STAtOR

periodiek van de VvS+OR jaargang 18, nummer 2, juli 2017

STAtOR

Jaargang 18, nummer 2, juli 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, Fred Steutel en Henk Tijms.

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

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

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

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

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

Ontwerp en opmaak
Pharos, Nijmegen

Uitgever

ISSN 1567-3383

Monter de vakantie in

Dit nummer van *STAtOR* zal voor veel lezers aan het begin van hun vakantie verschijnen. Dat komt goed uit, want het bevat artikelen van Roos en Hurkens die wat technischer zijn dan gewoonlijk. De leidraad voor *STAtOR* is het zoveel als mogelijk vermijden van veel en gecompliceerde formules, daar zijn we ditmaal bewust van afgeweken. We gedenken namelijk de onlangs overleden Jacques Benders die in OR-kringen wereldberoemd is geworden door de naar hem genoemde decompositie. En die decompositie is nu eenmaal niet goed uit te leggen zonder al die formules.

Voor hen die hun vakantie graag wat 'lichter' verpozen zijn er daarnaast genoeg bijdragen die met minder inspanning gelezen kunnen worden. Martijn van Ee behandelt de verdwaalde koe, Tim Lamballais Tessensohnt schrijft over robots in het magazijn en Richard Starmans behandelt het verband tussen statistiek en wetenschapsfilosofie. Voeg daarbij een verslag over de Hackathon tijdens de Dag voor Statistiek en OR, de bijdrage van de Young Statisticians en de columns van Henk Tijms en Fred Steutel en we hebben weer een zeer gevarieerd nummer.

Tijdens het persklaar maken van dit nummer kwam het bericht van het onverwacht overlijden van Fred, we gaan daar kort op in aan het einde van zijn column. In ons volgende nummer zullen we er meer uitgebreid bij stilstaan.

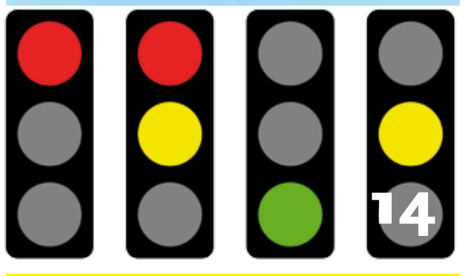
STATOR EN DE ENGELSE TAAL

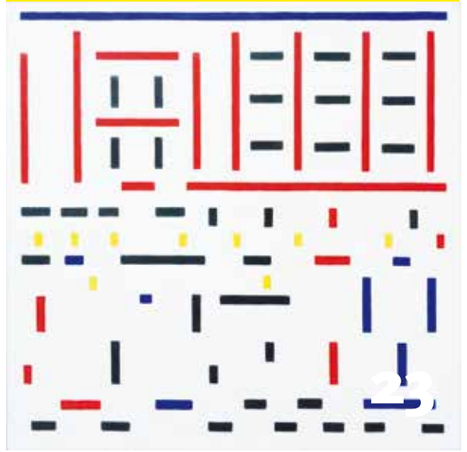
Vanaf het begin in 2000 is voor *STAtOR* het Nederlands de voertaal geweest. Dat was een bewuste keus van het VVS+OR bestuur en de redactie. *STAtOR* is namelijk, naast een blad voor onze leden, óók bedoeld om kennis over statistiek en OR in brede kring te verspreiden. Maar in dit nummer vindt u een artikel in het Engels. Gaat *STAtOR* een andere koers varen?

Het gebruik van de Engelse taal in *STAtOR* is vorig jaar door de redactie besproken met het VVS+OR bestuur, we hebben toen besloten om in principe het Nederlands als voertaal te blijven hanteren. Wel is geconstateerd dat het Engels meer en meer de overhand krijgt op onze universiteiten en dat we daarom, in voorkomende gevallen, ook Engelstalige artikelen zullen publiceren. Vandaar dat we in nr. 2016-3 het Engelstalige artikel van Yinyi Ma hebben opgenomen: zij was een PhD-studente die geen Nederlands beheerst. Ook het artikel van Kees Roos in dit nummer is in het Engels. Hij had dit al gereed en wij vonden het uitstekend geschikt om, naast het In Memoriam, te illustreren hoe belangrijk het werk van Benders is geweest.

Overigens; vanaf het begin van de Young Statisticians zijn hun mededelingen in het Engels. En ook de informatie over de Dag voor Statistiek en OR wordt al vele jaren in het Engels gepubliceerd. Dus: don't worry, in principe Nederlands, maar er kan worden afgeweken. De redactie wenst u veel zomers leesplezier!

De redactie





INHOUD

ROBOTS IN HET MAGAZIJN

Prestaties van robotsystemen modelleren met wachtrij-netwerken

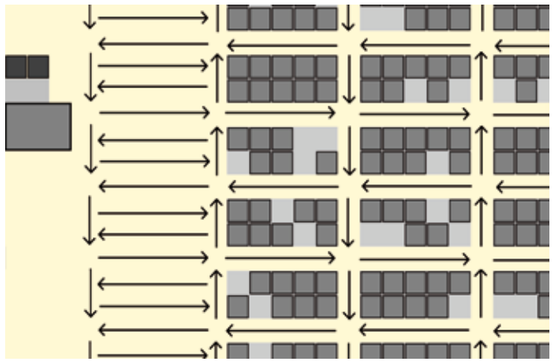
Tim Lamballais Tessensohn

Het verwerken van e-commerce bestellingen stelt magazijnen voor grote uitdagingen. Het assortiment aan producten is vaak groot, waarmee de afstanden in het magazijn ook groot worden. Daarnaast bestaan de bestellingen meestal uit één eenheid, in plaats van complete dozen of pallets die kostenefficiënt te verplaatsen zijn. Daarbovenop komt dat de frequentie waarmee bestellingen binnenkomen voor een bepaald product sterk kan variëren, wat een gunstige positionering van veelgevraagde producten verhindert. Hierdoor kunnen traditionele, handmatige systemen, waarbij werkers door een magazijn lopen, erg inefficiënt zijn voor e-commerce bedrijven.

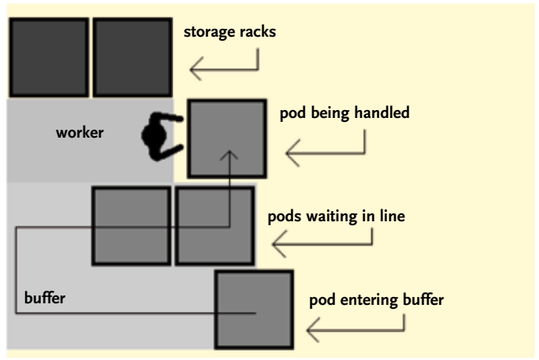
Zogeheten Robotic Mobile Fulfillment Systems (RMFS) vormen een nieuwe categorie van geautomatiseerde opslag- en bestellingenverzamelingssystemen, die speciaal ontwikkeld is om e-commerce bestellingen te verwerken. Deze RMFS worden op de markt gebracht door bedrijven als Amazon Robotics (voorheen bekend als Kiva Systems), Swisslog, Interlink, Grey-Orange, Scallog en Mobile Industrial Robots. Implementaties wijzen er tot nu toe op, dat de snelheid van het verzamelen van producteenheden kan verdubbelen in vergelijking met traditionele, handmatige systemen (Wulfraat, 2012).

Hoe het systeem werkt

De belangrijkste vernieuwing waarmee een RMFS zich onderscheidt van andere geautomatiseerde systemen, is de combinatie van robots en speciaal ontworpen kasten. Deze kasten bevatten de producten en worden door



Figuur 1. Bovenaanzicht van een deel van het magazijn



Figuur 2. Bovenaanzicht van een werkplek

de robots opgetild, verplaatst en neergezet. De robots transporteren de kasten tussen de opslagruimte en de werkplekken. Bij een werkplek aangekomen sluit de, met een kast beladen, robot aan in een wachtrij en wacht totdat hij aan de beurt is bij de werker. De werker haalt vervolgende de benodigde producten van de kast af (picken), of vult de kast bij met nieuwe producten (aanvullen). Figuur 1 laat een bovenaanzicht zien van een deel van het magazijn, met een werkplek aan de linkerkant en een deel van de opslagruimte aan de rechterkant. Overal geldt éénrichtingsverkeer, om blokkades en verkeersopstoppingen te voorkomen. Figuur 2 laat een wat gedetailleerder bovenaanzicht zien van een werkplek, inclusief de buffer waar kasten in de rij moeten wachten totdat ze bij de werker aan de beurt zijn.

Een RMFS is flexibel, in de zin dat een kast geen vaste locatie in de opslagruimte hoeft te hebben. In plaats daarvan kan een kast gedurende de dag op andere plekken neergezet kan worden, afhankelijk van hoe snel de kast in de toekomst nodig zal zijn (Wurman, D'Andrea & Mountz 2008). Voorraad kan dus dicht bij de werkplekken blijven wanneer het in de nabije toekomst nodig is, en verder weg worden gezet wanneer dat niet zo is.

De onderzoeksvraag

Een onbeantwoorde, maar belangrijke, vraag die hier om de hoek komt kijken is: over hoeveel kasten moet de voorraad van een bepaald product verspreid worden? Deze vraag is ook voor andere geautomatiseerde systemen nog niet bevredigend beantwoord. In een e-commerce magazijn is een van de belangrijkste prestatiemaatstaven hoe lang de wachttijd van een klant is en het gebruik van meerdere kasten heeft hier een grote invloed op. Als alle voorraad aan één kast wordt toegewezen, dan bestaat het risico dat dat product tijdelijk niet beschikbaar is, omdat het even kan duren voor een (bijna) lege kast aangevuld is. Aan de andere kant, als de voorraad over teveel kasten wordt verspreid, dan is de voorraad van een enkele kast sneller op en gebeurt het aanvullen vaker. Het wordt dan ook minder waarschijnlijk dat een grote bestelling met één enkele kast kan worden voldaan. Zowel met teveel als met te weinig kasten per product zullen bestellingen voor dat product worden vertraagd, wat de gemiddelde wachttijd van de klant verhoogd. Dit hangt ook samen met het zogeheten 'aanvullingsniveau'. Als het aantal eenheden van een product op een kast onder dit niveau daalt, dan wordt de kast naar een werkplek gebracht om aangevuld te worden met voorraad van dat product. Zeker met meerdere kasten kan het voordelig zijn om een kast aan te vullen voordat die compleet leeg is. Een hoger aanvullingsniveau betekent dat aanvullen vaker gebeurt en dus dat er extra transport door de robots nodig is en meer wachttijd bij de werkplekken ontstaat. Maar het betekent ook dat de gemiddelde voorraad op een kast hoger is en dus dat grotere bestellingen minder lang hoeven te wachten. De gemiddelde lengte van de wachtrij bij de werkplekken wordt bovendien ook beïnvloed door de verhouding van het aantal pick-werkplekken, waar producten worden verzameld voor bestelling, en het aantal 'aanvullingswerkplekken', waar producten op de kasten worden gezet. Een hoger aanvullingsniveau hoeft niet noodzakelijkerwijs te leiden tot meer wachttijd, als het aantal aanvullingswerkplekken ook hoger is. Als het aantal kasten per producttype en aanvullingsniveau niet optimaal is, kunnen lange en onnodige vertragingen optreden die een grote invloed op de klantwachttijd hebben. Als de verhouding tussen het aantal pick-werkplekken en het aantal aanvullingswerkplekken niet geoptimaliseerd is, kunnen sommige werkplekken onaanvaardbaar lage bezettingsgraden hebben, terwijl andere werkplekken kampen met lange wachtrijen.

Dit onderzoek bekijkt hoe de klantwachttijd geminimaliseerd kan worden door de volgende drie beslissingsvariabelen te optimaliseren: 1. het aantal kasten per product, 2. de verhouding tussen het aantal pick-werkplekken en het aantal aanvullingswerkplekken, en 3. het aanvullingsniveau.

Wachtrij-modellen en -netwerken

We modelleren de processen in het systeem met behulp van wachtrijen. Vanuit het perspectief van de kast, zijn er acht processen die van belang zijn en gemodelleerd moeten worden. De kast (1) wacht totdat er een bestelling binnenkomt waarvoor de kast nodig is, (2) wacht totdat een robot arriveert bij zijn opslaglocatie, (3) wordt door de robot opgetild en naar een pick-werkplek gebracht, (4) wacht in de rij samen met de robot totdat het zijn beurt is en ondergaat vervolgens het pick proces waarbij de werker producten van de kast afhaalt, waarna de kast of (5) terugkeert naar de opslagruimte of, als de voorraad van de kast onder het aanvullingsniveau is gedaald, (6) verder gebracht wordt naar een aanvullingswerkplek waar (7) de kast samen met zijn robot weer in de rij wacht op zijn beurt en dan wordt bijgevuld door de werker van de aanvullingswerkplek, om ten slotte (8) terug naar de opslagruimte gebracht te worden.

Elk van deze processen kan worden gemodelleerd als een wachtrij. Bij processen waar een robot een kast vervoert, kan de reistijd worden gemodelleerd als de servicetijd van een i*nfinite server* wachtrij, terwijl de processen waar een werkplek bij betrokken is, gemodelleerd kunnen worden als een normale wachtrij waarbij de servicetijd overeenkomt met de afhandelingstijd van de betreffende werkplek. Het matchen van een bestelling met een kast kan gemodelleerd worden met een zogeheten *synchronization station*. Op deze manier kan het netwerk aan processen omgezet worden in een wachtrijennetwerk. Dit is een zogeheten s*emi-open queuing netwerk*, omdat het open is in de zin dat orders het netwerk binnenkomen en verlaten, terwijl het ook gesloten is in de zin dat de kasten het systeem niet kunnen verlaten. Aangezien het RMFS speciaal ontworpen is met het oog op e-commerce, wordt aangenomen dat alle bestellingen voor slechts één product zijn, en niet voor meerdere producten. Als elke bestelling ook slechts één eenheid van een product vereist, kan het wachtrijmodel worden opgelost met de methoden beschreven in Buitenhek, van Houtum & Zijm (2000). We nemen echter aan dat een bestelling meerdere eenheden van een product kan vereisen. Dit maakt het analyseren en oplossen van het wachtrijmodel ingewikkelder, omdat de bestelling dan alleen gematcht kan worden met een kast die nog genoeg eenheden bevat. Om dat aspect mee te kunnen nemen, stellen we een Markov-keten op die het hele systeem omvat. Het analytisch uitrekenen van de gemiddelde tijd dat het systeem in elke toestand van de Markov-keten doorbrengt levert dan de gemiddelde klantwachttijd op, en andere statistieken zoals de bezettingsgraden van werkplekken en robots.

Resultaten en conclusie

De analyse van het netwerk levert drie inzichten op. Ten eerste is het zinvol om het aantal kasten per product te optimaliseren, te weinig of te veel kasten leidt tot een grote toename van de klantwachttijd. Oftewel, ook al passen alle eenheden van een product op één kast, dan kan het toch zinvol zijn om die eenheden over meerdere kasten te verspreiden. Een nadeel zou zijn dat dit kan leiden tot extra werk bij de aanvullingswerkplekken. Ten tweede is de optimale verhouding tussen het aantal pick-werkplekken en het aantal aanvullingswerkplekken sterk afhankelijk van het aanvullingsniveau. Ten slotte verslechteren de prestaties van het systeem ernstig als een kast pas aangevuld wordt als die compleet leeg is en niet eerder, zelfs als er meerdere kasten per product gebruikt worden.

Hoewel de mens nog lang niet uit de magazijnen verdreven is, hebben robotsystemen laten zien dat ze een belangrijk deel van het mensenwerk over kunnen nemen.

Literatuur

Buitenhek, R., Van Houtum, G.-J., & Zijm, H. (2000). AMVA-based solution procedures for open queueing networks with population constraints. *Annals of Operations Research, 93*, 15–40.

Wulfraat, M. (2012). *Is Kiva systems a good fit for your distribution center? An unbiased distribution consultant evaluation*. http://www.mwpvl.com/html/kiva_systems.html.

Wurman, P. R., D'Andrea, R., & Mountz, M. (2008). Coordinating hundreds of cooperative, autonomous vehicles in warehouses. *AI Magazine, 29*(1), 9–19.

Tim Lamballais Tessensohn is werkzaam als junior scientist bij TNO.
E-mail: tim.lamballaistessensohn@tno.nl

STATISTIEK EN WETENSCHAPSFILOSOFIE

over het nut van de ornithologie voor de grasparkiet

Hoewel veel wetenschapsfilosofen vrijwel onmiddellijk de relevantie van de probabilistische revolutie erkenden, voltrok zich paradoxaal genoeg een boedelscheiding, een verkaveling die tot op heden voortduurt. De daarbij optredende soms gekunstelde scheidslijnen bepalen ook de huidige relatie tussen statistiek en filosofie. In dit korte essay wordt getracht aannemelijk te maken dat het anders kan en moet.

Richard Starmans

Een kleine eeuw geleden was de verhouding tussen natuurkunde en filosofie nog betrekkelijk harmonieus. Dit in weerwil van fundamentele omwentelingen als relativiteitstheorie en kwantummechanica, die de aanschouwelijkheid van het wereldbeeld sterk ondermijnden en de oude en vertrouwde wijsgerige concepties van tijd, ruimte en causaliteit problematisch maakten. Fysici als Einstein, Heisenberg, Bohr en Schrödinger waren filosofisch zeer onderlegd en gaven daarvan veelvuldig blijk in woord en geschrift. Filosofen als Russell, Carnap, Reichenbach en Schlick volgden de actuele natuurkundige ontwikkelingen op de voet. Sommigen geloofden zelfs oprecht in de noodzaak, dan wel mogelijkheid van wisselwerking en kruisbestuiving of in ieder geval van een vreedzame co-existentie. Van dit alles is hedentendage weinig over. De natuurkundige en Nobelprijslaureaat Richard Feynman werd decennia geleden reeds veelvuldig geciteerd vanwege zijn dictum ‘Philosophy of science is about as useful to scientists as ornithology is to birds’. Steven Weinberg, eveneens winnaar van de Nobelprijs, hekelde in 1993 ‘the unreasonable ineffectiveness of philosophy’, met een knipoog naar Eugene Wigners bewondering voor ‘the unreasonable effectiveness of mathematics’. In (Weinberg, 1993) wijdde hij aan de problematiek een apart hoofdstuk, dat getooid ging met de veelzeggende titel ‘Against Philosophy’. De astronoom Stephen Hawking proclameerde enige tijd geleden zelfs de dood van de filosofie en meer recentelijk mengde ook de natuurkundige en beroemde popularisator van de fysicus Laurence Kraus zich in het debat. In 2015 gooide Weinberg met zijn boek *To Explain the World: The Discovery of Modern Science* opnieuw de knuppel in het hoenderhok. Het kwam hem te staan op een stevige reprimande door de historicus Stephen Shapin in zijn recensie ‘Why scientists shouldn’t write history’ (2015), gepubliceerd in *The Wall Street Journal*. De lijst kan moeiteloos worden uitgebreid, maar ook hun minder polemisch ingestelde collega’s bekennen zich doorgaans tot een eliminatief reductionisme, waarin vele filosofische concepten en ‘qualia’ van de hand worden gedaan of een specifieke abstracte of louter wiskundige duiding krijgen. Dat betreft uiteenlopende noties als ruimte, tijd, beweging, oorzakelijkheid, intentionaliteit, teleologie, maar ook betekenis, zingeving, geest, vrije wil, bewustzijn, persoonlijke identiteit en doorgaans alle metafysische of als religieus geduide denkbeelden.

Gelukkig wordt de relatie tussen statistiek en wetenschapsfilosofie niet gekenmerkt door een vergelijkbare animositeit of een bij tijd en wijle onversneden ressentiment. Toch verloopt de samenspraak tussen beide al decennia lang enigszins moeizaam. Ofschoon de genealogie van deze problematiek een aantal factoren kent, beperken we ons hier tot één cruciaal aspect ervan: sinds haar ontstaan rond 1925 heeft de wetenschapsfilosofie een radicaal onderscheid gemaakt tussen de *context of discovery* en de *context of justification*. In haar aanvankelijke, door het logisch positivisme geïnspireerde streven, een rationele reconstructie van (eenheids)wetenschap te geven en bijbehorende demarcatie-criteria te formuleren, richtte de wetenschapsfilosofie zich vooral op de context of justification, en niet op de praktijk van het onderzoek en een daarbij noodzakelijke constructieve methodologie. Hoewel vele wetenschapsfilosofen vrijwel onmiddellijk de relevantie van de probabilistische revolutie erkenden, voltrok zich paradoxaal genoeg een boedelscheiding, een verkaveling die tot op heden voortduurt. De ‘queeste naar de waarheid’ kreeg daarbij een gelaagde structuur, een triptiek bestaande uit vakdiscipline, (onderzoeks-)methodologie en kennisleer/wetenschapsfilosofie (zie ook Starmans, 2017a). Vele debatten uit de ideeëngeschiedenis en hedendaagse knelpunten en uitdagingen in de verhouding tussen wetenschap en filosofie kunnen vanuit deze ‘lagen’ en hun onderlinge relaties worden geduid. De daarbij optredende soms gekunstelde scheidslijnen bepalen ook de huidige relatie tussen statistiek en filosofie.

Dat het anders kan blijkt onder meer uit het werk van de Amerikaanse bayesiaanse statisticus Andrew Gelman. In zijn artikel ‘Philosophy and the practice of Bayesian statistics’ (2011, gereviseerd 2013) bepleit Gelman het eminente belang van filosofie voor zowel het statistisch onderzoek als de statistische praktijk. De auteur stelt dat in dit opzicht veelvuldig sprake is van *freewheelen* en dilettantisme van bepaalde statistici, waaronder hij ruimhartig ook zichzelf schaart. Ofschoon hij zich in voornoemd artikel beperkt tot iconen van de wetenschapsfilosofie als Popper en Kuhn, spoort Gelman zijn collega’s aan om vooral aansluiting te zoeken bij meer recent filosofisch onderzoek. Dat het daarnaast ook anders moet, wordt hier beargumenteerd aan de hand van ontwikkelingen binnen data science en big data. Als vertrekpunt kiezen we allereerst een notoir, doch veelvuldig geciteerd artikel van wetenschapsjournalist Chris Anderson die in 2008 in *Wired Magazine* feitelijk het einde van de wetenschappelijke methode en kennisleer proclameerde. Hij deed dit in zijn bijdrage ‘The end of theory: The data deluge makes the scientific method obsolete’. (*https://www.wired.com/2008/06/pb-theory/*) Daarna laten we de Google-onderzoekers Alon Havely, Peter Norvig en Fernando Pereira aan het woord, die op

analoge wijze de statistiek de maat nemen in hun artikel 'The unreasonable effectiveness of data', dat in 2009 verscheen in *IEEE Intelligent Systems*. De publicaties maken voldoende duidelijk dat een welbegrepen eigenbelang een belangrijke reden kan zijn voor een voorzichtige liaison tussen statistiek en wetenschapsfilosofie.

Gelman en de filosofie

Een recent voorbeeld van toenadering tot de wetenschapsfilosofie binnen de statistiek betreft zoals gezegd Gelmans artikel 'Philosophy and the practice of Bayesian statistics'. De auteur spoort niet alleen zijn collega's aan om vooral aansluiting te zoeken bij meer recent wetenschapsfilosofisch onderzoek, maar hij gaat zelf nog een stap verder. Gelman stelt zich expliciet tot doel een nieuwe filosofie voor de Bayesiaanse statistiek te ontwikkelen, die empirisch adequaat is en de bestaande statistische praktijk beter verdisconteert. In zijn enthousiasme stipuleert Gelman dat het in de hedendaagse statistiek gebruikelijk is om statistische benaderingen of uitgangspunten te relateren aan of zelfs te verankeren in filosofische paradigma's of posities. Hij verwijst daarbij met name naar het inductivisme en het hypothetico-deductieve-model (HD-model) als concepties van wetenschap. Vervolgens typeert hij diverse verworvenheden van de 'klassieke' statistiek als manifestaties van hypothetisch-deductief redeneren; de significantie-testen van Ronald Fisher, het hypothesetoetsen zoals ontwikkeld door Egon Pearson en Jerzy Neyman, en tot slot de theorie van betrouwbaarheidsintervallen, waarvan Neyman eveneens de grondlegger was. De Bayesiaanse statistiek wordt volgens Gelman in de statistiek algemeen beschouwd als een proeve van inductivisme. Beide stromingen worden door de auteur als tegenovergestelde, om niet te zeggen conflicterende visies gepresenteerd. Als een opmaat tot een nieuwe empirisch adequate filosofie voor de Bayesiaanse statistiek kritiseert de auteur de eerder geschetste vigerende typering van het veld en stelt dat de Bayesiaanse statistiek juist als een uitermate succesvolle instantiatie van het HD-model moet worden beschouwd en niet meer of minder inductief is dan de hierboven genoemde uitingen van de 'klassieke' statistiek. De meest succesvolle toepassingen van Bayesiaans redeneren overstijgen volgens Gelman de standaardopvatting waarbij redeneren wordt opgevat als het updaten van een prior-verdeling tot een posteriori-verdeling met behulp van een likelihood-functie op basis van nieuwe data, waarna vervolgens de 'graad van geloof' hieraan dient te worden aanpassen. Het gaat er veeleer om diverse modellen te ontwikkelen (*model fitting*), die vervolgens worden geëvalueerd (*model checking)* en aan een opeenvolging van strenge tests worden onderworpen, waarna het sterkste model overblijft. Omdat volgens Gelman (vrijwel) alle modellen toch 'fout' zijn, dat wil zeggen, niet de kansverdeling bevatten door welke de beschikbare data zijn gegenereerd, is er veeleer sprake van een trial-and-erroraanpak, die historisch gesproken uiteraard met Popper kan worden verbonden, maar in de recente literatuur in Mayo (Mayo, 1996; 2010) de meest prominente pleitbezorger vindt. Curieus is echter dat Gelman niet beseft dat inductivisme en het HD-model allerminst *claire et distinct* zijn en bezwaarlijk als tegengestelde concepties van wetenschap kunnen worden beschouwd, die een rol binnen de statistiek kunnen vervullen welke Gelman hen toebedeelt. Het onderscheid weerspiegelt veeleer de worstelingen van denkers met de redeneervormen en 'omkeringen' in hun zoektocht naar kennis (Starmans, 2013). Met name inductie heeft een lang ontwikkelingsproces doorlopen, kent verschillende verschijningsvormen en aspecten die sterk verbonden zijn met de wijsgerige posities van de betrokkenen. Dat gold voor Aristoteles, Bacon, Hume, Kant en Mill, maar evenzeer voor Popper en de Bayesianen. Het begrip heeft opmerkelijke verschuivingen ondergaan. Een strenge duiding, bij voorbeeld het naïeve inductivisme als *logic of discovery*, zal door niemand meer worden bepleit. Een gematigde en *sophisticated* versie, zoals die aan vele confirmatie-theorieën ten grondslag ligt, zal door weinigen als problematisch worden ervaren. Het HD-model kan niet bogen op een vergelijkbare respectabele geschiedenis en mag niet als een eenvoudige en succesvolle reparatiepoging of alternatief voor inductie worden beschouwd. Het is te algemeen en is feitelijk altijd met inductie verbonden geweest, zelfs in de naïeve versie van Francis Bacon. Als methode bevat het geen wezenlijk nieuwe elementen of inzichten, lost de problemen rond inductie niet op en kent diverse historische en systematische bezwaren, waaronder de beroemde Quine-Duhem these, die ruwweg stelt dat het niet mogelijk is om een cruciale test te ontwikkelen waarmee een afzonderlijke hypothese geïsoleerd wordt getoetst. De hypothese maakt immers deel uit van een netwerk van andere (hulp)hypothesen, begincondities en *ceteris paribus clausules*, zodat niet duidelijk is wat nu precies getoetst dan wel gefalsificeerd is. Hoe dan ook, bij het indelen van een persoon of idee in een van beide kampen liggen onverbiddelijk de drogreden van de foutieve oppositie en die van de stroman op de loer. Bovendien

Andrew Gelman

is de tijd van alomvattende epistemologische theorieën al lang voorbij en bij gevolg is het zoeken naar of postuleren van een één-op-één-correspondentie tussen statistische techniek en wetenschapsfilosofisch paradigma minder voor de hand liggend. Ondanks Gelmans niet zo geslaagde liaison met de filosofie, is zijn streven naar een alliage van methodologie en epistemologie van wezenlijk belang. Dat zijn inspanningen niet onopgemerkt blijven, vindt deels zijn verklaring in het feit dat de auteur zich niet alleen tot ingewijden richt, maar tracht een breder publiek te bereiken en op vele websites via blogs en commentaren actief is.

Het einde van de theorie

Dit alles wint a fortiori aan betekenis door recente ontwikkelingen binnen big data en data science, die verregaande consequenties voor wetenschapsfilosofie en methodologie hebben. Men denke in dit verband bij voorbeeld aan het artikel van wetenschapsjournalist Chris Anderson, die zoals gezegd reeds in 2008 het einde van de wetenschappelijke methode en kennisleer proclameerde. Zijn 'The end of theory: The data deluge makes the scientific method obsolete' is evenzeer voer voor filosofen als voor methodologen en statistici. Een kleine bloemlezing van de soms tamelijk onbezonnen opmerkingen van Anderson moge hier volstaan. Allereerst stelt de auteur dat '*Now Google and like-minded companies are sifting through the most measured age in history, treating this massive corpus as a laboratory of the human condition. They are the children of the Petabyte Age'*. Duidelijk is dat de auteur enig pathos niet schuwt. Hij is geen futuroloog, maar stelt onomwonden dat de geschetste situatie veeleer de status quo vormt dan een ver vooruitzicht. '*At the petabyte scale, information is not a matter of simple three- and four-dimensional taxonomy and order but of dimensionally agnostic statistics. It calls for an entirely different approach, one that requires us to lose the tether of data as something that can be visualized in its totality. It forces us to view data mathematically first and establish a context for it later.*' De auteur vervult zijn rol als apologeet met verve. '*Google's founding philosophy is that we don't know why this page is better than that one: If the statistics of incoming links say it is, that's good enough. No semantic or causal analysis is required.*' Dat alles betreft volgens de auteur elke vorm van menselijk handelen. '*Out with every theory of human behavior, from linguistics to sociology.*

Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.' Al met al kan onze hele visie op wetenschap op de schop. *'But faced with massive data, this approach to science – hypothesize, model, test – is becoming obsolete. Consider physics: Newtonian models were crude approximations of the truth (wrong at the atomic level, but still useful). A hundred years ago, statistically based quantum mechanics offered a better picture – but quantum mechanics is yet another model, and as such it, too, is flawed, no doubt a caricature of a more complex underlying reality.'* Daarna moet de biologie het ontgelden. '*The models we were taught in school about "dominant" and "recessive" genes steering a strictly Mendelian process have turned out to be an even greater simplification of reality than Newton's laws. The discovery of gene-protein interactions and other aspects of epigenetics has challenged the view of DNA as destiny and even introduced evidence that environment can influence inheritable traits, something once considered a genetic impossibility. In short, the more we learn about biology, the further we find ourselves from a model that can explain it*.' Andersons slotpleidooi laat evenmin aan duidelijkheid te wensen over. De nieuwe technologie '*offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all*'. Een jaar later publiceerden de Google-onderzoekers Alon Havely, Peter Norvig en Fernando Pereira hun evenzeer lucide maar door simplificaties evenmin onomstreden artikel 'The unreasonable effectiveness of data'. Wanneer het om mensen gaat en niet om elementaire deeltjes zijn eenvoudige wiskundige formules en elegante theorieën van beperkt nut, aldus de auteurs. Datzelfde geldt volgens hen voor parametrische statistische modellen. Om gedrag van mensen te begrijpen en te voorspellen kan men beter vertrouwen op vele petabytes van ruwe, ongeannoteerde, ongestructureerde, desnoods vervormde data. Deze worden dan via efficiënte datastructuren/architecturen gerepresenteerd en vervolgens door slimme, patroon herkennende algoritmen tot kennis gesmeed. In (Starmans, 2016; 2017b) wordt getracht meer recht te doen aan de visie van de drie auteurs, maar de hier enigszins gechargeerde weergave van hun ideeën volstaat ter illustratie van de problematiek, waarbij zowel de statistiek als de 'wetenschappelijke methode' en achterliggende filosofie onder druk komen te staan. Bovendien kan worden gesteld dat de visies van Anderson en Havely cum suis grosso modo al dan niet in afgezwakte vorm nagenoeg overal zijn terug te vinden in de populaire literatuur over big data. Dit impliceert uiteraard allerminst dat tegengeluiden en meer genuanceerde posities afwezig zijn. Dat geldt ook voor Google, en dan vooral in eigentijdse debatten over de risico's van AI.

Wahrheit und Methode

Enige opmerkingen tot slot. Beschouwingen als de onderhavige vooronderstellen dikwijls een common sense notie van wetenschap, een min of meer onproblematisch geacht standaardbeeld. Onderzoekers beschrijven en modelleren aspecten van de werkelijkheid, zoeken naar verklaringen, die idealiter uitmonden in c.q. deel uitmaken van een zo algemeen mogelijke, liefst unificerende theorie, die zowel verklarend adequaat als predictief succesvol is. Met het oog hierop verrichten de onderzoekers observaties, voeren experimenten uit, speuren naar causale verbanden, trachten (natuur)wetten te formuleren, bouwen modellen en toetsen hypothesen. Een wezenlijk aspect hiervan betreft uiteraard de subtiele relaties tussen empirie en theorie; de waarnemingen, de data, het 'bijzondere' en het concrete versus de wetten, de hypothesen, het 'algemene' en het abstracte. Betrouwbare en geldige conclusies dienen de overbrugging van de al dan niet vermeende kloof tussen empirie en theorie overbruggen. Hoe dan ook, dit alles laat onverlet dat de 'wetenschappelijke methode' vaker onder druk heeft gestaan en een ontwikkeling heeft doorgemaakt met een dynamische, verschuivende betekenis, variërend van Descartes' *Discours de la Methode* en Bacons *Novum Organon* tot de anarchistische, *anything goes* benadering van Paul Feyerabend, verwoord in zijn belangrijkste boek *Against Method* (1975). Of men denke aan een geheel andersoortige, geesteswetenschappelijke kritiek van de Duitse filosoof Hans-Georg Gadamer, die de hermeneutische cirkel zo interpreteerde dat waarheid niet langs methodische weg kan worden nagejaagd en nota bene de 19de eeuwse hermeneutici Schleiermacher en Dilthey verweet te zeer de natuurwetenschappelijke 'methode' te hanteren. Dit terwijl het juist Ditlhey was die het strikte onderscheid *erklären* versus *verstehen* introduceerde en zich nadrukkelijk tot de tweede categorie bekende. Feit is dat in het tijdperk van big data de kritiek op wetenschap, het zoeken naar waarheid en de gevolgde methode uit geheel andere hoek komt, daarbij veel radicaler de hier geschetste pijlers van het standaardbeeld ondermijnt, en een schisma zoals bij de fysica in het verschiet lijkt te liggen.

Ook vanuit historisch perspectief is Gelmans preoccupatie met een wijsgerige fundering van de statistiek niet vreemd. Met name in disciplines die juist één of meer grondslagencrises achter de rug hadden c.q. hiermee kampten en tot enige wasdom waren gekomen, bleek in het verleden zo'n oriëntatie manifest; onder meer in de fysica, biologie, psychologie, economie en sociologie. Onderzoekers zoeken dan expliciet aansluiting bij en verankering in grote thema's van de wetenschapsleer, de aldaar geïdentificeerde concepties van wetenschap, de gekoesterde of ontmantelde kennisidealen. Thema's als de aloude queeste naar zekere en objectieve kennis, kennisverwerving, waarneming en perceptie, waarheidsdefinities en waarheidscriteria, werkelijkheid en model, verklaringen, causaliteit en wetmatigheden, de structuur van wetenschappelijke theorieën, demarcatiecriteria, rationaliteit en vooruitgang, en de eenheid van de wetenschappen, et cetera. Maar ook issues als het wetenschappelijk realisme-debat, het inductieprobleem, confirmatietheorieën, et cetera. Zonder hier te vervallen in omineuze uitspraken dat de statistiek in zo'n grondslagencrisis verkeert, lijkt enige wisselwerking met de kennisleer hier toch wenselijk, gelet op het feit dat de statistiek door de probabilistische wending in nagenoeg alle wetenschappen vrijwel universeel wordt toegepast, maar de grondslagen ervan allerminst een volledig ontgonnen terrein betreffen, en de statistiek niet zelden wordt begrepen en toegepast als een *toolbox* van allerhande technieken en methoden, die onderling weinig samenhangen (Starmans, 2011) en in een andere publicatie van Gelman worden omschreven als een *bag of tricks*. Voor statistici kan een uitdaging liggen in de integratie en verzoening met computationele technieken, die niet zonder statistische inferentie en correct gespecificeerde modellen kunnen (Starmans, 2016).

Wetenschapsfilosofen mogen nu de boot evenmin missen en dienen aan te haken bij ontwikkelingen binnen data science. De uitdaging is te laten zien hoe noties als causaliteit, model, verklaring en theorie, unificatie en rationaliteit, interpretatie juist nodig blijven in het data science tijdperk, dat wetenschap als vorm van intentioneel handelen inzicht en verklaring zoekt, waarbij de bevindingen uiteindelijk in een coherent stelsel van uitspraken verwoord worden. Voorwaarde is wel dat de kloof met de bestaande onderzoeksmethodologie wordt geslecht, te meer daar deze methodologie zelf evolueert door nieuwe methoden en technieken van datavisualisering, *computational modeling*, simulaties of andere verworvenheden uit het tijdperk van data science. Het risico is dan dat wetenschapsfilosofische noties worden gebaseerd c.q. verondersteld worden ten grondslag te liggen aan een methodologie die nooit heeft bestaan of al lang niet meer bestaat. Tot slot, de vergelijking met de ornitholoog en de utiliteitsoverwegingen van Feynman. De geclaimde analogie lijkt weerbarstig en onvoldoende recht te doen aan de complexe relaties tussen statistiek en filosofie. Wellicht kan het echter louterend werken te beseffen dat de ornitholoog zich doorgaans met beide benen op de grond bevindt, terwijl het object van beschouwing zich vrijelijk ook in hoger sferen begeeft.

Literatuur

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology, 66*(1), 8–38, doi:10.1111/j.2044-8317.2011.02037.x

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems, 24*(2), pp. 8–12.

Mayo, D. (1996). *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Mayo, D. (2010). *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*. Chicago: University of Chicago Press.

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 Data* (pp. 1–20), New York: Springer.

Starmans, R. J .C. M. (2013). Idolen en Idealen; Francis Bacon, Inductie en het Hypothetisch-Deductieve Model, *Filosofie, 23*(2).

Starmans, R. J .C. M. (2016). The Advent of Data Science - some considerations on the unreasonable effectiveness of data. In P. Buhlmann, e.a. (Eds.), *Handbook of Big Data; Handbooks of Modern Statistical Methods*. New York: Chapman & Hall/CRC.

Starmans, R. J .C. M. (2017a). De triptiek van het Bayesiaanse Paradigma: confirmatie, inferentie en algoritmiek. *Filosofie, 27*(1).

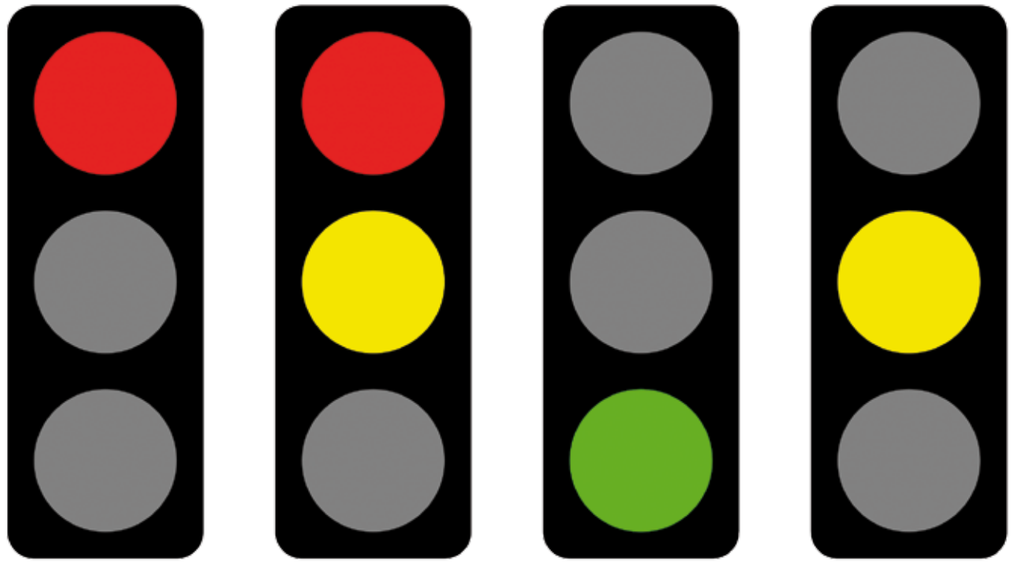
Starmans, R. J. C. M. (2017b). Het einde van de theorie of de onredelijkheid van de data. *Filosofie, 27*(2)

Weinberg, S. (1993). *Dreams of a Final Theory: The Search for the Fundamental Laws of Nature*. New York: Pantheon Books.

Weinberg, S. (2015). *To Explain the World: The Discovery of Modern Science*. New York: Harper/HarperCollins Publishers.

Wigner, E. (1960) The unreasonable effectiveness of mathematics in the natural sciences. *Communications in Pure and Applied Mathematics, 13*(1), 1–14

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



Optimaal stoppen

Elke beslissing is een afweging van risicofactoren. Wanneer het weer plotseling verslechtert moet de meteorologische dienst een beslissing nemen wanneer het tijd is om te stoppen met het afwachten van hoe het weer zich verder ontwikkelt en het moment gekomen is om een weerwaarschuwing te laten uitgaan. Wanneer Apple een nieuwe versie van de iPhone in voorbereiding heeft moet het beslissen wanneer het uittesten van het product gestopt moet worden en het juiste moment aangebroken is om het op de markt te brengen. Dit zijn voorbeelden van beslissingen die genomen moeten worden in een omgeving vol van onzekerheden en die grote gevolgen kunnen hebben wanneer de timing slecht is. Niet alleen grote bedrijven hebben te maken met dit soort beslissingen, ook in het dagelijks leven krijgt een ieder te maken met soortgelijke beslissingen zoals de bepaling van het juiste moment om een huis te kopen of om al of niet een aanbod voor een nieuwe baan te accepteren.

In de kansrekening staan beslissingsproblemen met als vraagstelling 'doorgaan of stoppen?' bekend als optimale stopproblemen. Het meest beroemde optimale stopprobleem is het datingprobleem. Voor kranten en tijdschriften een dankbaar onderwerp dat om de zoveel tijd weer volop aandacht krijgt. Stel dat je op een datingsite de profielen van een aantal niet-gelijkwaardige potentiele partners één voor één in een willekeurige volgorde te zien krijgt. Je mag met slechts één van hen een afspraak maken. Na een profiel van een kandidaat gezien te hebben moet je meteen beslissen een afspraak te maken of niet. Op een eerdere afwijzing kun je niet terugkomen. Als je een kandidaat te zien krijgt kun je alleen maar oordelen of deze kandidaat beter of minder goed is dan de voorgaande kandidaten die je gezien hebt. Anders gezegd, je kunt aan de kandidaten alleen *relatieve* waarderingen toekennen. Welke strategie geeft je in deze situatie de grootste kans de beste te vinden?

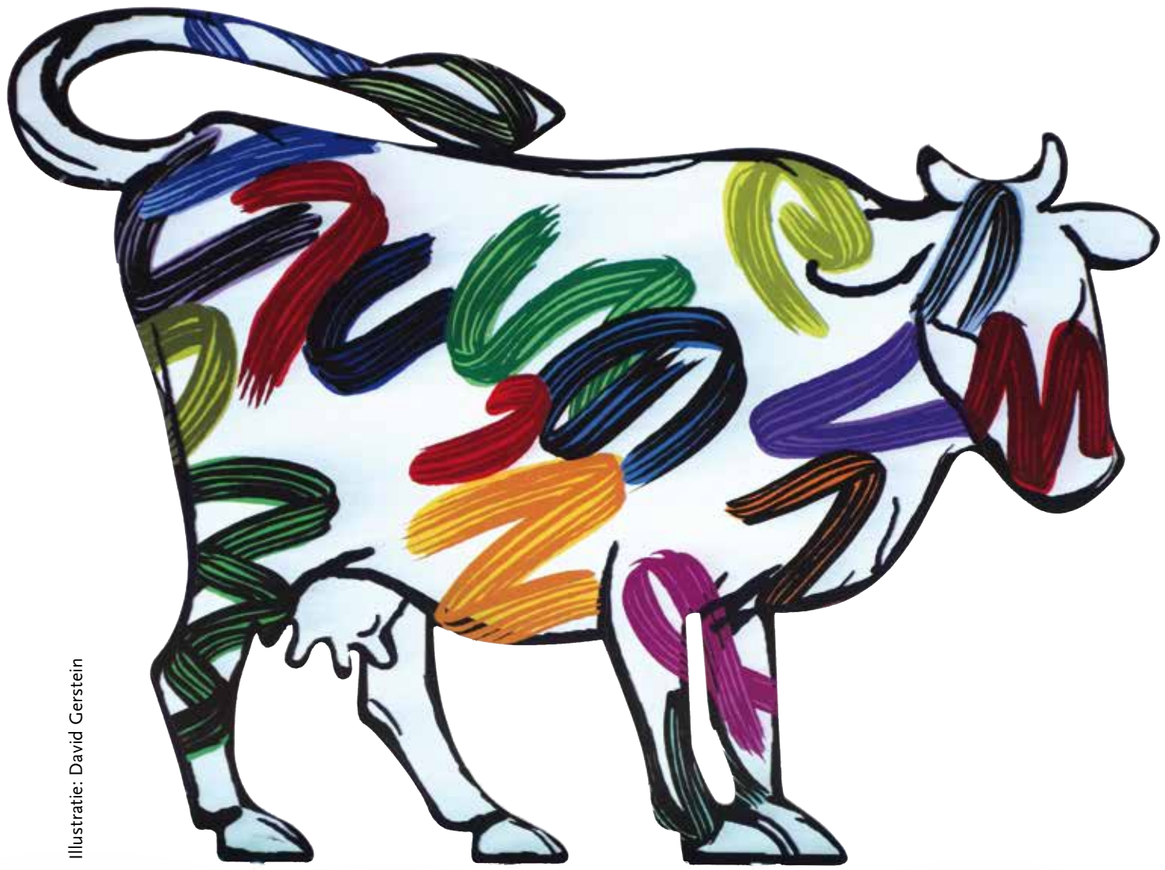
De wiskunde geeft antwoord op deze vraag. Wanneer het aantal kandidaten voldoende groot is, zeg 10 of meer, dan is de algemene vuistregel dat je ongeveer 37% van de kandidaten voorbij moet laten gaan om daarna de eerste te kiezen die beter is dan alle vorige kandidaten. Je hebt dan een kans van ongeveer 37% om de beste te vinden. Velen vinden het verrassend dat deze kans zo groot is en vrijwel niet afhangt van het aantal kandidaten. Zo verrassend is dit ook weer niet wanneer je bedenkt dat je al met een kans groter dan 25% de beste vindt wanneer je de eerste helft van de kandidaten laat passeren en daarna de eerste neemt die beter is dan elk van de anderen die je daarvoor gezien hebt. Onder deze simpele strategie vind je in elk geval de beste wanneer de op één na de beste in de eerste helft zit en de beste in de tweede helft. De kans op deze volgorde is $10/20 \times 10/19 \approx 0{,}263$. De ondergrens van 25% vind je met dezelfde redenering ook als je 1 miljoen kandidaten hebt! Een aardige variant van het datingprobleem is het Googol-spel voor twee spelers. Je laat iemand op een aantal papiertjes verschillende positieve getallen schrijven, waarbij geen beperking opgelegd is aan de grootte van de getallen en elk denkbaar getal opgeschreven mag worden. De papiertjes worden grondig door elkaar geschud, omgekeerd op tafel gelegd en daarna één voor één omgedraaid. Velen zullen de kans dat je het grootste getal zal vinden lager inschatten dan deze kans in werkelijkheid is en daarmee zou je een voor jou gunstige weddenschap kunnen afsluiten.

Verschillende uitbreidingen van het datingprobleem zijn onderzocht zoals de bepaling van de strategie om met maximale kans één van de twee beste te vinden. De maximale kans om te slagen stijgt dan tot ongeveer 57% en bij één van de drie beste tot ongeveer 71%, waarbij de optimale strategie door meerdere omschakelpunten wordt bepaald. Iemand die daar onderzoek over gedaan heeft was de Russische wiskundige Boris Berezovski die ook een gewiekst zakenman was. In de chaotische tijd in Rusland na de val van de Berlijnse muur vergaarde hij een fortuin en werd de rijkste man van Rusland. Berezovski kon doorgaan met zijn schimmige praktijken totdat hij publiekelijk stelling nam tegen Poetins plannen om de constitutionele macht van de president te vergroten. Deze kritiek leidde ertoe dat Berezovski in 2000 hals over kop naar Engeland moest vluchten waar hij harde kritiek bleef uiten op zijn voormalige protégé Vladimir Poetin. In maart 2013 werd Berezovski onder verdachte omstandigheden dood aangetroffen op zijn landgoed in het Engelse Ascot en tot op de dag van vandaag is de oorzaak van zijn dood onopgehelderd. De wiskundige Berezovski had toch als expert in de theorie van optimaal stoppen moeten weten wanneer te stoppen met zijn kritiek.

Een geheel andere situatie ontstaat wanneer voor elke kandidaat de waardering kan worden uitgedrukt in een getal tot achter de komma precies op een schaal van 0 tot 1. Laten we deze situatie in de volgende context beschouwen. Stel eens dat iemand met een random-generator een van te voren vastgelegd aantal van R getallen tussen 0 en 1 genereert. Deze persoon laat je de getallen één voor één zien en wel in een random volgorde. Je wilt met maximale kans het grootste getal identificeren. Elke keer dat je een getal te zien krijgt, moet je meteen beslissen of je dat getal als grootste getal aanmerkt en dan stopt of dat je het getal laat passeren en vraagt een volgende getal te laten zien. Je kunt niet op een eerdere beslissing terugkomen. Welke strategie moet je volgen? Een optimale strategie is vrij lastig te berekenen, maar je kunt een simpele beslisregel geven die vrijwel optimaal is. Stel dat je een getal, zeg getal a, te zien krijgt terwijl nog k getallen getoond kunnen worden en dit getal a is het grootste getal dat je tot dan gezien hebt. De simpele beslisregel zegt dan dit getal a te nemen alleen dan als $a^k \leq 0{,}5$, oftewel als de kans dat onder de k resterende getallen een nog groter getal zit niet meer dan 0,5 is. Onder deze heuristiek is de kans dat je het grootste getal vindt 60,5% als $R = 10$ en 58,9% als $R = 25$, zoals met simulatie kan worden nagegaan. De kans convergeert naar ongeveer 58% als R toeneemt.

Het datingprobleem heeft vanzelfsprekend ook de nodige aandacht getrokken van cognitieve psychologen. Zo kwam voor het datingprobleem waarin alleen relatieve waarderingen aan de kandidaten kunnen worden toegekend, de Duitse psycholoog Peter Todd met het 'magische' getal 12: bij honderd of meer kandidaten, laat de eerste 12 passeren en neem dan de eerste die beter is dan alle voorgaande wanneer je doel is om met maximale kans een partner uit de top 10% te vinden. Voor deze vuistregel hoef je niet het precieze aantaal kandidaten te weten. Voor het datingprobleem waarin absolute waarderingen aan de kandidaten kunnen worden toegekend, kwam onlangs de zogenaamde wortelregel volop in het nieuws: bij R kandidaten, laat de eerste $\sqrt{R}$ - 1 passeren en neem dan de eerste die beter is dan alle voorgaande. Niet bepaald een regel die aan te bevelen valt wanneer je met maximale kans de allerbeste partner wilt vinden: voor $R = 25$ bijvoorbeeld geeft deze vuistregel een kans van ongeveer 31% om de allerbeste partner te vinden, terwijl de maximale kans ongeveer 59% is. Zou er binnenkort een datingbureau met de naam Square-Root Dating, op de markt verschijnen, houdt u er verre van!

Henk Tijms is emeritus hoogleraar operations research aan de Vrije Universiteit en auteur van diverse leerboeken over operations research en kansrekening. E-mail: h.c.tijms@xs4all.nl



Illustratie: David Gerstein

DE VERDWAALDE KOE

Martijn van Ee

Een koe staat voor een hek en wil naar de andere kant. Er is maar één doorgang. De koe weet niet waar de doorgang is en het is ook nog donker. En het is wel een heel lang hekwerk. Wat is de snelste manier om de doorgang te vinden? Welke kant moet de koe oplopen? Wanneer moet ze draaien? Het zogenoemde Lost-Cow-probleem. De auteur geeft een karakterisatie voor symmetrische verdelingen waarvoor het optimaal is om niet te draaien. De uitdaging is om dit uit te breiden naar niet-symmetrische verdelingen. Ook is gekeken naar de approximatie.

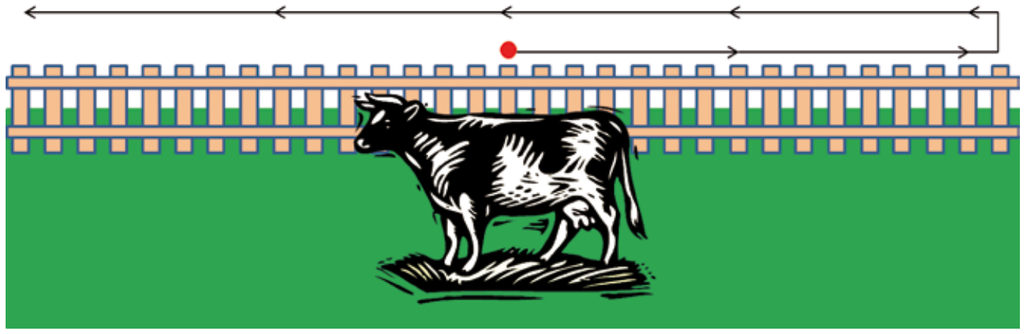
Het Lost-Cow-probleem kent twee versies. In de versie van de verdwenen koe is er een koe weggelopen van de boerderij. De koe bevindt zich ergens langs een oneindig lange weg. De boer wil de koe zo snel mogelijk vinden. In deze bijdrage benaderen we het Lost-Cow-probleem als het probleem van de verdwaalde koe. De situatie is als volgt. Er staat een koe voor een oneindig lang hek. In het hek zit een poort naar het andere weiland, waar de koe graag wil grazen. Helaas weet ze niet waar de poort zit. Bovendien is het mistig en kan ze alleen het stuk hek zien wat zich recht voor haar bevindt. Het doel van de koe is uiteraard om zo snel mogelijk de poort te vinden.

Keuzes

Ik heb tot op dit moment nog niets gezegd over hoe ik de koe ga beoordelen op haar keuzes. In de literatuur worden twee manieren aangedragen. In het eerste, waarbij het probleem meestal wordt genoemd als het Cow-Path-probleem, kiest een kwaadwillende tegenstander, na het waarnemen van de strategie van de koe, de positie van de poort (Baeza-Yates et al., 1993). De kwaliteit van de strategie wordt gemeten door de *competitive ratio*. Dit is de verhouding tussen de lengte van de wandeling tot het vinden van de poort en de afstand tussen het beginpunt en de poort. De koe kiest een strategie die deze ratio minimaliseert, terwijl de tegenstander met zijn keuze de ratio maximaliseert. Het blijkt dat een ratio van 9 het beste is dat de koe kan bereiken (Baeza-Yates et al., 1993). In de optimale strategie loopt de koe 1 meter naar rechts, keert terug naar het beginpunt, loopt 2 meter naar links, keert terug, loopt 4 meter naar rechts, enzovoort. De tegenstander plaatst de poort dan net na een van de keerpunten en ver weg van het beginpunt.

Bij de tweede manier, het *Linear Search* probleem (Bellman, 1963), heeft de koe de beschikking over een kansverdeling. Het doel is nu om de verwachte lengte van de wandeling tot het vinden van de poort te minimaliseren. In het geval dat de kansverdeling discreet is en een begrensd domein heeft, is het probleem op te lossen met behulp van dynamisch programmeren (Afrati et al., 1986). Verder zijn er resultaten bekend voor specifieke verdelingen zoals de uniforme verdeling, de normale verdeling en de driehoeksverdeling (Beck & Beck, 1984). In de literatuur is ook aandacht voor de *competitive ratio*. Sterker nog, Beck en Newman (1970) toonden al aan dat een factor 9 het best mogelijke resultaat is en gaven de eerdergenoemde verdubbelingsstrategie.

Als ik dit probleem aan anderen uitleg, krijg ik vaak de oplossing: ‘Loop naar het ene uiteinde van het hek en dan naar het andere uiteinde.’ Deze strategie is uiteraard alleen zinnig als het hek een eindige lengte heeft. Het deed mij afvragen voor welke kansverdelingen deze strategie optimaal is. Kella (1993) stelde zich dezelfde vraag en gaf een voldoende voorwaarde. Aangezien de



Figuur 1. Een niet-draaiende koe, waarbij de koe start bij het rode punt en de pijlen volgt

koe bij deze strategie geen extra draai maakt, maar alleen de noodzakelijke draai bij het uiteinde, zeg ik dat de koe hier niet draait. Zie figuur 1.

Welke koeien draaien niet?

Laten we ons direct beperken tot symmetrische verdelingen. We krijgen een functie $F(x)$, gedefinieerd voor $0 < x \leq 1$, die ons de kans geeft dat de poort rechts van het beginpunt en binnen een afstand x zit. Door symmetrie geldt dezelfde functie voor de linkerzijde en er geldt $2F(1) = 1$. Kella (1993) bewees dat als de functie $K(x) = x(1 - F(x))/2F(x)$ niet-stijgend is, het optimaal is om niet te draaien. De functie $K(x)$ maakt duidelijk een afweging tussen de afstand en de kans, iets wat je zou kunnen verwachten. Dit resultaat kan gegeneraliseerd worden naar algemene verdelingen en meerdere richtingen. Ik zal nu een karakterisatie geven voor symmetrische verdelingen en die generaliseren naar meerdere richtingen. Hierbij gebruik ik de notatie $E[D]$ om de verwachte afstand tussen het beginpunt en de poort weer te geven. Merk op dat F niet noodzakelijk een continue functie is.

Allereerst moet er gelden dat de beste oplossing met precies één draaipunt niet beter is dan de oplossing die niet draait. Als de koe draait op afstand y van het beginpunt, $0 < y < 1$, is de verwachte lengte van de wandeling $E[D] + y + (2y + 2)(1/2\text{-}F(y))$. De oplossing die niet draait heeft verwachte lengte $E[D] + 1$. Als we de bovengenoemde eis formeel opschrijven en herschrijven zien we dat er moet gelden dat $F(y) \leq y/(y + 1)$ voor alle $0 < y < 1$.

Merk op dat de functie $H(x) = x/(x+1)$ ook een geldige keuze voor $F(x)$ is en dat we dus kunnen kijken wat de koe moet doen bij deze verdeling. Omdat $H(x)$ tweemaal differentieerbaar is, kunnen we eerstejaars calculus gebruiken om dit te analyseren. Het blijkt dat we de verwachte waarde van een oplossing met m draaipunten kunnen uitdrukken als een functie van deze draaipunten. Als we nu de tweede afgeleide naar het m-de draaipunt nemen, zien we dat deze gelijk is aan 0. Dit betekent dat de optimale waarde, de verwachte lengte van de wandeling, een niet-dalende functie is van het aantal draaipunten. Hieruit kunnen we concluderen dat het voor $H(x)$ optimaal is om niet te draaien.

Als laatste moeten we bewijzen dat als we een functie $F(x)$ hebben waarvoor het optimaal is om niet te draaien, het ook optimaal is om niet te draaien voor functie $G(x)$ als geldt dat $G(x) \leq F(x)$ voor alle $0 < x \leq 1$. Anders gezegd, $G(x)$ heeft op elk moment een dikkere staart dan $F(x)$. Dit bleek eenvoudig te bewijzen met behulp van een bewijs uit het ongerijmde.

Uiteindelijk kunnen we dus concluderen dat het optimaal is om niet te draaien dan en slechts dan als $F(x) \leq H(x)$ voor alle $0 < x \leq 1$. In figuur 2 staan de functies $H(x)$ en de functies behorende tot de uniforme verdeling en de driehoeksverdeling getekend. Hieruit blijkt dat het voor de uniforme verdeling optimaal is om niet te draaien en dat dit niet zo is voor de driehoeksverdeling.

We kunnen de conditie $F(x) \leq H(x)$ herschrijven tot $K(x) \geq 1/2$. Merk nu op dat geldt $K(1) = 1/2$ en dat de eis van Kella, $K(x)$ moet niet-stijgend zijn, dus duidelijk een voldoende voorwaarde is. Dat deze voorwaarde niet noodzakelijk is, kan worden ingezien met behulp van het volgende voorbeeld. Laat $F(x) = x/2$ voor $0 < x \leq 1/2$, $F(x) = 1/4$ voor $1/2 < x \leq 3/4$ en $F(x) = x - 1/2$ voor $3/4 < x \leq 1$. In figuur 3 is nu duidelijk te zien dat geldt $K(x) \geq 1/2$, en dus ook $F(x) \leq H(x)$, maar dat $K(x)$ wel stijgend is op het interval $(1/2,3/4]$.

Uitbreidingen

De resultaten kunnen gegeneraliseerd worden naar meerdere richtingen. Als we het perspectief van de verdwaalde koe aanhouden, loopt de metafoor een beetje krom. Als we het perspectief van de verdwenen koe aanhouden, kan deze generalisatie geïnterpreteerd worden als de situatie waarin de boer op een kruispunt staat met k richtingen.

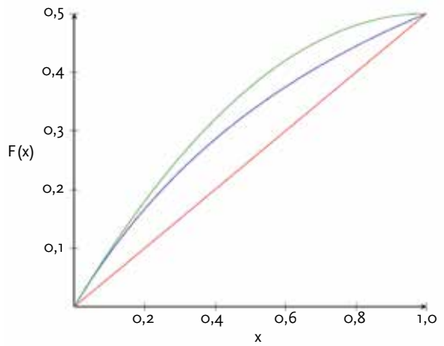
Op dezelfde manier als in de voorgaande paragraaf kan aangetoond worden dat het optimaal is om niet te draaien dan en slechts dan als $F(x) \leq x/(x + k - 1)$. Dit kan wederom herschreven worden in relatie tot de stelling van Kella. We krijgen nu $K(x) \geq (k - 1)/k$. Het valt op dat ook nu geldt dat het optimaal is om niet te draaien als geldt dat $K(x) \geq K(1)$. Momenteel onderzoek ik of vergelijkbare resultaten gelden voor meer algemene kansverdelingen.

Approximatie

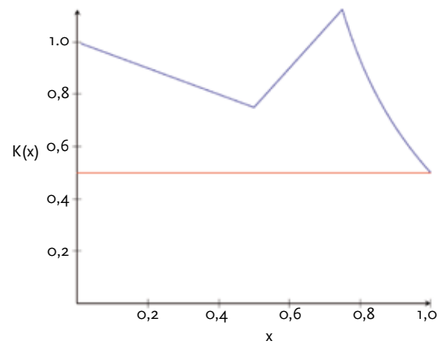
Alhoewel het voor de koe optimaal kan zijn om wel te draaien, geeft niet draaien voor sommige kansverdelingen een oplossing dichtbij het optimum. Theoretisch is daar het volgende over te zeggen. Het is duidelijk dat er geldt dat de verwachte lengte van de wandeling minstens $E[D]$ moet zijn. Met iets meer werk kan aangetoond worden dat $2E[D]$ een ondergrens is voor de optimale waarde indien de kansverdeling symmetrisch is. De verwachte waarde van niet draaien gelijk is aan $E[D] + 1$. De ratio tussen deze waarde en de ondergrens is dus

$$\frac{E[D]+1}{2E[D]} = \frac{1}{2}\left(1 + \frac{1}{E[D]}\right).$$

Hieruit blijkt dus dat als de poort maar ver genoeg staat, het niet heel erg is om niet te draaien.



Figuur 2. Functies H(x) (blauw), uniforme verdeling (rood) en driehoeksverdeling (groen)



Figuur 3. *K(x)* behorende bij beschreven *F(x)*

Literatuur

Afrati, F., Cosmadakis, S., Papadimitriou, C., Papageorgiou, G., & Papakostantinou, N. (1986). The complexity of the travelling repairman problem. *RAIRO Theoretical Informatics and Applications - Informatique Théorique et Applications, 20*, 79–87.

Baeza-Yates, R., Culberson, J., & Rawlins, G. (1993). Searching in the plane. *Information and Computation, 106*, 234–252.

Beck, A., & Beck, M. (1984). Son of the linear search problem. *Israel Journal of Mathematics, 48*, 109–122.

Beck, A., & Newman, D. (1970). Yet more on the linear search problem. *Israel Journal of Mathematics, 8*, 419–429.

Bellman, R. (1963). Problem 63-9, an optimal search. *SIAM Review, 5*, 274.

Kella, O. (1993). Star search – a different show. *Israel Journal of Mathematics*, *81*, 145–159.

Martijn van Ee is promovendus aan de Vrije Universiteit Amsterdam, begeleid door René Sitters en Leen Stougie. Zijn onderzoek richt zich op complexiteit en approximatie van problemen in de stochastische combinatorische optimalisatie. E-mail: m.van.ee@vu.nl

> mini_hackathon

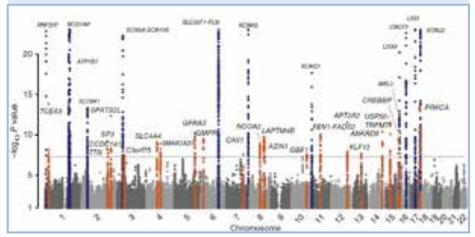
sectie data science.

OP ZOEK NAAR GENEN MET EFFECT OP HARTRITMESTOORNISSEN

Tijdens de jaarlijkse Dag voor Statistiek en OR van de VvS+OR had de onlangs gelanceerde Sectie Data Science (SDS) iets nieuws in petto voor een gezelschap data scientists. Zij werden uitgenodigd om met een fantastisch uitzicht vanaf de elfde verdieping van het stadskantoor in Utrecht – beschikbaar gesteld door de gemeente Utrecht – mee te doen aan de eerste SDS mini-hackathon. Dat betekent in één ochtend in teamverband creatieve oplossingen vinden voor een complex dataprobleem.

DE UITDAGING

Hartaanvallen zijn een veelvoorkomende doodsoorzaak, dus het voorkomen hiervan zou een belangrijke bijdrage aan de maatschappij betekenen. Door genetische data slim te combineren heeft het cardiovasculaire genetica team van het UMC een dataset gemaakt die eigenschappen van genen linkt aan een hartritmestoornis. Is het mogelijk om een patroon te achterhalen dat voorspelt of een stukje DNA belangrijk is voor deze stoornis zonder daar specifieke experimenten voor uit te hoeven voeren?



Volledig in lijn met het *Health care for the future* thema van de statistische dag werd de groep van cardioloog prof. dr. Folkert Asselbergs van het UMC Utrecht uitgenodigd een uitdaging op te stellen. Deze groep heeft ingezien dat er ontzettend veel data beschikbaar zijn en dat deze heel goed hergebruikt kunnen worden om onderzoek te doen naar de genetische fundamenten van hart- en vaatziekten. Samen met dr. Vinicius Tragante, hoofd van het datascience team uit deze groep, werd een grote genetische dataset geprepareerd om een brandende vraag te beantwoorden: kunnen we uit bestaande data genetische varianten identificeren die van invloed zijn op hartritmestoornissen?

Omdat niet iedereen een expert is op het gebied van cardiovasculaire genetica kregen de data scientists in de vroege morgen na een kopje koffie meteen een *crash course* in genetica van dr. Tragante. Zo werd iedereen ondergedompeld in de wereld van DNA, RNA, SNPs, eQTLs, en andere afkortingen (zie figuur).

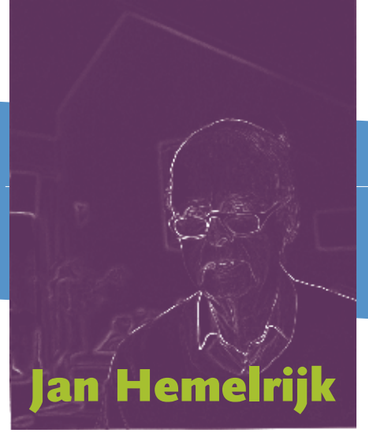
Hierna begonnen de teams vol goede moed aan hun uitdaging. Het doel: locaties in het genoom identificeren welke het meest relevant zijn voor de hartritmestoornis waar dr. Tragante onderzoek naar doet. Na uitgebreide dataverwerking, veel samenwerking, en – zoals het een hackathon betaamt – een korte lunch, konden de teams hun resultaten uploaden en vergelijken.

De gehele groep, zowel bestaande VvS+OR-leden als andere geïnteresseerden, werd vervolgens uitgenodigd om de rest van de Dag voor Statistiek en OR mee te beleven. De resultaten van de mini-hackathon werden hier nog eens in de schijnwerpers gezet door een plenaire presentatie van de SDS-voorzitter Daniel Oberski. De resultaten toonden uiteenlopende ideeën en methoden, van imputatie en enkelvoudige lineaire regressie tot Xgboost met *moving averages*.

Al met al heeft de mini-hackathon data scientists en statistici kennis laten maken met interessante nieuwe data en de VvS+OR. Daarnaast heeft het bijgedragen aan vervolgonderzoek door het cardiovasculaire genetica-team van het UMC Utrecht, dat geïnspireerd raakte door de resultaten. Een geslaagde start van de nieuwe sectie SDS, die de ambitie heeft haar rol binnen data science in Nederland in de nabije toekomst flink uit te breiden.

Erik-Jan van Kesteren

Met dank aan Martin Jansen (Gemeente Utrecht), Vinicius Tragante & Folkert Asselbergs (UMC Utrecht), Laurence Frank (VvS+OR) en alle deelnemers aan de mini-hackathon.



Jan Hemelrijk

Fred Steutel

column

Onlangs schreef ik een column over Van Dantzig onder de titel 'Professor van Dantzig'. Mijn verhouding met professor Hemelrijk was veel minder formeel, al geloof ik niet dat ik hem ooit Jan genoemd heb. Over Hemelrijk is al veel geschreven. Bij zijn overlijden in 2005 schreef Chris Klaassen een In Memoriam voor *Mededelingen* van het Korteweg-De Vries Instituut van de UvA. Voor *STAtOR* blikte Han Oud terug op zijn leven (*STAtOR* 2005, 1-2). Wikipedia schrijft kort en bondig: 'Jan Hemelrijk was een Nederlandse verzetsstrijder en statisticus. Tijdens de Tweede Wereldoorlog hield Hemelrijk zich bezig met het vervalsen van persoonsbewijzen en distributiebonnen.'

Hier geef ik alleen wat persoonlijke herinneringen. Eerst nog een curiositeit: Hemelrijk heeft een aantal dingen gemeen met mijn schoonmoeder. Zij regelde vanuit Laren (NH) adressen en bonkaarten voor joodse onderduikers. Zij kreeg het verzets-herdenkingskruis en er is voor haar een boom geplant in Israël. Vita en ik zien dagelijks onze namen op de servetringen die we als huwelijksgeschenk kregen van dankbare onderduikers. Ik ontmoette Hemelrijk voor het eerst toen ik begin twintig was, mijn vader was kort daarvoor overleden. Hemelrijk was geen tweede vader voor mij maar deze omstandigheid heeft mijn verhouding met hem toch enigszins gekleurd. Ik denk dat ik destijds op het Mathematisch Centrum (nu CWI) door hem ben aangenomen; dat ik niets van statistiek wist bleek geen bezwaar. Hij was lid van mijn promotiecommissie; ik zei iets doms tegen hem; niemand weet dat nog.

Hemelrijk was joods en had met goede redenen een hekel aan Duitsers. Ik herinner me dat een Duitse statisticus op bezoek kwam op het Mathematisch Centrum, toen een brandpunt op het terrein van de parametervrije (verdelingsvrije) statistiek. Het was een vrij jonge en heel vriendelijke jonge man. Hemelrijk wilde hem niet ontvangen. Ik heb nog mijn best gedaan, maar ik geloof niet dat het gelukt is. Ik denk ook niet dat hij ooit naar het Mathematisches Forschungsinstitut Oberwolfach in het Zwarte Woud is geweest, terwijl hij daarvoor ongetwijfeld uitgenodigd moet zijn.

Toen Hemelrijk met pensioen ging, heeft hij een grote stapel boeken weggegeven. Er waren veel proefschriften bij. Ik heb toen één boekje meegenomen dat ik nog af en toe tegenkom. Het is het proefschrift van J. Haeck, met de intrigerende titel *Colonization of the mole (talpa Europaea) in the IJsselmeerpolders.* Het heeft als ondertitel *Al vroetende komt men er door* – met een v, en met een plaatje van een mol. Promotor van Haeck is prof. dr. D. J. Kuenen, coreferent is dr. H. N. Kluyver, wiens (ere)promotor ik wel eens heb proberen te achterhalen – zonder succes. Dit boekje is mijn laatste tastbare herinnering aan Jan Hemelrijk.

Fred Steutel is emeritus hoogleraar kansrekening aan de TU Eindhoven.

Fred Steutel overleden

Onze vaste columnist en oud-redacteur Fred Steutel is op 1 juni j.l. onverwacht overleden. Hij was 85 jaar oud, maar dat was niet aan hem te merken. Nog geen twee weken daarvoor hadden we als redactie ons jaarlijkse etentje en waarschuwde hij ons voor de keus van cheese cake als dessert. Volgens hem had die term een suspecte betekenis: *Pictorial art or photograph of sexually attractive, nude or seminude women*. Dit was Fred ten voeten uit, een ongelooflijk rijke en soms bizarre feitenkennis die hij regelmatig in een grap verpakte. De familie plaatste op de rouwkaart de volgende tekst: 'Tot ons grote verdriet is helaas onverwacht maar vredig ingeslapen mijn lieve, erudiete, geestige, sportieve, zorgzame man en onze vader. Wij zullen hem zo vreselijk missen en hadden nog veel langer van hem willen genieten.' De redactie van *STAtOR* kan zich alleen maar volmondig bij deze kwalificaties aansluiten. Ook wij zullen hem, net als zijn familie, vreselijk missen. In het volgende nummer zullen we hem uitgebreid herdenken.

Jacques Benders (1924–2017)

Op 9 januari 2017 overleed op 92-jarige leeftijd de wiskundige Jacques Benders, hoogleraar Operations Research aan de Technische Universiteit Eindhoven. Zijn naam zal in de operationele research bekend blijven, vooral in de samenstelling *Benders decompositie*.

Jacques Benders werd geboren in Swalmen in 1924. Hij volgde de HBS aan het Bisschoppelijk College in Roermond. Uit interesse voor de wiskunde verwierf hij enkele onderwijsakten en gaf wiskundeles. Niet van plan in het bedrijf van zijn vader te werken, vertrok hij na de oorlog naar Utrecht, om daar in 1946 aan de universiteit te studeren. Daar was net Hans Freudenthal aangesteld als hoogleraar. Benders studeerde in 1952 af in de wiskunde, mechanica en natuurkunde.

Hij begon toen een carrière in de industrie bij de Rubbercompagnie in Delft, waar hij met statistische methoden de slijtage van rubber banden onderzocht. Later vond hij werk in het Shell laboratorium in Amsterdam en paste zijn kennis toe op logistieke problemen en de planning van de raffinage. Het was via de Amerikaanse tak van Shell dat Benders en zijn collega Zoutendijk op de hoogte kwamen van de ontwikkelingen in de operationele research. Hoewel collega-onderzoekers als Van Dantzig, Von Neumann en Koopmans veel publiceerden over de ontwikkelingen in de lineaire programmering, de *simplex method*, dualiteit, was daarvan in Nederland weinig doorgedrongen. Op bezoek bij hun collega's in California leerden Benders en Zoutendijk snel bij en begonnen ze hun eigen bijdragen te leveren aan de wetenschap. Zoutendijk hield zich bezig met continue niet-lineaire problemen, terwijl Benders zijn focus legde op problemen waarin geheeltalligheid van zekere variabelen een complicerende rol speelde.

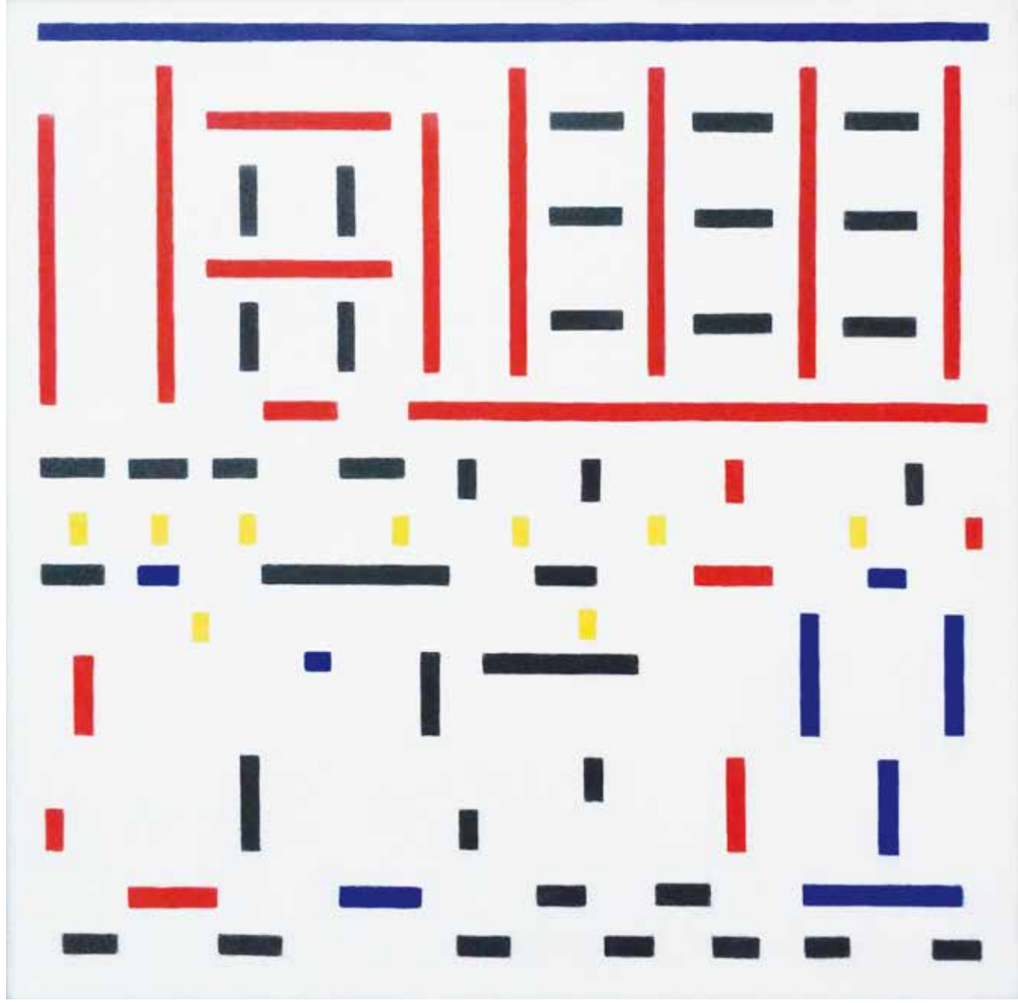
Terug in Nederland leidde dit werk tot een promotie, op 4 juli 1960, aan de universiteit van Utrecht, onder de supervisie van Hans Freudenthal. Het proefschrift van Benders was getiteld *Partitioning in Mathematical Programming* en werd gepubliceerd in *Numerische Mathematik* (1962, vol. 4, pp. 238–252). Toendertijd gold het OR-team van Shell in Amsterdam als een van de beste wereldwijd, en uniek voor Europa. Er bestond, ook vanuit Shell, behoefte om kennis van dit nieuwe gebied onder de aandacht te brengen in het wiskundecurriculum. Een eerste bijdrage werd door Benders gegeven in colleges in Utrecht in 1960–1961.

In Eindhoven werd tezelfdertijd gebouwd aan een technische universiteit. Jaap Seidel was daar onder andere bezig de onderafdeling der wiskunde te voorzien van een eigen opleiding voor de wiskundig ingenieur. Door zijn inspanningen werd Benders verleid tot het aannemen van de positie van de eerste hoogleraar in Nederland op het gebied van de mathematische programmering. In de zomer van 1964 verhuisde de familie Benders weer naar het zuiden des lands. In zijn intreerede kwalificeerde Benders *wiskundig gereedschap* (in de vorm van algoritmische concepten) van dezelfde statuur als moderne technologisch installaties. Bovendien stelde hij dat de wiskundig ingenieur van toegevoegde waarde zou zijn in technische ontwikkelingen, en dat hiervoor zowel creativiteit als kunde vereist waren.

Hij heeft zo'n 70 afstudeerders begeleid, honderd projecten uitgevoerd en vijf promovendi afgeleverd. Hij stak veel tijd en moeite in de opzet van het curriculum van de wiskundige ingenieur, en het overbrengen van kennis en inzicht aan collega's en studenten op het gebied van de operations research. Daarnaast stak hij veel energie in het overeind houden van de afdeling en de invoering van een goede ict-ondersteuning aan de universiteit. Het nadeel van al deze inspanningen is dat zijn bijdrage aan de literatuur verstopt is in de geschriften van zijn pupillen. De korte lijst van publicaties heeft wel geleid tot een groot aantal citaties. Vooral in de recente literatuur is er een groeiende aandacht, omdat gegeven de schaal van probleemgroottes en van computergeheugens de decompositiemethoden van Benders juist meer toepasselijk zijn.

De erkenning voor het werk van Benders is gelukkig op tijd gegeven in de vorm van de EURO Gold Medal die hem in 2009 is toegekend.

Cor Hurkens
E-mail: c.a.j.hurkens@tue.nl



Bart van der Leck, *Compositie 1917 no. 4 (uitgaan van de fabriek)*, 1917. Collectie Kröller-Müller Museum, Otterlo

INTRODUCTION TO BENDERS' DECOMPOSITION

Kees Roos

Benders was the first professor in The Netherlands in the field of Operation Research. Googling with the key 'Benders' decomposition' yields as many as 75.300 hits. Indeed, his decomposition method has been widely used in many applications. His path-breaking paper (Benders, 1962) on the method was reviewed by E.M.L. Beale who already recognized it as an *important but difficult paper that presents a principle that can be regarded as the dual to the Dantzig-Wolfe decomposition principle* (Beale, 1963). The paper was republished (Benders, 2005) with a foreword by István Maros (Maros, 2005) to honor Benders at the occasion of his 80th birthday. He was awarded the prestigious EURO Gold Medal at EURO 2009 in Bonn.

Benders became famous for the way he dealt in his paper with the problem

$$\max_{x,y} \{a^T x + f(y) : Ax + F(y) \leq c,\ x \geq 0,\ y \in S\}. \qquad (1)$$

Here S is an arbitrary subset of $\mathbb{R}^p$, with $\mathbb{R}^p$ denoting the p-dimensional Euclidean space. Furthermore, A is an $m \times n$ matrix, $f(y)$ a scalar function and $F(y)$ a vector function from S to $\mathbb{R}^m$, whereas $a \in \mathbb{R}^n$ and $c \in \mathbb{R}^m$ are fixed vectors.

In this paper we discuss Benders' method for the special case of (1) that arises when the functions F and f are linear and $S = \mathbb{R}^p_+$. By taking $F(y) = By$ for some $m \times p$

matrix B and $f(y) = b^T y$ for some vector b in R^p, (1) becomes the following linear optimization (LO) problem:

$$(P) \quad \max_{x,y} \{a^T x + b^T y : Ax + By \le c,\ x \ge 0,\ y \ge 0\}.$$

Of course, this problem can be solved straightforwardly by, e.g., the simplex method. But in many cases, especially when the matrices A and/or B have a special structure, Benders' decomposition method provides an attractive alternative for solving (P). This is certainly true if, e.g., optimization over the region

$$\boldsymbol{B} = \{u : B^T u \ge b,\ u \ge 0\} \tag{2}$$

is relatively easy. As Benders mentions, in the linear case his method is a special version of the decomposition method for linear optimization of Dantzig and Wolfe (1960). We provide a simple proof that Benders' method solves (P) in finitely many steps. When modelling a real-life problem as an LO problem there is something wrong with the model if it has no optimal solution, i.e., if the LO problem is infeasible or unbounded. So in principle one is interested only in LO problems that have an optimal solution. Interestingly, Benders' method uses subproblems that may lack this property. In a clever way, however, Benders' method explores the information provided by such subproblems. To understand this one needs to be familiar with duality theory for linear optimization. We therefore shortly recall this theory, in the next section.

Prerequisites

To start with we consider the LO problem

$$\min_x \{c^T x : Ax \ge b,\ x \ge 0\}, \tag{3}$$

where A, b and c are arbitrary and of appropriate sizes. Moreover, $x \ge 0$ means that the entries of vector x are non-negative and $Ax \ge b$ that $Ax - b \ge 0$. The *dual* problem of the minimization probem (3) is the maximization problem

$$\max_y \{b^T y : A^T y \le c,\ y \ge 0\}. \tag{4}$$

Problem (3) is called the *primal problem*. If x is feasible for (3) and y for (4), then $x \ge 0$, $y \ge 0$, $Ax \ge b$ and $A^T y \le c$. As a consequence we may write

$$b^T y \le (Ax)^T y = x^T (A^T y) \le c^T x.$$

In words, every dual objective value is a lower bound for every primal objective value. This property is known as *weak duality*. An immediate consequence is that if $b^T y$ attains arbitrarily large values then (3) must be infeasible. We say that (4) is unbounded in that case. Similarly, if $c^T x$ attains arbitrarily small values, i.e., if (3) is unbounded, then (4) must be infeasible.

The weak duality property implies that the so-called *duality gap* $c^T x - b^T y$ for any feasible pair (x,y) is nonnegative. Moreover, if it vanishes, then x is optimal for (3) and y is optimal for (4). The Duality Theorem for LO states that the converse also holds: if x is optimal for (3) and y is optimal for (4) then the duality gap vanishes, and hence $c^T x = b^T y$.

Finally, if (3) is infeasible then the system $Ax \ge b, x \ge 0$ has no solution. According to Farkas' lemma this happens if and only if there exists a vector $\bar{y} \ge 0$ such that $A^T \bar{y} \le 0$ and $b^T \bar{y} > 0$. Such a vector $\bar{y}$ is known as a *direction of recession*, or *infinite direction* of the dual problem (4). If (4) has a feasible solution y and an infinite direction $\bar{y}$ then $y + \lambda\bar{y}$, with $\lambda \ge 0$, will be feasible as well for (4); moreover, $b^T(y + \lambda\bar{y}) = b^T y + \lambda\, b^T\bar{y}$ goes to infinity if λ grows to infinity. We conclude that if (3) is infeasible then (4) has an infinite direction, and then (4) is either infeasible or unbounded. In a similar way it follows that if (4) is infeasible then (3) has an infinite direction, and (3) is either infeasible or unbounded.

If an LO problem has an optimal solution we call the problem *solvable*.

Key ingredients of Benders' decomposition method

We now describe Benders' decomposition method for the solution of the problem (P), which we write down once more:

$$(P) \quad \max_{x,y} \{a^T x + b^T y : Ax + By \le c,\ x \ge 0,\ y \ge 0\}.$$

The dual problem (D) of (P) is given by

$$(D) \quad \min_u \{c^T u : A^T u \ge a,\ B^T u \ge b,\ u \ge 0\}.$$

By the duality theorem for LO we have $z(P)=z(D)$ if (P) and (D) are solvable, where $z(.)$ denotes the optimal value of any LO-problem. If problem (P) is not solvable, it is either infeasible or unbounded. In that case we write $z(P)=NaN$, and similar for $z(D)$. With this notational assumption we always have $z(P)=z(D)$. We now may write

$$\begin{aligned} z(P) &= \max_{x\ge 0} \{a^T x + \max_{y\ge 0} \{b^T y : By \le c - Ax\}\} \\ &= \max_{x\ge 0} \{a^T x + \min_{u\ge 0} \{(c - Ax)^T u : B^T u \ge b\}\} \\ &= \max_{x\ge 0} \min_{u\ge 0} \{a^T x + (c - Ax)^T u : B^T u \ge b\}, \end{aligned} \tag{5}$$

where the second equality is due to the Duality Theorem for LO. Using the set $\boldsymbol{B}$ as defined in (2), we may rewrite (5) as follows:

$$z(P) = \max_{x\ge 0,\xi} \{\xi : a^T x + (c - Ax)^T u \ge \xi,\ \forall u \in \boldsymbol{B}\}. \tag{6}$$

The above maximization problem has a constraint for each $u \in \boldsymbol{B}$. So it is an LO problem with (usually) infinitely many constraints. As we will see in the next section, a relaxation of this problem occurs as a subproblem in Benders' method; the number of constraints in this relaxation is always finite. The other subproblem in Benders' method is an LO problem over the domain $\boldsymbol{B}$, with a changing objective function.

Obviously, if the set $\boldsymbol{B}$ is empty then the dual problem (D) of (P) is infeasible. It means that (P) has an infinite direction and hence (P) is either infeasible or unbounded. We therefore assume below that $\boldsymbol{B}$ is nonempty.

The algorithm

Benders' decomposition method can be described as in Algorithm 1. The algorithm uses a subset $\boldsymbol{S}$ of points in $\boldsymbol{B}$ and a set $\boldsymbol{R}$ of infinite directions of $\boldsymbol{B}$. Initially the set $\boldsymbol{R}$ is empty and $\boldsymbol{S}$ consists of an arbitrary point in $\boldsymbol{B}$. It is assumed that the optimal value of (P), if it exists, lies in the interval $[-M,M]$, where M is some large number, which may be taken ∞. We also use a vector $\bar{x}$ in the algorithm; initially this is an arbitrary non-negative vector.

The while loop starts with solving the first subproblem in the algorithm, which is formulated in line 3; it is the inner minimization problem in (5), with $x = \bar{x}$:

$$\min_u \{a^T\bar{x} + (c - A\bar{x})^T u : B^T u \ge b,\ u \ge 0\}. \tag{7}$$

1: INITIALIZE: $\mathcal{S} = \{u\}$ for some $u \in \mathcal{B}$; $\mathcal{R} = \emptyset$; $\underline{z} = -M$; $\bar{z} = M$; $\bar{x} \ge 0$
2: **while** $\bar{z} > \underline{z}$ **do**
3: Solve $\min_u \{a^T\bar{x} + (c - A\bar{x})^T u \ :\ u \in \mathcal{B}\}$
4: **if** this problem has an optimal solution u^* **then**
5: **if** $\underline{z} < a^T\bar{x} + (c - A\bar{x})^T u^*$ **then**
6: $\underline{z} := a^T\bar{x} + (c - A\bar{x})^T u^*$
7: **if** $\underline{z} = \bar{z}$ **then**
8: $\bar{x}$ is optimal solution for the x-part of (P); **return**
9: **else**
10: $\mathcal{S} = \mathcal{S} \cup \{u^*\}$
11: **else** it has an infinite direction v^*
12: $\mathcal{R} = \mathcal{R} \cup \{v^*\}$
13: Solve $\max_{x\ge 0,\xi} \{\xi \ :\ a^T x + (c - Ax)^T u \ge \xi,\ (c - Ax)^T v \ge 0,\ \forall u \in \mathcal{S},\ \forall v \in \mathcal{R}\}$.
14: **if** this problem has an optimal solution (x^*, ξ^*) **then**
15: $\bar{x} := x^*$
16: $\bar{z} := \xi^*$
17: **else if** it is unbounded **then**
18: find a feasible pair (x, ξ) such that $\xi \ge \bar{z}$ and put $\bar{x} = x$
19: **else** it is infeasible
20: (P) is infeasible or unbounded; **return**

Algorithm 1. Benders' decomposition method

Due to our assumption that $\boldsymbol{B}$ is nonempty, problem (7) either has an optimal solution u^* or it is unbounded.

If (7) has an optimal solution u^*, its value $a^T\bar{x} + (c - A\bar{x})^T u^*$ is a lower bound for the optimal value of (P), as follows from (5). If this lower bound is larger than the current lower bound $\underline{z}$ we replace $\underline{z}$ by $a^T\bar{x} + (c - A\bar{x})^T u^*$, as happens in line 10. If this makes the lower bound for the optimal value of (P) equal to the current upper bound we have found the optimal value of (P), and $\bar{x}$ is the x-part of an optimal solution for (P), as follows from (5). So in that case the algorithm stops in line 12. Otherwise we add u^* to the set $\boldsymbol{S}$, as in line 14.

On the other hand, in the unbounded case, its solution yields an infinite ray, i.e., a feasible point $\bar{u}$ and an infinite direction v^* such that[1]

$$a^T\bar{x} + (c - A\bar{x})^T(\bar{u} + \alpha v^*) = a^T\bar{x} + (c - A\bar{x})^T\bar{u} + \alpha(c - A\bar{x})^T v^*.$$

decreases to minus infinity if α increases to infinity. This implies $(c - A\bar{x})^T v^* < 0$. In line 7 the vector v^* is added to the set $\boldsymbol{R}$.

Next we solve the second subproblem, as given in line 13. As every LO problem it is either solvable, unbounded or infeasible. If it is solvable it has an optimal solution (x^*,ξ^*). We then replace $\bar{x}$ by x^* and the upper bound $\bar{z}$ by ξ^*, as in line 17–18. To justify this update recall from (6) that

$$z(P) = \max_{x\geq 0,\xi} \{\xi : a^T x + (c - Ax)^T u \geq \xi, \forall u \in \boldsymbol{B}\}. \qquad (8)$$

Since the set $\boldsymbol{B}$ is a polyhedron, its elements are obtained by taking the sum of any convex combination of its vertices and a nonnegative multiple of one of its infinite directions (Schrijver, 1986). In other words, every element of $\boldsymbol{B}$ can be written in the form $u + \alpha v$, where u is a convex combination of the vertices of $\boldsymbol{B}$, v an infinite direction of $\boldsymbol{B}$ and $\alpha \geq 0$. Hence, defining $S_{\boldsymbol{B}}$ as the convex hull of the vertices of $\boldsymbol{B}$ and $R_{\boldsymbol{B}}$ as the set of all infinite directions of $\boldsymbol{B}$, we may rewrite (8) as

$$\begin{aligned} z(P) = \max_{x\geq 0,\xi}\{\xi : a^T x + (c - Ax)^T(u + \alpha v) \geq \xi, \forall \alpha \geq 0, \\ \forall u \in S_{\boldsymbol{B}}, \forall v \in R_{\boldsymbol{B}}\}. \end{aligned}$$

$$\begin{aligned} = \max_{x\geq 0,\xi}\{\xi : a^T x + (c - Ax)^T u + \alpha(c - Ax)^T v \geq \xi, \\ \forall \alpha \geq 0, \forall u \in S_{\boldsymbol{B}}, \forall v \in R_{\boldsymbol{B}}\}. \qquad (9) \end{aligned}$$

If for some $v \in R_{\boldsymbol{B}}$ one has $(c - Ax)^T v < 0$ for all $x \geq 0$ then (9) is infeasible, because then $\alpha(c - Ax)^T v$, and hence also ξ goes to minus infinity if α grows to infinity and this implies that there is no feasible value for ξ. Hence, problem (9) will be solvable only if there exists a vector x for which $(c - Ax)^T v \geq 0$, for all $v \in R_{\boldsymbol{B}}$. This makes it clear that the problem in line 13 is a relaxation of (9) and hence also of (8). As a consequence, if the problem in line 13 it solvable its optimal value is an upper bound for $z(P)$.

If the problem in line 13 is unbounded, we may find x such that

$$a^T x + (c - Ax)^T u \geq \bar{z}, \forall u \in S$$

and we replace $\bar{x}$ by such an x, as in line 18.

Finally, the only reason for the problem in line 13 to be infeasible is that there exists a $v \in R_{\boldsymbol{B}}$ such that $(c - Ax)^T u < 0$ for all $x \geq 0$. As we established before, in that case the value of $z(P)$ becomes minus infinity, which means that (P) is infeasible or unbounded.

Analysis and finite convergence

We show in this section that the algorithm decides in a finite number of steps whether (P) is solvable, infeasible, or unbounded. In the first case it generates an optimal solution and in the two other cases it generates an infinite direction for the dual problem. Hereby it is assumed that the first subproblem is solved by the Simplex method, because then it yields either a vertex of $\boldsymbol{B}$ or an extreme infinite direction of $\boldsymbol{B}$.

First observe that if the algorithms stops this happens in line 2, in line 8 or in line 20. If it stops in line 2 or line 8 it provides the optimal value $z(P)$ and the x-part of an optimal solution for (P) and if it stops in line 20 then (P) is infeasible or unbounded.

We claim that if the algorithm does not stop during the execution of the while loop then it adds the vertex u^* of $\boldsymbol{B}$ to $\boldsymbol{S}$ if the first subproblem is solvable; if this problem is unbounded it adds the infinite direction v^* to $\boldsymbol{R}$. A crucial property of the algorithm is that it never adds an element to $\boldsymbol{S}$ that already belongs to $\boldsymbol{S}$, and the same holds for the set $\boldsymbol{R}$. This can be shown as follows.

If the first subproblem is solvable, and the algorithm does not stop in line 8, then we have $\underline{z} < z^-$. Due to this we may write

$$a^T\bar{x} + (c - A\bar{x})^T u^* \leq \underline{z} < \bar{z} \leq a^T\bar{x} + (c - A\bar{x})^T u, \ \forall u \in S.$$

This implies that $u^* \notin S$.

If the first subproblem (in line 3) is unbounded, then we have $(c - A\bar{x})^T v^* < 0$. However, due to the construction of $\bar{x}$ in line 15 or line 18, $\bar{x}$ is feasible for the second subproblem (in line 13). We therefore have $(c - A\bar{x})^T v \geq 0$ for every $v \in \boldsymbol{R}$. This proves that $v^* \notin \boldsymbol{R}$.

Thus we have shown that as long as the algorithm does not stop, in every iteration exactly one of the sets $\boldsymbol{S}$ and $\boldsymbol{R}$ is extended with a new element. From this property the finite convergence property easily follows, as we now show.

We focus on the first subproblem of the algorithm. If we use a Simplex method for solving this problem, in the solvable case it will generate a vertex u^* of the polyhedron $\boldsymbol{B}$, and add it to $\boldsymbol{S}$, and in the unbounded case an extreme infinite direction v^* of $\boldsymbol{B}$, adding it to $\boldsymbol{R}$. As we just showed no vertex or infinite direction will be generated twice. Since the number of vertices and the number of extreme infinite directions is finite, the sum of these numbers is an upper bound for the total number of iterations. Hence, when using a simplex method for solving the first subproblem, the finite convergence of the algorithm follows.

Example

We consider the case where

$$A = \begin{bmatrix} 2 & 3 \\ 3 & 2 \end{bmatrix},\ B = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix},\ a = \begin{bmatrix} 2 \\ 1 \end{bmatrix},\ b = \begin{bmatrix} 1 \\ 1 \end{bmatrix},\ c = \begin{bmatrix} 6 \\ 6 \end{bmatrix}. \qquad (10)$$

Below we use superscripts to number vectors and subscripts to number the entries in a vector.

ITERATION 1
We initialize with $\bar{x} = 0$, $M = 10$ and $S = \{u^1 = [1;1]\}$. We then have $a^T\bar{x} = 0$ and $c - A^T\bar{x} = c$. Hence the first subproblem becomes

$$\min_{x\geq 0} \left\{6u_1 + 6u_2 : \begin{matrix} u_1 + 2u_2 \geq 1 \\ 2u_1 + u_2 \geq 1 \end{matrix}\right\}.$$

Its optimal solution is $u^* = [1/3;\ 1/3]$ with $c^T u^* = 4$. So we get $\underline{z}$=4 and S = {u¹, u²} with u² = [1/3; 1/3].
We now have $a^T x = 2x_1 + x_2$ and

$$(c - Ax)^T u^1 = c^T u^1 - (u^1)^T Ax = 12 - 5(x_1 + x_2)$$
$$(c - Ax)^T u^2 = c^T u^2 - (u^2)^T Ax = 4 - 5/3(x_1 + x_2).$$

Hence, the second subproblem becomes

$$\max_{x\geq 0,\xi} \left\{\xi : \left\{\begin{matrix} (u^1) : 12 - 3x_1 - 4x_2 \geq \xi \\ (u^2) : 4 + 1/3x_1 - 2/3x_2 \geq \xi \end{matrix}\right\}\right\},$$

where the label (u^i) is used to indicate that the corresponding inequality is induced by the element u^i of S. This problem is solvable. Its optimal value is 24/5, which occurs at $x^* = [12/5;\ 0]$. Hence we put $\bar{x} = [12/5;\ 0]$ and $\bar{z} = 24/5$.

ITERATION 2
Using the results of the first iteration, we start the second iteration. We now have $a^T\bar{x} = 24/5$ and

$$c - A\bar{x} = \begin{bmatrix} 6 \\ 6 \end{bmatrix} - \begin{bmatrix} 2 & 3 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} 12/5 \\ 0 \end{bmatrix} = \begin{bmatrix} 6/5 \\ -6/5 \end{bmatrix}.$$

So the first subproblem becomes

$$\min_{u\geq 0} \left\{24/5 + 6/5u_1 - 6/5u_2 : \begin{matrix} u_1 + 2u_2 \geq 1 \\ 2u_1 + u_2 \geq 1 \end{matrix}\right\}.$$

This problem is unbounded. It admits

$$x = \begin{bmatrix} 1 \\ \alpha \end{bmatrix},\ \alpha \geq 0,$$

as a feasible solution, whose objective value equals 30/5 −6/5 α, which goes to minus infinity if α grows to infinity. Hence we add the vector $v^1 = [0;\ 1]$, which is an infinite direction, to the set $\boldsymbol{R}$. We have $c^T v^1 = 6$ and

$$(Ax)^T v^1 = (v^1)^T Ax = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 3x_1 + 2x_2.$$

Hence the second subproblem becomes

$$\max_{x\geq 0,\xi} \left\{\xi : \left\{\begin{matrix} (u^1) : 12 - 3x_1 - 4x_2 \geq \xi \\ (u^2) : 4 + 1/3\,x_1 - 2/3\,x_2 \geq \xi \\ (v^1) : 6 - 3x_1 - 2x_2 \geq 0 \end{matrix}\right\}\right\}.$$

This problem is solvable, with $x^* = [2;\ 0]$ as optimal solution and $\xi = 14/3$. We therefore put $\bar{x} = [2;\ 0]$ and $\bar{z} = 14/3$.

ITERATION 3
We now have $a^T\bar{x} = 4$ and

$$c - A\bar{x} = \begin{bmatrix} 6 \\ 6 \end{bmatrix} - \begin{bmatrix} 2 & 3 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

So the first subproblem becomes

$$\min_{u \geq 0} \left\{ 4 + 2u_1 : \begin{Bmatrix} u_1 + 2u_2 \geq 1 \\ 2u_1 + u_2 \geq 1 \end{Bmatrix} \right\}.$$

This problem is solvable. It has $u^* = [0; 1]$ as optimal solution, with objective value 4. Hence $\underline{z}$ remains 4, and we add the vector $u^3 = [0; 1]$ to S. We have

$$(c - Ax)^T u^3 = c^T u^3 - (u^3)^T Ax = 6 - [0 \ 1] \begin{bmatrix} 2 & 3 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 6 - 3x_1 - 2x_2.$$

Hence the second subproblem becomes

$$\max_{x \geq 0, \xi} \left\{ \xi : \begin{Bmatrix} (u^1) : 12 - 3x_1 - 4x_2 \geq \xi \\ (u^2) : 4 + 1/3 x_1 - 2/3 x_2 \geq \xi \\ (u^3) : 6 - x_1 - x_2 \geq \xi \\ (v^1) : 6 - 3x_1 - 2x_2 \geq 0 \end{Bmatrix} \right\}.$$

It turns out that $x^* = [3/2; 0]$ is optimal with value $\xi = 9/2$. Thus we update $\bar{x}$ and $\bar{z}$ accordingly: $\bar{x}$ $[3/2; 0]$ and $\bar{z} = 9/2$.

ITERATION 4
We now have $a^T\bar{x} = 3$ and

$$c - A\bar{x} = \begin{bmatrix} 6 \\ 6 \end{bmatrix} - \begin{bmatrix} 2 & 3 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} 3/2 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 \\ 3/2 \end{bmatrix}.$$

So the first subproblem becomes

$$\min_{x \geq 0} \left\{ 3 + 3u_1 + 3/2 u_2 : \begin{Bmatrix} u_1 + 2u_2 \geq 1 \\ 2u_1 + u_2 \geq 1 \end{Bmatrix} \right\}.$$

This problem is solvable. It has $u^* = [1/3; 1/3]$ as optimal solution, with objective value 9/2. Hence we obtain $\underline{z} = 9/2$ which coincides with $\bar{z}$. Hence the algorithm stops, telling us that $z(P) = 9/2$ and that the x-part of the optimal solution is $[3/2; 0]$.

Thus we have demonstrated Benders' decomposition method when the data of (P) is as in (10). We emphasize that the algorithm yields only the optimal objective value $z(P)$ of (P) and the x-part of an optimal solution. Finding the y-part of the solution is now easy. It can be found by solving the system

$$By \leq c - Ax^*, \quad b^T y = z(P) - a^T x^*, \quad y \geq 0$$

In the case of the example this leads to the system

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} y1 \\ y2 \end{bmatrix} \leq \begin{bmatrix} 3 \\ 3/2 \end{bmatrix}, \quad y_1 + y_2 = 3/2, \quad y_1 \geq 0, \quad y_2 \geq 0.$$

Though in general multiple solutions may exist, in the present case the solution is unique: $y_1 = 0$, $y_2 = 3/2$.

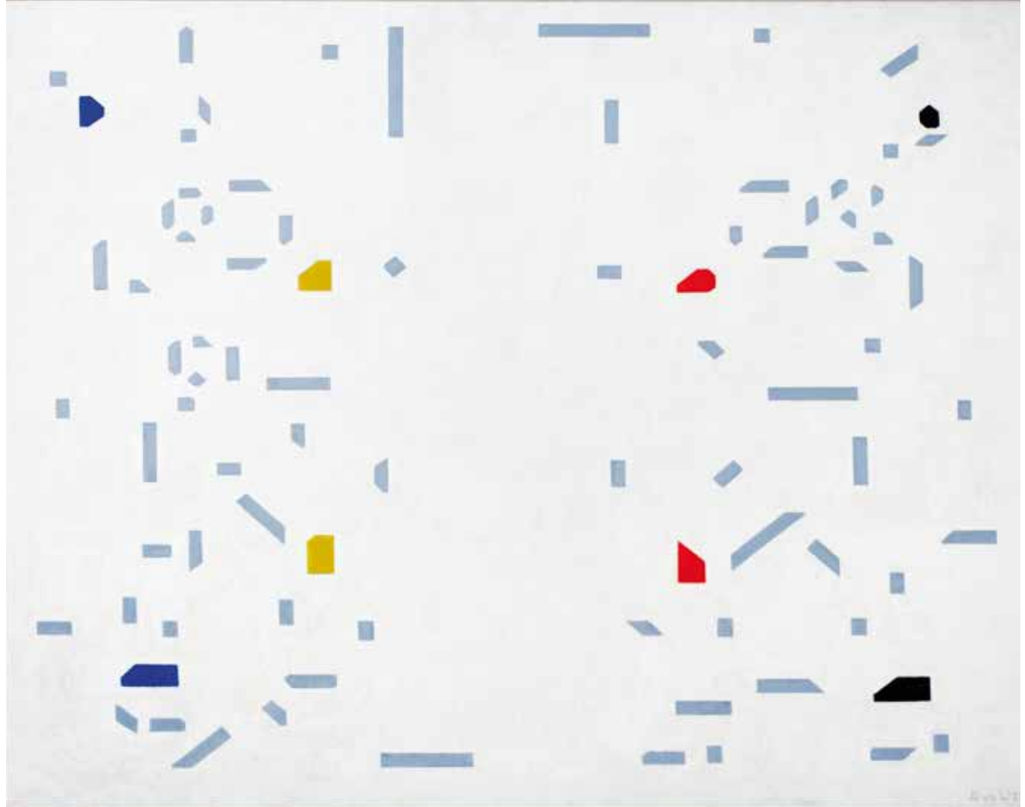
Acknowledgement
Thanks are due to dr. Bram Gorissen (VU) for his careful reading of an earlier version of this paper and his useful comments.

Note
1. In the Matlab implementation of the author he used the useful modelling package cvx, with Mosek as solver. This package yields in the unbounded case only an infinite direction. To obtain an feasible point, which may be needed when using Benders' method, one needs to solve an additional LO problem.

References

Beale, E. M. L. (1963) *Mathematical Reviews*, MR0147303 26 #4820.
Benders, J. F. (1962). Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik, 4*, 238–252.
Benders, J. F. (2005). Partitioning procedures for solving mixed-variables programming problems. *Computational Management Science, 2*(1), 3–19.
Dantzig, G. B., & Wolfe, Ph. (1960) Decomposition principle for linear programs. *Operations Research, 8*, 1, 101–111.
Geoffrion, A. M. (1972). Generalized Benders decomposition. *Journal of Optimization Theory and Applications, 10*, 237–260.
Hrouda, J. (1975). A modified Benders algorithm. *Ekonomomicko-Matematicke Obzor, 11*(4), 423–443
Johnson, E. L., & Suhl, U. H. (1980). Experiments in integer programming. *Discrete Applied Mathematics*, *2*, 1, 39–55.
Maros, I. (2005). Jacques F. Benders is 80. *Computational Management Science, 2*(1).
Rasmussen, R. V., & Trick, M. A. (2007) A Benders approach for the constrained minimum break problem. *European Journal of Operational Research, 177*(1), 198–213.
Schrijver, A. (1986). *Theory of Linear and Integer Programming.* New York: John Wiley & Sons, 1986.
Trick, M. A., & Yildiz, H. (2011). Benders' cuts guided large neighborhood search for the traveling umpire problem. *Naval Research Logistics, 58*(8), 771–781.

Kees Roos. Department of Electrical Engineering, Mathematics and Computer Science, Technical University Delft.
E-mail: c.roos@tudelft.nl



Bart van der Leck, *Houthakkers*, 928. Collectie Gemeentemuseum Den Haag

BENDERS' DECOMPOSITIE VOOR STATISTIEK

De Benders decompositie is een methode om lastige of grote problemen aan te pakken op basis van een geschikte partitie van de variabelen. In dit artikel beschouwen we een praktisch probleem waarbij de partitie zich vanzelf opdringt. Het betreft het zogenaamde secundaire cel-onderdrukkingsprobleem. Bij publicatie van geaggregeerde data over bedrijven of personen wil men soms voorkomen dat individuele gegevens uit de gepubliceerde data zijn af te leiden. Dit kan door betreffende getallen uit de tabel weg te laten. Maar er moeten vaak meer cellen weggelaten worden om te voorkomen dat de weggelaten data kunnen worden gereconstrueerd uit rij- en kolomtotalen. Het probleem is dus een minimale verzameling data op te offeren zodat gevoelige data alleen met een grote foutmarge kunnen worden benaderd.

Cor Hurkens

Introductie van decompositie in een data verhullingsprobleem

Benders heeft zijn decompositiemethode ontwikkeld rond de filosofie dat in een mathematisch programmeringsmodel de verzameling variabelen op een logische manier in twee delen uiteen kan vallen, waarbij het probleem een stuk eenvoudiger wordt wanneer de eerste verzameling variabelen gefixeerd is.

In het zogenaamde secundaire cel-onderdrukkingsprobleem is een tabel gegeven die numerieke data bevat die voortkomen uit surveys die bij bedrijven of personen zijn gehouden. In principe worden geaggregeerde data gepubliceerd voor het nut van iedereen. Zo kunnen

de rijen slaan op regio's, de kolommen op bedrijfstakken, en de waarden op jaaromzet in miljoenen euro's. Om te voorkomen dat uit de publicatie individuele data zijn te herleiden zijn cellen waarvan de bijdrage slechts door een of twee bedrijven is geleverd aan te merken als *gevoelig*. Zo'n gevoelige cel wordt dan in de tabel gemarkeerd met een asterisk * en de data worden weggelaten. Als rij-totalen en kolom-totalen zijn gegeven, en als we weten dat de getallen altijd niet-negatief zouden moeten zijn, is het wellicht mogelijk de weggelaten getallen te reconstrueren en alsnog de gevoelige data te kennen. Zelfs als we van niet-gevoelige cellen data weglaten bestaat de kans dat een gevoelige cel binnen een zekere range kan worden teruggerekend. Gegeven dat er K cellen zijn weggelaten is er een consistente invulling met waarden (x_1, ..., x_K), met x_1 minimaal, en een consistente invulling met waarden (y_1, ..., y_K), met y_1 maximaal. Als $y_1 - x_1$ maar groot genoeg is dan vinden we cel 1 genoeg *beschermd*, anders niet. Bepaling van minimum en maximum komt neer op het oplossen van een lineair programmeringsprobleem in $2K$ variabelen. Het aantal lineaire beperkingen hangt af van de posities van de K cellen in de tabel. Voor elke gevoelige cel $l = 1, ..., L$, moet zo'n lineair programma worden geformuleerd en opgelost. Laat voor de gevoelige cel l de vereiste marge gegeven zijn door μ_l.

Het weglaten van cellen in de gepubliceerde data vermindert de waarde voor de gebruiker, het publiek of de politiek. Als we aannemen dat er per cel (i,j) een waarde $\gamma_{i,j}$ is toe te kennen aan het verlies van publiek nut, dan is de beveiliger gesteld voor de vraag: wat is de minimum cost van weggelaten data zodanig dat een gepubliceerde tabel elke gevoelige cel beschermt?

Deze opzet valt precies in het stramien van de Benders decompositie: we hebben 0/1-variabelen $z_{(i,j)}$ voor elke cel (i,j) om aan te duiden of de cel wordt verbloemd, en we hebben variabelen $x^l_{i,j}$ en $y^l_{i,j}$ voor het beschrijven van de waarden in cel(i,j)voor de maximale marge voor de l-de gevoelige cel.

Een wiskundig model

We gaan uit van een tabel T met m rijen $i = 1, ..., m$ en n kolommen $j = 1, ..., n$, met te publiceren rijtotalen $T_{io} = \sum_j T_{ij}$ en kolomtotalen $T_{oj} = \sum_i T_{ij}$. Verder zijn gegeven L gevoelige cellen(i_l, j_l), met $l = 1,...,L$ die beschermd moeten worden met een marge μ_l. Weglaten van cel (i,j) geeft informatieverlies $\gamma_{i,j}$ tot gevolg. De volledige beschrijving in de vorm van een gemengd geheeltallig programmeringsprobleem is als volgt

$$\text{(SCO)} \quad \min \sum_{ij} \gamma_{ij} z_{ij}$$

onder voorwaarden

$$\begin{aligned}
\sum_j x^l_{ij} &= T_{io} && \forall l = 1,..., L;\ i=1,...,m \\
\sum_i x^l_{ij} &= T_{oj} && \forall l = 1,..., L;\ j=1,...,n \\
\sum_j y^l_{ij} &= T_{io} && \forall l = 1,..., L;\ i=1,...,m \\
\sum_i y^l_{ij} &= T_{oj} && \forall l = 1,..., L;\ j=1,...,n \\
y^l_{i_j i_l} - x^l_{i_j i_l} &\geq \mu_l && \forall l = 1,...,L \\
x^l_{ij} + T_{ij} z_{ij} &\geq T_{ij} && \forall l = 1,...,L;\ i = 1,..., m;\ j = 1,...,n \\
x^l_{ij} + (T_{ij} - T_{io}) z_{ij} &\leq T_{ij} && \forall l = 1,...,L;\ i = 1,..., m;\ j = 1,...,n \\
y^l_{ij} + T_{ij} z_{ij} &\geq T_{ij} && \forall l = 1,...,L;\ i = 1,..., m;\ j = 1,...,n \\
y^l_{ij} + (T_{ij} - T_{io}) z_{ij} &\leq T_{ij} && \forall l = 1,...,L;\ i = 1,..., m;\ j = 1,...,n \\
x^l_{ij}, y^l_{ij} &\geq 0 && \forall l = 1,...,L;\ i = 1,..., m;\ j = 1,...,n \\
z_{ij} &\in \{0,1\} && \forall i = 1,...,m;\ j = 1,...,n
\end{aligned}$$

Hierin zorgen de eerste beperkingen ervoor dat x^l en y^l consistente invullingen van de tabel zijn. De tweede helft van de restricties zorgen ervoor dat $x^l_{ij} = y^l_{ij} = T_{ij}$, *als* $z_{ij} = 0$. Voor $z_{ij} \neq 0$ zijn x^l_{ij} en y^l_{ij} effectief onbegrensd, gelegen tussen 0 en rij-totaal. De z-variabelen vormen de enige verbinding tussen de afzonderlijke problemen van maximering van de marge:

$$\text{(}LP_l(z)\text{)} \quad \max y^l_{i_j i_l} - x^l_{i_j i_l}$$

onder voorwaarden

$$\begin{aligned}
\sum_j x^l_{ij} &= T_{io} && \forall i = 1,...,m \\
\sum_i x^l_{ij} &= T_{oj} && \forall j = 1,...,n \\
\sum_j y^l_{ij} &= T_{io} && \forall i = 1,...,m \\
\sum_i y^l_{ij} &= T_{oj} && \forall j = 1,...,n \\
x^l_{ij} &= T_{ij} && \forall (i,j) | z_{ij} = 0 \\
y^l_{ij} &= T_{ij} && \forall (i,j) | z_{ij} = 0 \\
x^l_{ij}, y^l_{ij} &\geq 0 && \forall i = 1,...,m; j = 1,...,n
\end{aligned}$$

Hierin nemen de constraints voor $z_{ij} = 0$ de plaats in van de beperkingen

$$\begin{aligned}
x^l_{ij} &\geq T_{ij} - T_{ij} z_{ij} && \forall i = 1,...,m; j = 1,...,n \\
x^l_{ij} &\leq T_{ij} - (T_{ij} - T_{io}) z_{ij} && \forall i = 1,...,m; j = 1,...,n \\
y^l_{ij} &\geq T_{ij} - T_{ij} z_{ij} && \forall i = 1,...,m; j = 1,...,n \\
y^l_{ij} &\leq T_{ij} - (T_{ij} - T_{io}) z_{ij} && \forall i = 1,...,m; j = 1,...,n
\end{aligned}$$

Merk op dat het probleem $LP_l(z)$ altijd een oplossing heeft (neem $x = y = T$) en een begrensd optimum. Dus is de waarde $LP_l(z)$ gelijk aan de optimale waarde van het duale probleem

$$\begin{aligned}
\text{(}D_l(z)\text{)} \quad \min \ & \sum_i \alpha^x_i T_{io} + \sum_j \beta^x_j T_{oj} + \\
& \sum_i \alpha^y_i T_{io} + \sum_j \beta^y_j T_{oj} + \\
& \sum_{ij: z_{ij}=0} \delta^x_{ij} T_{ij} + \sum_{ij: z_{ij}=0} \delta^y_{ij} T_{ij}
\end{aligned}$$

onder voorwaarden

$$\alpha^x_i + \beta^x_j + \delta^x_{ij} \geq \begin{cases} -1 & \text{als } (i,j) = (i_l, j_l) \\ 0 & \text{anders} \end{cases}$$

$$\alpha^y_i + \beta^y_j + \delta^y_{ij} \geq \begin{cases} +1 & \text{als } (i,j) = (i_l, j_l) \\ 0 & \text{anders} \end{cases}$$

Hierin zijn alle duale variabelen α^x_i, β^x_j, α^y_i, β^y_j, δ^x_{ij}, δ^y_{ij} vrij van teken. Bij oplossing van het LP vinden we niet alleen de oplossing x,y maar ook het bewijs van optimaliteit gegeven door de duale variabelen α^x, α^y, β^x, β^y, δ^x, δ^y. De zwakke dualiteitsstelling zegt dat elke oplossing van $(LP_l(z))$ een waarde heeft kleiner of gelijk aan $D_l(z)$. Als $D_l(z) < \mu_l$ voor zekere l, dan is z geen aanvaardbare oplossing. De redenatie leunt met name op de ongelijkheid

$$y^l_{i_j i_l} - x^l_{i_j i_l} \leq \sum_j (\alpha^y_i + \beta^y_j + \delta^y_{ij}) y^l_{ij} + (\alpha^x_i + \beta^x_j + \delta^x_{ij}) x^l_{ij}$$

geldig voor elke niet-negatieve (x^l, y^l). Voor $\delta^y_{ij} \geq 0$ gebruiken we in de afschatting verder $y^l_{ij} \leq T_{ij} - (T_{ij} - T_{io}) z_{ij}$, terwijl we voor $\delta^y_{ij} < 0$ de afschatting $y^l_{ij} \geq T_{ij} - T_{ij} z_{ij}$ toepassen. Voor x^l_{ij} werkt de afschatting analoog.

Decompositie aan het werk

Het probleem (*SCO*) is op te lossen door het aan een MIP-solver aan te bieden, en dat gaat goed als het aantal beperkingen maar niet te groot is. Het aantal rijen in de constraint matrix is grofweg $4mnL$ en het aantal variabelen ongeveer $2mnL$. Voor een bescheiden voorbeeld met m=n=20 en L=10 geeft dat 16000 constraints. Nu is het aantal iteraties en deelgevallen die een MIP-solver nodig heeft moeilijk te voorspellen, maar het werk per LP-relaxatie schaalt grofweg met minstens het kwadraat van het aantal rijen. Daarnaast loopt het aantal deelgevallen (LP-relaxaties) ook gauw in de duizendtallen. Het kan zo gemakkelijk een 15 minuten rekentijd kosten.

We willen nu het probleem (*SCO*) oplossen zonder het in zijn volle glorie aan een MIP-solver te geven. De strategie is een probleem te formuleren alleen gesteld in z-variabelen. Een minimale verzameling condities zou kunnen zijn: $z_{i_j i_l} = 1$, $\sum_j z_{i_j j} \geq 2$, $\sum_i z_{i_j i} \geq 2$, voor $l = 1,...,L$ Dat lossen we op en vervolgens testen we voor elke l of $LP_l(\hat{z}) \geq \mu_l$, voor de gevonden oplossing $\hat{z}$. Is dat het geval dan zijn we klaar. Zo niet dan moeten we een aantal extra condities formuleren voor de z-variabelen, en opnieuw het compacte probleem oplossen.

Als $D_l(\hat{z}) < \mu_l$ voor zekere l, dan voegen we aan het stelsel condities op z een constraint toe van de gedaante:

$$\begin{aligned}
& \sum_i \alpha^x_i T_{io} + \sum_j \beta^x_j T_{oj} + \\
& \sum_i \alpha^y_i T_{io} + \sum_j \beta^y_j T_{oj} + \\
& \sum_{ij: z_{ij}=0} (\delta^x_{ij})^+ (T_{ij} - (T_{ij} - T_{io}) z_{ij}) + \\
& \sum_{ij: z_{ij}=0} (\delta^x_{ij})^- (T_{ij} - T_{ij} z_{ij}) + \\
& \sum_{ij: z_{ij}=0} (\delta^y_{ij})^+ (T_{ij} - (T_{ij} - T_{io}) z_{ij}) + \\
& \sum_{ij: z_{ij}=0} (\delta^y_{ij})^- (T_{ij} - T_{ij} z_{ij}) \geq \mu_l
\end{aligned}$$

Hierin staat $(t)^+$ voor max $\{0,t\}$, en $(t)^-$ = min$\{0,t\}$. De coëfficiënten α, β, δ zijn dus gevonden als duale variabelen bij het oplossen van $LP_l(\hat{z})$. Het is evident dat de oplossing $\hat{z}$ nu niet meer is toegelaten. Dat de condities valide zijn voor elke interessante oplossing z volgt uit duale toegelatenheid van de vectoren α, β, δ.

We blijven itereren tot het probleem is opgelost. Er zijn verschillende strategieën om te bepalen wanneer welke beperkingen worden toegevoegd. Dat kan eigenlijk bij elke kandidaat oplossing $\hat{z}$. De hoop is dat het aantal toe te voegen constraints niet al te hard oploopt. Daar is theoretisch niet veel over te zeggen maar in de praktijk is het aantal constraints nodig om het probleem in alleen maar z-variabelen te beschrijven veel kleiner dan de expliciete volledige beschrijving.

We werken dus met een MIP met mn variabelen, en initieel orde m+n contraints. We lossen de MIP op en voegen dan maximaal steeds L constraints toe. Voor het vinden van de constraints moeten we L LP's oplossen met $2(m+n)$ constraints elk. Naast het werken met een MIP (mixed integer problem) van aanzienlijk kleinere omvang dan het oorspronkelijke (*SCO*), is er hier een potentieel voordeel van parallellisme. Immers, voor vaste $z = \hat{z}$

moeten steeds L onafhankelijke LP's worden opgelost. Met tegenwoordige hardware kun je dan goed gebruik maken van meerdere processoren. Het is heel goed denkbaar dat de rekentijd met een factor 10 omlaag gaat, ook al zou je bijvoorbeeld wel 100 maal je MIP met 10 constraints uitbreiden. Het herhaaldelijk herformuleren van het probleem wordt ruimschoots gecompenseerd door de rekentijden van de deelproblemen die vele ordegrootten kleiner zijn. In een rekenvoorbeeld met m=30,n=25 en L=20 werd zo de rekentijd teruggebracht van 4,5 uur tot 1 minuut.

Conclusie

Het beschreven tabel-beveiligingsprobleem is een (gesimplificeerd) praktisch probleem waarbij een fractie van de variabelen een cruciale rol speelt. Het probleem is te vertalen naar een beschrijving in alleen die cruciale variabelen, zij het door toevoeging van een potentieel grote verzameling additionele beperkingen, die het bestaan en de optimaliteit van de rest van de variabelen garandeert. In de praktijk van statistieken zullen sommige cellen niet aanpasbaar zijn omdat ze leeg zijn. Wellicht kunnen ook de marginalen van de tabel verbloemd worden. Ook zijn er meer complexe situaties denkbaar omdat vaak verschillende tabellen, gebaseerd op dezelfde survey data, gepubliceerd worden.

Door de verbanden te gebruiken tussen gegevens in verschillende tabellen kan meer worden teruggerekend, en is het dus moeilijker data te verbloemen. Des te meer heeft een Benders decompositie aanpak dan juist zin.

Een variant op dit probleem is het instellen van parameters voor productie of verkoop, waarbij de verwachte winst is gebaseerd op een groot aantal scenario's. Ook dan is een iteratief proces denkbaar: een instelling z wordt getoetst aan een groot aantal scenario's l beschreven in variabelen x^l, y^l.

De splitsing van de variabelen is een creatief proces, dat erg leunt op inzicht in de aard van het onderhavige probleem. Het is daarom lastig een algemeen recept te geven. Maar het moge duidelijk zijn uit het gegeven voorbeeld dat de mogelijkheden legio zijn. We mogen er trots op zijn dat deze methodologie zijn oorsprong heeft gevonden in onze directe nabijheid, in de persoon van Jacques Benders.

Cor Hurkens studeerde in 1984 af bij professor Benders en promoveerde vijf jaar later in de Combinatorische Optimalisering onder Lex Schrijver. Sindsdien werkt hij als universitair docent aan de leerstoel die Benders bij zijn emeritaat in 1989 verliet. Aandachtsgebieden zijn voornamelijk toepassingen van scheduling en geheeltallige programmering in een industriële context.
E-mail: c.a.j.hurkens@tue.nl

THE YOUNG STATISTICIANS HAVE BEEN BUSY!

During the last few months there were three exciting events for the Young Statisticians. We were present at the annual VvS+OR Statistics Day in Utrecht. We went on a company visit to Shell in Amsterdam and we organized the first Statistics Café of 2017.

23 March: Young Statisticians at VvS+OR Statistics Day

The VvS+OR Statistics Day was held on March 23rd at the Jaarbeurs in Utrecht. Many of our members were present to gain insight in the role of statistics in health care by either listening to the numerous interesting talks or by participating in the hackathon, that was introduced by the Data Science section. During our luxurious lunch at Restaurant Ruig, just a few minutes walking from the venue, there were plenty of opportunities to get to know each other and engage in discussion.

10 April: Company visit to Shell

On April 10th, the company visit to Shell took place. Our members were given several interesting presentations about the role of statistics in this multinational. Not everything could be covered during the presentations, but our curiosity about the challenges that statisticians could tackle whilst working there was completely satisfied during the drinks.

17 May: Statistics Café

Causality was the topic of this year's first Statistics Café that was held on May 17th at Bar Beton CS, in Utrecht. It was one of the warmest evenings of that week, so we were very pleased to see that so many of our members showed up to listen to our three amazing speakers. Jan-Willem Romeijn (University of Groningen) spoke about the metaphysics and epistemology of causality. Next, Joris Mooij (University of Amsterdam) went into joint causal inference and lastly Daniele Marinazzo (University of Ghent) spoke about causality analysis applied to real world problems. It was an amazing event and we are really glad that the speakers took the time to share their insights on causality with us.

Our next events will take place after the summer holidays, we hope to see you there! Please consult our website youngstatisticians.nl for the latest news.

www.youngstatisticians.nl
contact@youngstatisticians.nl