinations. The observed data together with errors update the prior probability and end up to the most likely combination.

Multiplicative errors

Three types of errors are considered: prior errors, estimation errors and observation errors. The dissertation creatively established a new method to estimate these errors. The observation or measurement error in most relevant research is additive, assuming the error as a constant. In my research, the multiplicative error is proposed, which is a first time application in this field. This innovative approach of measurement error estimation increases the modeling performance and the prediction accuracy.

Applications

In this dissertation a novel statistical approach is established to estimate transportation demand, integrating traditional survey data and modern monitoring devices generating big data. The results can be used by government agencies, like the United States Department of Transportation or Statistics Netherlands, to produce more accurate and timely statistics and to reduce the response burden on transport companies.

Reference

Ma, Y. (2016). *The Use of Advanced Transportation Monitoring Data for Official Statistics.* Doctoral dissertation. Erasmus University, Rotterdam. Free full text at <http://iturl.nl/snFnmIY>.

Yinyi Ma holds a bachelor degree of Transportation Engineering in China. She studied the Transportation Infrastructure and Logistics major at Delft University of Technology. She recently obtained her PhD from Erasmus University, funded by Statistics Netherlands (CBS). Her research interests include: transportation demand management, travel demand forecasting, transportation performance evaluation and econometrics modeling.
E-mail: yinyimanl@gmail.com

Illustratie: Vectorpector

STATISTIEK EN BIG DATA een samenwerkingsmodel

Mobiele-telefoondata bieden commerciële kansen voor netwerkoperatoren, maar zijn ook een veelbelovende big-databron voor officiële statistieken. In december 2015 hebben Statistics Belgium, Eurostat en Proximus, de grootste netwerkoperator in België, een project opgestart om gezamenlijk hun potentieel te evalueren. Dit heeft niet alleen de kennis vergroot, maar suggereert ook een algemeen zakenmodel voor de samenwerking tussen statistische instituten en privébedrijven die big data bezitten. Het bundelen van datasets en van middelen levert resultaten op die geen van de partners alleen bereiken kan. Verder werden uit de concrete samenwerking een aantal 'beste praktijken' afgeleid die ook elders toegepast kunnen worden.

Marc Debusschere & Freddy De Meersman

Maatschappelijke en technologische veranderingen hebben recent geleid tot wat men big data noemt: de explosieve toename van immense volumes ongestructureerde data die continu gecreëerd worden door sensoren en camera's, satellieten, *machine-to-machine* communicatie, e-business, elektronische betalingen en afhalingen, allerlei activiteit op het internet, sociale media: de derde datarevolutie. Deze stortvloed bevat waardevolle

informatie die gebruikt kan worden voor operationele en commerciële doeleinden, maar ook om er officiële statistieken mee te produceren.

Mobiele-telefoondata vormen een veelbelovende big-data-bron voor statistieken, vooral in de domeinen bevolking, migratie, toerisme en mobiliteit. Hun gebruik zal naar verwachting leiden tot snellere en zelfs onmiddellijke statistieken, in veel meer detail, met bijna volmaakte dekking, zonder antwoordbias, tegen lagere kosten en zonder de noodzaak burgers en ondernemingen lastig te vallen. Bovendien geven ze wellicht directe ogenblikkelijke toegang tot fenomenen die tot nu onmeetbaar waren (zoals reëel aanwezige versus geregistreerde bevolking of gedetailleerde pendelpatronen naargelang de werkdag, weersomstandigheden, ...).

De toegang en het gebruik worden echter bemoeilijkt door tal van factoren. Statistische instituten zijn geen eigenaars van de data en hebben er daardoor nauwelijks kennis over, ze beschikken niet over in de IT-infrastructuur om dergelijke grote volumes te behandelen en wettelijke regelingen rond statistische exploitatie zijn ontoereikend. Bovendien is het problematisch om netwerkoperatoren te dwingen toegang te geven, want mobiele-telefoon-data liggen niet zomaar klaar voor gebruik; er is integendeel een aanzienlijke begininvestering vereist om netwerksignalen om te zetten in bruikbare ruwe data.

Positief is dan weer dat de meeste netwerkoperatoren intussen beseffen dat de data ook voor hen potentieel grote waarde hebben, en dat ze begonnen zijn met die investering. Hun belangrijkste doel is daarbij netwerkoptimalisatie, maar steeds meer operatoren zien ook het commercieel potentieel. De initiële investering is echter niet de enige uitdaging bij het commercialiseren. Vanuit hun kerntaak, het beheer van een mobiel netwerk, vormden deze data voor operatoren louter 'afval'. De omzetting ervan naar geldige en accurate informatie is veel minder evident dan ze zich doorgaans realiseerden en vereist capaciteiten die ze niet in huis hebben. Bovendien beschikken de operatoren alleen maar over mobiele-telefoon-data, rijk maar met beperkingen die spoedig duidelijk worden.

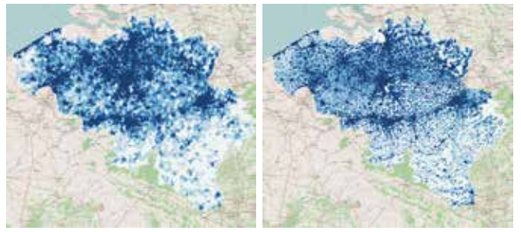
Deze situatie waarin zowel statistische instituten als netwerkoperatoren geconfronteerd worden met uitdagingen die beter met partners dan alleen aangepakt kunnen worden, biedt uiteraard een stevige basis voor wederzijds voordelige samenwerking.

Innovatieve analyse van netwerksignalen

Statistics Belgium, Eurostat en de belangrijkste netwerkoperator in België, Proximus, hebben in december 2015 een project gestart om gezamenlijk de mobiele-telefoondata van Proximus te onderzoeken op hun informatie-inhoud en mogelijke toepassingen, zowel commercieel als statistisch. Bij het begin werd, omwille van het exploratieve karakter, het bereik met opzet open gelaten, maar de ambitie was toch om al concrete resultaten te hebben tegen april 2016, en dus werden duidelijke onderzoeksdoelstellingen geformuleerd. De projectgroep, verder aangevuld met het Joint Research Centre (JRC) van de Europese Commissie, bracht elf specialisten samen die alle mogelijke invalshoeken afdekten en expertise leverden op technisch, statistisch, datawarehouse, IT, GIS, commercieel en inhoudelijke vlak.

Het project werd opgezet vanuit een stapsgewijze logica, beginnend met het schatten van de *werkelijk aanwezige bevolking*, met de bedoeling daarna te focussen op de *ingezeten bevolking* (na bepaling van de 'gebruikelijke woonplaats') en in een later stadium op *arbeidspendel*, *arbeidsmobiliteit* op korte of lange termijn, *toerisme* en *migratie* (na het vastleggen van de werkplaats en alle andere aspecten van de 'gebruikelijke omgeving'). Om privacykwesties en de daardoor eventueel veroorzaakte vertragingen te vermijden, werden in het eerste stadium alleen geaggregeerde data gebruikt, zonder tracering van individuele mobiele toestellen.

De eerste analyse, van de werkelijke en ingezeten bevolking en de bevolkingsdichtheid, was voor belangrijke aspecten innovatief. In tegenstelling tot de meeste voorgaande studies die een toestel in tijd en ruimte lokaliseren via *call detail records* (CDR's), die nodig zijn om het gebruik van een toestel te kunnen factureren, is deze studie gebaseerd op netwerksignalen, met een frequentie die zowat tienmaal hoger ligt dan CDR-registraties.



Figuur 1. Bevolkingsdichtheid per km² gebaseerd op mobiele-telefoon-data (links) en Census 2011 (rechts)

Ook combineert de analyse mobiele-telefoon-data met statistische datasets, wat het mogelijk maakt om deze tegenover elkaar te valideren en volledig nieuwe informatie te genereren. Om beide te koppelen moesten de geografische eenheden van de mobiele-telefoondata (de 11.000 mobiele-netwerkcellen op het Belgisch grondgebied, veelhoeken rondom een antenne) omgezet worden in de standaard roostervierkanten van 1 km² van de statistische datasets, en vice versa. In toekomstige studies zal de precisie van de lokalisatie nog toenemen door het gebruik van kleinere vierkanten en zelfs triangulatie van individuele toestellen.

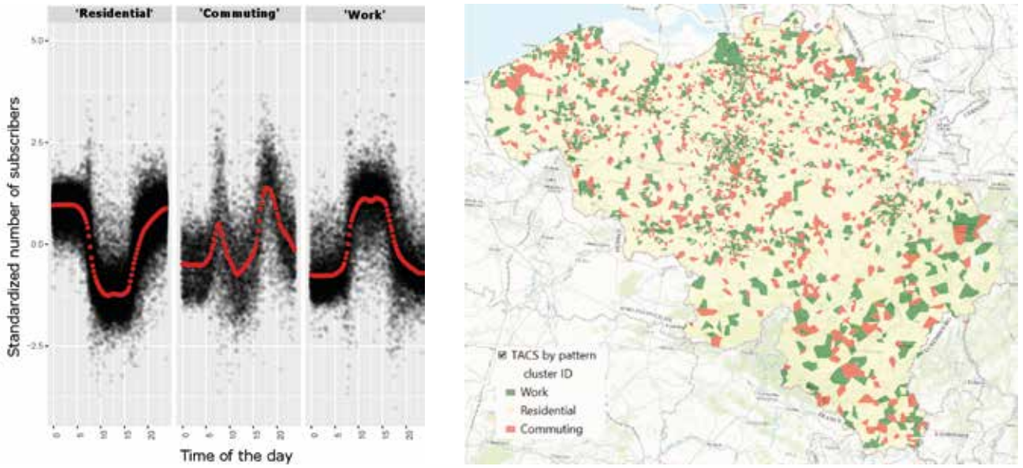
De eerste resultaten werden zoals gepland afgewerkt en gepubliceerd (De Meersman e.a., 2016). Ze kunnen via enkele figuren geïllustreerd worden. De eerste (figuur 1) toont de bevolkingsdichtheid aan de hand van de mobiele-telefoon-telling om 4 uur 's morgens op donderdag 8 oktober 2015 (links) en de Census van 2011 (rechts). De Pearson-correlatie tussen beide datasets bedraagt 0,85, een duidelijke aanwijzing dat mobiele-telefoondata een geldige en accurate maat bieden voor de bevolkingsdichtheid.

Figuur 2 toont de resultaten van een clusteranalyse op de waarnemingen van de werkdag, met een typisch 'residentieel' patroon (mobiele telefoons die 's morgens vertrekken en 's avonds terugkeren), 'pendel'-patroon (twee pieken in de spits) en 'werk'-patroon (mobiele telefoons die 's morgens aankomen en 's avonds vertrekken). Een geografische representatie daarvan lijkt logisch en zinvol. Figuur 3 biedt een meer gedetailleerde voorstelling van de clusteranalyse op kaart.

Het project loopt verder en richt zich nu op het toevoegen van tijdruimtelijke datasets (bv. bodemgebruik, stedelijk versus landelijk gebied) die de errorvariatie naar beneden kunnen halen en zo de correlaties verhogen, en op nieuwe mobiele-telefoon-datasets om verdere onderzoeksvragen aan te pakken.

Een zakenmodel tekent zich af

De nauwe samenwerking tussen een statistisch instituut en een mobiel-netwerkoperator heeft in de praktijk aangetoond dat samen meer waarde gecreëerd kan worden



Figuur 2. Werkdagcellen aangeduid als 'residentieel', 'pendel' of 'werk', met voorstelling op kaart

Figuur 3. Geografische voorstelling van de clusters in de buurt van Brussel en Leuven

dan apart. De voordelen voor de officiële statistiek zijn duidelijk. Zonder toegang tot de data zijn statistieken eenvoudigweg niet mogelijk, maar de metadata, technische expertise, infrastructuur voor dataopslag en -verwerking en de praktijkgevallen die operatoren kunnen leveren zijn evengoed essentiële bijdragen.

Maar ook mobiele-netwerk-operatoren winnen hierbij. De gegeocodeerde statistische datasets op maat die ze kunnen combineren met hun eigen data om zo de commerciële waarde daarvan drastisch te verhogen, zijn wellicht het meest tastbare voordeel. Maar ze kunnen ook profiteren van de statistische en domeinexpertise die statistische instituten in huis hebben, verworven na ruim 200 jaar ervaring met het extraheren van geldige en betrouwbare informatie uit ruwe data! Bovendien zet samenwerking met officiële statistici een kwaliteitsstempel op de datasets die operatoren leveren aan commerciële klanten. Tenslotte, en niet in het minst, biedt het hen de mogelijkheid zich te profileren als een sociaal verantwoordelijke onderneming die vrij bijdraagt aan het algemeen goed.

Samenvattend suggereert dit een zakenmodel waarbij mobiele-netwerk-operatoren en statistische instituten beide investeren en samenwerken om hun respectievelijke en niet-competitieve commerciële en statistische doelstellingen te verwezenlijken. In het ideale geval leidt deze publiek-private samenwerking dan tot een overeenkomst op lange termijn die een statistisch instituut toelaat om op duurzame wijze statistieken te produceren die geheel of gedeeltelijk op mobiele-telefoondata gebaseerd zijn, terwijl de mobiel-netwerk-operator die ze levert in ruil aanvullende datasets en statistische ondersteuning krijgt om deze beter te commercialiseren.

Wat we geleerd hebben

De concrete ervaring met het gezamenlijk exploiteren van mobiele-telefoon-data en andere datasets door Statistics Belgium, Proximus en Eurostat heeft inzichten opgeleverd die ook in andere soortgelijke contexten nuttig kunnen blijken.

1. ***Grijp de kans***
   Zoek als statistisch instituut dat big data wil exploiteren contact met een data-eigenaar die daar ook brood in ziet.
2. ***Verzamel de juiste mensen***
   Bij een netwerkoperator wil de juridische dienst risico vermijden, onderzoeksunits hebben vaak geen link met de productie en Sales & Marketing wil verkopen, maar Business Development of Innovation zoekt nieuwe producten op basis van big data. Een projectteam moet allerlei technische, IT, statistische en domeincompetenties bundelen.
3. ***Vind bondgenoten en partners***
   In statistiek, de privésector, academia. Ook internationaal!
4. ***Definieer een concreet project en begin klein***
   Start laagdrempelig met op korte termijn haalbare concrete doelen.
5. ***Ken je eigen limieten en zie wat de ander nodig heeft***
   De officiële statistiek heeft mobiele-telefoon-data nodig, operatoren missen de extra data en expertise om deze ten volle te exploiteren. Om een joint venture te doen slagen, moeten statistici de operator helpen geld te verdienen en moeten operatoren een duurzame productie van statistieken mogelijk maken.
6. ***Investeer en maak vervolgens bekend hoe waardevol je bent***
   Statistische instituten moeten investeren in gegeocodeerde datasets en data science, operatoren in het omzetten van netwerksignalen in data. Beide moeten de ander vertellen welke waardevolle activa ze te bieden hebben.
7. ***Garandeer absolute vertrouwelijkheid***
   Mobiele-netwerkoperatoren willen hun data en know-how niet bij de concurrentie zien belanden. Vertrouwelijkheid is een wettelijk verankerde kernwaarde voor statistische instituten, maar operatoren zijn daar misschien onvoldoende van op de hoogte.
8. ***Wees aandachtig voor wettelijke en privacykwesties***
   Echte of vermeende privacyproblemen zullen elk gebruik van big data ondermijnen en vormen een serieus reputatierisico voor zowel statistische instituten als data-eigenaars.

Besluit

Diverse modellen om statistieken te produceren op basis van mobiele-telefoon-data zijn mogelijk: operatoren kunnen wettelijk verplicht worden om data of resultaten te leveren, statistische instituten kunnen de data kopen, een integrator kan de data van diverse operatoren centraliseren, verwerken en doorspelen naar een statistisch instituut of direct naar de gebruikers. Elk van deze modellen vertoont belangrijke gebreken die ze in de praktijk niet werkbaar en/of wenselijk maken.

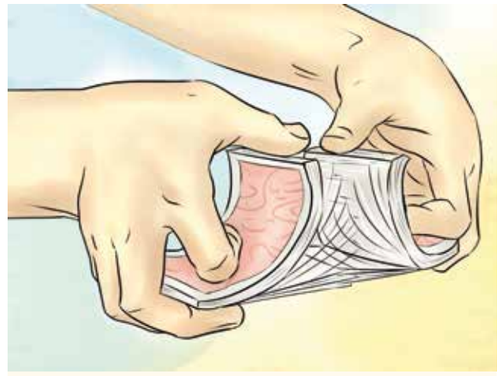
Het samenwerkingszakenmodel dat uitgetest werd door Statistics Belgium, Proximus en Eurostat biedt een aansporing voor zowel statistische instituten als privépartners om data en middelen op gelijkwaardige wijze te bundelen en zo waarde voor beide partijen te creëren, op een manier die wettelijk, operationeel en zakelijk hanteerbaar is. Bovendien heeft de samenwerking het potentieel om uit te groeien tot een relatie op langere termijn, essentieel voor het produceren van statistische tijdreeksen.

Literatuur

De Meersman, F., Seynaeve, G., Debusschere, M., Lusyne, P., Dewitte, P., Baeyens, Y., Wirthmann, A., Demunter, C., Reis, F., & Reuter, H.I. (2016). *Assessing the quality of mobile phone data as a source of statistics*. Paper of the European Conference on Quality in Official Statistics <www.ine.es/q2016/docs/q2016Final00163.pdf>.

Marc Debusschere is Coordinator Administrative & Big Data bij Statistics Belgium en betrokken bij diverse Europese projecten over het gebruik van big data voor statistiek.
E-mail: marc.debusschere@economie.fgov.be

Freddy De Meersman is actief in de unit Business Development & Innovation van Proximus ('incumbent' netwerkoperator in België met een mobiel marktaandeel van iets meer dan 40%).
E-mail: freddy.demeersman@proximus.com



KANSREKENING VAN ALLEDAG

De lezers van *STAtOR* kennen Henk Tijms als de schrijver van columns over allerlei aspecten van de kansrekening. Minder bekend zal zijn dat hij ook regelmatig bijdraagt aan de Numberplay-blog van de *New York Times* en daar een zekere reputatie heeft opgebouwd.

Recent is een boekje verschenen waarin 16 van zijn columns zijn gebundeld en ook voor iemand die de afgelopen jaren steeds aandachtig de *STAtOR*-columns las is het een verrassend weerzien.

De onderwerpen bestrijken een een breed terrein. De titel wijst op het voorkomen van de kansrekening in het alledaagse leven, dat is enigszins misleidend tenzij dat leven zich grotendeels afspeelt in gokhallen en loterijkantoren.

De heldere stijl, die de *STAtOR*-lezers van Henk gewend zijn, zorgt voor een prettig leesbare inhoud. Ook de onderkoelde humor die in veel columns aanwezig is draagt daaraan bij. Wel zal menigeen zich hier en daar enige moeite moeten getroosten om de formules te doorzien en te begrijpen. Maar dat mag, zeker voor onze lezers, geen bezwaar vormen.

Op het gevaar af beschuldigd te worden van een gevalletje 'WC-eend', kan ik de aanschaf van dit boekje ten zeerste aanraden. De komende feestdagen kunnen daar een goede aanleiding voor zijn.

Gerrit J. Stemerdink

Henk Tijms, *Kansrekening van alledag, een wereld vol verrassingen*. Amsterdam: Epsilon Uitgaven, 2016. Isbn 9789050411585, € 12,-, 76 blz.

Een gokje in de loterij

Binnenkort vindt de oudejaarstrekking van de Staatsloterij weer plaats. Kranten, radio en televisie zijn dan weer uit op pakkende uitspraken over de kans om de hoofdprijs te winnen in de verschillende loterijen. Voor de Staatsloterij is geen hogere wiskunde nodig om de winkans te geven. Om en nabij de vier miljoen loten worden verkocht voor de oudejaarstrekking en dat betekent dat de kans om met een enkel lot de jackpot te winnen ongeveer 1 op de vier miljoen is, een kans die te vergelijken is met de kans om met een zuivere munt 23 keer keer achterelkaar kop te gooien. Sommige wiskundigen doen op grond van een dergelijke zeer kleine kans de uitspraak dat meespelen in de Staatsloterij irrationeel is en voor de dommen zou zijn. Ik zie dit echter anders. Beter een staatslot in de hand dan een kapitaal aan vuurwerk in de lucht. Met een staatslot kun je dagenlang dromen over wat je met al het geld gaat doen als je op oudejaarsnacht miljonair zou worden. Een uitgave voor een staatslot is als een uitgave in een casino waar je heengaat om een plezierige avond te hebben en niet met het vaste doel om geld te winnen.

In de Nederlandse lotto ligt de vraag of meespelen verstandig is anders dan bij de Staatsloterij. In de Nederlandse lotto is de kans om de jackpot te winnen met een enkel ingevuld rijtje ongeveer 1 op de 49 miljoen. Je moet zes verschillende getallen getrokken uit de getallen 1 tot met 45 goed hebben alsmede de jackpotkleur die getrokken wordt uit zes verschillende kleuren. Het aantal mogelijke combinaties van de zes verschillende getallen wordt gegeven door de binomiaalcoefficiënt $\binom{45}{6} = 8\,145\,060$. Dus de kans dat je met een enkel rijtje de winnende zes getallen en de kleur goed hebt is gelijk aan

$$w = \frac{1}{8\,145\,060} \times \frac{1}{6} = \frac{1}{48\,870\,360}.$$

Hoe onvoorstelbaar klein de winkans op de jackpot is, is als volgt te illustreren. De kustlijn van Hoek van Holland naar Den Helder wordt overbrugd door net zoveel muntstukken van 1 euro achterelkaar te leggen als er mogelijke uitkomsten van de lottotrekking zijn. Stel dat jouw gemerkte euro daar ergens tussen ligt en dat een strandjutter een door het lot bepaalde euro opraapt, dan is de kans dat de jutter precies jouw euro opraapt hetzelfde als de kans om met één ingevuld rijtje zes goed en de kleur goed te hebben. Tel uit je winst! Eerlijkheidshalve moet hierbij gezegd worden dat de kans om de jackpot te winnen met een eenmalige investering van 2 euro in één rijtje in theorie ongeveer gelijk is aan 1 op de 41,5 miljoen. Dit vanwege het feit dat je een gratis lot krijgt bij precies twee van de zes getrokken getallen goed, daarbij wel aannemend dat je elke keer het gratis lot inlevert.

In 2002 was de deelnameprijs aan de lotto nog 1 euro per ingevuld rijtje. In die tijd was het aantal ingevulde rijtjes gemiddeld 3 miljoen per trekking en viel de jackpot gemiddeld één keer in de 17 weken. In 2014 werd de deelnameprijs per ingevuld rijtje drastisch verhoogd tot 2 euro. Deze prijsverhoging leidde al snel tot een teruglopende deelname aan de lotto. Verwonderlijk is dit natuurlijk niet wanneer de deelnameprijs van 2 euro in het licht wordt gezien van de uiterst kleine kans om de jackpot te winnen. De aanpassing dat behalve zes getallen ook de jackpotkleur goed geraden moet worden om de jackpot te winnen, was gedaan om te bereiken dat de jackpot vaker zou blijven staan en dusdanig zou groeien dat een jackpot van 20 miljoen niet ongewoon zou worden. Vraagtekens kunnen worden gezet bij de vraag of dit een verstandig beleid is. Mijn inschatting is dat lottospelers in Nederland een voorkeur hebben voor een jackpot die vaker valt en dan maar wat minder hoog is. Een jackpot van vijf miljoen is tenslotte ook heel veel geld. Zoals eerder opgemerkt, rond de tijd van de invoering van de euro in 2002 werden gemiddeld ongeveer 3 miljoen rijtjes per trekking ingevuld. Hoe zou dit zijn in de huidige situatie? De lotto geeft de informatie niet. Als we zouden weten hoeveel trekkingen gemiddeld genomen nodig zijn voordat de jackpot valt, dan zouden we een schatting kunnen maken van het gemiddeld aantal rijtjes dat per trekking wordt ingevuld sinds de prijsverhoging tot 2 euro. Zo'n schatting kunnen we vinden met enige eenvoudige kansrekening. Daartoe maken we gebruik van een basisresultaat uit de kansrekening dat stelt dat bij n onafhankelijke experimenten elk met kans p op succes de kansverdeling van het totale aantal successen bij goede benadering Poisson-verdeeld is met verwachtingswaarde np wanneer n heel groot is en p heel klein. In het bijzonder, de formule $1 - e^{-np}$ is van toepassing op de kans op tenminste één succes. Dit leidt tot een simpele relatie tussen r en a, waarbij

Vincent van Gogh, *De armen en het geld; trekking van de Staatsloterij in Den Haag*, 1882. Collectie: Van Gogh Museum, Amsterdam

r = gemiddeld aantal rijtjes per trekking
a = gemiddeld aantal trekkingen nodig voor de jackpot.

Voor het al of niet vallen van de jackpot is sprake van een zeer groot aantal rijtjes elk met een zeer kleine succeskans. Dus de kans dat de jackpot valt bij een lottotrekking met r ingevulde rijtjes is bij goede benadering gelijk aan $1 - e^{-rw}$, waarbij w de bovengenoemde winkans is op de jackpot met een enkel rijtje. Dit resultaat impliceert weer dat gemiddeld genomen $1/(1 - e^{-rw})$ trekkingen nodig zijn voor het vallen van de jackpot. Dit geeft de benaderingsrelatie $a \approx 1/(1 - e^{-rw})$. Dit kunnen we nog inzichtelijker maken. Voor x dicht bij 0, geldt $1 - e^{-x} \approx x$. Derhalve vinden we de intuïtief voor de hand liggende benaderingsregel

$$r \approx \frac{1}{wa}.$$

Uit gepubliceerde gegevens van de lotto blijkt dat in de periode van januari 2014 tot december 2016 de jackpot slechts 3 keer viel in zo'n 185 trekkingen, waarbij in één geval de jackpot maar werd uitgekeerd bij 5 van de 6 en de kleur goed (in 2013 viel de jackpot 3 keer). Nemen we de voorzichtige schatting $a = 40$, dan komen we tot gemiddeld ongeveer 1,22 miljoen ingevulde rijtjes per trekking. Dit is meer dan een halvering van het gemiddeld aantal ingevulde rijtjes per trekking zo'n 15 jaar geleden. De Nederlandse lotto zou er verstandig aan doen de deelnameprijs per rijtje weer op 1 euro te zetten (in de Duitse lotto is het ook 1 euro) en het aantal jackpotkleuren te halveren van 6 naar 3. Een duur lot en een jackpot die heel zelden valt zijn niet bevorderlijk voor de deelname. Misschien doet de lotto er verstandig aan de klant eens om zijn mening te vragen!

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

SLIM SAMENWERKEN AAN EEN EVENWICHTIGE BEDBEZETTING

Saskia Cornelissen, Maartje van de Vrugt, Eric Smits & Thom Timmerhuis

Een aantal verpleegafdelingen van het Jeroen Bosch Ziekenhuis (JBZ) in 's Hertogenbosch heeft in 2012 en in het eerste kwartaal van 2013 te maken gehad met een onevenredige bedbezetting. Zowel een (te) hoge als lage bedbezetting is een risico voor de kwaliteit van zorg en arbeid, zoals verderop in dit artikel toegelicht wordt. In opdracht van de specialisten en de manager bedrijfsvoering van het JBZ, en ondersteund door een wiskundige en organisatiedeskundige, is een aantal mogelijke interventies voor het beddenhuis van het JBZ prospectief cijfermatig geanalyseerd. Dit resulteerde in een adequate interventie voor optimalisering van de bedbezetting voor de betrokken specialismen, waardoor de kwaliteit van zorg en arbeid aantoonbaar verbeterd is. In dit artikel worden de gebruikte methoden en de resultaten van het onderzoek kort toegelicht. Zie voor een uitgebreid verslag van het onderzoek hoofdstuk 3 van het proefschrift van Van de Vrugt (Van de Vrugt, 2016)

Probleemstelling

Door verkorting van de ligduur van neurologische patiënten waren er, voor aanvang van het onderzoek, op de afdeling Neurologie relatief veel onbenutte bedden. De specialismen Interne geneeskunde en Longgeneeskunde ervoeren daarentegen overbezetting van hun verpleegafdelingen. In de praktijk betekende dit dat er patiënten Interne geneeskunde verspreid door het hele beddenhuis lagen en dat de afdeling Neurologie patiënten van veel verschillende specialismen op de afdeling ontving. Daarnaast werden patiënten vaak, wanneer er op de eigen afdeling weer een bed vrij kwam, teruggeplaatst naar de eigen afdeling. Dit resulteerde in een extra overdrachtsmoment hetgeen de kwaliteit van zorg negatief kon beïnvloeden. Ook de werkbeleving van verpleegkundigen werd negatief beïnvloed door zowel de piek- en dalbelasting als de verpleegkundige zorg voor patiënten van andere specialismen. Tot slot kostte de artsenvisite meer tijd wanneer patiënten over meerdere afdelingen verspreid lagen.

Het JBZ heeft daarom een projectgroep ingericht bestaande uit een neuroloog, een internist, een longarts, een afgevaardigde van het bureau opname, de manager bedrijfsvoering, het unithoofd spoedeisende hulp en de unithoofden van de drie betrokken verpleegafdelingen. De projectgroep werd ondersteund door een adviseur patiëntenlogistiek en een wiskundige. Doelstelling van de projectgroep was het verbeteren van de kwaliteit van zorg en arbeid op alle betrokken afdelingen, zonder daarbij de weigeringskans (kans dat een patiënt niet opgenomen kan worden omdat alle bedden bezet zijn) te verhogen.

Mogelijke interventies

Het totale beddenhuis van het JBZ bestaat uit 730 bedden, waarvan 64 bedden voor Neurologie, 32 voor Longgeneeskunde en 96 voor Interne geneeskunde. Neurologie vormde al een samenwerkingsafdeling met Longgeneeskunde. Dit wil zeggen dat er standaard longpatiënten op de afdeling Neurologie geplaatst konden worden en dat de afdeling hier op voorbereid was. Uit een analyse van data uit het Ziekenhuis Informatie Systeem bleek dat de gezamenlijk beschikbare beddencapaciteit voor de drie betrokken afdelingen grotendeels voldoende was om aan de vraag te voldoen.

De historische bedbezetting van deze afdelingen is weergegeven in figuur 1. Opmerkelijk in figuur 1 is de piek begin 2013; door de strenge winter werden in die periode relatief veel patiënten met een longontsteking opgenomen.

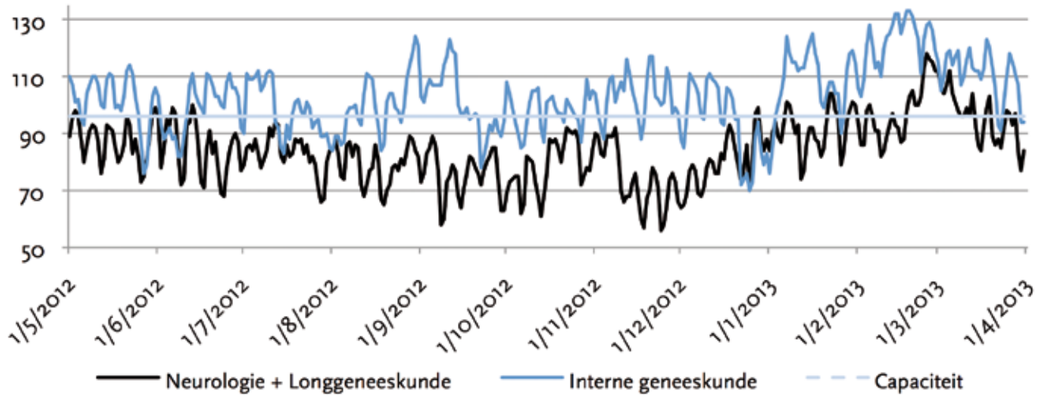
Met een Erlang-wachtrijmodel (M/G/c/c) is doorgerekend wat de minimaal benodigde capaciteit is om maximaal 5% weigeringskans te garanderen voor ieder specialisme. Hieruit volgde dat Interne geneeskunde 99 bedden nodig had, Longgeneeskunde 43 en Neurologie 47 (in totaal dus 189 bedden). Omdat het totaal aantal beschikbare bedden 192 was, bleek ook hieruit dat capaciteit toevoegen aan het beddenhuis niet nodig was. Er is daarom gezocht naar interventies om de bedden te herverdelen of poolen voor de betrokken specialismen. De projectgroep zag drie mogelijke oplossingen:

- het openen van een opnameafdeling;
- een interne verhuizing;
- een vaste samenwerkingsafdeling voor Interne geneeskunde.

Deze oplossingen worden achtereenvolgens kort toegelicht.

Openen van een opnameafdeling

Een opnameafdeling is een bufferafdeling voor spoedpatiënten die indien nodig overgeplaatst worden naar hun definitieve afdeling. Het verschil tussen een opnameafdeling en een gewone verpleegafdeling is dat op een opnameafdeling patiënten een korte ligduur hebben (vaak maximaal 48 uur), er vaker een artsenvisite is om



Figuur 1. Bedbezetting van de betrokken specialismen, met rechte lijn op 96 bedden (capaciteit)

24U/DAG OPNAMES			ALLEEN 'S NACHTS OPNAMES		
#bedden	bedbezetting	weigeringskans	#bedden	bedbezetting	weigeringskans
5	82%	46%	3	61%	37%
8	73%	20%	4	55%	21%
10	66%	9%	6	43%	5%
12	58%	3%			

Tabel 1. Resultaten Erlangmodel voor opnameafdeling waar patiënten maximaal 48 uur mogen verblijven

snel ontslag te bevorderen en medewerkers breed opgeleid zijn om acute zorg te verlenen. De werking van een opnameafdeling voor alle beschouwende specialismen is, voor verschillende scenario's met betrekking tot verblijfduur en overplaatsingsbeleid, geschat met behulp van een Erlangmodel met tijdsafhankelijke aankomsten en ontslagen en de Modified Offered Load approximatie (Massey & Whitt, 1994). Input voor dit model is de aankomsttijdenverdeling van alle spoedpatiënten en een schatting van het aantal patiënten dat gebruik gaat maken van de opnameafdeling, welke gemaakt is met behulp van een lijst opname-indicaties die volgens de specialisten geschikt zijn voor opname op een opnameafdeling. Er zijn verschillende scenario's met betrekking tot verblijfduur en overplaatsingsbeleid onderzocht, waarvan er twee in tabel 1 zijn weergegeven.

De opnameafdeling blijkt te klein voor een efficiënte bedbezetting. Het JBZ streeft naar 80%-90% bedbezetting (niet in klinische verpleegdagen, maar gerealiseerde opname- en ontslagtijd) en weigeringskans ≤5%; de onderzochte scenario's presteren veel lager dan deze norm.

Interne verhuizing

Bij de tweede oplossing, een interne verhuizing, is gezocht naar mogelijke combinaties van beddenunits en specialismen. Omdat het aantal mogelijkheden beperkt was, is dit met de hand gedaan, waarbij toegewerkt werd naar het aantal benodigde bedden zoals dat met de M/G/c/c-wachtrij berekend was. Dit leverde geen oplossing op waarin alle specialismen een voldoende aantal bedden toegewezen kregen zonder de bestaande beddenunits te versnipperen (één specialisme verspreid over meer dan één verpleegvloer) of verbouwen.

Vaste samenwerkingsafdeling voor Interne geneeskunde

Voor de derde oplossing is onderzocht of Neurologie de samenwerkingsafdeling voor patiënten van Interne geneeskunde kan worden. Hiertoe hebben de internisten vastgesteld welke patiëntencategorieën medisch gezien geschikt zijn om op de samenwerkingsafdeling te plaatsen en is het aantal benodigde bedden voor deze patiënten geschat. De capaciteit op de afdeling Neurologie bleek toereikend om deze patiënten te plaatsen, op basis van het verwachte aantal benodigde bedden uit historische data van de bedbezetting. Vervolgens zijn de scholing- en trainingsbehoefte geïnventariseerd en zijn er richtlijnen opgesteld over het visitelopen, de medische supervisie en implementatie-ondersteuning. Deze oplossing is in september 2013 geïmplementeerd.

Effectmeting

In januari 2014 is een effectmeting uitgevoerd voor twee prestatiematen: (1) bedbezetting en (2) het aantal overplaatsingen van 'vreemd specialisme'-patiënten van de afdeling Neurologie naar een andere afdeling. Daarnaast zijn de resultaten voor het betrokken personeel geëvalueerd op de vergaderingen van de projectgroep.

GEMIDDELDEN	VOOR	NA
Bedbezetting afdeling Neurologie	45,25	44,87
waarvan Neurologie	37,98	36,82
waarvan Longgeneeskunde	5,80	5,14
waarvan Interne geneeskunde	0,77	2,55 (p=0,00)
waarvan overige specialismen	0,71	0,36 (p=0,03)
Bedbezetting afdeling Interne geneeskunde	80,77	80,88
Bedbezetting afdeling Longgeneeskunde	24,85	24,53

Tabel 2. Bedbezetting voor en na interventie, p-waarden uit de wilcoxonrangtekentoets

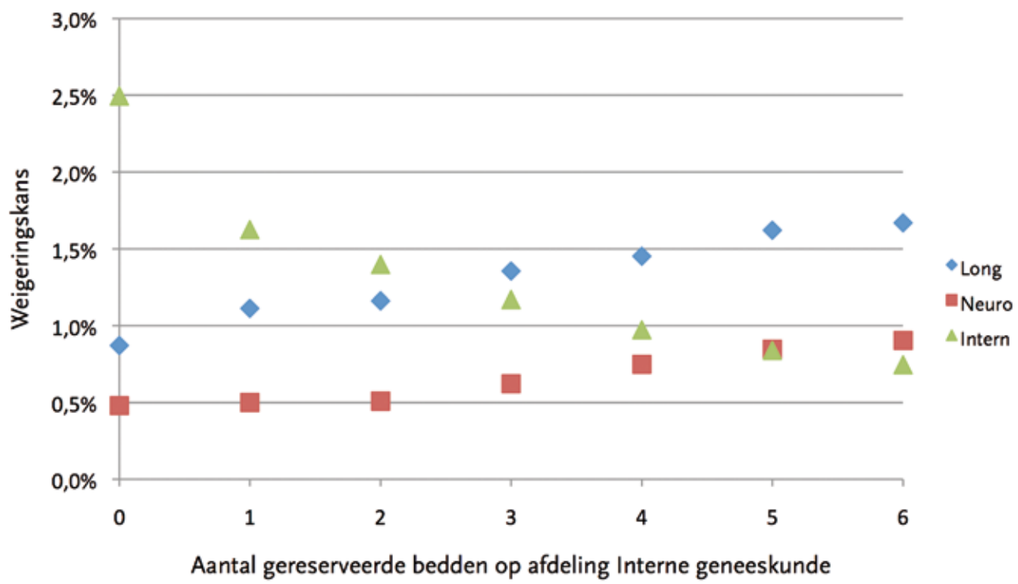
SPECIALISME	VOOR INTERVENTIE	NA INTERVENTIE	VERSCHIL
Interne geneeskunde	0,62	0,25	-60%
Overig	0,10	0,05	-50%

Tabel 3. Aantal overplaatsingen per patiënt opgenomen op afdeling Neurologie

Hoewel de verandering in bedbezetting klein is, zijn er significant meer bedden door Interne geneeskunde benut op Neurologie (tabel 2). Tegelijkertijd is het aantal bedden gevuld door overige specialismen daar afgenomen. Uit tabel 3 blijkt dat er minder patiënten worden overgeplaatst van Neurologie naar andere afdelingen. Uit de evaluatie blijkt dat Internisten minder tijd kwijt zijn aan visite lopen en meer vertrouwen hebben in de kwaliteit van zorg op de uitwijkafdeling. De afdeling Neurologie ervaart minder piek- en dalbelasting en voelt zich bekwaam voor de aanwezige patiëntencategorieën. Aandachtspunt is de verzwaring van de complexiteit van zorg op afdeling Interne geneeskunde, doordat patiënten met relatief lichte indicaties ('makkelijke' patiënten) uitwijken naar Neurologie.

Beslisregel

Na de implementatie ontstond er behoefte aan een beslisregel om preciezer te bepalen hoe de afdelingen moeten samenwerken. Hiertoe is een simulatie uitgevoerd, waarin verschillende beslisregels zijn onderzocht. Input voor de simulatie was de opname- en ontslagdata van de patiënten van de betrokken afdelingen, waarbij een fractie van de patiënten willekeurig aangewezen werd als 'geschikt voor de uitwijkafdeling' omdat de diagnosecodes niet beschikbaar waren in de dataset. Alle onderzochte beslisregels zijn op basis van drempelwaarden: patiënten wijken uit wanneer er nog *x* of minder beschikbare bedden zijn op de afdeling Interne geneeskunde. Omdat het aantal patiënten van Interne



Figuur 2. Weigeringskans wanneer er x bedden gereserveerd worden op Interne geneeskunde én twee bedden gereserveerd op Neurologie/Longgeneeskunde voor Neurologie en Longpatiënten

geneeskunde groter is dan het aantal patiënten van de andere specialismen zijn er in de beslisregels ook bedden gereserveerd voor Neurologie en Longgeneeskunde; de laatste *y* beschikbare bedden worden vrij gehouden voor Neurologie en Longgeneeskunde. Wanneer een patiënt niet opgenomen kan worden op de eigen afdeling of uitwijkafdeling, wordt deze geweigerd.

Uit figuur 2 blijkt dat de weigeringskans voor Interne geneeskunde verbetert, terwijl voor Neurologie en Longgeneeskunde weinig verandert. In deze figuur refereert '0' op de horizontale as aan de situatie zonder georganiseerde uitwijk. De projectgroep heeft op basis van deze resultaten ervoor gekozen om drie bedden te reserveren op Interne geneeskunde (x=3).

Conclusie

Bij aanvang van dit project was er sprake van disbalans in de bedbezetting van de betrokken verpleegafdelingen; op de afdeling Interne geneeskunde waren er te weinig bedden en op Neurologie juist te veel bedden. Zowel het hebben van te weinig als te veel bedden is nadelig voor de kwaliteit van zorg en arbeid. Het totale aantal benodigde bedden is geschat met behulp van een Erlangmodel. De gezamenlijk beschikbare bedden bleken toereikend voor de betrokken specialismen. Er zijn drie interventies cijfermatig en medisch-inhoudelijk onderzocht, waarna één interventie gekozen is. In januari 2014 is er een samenwerkingsafdeling voor patiënten van Interne geneeskunde op afdeling Neurologie ingericht. Deze interventie heeft geresulteerd in een meer evenwichtige bedbezetting en een reductie van het aantal ongewenste overplaatsingen. Hierdoor zijn zowel de kwaliteit van zorg als de kwaliteit van arbeid op de afdelingen verbeterd. Het betrokken personeel is tevreden over de interventie, maar noemt als aandachtspunt de toegenomen werkdruk op de afdeling Interne geneeskunde.

Het is voor projectgroepen vaak lastig om écht iets in zorgprocessen te veranderen. De gehanteerde veranderaanpak blijkt een sleutel tot succes in dit project. De zorgprofessionals hebben zelf oplossingen gezocht en zijn daarbij inhoudelijk ondersteund. Deze inhoudelijke ondersteuning bestond vooral uit data-analyse door een wiskundige. Uitgangspunt was dat we door samenwerking de zorg voor patiënten willen verbeteren. Door samen op cijfermatige basis oplossingen te verkennen, is er nauwelijks sprake geweest van domeinstrijd.

De auteurs bedanken graag prof. dr. Richard J. Boucherie en alle medewerkers van de beschouwende afdelingen van het Jeroen Bosch Ziekenhuis.

Literatuur

Massey, W.A., & Whitt, W. (1994). An Analysis of the Modified Offered-Load Approximation for the Nonstationary Erlang Loss Model. *The Annals of Applied Probability, 4*(4), 1145-1160

Van de Vrugt, N.M. (2016). Efficient healthcare logistics with a human touch. PhD thesis, University of Twente. CTIT Ph.D.-thesis series No. 16-389.

Saskia M. Cornelissen is werkzaam als manager Kwaliteit & Veiligheid van het Deventer Ziekenhuis. Ten tijde van het project was zij werkzaam als adviseur Kwaliteit en Veiligheid in het Jeroen Bosch Ziekenhuis. Aandachtsgebieden zijn procesoptimalisatie, patiëntveiligheid, patiëntgerichtheid en continue verbeteren.
E-mail: S.Cornelissen@dz.nl

Maartje van de Vrugt is afgestudeerd en gepromoveerd in Technische Wiskunde aan Universiteit Twente. Haar promotieonderzoek naar efficiënte, zorgverlener- en patiënt-vriendelijke zorglogistiek heeft zij uitgevoerd bij het Jeroen Bosch Ziekenhuis. Tegenwoordig werkt zij bij het Leids Universitair Medisch Centrum, als adviseur integrale patiëntenlogistiek, en als postdoc bij Universeit Twente.
E-mail: n.m.vandevrugt@utwente.nl

Eric A.A. Smits Eric is sinds 2008 werkzaam als manager bedrijfsvoering Interne Geneeskunde – MDL, Geriatrie, Reumatologie, Neurologie en Longgeneeskunde in het Jeroen Bosch Ziekenhuis. Ook is hij verantwoordelijk voor dure geneesmiddelen, planning en zorglogistiek, capaciteitsmanagement en kwetsbare ouderen.
E-mail: E.Smits@jbz.nl

Thom P.J. Timmerhuis is neuroloog sinds 1999 en werkzaam in het Jeroen Bosch Ziekenhuis. Hij is medisch manager sinds 2011 van de RVE neurologie. Als aandachtsgebieden heeft hij epilepsie en slaapstoornissen.
E-mail: t.timmerhuis@jbz.nl

Fred Steutel

column

Professor Van Dantzig

David van Dantzig werd op 23 september 1900 in Amsterdam geboren. Hij stierf in Amsterdam op 22 juli 1959.

Ik ben een van de laatsten die deze veelzijdige man gekend heeft. Curiositeit: Van Dantzig en ik hadden drie dingen gemeen: in militaire dienst gewoon soldaat gebleven, begonnen met scheikundestudie en we schreven allebei wel eens een vers – ik vaak een limerick, hij wat serieuzer. Van Dantzig componeerde ook.

Er is veel over hem geschreven. In de *NRC* van 27 juli 1959 staat een lang in memoriam door Hans Freudenthal; het beslaat – in het A3-formaat van toen - ongeveer een zevende pagina. Het CWI bracht drie boekjes uit; twee geschreven door Gerard Albers en een door Constance van Eeden; zij geeft in de Scientific Family Tree de vier promovendi van Van Dantzig en hun wetenschappelijk nageslacht. De naam van Klaas van Harn met promotoren Doornbos en mijzelf ontbreekt door een misverstand. Hier haal ik persoonlijke herinneringen op.

In september 1956 meldde ik me bij het Mathematisch Centrum, waar assistenten werden gevraagd; ik was door gebrek aan resultaten mijn studiebeurs kwijtgeraakt en moest wat verdienen. Er was alleen nog plaats bij de afdeling Statistiek. Dat ik niets van statistiek wist was geen bezwaar. Er werd wel verwacht dat ik het statistiekcollege van professor van Dantzig zou volgen.

Statistiek was een nieuw vak op de Nederlandse universiteiten; het werd alleen in Wageningen gegeven, niet echt als wiskundevak, en dus nu in Amsterdam.

Bij de colleges van Van Dantzig zaten we met ons vieren: Fabius, Van Zwet uit Leiden, ikzelf en nog iemand, wiens naam mij is ontschoten. Van Dantzig kwam bijna altijd te laat, soms een half uur, soms drie kwartier. Je kreeg toch wel waar voor je geld, want hij bleef soms langer dan twee uur aan het woord. Dit veranderde toen zijn zoon het college ging volgen; die sloeg een krant open als hij vond dat het tijd was.

Ik maakte ijverig aantekeningen, wat niet eenvoudig was, omdat Van Dantzig zijn onderwijs niet echt voorbereidde, zodat het voor kon komen dat hij halverwege een bewijs vastliep en dan aankondigde dat het volgende week verder zou gaan. De gemaakte aantekeningen werkte ik de volgende dag zo goed mogelijk en keurig geschreven uit. Ik deed de vier of vijf volgeschreven vellen in een grote envelop en deed die bij de professor in de bus van zijn huisadres; ik meen in de Jacob Obrechtstraat in de buurt van het concertgebouw.

Bij het begin van het volgende college haalde Van Dantzig de nog gesloten envelop uit zijn tas, maakte hem open en bekeek korte tijd de inhoud. Daarna ging het college verder.

Ik heb nooit tentamen gedaan bij Van Dantzig; hij overleed een paar weken voor de datum die ik daarvoor afgesproken had. Het vak werd overgenomen door Constance van Eeden en Theo Runnenburg. Bij hen heb ik toen tentamen gedaan: een zeven (7).

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

Speciaal voor de Gay Pride 2014 ontworpen ING PINKautomaat. Foto: Bas Arps

DE ENERVERENDE WERELD VAN CASH OPERATIONS

Roel van Anholt

'Roel, hoe zou je het vinden om verder te gaan met academisch onderzoek door een promotieonderzoek te doen?' vroeg Iris Vis, hoogleraar Industrial Engineering aan de Rijksuniversiteit Groningen en mijn scriptiebegeleider, me vlak voordat ik mijn master in Bedrijfskunde met de specialisatie logistiek en supply chain management afrondde. De PhD-plek bood een mooie kans om me verder in het cash management te gaan verdiepen. De positie was volstrekt anders dan andere promotieonderzoeken omdat deze gefinancierd werd door een 'derde geldstroom', namelijk door ING Bank, Rabobank en de ABN Amro. In plaats van te beginnen van een literatuurstudie, ging ik me het eerste jaar bezighouden met de ontwikkeling van een bevoorradingsstrategie voor geldautomaten om de prestaties van de toenmalige bevoorradingsmethoden van de banken met elkaar te vergelijke, te benchmarken. Bij een aanzienlijke potentiële kostenbesparing zouden de drie banken samenwerking op dit gebied overwegen. Door de financiële crisis en de stagnerende groei van het aantal transacties met cash geld moesten de banken wegen zoeken om de kosten te verlagen. De praktische relevantie en de enorme uitdagingen waren redenen om er met veel enthousiasme aan te beginnen. Voor ik het wist zat ik er middenin: één jaar als consultant een interbancair project leiden en vervolgens drie jaar als onderzoeker analytische modellen ontwikkelen voor zowel theoretisch als praktisch problemen. In dit artikel beschrijf ik twee van de, wat mij betreft, meest intrigerende problemen in cash operations en de daarvoor ontwikkelde oplossingsmethodieken.

Geldlogistiek en geldverwerking

Cash operations is een verzamelbegrip voor twee hoofdactiviteiten: geldlogistiek en geldverwerking. Geldlogistiek betreft het met verschillende coupures bevoorraden van geldautomaten tegen de laagst mogelijke kosten door efficiënte bevoorradingsmomenten en -hoeveelheden te kiezen. Geldverwerking omvat het sorteren, tellen en verpakken van bankbiljetten en munten in zogenaamde cashcenters. Door de uniformiteit en hoge waardedichtheid van cash, kunnen zowel geldlogistiek als geldverwerking profiteren van geavanceerde OR-technieken. De uniformiteit (dat wil zeggen dat er slechts enkele denominaties per valuta beschikbaar zijn) maakt een hoge mate van standaardisatie en automatisering mogelijk en de hoge waardedichtheid van cash zorgt ervoor dat het investeren in efficiënte oplossingen zich op termijn terugbetaalt. Voor zowel geldlogistiek als geldverwerking kunnen talloze interessante onderzoeksproblemen worden bedacht, maar nu beperk ik me tot twee uitdagingen van geldlogistiek:

1. het voorspellen van de klantvraag bij geldautomaten;
2. het gecombineerde voorraad- en routeringsprobleem waarbij cash zowel wordt opgehaald als afgeleverd.

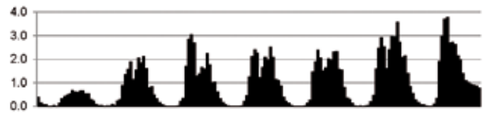
Voorspellen van de klantvraag bij geldautomaten

Om geldautomaten te bevoorraden tegen zo laag mogelijke kosten, moet nieuwe voorraad just in time worden geleverd om leegstand te voorkomen. Just in time leveringen zijn lastig te plannen vanwege de levertijd tussen het verpakken van een cash order in een cash center en het daadwerkelijk bijvullen van de geldautomaat. Het is belangrijk om de klantvraag gedurende de levertijd accuraat in te schatten. Zonder een accurate inschatting bestaat de kans dat geldautomaten te laat worden bijgevuld, wat leidt tot lege geldautomaten, of te vroeg bijgevuld, waardoor teveel kostbare voorraad wordt aangehouden. Het voorspellen van klantvraag bij een geldautomaat kan gezien worden als het voorspellen van een speciaal type multivariate tijdreeks. Een speciaal type doordat deze wordt gekenmerkt door seizoensinvloeden (d.w.z. verschillende patronen per weekdag-uur, maand, jaar, feestdag en evenement), plotselinge hoogte verschuivingen, onjuiste inputdata en ontbrekende inputdata. Figuur 1 geeft een beeld van het weekpatroon waarbij duidelijk zichtbaar is dat de klantvraag ieder uur van de week anders is. De klantvraag bij een geldautomaat hangt tevens af van de beschikbaarheid van nabij gelegen geldautomaten. Indien één geldautomaat niet beschikbaar is, zal een deel van de klantvraag worden verplaatst naar nabijgelegen automaten.

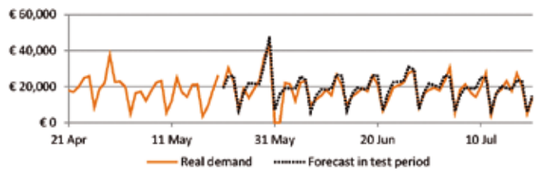
Er is een voorspellingsprocedure ontwikkeld die de toekomstige klantvraag voor meerdere periodes voorspelt (zie hoofdstuk 3 in Van Anholt, 2014). Deze procedure geeft een hogere voorspellingsnauwkeurigheid dan – voor zover wij weten – alle andere statistische en machine learning methodes. Dat onze oplossing beter presteert valt toe te schrijven aan een aantal unieke eigenschappen: een combinatie van de beste technieken voor ieder voorspellingselement, een nieuwe *smoothing* aanpak (d.w.z. het gladstrijken van de data) voor maandelijkse en jaarlijkse patronen en de efficiënte wijze waarop ontbrekende en onjuiste data wordt verwerkt. Figuur 2 laat zien dat de werkelijke klantvraag wordt benaderd door onze voorspellingen.

Gecombineerde voorraad- en routeringsprobleem met haal- en leveropdrachten

De basisgedachte achter deze tweede uitdaging is eenvoudig: een aantal geldautomaten die dicht bij elkaar staan zouden op dezelfde dag van de week bijgevuld moeten worden. Het is overduidelijk dat er veel kilometers bespaard kunnen worden als leveringen aan meerdere nabij gelegen geldautomaten gecombineerd worden in één route, vooral als de automaten in afgelegen gebieden staan. Maar hoe regel je de voorraad en routes voor deze geldautomaten op een kostenefficiënte manier, waarbij rekening wordt gehouden met de kosten voor transport, voorraad en bijvullen/leeghalen? Het antwoord moet gezocht worden in het simultaan oplossen van het voorraad- én het routeringsprobleem. Dit is een uiterst complex probleem



Figuur 1. Voorspellingen van het weekpatroon (van zondag tot en met zaterdag) voor de klantvraag in euro's per uur (genormaliseerd) bij één willekeurige Nederlandse geldautomaat



Figuur 2. Gerealiseerde en voorspelde dagelijkse vraag in euro's bij één willekeurige geldautomaat

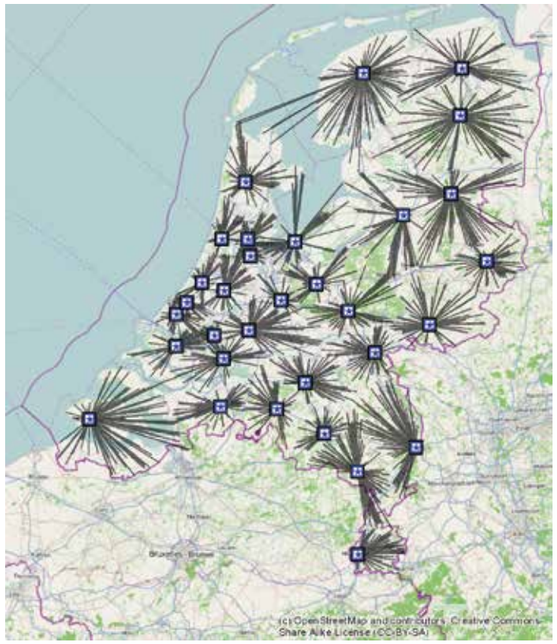
wat enkel voor kleine instanties optimaal kan worden opgelost.

Tegen het eind van mijn PhD kreeg ik de kans om bij het onderzoeksinstituut CIRRELT in Montreal samen te werken met prof. G. Laporte, prof. L.C. Coelho en prof. I.F.A. Vis. Samen hebben we een oplossing ontwikkeld voor dit complexe probleem dat zich onder andere voordoet bij recirculatie geldautomaten (Van Anholt et al., 2015). Via recirculatie-geldautomaten kan zowel geld worden opgenomen als worden gestort. Omdat vraag en aanbod van cash bij één automaat zelden in balans zijn, moet cash worden opgehaald of bijgevuld om lege en volle geldautomaten te voorkomen. Ondanks dat de huidige praktijk nog niet zo werkt wegens opgelegde restricties, hebben we aangenomen dat de waardevervoerder tijdens één route cash herverdeelt tussen recirculatie-geldautomaten om cash overschotten en cash tekorten te effenen. Momenteel wordt al het cash van en naar de cash centers vervoerd, zonder deze tussen automaten te herverdelen. De kwaliteit van onze oplossing hebben we getoetst met zowel numerieke data als met werkelijke klantvraag/-aanbod bij geldautomaten. Ons onderzoek heeft uitgewezen dat substantiële kostenbesparingen mogelijk zijn, dus wellicht zullen de betrokken partijen hun huidige methodes heroverwegen. We hebben dit probleem geformuleerd als een gemixt-geheeltallig lineair programmering model en opgelost met een exact branch-and-cut algoritme. Figuur 3 toont de clustering van geldautomaten die is uitgevoerd voordat het branch-and-cut algoritme is toegepast.

Cash operations in opkomst als onderzoeksgebied

Het aantal wetenschappelijke papers in toonaangevende *journals* dat gerelateerd zijn aan cash operations neemt snel toe. Over de hele wereld ontdekken onderzoekers dit uiterst interessante gebied. Mij zou het niet verbazen als steeds meer operations research conferenties een gehele stroom aan dit onderwerp zouden wijden. Wellicht komt er een dag waarop cash operations net zoveel aandacht krijgt als, bijvoorbeeld, containerterminal operations.

In het begin van 2015 zijn drie promovendi (M. Hoogeboom, A. Baller en C. Orlis), aan de Vrije Universiteit Amsterdam begonnen met vervolgonderzoek in cash operations. Zij bestuderen de veiligheid van de transportroutes, de mogelijkheden tot het combineren van bevoorradingen in routes en de efficiëntie en effectiviteit van contractuele verhoudingen tussen belangrijkste spelers in de geldketen.



Figuur 3. Clustering van geldautomaten

Literatuur

Van Anholt, R.G., Coelho, L.C., Laporte, G., & Vis, I.F.A. (2015). An Inventory-Routing Problem with Pickups and Deliveries Arising in the Replenishment of Automated Teller Machines. *Transportation Science*, <http://dx.doi.org/10.1287/trsc.2015.0637>

Van Anholt, R.G. (2014). *Optimizing Logistics Processes in Cash Supply Chains.* PhD proefschrift, Vrije Universiteit Amsterdam, Faculteit der Economische Wetenschappen en Bedrijfskunde.

Roel van Anholt behaalde zijn doctorsgraad bij de Faculteit der Economische Wetenschappen en Bedrijfskunde van de Vrije Universiteit in december 2014. Roel werkt sinds september 2012 bij Geldservice Nederland, eerst parttime en na afronden van het promotieonderzoek fulltime als strategie & business development consultant. De benchmarkstudie die Roel uitvoerde in opdracht van ING Bank, Rabobank en ABN AMRO Bank in 2009 liet een substantiële potentiële kostenbesparing zien. Dit resultaat heeft bijgedragen aan de oprichting van de joint venture Geldservice Nederland in september 2011.
E-mail: ranholt@feweb.vu.nl

Gerrit Stemerdink

column

VERSCHILLEN IN CULTUUR

Ik heb niet de illusie dat iemand het heeft opgemerkt, maar in het voorstel om mij erelid van de vereniging te maken zat een foutje. In het artikel in *STAtOR* 2015-2 hierover kon ik dat als eindredacteur corrigeren. Kennelijk had een bestuurslid Google en LinkedIn geraadpleegd om wat details rond mij en mijn bezigheden aan het voorstel toe te voegen. Het gevolg was dat in het voorstel stond dat ik geboren was op 8 maart 1941, terwijl dat 3 augustus 1941 is. Zelfs op die leeftijd voelt het vreemd als je ineens bijna een half jaar ouder wordt verklaard.

Het is duidelijk wat er is misgegaan: ergens op het internet stond mijn geboortedatum als 3-8-41 en dat is nu eenmaal voor tweeërlei uitleg vatbaar. In ongeveer de helft van de wereld komt eerst de maand en dan de dag en in het andere deel is het andersom. Zeker als je met internationale groepen werkt moet je daar altijd alert op zijn.

Zo heb ik in 2003 een historische lijst samengesteld van iedereen die ooit Elected Member was van het International Statistical Institute (ISI). Daarvoor heb ik maandenlang de archieven geraadpleegd waarin alle aanmeldingen waren bewaard. In het begin prachtige handschreven brieven, later werd gebruikt gemaakt van voorgedrukte formulieren. Gelukkig was er bij de oudste aanmeldingen vrijwel nooit een probleem, iedereen schreef de naam van de maand voluit in het Frans of het Engels – de beide officiële talen van het ISI. Maar toen de formulieren in zwang kwamen was het soms lastig. Zeker toen op een gegeven ogenblik die formulieren voorgedrukte hokjes kregen, zonder daarbij aan te geven wat er in die hokjes moest komen. Ik heb het probleem zo goed als mogelijk was in een beperkte tijd aangepakt: data zijn maar voor een uitleg vatbaar, als er een getal groter dan 12 in voorkomt kan dat niet anders dan de dag zijn. Dus bij een gelijkmatige verdeling van geboortedagen is er slechts bij rond de 40% (144/365) van de aanmeldingen kans op een verwisseling. Bij die data ben ik ervan uitgegaan dat alle aanmeldingen uit Nederland de volgorde d-m-j hadden en alle Amerikaanse de volgorde m-d-j. Bij de rest heb ik andere bronnen geraadpleegd. Zo heb ik allereerst gekeken of er bij een probleem meer aanmeldingen uit het zelfde land en dezelfde tijd waren waarbij wél duidelijk was welke volgorde werd gebruikt. En van veel van de oudere, soms zeer beroemde, leden waren er biografieën en herdenkingsartikelen te vinden die uitsluitsel gaven. Maar toch heb ik in het rapport vermeld dat ik niet altijd kon instaan voor de correcte vermelding van geboorte- en sterfdata.

Niet alleen geboortedata, ook allerlei afmetingen etc. worden in verschillende culturen in verschillende systemen vastgelegd. Het verschil tussen pounds en ponden, graden Fahrenheit en Celsius en inches en centimeters kennen we allemaal. Maar wist u dat er fatale ongelukken zijn gebeurd omdat één onderdeel van een ingewikkelde machinerie, zoals een vliegtuig, een ander systeem hanteerde dan de rest? Naar verluidt was dit ook de oorzaak van de aanvankelijk mislukte wazige beelden van de Hubble ruimtetelescoop.

Ook in de statistiek moeten we alert blijven: als we bijvoorbeeld lengte van personen in klassen als 165–169 cm etc. indelen, zullen we een afwijking kunnen krijgen als we gegevens die gegroepeerd waren als 66–68 inch ‘vertalen’ naar onze metrische groepering. Het is daarom verstandig gegevens van elders nooit in geaggregeerde vorm over te nemen, maar altijd te proberen die gegevens zo basaal mogelijk te verkrijgen.

Trek nu niet de conclusie dat ik van die verschillen af wil. Ik vind het véél te leuk om in Engeland het verbruik van mijn auto van kilometers per liter naar miles per gallon om te rekenen of om een pint bier te bestellen en dan te denken dat ik matig drink omdat een pint tenslotte maar iets meer dan halve liter is.

Maar laten we wél alert blijven, er kan wel eens iets mis gaan!

Gerrit Stemerdink is eindredacteur van *STAtOR*.
E-mail: gjstemerdink@hotmail.com

CQM EN SMART SOCIETY – AN ETERNAL GOLDEN BRAID*

De term Smart Society wordt vaak gekoppeld aan een combinatie van Smart Systems, Smart People, en Smart Organizations om de Systems en People te verbinden. Dit is onlosmakelijk verbonden met Operations Research en Statistiek. Vaak omdat praktische toepassingen daarvan een positief effect hebben op duurzaamheid, leefbaarheid, en de inbedding van slimme technologie in ons dagelijks leven om dat comfortabeler te maken. Maar zeker ook omdat een slimme maatschappij effectief en efficiënt omgaat met haar publieke middelen.

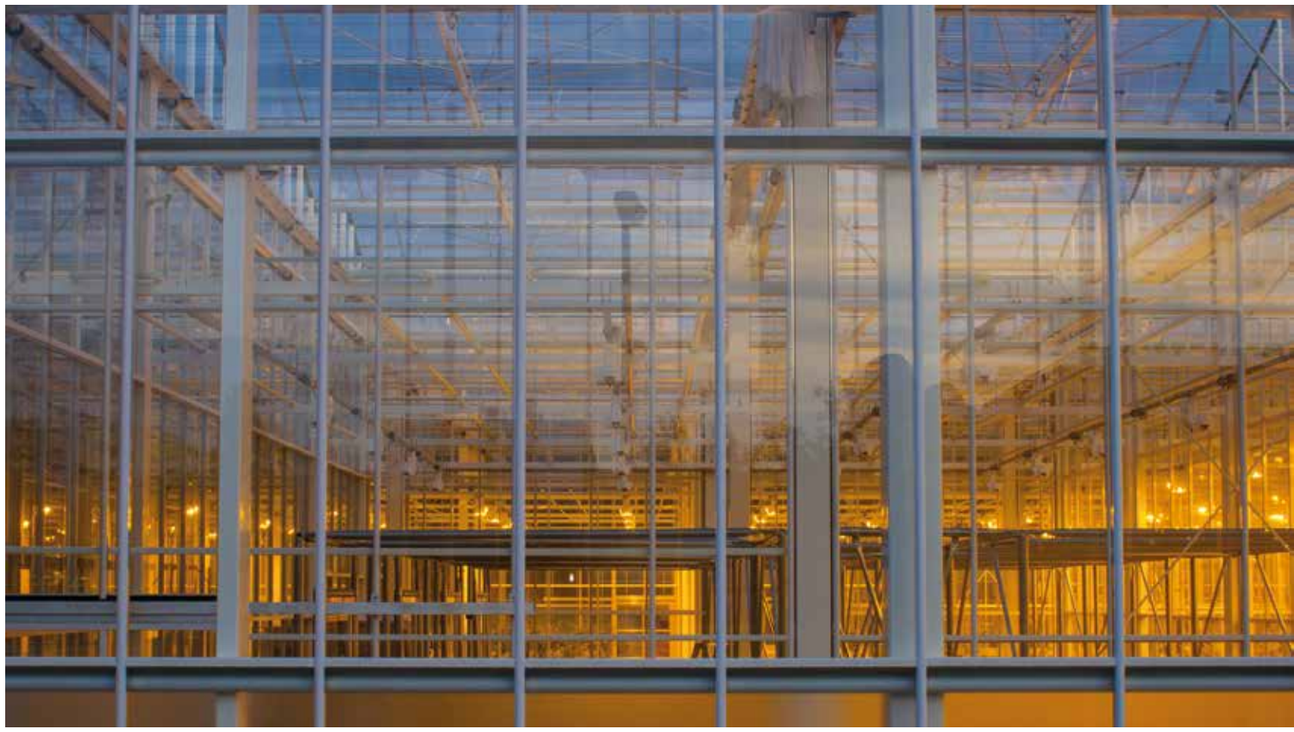


Foto: Pieter Bosch

Marnix Zoutenbier

In de projecten die we bij CQM uitvoeren spelen technieken uit de Operations Research, Statistiek, en Information Science bijna altijd een belangrijke rol. Voor veel projecten geldt ook dat ze een Smart Society component hebben. Daarom een bloemlezing uit recente projecten om dit te illustreren.

Productontwikkeling: comfort door te ontwikkelen wat een consument écht wil

Productontwikkeling wordt traditioneel door marketing gedreven vanuit de perceptie van klantwensen. Grootschalig gericht consumentenonderzoek ligt daaraan vaak ten grondslag en dan nog blijkt het erg lastig om te ontdekken wat consumenten echt belangrijk vinden aan producten.

Tegenwoordig zijn we in staat om alle online-uitingen van consumenten, bijvoorbeeld in *reviews* en *ratings* bij internetwinkels, voor zowel onze klant als hun concurrenten te analyseren: interpretatie van *free* tekst en statistische modellen die de link leggen tussen de betekenis van de tekst en de ratings, geven ontwikkelaars inzicht in welke verbeteringen een groot effect gaan hebben op de klantwaardering en welke niet. Interessante bijvangst is inzicht in de positionering volgens klanten ten opzichte van de concurrentie.

Maar ook *loggings* van consumentenapparaten worden gebruikt om producten beter te maken of beter te laten functioneren. Voordeel is dat een klant niet hoeft te vertellen over de manier waarop hij een apparaat dénkt te (gaan) gebruiken, maar het werkelijke gebruik in de praktijk wordt gebruikt voor verbetering. Denk hierbij aan de wake up light, baby monitoren, chips, en machines voor post- en pakketverwerking. Maar ook een altijd pijnlijk proces van lichaamsontharing wordt verbeterd door automatische detectie van het huidtype van de consument.

Voor alle voorbeelden geldt dat goedkope en kwalitatief betere informatie gebruikt wordt om ons leven comfortabeler te maken.

Planning en scheduling: minder CO_2, uitgerust personeel bij Spoedeisende Hulp, en iedereen op tijd zijn medicijnen in het ziekenhuis

Wij werken al jaren voor grote transporteurs die multimodaal willen plannen. Niet alleen de vrachtwagen, maar ook boot en trein om zo snel en goedkoop mogelijk goederen op de juiste plaats te krijgen. De winst in CO_2-uitstoot is enorm door betere routes. Maar de kans op ongelukken is ook een stuk kleiner door een goede balans in rij- en rusttijden binnen de wettelijke kaders.

Het vinden van optimale personeelsroosters voor de afdeling spoedeisende hulp (SEH) in een ziekenhuis is een ander traditioneel Operations Research-vraagstuk. Door een fusie tussen twee SEH-afdelingen en daarmee samenhangend een kleinere bezetting was het zeer twijfelachting of het mogelijk zou zijn om binnen de Arbo eisen tot een goede bezetting te komen. Tot vreugde van medewerkers van de afdeling en management bleek het mogelijk om tot een kwalitatief goede bezetting te komen zonder overbelasting van de medewerkers.

Bij het ontwerp van een nieuw ziekenhuis in Canada wordt gebruik gemaakt van Automated Guided Vehicles (AGV's). Deze AGV's zijn cruciaal voor de logistiek in een ziekenhuis: mensen moeten op tijd hun medicijnen en voeding hebben en die worden met de AGV's door het ziekenhuis getransporteerd. Minder belangrijk, maar ook nodig, het vervoer van linnengoed en afval. Liften voor de AGV's vormen potentieel een bron van verkeerscongestie. Wij hebben de algoritmiek, en onderbouwing daarvan, van de aansturing van deze liften ontwikkeld om een optimale doorstroming van AGV's te garanderen.

Planning, scheduling, en simulatie dragen hiermee bij aan een beter milieu, betere arbeidsomstandigheden, en een effectief en aangenaam verblijf in het ziekenhuis.

Effectief en efficiënt gebruik van publieke middelen

Samen met de zwembond werken we aan een zogenaamde TalentCalculator. Uiteindelijk moet dat een app worden waarmee niet alleen leeftijd en gezwommen tijd bepalend zijn voor de talentkwalificatie, maar ook andere meetbare factoren die samen de potentie van een jeugdige zwemmer beter beschrijven.

We ontwikkelen een statistisch model om zwemmers van verschillende leeftijden, lengtes, biologische leeftijden[1], etc. te kunnen vergelijken op hun potentieel als volwassen zwemmer. En het mes snijdt hier ook aan twee kanten: de kans op zowel fouten van de eerste als de tweede soort worden hiermee kleiner. Dus niet alleen meer medailles, maar ook een betere besteding van maatschappelijke middelen omdat met grotere kans geïnvesteerd wordt in de juiste talenten vanuit de altijd krappe sportbudgetten.

Publieke middelen worden ook aangewend om het

spoor in goede conditie te houden. We zijn tegenwoordig in staat om kleine defecten op het spoor automatisch te detecteren alvorens ze tot problemen leiden. Grote hoeveelheden handmatig werk kunnen daarmee geautomatiseerd worden. De algoritmiek hiervoor, *Deep Learning*[2], is niet nieuw, maar door de ontwikkelingen op hardware-gebied is het nu mogelijk om de algoritmiek real time werkend te krijgen in praktische toepassingen.

Een andere spoortoepassing betreft het ontsluiten van ongeveer één miljoen tekeningen van het spoor. Google heeft ons geleerd wat er mogelijk is op het gebied van het zoeken in grote hoeveelheden gegevens. Wij ontwikkelen nu samen met RIGD-LOXIA een specifieke zoekmachine voor ProRail om in grote hoeveelheden tekeningen snel de juiste te vinden. De betreffende technische tekeningen zijn momenteel lastig vindbaar via traditionele zoekmethoden gebaseerd op metadata. De metadata bevat soms fouten en beperkt de zoekmogelijkheden. Door deze tekeningen, op een efficiënte en geautomatiseerde wijze, sneller vindbaar te maken, kunnen verstoringen aan het spoor door monteurs sneller worden opgelost en ondervindt het treinverkeer minder last van hinder. Een volgende stap zal zijn om ook tekeningen met elkaar te associëren zonder specifieke zoektermen.

Duurzaamheid door energiebesparing

In de glastuinbouw is heel veel warmte nodig. Bij de productie daarvan wordt door veel kwekers ook elektriciteit opgewekt die ze kunnen verkopen op de APX-elektriciteitsmarkt. Voor Agro Energy, dé energiespecialist in de voor de Nederlandse agrarische sector, reden om ons te vragen kwekers te ondersteunen bij het nemen van optimale beslissingen rondom koop en verkoop van energie. Inhoudelijk een zeer uitdagende dynamiek van vraag, aanbod, prijsvorming en real time beslissingen die ondersteund moeten worden. Dat vraagt om Information Science om dit in een geautomatiseerde omgeving 24/7 te laten draaien, statistische modellen om vraag en aanbod te voorspellen, en snelle optimalisatie om tot goede beslissingen te komen. Maar vooral van groot maatschappelijk belang om overproductie van warmte tegen te gaan.

Zonder goede gegevens geen Smart Society

Een Smart Society maakt veel gebruik van data, vaak in grote hoeveelheden. Voor veel toepassingen is het niet zo erg als die gegevens niet exact kloppen. Als het om levensmiddelen gaat ligt dat echter anders.

We vinden het vanzelfsprekend dat het etiket van een product in het schap van de retailer ons exact vertelt wat er in het product verwerkt is. Het toenemend gebruik van internet om boodschappen te doen vraagt veel van de kwaliteit van de online etiketinformatie. GS1 is een not-for-profit organisatie die zorgt voor een betrouwbare uitwisseling van productinformatie tussen de meer dan 1000 toeleveranciers en grote retailers van Nederland die zijn aangesloten. Voor hen is het een forse uitdaging om die kwaliteit te bereiken en te behouden. Wij hebben daar op verschillende manieren aan bijgedragen. Iedereen met een allergie zal de maatschappelijke baten direct herkennen!

Het hart van een Smart Society: Smart People

En een Smart Society kan natuurlijk alleen maar bestaan bij de gratie van Smart People. Ze vormen het hart van CQM. Door samen te werken met kennisinstellingen maar ook door bijvoorbeeld hier in de regio Eindhoven de 'biggest brains' zelf te trainen in goed gebruik van Data Science methoden dragen we ook bij aan de ontwikkeling van deze Smart People.

Maar wellicht doen we dat nog veel meer door een omgeving te creëren waarin slimme mensen graag bij ons werken en, staande op de schouders van onze oud-collega's, telkens weer, nog mooiere toepassingen weten te vinden voor ons vak. Door het enthousiasme voor ons vak af te laten stralen op ons werk en altijd in verbinding met klanten die meestal geen wiskundige zijn. Daarmee laten we breed in de maatschappij zien wat de waarde en het plezier kan zijn van slimme mensen in combinatie met mooie wiskunde en uitdagende vraagstellingen.

Dank voor de hulp van vele collega's.

* Vrij naar Douglas Hofstadters *Gödel, Escher, Bach, an eternal golden braid*

1. Met biologische leeftijd wordt bedoeld in welke mate een jeugdige zwemmer al volgroeid is. Als twee zwemmers van 14 jaar dezelfde tijd zwemmen maar de éne moet nog in de puberteit komen terwijl de ander biologisch al 17 is, dan heeft de eerste zwemmer meer potentieel.
2. Zie bijvoorbeeld http://www.deeplearningbook.org/ of https://en.wikipedia.org/wiki/Deep_learning.

Marnix Zoutenbier werkt als Principal Consultant bij CQM en heeft 250 projecten uitgevoerd op het gebied van Data Science in de volle breedte van het leven.
E-mail: zoutenbier@cqm.nl

Bayesiaanse methoden voor longitudinale modellen een nieuwe kijk op ontwikkeling van jonge kinderen

In 2015 werden er ruim 170 duizend kinderen geboren in Nederland en in de periode 2011-2013 bezocht ongeveer 99 procent van alle eenjarigen ten minste eenmaal het consultatiebureau in de voorgaande twaalf maanden (CBS). Onder vierjarigen is dit nog steeds 85%. Naast grootheden als lengte en gewicht zijn 'kan [wel/niet]'-testen van motorische en cognitieve vaardigheid een bron voor diagnose en eventuele doorverwijzing. Statistische technieken kunnen daarbij een belangrijke rol spelen, zowel in de analyse als in de communicatie richting medisch specialisten en ouders.

Kevin Duisters

Om de motorische en cognitieve vaardigheden van jonge kinderen in kaart te brengen neemt ieder consultatiebureau het Van Wiechen schema af. Dit is een systematisch onderzoek waarin 75 'kan [wel/niet]'-testen (tweemaal) verspreid worden over veertien bezoeken in de eerste vier levensjaren. Het doel is een individuele ontwikkelingscurve te schatten uit deze data om zo achterblijvende pupillen te kunnen voorzien van verdere diagnose en behandeling. TNO Life Sciences houdt zich bezig met dit vraagstuk (Van Buuren, 2014) in dienst van diverse consultatiebureaus. Vanuit deze partners zijn enkele verbeterpunten met betrekking tot interpretatie, robuustheid, gladheid en correctie voor achtergrondvariabelen (zoals geslacht en zwangerschapsduur) van de huidige ontwikkelingscurves naar voren gekomen. Overigens, dit type gegevens komt voor in een breed scala aan toepassingsgebieden zoals de industriële statistiek, gedragswetenschappen en biomedische studies.

Toegevoegde waarde van statistiek

De analyse van deze data kent tenminste vier vraagstukken waarbij statistiek van belang kan zijn:
A. Wat betekent een 'kan wel' op een specifieke test voor

de ontwikkeling van een kind en hoe leiden verschillende 'kan [wel/niet]'-uitkomsten tot een onderliggende, continue 'ontwikkelingsscore' op moment *t*?

B. Wat is het effect van achtergrondvariabelen op deze score?

C. Hoe om te gaan met een laag aantal waarnemingen per bezoek en verschil in spreiding van bezoeken per individu?

D. Wanneer wijkt een ontwikkelingscurve van een individueel kind af?

Vraagstuk A is een goed onderzocht probleem in de psychometrie (Rasch en Item Response Theorie (IRT) modellen, zie o.a. Fox & Glas, 2001), veelal toegepast op testdata verzameld op één enkel moment. In bredere, statistische context kan men deze methoden vergelijken met Generalized Linear Models (GLM) en uitbreidingen richting data verzameld over tijd zijn voorhanden (GL Mixed M). In dit raamwerk is vraagstuk B een eenvoudige toevoeging.

Drie formules

We beschrijven een deel van het schattingsmodel voor kind *i* getest met item *j* op tijdstip *t*.

[Wel/niet](*ijt*) volgt een Bernoulli (**Kans(*ijt*)**) verdeling. (1)

linkfunctie(**Kans(*ijt*)**) = **Ontwikkeling(*it*)** – Moeilijkheidsgraad(*j*) (2)

Ontwikkeling(*it*) = functie(Achtergrond(*i*)) + **Referentie(*it*)** + fout(*it*) (3)

Niveaus (1) en (2) zijn standaard in de IRT literatuur, waarbij de 'logit' de rol van linkfunctie vervult voor de kenner. Deze linkfunctie zorg ervoor dat de discrete 'kan [wel/niet]'-data via kansen getransformeerd wordt tot een continue ontwikkelingsscore. Deze score, specifiek voor individu *i* op tijdstip *t*, wordt gecorrigeerd voor de moeilijkheidsgraad van het bevraagde item *j*.

Onze bijdrage komt tot uiting in niveau (3). Door Ontwikkeling(*it*) hiërarchisch verder te modelleren is het op de eerste plaats eenvoudig een lineaire functie van achtergrondvariabelen toe te voegen (vraagstuk B). Uitbreidingen middels interacties van achtergrondvariabelen en tijd zijn bovendien mogelijk. Ten tweede is het mogelijk de verdeling van fout(*it*) zodanig te specificeren dat verschil in spreiding van bezoeken per individu toelaatbaar is (bijvoorbeeld middels een Gaussisch proces). Ten derde wordt hier een verband met een referentiewaarde geïntroduceerd. Deze referentiewaarde voor Ontwikkeling van kind *i* op moment *t* wordt afgelezen uit een referentiecurve, beschreven als een polynoom van tijd (bijvoorbeeld kwadratisch of kubisch). Dit vindt plaats op lagere, niet weergegeven niveaus in de hiërarchie.

Bayesiaanse schattingsmethode

De Bayesiaanse statistiek kenmerkt zich door parameters te voorzien van een kansverdeling. Er bestaat niet één universele waarheid, zoals de Maximum Likelihood (ML) estimate nastreeft, maar een verdeling waarin verschillende waar(he)den een kans krijgen. De onderzoeker begint met een verdeling a priori en weegt deze middels nieuwe data tot een verdeling a posteriori.

Er zijn twee redenen om in dit longitudinale model voor een Bayesiaanse schattingsmethode te kiezen. Ten eerste bieden moderne algoritmen (zoals MCMC) gelegenheid tot rijkere modellen door het omzeilen van integreren in hoge dimensie; dit maakt het schatten van de referentiecurve haalbaar. Ten tweede is het mogelijk deze referentiecurve zowel robuust als subject-specifiek te maken, met afhankelijkheid van *i* en van *t*.

Het bepalen van de referentiecurve

De Bayesiaanse mechaniek wordt gebruikt om Ontwikkeling(*it*) in meer of mindere mate richting de Referentie(*it*) te bewegen, met name meer in het geval van weinig data (vraagstuk C). Hierdoor is de ontwikkelingscurve minder onderhevig aan de grilligheid die het verschil tussen drie en vier 'kan wel'-uitkomsten met zich meebrengt bij een bezoek dat door (willekeurige) omstandigheden slechts vijf testen bevat. Dit is van belang in de communicatie richting ouders en medisch specialisten. Bovendien is deze methode niet gevoelig voor verschil in spreiding van bezoek (ieder punt *t* is af te lezen uit de referentiecurve).

Stap I

Om dit mogelijk te maken heeft deze techniek a priori een referentiecurve nodig, zonder enige 'kan [wel/niet]'-data van kind *i*. Om een zodanige curve en achtergrond effecten te schatten, wordt een steekproef van ruim tweeduizend Nederlandse kinderen gebruikt (zie resultaten). Empirische Bayesiaanse methoden die parameters diep in de hiërarchie door middel van ML bepalen (in plaats van priors op lastig interpreteerbare parameters in de populatie), beperken de subjectiviteit in dit stadium (Casella, 2001). Deze 'diepe' parameters, denk aan variantie matrices en parameters in de foutverdeling, vormen samen met de (mean posterior) curve en achtergrondeffecten de prior voor het subject-specifieke proces. Stap I is computerintensief (enkele uren op een standaard machine), maar hoeft slechts eenmaal berekend te worden.

Stap II

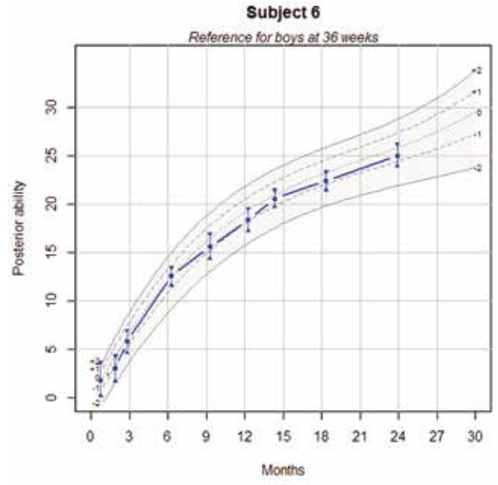
Door het sequentiële karakter van de data kan na ieder bezoek de subject-specifieke referentiecurve worden bijgesteld binnen enkele seconden. Deze procedure gebruikt dezelfde computationele techniek, maar nu volledig Bayesiaans (zonder ML tussenstap) met de prior uit stap I en alle informatie die reeds verzameld is over het betreffende kind. De referentiecurve wordt almaar meer subject-specifiek naarmate er meer en vaker data van het kind worden verzameld.

Door alle computationele kracht op één individu te concentreren, wordt het mogelijk een posterior rond relevante parameters te visualiseren. De voornaamste toepassing vindt plaats in betrouwbaarheidsschatting: op ieder bezoek wordt er een verdeling gemaakt rondom Ontwikkelingscurve(*it*). Kwantielen in deze verdeling kunnen worden weergegeven als 'geloofwaardigheidsinterval', de Bayesiaanse analoog van het betrouwbaarheidsinterval.

Resultaten: ieder kind is uniek

Om het resultaat van deze nieuwe methode te illustreren wordt een aan TNO beschikbaar gestelde dataset gebruikt. De SMOCK-data (Sociaal-Medisch Onderzoek bij Consultatiebureau Kinderen, Herngreen et al., 1992) bevat uitkomsten van het Van Wiechen schema en achtergrondinformatie voor ruim tweeduizend kinderen geboren in 1988-1989 en wordt hier ingezet om de referentiecurve te bepalen. Dit onderzoek beperkt zich tot de eerste twee levensjaren met 57 'kan [wel/niet]'-testen (tweemaal) verspreid over negen bezoeken, samen goed voor ruim 163 duizend waarnemingen.

Merk op dat naast het subject-specifieke geloofwaardigheidsinterval van kind *i* op moment *t* uit stap II, de informatie uit stap I (SMOCK) ook gebruikt kan worden voor visualisatie. De vraag (D) of een ontwikkelingscurve afwijkt kan beantwoord worden door het kind op een (voor achtergrondvariabelen gecorrigeerd) referentiediagram weer te geven. De informatie uit stap I wordt samengevat door empirische kwantielen te schetsen voor de ontwikkeling van een gemiddeld kind (met dezelfde achtergrond), continu over tijd. Zo wordt communicatie richting ouders en medisch specialisten aangescherpt door schattingen én referentiediagrammen gepersonaliseerd voor ieder kind.



Ontwikkelingscurve door negen bezoeken inclusief 95% geloofwaardigheidsinterval, weergegeven op een referentiediagram voor jongens geboren na 36 weken, inclusief 1 SD (68%) en 2 SD (95%) banden

Een punt van aandacht richting de toekomst is het statistisch kwantificeren van afwijking tussen een individuele curve en het referentiediagram. Is er reden voor doorverwijzing na één 'slechte' dag? Met behulp van medische diagnoses op latere leeftijd kunnen pragmatische beslisregels opgesteld worden voor vroegtijdige signalering. Denk bijvoorbeeld aan oppervlakte tussen de ontwikkelingscurve en een kwantiel in de referentie, of een aantal tegenvallende bezoeken. Zonder deze aanvullende analyse moet men waakzaamheid betrachten in het vergelijken van schattingen: er ligt multiple testing op de loer.

Literatuur

Casella, G. (2001). Empirical Bayes Gibbs sampling. *Biostatistics, 2*(4), 485-500.

Fox, J-P., & Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*(2), 271-288.

Herngreen, W.P., Reerink, J.D., Noord-Zaadstra, B.M. van, Verloove-Vanhorick, S.P., & Ruys, J.H. (1992). SMOCC: design of a representative cohort-study of live-born infants in The Netherlands. *European Journal of Public Health, 2*(2), 117-122.

Van Buuren, S. (2014). Growth charts of human development. *Statistical Methods in Medical Research, 23*(4), 346-368.

Kevin Duisters is promovendus aan het Mathematisch Instituut, Universiteit Leiden, begeleid door Jacqueline Meulman en Aad van der Vaart. Zijn onderzoek focust zich op hoog-dimensionele data, betrouwbaarheidsschattingen en statistische leertechnieken. Hij ontving de Hemelrijk Award 2015 voor zijn master-scriptie in Statistical Science (onder begeleiding van Elise Dusseldorp (UL), Stef van Buuren (TNO, UU) en Aad van der Vaart (UL)).
E-mail: k.l.w.duisters@math.leidenuniv.nl

Ontbrekende waarden bij het schatten van het aantal inwoners van Nederland

Susanna Gerritse, Bart Bakker & Peter van der Heijden

Statistische bureaus zoals het Centraal Bureau voor de Statistiek (CBS) houden periodiek een volkstelling. De meeste landen houden nog een traditionele volkstelling waarbij alle inwoners van een land een vragenlijst invullen (Schulte Nordholt, 2016). In Nederland maakt het Centraal Bureau voor de Statistiek gebruik van een populatieregister gebaseerd op administratieve data van gemeenten om het aantal inwoners te schatten. Vóór 1 januari 2014 was de Gemeentelijke basisadministratie (GBA) het Nederlandse Populatieregister, daarna is deze vervangen door de Basisregistratie Personen (BRP). Aan het Populatieregister worden dan andere registraties en een enquête gekoppeld om de volledige informatie van de volkstelling te kunnen leveren.

Vangst-hervangst-methodologie

Het aantal inwoners van een land wordt in het Engels aangeduid met *usual residents*, ofwel de 'gewoonlijk verblijvende bevolking', welke wij in het vervolg tevens aanduiden met 'inwoners'. Volgens EU Regulation No 1260/2013 worden de inwoners gedefiniëerd als mensen die minstens 12 maanden in Nederland verblijven, of deze intentie hebben. Het Populatieregister omvat een aanzienlijk deel van het aantal inwoners. Echter het Populatieregister alleen is niet voldoende om het aantal inwoners te schatten, aangezien er ook mensen zijn die wel in Nederland verblijven maar die zich niet hebben ingeschreven. Twee zulke grote groepen zijn voormalige asielzoekers die uitgeprocedeerd zijn, en Europese arbeidsmigranten die wegens vrij verkeer van personen legaal in Nederland verblijven, ook als ze zich niet inschrijven in het Populatieregister. Als het Populatieregister gebruikt wordt om het aantal inwoners in Nederland te schatten, is er sprake van onderdekking.

Vangst-hervangst-methodologie kan worden gebruikt om de onderdekking te schatten van het Populatieregister (Bishop, Fienberg en Holland, 1975; International Working Group for Disease Monitoring and Forecasting, 1995). In het simpelste voorbeeld worden twee registers gebruikt, waarbij mensen in ieder van de registers wel of niet zijn opgenomen. Hierdoor kunnen aantallen mensen worden geclassificeerd in een 2 bij 2-tabel. Het aantal mensen dat in beide registers niet voorkomt maar wel tot de bevolking behoort, is per definitie onbekend en dient geschat te worden. Dit aantal kan worden geschat met een loglineair model voor de 2×2-tabel als wordt aangenomen dat de insluitkansen van de registers niet samenhangen.

In dit onderzoek worden twee registers gekoppeld aan het Populatieregister: het HerKenningsdienst Systeem (HKS) van de politie en het werknemersdeel uit de Polisadministratie, dat het WerkNemersBestand (WNB) is genaamd. Het HKS is een register dat bijgehouden word door de politie, waarbij alle verdachten van bij de politie bekende procesverbalen geregistreerd staan. De Polisadministratie is een groot register op het CBS waarin allerlei werknemers- en werkgeversinformatie geregistreerd staat. Voor het WNB hebben wij enkel het deel van de werknemers uit de Polisadministratie gehaald. Er zijn twee niveaus per register: mensen kunnen wel of niet in een register zitten en dus krijgen we een 2×2×2-tabel, met 1 structurele nul. Door middel van de AIC kan er gezocht worden naar het best passende loglineaire model. Vanuit dit model wordt de structurele nul geschat.

Verblijfsduur bepalen

Om de vraag te beantwoorden hoeveel inwoners Nederland heeft, dient eerst deze vraag te worden beantwoord: hoeveel personen worden gemist door deze drie gekoppelde registers. Verblijfsduur wordt in het loglineaire model als covariaat opgenomen en is van cruciale betekenis om te schatten hoeveel inwoners er zijn, omdat de EU vereist dat de verblijfsduur minimaal 12 maanden is. Echter, de verblijfsduur was enkel te bepalen voor het Populatieregister en het WNB en niet voor het HKS.

Voor het Populatieregister werd de verblijfsduur afgeleid als het verschil tussen de dag van inschrijving en het peilmoment. Door gebruik te maken van het WNB konden we voor het Populatieregister ook gebruik maken van de duur van banen. Als deze duur langer was dan de lengte van inschrijving in het Populatieregister dan werd deze duur gebruikt. Voor personen die alleen in het WNB voorkwamen konden we de duur van banen gebruiken om een verblijfsduur af te leiden, onder de aanname dat iemand die in Nederland een baan heeft ook hier verblijft. Voor de mensen in het HKS die niet gekoppeld konden worden aan het Populatieregister of het WNB was er voor verblijfsduur een ontbrekende waarde. Voor het HKS is besloten deze waarden te imputeren. Verschillende methoden zijn hiervoor beschikbaar en het is onbekend welke methode het beste is. Daarom is besloten een sensitiviteitsanalyse uit te voeren voor de uiteindelijke vangst-hervangst schattingen. Voor de ontbrekende gegevens zijn via drie methoden waarden ingevuld en de resulterende populatieschattingen zijn met elkaar vergeleken. Twee van de drie scenario's gebruikten het Expectation Maximization-algoritme (EM; Schafer, 1997), en het laatste scenario maakte gebruik van Predictive Mean Matching (PMM; Van Buuren, 2012).

We geven hieronder een voorbeeld van personen met een Poolse nationaliteit in 2010. In tabel 1 is de kruistabel gegeven van het al dan niet voorkomen in de drie registers, uitgesplitst naar korter of langer verblijvend dan 12 maanden. In de tabel zien we het aantal Polen in Nederland in 2010 die in minstens 1 van de drie registers staan geregistreerd. Zo zitten er 32 Polen in zowel de PR, HKS als de WNB met een verblijfsduur van korter dan 12 maanden. Maar bijvoorbeeld, 3.523 Polen zitten wel in de WNB en de PR, maar niet in de HKS met een verblijfsduur korter dan 12 maanden. Er zijn twee (structurele nullen): deze nullen verwijzen naar de mensen die in geen van de drie registraties voorkomen. Daarnaast zijn er mensen die alleen in de HKS voorkomen en niet in de andere twee registers. Voor hen is de verblijfsduur niet af te leiden uit het Populatieregister of het WNB. Dat zijn de twee cellen waar 'Ontbreekt' staat. Het totaal van deze twee cellen is een aantal van 1.043 Polen. Echter, hoeveel daarvan korter verblijven dan 12 maanden en hoeveel ervan langer verblijven dan 12 maanden is niet bekend. Er is dus sprake van ontbrekende waarden, en dit probleem moet als eerste opgelost worden. Hiervoor worden twee manieren gebruikt: het EM-algoritme en PMM.

			HKS = Ja	HKS = Nee
Verblijfsduur <12 maand	PR = Ja	WNB = Ja	32	3.523
		WNB = Nee	34	3.225
	PR = Nee	WNB = Ja	149	60.190
		WNB = Nee	Ontbreekt	0
Verblijfsduur >12 maand	PR = Ja	WNB = Ja	183	21.309
		WNB = Nee	195	14.052
	PR = Nee	WNB = Ja	81	21.216
		WNB = Nee	Ontbreekt	0

Tabel 1. De geobserveerde waardes van mensen met een Poolse nationaliteit die korter of langer dan 12 maanden in Nederland verblijven en die in minstens 1 van de 3 registers staan geregistreerd (PR=Populatieregister, HKS=HerKenningsdienst Systeem, WNB=WerkNemersBestand)

EM-algoritme en PMM

Het EM-algoritme is een iteratief algoritme voor het maken van *maximum likelihood* schattingen van een model wanneer er ontbrekende waarden zijn (Little and Rubin, 2002). Iedere iteratie kent een Expectation (E) en een Maximization (M) stap. In de E-stap worden de verwachtingen van de ontbrekende waarden ingevuld op basis van de geobserveerde waarden (hier het geobserveerde aantal 1.043 en de andere waarden in de bovenstaande tabel) en de schattingen afkomstig uit de M-stap. In de M-stap worden parameters geschat op basis van de gegevens verkregen in de E-stap. Na de M-stap volgt weer een E-stap en dit gaat zo door totdat de parameterschattingen geconvergeerd zijn. In de M-stap wordt steeds een loglineair model geschat. Parameter schattingen van het loglineaire model kunnen gebruikt worden voor het schatten van het aantal personen dat door alle drie registers is gemist.

Twee scenario's gebruiken het EM-algoritme voor completering. In het eerste scenario (EM-A) is een verzadigd model gebruikt om de data te completeren. Op deze gecompleteerde data is vervolgens een spaarzaam loglinear model geschat dat de data het beste beschrijft (volgens het Akaike Informatie Criterium) en op basis hiervan is een schatting gemaakt van het aantal inwoners. In scenario twee (EM-B) is een best passend, spaarzaam loglineair model gebruikt voor completering van de data, waarna de parameters van dit laatste model zijn gebruikt om tot een schatting te komen van het aantal inwoners.

Een derde scenario maakt gebruik van multiple imputatie met *Predictive Mean Matching* (PMM). Kandidaat donoren worden gekozen uit de geobserveerde personen die vergelijkbare achtergrondkenmerken hebben als personen met een ontbrekende waarneming. Een willekeurige donor wordt dan gekozen uit alle kandidaten, die dan als donor dient voor de ontbrekende waarneming (hier: van de waarde voor verblijfsduur).

In de context van het huidige onderzoek is er een belangrijk verschil tussen EM en PMM als datacompleteringsmethode. In ons onderzoek is het niet logisch om de gehele populatie te nemen als donoren. De mensen die geregistreerd staan in het HKS en niet ingeschreven zijn in het Populatieregister lijken qua verblijfsduur veel meer op de mensen die een baan hebben in Nederland maar niet zijn ingeschreven zijn in het Populatieregister. Met PMM kunnen we de donoren beperken tot deze subcategorie, terwijl het EM-algoritme de ontbrekende getallen invult aan de hand van de gegevens van de gehele populatie.

Schatting van de onderdekking van het Populatieregister

De schattingen uit de drie scenario's zijn met elkaar vergeleken om tot de beste schatting van de onderdekking van het Populatieregister te komen. Om tot dit besluit te komen is er allereerst literatuuronderzoek gedaan naar eerder onderzoek met vergelijkbare schattingen (Hoogteijling, 2002; Bakker, 2009; Van der Heijden et al., 2012), waaruit een verwachting kon worden afgeleid voor de schatting voor de cijfers uit 2010. Daarnaast gebruikten we ook informatie over de asielaanvragen in 2010 en de jaren ervoor. De verwachte schatting voor 2010 zou mogelijk kunnen liggen tussen 175 en 225 duizend personen.

GEMIST DOOR ALLE DRIE DE REGISTERS		
SCENARIO	schatting × 1000	betrouwbaarheidsinterval
EM-A	139	120 – 176
EM-B	129	111 – 170
PPM	179	121 – 237

Tabel 2. Schattingen voor de 3 scenario's vanuit de vangst-hervangst voor de drie scenario's; tevens zien we de betrouwbaarheidsintervallen

ONDERDEKKING POPULATIEREGISTER		
SCENARIO	schatting × 1000	betrouwbaarheidsinterval
EM-A	172	153 – 209
EM-B	162	143 – 203
PPM	212	154 – 270

Tabel 3. Schattingen voor de onderdekking van het Populatieregister (let op: dit zijn de schattingen van tabel 2, opgehoogd met 33 duizend mensen)

Tabel 2 toont de resultaten. In de tweede kolom is de vangst-hervangst uitkomst te zien, en de bijbehorende betrouwbaarheidsintervallen staan in kolom drie. Deze twee kolommen beslaan het aantal mensen die door alle drie de registers worden gemist. Echter in het HKS en WNB bleken er 33.000 mensen die niet in het Populatieregister waren geregistreerd maar wel langer dan 12 maanden verbleven. Deze behoren tot het aantal inwoners en behoren dus tot de onderdekking van het Populatieregister. De schattingen moeten dus met 33 duizend worden opgehoogd. Dit is te zien in tabel 3. Kolom 2 van tabel 3 geeft de met 33.000 personen opgehoogde schatting en kolom 3 van tabel 3 de opgehoogde betrouwbaarheidsinterval.

De verschillende scenario's leiden tot verschillende uitkomsten. Onze voorkeur gaat uit naar de schatting met de PMM, omdat de resulterende schatting inhoudelijk het beste aansluit bij de ontbrekende data in de HKS. In totaal gaat het dan om 212.000 inwoners die gemist worden in de BRP. Dit aantal ligt bovendien binnen de range van het referentiekader en is derhalve plausibel. Let wel, deze schatting had alleen betrekking op mensen met een leeftijd tussen de 15 en 65, vanwege de begrenzing van de registers. De schatting moet nog worden aangevuld met een schatting van mensen jonger dan 15 en ouder dan 65. Al met al is er dan maar een onderdekking van het Populatieregister van 0,5 tot 1,1%.

Literatuur

Bakker, B.F.M. (2009). *Trek alle registers open!* (Open all registers!). Amsterdam: Vrije Universiteit Amsterdam.

Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. (1975). *Discrete multivariate analysis*. MIT press, Cambridge, MA.

Buuren, S. van. (2012). *Flexible imputation of missing data*. Boca Raton: CRC Press.

Gerritse, S.C., Bakker, B.F.M., & Heijden, P.G.M. van der. *(2015)*. Different methods to complete datasets used for capture-recapture estimation: estimating the number of usual residents in the Netherlands. *Statistical Journal of the IAOS, 31,* 613-627.

Hoogteijling, E.J.M. (2002). *Raming van het aantal niet in de GBA geregistreerden* (technical report). Heerlen: Centraal Bureau voor de Statistiek.

International Working Group for Disease Monitoring and Forecasting (1995). Capture-recapture and multiple-record systems estimation I: History and theoretical development. *American Journal of Epidemiology, 142*(10), 1047-1058.

Little, R.J., & Rubin, D.B. (2002). *Statistical analysis with missing data*. Hoboken, New Yersey: John Wiley and Sons.

Schafer, J.L. (1997). *Analysis of incomplete multivariate data*. Chapman and Hall.

Schulte Nordholt, E. (2016). Volkstellingen in Nederland en elders, *STAtOR, 17*(1), 8-13.

Heijden, P.G.M. van der, Whittaker, J., Cruyff, M.J.L.F., Bakker, B.F.M., & Vliet, H.N. van der. (2012). People born in the Middle East but residing in the Netherlands: Invariant population size estimates and the role of active and passive covariates. *The Annals of Applied Statistics, 6*, 831-852.

Zwane, E.N., & Heijden, P.G.M. van der. (2007). Analysing capture-recapture data when some variables of heterogeneous catchability are not collected or asked in all registries. *Statistics in Medicine, 26*, 1069-1089.

Susanna Gerritse studeerde psychologie aan de Universiteit van Amsterdam en heeft een promotietraject afgerond aan de Universiteit Utrecht. Deze promotie was in samenwerking met het CBS, waar het bovenstaande onderzoek uit voortvloeide. Momenteel is ze statistisch onderzoeker bij het Centraal Bureau voor de Statistiek.
E-mail: sc.gerritse@gmail.com

Bart F.M. Bakker is hoogleraar Methodologie van Registerdata voor sociaalwetenschappelijk onderzoek aan de VU Amsterdam en is Hoofd van de afdeling Methodologie van het Centraal Bureau voor de Statistiek in Den Haag.
E-mail: Bfm.bakker@cbs.nl

Peter G.M. van der Heijden is hoogleraar Statistiek t.b.v. de sociale wetenschappen aan de Universiteit Utrecht en Professor of Social Statistics at the University of Southampton.
E-mail: p.g.m.vanderheijden@uu.nl

KLASSIEKERS

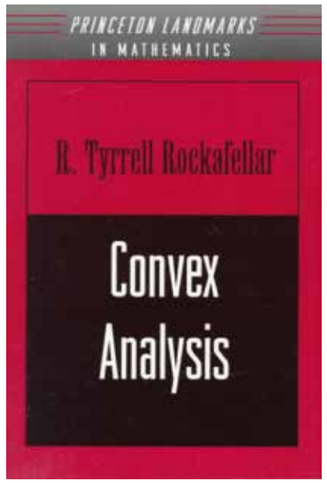
OVER ROCKAFELLAR: CONVEXITEIT, OPTIMALISERING, DUALITEIT EN ONZEKERHEID

Wim Klein Haneveld

In de jaren zestig van de vorige eeuw studeerde ik toegepaste wiskunde in Groningen met naast de vakken kansrekening en statistiek ook operations research (OR). Toen een nieuw vak bij econometrie, dat verzorgd werd door Cor van de Panne. De kennismaking beviel heel goed. En zo werd ik na mijn afstuderen medewerker OR. Dat bleef ik toen Van de Panne zelf naar Calgary vertrok en in 1969 werd opgevolgd door Jaap Ponstein. Lineaire programmering (LP) was uiteraard belangrijk, en daarin stond de Simplexmethode weer centraal. Maar de fundamentele oorsprong van het duale probleem en de conceptuele betekenis van schaduwprijzen bleven enigszins vaag en gaven indertijd regelmatig aanleiding tot felle discussies.

Convexiteit

De wiskundige R.T. (Terry) Rockafellar heb ik voor het eerst horen spreken in 1970 op het Mathematical Programming Symposium in Den Haag. Zijn presentatie was helder en overtuigend. Hij behandelde onderwerpen uit zijn nieuwe boek *Convex Analysis* (Rockafellar, 1970). Dat heb ik direct aangeschaft. Ik wist toen nog niet dat dit boek van 470 pagina's een uitwerking was van zijn proefschrift aan Harvard (1963). Het boek zag er wat saai uit, geen plaatjes bijvoorbeeld, maar het bleek een zeer goed gestructureerde aanpak te bevatten van convexiteit in $\mathbb{R}^n$ in de context van optimaliseren en differentiëren. Uit de klassieke analyse weet iedereen, dat bij convexe functies een locaal minimum tevens globaal is, en dat in een locaal minimum de afgeleide nul is. Tenminste als de functie differentieerbaar is, en er geen nevenvoorwaarden zijn. Rockafellar pakt het omgekeerd aan. De geldigheid van de optimaliteitsconditie staat centraal, en het begrip afgeleide moet zich aanpassen. Voor hetzelfde doel moeten het domein van de functie en potentiële nevenvoorwaarden de aandacht niet afleiden. Zo wordt de functiewaarde $+\infty$ voor de doelstellingsfunctie geïntroduceerd voor ontoelaatbare waarden van de beslissingsvariabelen, en het begrip afgeleide voor convexe functies gegeneraliseerd tot subgradiënt. Zo krijgt de optimaliteitsconditie een algemene setting. Natuurlijk heb je daar in de praktijk alleen wat aan als er een passende calculus wordt ontworpen, en dat is precies wat er in Convex Analysis gebeurt. Vanaf het begin. Convexe functies zijn functies met een convexe epigraaf (grafiek met alle punten erboven er aan toegevoegd), en hun eigenschappen volgen dus uit die van convexe verzamelingen. Dus wordt eerst betrekkelijk volledig de calculus van convexe verzamelingen behandeld, en daarna toegepast op (uitgebreid-reële) convexe functies op R^n en op convexe optimaliseringproblemen.



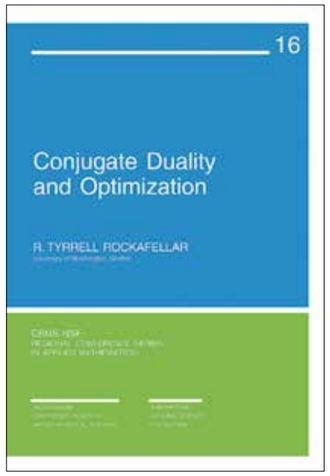
Dualiteit

Ook dualiteit vindt hier zijn basis. Een (gesloten) convexe verzameling heeft een duale representatie: ze valt samen met de doorsnijding van alle gesloten halfruimtes die de verzameling omvat. Deze collectie halfruimtes is zelf ook voor te stellen als een gesloten convexe verzameling, in de duale ruimte. En de duale van de duale is weer de oorspronkelijke verzameling! Toegepast op convexe functies leidt deze dualiteit tot geconjugeerde convexe functies. Convexe minimaliseringproblemen blijken een veelvoud van duale problemen te kunnen genereren, als volgt. Introduceer verstoringvariabelen in het gegeven minimaliseringprobleem, en open de mogelijkheid om zulke storingen te kopen tegen nog te bepalen prijzen. Het bijbehorende duale probleem beoogt deze prijzen, de duale variabelen, zodanig te bepalen, dat het gebruik van storingen maximaal ontmoedigd wordt. Onder zwakke regulariteitvoorwaarden is bij optimale prijzen de aankoop van storingen niet winstgevend. Primaire en duale optimale oplossingen worden dan gekarakteriseerd als zadelpunt van de bijbehorende Lagrangiaan. Dit dualiteitsschema generaliseert de klassieke theorie, zoals gebruikt in LP, waar de verstoring altijd een verandering van de rechterleden van de beperkingen betreft, leidend tot Lagrange multiplicatoren als duale variabelen.

Uitgespit

Met de vakgroep OR en andere wiskundigen van de Interfaculteit Econometrie hebben we indertijd het boek volledig uitgespit. Het werd onze standaard voor convexe programmering en de bijbehorende dualiteitstheorie. Daar viel niets aan toe te voegen; als je iets nodig had, zocht je het gewoon op. Alleen langs de lijn van generalisatie was uitbreiding mogelijk. En dat probeerden we dan ook. Er bestaan natuurlijk lineaire ruimtes, algemener dan $\mathbb{R}^n$, zoals functieruimtes, waarin prachtige optimaliseringproblemen hun plek vinden. Het schema van Rockafellar blijft in principe ook daar goed toepasbaar. Maar de keuze van de topologie is dan cruciaal, en de functionaalanalytische consequenties zijn verre van triviaal. Veel optimaliseringproblemen passen bijvoorbeeld goed in een L^∞ ruimte (begrensde functies), maar de norm-duale ruimte hiervan is onhanteerbaar (eindig-additieve verzamelingsfuncties). Jaap Ponstein stortte zich in het diepe en deed daar leuke dingen. Dat probeerde ik ook. Hans Nieuwenhuis richtte zich op generalisaties van meervoudige doelstellingen, en Gerard Sierksma ging na hoe convexiteit in ruimtes zonder lineaire structuur gedefinieerd zou kunnen worden.

Rockafellar zat zelf ook niet stil. In 1974 verscheen zijn verrassend dunne maar inhoudsrijke werk *Conjugate duality and optimization* (Rockafellar, 1974). Het bevatte precies de dualiteitstheorie in oneindig-dimensionele lineaire ruimten waarnaar ik op zoek was, met talloze toepassingen. Voor mij was hiermee de 'convexiteitsbijbel' compleet. Naast het 'Oude Testament', waar de principes in $\mathbb{R}^n$ definitief hun plek kregen kwam nu ook het 'Nieuwe Testament' ter beschikking waarin precies werd uitgewerkt wat deze principes in algemenere ruimtes konden betekenen. Vanaf dat moment werd ik een enthousiaste toepasser van deze theorie.



Kansmaten als beslissingsvariabelen

Met name allerlei problemen waarin kansmaten als beslissingsvariabelen optreden vind ik een leuk toepassingsgebied, onder meer omdat doelstelling en beperkingen veelal lineair zijn. Het gaat bijvoorbeeld om de verwachtingswaarde van een functie van een stochastische vector (een zogenaamd gegeneraliseerd moment), en de achterliggende integraal is niet alleen lineair in de functie bij vaste maat, maar ook lineair in de gehanteerde maat bij een vaste functie. Vaak gaat het over worst-case verdelingen, waarbij sommige gegeneraliseerde momenten, bijvoorbeeld de verwachtingswaarde van de stochast zelf, worden voorgeschreven of naar boven begrensd. Of waarbij marginale verdelingen gefixeerd zijn maar afhankelijkheden vrij zijn. Als resultaat krijg je een abstract LP-probleem, met beslissingsvariabelen die natuurlijk niet-negatief zijn omdat kansen dat zijn. Een van de charmes van lineaire programmering is, dat de primaire variabelen niet optreden in het duale probleem, en omgekeerd, wanneer de dualiteit gebaseerd is op de traditionele storingen op de rechterleden van de beperkingen. Veelal komt het duale probleem neer op een functie approximatie. Voor de optimisten is er een eenvoudige aanpak: als het lukt om aan *complementaire speling* te voldoen (dat betekent hier: de kansmaat beperken tot die punten waar de functie approximatie exact is), heb je de optimale oplossing te pakken. Dat lukt in allerlei specifieke gevallen, doordat de functie approximatie goed te doen is. Pas als je op zoek bent naar voldoende voorwaarden die garanderen dat er geen dualiteitskloof is, en dat primair en/of duaal probleem een optimale oplossing hebben komen de diepere resultaten van Rockafellar goed van pas. Ik heb dit schema vaak toegepast, ook op bekende problemen. Je komt zo met standaard dualiteitstheorie tot bekende optimaliteitscondities. Dat vind ik prachtig. Ook al kunnen specialisten

bij specifieke gevallen het resultaat soms met iets zwakkere onderstellingen bewijzen, ik vind algemeen inzicht onmisbaar. De betrokkenheid van de vakgroep OR bij convexe analyse heeft ertoe geleid dat Terry Rockafellar in 1984 een eredoctoraat van de RuG heeft ontvangen. Ik herinner me nog de bevlogen laudatio, die Jaap Ponstein als erepromotor bij de officiële plechtigheid in de Martinikerk uitsprak. Met terugwerkende kracht had deze onderscheiding trouwens een voorspellende waarde: later werd Rockafellar herhaaldelijk onderscheiden met eredoctoraten en prestigieuze prijzen.

Stochastische programmering

Terry en ik hebben meer gemeenschappelijke interesses, waaronder stochastische programmering. Het gaat daarbij om modellen voor beslissingen onder onzekerheid. In concreto om LP-problemen in $\mathbb{R}^n$, waarbij sommige parameters geen gegeven waarde hebben maar gerepresenteerd worden door stochasten. Dat vraagt om herformulering van het deterministische model, omdat de beslissingen genomen moeten worden terwijl de realisaties van de stochasten (nog) niet bekend zijn wanneer de beslissingen moeten worden genomen. De gebruikelijke specificatie van toelaatbaarheid (aan alle beperkingen is voldaan) en optimaliteit (er bestaat geen toelaatbare oplossing met een betere waarde van de doelstellingsfunctie) wordt zinloos als beperkingen en doelstelling afhangen van onbekende parameters. Beslissingen moeten, vaker wel dan niet, niet-anticipatief zijn: ze mogen niet afhangen van de (vooralsnog) onbekende uitkomst van de stochasten.

Er zijn talloze praktische problemen die in deze context passen. Wat een geschikte herformulering is, hangt af van het toepassingsgebied. Bij serieuze onzekerheid is de aanpak die stochastische parameters vervangt door de verwachtingswaarde ervan nogal naïef. Immers, in LP-modellen stellen de beperkingen vaak harde eisen of wensen voor, waar je in geen geval van wilt afwijken ongeacht wat dat in termen van doelstelling kost. De naïeve aanpak leidt evenwel veelal tot een riskant resultaat: gemiddeld zit je misschien wel goed, maar feitelijk toch vaak fout en soms zelfs heel erg fout. Een klassieke manier om hier iets aan te doen is het vervangen van de stochastische beperking door de beperking, dat aan de eis of wens toch zeker met een kans van minstens 95% moet zijn voldaan. Stochastische programmeurs noemen dit van de statistiek afgekeken concept een kansbeperking. In de financiële wereld staat de achterliggende risicomaatstaf bekend als VaR (Value-at-Risk). De oplettende lezer zal opmerken, dat wij hier impliciet zijn uitgegaan van een ongelijkheidbeperking. En dat bij gelijkheidbeperkingen, in LP niet ongebruikelijk, de kansbeperking zijn betekenis verliest. Immers, hoe de beslissingsvariabelen ook worden gekozen, de kans dat de realisaties van de stochastische coëfficiënten voor de gelijkheid in de beperking zorgen is klein, zeker maar niet alleen als er continue verdelingen achter zitten. Het is dan ook geen wonder, dat de meest gebruikte modelvorm anders met risico omgaat. Bij zogenaamde *recourse* modellen worden afwijkingen als feitelijkheid geaccepteerd, maar het model wordt uitgebreid met herstelbeslissingen achteraf, die ervoor zorgen dat aan de achterliggende eisen/wensen alsnog wordt voldaan. De oorspronkelijke beslissingen houden rekening met de ongewenste effecten achteraf, doordat de (verwachte) kosten van herstel in de doelstelling worden meegenomen.

Op de groepsfoto van de ICSP in 2016 in Brazilië staat Rockafellar naast Maarten van der Vlerk en Ward Romeijnders (in 2016 gepromoveerd bij Maarten van der Vlerk en mij). Tijdens de ICSP 2016 werd Ward onderscheiden voor zijn artikel als (bijna) het beste van aankomende onderzoekers op ons vakgebied in de afgelopen drie jaren. Als lid van de promotiecommissie heeft Rockafellar waardevolle input gegeven voor het vervolg van het onderzoek op het gebied van de stochastische gemengd-geheeltallige programmering.

Wetenschappelijke hobby

Naast convexiteit en dualiteit werd de stochastische programmering mijn permanente wetenschappelijke 'hobby'. Zo'n veertig jaar geleden gaf ik voor het eerst een cursus over dit vakgebied. Deze cursus loopt overigens nog steeds. Tegenwoordig verzorgt mijn opvolger Maarten van der Vlerk voor het LNMB een steeds betere versie ervan in Groningen en in Utrecht. Een van mijn inspiratoren op dit gebied was Roger Wets. Hij heeft als een van de eersten recourse-modellen grondig geanalyseerd, en algoritmen ervoor ontworpen. Bij hem heb ik in de jaren zeventig een jaar doorgebracht, aan de University of Kentucky in Lexington.

Nu wil het geval, dat Terry door Roger ook in de stochastische programmering geïnteresseerd is geraakt, en samen hebben ze er menig artikel over geschreven. Fundamentele verhandelingen, waar anderen zoals ik veel van leerden. Zoals over *Lagrange multipliers voor nonanticipativity constraints*, waar dualiteit en stochastisch programmeren samenkomen. Ik ontmoette beide creatieve onderzoekers minstens eens in de drie jaar op de ICSP (International Conference on Stochastic Programming). Terwijl ik tegenwoordig van mijn pensioen geniet, gaan beide heren ondanks hun 80+ leeftijd gewoon door soms, maar niet altijd, als tandem. In juni 2016 heeft Terry naar verluidt nog een prachtig verhaal gepresenteerd op de laatste ICSP in Brazilië.

Conditional Value-at-Risk

De mooiste bijdrage van Terry Rockafellar aan de stochastische programmering, in ruime zin, vind ik de ontwikkeling van het concept CVaR Conditional Value-at-Risk). Enerzijds is dit een moderne risicomaatstaf, die aan allerlei gewenste eigenschapen voldoet (coherentie). In die zin heeft de CVaR conceptueel gezien betere papieren dan de alom gehanteerde VaR. En Rockafellar propageert dit concept met verve in de wereld van risicomanagement in de financiële wereld. Anderzijds is het begrip goed te implementeren in recourse modellen, omdat het de convexiteit in het model niet nodeloos verstoort, zoals wel het geval is bij kansbeperkingen. En dat is weer belangrijk voor het vinden van efficiënte algoritmen. In het verleden heb ik om die reden het begrip geïntegreerde kansbeperkingen geïntroduceerd, waar in wezen niet de kans op een overschrijding wordt beperkt, maar de verwachte omvang ervan. Achteraf gezien was ik op zoek naar wat Rockafellar later vond, en in een overtuigende context plaatste. Niet voor het eerst, en dat verklaart allicht de grote bewondering die ik voor hem heb. Als u dat geen overtuigend argument vindt, neem dan kennis van zijn homepage (Rockafellar, 2016) met een gigantische lijst van publicaties.

Op de 10e ICSP in 2004 in Tucson (AZ) werd ik tot mijn verbazing geëerd als een van de twaalf 'pioniers' van de stochastische programmering, naast mensen als bijvoorbeeld George Dantzig, Terry Rockafellar en Roger Wets die wat mij betreft tot de buitencategorie behoren.

Verder dan convexe analyse

In 2014 bestond de RuG 400 jaar. Als eredoctor van 30 jaar geleden was Terry een van de eregasten van het feest. Tussen alle festiviteiten door was er ook nog gelegenheid tot ontspanning. Ooit had Terry vanuit zijn thuisbasis in Seattle een bergwandeling georganiseerd in de Rocky Mountains, waaraan Maarten van der Vlerk en ik meededen. In Groningen konden we de rekening vereffenen door met zijn drieën een wadlooptocht te maken. Typerend vond ik, dat Terry niet aarzelde om een pad te zoeken dat afweek van wat de gids had aangeraden, maar toch heelhuids uit de blubber te voorschijn kwam. Ook de wetenschap werd niet vergeten. Voor onze Faculteit Economie en Bedrijfskunde presenteerde hij een boeiende vergelijking van risicomaatstaven in de optimalisering en de statistiek. Voor de Faculteit Wiskunde en Natuurwetenschappen hield hij een ander verhaal, over variationele analyse. Rockafellar en Wets hebben daar een dik standaardwerk over geschreven (Rockafellar, 1997) waarvan ondertussen al de derde verbeterde druk is verschenen. Zelf heb ik het nooit gelezen, omdat het veel verder gaat dan convexe analyse. Misschien wel jammer, want na zijn verhaal bleek het onderwerp toch interessanter te zijn dan ik had gedacht, op basis van vroeger werk over gegeneraliseerde afgeleiden. Mijn buurman tijdens het colloquium, een wiskundige die ik niet kende, vertelde dat voor hem dit boek als bijbel functioneert: 'Het ligt op mijn bureau, dan kan ik het naslaan als ik iets nodig heb.' Dan zwijg ik in gepaste stilte en bewondering.

W. Klein Haneveld en R.T. Rockafellar (rechts)

Literatuur

Rockafellar, R.T. (1970). *Convex Analysis*. Princeton University Press.

Rockafellar, R.T. (1974). *Conjugate Duality and Optimization. SIAM Publications*.

Rockafellar, R.T. www.math.washington.edu/~rtr/mypage.html. Geraadpleegd juli 2016.

Rockafellar, R.T., & Wets, R.J-B (1997, third improved printing 2005). *Variitional Analysis*. Springer Verlag.

Wim Klein Haneveld is emeritus hoogleraar Operations Research bij de vakgroep Econometrie en Operations Research van de Rijksuniversiteit Groningen.
E-mail: w.k.klein.haneveld@rug.nl

Oud-voorzitter prof. dr. ir. Paul van der Laan (1937–2016)

Paul van der Laan werd in Utrecht geboren op 16 februari 1937. Hij ging naar de Lagere School in Rotterdam en doorliep de 1e Christelijke HBS aan het Henegouwerplein te Rotterdam en behaalde in 1954 het eindexamen van de B-afdeling. Daarna studeerde hij wis- en natuurkunde aan de Universiteit van Leiden waar hij in 1957 het kandidaatsexamen en in 1960 het doktoraalexamen aflegde. Van 1958 tot 1960 was hij verbonden als assistent statistiek bij prof. dr. N. H. van der Vaart van het Instituut voor Theoretische Biologie van de Universiteit van Leiden. Van 1960 tot 1965 was hij medewerker van de afdeling Mathematische Statistiek van het Mathematisch Centrum te Amsterdam.

Hij trouwde op 5 september 1961 met An de Vries. Uit dit huwelijk kwamen vier kinderen voort: Robert, Annette, Mark en Marianne.

Sinds 1965 was hij werkzaam bij de NV Philips' Gloeilampenfabrieken te Eindhoven, tot 1967 op het Natuurkundig Laboratorium en vanaf 1967 bij de hoofdgroep Informatie Systemen en Automatie, afdeling Research, als Leider van de groep Statistiek en Waarschijnlijkheidsrekening. Op 26 juni 1970 promoveerde hij aan de Technische Hogeschool Eindhoven; zijn dissertatie was getiteld: *Simple Distribution-Free Confidence Intervals for a Difference in Location*. Zijn promotoren waren prof. dr. ir. L. C. A. Corsten (Landbouwhogeschool Wageningen) en prof. dr. R. Doornbos (Technische Hogeschool Eindhoven).

In 1972 werd hij benoemd tot hoogleraar in de Wiskundige Statistiek aan de Landbouwhogeschool (LH) Wageningen. Zijn inaugurele rede – in kapitalen – was getiteld *ZEKERHEID ONDANKS ONZEKERHEID*. Zijn wetenschappelijke belangstelling was de Verdelingsvrije Methoden en Selectiemethoden. Aan de LH was hij zeer gezien hetgeen resulteerde dat hij in 1981 benoemd werd in het College van Decanen als lid en van 1984 tot 1987 was hij secretaris van het College.

In 1990 aanvaardde hij het hoogleraarschap Wiskundige en Toegepaste Statistiek aan de Technische Universiteit Eindhoven (TUE). Hier was hij van 1994 tot 1996 decaan van de faculteit Wiskunde en Informatica. Op 31 augustus 2001 hield hij aan de TUE zijn afscheidscollege *zekerheid dankzij onzekerheid*. Uit de titel van zijn inaugurele rede aan de LH in hoofdletters en de titel van zijn afscheidscollege aan de TUE in kleine letters blijkt dat hij deemoediger was geworden.

Hij was promotor van vijf promovendi aan de LH Wageningen en van zeven promovendi aan de TU Eindhoven.

Van 1987–1991 was hij voorzitter van de VvS + OR (Netherlands Society for Statistics and Operations Research). Voor zijn emeritaat was hij Lid van de commissie 'Veiligheidsbeoordeling nieuwe voedingsmiddelen' van de Gezondheidsraad en lid van de Medisch Ethische Commissie van het St. Anna Ziekenhuis in Geldrop.

Op 9 juli 2016 overleed Paul te Eindhoven. De statistische genen van Paul zijn bewaard gebleven bij zijn zoon Mark, die nu hoogleraar Biostatistics and Statistics is aan de University of California, Berkeley School of Public Health.

Wij gedenken Paul met grote eerbied en respect en wensen zijn familie en vrienden sterkte toe bij dit verlies.

Fred Steutel, Rob Verdooren & Frank Coolen

PERSOONLIJKE HERINNERINGEN AAN PAUL VAN DER LAAN

Ik heb het grote genoegen gehad mijn promotieonderzoek aan de Technische Universiteit Eindhoven (TUE) te doen onder begeleiding van Paul. Ik begon in mei 1990, enkele maanden na Pauls komst naar de TUE, en wat mij in het bijzonder aantrok was zijn bereidheid mij volledig mijn eigen gang te laten gaan. Mijn promotieonderzoek, naar methodes voor het gebruik van expert kennis, was ver verwijderd van Pauls onderzoeksgebieden, maar hij was zeer geïnteresseerd en geboeid door het onderwerp en onderzoek, en stelde steeds zeer goede vragen. Hij was tevens enthousiast en zorgzaam, ook waar het aangelegenheden in het leven buiten het onderzoek aanging.

Na mijn promotie, en vertrek naar de Universiteit van Durham in Engeland, begonnen we daadwerkelijk samen te werken, eerst aan het combineren van twee verschillende klassieke statistische selectie methoden, daarna aan gerelateerde niet-parametrische methoden. Deze samenwerking en verscheidene bezoeken van Paul aan Durham, waren bijzonder plezierig, met goed werk afgewisseld met interessante gesprekken over allerlei zaken, waardoor Paul mijn visie zowel op onderzoek als andere, belangrijkere, zaken in het leven zeer positief beïnvloed heeft.

De bovengenoemde onderwerpen, statistische selectie methoden en niet-parametrische statistiek, boeiden Paul en hij heeft belangrijke bijdragen geleverd door zijn onderzoek op deze gebieden, met name ook middels langdurig en succesvol samenwerken met Constance van Eeden en Subhabrata Chakraborti. In presentaties over zijn onderzoek benadrukte Paul altijd het belang van de juiste statistische theorie en methoden om praktische problemen aan te pakken, de wisselwerking tussen academisch onderzoek en praktische toepassingen was belangrijk voor hem.

Frank Coolen, frank.coolen@durham.ac.uk

Met Paul en An maakte ik in 1966 kennis op de European Meeting of Statisticians te Londen. In 1972 werd Paul benoemd als hoogleraar aan de Landbouwhogeschool (LH) Wageningen met de leeropdracht Wiskundige Statistiek, naast prof. dr. ir. L. C. A. Corsten, die sinds 1964 dezelfde leeropdracht had. Als medewerker werd ik aan Paul toegewezen. Paul wilde graag kennis maken met de vakgroepen die veel statistiek toepasten in hun onderzoek. Op de fiets vergezelde ik hem met deze bezoeken om hem te introduceren. Bij de vakgroep Zoötechniek zei de hoogleraar: 'Dit is de eerste keer dat een hoogleraar Wiskundige Statistiek ons bezoekt.' Paul werd zeer geliefd bij de LH, dat bleek omdat hij in 1981 gevraagd werd om als eerste hoogleraar Wiskundige Statistiek lid te worden van het College van Decanen.

Paul en ik bezochten samen diverse conferenties in het buitenland. Daar Paul belangstelling had in de Verdelingsvrije Methoden hielden wij samen in 1986 in de DDR de voordracht 'Classical analysis of Variance Methods and Nonparametric Counterparts', dat gepubliceerd werd in de *Biometrical Journal* (1987). In 1988 promoveerde ik bij Paul, waarbij hij mij zeer stimuleerde met kritische opmerkingen. Zijn belangstelling in de Selectiemethoden resulteerde in een aantal artikelen en twee dissertaties onder zijn leiding. Paul vroeg mij om samen met hem een overzichtsartikel te schrijven dat in 1989 in de *Biometrical Journal* verscheen 'Selection of populations: an overview and some recent results'.

Na zijn vertrek naar de Technische Universiteit Eindhoven kwamen mijn vrouw en ik, en An en Paul, regelmatig op elkaars verjaardagen. Na zijn emeritaat ging Paul zich meer toeleggen op filosofie. Hij hield daar ook lezingen over voor Probus Eindhoven.

Rob Verdooren, l.r.verdooren@hetnet.nl

Paul en ik leerden elkaar kennen in 1960 op het Mathematisch Centrum (MC) – nu CWI – in de Tweede Boerhaavestraat in Amsterdam. Hij was in Leiden afgestudeerd, ik had na mijn diensttijd net kandidaatsexamen gedaan. In de middagpauze werd er geschaakt. Ik deed maar sporadisch mee, want ik kon eigenlijk niet schaken. Toch heb ik Paul, die wel kon schaken, een keer verslagen; hij was onvoorzichtig: 'mat op de onderste lijn'. We zijn desondanks vrienden geworden.

Na het MC gingen we ieder onze eigen weg; Paul naar Wageningen, ik naar Twente. We kwamen elkaar af en toe tegen op een congres.

Veel later was er een vacature voor een hoogleraarschap statistiek in Eindhoven. Er waren drie gegadigden, waarvan er al snel één afviel. Paul was een van de overblijvende twee. Over de benoemingsprocedure was verwarring ontstaan. Mij werd door de benoemingscommissie gevraagd deze verwarring weg te nemen. Daartoe troffen Paul en ik elkaar op het terras van de uitspanning Kievitsdel in Doorwerth, niet ver van Wageningen. Ik weet niet meer hoe lang het geduurd heeft en wat we gegeten of gedronken hebben; ik houd het op koffie en na de goede afloop een pilsje. Het resultaat was dat Paul naar Eindhoven kwam als opvolger van Roel Doornbos, die tweede promotor was geweest bij zijn doctoraat.

Onze specialismen lagen ver uiteen en we hebben nooit samen wiskunde gedaan; we troffen elkaar in de koffiekamer en af en toe in een promotiecommissie.

We zagen elkaar wel vaak in de schouwburg of de concertzaal. We hadden jarenlang gezamenlijke abonnementen voor muziek en toneel. Het ene stel haalde dan het andere stel met de auto op en bij het 'andere' werd dan na afloop een gezellig glaasje wijn gedronken en een wereldprobleem opgelost. Ik weet niet waarom we met de cultuurevenementen gestopt zijn. We werden wat ouder en gingen niet graag in het donker de weg op.

We kwamen wel regelmatig bij elkaar op bezoek voor koffie of thee en een glaasje. Ongeveer een half jaar geleden zaten Vita en ik nog op een zonnige middag bij de Van der Laans in hun mooie tuin herinneringen op te halen. Kort daarna ging het snel slechter met Paul, en was het, in ieder geval voor ons, onverwacht snel afgelopen.

Paul was altijd gelovig gebleven en er werd afscheid genomen in een Christelijke kerk. Oud-promotor Roel Doornbos kon tot zijn spijt niet bij het afscheid zijn, maar collega Willem Schaafsma was helemaal uit Groningen gekomen. De receptie vond plaats bij de uitspanning Karpendonkse Hoeve. Het was prachtig weer en de receptie werd een soort feestje. Ik denk dat Paul daar wel vrede mee gehad zou hebben.

Fred Steutel, fsteutel@xs4all.nl

Prof. dr. Leo Kroon
(1958–2016)

Op 14 september 2016 is Leo Kroon (*4 november 1958) onverwacht overleden. Leo was hoogleraar kwantitatieve logistiek aan de Rotterdam School of Management, Erasmus Universiteit, en daarnaast part-time werkzaam als logistiek adviseur bij de Nederlandse Spoorwegen (NS).

Leo studeerde af in de wiskunde aan de VU Amsterdam (cum laude), op een scriptie bij prof. Harm Bart, in 1982. Vanaf 1984 was hij werkzaam aan de Rotterdam School of Management. Hij promoveerde in 1990 bij Prof. Jo van Nunen, op een proefschrift over *Job scheduling and capacity planning in aircraft maintenance.*

Leo bestudeerde beslissingsondersteuning voor de planning en real-time besturing van logistieke systemen. Hij was vooral geïnteresseerd in het openbaar vervoer, voornamelijk per trein. In 1996 is hij daarom ook part-time gaan werken bij de NS. Die eerste jaren bij NS waren echte pionierstijden: mouwen opstropen en met studenten rekenen aan een betere dienstregeling, materieel- en personeelsplanning. Zo heeft Leo onder andere gerekend aan het Lusten & Lasten-model en op die manier bijgedragen aan het oplossen van de arbeidsonrust tegen het 'rondje om de kerk'.

Zijn oratie uit 2001 was getiteld *Opsporen van sneller en beter*. In zijn oratie gaf hij aan wat hij wilde bereiken op het gebied van spooronderzoek. Dat was veel. Hij identificeerde een groot aantal logistieke problemen op het spoor: planning van treinen in stations, ontwerp van het spoorboekje, de planning en bijsturing van materieel en personeel. Dit heeft geresulteerd in meer dan 100 publicaties in tijdschriften als *Operations Research, Transportation Science*, *Transportation Research B*, en *Interfaces*. Daarnaast heeft hij vele promovendi en afstudeerders begeleid die in de wetenschap en het bedrijfsleven terecht zijn gekomen.

Leo (met zijn team) heeft veel blijken van erkenning gekregen voor zijn onderzoek. De belangrijkste erkenning was het winnen van de Franz Edelman Award in 2008 voor de ontwikkeling van een totaal nieuwe dienstregeling op het Nederlandse spoor. In de totstandkoming van deze dienstregeling, ingegaan in december 2006, en de bijbehorende materieel- en personeelsplanning is gebruik gemaakt van verschillende OR modellen en technieken. Met als resultaat dat niet alleen veel afstudeerders en promovendi geprofiteerd hebben van Leo zijn werk, maar ook al die miljoenen Nederlanders die regelmatig met de trein reizen. Immers, sinds die nieuwe dienstregeling rijden er meer treinen op tijd en is de klanttevredenheid van treinreizigers enorm toegenomen. Ook de nieuwe dienstregeling die onlangs is ingegaan, waarbij getracht is klantreistijden te verkorten draagt mede zijn stempel.

Tevens ontving Leo de ERIM impact award in 2008 en de Best paper award van het World Congress on Railway Research in 2011 (thema *even more trains, even more on time*). Daarnaast won hij een Strategic University Relationship (SUR) samenwerking met IBM in 2013, voor innovatief onderzoek in openbaar vervoer.

Leo was een begenadigd onderzoeker en een geweldige coach voor zijn PhD-studenten (tweemaal winnaar van de TRAIL best PhD supervisor award). Leo was bescheiden, bedachtzaam en kweet zich naar eer en geweten van zijn verantwoordelijkheden in de vakgroep en als voorzitter van de examencommissie van de faculteit. Hij is onder grote belangstelling begraven op 21 september 2016. In onze gedachten leeft hij voort

René de Koster, Sectieleider Supply Chain Management, Rotterdam School of Management, Erasmus Universiteit. E-mail: rkoster@rsm.nl
Dennis Huisman, Expertise Manager Logistieke Processen, Proceskwaliteit & Innovatie, NS en hoogleraar openbaar vervoer optimalisatie bij het Econometrisch Instituut, Erasmus Universiteit. E-mail: huisman@ese.eur.nl

Installatie van de kunstenaarsgroep Luminarie De Cagna, te zien tijdens Glow 2013 in Eindhoven

DE NACHT VAN EINDHOVEN – een nacht vol wiskunde en statistiek

Pleuni Naus

Afgelopen juni organiseerden we als CQM de Nacht van Eindhoven, dat deden we dit jaar voor de tiende keer, dus reden voor een extra uitdagende editie! De Nacht van Eindhoven is een wedstrijd waarbij negen teams bestaande uit drie (overwegend) masterstudenten wiskunde, econometrie en informatica van verschillende universiteiten de strijd met elkaar aangaan. Gedurende één nacht wordt op ons kantoor hard gewerkt aan opdrachten in toegepaste statistiek, operations research en optimalisatie. Deze opdrachten zijn door CQM'ers gemaakt en geïnspireerd op cases uit de praktijk. Kortom opdrachten die wij zelf ook leuk vinden en waar we goed in zijn.

De studenten werkten gedurende nacht aan vier verschillende opdrachten, elk met eigen deadlines. Hierdoor moesten de studententeams goed samenwerken en plannen om alle opdrachten op tijd in te leveren. Even indutten konden de studenten zich dus niet veroorloven.

Het programmeren van een lift

In deze opdracht gaat het om het aansturen van de liften van een van de negen vestigingen van een grote organisatie. Het doel is de wachttijd van personeel zo kort mogelijk te houden. Dit is echter niet de reële wachttijd, maar de gevoelsmatige tijd die men staat te wachten; mensen zijn namelijk gevoelig voor de sfeer in de lift. Deze sfeer wordt bepaald door de muziek die in de lift aanstaat, hierdoor kan een ritje met de lift gevoelsmatig korter duren. De muziekjes van de negen vestigingen worden echter geregeld vanuit één centraal systeem, dit zorgt ervoor dat verschillende aanvragen van muziek door elkaar heen kunnen gaan lopen. Voor de mensen in de lift klinkt dit dan zeer onprettig, waardoor de tijd gevoelsmatig langer duurt! Elk team stuurt de liften in een van de vestigingen aan en moet daarbij slim gebruikmaken van het instrument muziek.

Het bleek voor de studenten redelijk lastig te zijn om de lift goed te laten opereren; alle mensen naar de gewenste verdieping brengen zonder lange wachttijd. De aanpak die CQM zou kiezen is het maken van een simpel en constructief algoritme dat snel oplost. In dit algoritme kun je verschillende keuzes random maken, hierdoor kun je veel oplossingen doorrekenen en de beste selecteren.

Consultant voor een nacht

Deze opdracht gaat over het verkrijgen van informatie uit reviews van vluchten. Hierbij stonden de studenten een nacht lang in onze schoenen als consultants. Ze moesten vragen beantwoorden als: Hoe beoordelen reizigers vliegtuigmaatschappijen en welke factoren beïnvloeden de tevredenheid van reizigers? Hoe kan de vliegtuigmaatschappij zich (met zo min mogelijk kosten) onderscheiden van andere maatschappijen? En hoe kan de vliegtuigmaatschappij klanten het best benaderen? Een flink staaltje data-analyse dus.

In de ochtend hadden de verschillende teams een 'afspraak' met de klant (lees: CQM-collega's), waarbij de studenten een pitch moesten houden over hun aanpak en resultaten. Ze werden hierbij niet alleen op de inhoud beoordeeld, maar ook op hun presentatie-skills; het is pas een goed idee als je het goed kunt overbrengen. Dit ging voor sommige teams beter dan voor andere, maar dat is niet vreemd na een nacht hard werken zonder slaap, en daarbij de spanning van presenteren voor de klant!

Voetbaluitslagen voorspellen

De Nacht van Eindhoven vond net voor het EK 2016 plaats, daarom leek het ons gepast een opdracht over voetbal te maken. Deze opdracht heeft als doel om de verschillende wedstrijden in het kampioenschap zo goed mogelijk te voorspellen. De deelnemende voetbalteams zijn geen landen, maar de 24 studieverenigingen in Nederland die mee kunnen doen aan de Nacht van Eindhoven; baseer je de voorspelling als team enkel op je statistiek-skills of kun je het niet laten om je eigen studievereniging een ronde verder te laten komen?

De teams kregen data met verschillende historische gegevens zoals wedstrijduitslagen, trainers, en omzet van teams. Het was belangrijk om het kwaliteitsverschil tussen teams mee te nemen en daarbij te corrigeren voor het thuisvoordeel dat in de historische wedstrijden (gespeeld bij een van de teams thuis) een rol speelt, maar tijdens het kampioenschap in Eindhoven niet (behalve voor de verenigingen Gewis en Industria van de TU/e). Daarnaast hadden rode kaarten meer effect op de uitslag voor bepaalde teams, deze teams zijn daardoor minder ver gekomen in het toernooi dan je aanvankelijk zou verwachten. Studenten moesten ook kritisch kijken naar de verschillende variabelen, omzet was bijvoorbeeld door ons verzonnen, ze zouden dus geen relatie moeten vinden tussen omzet en wedstrijduitslag.

Pakketbezorging

De laatste opdracht gaat over het bezorgen van pakketten; hoe kun je zo snel mogelijk alle pakketjes bezorgen en tegelijkertijd aan alle voorwaarden van vervoer te doen? Zo hebben laders een beperking in gewicht en omvang in combinatie met de af te leggen afstand, er is maar één wereldwijd distributiepunt, en als een pakket eenmaal geladen is blijft het geladen tot aan de eindbestemming.

Leuke bijkomstigheid bij deze opdracht; oplossingen konden gedurende de hele nacht ingeleverd worden en werden voor iedereen zichtbaar getoond op een scherm. De punten voor deze opdracht werden bepaald op basis van de ranking ten opzichte van andere teams, hierin speelde niet alleen de eindnotering een rol, maar ook de verschillende tussenresultaten. Het snel inleveren van een eerste oplossing, hoe simpel ook, kon je dus veel punten opleveren. De aanpak die CQM zou kiezen is het snel genereren van een startoplossing met een constructieheuristiek en deze oplossing daarna verbeteren via *local search*.

Resultaat

Hoe hebben de teams al deze opdrachten volbracht? Het was een spannende race tussen studieverening A-Eskwadraat (wiskunde) uit Utrecht en studieverening Industria (technische bedrijfskunde) uit Eindhoven. Het eerste team wist zeer snel een goede oplossing in te dienen voor zowel de planning van liften als de pakketbezorging. Industria bleek vooral sterk in het maken van goede voorspellingen voor de andere twee opdrachten, daarnaast wisten zij in het klantgesprek de resultaten ook heel goed over te brengen. In deze tweestrijd kon maar één team de winnaar zijn en hierbij trok A-Eskwadraat aan het langste eind. Gefeliciteerd met deze goede teamprestatie!

Wij als CQM denken alweer na over opdrachten voor de volgende Nacht van Eindhoven. Hierbij putten we uit veel ervaring met verschillende klanten en projecten. De Nacht van Eindhoven is een kans om de consultancy praktijk van kwantitatieve studies te laten zien aan bijna afgestudeerden. Wie weet hebben we dit jaar weer een aantal studenten enthousiast gemaakt voor dit mooie vak.

Wil je meer weten over de Nacht van Eindhoven 2017? Stuur dan een e-mail naar <nachtvaneindhoven@cqm.nl> of neem telefonisch contact op met Pleuni Naus, Wieke Selten of Lianne van Iersel (040 750 23 23).
Voor meer informatie kijk op www.nachtvaneindhoven.nl of www.facebook.com/nachtvaneindhoven

YOUNG STATISTICIANS

This is the first time that we, from the new board, submit a contribution to the STAtOR. We'll shortly introduce ourselves and then we'll discuss the previous event, the company visit to Rabobank.

The board

We're a mix of master and PhD students from various scientific backgrounds, from Biology and Psychology to Mathematics, but we all have at least one thing in common: our passion for Statistics. We've already had a discussion about Bayesians versus Frequentists, and even though there were some slight preferences, it's too soon to tell where we'll end up. We started off by defining our goals for this year, which can be found in the revised mission statement. This has also resulted in a revision of our visual identity, with a new website and updated social media channels. This is just the beginning. We've got many interesting events lined up for you, so make sure to check our social media channels or our website. If you want to contact us, don't hesitate to send us an e-mail. YS-board:

Laura Verkerk
Kees Mulder
Machiel Visser
Jonas Haslbeck (chair)
Erik-Jan van Kesteren
Annette Emerenciana

Fraud and communication

There was a time in which banks were mainly local financial service providers. Nowadays, the advances in IT have changed and improved the way we perform financial transactions and how finances are managed. They have also led to an increase in the amount of financial data that is generated, that can be used to improve the services provided by the banks. For instance, by predicting the credit credibility of a customer, the bank can make a data driven decision on providing a loan.

The Rabobank is a financial services provider that operates internationally and has its roots in the agricultural and food sector. It was founded in 1972, resulting from a merger of the first agricultural cooperative banks founded by farmers.

On May 30th, a group of young statisticians visited the Rabobank division in Utrecht. The company visit was mostly organised by Edwin Thoen, a Leiden University alumnus and enthusiastic young statistician who is now practicing data science in the financial sector. After the master's programme, he went into consultancy, and shortly after he started working at the Rabobank.

In financial transactions, the safety of transactions is the responsibility of both the bank and the customer. The role that banks can play mostly concern cyber security, by investing in fraud detection. As can be imagined, we didn't go into too much detail on this topic, because of its sensitive nature, but Jonatan Samocha was still able to give us a fairly good idea of what it takes for fraud detection to take place autonomously. But what if you have a great, but somewhat complicated idea to improve the service provision of a bank? In that case, communication is key, as Willem van Asperen explained during his talk. After our thirst for knowledge had been satiated, it was time to socialize as several other employees joined for drinks. Thanks again, Edwin, for hosting this fun and informative event!

UPCOMING EVENTS IN 2017

- →February > Company Visit Alliander
- →March > SAS Workshop
- →April > Company Visit Google
- →May > Science Café

www.youngstatisticians.nl
contact@youngstatisticians.nl