STAtOR

periodiek van de VvS+OR jaargang 18, nummer 4, december 2017

STAtOR

Jaargang 18, nummer 4, december 2017

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 4 keer per jaar.

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

Kopij en reacties richten aan
Prof. dr. J.A.S. Gromicho (hoofdredacteur), Faculteit der Economische Wetenschappen en Bedrijfskunde, afdeling Econometrie, Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, 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, oktober en december.

Ontwerp en opmaak
Pharos, Nijmegen

Uitgever

ISSN 1567-3383

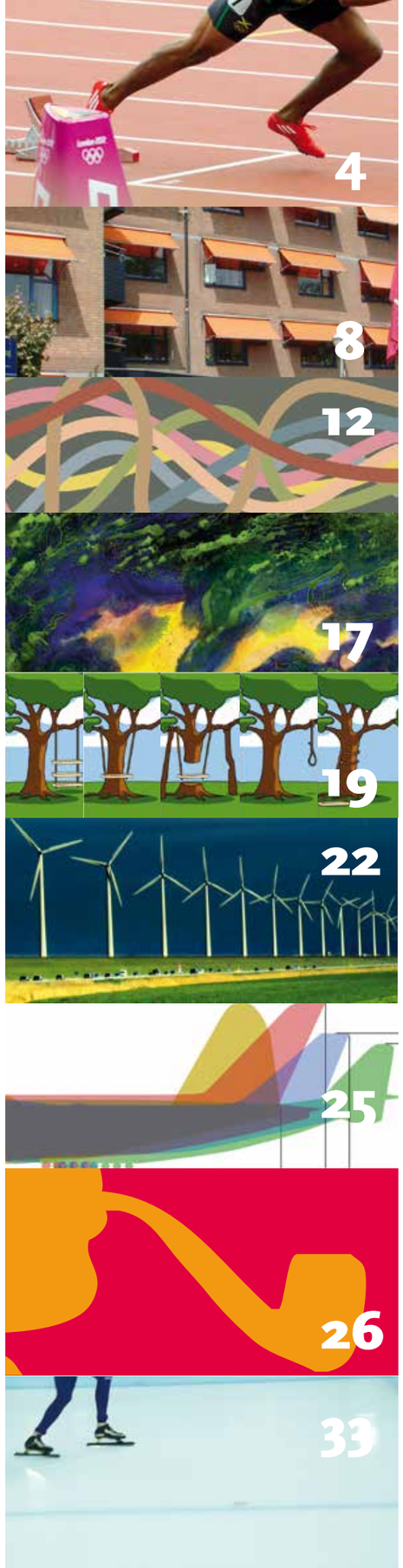
Oud en nieuw

Dit nummer verschijnt rond de kerst, het seizoen voor goede wensen. Laten we daarom beginnen met al onze lezers, namens de gehele redactie van *STAtOR*, een voorspoedig en gezond 2018 toe te wensen.

Het is ook het seizoen om goede voornemens voor het komend jaar te maken, maar eigenlijk kun je zoiets alleen maar doen als je eerst terugblikt op het aflopende jaar. Laten we dat dan ook maar doen. Dit jaar besloten we om 4 nummers uit te brengen. Maar dat maakte het wel spannend voor de redactie: zou het ons lukken 4 gevarieerde nummers samen te stellen? Wij hopen dat u het met ons eens bent dat dit gelukt is. We hebben in totaal 144 pagina's kunnen produceren dankzij de enthousiaste medewerking van maar liefst 39 auteurs en van onze immer productieve vaste columnisten. Aan allen onze hartelijke dank voor de vele en zeer afwisselende bijdragen! We durven dan ook met een gerust hart ons goede voornemen kenbaar te maken om ook in het komende jaar 4 zeer gevarieerde nummers samen te stellen.

Over dit nummer kunnen we melden dat er weer van alles te lezen valt. Er zijn enkele directe toepassingen zoals het binnen/buitenbaan-probleem in de atletiek, de selectie van schaatsers voor grote toernooien (met de unieke kans een paar antieke schaatsen te winnen) en het zodanig organiseren van zorg dat verpleeghuisbewoners niet te lang op hulp bij toiletbezoek hoeven te wachten. Daarnaast een uitgebreid, bijzonder interessant artikel van Mark van der Laan waarin hij een verbinding legt tussen Statistiek en Data Science, maar ook een opsomming van valkuilen bij het toepassen van OR, een uitleg over de wetenschap die de speurders in het tv-programma *Hunted* gebruiken en een artikel waarin geconstateerd wordt dat onzekerheid vóórdelen kan hebben bij het oplossen van optimaliseringsproblemen. Extra aandacht verdient de column van Henk Tijms: u kunt tijdens het kerstdiner of tussen de oliebollen door uw gasten laten denken dat u voorspellende gaven bezit. Zelden was statistiek zó leuk!

De redactie wenst u veel leesplezier!



INHOUD

Yohan Blake op de Olympische Spelen van 2012 in Londen op de 200 meter. Foto: Nick Webb (CC 2.0)

EFFECT VAN DE STARTBAAN BIJ DE 200-METERATLETIEK

Miriam Loois

Bij de 200-meteratletiek lopen atleten die in de binnenste banen starten een krappere bocht dan degenen die in de buitenste banen starten. Het is bekend dat het nadelig is om een scherpere bocht te lopen, maar hoe groot is het verschil? In dit artikel kijken we naar data van Wereldkampioenschappen en Olympische spelen, en maken een vergelijking met meer theoretische natuurkundige modellen. We zullen zien dat het effect op kan lopen tot 0,2 seconden, op een afstand waar het om honderdsten van seconden gaat. En dat dit Daphne Schippers wellicht het laatste zetje gaf dat ze nodig had in Londen om goud te pakken op het WK.

De 200-meteratletiek

Bij baanatletiek lopen atleten op een 400-meterbaan. De baan bestaat uit twee rechte stukken van ongeveer 100 meter, en twee bochten. De atleten lopen naast elkaar in 4 tot 9 banen. De atleet in de binnenste baan, baan 1, loopt een kortere bocht dan die in de buitenste baan. Daarom starten de atleten op de 200-metersprint niet naast elkaar maar mag de buitenste baan verder naar voren starten zodat ze, als ze over de finish komen, allemaal precies 200 meter hebben afgelegd.

Op een groot toernooi worden er eerst series gelopen, waar de sprinters zich kunnen plaatsen voor de halve finales en de finale. In de eerste serie krijgen de atleten willekeurig een baan toegewezen. Atleten die de snelste tijd hebben gelopen in de series, starten in de volgende ronde in de middelste banen (4 en 5). Het is voordelig om een ruimere bocht te lopen, want dit kost minder kracht. Het is dus nadelig om in baan 1 te starten. Bij indooratletiek is de baan 200 meter in plaats van 400 meter, en lopen de atleten dus een volle ronde. Het nadeel van een krappere bocht is hier zo groot, dat deze afstand bij indoortoernooien van het programma is gehaald.

Er zijn al meerdere onderzoeken gedaan naar dit effect. Zo heeft Jonas Mureika een model gemaakt waarin het effect van wind, hoogte en startbaan wordt gemodelleerd. Mike Quinn heeft gekeken naar de invloed van de vorm van de atletiekbaan op het nadeel om in de binnenste baan te starten. Niet alle atletiekbanen hebben namelijk dezelfde vorm. De ene baan kan wat langere rechte stukken hebben en een krappere bocht, bij de andere baan is de bocht ruimer en zijn de rechte stukken wat korter. Ook kan de vorm van de bocht verschillen. In beide artikelen wordt een natuurkundig model gebruikt om het effect van de startbaan te berekenen. De basis is de wet van Newton, F=mxa, kracht = massa × versnelling. In de bocht moet de atleet een kracht uitoefenen om de bocht naar links te maken. Hoe krapper de bocht, hoe nadeliger dit is voor de eindtijd. Quinn vindt een nadelig effect tussen de 0,15 en 0,23 seconden voor baan 1 ten opzichte van baan 8. Mureika geeft geen concreet getal, maar op basis van de parameters die hij in zijn artikel geeft kun je afleiden dat het verschil tussen baan 1 en 9 ongeveer 0,17 seconden is. Als je bedenkt dat de 200 meter vaak een spel van honderdsten van seconden is, is dit een enorm effect!

Van theoretisch model naar de praktijk

In dit artikel kijken we of we het effect van de startbaan ook terugzien in data van Wereldkampioenschappen en Olympische Spelen 1991 tot en met 2015 voor mannen en vrouwen. Wellicht spelen 'in het echt' andere effecten een rol die niet in een natuurkundig model te vatten zijn. Zo wordt er vaak gesproken over het psychologische voordeel van het kunnen starten in de middelste banen, omdat je dan goed zicht heb op je tegenstanders. We nemen in de analyse alleen de eerste rondes mee, omdat de atleten daar een willekeurige baan krijgen toegewezen. Atleten die niet finishen of een tijd boven de 25 seconden (bij de mannen) en 28 (bij de vrouwen) hebben zijn uit de data gehaald. Alleen atleten van wie het persoonlijk record bekend is zijn meegenomen.

We schatten het effect van de startbaan aan de hand van een lineair regressiemodel. We nemen aan dat de eindtijd op de 200 meter afhangt van het persoonlijke record op het moment van de race, de omstandigheden op het toernooi (bijvoorbeeld door de hoogte waarop de baan ligt) en de startbaan.

$$T_{i,j,k} = \alpha \cdot pb_{i,j} + c_j + l_{j,k} + \epsilon_{i,j,k}$$

$T_{i,j,k}$ = tijd op 200 meter van atleet i in baan k op toernooi in jaar j

$pb_{i,j}$ = persoonlijk record van atleet i op het moment van het toernooi in jaar j

c_j = constante bij toernooi in jaar j

$l_{j,k}$ = effect van startbaan k ($k = 1..9$) in jaar j

$\epsilon_{i,j,k}$ =normaal verdeelde error met standaarddeviatie σ_j

We modelleren het effect van de startbaan als factor. We schatten het effect dus apart voor elke startbaan, en veronderstellen geen verband. We schatten het model eerst per jaar. Dit levert voor elk jaar een schatter voor het baaneffect op, met een bijbehorende standaardafwijking. Vervolgens worden de effecten geaggregeerd. Voor het effect van de startbaan leidt dit tot:

$$l_k = \frac{\sum \frac{l_{j,k}}{\sigma l^2_{j,k}}}{\sum \frac{1}{\sigma l^2_{j,k}}} \text{ en } \sigma l_k = \sqrt{\frac{1}{\sum \frac{1}{\sigma l^2_{j,k}}}}.$$

We doen hetzelfde voor het effect van het persoonlijk record. Die blijkt zowel voor de mannen als voor de vrouwen erg significant. Zie tabel 1.

	α	standaardfout
mannen	0,76	0,02
vrouwen	0,74	0,02

Tabel 1. Schatter en standaarddeviatie van het effect van persoonlijk record op eindtijd

	β	standaardfout
mannen	-0,020	0,005
vrouwen	-0,010	0,007

Tabel 2 Schatter en standaarddeviatie van het effect van startbaan op eindtijd bij aanname lineariteit

In de figuren 1 en 2 is het effect van de startbaan voor de mannen en de vrouwen weergegeven. De gestippelde lijn geeft het voorspelde effect van Mureika weer. Het effect is voor baan 1 en 2 significant bij de mannen, bij de vrouwen net niet. De schatter voor baan 1 is een effect van bijna 2 tiende van een seconde achterstand op baan 5. De resultaten van Mureika vallen binnen de betrouwbaarheid, echter, zijn resultaat zit wel aan de onderkant van het betrouwbaarheidsinterval van baan 1 bij zowel de mannen als bij de vrouwen. Wellicht is het nadeel in baan 1 nog wat groter dan uit het natuurkundige model volgt. De uitkomsten bij baan 9 zijn zowel bij de mannen als de vrouwen opvallend. Ten eerste is het betrouwbaarheidsinterval een stuk breder. Dit komt omdat niet op alle toernooien in baan 9 werd gelopen. Soms wordt er juist niet in baan 1 gelopen als er minder atleten zijn dan banen, maar dat komt minder vaak voor. Maar het meest opvallend is dat het heel gunstig lijkt om in baan 9 te starten, terwijl dit niet terug is te zien bij baan 6, 7 en 8. Zonder de resultaten van baan 9 zou je kunnen denken dat het min of meer lineaire effect dat Quinn en Mureika vinden niet klopt, dat het alleen nadelig is om in baan 1 en 2 te starten, en dat het verder niet veel uitmaakt. Met de resultaten van baan 9 erbij lijkt het toch of er min of meer een constant nadelig effect is, elke keer als je een baan naar binnen opschuift.

Het betrouwbaarheidsinterval rond het effect van de startbaan is vrij groot. Dit komt mede omdat we het effect van elke baan apart modelleren. Quinn en Mureika vinden allebei een effect dat bij benadering lineair is. Het nadeel tussen baan 1 en 2, 2 en 3, ..., 8 en 9 is steeds ongeveer even groot. We kunnen dit lineaire verband opleggen aan ons model door de startbaan niet als factor maar als numerieke waarde te modelleren. Het model verandert dan in:

$$T_{i,j,k} = \alpha \cdot pb_{i,j} + c_j + \beta \cdot k + \epsilon_{i,j,k}$$

Het voordeel hiervan is dat het nadelige effect per baan nauwkeuriger te bepalen is. Het nadeel is dat, doordat je het verband oplegt, een eventueel psychologisch voordeel van baan 4 en 5 niet meer zichtbaar is. Zie voor de resultaten tabel 2.

Bij de mannen is het effect 0,02 seconde per baan, met een standaardfout van 0,005, dus significant. Het verschil tussen baan 1 en 9 is dan 0,16 seconde. Bij de vrouwen is het effect 0,01 seconde per baan, met een standaardfout van 0,007, dus niet significant op basis van een 95% betrouwbaarheidsinterval. De uitkomsten van de mannen sluiten goed aan bij Mureika en Quinn, die ook rond de 0,02 seconden per baan uitkomen.

Conclusie

Starten in baan 1 en 2 levert bij de mannen een significant nadeel op. Het min of meer constante nadeel van 0,02 seconde per baan dat uit de natuurkundige modellen volgt is niet in tegenspraak met de data. Als we uitgaan van een lineair verband vinden we een effect van 0,02 seconde per baan bij de mannen, en 0,01 bij de vrouwen. In de data is geen psychologisch voordeel van de middelste banen te zien. Echter, de betrouwbaarheidsintervallen zijn relatief groot. Om met meer zekerheid te kunnen zeggen of het verschil tussen elke baan ongeveer hetzelfde is, of dat het alleen nadelig is om in baan 1 en 2 te starten, zijn meer data nodig.

Een effect van tegen de 0,2 seconden tussen de binnenste en buitenste baan is enorm. Het kan dus zeker gebeuren dat iemand die de series in baan 1 start en zich niet plaatst voor de volgende ronde, zich wel had geplaatst als hij in baan 9 was gestart. Heeft het ook effect op wie de gouden medaille wint aan het einde van het toernooi? De beste atleten hebben meestal een vrij grote voorsprong in de series. In de finale starten de favorieten vaak in de middelste banen. Het effect zal dus minder groot zijn. Afgelopen WK zagen we echter wel weer dat elke honderdste van een seconde telt. Daphne Schippers won goud met een verschil van 0,03 seconde. De tegenstandster die zilver veroverde startte in baan 4, Schippers in baan 6. De paar honderdsten voordeel die dat volgens de modellen geeft hebben haar misschien het goud opgeleverd.

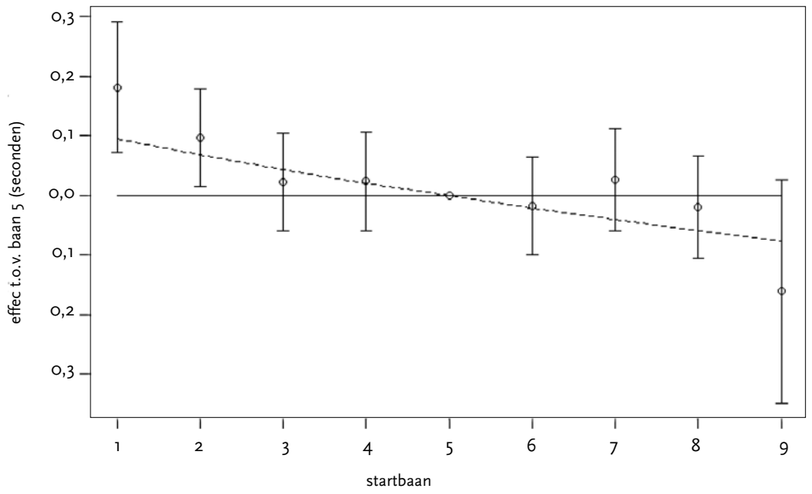
Literatuur

Mureika, J. R. (2003). Modeling wind and altitude effects in the 200 m sprint. *Canadian Journal of Physics*, *81*(7), 895–910.

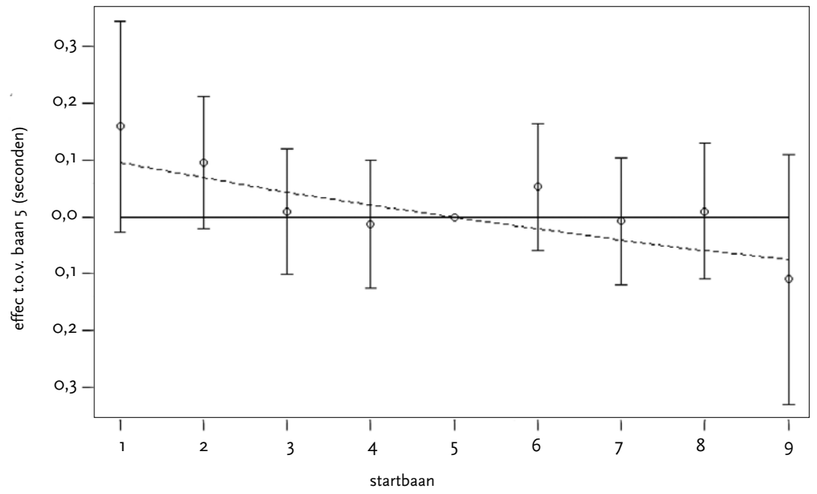
Quinn, M. D. (2009). The effect of track geometry on 200- and 400-m sprint running performance. *Journal of sports sciences*, *27*(1), 19–25.

Data: www.iaaf.org.

Miriam Loois heeft de Master Theoretische Natuurkunde gevolgd aan de Universiteit Utrecht en de Master Actuarial Science and Financial Mathematics aan de Universiteit van Amsterdam. Na 7 jaar in de financiële sector te hebben gewerkt bij Delta Lloyd en PGGM werkt ze nu als docent Toegepaste Wiskunde op de Hogeschool van Amsterdam. Als hobby schrijft ze daarnaast artikelen over statistiek op miriamenstatistiek.wordpress.com.
E-mail: miriamloois@gmail.com



Figuur 1. Effect startbaan mannen t.o.v. baan 5



Figuur 2. Effect startbaan vrouwen t.o.v. baan 5

't Seghe Waert in Zoetermeer

OP TIJD NAAR HET TOILET DOOR DATA-GEDREVEN CAPACITEITSPLANNING

Dennis Moeke, René Bekker & Ger Koole

In Nederland wonen zo'n 130.000 mensen in een verpleeg- of verzorgingshuis verdeeld over meer dan 2.000 locaties. Bewoners van een verpleeg- of verzorgingshuis zijn langdurig –of in veel gevallen zelfs structureel– afhankelijk van zorg bij onder meer het opstaan, naar bed gaan, persoonlijke hygiëne, aan- en uitkleden, eten en het toedienen van medicatie. Om het voor bewoners mogelijk te maken om hun leven te leiden zoals zij dat willen, is het cruciaal dat de benodigde zorg zoveel mogelijk op het door hen gewenste tijdstip wordt geleverd. In de praktijk blijkt dit soms lastig. 'De Patiëntenfederatie Nederland hoort het ook geregeld: ouderen die uren moeten wachten voor ze naar de wc kunnen; mensen die te lang in een natte luier zitten. Woordvoerder Thom Meens vertelt een schrijnend verhaal van een vrouw in een verpleeghuis die door de verpleegsters altijd om acht uur 's avonds op de wc werd geholpen. Als de vrouw om negen uur zou gaan, zou ze de nacht droog doorkomen. Maar dat viel bij het verzorgingshuis niet te regelen' (Vasterman & Dekker, 2016).

Mede onder druk van krimpende budgetten wordt het een steeds grotere uitdaging om de zorg tijdig te kunnen leveren. In de praktijk merk je bijvoorbeeld dat, met name tijdens piekdrukte, zorgmedewerkers (steeds vaker) moeite hebben om aan de voorkeurstijden van de bewoners te kunnen voldoen. Een belangrijke uitdaging is om de personele capaciteit aan te laten sluiten bij de zorgvraag. Ofwel, hoeveel zorgmedewerkers zijn er nodig en wanneer moeten deze worden ingezet? In de huidige situatie gebeurt dat nog te vaak op basis van een onderbuikgevoel. Zo zijn er recentelijk bezettingsnormen geformuleerd die voorschrijven hoeveel medewerkers er minimaal moeten worden ingezet om de kwaliteit te kunnen garanderen (Centraal Planbureau, 2017). Deze normen zijn echter niet gestoeld op een uitvoerige analyse van de zorgvraag. Daarnaast maken zorgaanbieders op dit moment geen gebruik van prestatie-indicatoren die betrekking hebben op wachttijd. In de huidige situatie wordt er, wat betreft 'tijdigheid van de zorg', gebruik gemaakt van subjectieve maatstaven zoals cliënttevredenheid.

In dit artikel laten we zien wat de meerwaarde is van data-gedreven capaciteitsplanning. Dit wordt gedaan aan de hand van een compilatie van een aantal studies die gedaan zijn in het kader van een promotieonderzoek (Moeke, 2016). De volgende generieke onderzoeksvraag staat hierbij centraal: hoe kan wachten op zorg, in een verpleeghuissetting, worden verkort door een betere capaciteitsplanning?

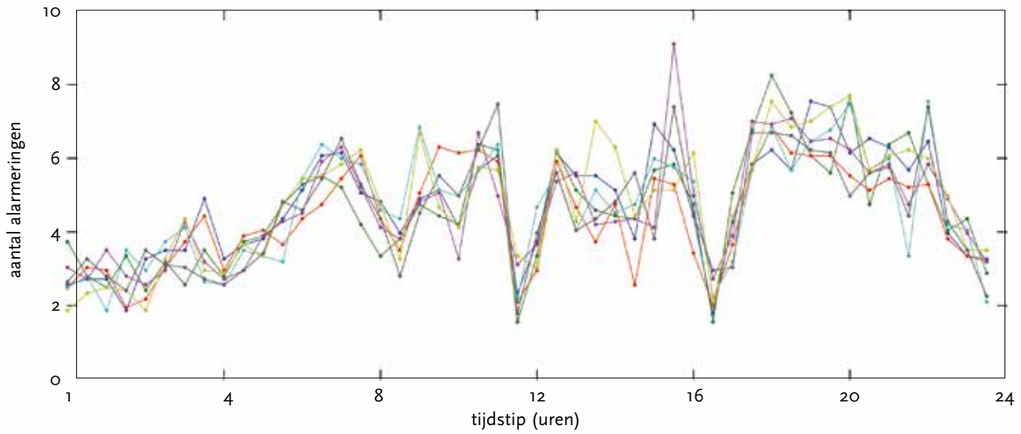
Kenmerken van verpleeghuiszorg

Het bepalen van benodigde hoeveelheid personele capaciteit is een gecompliceerd onderwerp, waarbij zowel de vraag als het aanbod specifieke kenmerken kent. Aan de vraagzijde kunnen er twee typen zorgactiviteiten worden onderscheiden. Voor sommige zorgactiviteiten is het mogelijk om, op basis van de behoeften en voorkeuren van de bewoners, een getailleerde planning te maken. Voorbeelden van dit type zorgactiviteiten zijn 'het toedienen van medicijnen' en 'hulp met opstaan en/of naar bed gaan'. Dit type zorg wordt ook wel planbare zorg genoemd. Daarnaast zijn er zorgactiviteiten die worden uitgevoerd als reactie op 'random' vraag. Een voorbeeld van dit type vraag is 'hulp bij naar het toilet gaan'. Dit type zorg wordt ook wel niet-planbare zorg genoemd. Deze typen zorg karakteriseren zich door een verschillende mate van onzekerheid. Daarnaast zijn er nog fluctuaties en onzekerheden in de duur en het tijdstip van de zorg. Aan de aanbodzijde spelen met name zaken als kwalificatieniveaus, allocatiesystematiek en arbeidsvoorwaarden een belangrijke rol. Tot slot zijn er veranderende opinies met betrekking tot zorgconcepten die van invloed zijn. Denk bijvoorbeeld aan toenemende populariteit van kleinschalige zorg.

Niet-planbare zorg

Uit het onderzoek blijkt dat zowel de vraag naar planbare als niet-planbare zorg fluctueert in de tijd en gedurende het verloop van een dag. Wel blijken de fluctuaties een voorspelbaar karakter te hebben. Om inzicht te krijgen in de karakteristieken van de niet-planbare zorgvraag is gebruik gemaakt van persoonsalarmering data. In noodsituaties moeten kwetsbare bewoners erop kunnen vertrouwen dat er zo snel mogelijk hulp ter plekke is. In de praktijk wordt hiervoor vaak gebruik gemaakt van persoonsalarmering. De bewoner heeft een alarmknop aan een halsketting of naast het bed. In geval van een noodsituatie kan de bewoner hulp inroepen door op de alarmknop te drukken. Onderzoek wijst uit dat in het gebruik van de alarmknop een duidelijke dagpatroon zichtbaar is. Figuur 1 laat het belgedrag van bewoners in een Belgisch verzorgingshuis zien. Voor iedere dag van de week (de 7 verschillende lijntjes) wordt het gemiddeld aantal belletjes per kwartier weergegeven. Daarnaast blijkt het totaal aantal alarmeringen van week tot week redelijk gelijk te zijn. Daarentegen blijkt de duur van niet-planbare zorgactiviteiten nauwelijks af te hangen van het tijdstip.

Op basis van een analyse van de persoonsalarmering-data kan een geschikt wachtrijmodel worden bepaald (zie bv. Van Eeden, Moeke, & Bekker, 2016). Dit model kan ondersteuning bieden bij het bepalen van het aantal zorgmedewerkers dat nodig is om aan de vraag naar niet-planbare zorg te kunnen voldoen. Dergelijke wachtrijmodellen zouden volgens ons gebruikt moeten



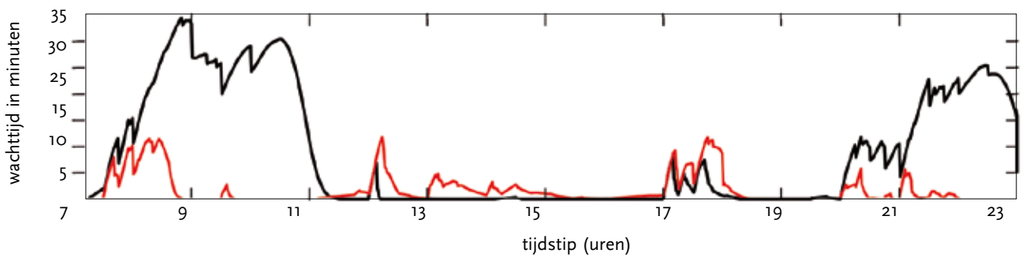
Figuur 1. Gemiddeld aantal belletjes per kwartier gedurende de verschillende dagen van de week

worden om bestaande bezettingsnormen tegen het licht te houden.

Planbare zorg

In tegenstelling tot niet-planbare zorg zou het moment waarop de zorgvraag ontstaat bij planbare zorg (in theorie) volledig voorspelbaar moeten zijn. Echter, de meeste zorginstelling hebben geen inzicht in de voorkeurstijden van de bewoners. De resultaten van een onderzoek naar de planbare zorgvraag van drie Nederlandse verpleeghuisafdelingen laten zien dat de voorkeurstijden zich concentreren gedurende de ochtend en – in mindere mate – de avond (Bekker, Moeke, & Schmidt, 2017). Dit volgt logisch uit het natuurlijke dagritme van de bewoners. Uit de wachttijden in de huidige situatie blijkt dat de personele bezetting onvoldoende 'mee ademt' met de vraag. Zo kunnen de wachttijden in de ochtend oplopen tot wel 40 minuten.

Het wachten op zorg kan aanzienlijk worden gereduceerd door zorgmedewerkers beter te verdelen over de dag. Het voorspelbare vraagpatroon gecombineerd met de mogelijkheid om zorg vooruit te schuiven maakt het gebruik van wachtrijmodellen in deze context lastig. De inzet van personeel is daarom gevangen in een Mixed-Integer Linear Programming (MILP), waarbij de stochastiek wordt meegenomen door het toelaten van verschillende scenario's. De wachttijden in opeenvolgende perioden zijn aan elkaar gerelateerd met behulp van een Lindley-achtige vergelijking. In figuur 2 wordt voor één van de afdelingen de oorspronkelijke (in zwart) met de optimale situatie (in rood) vergeleken. Met name tijdens drukke momenten op de dag is een significante reductie van de wachttijden zichtbaar. Tijdens niet drukke momenten neemt de wachttijd slechts zeer beperkt toe.



Figuur 2. De wachttijd gedurende dag (in minuten) voor één verpleeghuisafdeling

Schaalgrootte

Een algemeen wiskundig principe is dat een grotere schaalomvang over het algemeen leidt tot een betere prestatie bij gelijkblijvende belasting. Vanuit stochastisch perspectief zien we dit reeds bij klassieke wachtrijsystemen, waarbij *square-root staffing* principes een rode draad vormen. Vanuit een scheduling perspectief geeft schaalgrootte flexibiliteit vanwege de variatie in voorkeurstijden tussen bewoners. Dit zien we ook terug in onze studies rond verpleeghuiszorg (Lieder, Moeke, Koole, & Stolletz, 2015; Moeke, Koole, & Verkooijen, 2014; Moeke, Van de Geer, Koole, & Bekker, 2016), waaruit blijkt dat het kleinschalig organiseren van zorg kan leiden tot 'langer wachten'. Hiermee kunnen vraagtekens worden geplaatst bij de huidige politieke ontwikkelingen. Echter, door zorgmedewerkers flexibel over woningen in te zetten kunnen de schaalnadelen worden ingeperkt. Uiteraard kunnen dergelijke organisatievormen kwantitatief worden onderbouwd, maar laten de complexe vraagpatronen en flexibele capaciteitsvormen vaak niet veel andere keus dan het ontwikkelen van een simulatiemodel.

Groter is niet per definitie beter. Zo geeft de fysieke inrichting van de instelling de nodige beperkingen. Bij een toenemende schaalgrootte nemen de afstanden die overbrugd moeten worden tussen de bewoners eveneens toe, waarmee schaaleffecten teniet gedaan kunnen worden.

Conclusies

Uit ons onderzoek blijkt dat data-gedreven capaciteitsplanning een belangrijke bijdrage kan leveren aan het verminderen van het 'wachten op zorg' en het bepalen van de juiste bezettingsnormen. Nauwkeuriger inzicht in de vraag(patronen) en de toepassing van data-gedreven optimaliseringsmethoden maken het mogelijk om zorgvraag en de inzet van zorgmedewerkers beter op elkaar af te stemmen. Tevens kan geconcludeerd worden dat beleidsmakers en managers meer oog zouden moeten hebben voor de nadelen van het kleinschalig organiseren van zorg. Uit de onderzoeken blijkt dat het kleinschalig organiseren van zorg kan leiden tot 'lang wachten', met name tijdens drukke periodes gedurende de dag.

Tot slot zouden we willen benadrukken dat een meer data-gedreven benadering eisen stelt aan de informatiehuishouding van aanbieders van verpleeghuiszorg. Zo is er bijvoorbeeld doorlopend inzicht nodig in de individuele vraag van bewoners (zowel planbare als niet-planbare vraag). Dergelijke data worden op dit moment veelal niet (systematisch) verzameld, maar zijn wel cruciaal voor verdere kwantitatieve analyses.

Literatuur

Vasterman, J., Dekker, M. (2016, November 3). Verpleeghuizen worstelen met toilet-dilemma. *NRC Handelsblad*. Geraadpleegd van: https://www.nrc.nl.

Centraal Planbureau (2017). *Bezettingsnormen voor de verpleeghuiszorg*. Geraadpleegd van: https://www.cpb.nl/sites/default/files/omnidownload/CPB-Notitie-16-feb2017-Bezettingsnormen-voor-de-verpleeghuiszorg.pdf

Moeke, D. (2016). *Towards high-value(d) nursing home care: providing client-centred care in a more efficient manner* (Proefschrift). Amsterdam: Vrije Universiteit Amsterdam.

Van Eeden, K., Moeke, D., & Bekker, R. (2016). Care on demand in nursing homes: a queueing theoretic approach. *Health Care Management Science*, *19*(3), 227–240.

Bekker, R., Moeke, D., & Schmidt, B. (2017). Keeping pace with the ebbs and flows in daily nursing home operations. *Submitted*.

Lieder, A., Moeke, D., Koole, G., & Stolletz, R. (2015). Task scheduling in long-term care facilities: A client-centered approach. *Operations Research for Health Care*, *6*, 11–17.

Moeke, D., Koole, G., & Verkooijen, L. (2014). Scale and skill-mix efficiencies in nursing home staffing: inside the black box. *Health Systems*, *3*(1), 18–28.

Moeke, D., Van de Geer, R., Koole, G., & Bekker, R. (2016). On the performance of small-scale living facilities in nursing homes: A simulation approach. *Operations Research for Health Care*, *11*, 20–34.

Dennis Moeke is senior onderzoeker bij het KennisDC Logistiek van de Hogeschool van Arnhem en Nijmegen (HAN). Hij doet onderzoek op het gebied van Zorglogistiek en coördineert de minor Slim Plannen en Organiseren in de Zorg. Daarnaast doceert hij aan de Arnhem Business School van de HAN. In 2016 promoveerde hij aan de Vrije Universiteit op het proefschrift *Towards high-value(d) nursing home care: providing client-centred care in a more efficient manner*.
E-mail: dennis.moeke@han.nl

René Bekker is universitair docent aan de afdeling Wiskunde van de Vrije Universiteit Amsterdam. Zijn onderzoek richt zich op wachtrijtheorie en op OR toepassingen binnen de zorg.
E-mail: r.bekker@vu.nl

Ger Koole is hoogleraar optimalisatie van bedrijfsprocessen aan de Vrije Universiteit. Hij doet onderzoek naar stochastische optimalisatie en toepassingen daarvan in klantcontact, zorg en revenue management. Daarnaast heeft hij deeltijdaanstellingen bij het call center-planningsbedrijf CCmath en het VU medisch centrum. Qua onderwijs is hij actief betrokken bij de bachelor, master en post-graduate opleidingen business analytics.
E-mail: ger.koole@vu.nl

TARGETED LEARNING: THE LINK FROM STATISTICS TO DATA SCIENCE

Mark van der Laan

The foundations of statistical learning from data are based on the following important concepts: data generating probability distribution, statistical model, target estimand, estimator, and sampling distribution of estimator. Firstly, a statistician acknowledges that data can only be interpreted if one understands the experiment that generated the data. The data are viewed as a realization of a random variable with a certain probability distribution, which is often referred to as the data generating distribution.

Secondly, we need to find out as much as possible about this data generating distribution to restrict the set of possible probability distributions that might have generated the data. This latter set is called the statistical model for the data distribution, which represents our statistical knowledge about the data generating experiment.

Thirdly, the statistician has to specify a so called target parameter mapping from this model to the parameter space (e.g., real line), which specifies for a particular data distribution the feature we aim to learn from the data. That is, one will talk to the scientific collaborator to determine the estimand that represents the best approximation to the scientific question of interest, where the estimand is defined by the target parameter mapping applied to the true data distribution. Determining the statistical model and target estimand thus requires strong interaction with the scientists and data collectors, so that a statistician is naturally part of the scientific team. The statistical estimation problem is now defined.

Fourthly, one has to determine an estimator that maps the data set into an estimate of the target estimand. For example, one might use maximum likelihood estimation to estimate the data probability density and subsequently plug this density estimate in the target parameter mapping to obtain the desired estimate of the target estimand. Finally, we recognize that the estimator is itself a random variable and therefore has a probability distribution, called its sampling distribution. The spread of this sampling distribution represents the uncertainty of the estimator, and an estimator of this sampling distribution can now be used to construct a confidence interval centered at the estimator that will contain with high probability the true estimand.

The erosion of the notion ‘statistical model’

This beautiful foundation of statistical learning appears to have been lost in most of statistical practice. Obviously, the choice of model is crucial since it is supposed to contain the true data distribution. Misspecification of the model, i.e. making assumptions about your data generating distribution that are wrong, will guarantee that the target estimand does not represent the scientific question of interest, and thereby that the resulting confidence interval will most likely not contain the answer to the question of interest.

Typically, we have no knowledge about functional stochastic relations between the different variables we observe. For example, the probability of death will not be a logistic linear function of the baseline measurements on the subject. Nonetheless, almost all statistical methods are based on regression models such as logistic linear regression, linear regression, Cox-proportional hazard regression, and so on, that precisely assume highly simplistic linear relations between outcomes of interest and covariates.

These models are guaranteed to be wrong. An in-depth philosophical and historical perspective on the erosion of ‘truth’ in statistics, and its dire consequences for the field and science in general, is provided in (Starmans, 2011).

The dramatic implications for the practice of statistics: The Wild-West of statistics

Suppose a scientist consults two different statisticians with a data set, a specification of the question of interest, and description of the data generating experiment.

For example, one might observe on a sample of patients baseline covariates, a binary treatment, and an indicator of heart disease a year after treatment was initiated, as part of an observational study aiming to assess the effect of the treatment on heart disease. Most likely, these two statisticians will only care about knowing the format of the data and will quickly decide that this is a logistic regression problem. Each one will specify a particular form for the logistic linear regression. Of course, there is no reason they would select the same model, so most likely both select quite different linear logistic regression models (may be one includes certain interactions, while the other did not).

They will probably report the coefficients of their logistic model fit with p-values and confidence intervals. Presumably, the focus will be on the coefficient in front of treatment, even though that gets quite complicated when the model would include interactions of treatment and covariates, which is a reason for most practitioners to not include interactions with treatment (and thereby force additional misspecification!). Clearly, for large enough sample size, the answers reported will not only be different but even statistically significantly different: the different choice of model will imply a different ‘true’ coefficient in front of treatment, so that the two confidence intervals corresponding with the two statisticians will not overlap each other for large enough sample size. That is, our scientist report random output just depending on what statistician is consulted.

In fact, real practice will typically involve fine tuning the model choice based on interactions with the collaborators till all are satisfied with the reported results and at that point one uses the output in terms of p-values and confidence intervals even though these assumed that the final data and human-adaptively selected model were a priori specified.

As a result, there is no well-defined estimator and

thereby a scientifically sound way to determine the uncertainty in the estimator: in particular, one cannot run a bootstrap that reproduces the estimator applied to a random sample from the actual data set, due to these human interventions.

Another painful by-product of this approach to statistics is that the choice of regression model also implies the choice of target estimand, so that also the estimand is selected without regard to what the actual question of interest is. Therefore, this approach minimizes scientific communication between the statistician and the scientific collaborator. In reality, the scientific collaborator might have taught us that he/she wants to know what the difference in probability of heart disease would have been in a treatment and control arm if one would run a randomized trial. Clearly, these coefficients in the misspecified conditional logistic regression models are not even close in answering this question.

One needs to conclude that statistics has become an art involving human intervention with all its natural biases instead of a science, and it represents an approach that is destined to generate an epidemic of false positives and false claims.

Going back to our foundations of the scientific approach

Let's revisit the above example, but now respecting the foundations of statistical learning. Firstly, we would want to know how the data were generated. We might learn that it is fair to assume that the data on each patient were independently generated and that the data can be represented as a repetition of n independent and identical experiments. We would also want to know how the medical doctor made its treatment decision. We might learn that the medical doctor only takes into account two biomarkers and the age of the patient. Therefore, we might feel comfortable concluding that treatment is conditionally independent of all other baseline measurements, given these three variables. One might conclude that there is no meaningful knowledge about the probability distribution of the covariates and the conditional probability on heart disease as a function of treatment and the covariates. Or may be, we will learn that the outcome is a rare event and that it is perfectly reasonable to assume that the conditional probability on heart disease is always smaller than 0.03,

We can then commit to a statistical model for the probability distribution of the data on a randomly selected patient defined by only assuming the above conditional independence assumption on the conditional distribution of treatment, given the covariates, and (possibly) the bound on the conditional probability of heart disease. Note that this model is not specifying a functional form for the relation between the variables.

Secondly, we would talk to our collaborator, provide causal language using the notion of hypothetical experiments and potential outcomes or nonparametric structural equation models, and determine that our collaborator wants to know the average causal effect defined by the difference in probabilities on heart disease of the two treatment arms in a randomized controlled trial with infinite sample size.

We would then conclude that a best approximation of the answer to this scientific question of interest is the expectation w.r.t. population distribution of covariates of the difference of the conditional probability of heart disease under treatment and control and covariate vector. The latter now represents the estimand of interest, which also defines a mapping from the model to the real line. We can show that this target estimand precisely equals the desired average causal effect if indeed the treatment decision of the doctor was only based on the three variables mentioned and independent noise. We have now defined a realistic statistical model for the data distribution and a target estimand. The statistical estimation problem is defined.

The remaining challenge is now to construct an estimator of the target estimand and estimate its sampling distribution so that we can also construct a confidence interval. Keep in mind though, that after having been so careful in defining the estimation problem, we cannot just return to using parametric regression models to fit this estimand, but, instead, we need to find a way to learn this estimand whatever the true data distribution might be in our statistical model.

Super-learning to learn the unknown functional stochastic relations

The only key stochastic relation we will have to learn from the data is the probability of heart disease as a function of treatment and covariates. The expectation over the covariate distribution can be estimated with an empirical average across the patients in our sample.

Since we do not know the functional form of the probability of heart disease as function of treatment and covariates, we do not want to bet on one particular logistic regression estimator. Instead we build an extensive library of candidate logistic regression estimators. This library can include a large number of linear logistic regression models that vary in the choice of main terms and interactions and possibly functional transformations of various baseline covariates of interest. However, we should also include machine learning algorithms that aim to flexibly learn the functional form. A large variety of such machine learning algorithms have been developed in the machine learning and data science community, such as Lasso regression, random forest, polynomial regression, wavelet regression, neural networks and deep neural networks, to name a few (Hastie et al., 2001).

Each of these algorithms depends on the choice of various tuning parameters, so that one might incude many versions of it, or create a tuned version of the algorithm that internally searches the best tuning parameter. Recently, we proposed a new machine learning algorithm named Highly Adaptive Lasso (HAL) (van der Laan, 2015; Benkeser & van der Laan, 2016). We would add this algorithm as well since it has been shown to consistently estimate the true functional relation at a rate faster than $n^{-1/4}$ as a function of sample size n, whenever the true target function has finite variation norm. The latter assumption can be included in our statistical model without any concern for misspecification, since it is unlikely that the true functional relations have infinite variation as a function of the different covariates.

Instead of betting on one of these algorithms, we carry out a competition by training each of these algorithms on 9/10 of the data, and evaluating its performance in predicting heart disease status of the patient on the 1/10 left-out patients, and apply this competition to a number of random splits of the data in training and validation sample. We evaluate the performance of each algorithm by its average prediction error across the different sample splits, which is called the cross-validated risk of the estimator (e.g. log likelihood risk or squared-error risk). We now select the best algorithm and rerun that algorithm on the complete data set.

This defines a new machine learning algorithm, a so called ensemble algorithm, that combines all these algorithms in the library into a new more powerful and adaptive algorithm. We have named this ensemble algorithm a super-learner, due to its theoretical property that it is asymptotiallly as good as the oracle selector that chooses the best algorithm when given an infinite validation data set, even when the number of candidate algorithm grows polynomial in sample size (van der Laan and Dudoit, 2003; van der Laan et al., 2007; van der Laan and Rose, 2011; van der Vaart et al., 2006).

Without much additional effort, one can also compute the cross-validated risk of each weighted convex combination of the candidate algorithms and determine the optimal weighted combination, so that one ends up selecting some weighted combination of the different algorithms in the library.

Targeted learning to tune the fit of the data distribution towards the target estimand, and obtain an approximate normal sampling distribution

The super-learning fit of the conditional probability of heart disease is optimized to fit this whole function, and as a result spreads its approximation error across all covariate profile configurations of the population of patients. However, we only care about fitting a specific function/summary measure of this conditional probability function, namely the target estimand. Therefore, the spread of error is not optimized for the target estimand. Another more statistical way of saying this is that the super-learner optimally trades off bias and variance w.r.t. the whole function, while we need to optimally trade off bias and variance of the corresponding estimator of the target estimand (just a real number). The plug-in estimator of the target estimand based on the super-learner will have variance $1/n$, while its bias will be of the same order as the bias of the super-learner, and thus be larger than $n^{-1/2}$ (e.g. around $n^{-1/4}$). As a consequence, the plug-in estimator based on the super-learner will not even be asymptotically normally distributed.

This is resolved with the so called targeted maximum likelihood estimator (TMLE) which creates a fluctuation of the super-learner fit where the amount of fluctuation is a coefficient we will name ϵ. At zero fluctuation $\epsilon = 0$, this fluctuation returns the super-learner fit, and if we move ϵ away from zero it will fluctuate the super-learner fit in such a way that it maximizes the square change of the target estimand per increase in log-likelihood, at least locally around $\epsilon = 0$. Such a fluctuation strategy is solved by a mathematical optimization problem, and it

is referred to as the least favorable parametric model in the efficiency theory literature (Bickel et al., 1997). We now fit ϵ with standard maximum likelihood estimation for this one-dimensional parametric model, resulting in an updated fit of the super-learner. This fitting step is using all the data to purely fit the target estimand and, as a consequence, it removes bias and optimizes the mean squared error for the target estimand. One now plugs this targeted super-learner fit into the estimand representation to obtain the desired TMLE of the target estimand.

This TMLE can now be shown to be behave as an empirical mean of an efficient influence curve transformation of the unit data structure, and is thus approximately normally distributed, and has minimal asymptotic variance (van der Laan, 2015). As a consequence, a variance estimator (e.g., a bootstrap or Wald-type confidence interval) provides an approximate 95% confidence interval for the target estimand.

Above we demonstrated a TMLE for the sake of estimation of the ATE. TMLE is a general method for construction of a two stage estimator of any target estimand for any statistical model, where the first stage uses super-learning to obtain an initial estimator of the data distribution, while the second stage consists of carrying out a TMLE-update step based on a least favorable parametric submodel (van der Laan and Rose, 2011).

Data science and targeted learning

Data science is a flourishing field with growing amount of resources. The machine learning world has played a fundamental role in the definition and growth of data science. Typically, the proposed approaches lack statistical formulation, the recognition of an underlying experiment, the careful definition of a target estimand answering the question of interest and construction of a corresponding theoretically grounded estimator, and above all it lacks assessment of statistical uncertainty. It often appears that statistical thinking is under-appreciated and is not receiving the place it deserves. However, simultaneously, there is a clear and growing recognition that one should not only be concerned with prediction but also assessment of (e.g., causal) effects of interventions on outcomes of interest with proper statistical inference. Due to the massive amounts of data one is often confronted with, standard parametric regression models are not even applicable, so that this community realizes that one will have to use data adaptive estimation and machine learning to make progress.

Therefore, targeted learning, a pure statistical approach that incorporates the state of the art in machine learning targeting the estimand of interest, while preserving formal statistical inference, brings the statistical foundations and the enormous advances that have been made in our field (e.g., causal inference, empirical process theory, efficiency theory, missing data, censored data, biased sampling, etc) to the forefront in data science. In this manner, statistics deserves its place and can flourish as well as data science as a whole, while moving science forward in the process.

References

Benkeser, D, & van der Laan, M.J. (2016). *The highly adaptive lasso estimator. Proceedings of the IEEE Conference on Data Science and Advanced Analytics.* To appear.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y., & Wellner, J. (1997). *Efficient and adaptive estimation for semiparametric models*. Berlin Heidelberg New York: Springer.

Hastie, T.J., Tibshirani, R.J., & Friedman, J.H. (2001). *The elements of statistical learning: data mining, inference, and prediction*. Berlin Heidelberg New York: Springer.

Starmans. R.J.C.M. (2011). Models, inference and truth: Probabilistic reasoning in the information era. In M.J. van der Laan & S. Rose (Eds.). *Targeted Learning: Causal Inference for Observational and Experimental Studies*, pp. 1–20. New York: Springer.

Van der Laan, M.J. (2015). *A generally efficient targeted minimum loss-based estimator.* Technical Report 300, UC Berkeley, 2015. http://biostats.bepress.com/ucbbiostat/paper343. To appear in *International Journal of Biostatistics*.

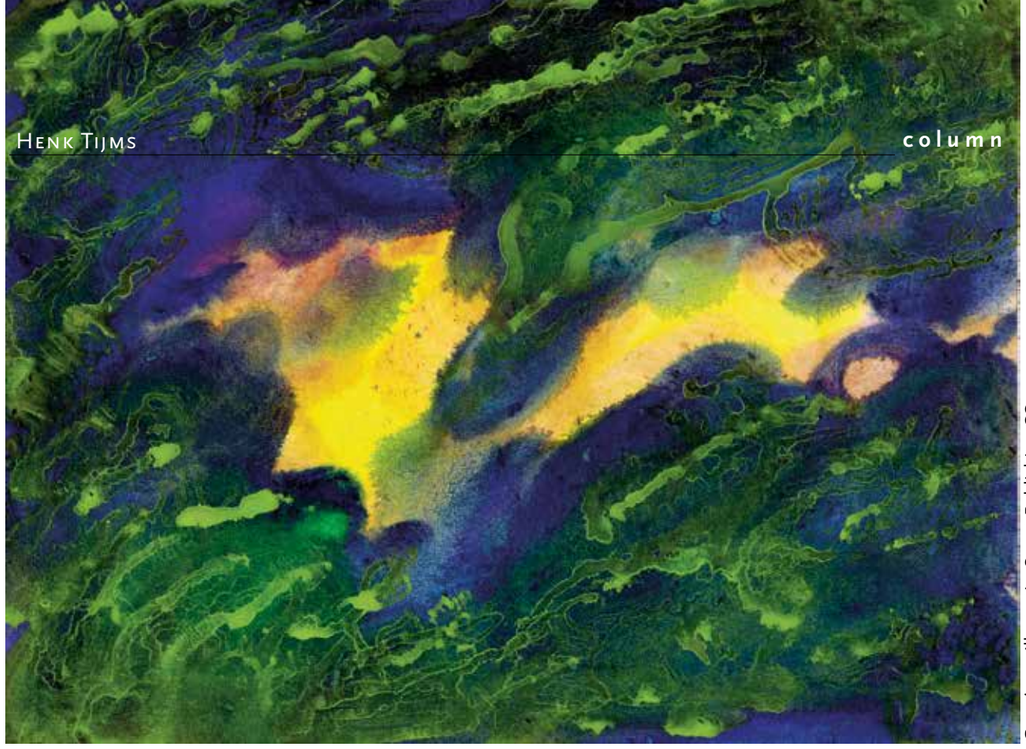
Van der Laan, M.J., & Dudoit, S. (2003). *Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples.* Technical Report 130, Division of Biostatistics. Berkeley: University of California.

Van der Laan, M.J., & Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data.* Berlin Heidelberg New York: Springer.

Van der Laan, M.J., Polley, E.C., & Hubbard, A.E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology,* 6(1).

Van der Vaart, A.W., Dudoit, S., & van der Laan, M.J. (2006). Oracle inequalities for multi-fold cross-validation. *Statistics and Decisions,* 24(3), 351–371.

Mark van der Laan, Ph.D., is a Professor of Biostatistics and Statistics at UC Berkeley. His research interests include statistical methods in genomics (i.e., computational biology), survival analysis, censored data, targeted maximum likelihood estimation in semiparametric models, causal inference, data adaptive loss-based super learning, and multiple testing. He is the recipient of the VvS-OR Van Dantzig Award 2005.
E-mail: laan@berkeley.edu



Genesis 1-2. Illustratie: Sweet Publishing CC 3.0

Kruskal telling en de bijbel

In deze column wil ik aandacht besteden aan een fascinerende kaarttruc. Deze kaarttruc staat bekend als de Kruskal telling en heeft zijn oorsprong in het werk van de Amerikaanse fysicus Martin Kruskal uit de jaren 1970. Het werk van Kruskal werd gepopulariseerd door Martin Gardner, de beroemde Amerikaanse schrijver van populair-wetenschappelijke artikelen over wiskundige raadsels en spellen. De kaarttruc gaat als volgt. Een toeschouwer wordt door een goochelaar uitgenodigd een standaard pak van 52 kaarten grondig te schudden. De kaarten worden daarna achterelkaar op tafel gelegd met de beeldzijde naar boven. Elk van de kaarten representeert een getalwaarde, stel de azen en de plaatjeskaarten (heer, vrouw, boer) elk 1 punt en de andere kaarten het getal dat op de kaart staat. De toeschouwer wordt gevraagd een getal tussen 1 en 10 in gedachten te nemen. De goochelaar vertelt dan dat de kaart in de positie van dit geheime getal de eerste 'sleutelkaart' van de toeschouwer wordt en legt de toeschouwer uit dat de waarde van de sleutelkaart aangeeft hoeveel kaarten verder in de rij de nieuwe sleutelkaart ligt. Dus stel dat het geheime getal van de toeschouwer 7 is, dan wordt de 7de kaart in de rij de eerste sleutelkaart van de toeschouwer en als deze kaart een vier is, dan wordt de 11de kaart in de rij de nieuwe sleutelkaart van de toeschouwer, en als die 11de kaart een boer is, dan wordt de 12de kaart in de rij de nieuwe sleutelkaart van de toeschouwer, etc. De toeschouwer telt in stilte totdat hij een sleutelkaart bereikt waarvan de getalwaarde zodanig is dat tellen tot een volgende kaart niet mogelijk is omdat er niet meer voldoende kaarten zijn. Deze sleutelkaart is de eindkaart van de toeschouwer. Bij het bereiken van deze eindkaart voorspelt de goochelaar wat voor kaart de eindkaart is. Vrijwel altijd zal de goochelaar het goed hebben! Wat is de truc? Deze is verbluffend simpel. De goochelaar kiest ook een geheim getal tussen 1 en 10 en telt dan op dezelfde wijze mee als de toeschouwer. Hoewel de beginkaarten van de toeschouwer en de goochelaar niet hetzelfde hoeven te zijn, komt met grote waarschijnlijkheid een moment waarop beiden op eenzelfde kaart in de rij landen en vanaf dat moment volgen ze hetzelfde pad. Zoals Sherlock Holmes, vooruitlopend op de Kruskal

telling, al stelde in het verhaal 'The disappearance of Lady Frances Carfax' uit de bundel *His Last Bow* geschreven door Arthur Conan Doyle: 'When you follow two separate chains of thought, Watson, you will find some point of intersection which should approximate to the truth.' Nemen we aan dat de toeschouwer en de goochelaar onafhankelijk van elkaar elk blindelings een getal tussen 1 en 10 kiezen, dan is de kans dat de goochelaar de eindkaart van de toeschouwer correct voorspelt ongeveer 93,1%. Zou de goochelaar twee pakken van 52 kaarten gebruiken, dan stijgt deze succeskans tot ongeveer 99,5%. Geloof je het niet? Probeer het dan zelf maar eens uit met een gewillig slachtoffer of schrijf een simulatieprogramma! Een exacte formule voor de succeskans van de goochelaar in de Kruskal telling is niet bekend, maar wel de benaderingsformule

$$1-\left(1-\frac{1}{a^2}\right)^N,$$

waarbij N het aantal kaarten is en a de gemiddelde waarde per kaart. In het beschouwde geval waarbij niet alleen de azen maar ook de heer, vrouw en boer voor 1 tellen, is a gelijk aan 1/13(1+1+1+1+2+3+ ... +10)= 58/13. Simulatie laat zien dat de benaderingsformule bijzonder goede resultaten geeft. De benaderingsformule oogt simpel, maar is minder simpel te beargumenteren. Markov-keten theorie ligt ten grondslag aan de kanstheoretische analyse van het kaartspel. Zonder Markov-ketens te gebruiken, kan een andere uitstekende benaderingsformule worden afgeleid. Dit gaat met het volgende heuristische argument. De kans is 9/10 dat de goochelaar een andere begingetal kiest dan de toeschouwer. Voor zowel de goochelaar als de toeschouwer is de verwachtingswaarde van de positie van de eerste sleutelkaart in de rij van N kaarten gelijk aan $1/10(1+2+ \ldots +10) = 5{,}5$. Gemiddeld genomen heeft de eerste sleutelkaart de waarde a voor zowel de goochelaar als de toeschouwer. Dit is ook de gemiddelde lengte van de stappen bij het doorlopen van de rij van neergelegde kaarten. Dit betekent dat binnen de rij van N kaarten de dichtheid van de sleutelkaarten van de toeschouwer gezien kan worden als 1/a. De goochelaar zal gemiddeld genomen $(N-5{,}5)/a$ sleutelkaarten gebruiken na de eerste sleutelkaart. Elk van deze sleutelkaarten heeft een kans van $1/a$ om een sleutelkaart van de toeschouwer te treffen op dezelfde positie binnen de rij van neergelegde kaarten, ofwel de kans is $1-1/a$ dat dit niet gebeurt. Een en ander maakt plausibel dat de kans dat noch de eerste sleutelkaart van de goochelaar noch elke volgende sleutelkaart van de goochelaar een sleutelkaart van de toeschouwer ontmoet in eenzelfde positie binnen de rij van neergelegde kaarten benaderd kan worden door $9/10(1-1/a)^{(N-5,\,5)/a}$. Dus een alternatieve benaderingsformule voor de succeskans van de goochelaar is

$$1-\frac{9}{10}\left(1-\frac{1}{a}\right)^{(N-5{,}5)/a}.$$

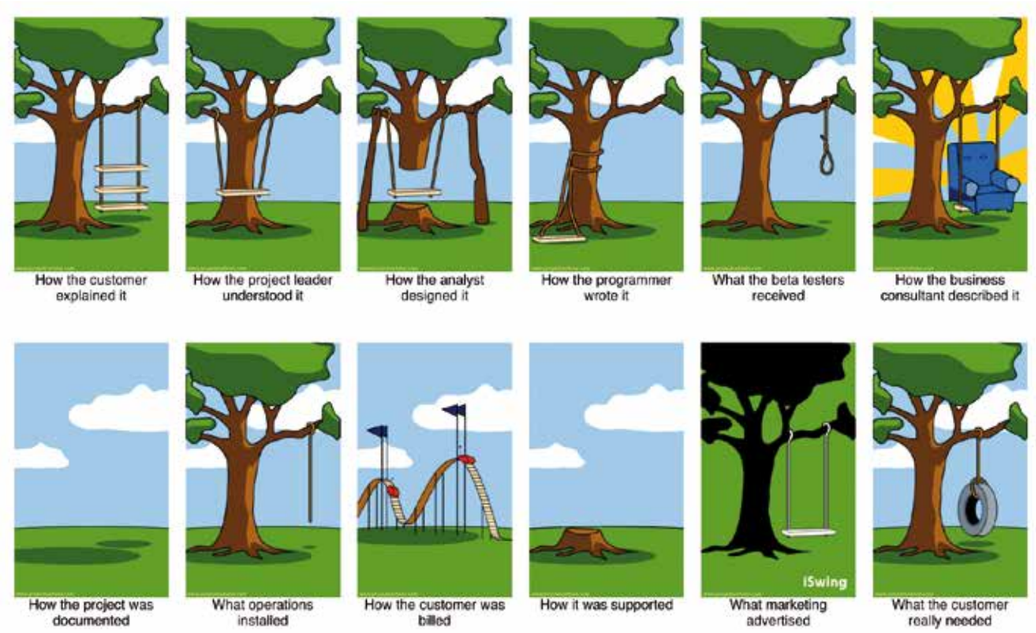
Voor één pak kaarten ($N = 52$) respectievelijk twee pakken kaarten ($N = 104$) met de waarden 1 tot en met 10 voor de kaarten, geeft deze benaderingsformule de succeskansen 93,6% en 99,7%, vrijwel dezelfde benaderingswaarden als bij de eerst gegeven benaderingsformule. In Kruskals oorspronkelijke versie van het kaartspel telden de heer, vrouw en boer niet voor 1 maar voor 5. In dat geval is $a = 70/13$ en is de succeskans van de goochelaar ongeveer 85% bij 52 kaarten en ongeveer 98% bij 104 kaarten.

De Kruskal telling kan ook worden toegepast op tekstpassages. Een illustratie, die sommigen als bijna mystiek zullen ervaren, is de volgende. Neem uit het bijbelboek Genesis de eerste drie verzen:

1. In the beginning God created the heaven and the earth.
2. And the earth was without form, and void; and darkness was upon the face of the deep. And the Spirit of God moved upon the face of the waters.
3. And God said, Let there be light: and there was light.

Kies een woord in het eerste vers. Tel hoeveel letters dit woord heeft. Als het woord L letters heeft, ga dan L woorden verder. Ga door dit procedé te volgen totdat je in het derde vers belandt. Stop dan. Ongeacht welk woord je als beginwoord neemt, je zult altijd met hetzelfde woord eindigen. Niet alleen voor deze tekstpassage werkt de truc, maar ook voor vele andere tekstpassages. Des te langer de tekst, des te groter de kans. Probeer het maar eens met een tekstpassage uit je favoriete boek en verbaas je vrienden!

Henk Tijms is emeritus hoogleraar Operations Research aan de Vrije Universiteit en auteur van diverse leerboeken over operations research en kansrekening. Zijn meest recente boek is *Kansrekening van alledag, een wereld vol verrassingen*.
E-mail: h.c.tijms@xs4all.nl



BELANGRIJKE VALKUILEN BIJ HET TOEPASSEN VAN OR

Adriaan Tas

Het is een bekend feit dat projecten dikwijls anders uitpakken dan dat ze gepland waren. Onduidelijkheden, gebrekkige omschrijvingen en verschillende percepties. De cartoon hierboven illustreert dat op ludieke wijze.

In dit artikel ga ik in op de valkuilen die ik ben tegengekomen bij het uitvoeren van projecten voor klanten van ORTEC. Sinds 1990 ben ik betrokken bij een heel aantal kleine en grote projecten, waarbij altijd een wiskundig modelelement in het project zat. Projecten duren soms een paar weken, soms 2 jaar. Qua onderwerpen loopt dit van routeringsproblemen voor mengvoeders, via productie-modellen voor gasvelden en netwerk-design studies voor logistieke dienstverleners naar *inventory* managementmodellen voor airlines. Niet alle projecten zijn uitgevoerd zoals bij het begin voorzien was, en eigenlijk zien we een beperkt aantal redenen waarom projecten anders lopen dan gepland.

Hoge ambities

De allerbelangrijkste reden hiervoor is dat de *ambitie* van de organisatie voor het toepassen van *analytics* veel ho-

ger is dan het startpunt waarop de organisatie zich bevindt. Hierdoor worden stappen overgeslagen en besluiten we gedurende het project het ambitieniveau terug te schroeven. Voorbeelden hiervan zijn het niet beschikbaar zijn van benodigde informatie en onbetrouwbare informatie als deze wel beschikbaar is. Hiermee samenhangend zien we als reden ook de *gebruikers* van het ontwikkelde model. Deze zijn vaak niet gewend te werken met applicaties die beslissingen ondersteunen. De ontwikkelde rekenmodule voor het oplossen van het integrale probleem wordt niet gebruikt, maar er is veel tevredenheid over de ondersteuning in termen van meldingen van schendingen van regels en KPI-informatie. Veelal wordt de applicatie gebruikt voor handmatige planningen, waarbij als voordeel wordt gegeven dat er minder fouten worden gemaakt bij het maken van een planning. Op zich een duidelijke verbetering voor de klant, maar feitelijk niet het doel van veel van de gestarte projecten.

Enthousiast management

Naast deze omgevingsoorzaken hebben we ook te maken met het *klimaat* bij onze klanten. Vaak is iemand in het management enthousiast over het toepassen van OR en is daarmee een belangrijke sponsor van het project. Maar deze trekker wisselt op een gegeven moment van baan, krijgt promotie of krijgt andere verantwoordelijkheden. Als dat midden in het project gebeurt, kan dat tot gevolg hebben dat het project opeens veel minder aandacht en middelen krijgt vanuit de klantorganisatie. En dan ligt een teleurstellend resultaat op de loer. Van belang is dan ook om te zorgen voor een goede inbedding van het project in de organisatie en het betrekken van voldoende mensen bij de klant om dit soort personele wisselingen op te vangen. Projectmanagementmethodes haken hierop in, bijvoorbeeld als er *agile* wordt ontwikkeld. Daarbij wordt al snel in het traject bruikbare software opgeleverd en de functionaliteit wordt in korte sprints uitgebouwd tot het eindresultaat. Hiermee wordt de betrokkenheid van eindgebruikers vergroot. Daarnaast wordt het voor de sponsors van het project ook zichtbaar dat er resultaten worden geboekt.

Dynamische omgeving

Nog weer een andere valkuil is de *timing* waarop de inputdata beschikbaar komen. Traditioneel bouwen we modellen die voor een input dataset, de opgegeven doelstelling en alle beperkingen een oplossing opleveren. We bepalen bijvoorbeeld alle distributieroutes om de producten morgen bij de klanten af te leveren. Of we bepalen de dagplanning van servicemonteurs die controlebezoeken moeten afleggen op een aantal locaties. Deze modellen maken een goed of zelfs optimaal plan, gegeven de vaste input. Echter, in de praktijk zien we dat deze input helemaal niet vast is. Er worden opdrachten geannuleerd, hoeveelheden worden aangepast en spoedorders worden nog na de officiele deadline geplaatst. In deze omgeving is de waarde van ons algoritme niet zo groot, want alle wijzigingen zijn moeilijk in te passen in een planning die vanwege de doelstelling weinig slack bevat. In een dynamische omgeving moeten we heel goed kijken naar de werkwijze en aanpak van de planner. Hoe gaat hij om met de onzekerheid? Wanneer maakt hij zijn basisplan? Heeft hij kennis van toekomstige orders die hij al gebruikt bij zijn plan? (‘Klant x bestelt altijd voor woensdag, dus ga er maar vanuit dat die order nog komt’; ‘Klant y vindt het ook goed als we 20% meer brengen, dus daarmee benut ik die ritcapaciteit wel volledig’).

Moeten we in deze situatie met de opdrachtgever niet méér kijken naar het verkleinen van de onzekerheid, het terugdringen van de spoedorders, in plaats van met een algoritme een zich telkens wijzigend probleem heel goed oplossen? Of moeten we bij de oplossing juist een robuste oplossing nastreven? Of moet juist de planner ondersteund worden in het huidige proces van plannen met voorstellen waar een spoedorder ingepland kan worden met weinig wijzigingen op de bestaande planning? We hebben in het verleden dit aspect soms niet in de gaten gehad en bij de implementatie moeten constateren dat we een prima algoritme hadden gemaakt, maar dat het probleem van de opdrachtgever eigenlijk een ander probleem was.

Haalbaarheid

Omdat dit artikel zich concentreert op de valkuilen, kan de indruk ontstaan dat het toepassen van OR in de praktijk vaak mislukt of in elk geval slechts ten dele lukt. Dat is echter alleen zo als we OR willen toepassen als aan de voorwaarden voor een succesvolle implementatie niet is voldaan. Dit artikel moet dan ook vooral worden gezien als een aanmoediging om goed te kijken hoe het te ontwikkelen model in de organisatie gebruikt zal en kan worden. Daarvan moet een helder beeld zijn, zowel bij opdrachtgever als leverancier. Een goede methode hiervoor is het model van Davenport, waarbij op een aantal onderdelen gekeken wordt op welk *level* de opdrachtgever zich bevindt. Aan de hand van deze uitgangssituatie kan dan gekeken worden of de doelstellingen haalbaar zijn, dat wil zeggen zich qua ambitieniveau iets boven het huidige niveau bevinden. Als de te nemen stap groter is dan 1 niveau op een van de success factoren, moet dit als een groot risico voor het uit te voeren project worden beschouwd. Als zowel de opdrachtgever als leverancier de verschillende *use cases* als realistisch beschouwen, kan de focus worden gericht op de op te lossen puzzel. De grootste valkuil is dus dat we de geschetste puzzel zo interessant vinden, dat we vergeten naar de omgeving van de puzzel te kijken, de puzzel oplossen en er dan achterkomen dat we de verkeerde puzzel hebben opgelost.

SUCCESS FACTOR		MOVING TO:			
	Stage 1: Analytically Impaired	**Stage 2: Localized Analytics**	**Stage 3: Analytical Aspirations**	**Stage 4: Analytical Companies**	**Stage 5: Analytical Competitors**
Data	Inconsistent, poor quality, poorly organized	Data useable, but in functional or process silos	Organization beginning to create centralized data repository	Integrated, accurate, common data in central warehouse	Relentless search for new data and metrics
Enterprise	n/a	Islands of data, technology, and expertise	Early stages of an enterprise-wide approach	Key data, technology and analysts are centralized or networked	All key analytical resources centrally managed
Leadership	No awareness or interest	Only at the function or process level	Leaders beginning to recognize importance of analytics	Leadership support for analytical competence	Strong leadership passion for analytical competition
Targets	n/a	Multiple disconnected targets that may not be strategically important	Analytical efforts coalescing behind a small set of targets	Analytical activity centered on a few key domains	Analytics support the firm’s distinctive capability and strategy
Analysts	Few skills, and these attached to specific functions	Isolated pockets of analysts with no communication	Influx of analysts in key target areas	Highly capable analysts in central or networked organizations	World-class professional analysts and attention to analytical amateurs

Bron: Thomas H. Davenport, Jeanne G. Harris & Robert Morison, Analytics at Work: Smarter Decisions, Better Results. Boston: Harvard Business School Publishing Corperation, 2010

Adriaan Tas studeerde bedrijfseconometrie aan de Erasmus Universiteit en is sinds 1989 werkzaam bij ORTEC. Vanaf 2015 is hij Director Operations Research bij de business unit Consulting.
E-mail: adriaan.tas@ortec.com

Voordelen van onzekerheid in stochastische optimaliserings-problemen met geheeltallige variabelen

Swifterbant, 2011. Foto: Janneman (CC 3.0)

Ward Romeijnders

In de praktijk moeten veel beslissingen al genomen worden terwijl belangrijke informatie nog onbekend is. In de zorg- en energiesector bijvoorbeeld. Daar worden operaties toegewezen aan operatiekamers voordat bekend is hoe lang deze operaties zullen duren en worden kolencentrales aan- en uitgeschakeld voordat bekend is hoeveel wind- en zonne-energie er de rest van de dag zal worden opgewekt. De beslissingen in deze twee uiteenlopende toepassingen kunnen beiden worden gemodelleerd met behulp van stochastische gemengd-geheeltallige lineaire optimaliseringsmodellen (SMILPs). Deze modellen worden gekenmerkt door stochastische parameters en geheeltallige beslissingsvariabelen. In de voorbeelden uit de zorg- en energiesector bijvoorbeeld kunnen de duur van de operaties worden gemodelleerd als stochastische parameters en het aan- en uitschakelen van de kolencentrales met geheeltallige (of binaire) variabelen. Deze modellen zijn zeer generiek en hebben ook toepassingen in bijvoorbeeld de logistieke en financiële sector. Tegelijkertijd is het extreem moeilijk om de optimale oplossing exact te berekenen.

Stochastische optimaliseringsproblemen

In de twee-stadia variant van het model worden twee typen beslissingen onderscheiden: eerste-stadium beslissingen, die gemaakt worden terwijl de realisaties van de stochastische parameters nog onbekend zijn, en tweede-stadium beslissingen, die gemaakt worden als alle informatie beschikbaar is. Het doel is om de verwachte kosten te minimaliseren, waarbij we aannemen dat de kansverdeling daarvan bekend is, bijvoorbeeld op basis van historische data. Het model dat we dan beschouwen is

$$\min_{x \in X} \{ c^T x + Q(x) \}$$

waarbij Q gedefinieerd is als de verwachtingswaarde van het tweede-stadium LP-probleem

$$Q(x) = \mathbb{E}_\omega \left[\min_{y \in Y} \{ q(\omega)^T y : Wy = T(\omega)x - h(\omega) \} \right].$$

We gaan hier niet dieper in op de specificatie van het model, maar zijn geïnteresseerd in wiskundige eigenschappen van de functie Q die gebruikt kunnen worden om het probleem efficiënt op te lossen. Aangezien $Q(x)$ geïnterpreteerd kan worden als de verwachte toekomstige kosten van een eerste-stadium beslissing x, noemen we Q de verwachte-waarde-functie.

Complexiteit en convexiteit

Als alle tweede-stadium beslissingen continu zijn, dan is de verwachte-waarde-functie Q convex. Zulke functies laten zich makkelijk minimaliseren. Helaas gaat deze convexiteit bijna altijd verloren wanneer (een deel van) de tweede-stadium beslissingsvariabelen geheeltallig zijn. Dit verklaart voor een belangrijk deel de algoritmische complexiteit van zulke SMILPs. Ook verklaart het waarom deze modellen zoveel moeilijker op te lossen zijn dan dezelfde modellen met continue in plaats van geheeltallige beslissingsvariabelen. Traditionele oplossingstechnieken voor SMILPs combineren technieken uit de deterministische gemengd-geheeltallige optimalisering met die uit de stochastische continue optimalisering, maar in het algemeen kunnen deze traditionele methoden geen probleeminstanties van realistische grootte oplossen; een aantal speciale toepassingen daargelaten.

Oplossingsfilosofie

Wij zullen hier een fundamenteel andere oplossingsmethodologie beschrijven, waarin we in tegenstelling tot traditionele methodes wél gebruik kunnen maken van convexe optimaliseringstechnieken. In plaats van met de exacte verwachte-waarde-functie Q te werken, gebruiken wij namelijk een convexe benadering $\hat{Q}$. Op deze manier krijgen we niet de exacte optimale oplossing x^*, maar wel een oplossing $\hat{x}$. die zeer goed is als het verschil tussen Q en $\hat{Q}$. klein is. Om de kwaliteit van de benaderingsoplossing $\hat{x}$ te waarborgen leiden we bovengrenzen af op

$$\|Q - \hat{Q}\|_\infty := \sup_x |Q(x) - \hat{Q}(x)|.$$

In het vervolg van dit artikel beschrijven we deze convexe benaderingen $\hat{Q}$ en bovengrenzen op $\|Q - \hat{Q}\|_\infty$. Eerst zullen we motiveren waarom deze aanpak werkt door de verwachte-waarde-functie Q in een numeriek voorbeeld te beschouwen.

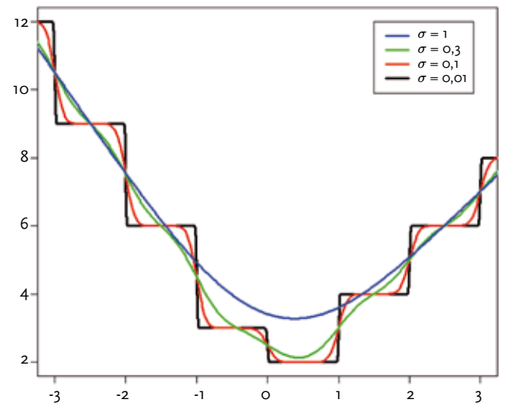
Numeriek voorbeeld

We nemen een eenvoudig voorbeeld van een verwachte-waarde-functie Q. Namelijk,

$$Q(x) = \mathbb{E}_\omega \left[3 \lceil \omega - x \rceil^+ + 2 \lceil \omega - x \rceil^- \right], x \in \mathbb{R}$$

Hier is x een één-dimensionale eerste-stadium beslissingsvariabele en ω een één-dimensionale stochast. Het doel is om x te selecteren (voordat de realisatie van ω bekend is) zodat $x = \omega$. Afwijkingen van dit doel worden in het tweede stadium naar boven afgerond en dan bestraft. Dus voor $x < \omega$ hebben we $\lceil \omega - x \rceil^+ = \max \{ \lceil \omega - x \rceil, 0 \}$en voor $x > \omega$ hebben we $\lceil \omega - x \rceil^- = \max \{0, -\lfloor \omega - x \rfloor\} = -\lfloor \omega - x \rfloor = \lceil x - \omega \rceil$. In dit voorbeeld gaat de afwijking $x < \omega$ met hogere kosten gepaard dan $x > \omega$.

Figuur 1 schetst de verwachte-waarde-functie Q voor normaal verdeelde stochasten ω met verwachting $\mu = 0$ en standaarddeviatie $\sigma = 0{,}01$, 0,1, 0,3 en 1. Voor $\sigma = 0{,}01$ lijkt Q op een trapfunctie vanwege de afrondingen in het model. Deze afrondingen zijn het gevolg van geheeltallige tweede-stadium beslissingsvariabelen en zorgen ervoor dat Q niet convex is. Wat opvalt in figuur 1 is dat naarmate de standaarddeviatie toeneemt de verwachte-waarde-functie Q 'convexer' lijkt. We zien hier het gladstrijkende effect van de verwachte-waarde operatie in Q. Dit is een onverwacht voordeel van de onzekerheid in het model, waarbij het effect sterker is voor stochasten met hogere variabiliteit.



Figuur 1. Verwachte-waarde-functie Q uit het numerieke voorbeeld voor ω normaal verdeeld met $\mu = 0$ en $\sigma = 0{,}01$, 0,1, 0,3 en 1

Enkelvoudig geheeltallige recourse modellen

De verwachte-waarde-functie Q uit figuur 1 is een voorbeeld van een verwachte-waarde-functie van een zogenaamd enkelvoudig geheeltallig *recourse* model (een speciaal soort SMIP). Deze modellen zijn uitvoerig bestudeerd door Klein Haneveld et al. (2006). Zij hebben een klasse van stochasten ω geïdentificeerd waarvoor Q daadwerkelijk convex is. Voor alle overige stochasten construeren ze een convexe benadering $\hat{Q}$ van Q door op een slimme manier een benaderde stochast $\hat{\omega}$ uit bovengenoemde klasse te selecteren. Ook leiden ze een bovengrens af op $\|Q - \hat{Q}\|_\infty$. In overeenstemming met figuur 1 is deze bovengrens kleiner als de variabiliteit van de stochast ω groter is. De bovengrens hangt echter niet van de standaarddeviatie σ van ω af maar van de totale variatie $|\Delta| f$ van de kansdichtheidsfunctie f van ω. Deze totale variatie $|\Delta| f$ is gedefinieerd voor één-dimensionale functies als

$$\sup_P \sum_{i=1}^{N} |f(x_i + 1) - f(x_i)|$$

waarbij $P := \{x_1,\ldots,x_{N+1}\}$ een partitie in $\mathbb{R}$ is met $x_1 < \ldots < x_{N+1}$. Romeijnders et al. (2016b) bewijzen dat voor $Q(x) = \mathbb{E}_\omega [\lceil \omega - x \rceil^+]$ en de bijbehorende convexe benadering $\hat{Q}$ van Klein Haneveld et al. (2006):

$$\|Q - \hat{Q}\|_\infty \leq h(|\Delta| f) := \begin{cases} |\Delta| f/8, & |\Delta| f \leq 4, \\ 1 - 2/|\Delta| f, & |\Delta| f \geq 4. \end{cases}$$

Voor de normale verdeling geldt inderdaad dat de totale variatie $|\Delta| f$ kleiner wordt naarmate de standaarddeviatie σ toeneemt en dat dus de bovengrens op $\|Q - \hat{Q}\|_\infty$ afneemt (cf. figuur 1). In het algemeen geldt: hoe 'breder en platter' de kansdichtheidsfunctie f van ω, des te beter de benadering. Bijvoorbeeld voor unimodale f is de benadering dus beter als de top van f lager ligt.

Nieuwe ontwikkelingen

De oplossingsfilosofie is niet alleen toegepast op enkelvoudige geheeltallige recourse modellen, maar ook op SMILPs waarbij de matrix van coefficiënten totaal unimodulair is (Romeijnders et al., 2015) en op generieke twee-stadia SMILPs (Romeijnders et al., 2016a). In het laatste geval maken we gebruik van zogenaamde Gomory relaxaties voor deterministische MILP problemen om te laten zien dat op bepaalde convexe deelverzamelingen van het domein de onderliggende waardefunctie van Q gelijk is aan de som van een lineaire en periodieke functie. We buiten deze periodiciteit uit om een convexe benadering $\hat{Q}$ te construeren en om een bovengrens op $\|Q - \hat{Q}\|_\infty$ af te leiden. Net als voor enkelvoudig geheeltallige recourse modellen convergeert deze bovengrens naar nul als de totale variaties van de kansdichtheidsfuncties van de stochastische parameters in het model naar nul convergeren. Dus ook voor generieke modellen geldt: als de variabiliteit van de stochasten in het probleem toeneemt wordt de verwachte-waarde-functie Q 'convexer'. Dit blijkt ook uit numerieke experimenten op routerings- en planningsproblemen onder onzekerheid in Romeijnders et al. (2017).

Discussie

Uit deze resultaten kunnen we concluderen dat de variabiliteit van de stochasten in het model een onverwacht positief effect heeft. Naarmate deze variabiliteit toeneemt wordt namelijk het negatieve effect van de geheeltallige variabelen tenietgedaan en gedraagt het model zich als een convex model. Traditionele methoden maken echter geen gebruik van dit convexificerende effect. Een richting voor toekomstig onderzoek is daarom om dit effect in de traditionele methoden mee te nemen zodat ze realistische instanties van stochastische gemengd-geheeltallige optimaliseringsproblemen, zoals in de zorg- en energiesector, efficiënter kunnen oplossen.

Literatuur

Klein Haneveld, W.K., Stougie, L., & Van der Vlerk, M.H. (2006). Simple integer recourse models: convexity and convex approximations. *Mathematical Programming, 108,* 435–473.

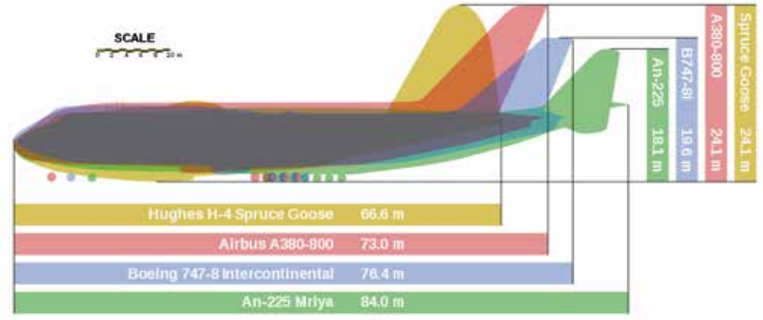
Romeijnders, W., Morton, D.P, & Van der Vlerk, M.H. (2017). Assessing the quality of convex approximations for two-stage totally unimodular integer recourse models. *INFORMS Journal on Computing, 29*, 211–231.

Romeijnders, W., Schultz, R., Van der Vlerk, M.H., & Klein Haneveld, W.K. (2016a). A convex approximation for two-stage mixed-integer recourse models with a uniform error bound. *SIAM Journal on Optimization, 26*, 426–447.

Romeijnders, Van der Vlerk, M.H., & Klein Haneveld, W.K. (2015). Convex approximations of totally unimodular integer recourse models: A uniform error bound. *SIAM Journal on Optimization, 25*, 130–158.

Romeijnders, Van der Vlerk, M.H., & Klein Haneveld, W.K. (2016b). Total variation bounds on the expectation of periodic functions with applications to recourse approximations. *Mathematical Programming, 157*, 3–46.

Ward Romeijnders is universitair docent bij de vakgroep Operations Research aan de Rijksuniversiteit Groningen. Voor zijn proefschrift ontving hij de Willem R. van Zwet Award. In 2017 ontving hij een NWO VENI-beurs; hij gaat de consequenties van het convexificerende effect voor traditionele en nieuwe oplossingsmethoden onderzoeken.
E-mail: w.romeijnders@rug.nl



Theorie én data

Soms kom je iets tegen waarvan je denkt 'Hoe verzinnen ze het?'. Onlangs had Arjen Lubach zijn uitzending gewijd aan de geluidsproblemen bij de groei van Schiphol en de uitbreiding van Lelystad Airport. Een deskundige werd getoond die in volle ernst uitlegde dat men met modellen werkte, want als men van metingen uit zou gaan zou de te verwachten geluidoverlast significant hoger zijn.* Hopelijk is het fragment dat werd getoond deel van een uitgebreider verhaal met meer nuance, ik kan me namelijk niet voorstellen dat een exacte wetenschapper een dergelijke uitspraak zonder meer doet.

Kennelijk klopten de data niet bij de modellen, dat is niet zo vreemd, dat komt vaker voor. Als de data correct zijn zit er dan niets anders op dan het model aan te passen. In feite is deze procedure de motor achter de ontwikkelingen in bijvoorbeeld de natuurkunde. Met de mechanica van Newton konden we prima uit de voeten, onder andere planeetbanen konden met grote precisie worden berekend. Maar naarmate de metingen nauwkeuriger werden kwamen er afwijkingen met de theorie tevoorschijn. Zo week de precessie van het perihelion van Mercurius miniem af van dat wat door Newtons mechanica werd voorspeld. Het is een van de eerste zaken die Einstein berekende in zijn algemene relativiteitstheorie: deze bleek wél exact de waargenomen waarde van de precessie te voorspellen.

Ook in ons vakgebied hebben we vaak te maken met modellen, we proberen bijvoorbeeld de variantie in een afhankelijke variabele op te splitsen in de varianties veroorzaakt door enkele verklarende variabelen en hun interacties. Het restant wordt dan traditioneel aangeduid met *error*. Ik heb altijd een aversie gehad tegen die term error in een model. Liever spreek ik van 'niet door het model verklaarde variantie'. Dat geeft volgens mij beter weer dat een model niet meer is dan slechts een model en dat de werkelijkheid groter kan zijn. De term error hoort wat mij betreft gereserveerd te blijven voor zaken als meetfouten etc. en zelfs die kun je als deel van het model zien.

Theorie en data kun je niet los van elkaar zien is mijn vaste overtuiging. Data gebruik je om theoretische modellen te toetsen en daarna eventueel aan te passen en uit de theorie wordt duidelijk welke data je zou moeten meten. Onlangs vond ik bij toeval (tja, je bent statisticus of niet) een prachtig onder woorden gebrachte uitspraak met dezelfde strekking.

Sinds mijn gedwongen vervroegde pensionering in 1994 werk ik een dag per week als vrijwilliger bij het Permanent Office van het International Statistical Institute (ISI) in Den Haag. Ik doe daar allerlei klussen waar men met de normale bemensing niet aan toe komt. Zo kreeg ik een verzoek vanuit Singapore om een kopie van een lezing die was gepresenteerd op het tweejaarlijkse ISI congres in Parijs in 1961. Het 19 pagina's tellende artikel 'Multiple Classifications in Social Accounting' was van Sir Richard Stone (1913-1991) die in 1984 de Nobelprijs voor Economie kreeg. Nu ben ik altijd nieuwsgierig, volgens mijn schoonmoeder lees ik álles, zelfs 'de puntzak van De Gruyter'. Dus ook dit artikel heb ik grotendeels gelezen en daarin kwam ik de volgende eindconclusie tegen waar ik het van harte mee eens ben.

> '[...] All of this goes to show that it is very difficult to separate facts and theories. Those who insist on being guided by facts alone may find that they get very little guidance and those who insist on the distinctions of pure theory may confine themselves to a world which most of their fellow men consider uninteresting.'

* Voor *Zondag met Lubach* zie: https://www.youtube.com/watch?v=ftD9SZYviqg.

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

SLIMME OPSPORING MET QUIN

Bob Van der Vecht, Freek van Wermeskerken & Selmar Smit

> *'I have a certain friend – his name is Mr. Quin, and he can best be described in terms of catalysis. His presence is a sign that things are going to happen, because when he is there strange revelations come to light, discoveries are made.' Agatha Christie – The Mysterious Mr Quin*

Voortvluchtigen doen alle moeite om uit handen van opsporingsteams te blijven. Echter, ze laten toch vaak kleine sporen achter; een pintransactie, een telefoontje met een bekende. Dit is de context van het AvroTros-programma *Hunted*. Hierin moeten deelnemers zich gedurende drie weken schuilhouden en uit handen zien te blijven van een team dat werkt volgens de principes van opsporingsonderzoeken bij de politie.

Tijdens opsporingsonderzoek is de hoeveelheid informatie vaak zowel enorm beperkt als overweldigend. Van de zaak zelf is maar weinig bekend. In deze tijd van sociale media en mobiele telefonie zijn er tegelijkertijd enorm veel data beschikbaar over de persoon in kwestie. Denk aan sociale netwerken, alle Facebook-vrienden. Mogelijk geeft één van hen hem onderdak. Maar hoe haal je uit al die mogelijkheden datgene wat er echt speelt?

Dit jaar krijgen de rechercheurs in de serie *Hunted* hulp van zo'n Mr. Quin. QUIN (QUestion and INvestigate, tevens geïnspireerd op het personage Mr. Quin van Agatha Christie) is namelijk een methode en ondersteuningstool voor analisten en rechercheurs die beoogt richting te geven binnen opsporingsonderzoeken (Smit, 2016). Het uitgangspunt is: een zaak lijkt altijd op iets dat al een keer eerder is gebeurd. QUIN maakt gebruik van historische zaken, en gaat na welke het meest lijken op de huidige. In dit geval gaat het om de gedragingen van kandidaten uit eerdere series uit binnen- en buitenland. Die informatie wordt omgezet naar de huidige zaak: welk gedragspatroon past er bij het huidige bewijs? Welke scenario lijkt het meest passend? Deze informatie kan de rechercheur gebruiken om het onderzoek te sturen.

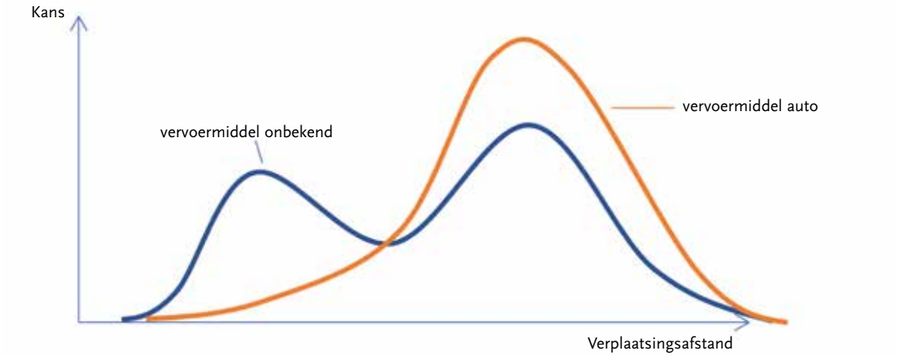
Scenario's

QUIN is een voorbeeld van hoe kwantitatieve analyses op historische data kunnen bijdragen aan opsporing en inlichtingen. Aan de basis van de aanpak van QUIN ligt een bekende criminologische aanpak genaamd *crime scripting* (Cornish, 1994), waarin crimineel gedrag wordt vastgelegd in scenario's. Een script of scenario deelt een criminele activiteit op in scenes. Binnen die scenes zijn er actoren die handelingen verrichten, waarbij ze gebruik maken van bepaalde middelen, et cetera. Het gedrag van daders, criminele organisaties of deelnemers aan *Hunted* kan in een dergelijk format beschreven worden. Vervolgens kunnen de scripts geanalyseerd worden om gedragspatronen te herkennen.

Elke scene is opgebouwd uit componenten voor subject, tijdsperiode, locatie(s), middelen en (menselijke) ondersteuning. Vervolgens zijn aan al deze componenten weer attributen gehangen voor onderscheidende eigenschappen. Daarmee wordt bijvoorbeeld van een tijdsperiode het dagdeel vastgelegd. Bij locatie zijn de karakteristieken vastgelegd in termen van afstand tot een stad, openbaar vervoer, water, bos. Bij support (een persoon die de voortvluchtige helpt) gaat het om geslacht en leeftijd, maar ook om de relatie die hij/zij heeft met de voortvluchtige. Gaat het hier om een bekende? Een bekende van een bekende? Of een volslagen vreemdeling?

Belangrijk om op te merken is dat bij het maken van de database alleen de afgeleide attributen opgeslagen worden. Hierdoor wordt geabstraheerd van de specifieke individuele casus en bevat de database geen informatie over exacte locaties of personen, maar alleen de karakteristieken en eigenschappen hiervan. Voor locaties kun je denken aan de afstand tot verschillende soorten point-of-interest, zoals restaurants, openbaar vervoer voorzieningen en stadscentra. Voor personen kun je denken aan geslacht, leeftijd, en relatie tot het subject. Op deze wijze kun je ook voor een nieuwe casus bepalen wat de meest waarschijnlijke invulling zal zijn, want als de eerste verplaatsing altijd wordt gedaan door de echtgenoot van de voortvluchtige, dan zal dat bij een nieuwe casus waarschijnlijk ook wel zo zijn. Zie ook figuur 1.

Figuur 1 Screenshot van QUIN toont scenario's op volgorde van waarschijnlijkheid; de onderste rij toont de geselecteerde scene en geeft inzicht in de voorspelde attribuutwaarden van de componenten



Figuur 2. Kernel Density Estimation (KDE) van attribuut verplaatsingsafstand in een verplaatsingsscene zonder bewijs over vervoersmiddel (blauw) en gegeven dat het vervoersmiddel een auto is (oranje)

Redeneren

Dit is de kern van het redeneerproces van QUIN, namelijk het uitrekenen hoeveel een huidige casus lijkt op historische zaken uit de database. De uitkomst van deze vergelijking is een getal, *scenario similarity* genaamd.

Om de *scenario similarity* van een huidige casus met een historische casus te bepalen wordt eerst gekeken of het bewijs kan passen in het corresponderende *scenario template*. Vervolgens wordt voor alle attributen in het bewijs de *attribute similarity* bepaald, waarvan het product wordt genomen.

De hoogte van attribute similarity geeft aan hoeveel attribuut-waarden op elkaar lijken, bijvoorbeeld de leeftijden 38 (uit het bewijs) en 51 (uit historische data). De similarity schaal van een attribuut kan handmatig gedefinieerd worden, bijvoorbeeld hoeveel lijkt een broer op een neef? Voor continue waarden kan dit ook automatisch worden bepaald uit de historische data. Hierbij wordt er eerst de Silverman's Bandwidth (Silverman, 1986) bepaald op basis van alle waarden van de attributen. Hoewel deze niet de meest nauwkeurige Kernel Density Estimation (KDE) oplevert, is deze wel zeer efficiënt te berekenen bij grote hoeveelheden data. De Silverman's Bandwidth schrijft ook het gebruik van een normaalverdeling als basisfunctie voor de KDE voor. En met behulp van deze basisfunctie, de bandwidth, en alle attribuutwaarden kan dan een KDE gemaakt worden. De similarity is vervolgens uit te rekenen door de gemaakte verdeling te evalueren op de waarde van het attribuut uit het bewijs.

Als het verkregen bewijs op meerdere scenario-templates past kunnen we met de scenario-similarity een uitspraak doen. Hierbij wordt gecorrigeerd voor het aantal attributen, omdat niet elk scenariotemplate evenveel attributen bevat.

Wanneer bepaald is wat de similarity is tussen de huidige casus en elke historische zaak, kan het systeem uitspraken doen over de verwachte waarden van onbekende informatie voor de huidige casus. Hiertoe wordt een *posterior* KDE opgesteld, door de verdeling van alle mogelijke attribuutwaarden te bepalen uit de historische zaken gewogen met de scenario similarity. Een voorbeeld hiervan is te zien in figuur 2.

Analyseren

Een analist heeft de mogelijkheid om bewijs (scenes, attribuutwaarden en (tijds)relaties) in te voeren in QUIN. Vervolgens heeft hij de mogelijkheid om de resultaten van QUIN-redeneringen te inspecteren. Hieronder bespreken we enkele visualisaties. Daarna bespreken we een voorbeeld van een werkwijze.

De *scenario view* (figuur 1) geeft een overzicht van mogelijke scenario's en verwachte invullingen daarvan. Het meest waarschijnlijke scenario template wordt bovenaan getoond en toont de scenes met een korte beschrijving ervan. Wanneer een scene wordt geïnspecteerd is er een overzicht te zien van de componenten en attributen in die scene. Wanneer de waarde van een attribuut rechtstreeks uit het bewijs volgt is de waarde van het bewijs te zien. Als dat niet het geval is, is de verdeling van attribuutwaarden (de posterior KDE) te zien die uitgerekend is met behulp van de scenario similarities. Er zijn verschillende visualisaties voor de verwachte

Figuur 3. Een heatmap is de visualisatie die gebruikt wordt om de waarschijnlijkheid van locaties weer te geven

waarde van een attribuut, afhankelijk van het type attribuut. Voor een continue attribuut wordt een distributie plot gemaakt en voor een locatie wordt een kaart met een extra gekleurde laag getoond (figuur 3).

Door middel van de TRY-functie kan een analist hypotheses toetsen om *soft evidence* of aannames in te voeren, en die eenvoudig weer ongedaan maken. Hiermee kan ook de meerwaarde van (nog) onbekende informatie inzichtelijk gemaakt worden en kunnen ook hypotheses van bijvoorbeeld een profiler gebruikt worden om de voorspellingen aan te scherpen.

Deze visualisaties worden door de analist gebruikt als ondersteuning van de analyses en conclusies. Het is geen data-analyse die bewijs levert, maar het versterkt de argumentatie van een analist. Binnen het opsporingsteam van *Hunted* (figuur 4) is QUIN ingezet om te voorspellen waar de voorvluchtige zich schuil houdt: slaapt hij in het wild in de bossen, of logeert hij bij een vriend? Is hij bij een waarneming op doorreis of maakte hij een ommetje? Waar naartoe is hij op weg? En hoe lang gaat hij daar blijven? De analyses met QUIN bepalen mede hoe de opsporingscapaciteit wordt ingezet.

Voor het opsporingsteam van *Hunted* zat een belangrijke kracht van QUIN in het samenbrengen van gedragspatronen uit historische zaken met geografische analyse. Dit type analyse leent zich ook voor de toepassing van voortvluchtigen. Een dergelijke combinatie kan echter ook gemaakt worden tussen gedragspatronen en sociale relaties van betrokkenen, wat wellicht bij georganiseerde misdaad meer van belang is.

QUIN is een ondersteuningstool, en geen vervanger van een analist of rechercheur. De grootste meerwaarde

Figuur 4. QUIN in gebruik tijdens het AvroTros programma *Hunted*

valt te behalen als de uitkomsten goed kunnen worden geduid: Waar komt deze voorspelling vandaan (welke casussen hebben de grootste similarity)? En klopt dat? Ook het maken van hypotheses, het bepalen van *soft evidence* en het beoordelen van de betrouwbaarheid van informatie is iets dat juist de mens heel goed kan, en dat moet een dergelijke tool faciliteren, niet proberen te vervangen.

Conclusies

De gepresenteerde methode van QUIN laat zien hoe historisch gedrag in scenario's vastgelegd kan worden, waardoor je nieuwe situaties kunt vergelijken met historische zaken. Het gaat om duiden (wat is er aan de hand?) en voorspellingen doen over wat je nog niet weet (hypotheses stellen). Op deze manier kan QUIN ingezet worden als instrument om te bepalen waarop je je opsporingsmiddelen het beste kunt richten. De aanpak via similarity en Kernel Density Estimators is generieker toepasbaar dan in alleen opsporingsonderzoeken of scenario-analyse.

Literatuur

Smit, S., Van der Vecht, B., Van Wermeskerken, F., & Streefkerk, J.W. (2016). QUIN: Providing Integrated Analysis Support to Crime Investigators. In *Proceedings of Intelligence and Security Informatics Conference (EISIC)*, pp. 120–123. IEEE.

Cornish, D.B. (1994). The procedural analysis of offending and its relevance for situational prevention. *Crime prevention studies*, *3*, 151–196.

De Kock, P.A.M.G. (2014). *Anticipating criminal behaviour: Using the narrative in crime-related data*. Tilburg University

Silverman, B.W. (1986). *Density estimation for statistics and data analysis*. CRC press.

Bob van der Vecht studeerde kunstmatige intelligentie aan de Rijksuniversiteit Groningen en is hierin in 2009 gepromoveerd aan de Universiteit Utrecht. Hij werkt sindsdien als onderzoeker bij TNO op het gebied van operations research.
E-mail: bob.vandervecht@tno.nl

Freek van Wermeskerken studeerde wiskunde aan de Vrije Universiteit en behaalde zijn master in 2015 op het gebied van modelleren en optimalisatie. Na zijn afstuderen werkt hij bij TNO aan soortgelijke vraagstukken.
E-mail: freek.vanwermeskerken@tno.nl

Selmar Smit is aan de Vrije Universiteit gepromoveerd op het onderwerp machine learning, en sindsdien werkzaam als onderzoeker en adviseur op het gebied van artificial intelligence bij TNO.
E-mail: selmar.smit@tno.nl

OPROEP

UITMUNTENDE MASTER'S OF PH.D. THESIS?
Oproep om kandidaten te nomineren voor de Jan Hemelrijk en de Willem R. van Zwet Award 2017

Supervisors (begeleiders) worden opgeroepen om een uitmuntende afstudeerscriptie (Master) of dissertatie (Ph.D.) te nomineren voor de Jan Hemelrijk dan wel Willem R. van Zwet Award 2017. Genomineerd kunnen worden personen die tussen september 2015 en september 2017 zijn afgestudeerd respectievelijk gepromoveerd en die nog niet eerder zijn genomineerd.

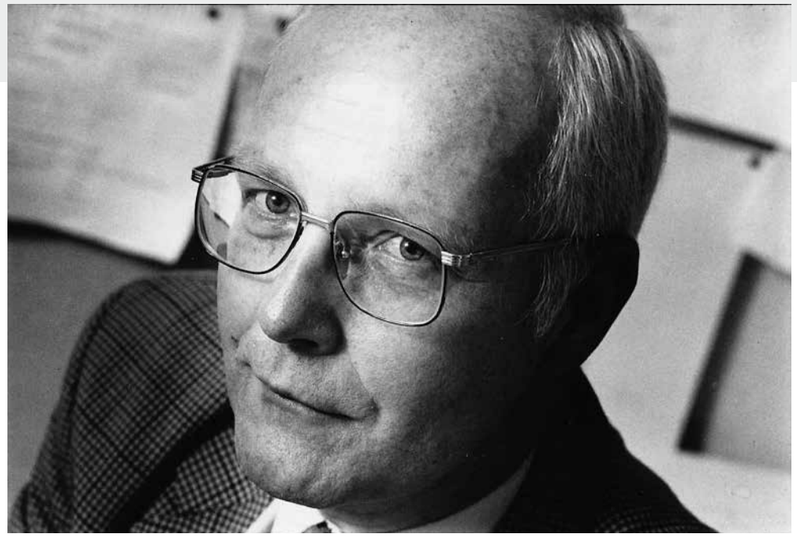
De indiening van een nominatie dient vergezeld te gaan van een aanbevelingsbrief van de supervisor van de genomineerde. De precieze procedure voor beide prijzen, alsmede de reglementen en het nominatieformulier zijn te downloaden op de website van de VvS+OR http://www.vvs-or.nl.

Al sinds 1989 looft de VvS+OR een scriptieprijs uit ter bekroning van een uitzonderlijke afstudeerprestatie aan een Nederlandse instelling voor wetenschappelijk onderwijs of hoger beroepsonderwijs. 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 tijdens de Annual Meeting van de VvS+OR, op woensdag 28 maart 2018.

De nominatie dient uiterlijk 21 januari 2018 binnen te zijn.

Namens de VvS+OR,
Prof. dr. Eric Cator | *jurivoorzitter Jan Hemelrijk Award*
Prof. dr. Jelle Goeman | *jurivoorzitter Willem R. van Zwet Award*
Drs. Sander Scholtus | *secretaris der beide jury's*



Douwe van der Sluis (1942–2017)

Op 31 oktober 2017 is Douwe van der Sluis plotseling overleden. Douwe werd op 9 augustus 1942 geboren in Aegum, een klein gehucht van zo'n 30 inwoners tussen Leeuwarden en Grouw.

Hij studeerde af aan de Universiteit Groningen in juli 1971 als numeriek wiskundige. Daarna trad hij in dienst bij het Rekencentrum van de RUG. Douwe heeft bijna 25 jaar op dat Rekencentrum gewerkt en daarna nog 10 jaar bij de faculteit PPSW (Psychologische Pedagogische en Sociologische Wetenschappen). Douwe heeft zich gedurende zijn gehele loopbaan vooral bezig gehouden met statistische pakketten, waarbij in de loop der tijd het zélf maken van programmatuur werd vervangen door het geven van cursussen in het gebruik van standaardprogrammatuur. Twee grote projecten springen daar uit, namelijk WESP (Waarlijk Eenvoudig Statistisch Pakket) en POSCON (POSterior CONfidence intervals). WESP is in Groningen vele jaren hét pakket voor de verwerking van statistische gegevens geweest, duizenden studenten en medewerkers hebben het gebruikt. Na 1980 en vooral na 1990 werd SPSS het belangrijkste statistisch pakket. Douwe maakte in 2004 gebruik van de FPU regeling en ging met vervroegd pensioen.

Douwe is actief geweest in een groot aantal commissies en werkgroepen, zowel binnen als buiten de Universiteit. Hij heeft een bijzonder grote rol gespeeld in de voormalige Sectie Statistische Programmatuur van de VvS+OR. Hij was jarenlang de coördinator van de werkgroep SPSS van deze sectie en een zeer actief lid van veel andere werkgroepen. Hij zat in vrijwel alle organisatiecomités voor de tweejaarlijkse Symposia Statistische Software die de SSP zo'n 20 jaar lang organiseerde. Ook maakte hij, namens de VvS+OR, deel uit van het organisatiecomité van het IASC congres Compstat 2000 in Utrecht.

Als laatste secretaris van de SSP heeft hij, na het opheffen van de sectie, er voor gezorgd dat het uitgebreide archief is overgedragen aan het Rijksarchief te Haarlem. Deze instelling fungeert als bewaarplaats voor vrijwel alle wetenschappelijke archieven in Nederland en beheert nu, dankzij Douwe, dus ook de nalatenschap van een belangrijk onderdeel van de VvS+OR-geschiedenis.

Bij de gedachtenisbijeenkomst op 5 november werd door veel sprekers ingegaan op de grote inzet, betrokkenheid en behulpzaamheid van Douwe, iets waarvan ook de VvS+OR in hoge mate heeft mogen profiteren.

Gerrit Stemerdink

Gerard Sierksma

column

Sven_Kramer op de 10 kilometer tijdens de wereldkampioenschappen in Sotchi, 2013. Foto: Sasha Krotov (CC 3.0)

BHAIPARTHAIT EN PYEONGCHANG

OLYMPISCHE SCHAATSSTATISTIEK MET PRIJSVRAAG

Pyeongchang. Ik heb geen idee hoe je dat moet uitspreken. De echte schaatsliefhebbers weten het vast wel. In ieder geval is hen bekend dat hier de Olympische Winterspelen worden gehouden. Straks in februari met een breed palet aan disciplines, van kunstschaatsen tot curling. In Pyeongchang dus, een van de noordelijkste provincies van Zuid-Korea. Voor ons telt eigenlijk alleen het schaatsen, wellicht inclusief dat maffe shorttrack. Met dat shorttracken erbij halen we vast wel weer de top-5 in het landenklassement van Olympische medailles. En dat zou knap zijn. Maar voordat de Oranje-schaatsers oostwaarts trekken zijn er eerst nog een paar hobbels te nemen, zoals besluiten wie er gaan. Nederland telt een groot aantal medaillekanshebbers. Maar omdat er van het IOC maar tien mannen en tien vrouwen – ik heb het nu over het echte schaatsen – mogen starten per land, moet er zeker een aantal toppers thuisblijven. Tja, hoe pak je die selectie dan aan?

Realisaties en prognoses: maar al te vaak dekken ze elkaar niet en krijgen statistici de hoon van volk en media over zich heen als hun voorspellingen niet uitkomen. Zou dit ook te maken kunnen hebben met de term 'voorspellen'? Die zou namelijk de suggestie kunnen wekken dat statistici de glazen bol begluren om de toekomst te kennen. De termen verwachting en prognose hebben die connotatie niet. Ook bij selecteren gaat het om verwachtingen: nu beslissen in de verwachting dat later de medailles komen. Acht jaar geleden alweer, zo ergens in november voor de Vancouver-spelen, kreeg ik het idee om het selecteren van schaatsers voor Olympische Spelen volledig door de computer te laten uitvoeren. Te berekenen dus, in plaats van te 'beredeneren'.

Wat was dan wel dat idee? Maar daartoe eerst wat wiskundig jargon. Centraal stond een *weighted complete bipartite graph*. Tussen haakjes, ik krijg de Nederlandse vertaling hiervan niet goed uit mijn pen: 'Graaf' dat klinkt toch nergens naar in dit verband, om niet te spreken over 'bipartiet' of 'tweedelig'. 'Bhaiparthait', dat klinkt als een klok in de collegezaal. OK. De *nodes* 'links' in de *graph* betreffen de schaatsers en de *nodes* 'rechts' de Olympische schaatsstartplekken. Daartussen liggen de *edges* met bij elke *edge* een getal, *weight* genaamd, dat aangeeft wat de kans is dat de schaatser links op die *edge* een medaille wint op de startplek rechts op die *edge*. Zo was er, voor wat betreft de Vancouver-spelen, een *edge* tussen de *node* van Kramer naar een startplek (er waren er toen drie) op de 5 kilometer met een *weight*-kans van 100%, of beter gezegd 1. De *edge* van Kramer naar een startplek op de 500 meter bevatte de kans 0. *Mutatis mutandis* geldt dit verhaal natuurlijk ook voor de schaatssters.

Dat *bipartite*-gedoe hebben we omgezet in een *integer linear optimization model* met de benodigde beperkingen, zoals het quotum van maximaal 10 schaatsers en de aantallen startplekken op de diverse afstanden. Ook het oplossen ervan is nauwelijks schokkend voor een doorgewinterde OR-er, hoewel een tweedejaars student OR hier toch gauw het *if-then constraint*je vergeet. En dat studenten het dan vaak hebben over 'integur programmerun' neem ik dan maar voor lief, al klinkt het nergens naar.

Een maand later, tijdens het OKT (Olympisch Kwalificatietoernooi) in december belde ik Arie Koops, de Technisch Directeur van de KNSB en verantwoordelijk voor de gang van zaken bij de Olympische schaatsselecties langebaan. Een dag later al zat ik bij hem aan tafel en een week later hadden we met terugwerkende kracht de selectie voor de Winterspelen van Turijn berekend. Onze computerberekening week op twee plaatsen af van die van de KNSB. Zo zou de computer Gerard van Velde laten starten op de 500 meter in plaats van Beorn Nijenhuis die slechts 35ste werd, en Mark Tuiter op de 1500 meter in plaats van Simon Kuiper die teleurstelde met een vierde plaats. Koops was verrast maar niet onthutst. Jac Orie, de zeer succesvolle 'wetenschappelijke' schaatscoach van beide niet-geselecteerde heren, stak beide armen omhoog toen de Groningse computer hem alsnog in het gelijk stelde. Van Velde stopte daarna met wedstrijdschaatsen en Tuitert won vier jaar later in Vancouver goud op de 1500 meter.

De *bipartite graph* was de start van een nauwe samenwerking tussen RUG en KNSB, waarbij later ook ORTEC-sports werd betrokken. Voor de Spelen van Vancouver hebben we op eigen houtje de computerselectie bepaald, maar mochten die niet openbaar maken. Buiten de mediaschijnwerpers hebben we de kansen in de *bipartite graph* berekend en genoteerd in de Prestatiematrix, een voor de vrouwen en een voor de mannen. Op basis van de beide Prestatiematrices werden vervolgens de zogenaamde Selectievolgordes vastgesteld. Zo'n Selectievolgorde is niks meer dan een lijst van alle individuele startplekken gerangschikt van grootste naar kleinste medaillekans van die plekken. Zo stond bij de Spelen van Sotchi, vier jaar geleden, de 10 kilometer bij de mannen bovenaan en bij de vrouwen was dat de 1.500 meter.

De Prestatiematrix is inmiddels omgedoopt tot Kansenmatrix en de term Selectievolgorde is ver-acronymd tot SeVo. Zoals gezegd bevat de SeVo alle individuele startplekken. Dat zijn er drie voor de 500, de 1.000 en de 1.500 meter. Ook de 3.000 vrouwen en de 5.000 mannen kennen beide drie startplekken. Daarentegen zijn er voor de 5.000 vrouwen en de 10.000 mannen slechts twee startplekken per land beschikbaar. Dit jaar staat ook de Mass Start op het programma, die goed is voor twee extra startplekken op de SeVo. Totaal dus 16 individuele startplekken voor beide seksen.

De grote slag vindt plaats op het OKT in Heerenveen. Het evenement duurt vijf dagen en begint op tweede kerstdag. Op basis van de uitslagenlijsten van dat OKT worden tien namen ingevuld op de SeVo. Zodra de tiende man en vrouw zijn genoteerd, worden de resterende startplekken ingevuld met de reeds genoteerden, ook hier in de volgorde van de OKT-uitslagen. Voor de Mass Start gelden andere invulregels; zie de KNSB-site. Ten slotte hebben we nog de Team Pursuit, waarvan de deelnemers gerekruteerd moeten worden uit de tien schaatsers die op de SeVo zijn genoteerd.

Tot slot de prijsvraag. Ik heb bewust in bovenstaande niet verteld hoe en met welke data we de kansen hebben berekend. Er is behoefte aan een frisse kijk op de selectieprocedure en daarom vragen we studenten om een nieuwe slimme selectieprocedure te ontwerpen, die aan de voorwaarden van het IOC voldoet. Die voorwaarden staan op de site van de KNSB. Daar komt bij dat het NOC*NSF en de KNSB per se in december, voorafgaand aan de Spelen, een OKT willen organiseren. Voor de ontwerper van de meest innovatieve oplossing stel ik een paar bijna honderd jaar oude Koningin Wilhelmina-schaatsen beschikbaar; zie foto.

P.S. U vraagt zich wellicht af wat die h's moeten in bhaiparthait. Dat zit zo. Een opmerkzame lezer van een vorige versie van dit verhaal maakte mij erop attent dat ik op college het woord *bipartite* op zich wel goed uitsprak, maar dat mijn Neder-Saksische roots... Afin duidelijk. Bhaiparthait: heerlijk.

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

DAG VOOR STATISTIEK EN OR 2018

28 en 29 maart 2018 in Utrecht

thema CLIMATE CHANGE

Het programma 28 maart begint om 14:30 uur en omvat een hackathon, lezingen, prijsuitreikingen, de ALV van de VvS+OR en borrel. Na de borrel is er een diner en is er feest.

Het programma van 29 maart begint om 9:15 uur en bestaat uit lezingen van sprekers die aangedragen zijn door de secties van de vereniging. De dag wordt om 16:30 uur afgesloten met een borrel. Zie voor uitgebreide informatie het eerstvolgende nummer van STAtOR en de website van de VvS+OR.

Reserveer alvast de data!

NGB/LNMB seminar op 18 januari 2018 in het Congrescentrum De Werelt in Lunteren

WAT VERANDERT DATA SCIENCE AAN OPERATIONS RESEARCH?

De Operations Research in Nederland staat wereldwijd bekend om zijn goede kwaliteit, zowel wanneer het gaat om onderzoek als toepassingen in de praktijk. Men is verrast dat ons land zo veel operations-researchconsultancybedrijven kent. Ook het aantal Franz Edelman Award-winnaars en finalisten uit Nederland is verbazingwekkend hoog.

Aan de andere kant lijkt het erop dat de Nederlandse Operations-Researchgemeenschap niet staat te springen om Data Science te omarmen als waardevolle toevoeging op Operations Research (zowel in de wetenschap als in het bedrijfsleven). Echter, in het bedrijfsleven lijkt men steeds meer op zoek te zijn naar Data Scientists die breder zijn opgeleid dan operations researchers. Deze observatie triggert verschillende vragen. Zouden huidige operations researchers beoefenaars beter bekend moeten raken met de ontwikkelingen in data science, en hun vaardigheden uitbreiden? Zouden data-sciencevakken moeten worden opgenomen in onderwijsprogramma's? Wat zijn de wetenschappelijke en praktijkuitdagingen wanneer we operations research combineren met data science? Is het tijd om uit onze comfortzone te stappen?

sprekers

- Emile Aarts (rector magnificus, Tilburg University)
- Dimitris Bertsimas (Professor OR/Statistics, MIT)
- Merwin de Jongh (Founder & CTO, Building Blocks)
- Gertjan de Lange (SVP, AIMMS)
- Patrick Hennen (COO, ORTEC Data Science)
- Han Hoogeveen (Utrecht University)
- Aziz Mohammadi (Dir. Advanced Analytics, VodafoneZiggo)
- Seppo Pieterse (Fellow and Founder, Quintiq)
- Alexander Rinnooy Kan (UvA, Big Data Alliance)
- Arjen Vestjens (Managing Partner, CQM)
- Bernard Vroom (OR Group Manager, Air France-KLM)

Bezoek het seminar en leer van de experts over de impact van Data Science op Operations Research!

Voor meer informatie en registratie
zie http://tinyurl.com/NGB-LNMB2018

Deze conferentie wordt mede mogelijk gemaakt door ORTEC

THE YOUNG STATISTICIANS

HAPPY 2018 & EXPECTATIONS

While writing this, I was wondering about the word expectation. It is of course a known statistical concept, the expected value or outcome of a certain random variable or process. But there is more. People form expectations every day, for example about the weather. Moreover, Dutch people have a tradition to start a new year with several clearly specified good intentions. However, the expected value of the success probability of these expectations is unfortunately not equal to one. As Young Statisticians, we have high expectations of the coming year. We want to continue with the organization of Pub Quizzes, Statistics Cafés and Company Visits. But first an overview of the events at the end of 2017.

Monday 18 September: Company Visit at Alliander

The kick-off event of this academic year was a company visit to network provider Alliander. At their office in Arnhem, we learned how their 'Data & Insight' department learns from the past, monitors the present and predicts the developments in future energy. We listened to an interactive case in which machine learning was used to predict waiting times, joined a nice energy quiz and had time for drinks and networking.

Wednesday 1 November: Pub Quiz

We organized a statistical pub quiz in Leiden for all Young Statisticians at the clubhouse of student association VSL Catena. There were questions about heroes in statistics, R-code, conjugate priors and many more. All people had a lot of fun and enjoyed the questions and music.

Tuesday 28 November: Statistics Café

The topic of this symposium was 'Sports & Statistics'. We listened to inspiring and interesting talks about the ways of applying statistics and data science to sports. Arie-Willem de Leeuw, researcher at Leiden University, explained how he builds models and visualizations based on sports data to help improve the prestations of sport teams. Also, the company SciSports told us how they use soccer data to get rankings of the best players, biggest talents, the fit between a possible new player and a team and prediction of match results. An inspiring evening!

The most famous soccer model is (not) surprisingly based on expectations. Based on soccer data, an expected goals model can be formed. This is a model in which the probability of scoring a goal is predicted by variables like distance from the goal, part of the body used, pass before the attempt, distance to closest player. In this way, a probability of scoring a chance can be estimated and players can be ranked and graded based on this model.

The Young Statisticians Board hopes to see you again at all activities we want to organize in 2018. We advise to pay extra attention to the VvS+OR-days in March. These two days will consists interesting speakers and fun activities, which is a fruitful combination. Also, we are searching for a new board member. So if you like to organize events and want to join the board or gather more information, you can email us at contact@youngstatisticians.nl. Finally, we want to wish everyone a happy New Year with a lot of statistics in it.