STAtOR

Periodiek van de VVSOR jaargang 24, nummer 1, maart 2023

STAtOR

Jaargang 24, nummer 1, maart 2023

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

Kopij en reacties richten aan

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

Leden- en abonnementenadministratie van de VVSOR

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

Voor advertenties

Prof. dr. J.A.S. Gromicho, j.a.s.gromicho@uva.nl
STAtOR verschijnt in maart, juni, september en december.

Uitgever

ISSN 1567-3383

STAtOR 2.0

Dit eerste nummer van de 24e jaargang van ons tijdschrift is het eerste van een tweede reeks. Het eerste nummer dat we zonder de professionele hulp van Monique van Hootegem vormgegeven. Nu weten we pas echt wat een werk zij de afgelopen 23 jaar voor STAtOR verricht heeft. De opmaak wordt nu gedaan door enkele redactieleden met LaTeX/Overleaf als gereedschap. We hadden het geluk in contact te komen met Esger Renkema die de LaTeX-templates hiervoor ontwikkelde. Ongetwijfeld zult u in dit nummer onvolkomenheden aantreffen die te wijten zijn aan ons gebrek aan ervaring, of aan complicaties die we niet hadden voorzien. Maar we hebben het vertrouwen dat we snel aan de nieuwe wijze van werken zullen zijn gewend.

Dit is een nummer vol actualiteit, allereerst door de informatie over de Annual Meeting, die plaats zal vinden op donderdag 23 maart. Na de ledenvergadering zullen dit jaar vier spekers hun visie met ons delen over de toepassing van statistiek en operations research om robuuste beslissingen. Aanmelden is nog mogelijk. Graag voor 16 maart.

Veel artikelen gaan over een actueel onderwerp, zoals dat van Iris van Beuzekom over Optimalisatie van multi-energiesystemen, iets dat buitengewoon belangrijk is bij de huidige en komende transitie. En wat te denken van de bijdrage van Jelke Bethlehem over de (on)mogelijkheden van peilingen in het huidige Rusland?

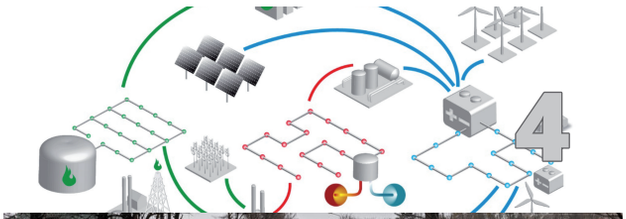
Overal in het straatbeeld ziet men tegenwoordig AED-apparaten. Vaak zit hier een groep vrijwillige burgerhulpverleners achter waardoor hulp bij hartinfacten zeer snel gegeven kan worden. Caroline Jagtenberg en haar collega-auteurs hebben onderzoek gedaan naar een optimale meldingsstrategie hiervoor. Ook dit is heel actueel.

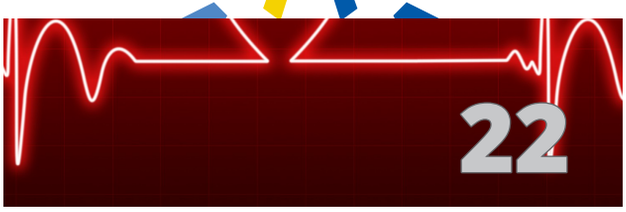
Vincent Warmerdam schrijft over het gevaar van foutieve labeling van data die voor machine-learning worden gebruikt. Gezien de groeiende rol van AI is ook dit een actueel onderwerp.

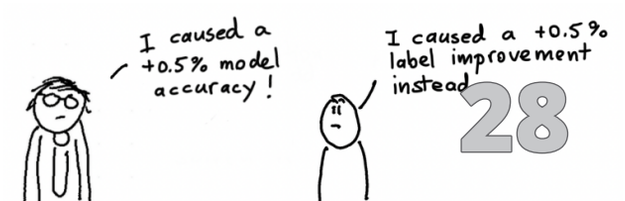
Verder vindt u in dit nummer het nieuws van de Young Statisticians en een drietal columns.

Wij wensen u veel leesplezier met het eerste nummer van STAtOR dat we op een nieuwe toekomstbestendige manier gemaakt hebben.

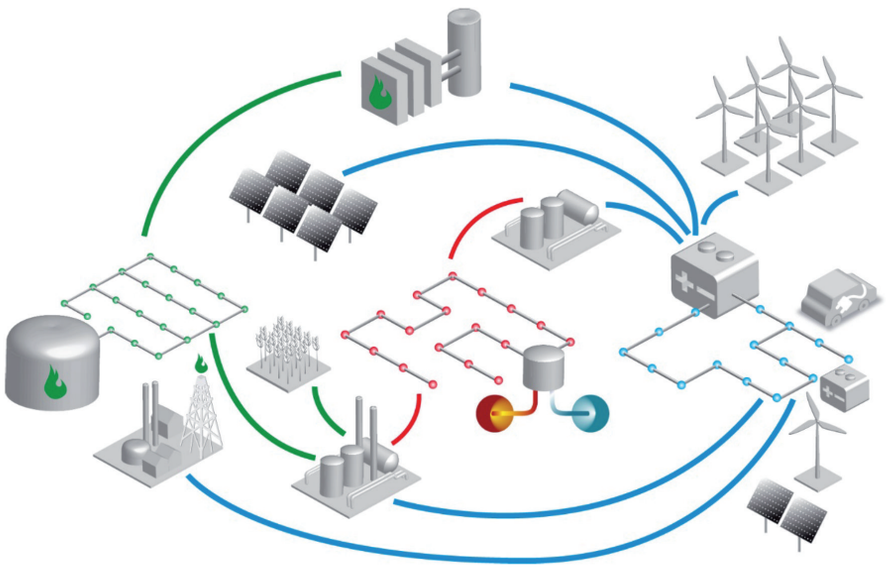
De STAtOR-redactie







INHOUD



Optimalisatie van multi-energiesystemen

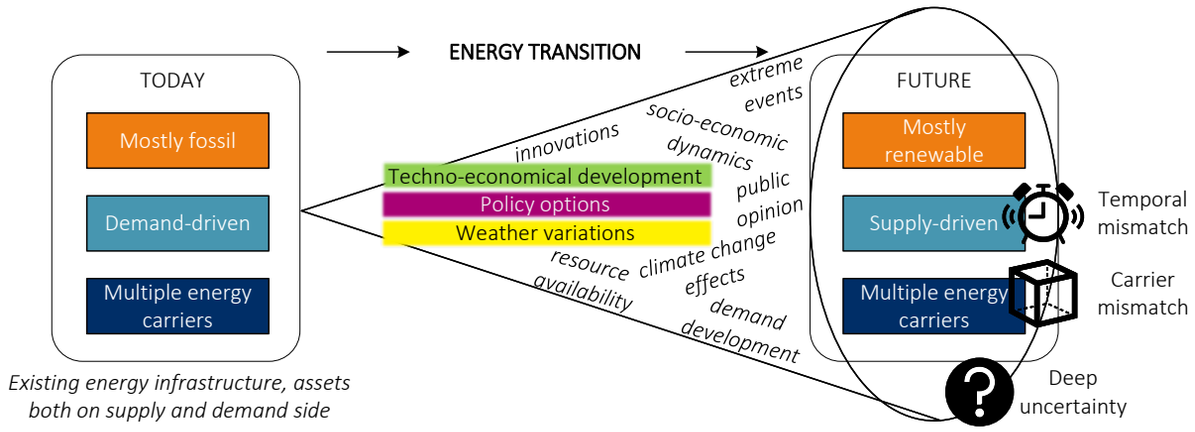
Optimalisatie van investeringen in geïntegreerde multi-energiesystemen om steden te helpen met het ontwerpen van een energietransitieplan

Iris van Beuzekom

Sinds de klimaatconferentie van de Verenigde Naties in Parijs in 2015 is er geen twijfel meer dat er dringend actie nodig is om verdere uitstoot van broeikasgassen versneld te verminderen en zo de ergste gevolgen van klimaatverandering te voorkomen. Vooral steden staan hierin voor grote uitdagingen, omdat ze zowel de grootste bijdrage leveren aan broeikasgasemissies, als de grootste gevolgen ondervinden van klimaatverandering. Zodoende hebben veel stedelijke beleidsmakers strenge klimaatdoelen opgesteld, vaak voorlopend op nationale doelen. Het is echter enorm lastig om een plan te ontwikkelen van vandaag naar die toekomstige doelen. Veel toekomstvisies kijken alleen naar het elektriciteitssysteem, omdat de meeste duurzame energiebronnen elektriciteit genereren. Echter is het huidige energiesysteem grotendeels niet elektrisch, wat zorgt voor een zogenoemde *carrier mismatch*: een verschil in de energiedrager die geleverd wordt en degene die gevraagd wordt (zie Figuur 1). Daarnaast zijn veel duurzame energiebronnen afhankelijk van weersomstandigheden en fluctueert daarmee de elektriciteitsgeneratie dagelijks, maandelijks en zelfs jaarlijks. Dit correspondeert niet altijd met de vraag naar energie, wat zorgt voor een *temporal mismatch*. Tot slot is de benodigde transitie van een voornamelijk fossiel naar een duurzaam energiesysteem enorm onzeker en afhankelijk van lokale, nationale, en zelfs internationale ontwikkelingen. Bovendien zijn experts het niet eens over de verdeling van deze onzekerheden, daarmee hebben we te maken met zogenoemde *deep uncertainty*.

Methode

In dit onderzoek worden deze drie uitdagingen aangepakt door toepassing van een systeemperspectief op alle energiedragers, gecombineerd met een investeringsplanning methode voor lange termijn met meerdere tijdsperioden (Beuzekom, 2022). Om steden te ondersteunen met het ontwikkelen van een energietransitieplan van vandaag naar een duurzame toekomst wordt een raamwerk voor de optimalisatie van

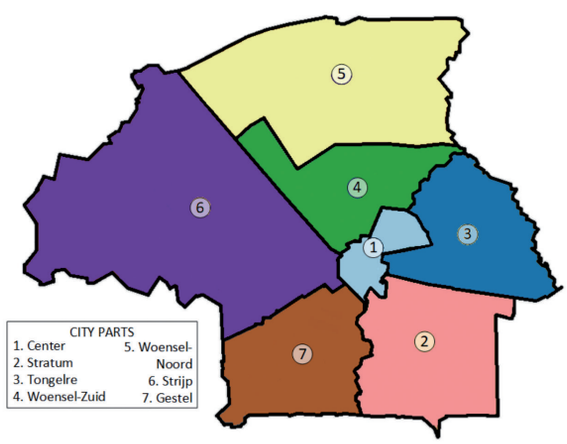


Figuur 1: Uitdagingen in de energietransitie, Bron: van Beuzekom, 2022

geïntegreerde *multi-energy* systemen voorgesteld (Beuzekom, B.-M. Hodge en H. Slootweg, 2021). Dit raamwerk bevat investeringsbeslissingen voor energienetwerken, energieconversie, energieopslag, en energieopweksystemen voor alle relevante energiedragers in het stedelijk gebied en zorgt dat klimaatdoelen behaald worden, terwijl het systeem als geheel in balans blijft. Het resulterende optimalisatieprobleem wordt geformuleerd als mixed integer linear program en vertaalt zich naar een nieuwe toepassing van een *capacitated facility location network design problem*; een combinatie van twee NP-hard problemen (Bonenkamp, 2020).

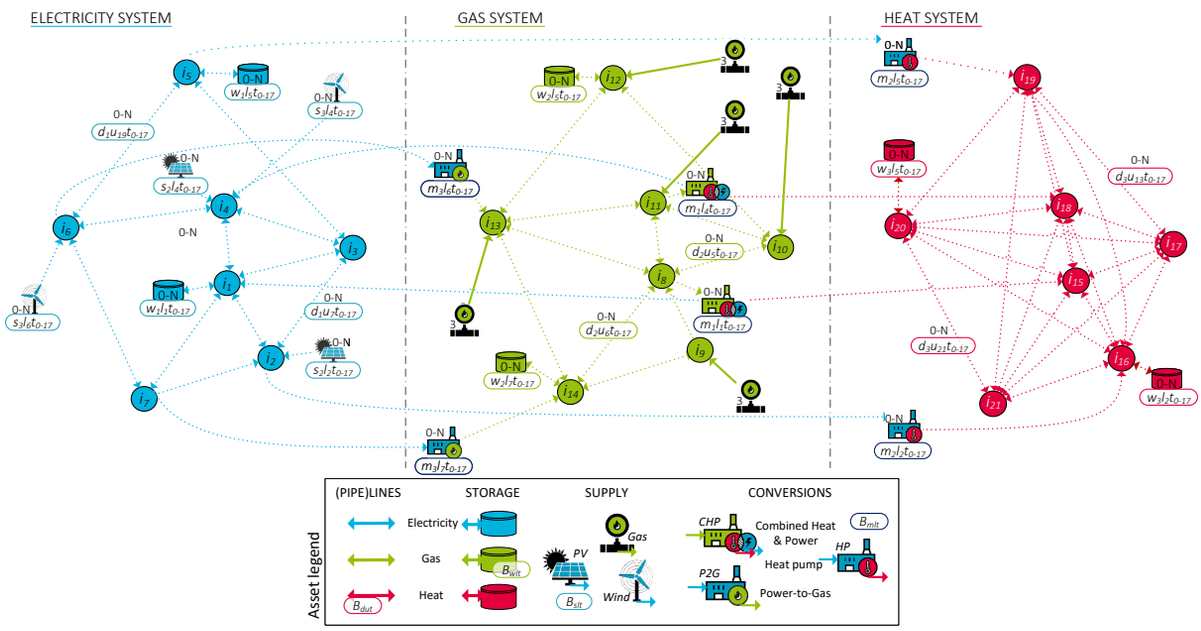
Data

Om dit raamwerk te valideren en demonstreren is een case study opgesteld van de gemeente Eindhoven (Figuur 2). Deze case bevat alle energienetwerken aanwezig in de stad: elektriciteit, warmte en gas. Daarnaast worden de klimaatdoelen van de stad gebruikt: 95% CO_2-reductie in 2050 ten opzichte van 1990. Verschillende datasets worden gecombineerd, onder andere over de ontwikkelingen van de energievraag in de stad, inclusief woninggebieden, commerciële en industriële gebieden en lokaal transport (Gemeente Eindhoven, 2016), technische data en technologische ontwikkelingen (ETSAP, g.d.), alsook socio-economische ontwikkelingen (Centraal Planbureau, 2016). In tweejaarlijkse tijdstappen van vandaag tot aan 2050 worden de benodigde investeringen in de energie-infrastructuur van de stad geoptimaliseerd en het energietransitieplan ontworpen. De case is geaggregeerd op niveau van de stadsdelen



Figuur 2: Eindhoven stadsdelen

(7 locaties) vanwege de complexiteit van het model. Elke locatie telt zoveel nodes als dat er energiedragers zijn. Figuur 3 geeft een voorbeeldoplossing van de 21 nodes die in deze case worden meegenomen. In dit figuur zie je de verschillende *assets* die meegenomen worden in de case: energienetwerken en -opslag voor elke energiedrager, drie soorten energievoorziening (gas, zonne- en windenergie) en drie soorten

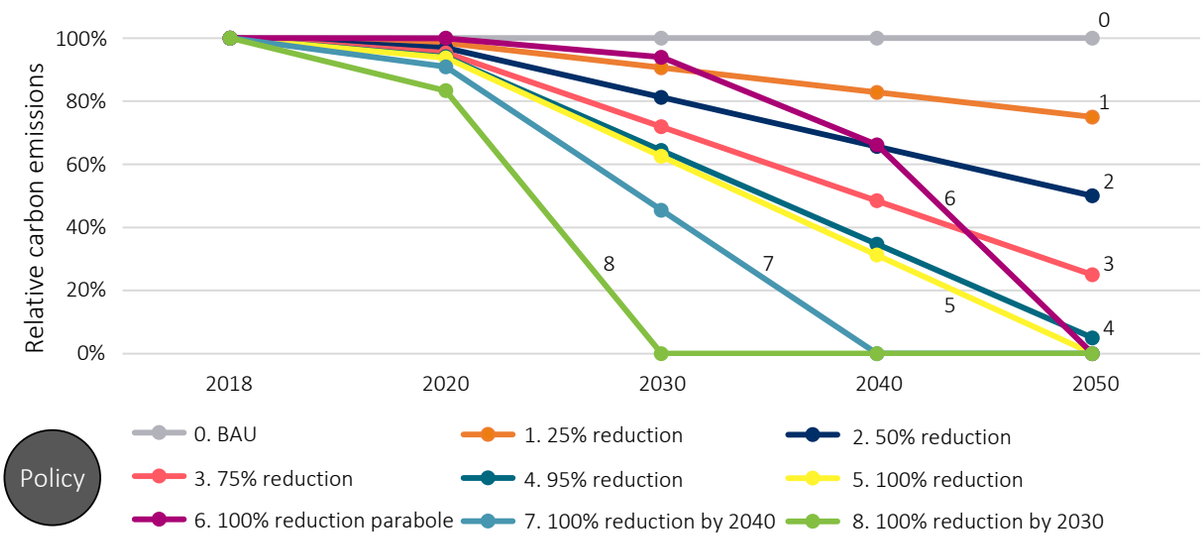


Figuur 3: Voorbeeld van de oplossingsmogelijkheden in een systeem met 21 nodes

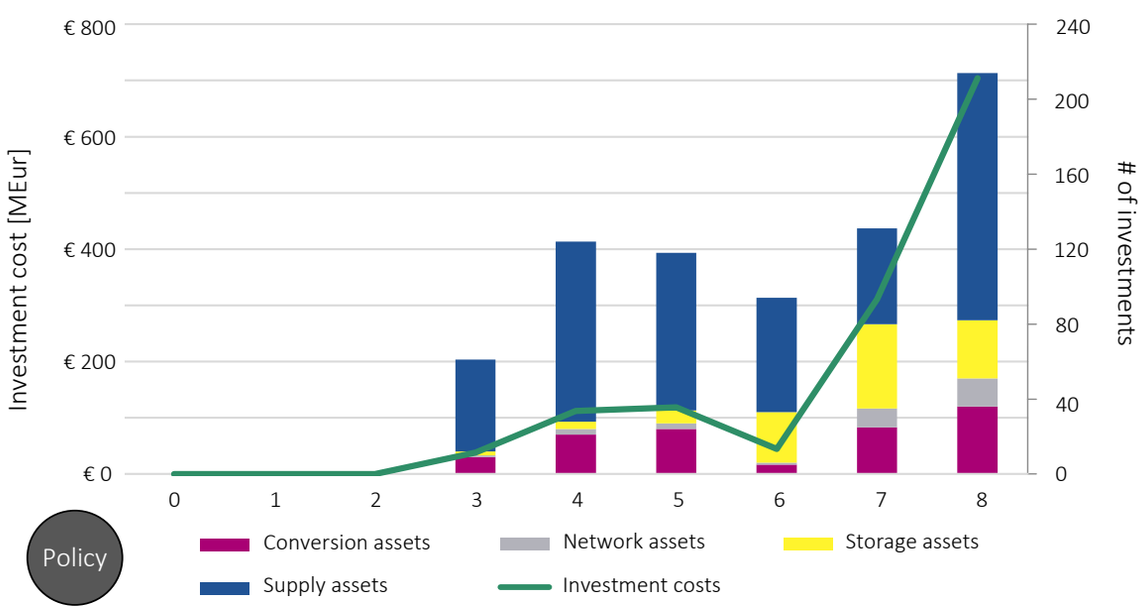
energieconversies die de individuele netwerken aan elkaar koppelen: een warmtekrachtkoppeling (*Combined Heat & Power, CHP)*, een warmtepomp (*Heat Pump, HP)* en een *Power-to-Gas* (P2G) asset. Een belangrijke kanttekening is dat de gasvoorziening al bestaat en juist moet worden gereduceerd, dus hier kan niet in worden geïnvesteerd.

Resultaten *what-if* scenario's

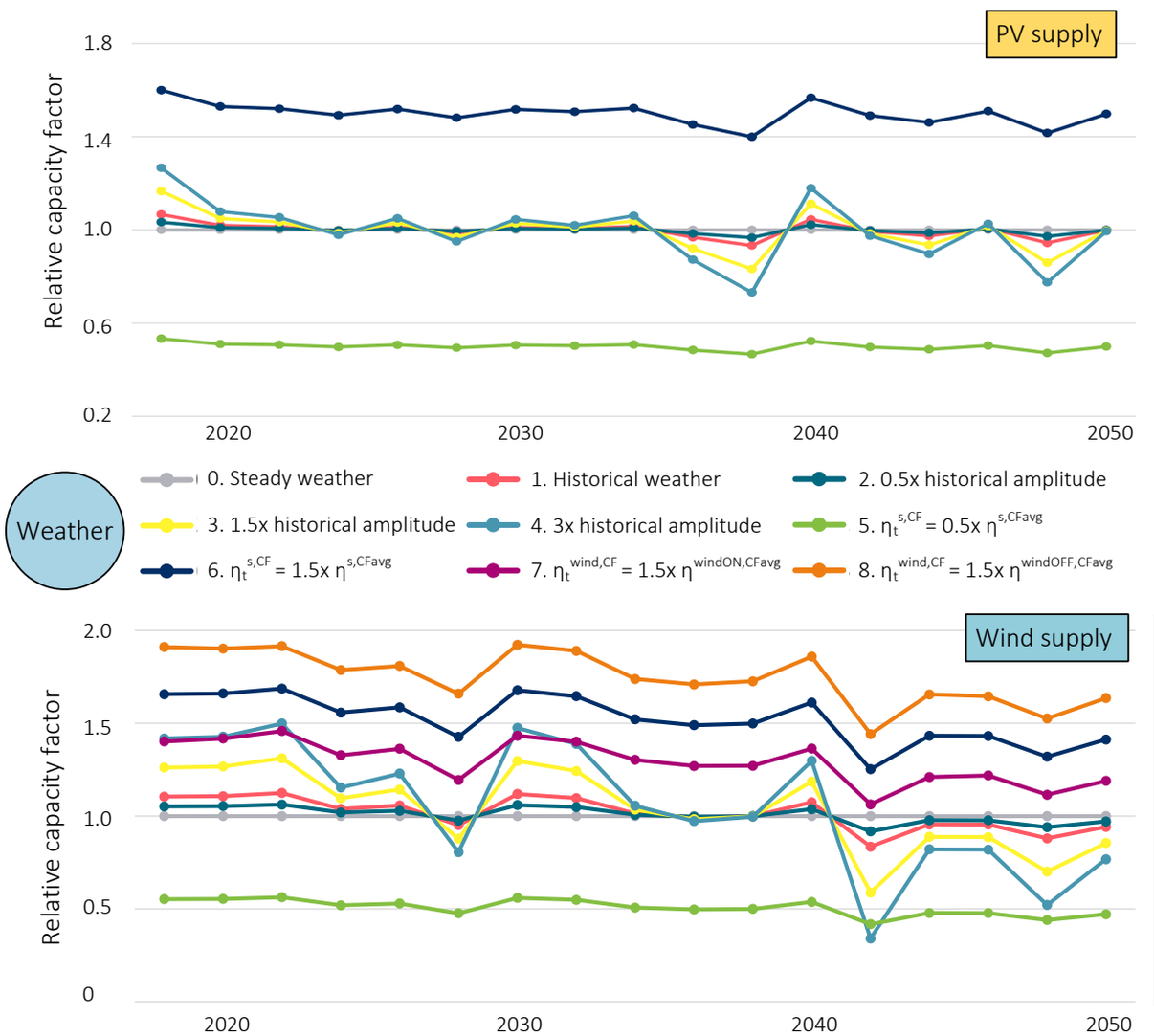
Het raamwerk is eerst getest op variaties in klimaatbeleid en jaarlijkse weersvariaties. Acht klimaatbeleid scenario's zijn gegenereerd: van een 'business-as-usual' (BAU) scenario zonder CO_2-emissie reducties, tot een scenario waar de emissies al in 2030 teruggebracht zijn naar nul (Figuur 4). Scenario 4 komt overeen met het klimaatdoel van Eindhoven. Figuur 5 laat de overkoepelende resultaten zien van alle scenarios vergeleken met het BAU scenario. De figuur geeft een vergelijking op basis van de totale kosten en de investeringen geaggregeerd per type asset (energieconversie, -netwerk, -opslag, en -opwek). Een strenger beleid leidt tot hogere kosten, echter met significant minder cumulatieve CO_2-emissies. Uitgesteld beleid in Scenario 6, met een parabolisch reductiepad, drukt de totale investeringskosten, echter zorgt dit voor enorme toename in de cumulatieve CO_2-emissies (vergelijkbaar met Scenario 2). Dit heeft serieuze consequenties voor de globale temperatuurstijging met mogelijk veel hogere kosten door klimaatverandering als gevolg. Dit is een belangrijk resultaat wat de toegevoegde waarde van het raamwerk toont, alsook de afwegingen die beleidsmakers moeten maken. De scenario's met weersvariaties zijn gebaseerd op historische data en de resulterende variaties in energieopwekking zijn veranderd in amplitude en absolute hoeveelheid. Figuur 6 laat de verschillende scenario's zien voor elektriciteitsopwekking via zonne-energie (PV supply) en vanuit wind. De absolute hoeveelheid opwek relateert aan de *capacity factor* $\eta^{s,CF}$ van de wind- en zonne-energie, een factor die aangeeft hoeveel uren van het jaar een energiecentrale elektriciteit genereert. Voor zon in Nederland ligt dit rond de 12%, voor wind op land rond de 27% en op zee rond de 35% (Renewables Ninja, 2021). Wederom leiden de uitdagendere scenario's tot meer investeringen, zie Figuur 7. Echter leidt dit niet altijd tot hogere kosten, omdat sommige investeringen pas in latere tijdsperiodes



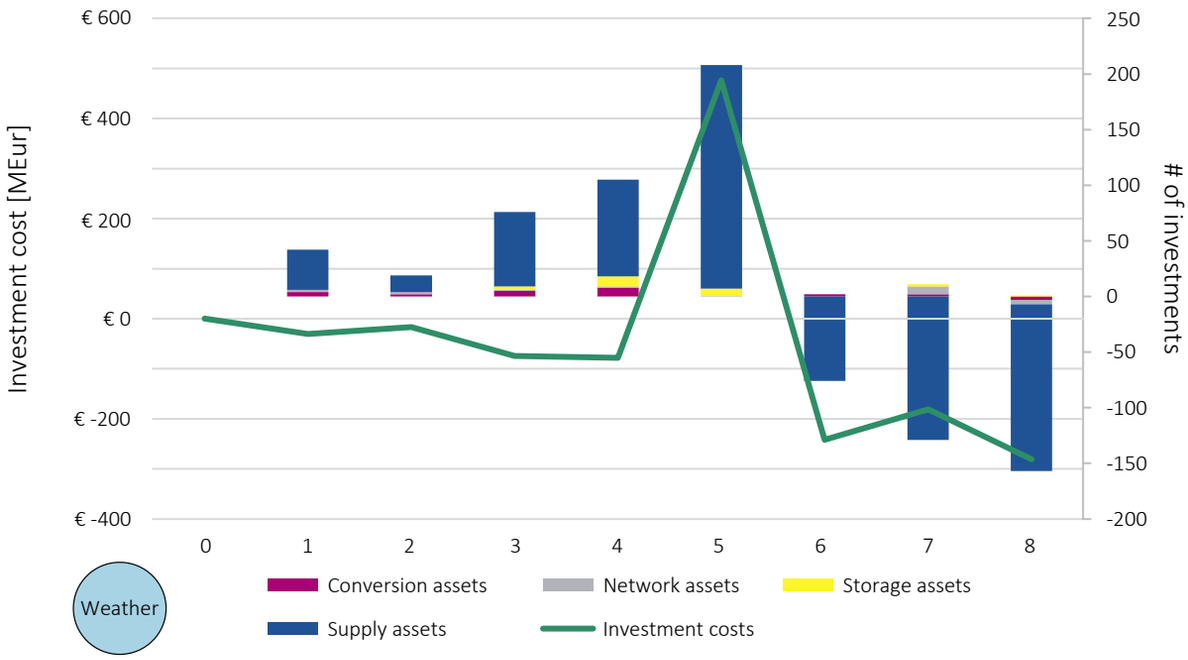
Figuur 4: Klimaatbeleidscenario's met CO_2-reductiedoelen



Figuur 5: Resultaten klimaatbeleidscenario's t.o.v. business-as-usual (BAU Scenario 0)



Figuur 6: Lange termijn weersfluctuatiescenario's met relatieve capacity factor voor elektriciteitsopwekking via zonne- (PV) en windenergie (PV = photovoltaic, $\eta^{s,CFavg}$= gemiddelde capacity factor, windON = wind op land, windOFF = wind op zee)



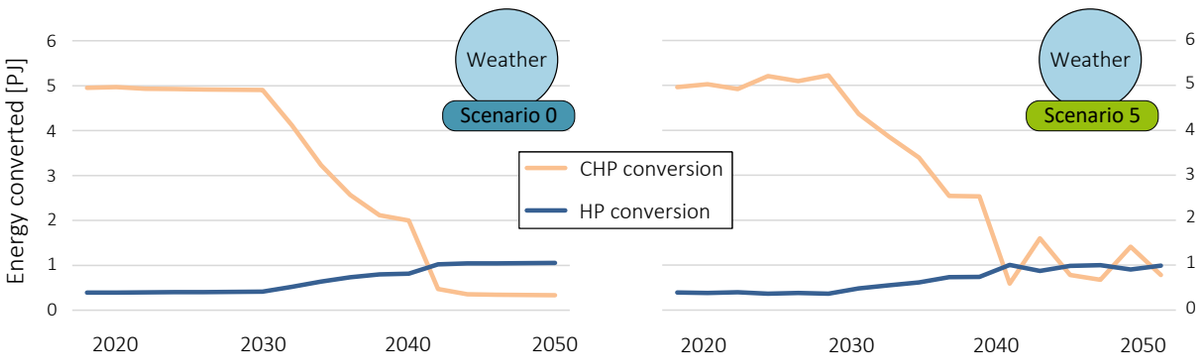
Figuur 7: Resultaten lange termijn weersfluctuaties t.o.v. stabiel weer (Scenario 0)

benodigd zijn, op het moment dat de disconteringsvoet investeringen relatief goedkoper maakt. Ook anders dan bij de klimaatbeleidscenario's, tonen de resultaten nu veel meer respons binnen de operatie van het energiesysteem. De weersvariaties leiden tot variaties in de opwek van elektriciteit, waarop warmte- en gas-gerelateerde conversie- en opslagsystemen reageren (zie Figuur 8). Dit bevestigt de potentie van een *multi-energy* systeem om weersfluctuaties op te vangen, alsook het belang van het meenemen van weerseffecten op deze lange termijn.

Methode en resultaten *deep uncertainty*

Overigens manifesteert onzekerheid zich meestal niet in geïsoleerde parameters, maar in meerdere parameters tegelijk. Zeker tijdens de energietransitie is er sprake van zogenoemde diepe onzekerheid. Dit is aan de orde wanneer meerdere parameters die invloed hebben op het energiesysteem, tegelijkertijd onzeker zijn en experts het niet eens worden over de mate van deze onzekerheid. Om dit te onderzoeken is een *exploratory modeling* methode toegevoegd aan het raamwerk (Beuzekom, B. Hodge en J. Slootweg, 2022). In de case worden nu de ontwikkelingen in vraag naar energie, alsook de technologische en socio-economische ontwikkelingen tegelijk gevarieerd. *Latin Hypercube Sampling* wordt toegepast omdat de distributie van de onzekerheden onbekend is en bovendien wordt met deze vorm van sampling de gehele onzekerheidsruimte structureler onderzocht dan bij Monte Carlo sampling. Vervolgens worden verschillende *data science*-technieken toegepast, waaronder een Extra Trees classifier en een agglomeratief clusteralgoritme, om de resultaten te analyseren en te vergelijken met een base case zonder onzekerheid.

Tabel 1 geeft een vergelijk tussen de *base case* zonder onzekerheid, en de resultaten van de 800 experimenten inclusief onzekere parameters. Naast de totale kosten en hoeveelheid investeringen, laat de tabel de investeringen per asset in energiecapaciteit in petajoules (PJ) zien. In de eerste kolom is ook het geïnstalleerd vermogen in megawatt (MW) meegenomen om het effect van de capacity factors per asset te duiden. Omdat de distributie van verschillende resultaten niet Gaussian, maar discreet is,



Figuur 8: Vergelijk operationele respons tussen Scenario 0 (stabiel weer) en Scenario 5 ($0.5x\eta^{s,CFavg}$)

is de standaarddeviatie niet geschikt om te resultaten te duiden. In plaats daarvan worden de mediaan, de Inter Quantile Range (IQR), het minimum en maximum gebruikt.

Result value	Base case	Median	Q1	Q3	IQR	Min	Max
Total costs [MEur]	700.75	654.68	617.43	702.47	85.04	581.47	966.18
No. of investments	328	330	294	422	128	220	914
Total capacity [PJ]	(MW)						
Electricity network	1.79 (57)	1.97	1.61	2.33	0.72	1.07	4.48
Gas network	3.94 (125)	4.55	4.31	4.67	0.37	3.81	6.15
Heat network	0.57 (18)	0.57	0.57	0.57	-	0.57	0.85
CHP conversion	4.97 (175)	4.97	4.97	4.97	-	4.97	5.53
HP conversion	1.15 (43)	1.07	0.91	1.36	0.45	0.59	1.70
P2G conversion	- (0)	0.40	-	0.79	0.79	-	7.54
PV supply	1.50 (389)	0.95	0.87	2.54	1.68	0.87	8.41
Wind supply	3.96 (438)	3.69	2.71	4.56	1.84	2.17	9.55
Electricity storage	-	-	-	-	-	-	
Gas storage	2.52	2.16	1.44	3.24	1.80	0.72	16.20
Heat storage	-	-	-	-	-	-	0.01

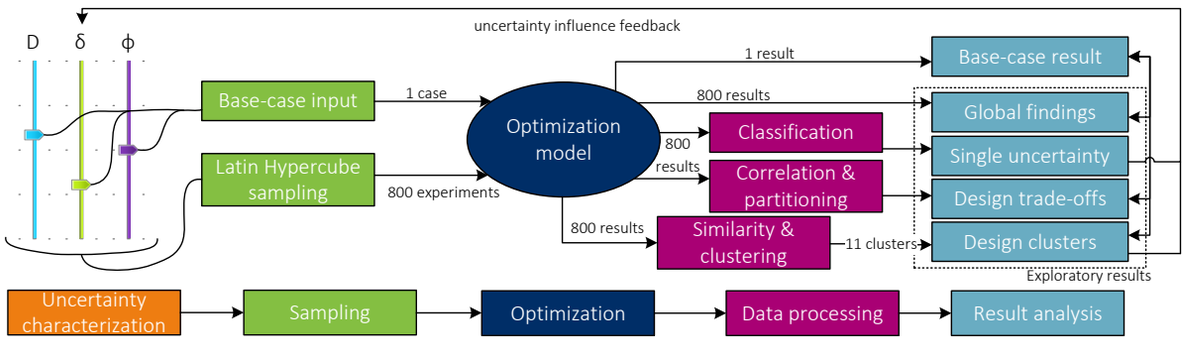
Tabel 1: Kwantitatieve resultaten van de base case in vergelijking met alle 800 experimenten

Zoals verwacht heeft deze diepe onzekerheid een enorm effect op de spreiding van de mogelijke oplossingen, wat direct een validatie is van de exploratory modeling methode. De spreiding van de totale kosten is bijvoorbeeld -17%, +38%. En waar Power-to-Gas assets in de base case geen rol spelen, zijn er experimenten waarbij het een van de relevantste assets is. De brede kijk op de gevoeligheid van een model geeft beleidsmakers de mogelijkheid om investeringstrends te ontdekken en effectief beleid te bepalen om bepaalde toekomstbeelden mogelijk te maken. Zo zijn de duurste ontwerpen diegenen met een hoge gasvraag in 2050. Dit toont de waarde van beleid gericht op versnelde elektrificatie en hogere energie-efficiëntie.

Om de resultaten verder te duiden is een hierarchisch, agglomeratief clusteralgoritme toegepast op basis van de cosinusafstand tussen de resultaten. Er zijn meerdere analyses uitgevoerd door de resultaten per cluster te aggregeren op verschillende assen: per asset, per locatie en per tijdsperiode. De aggregatie per asset laat zien dat bijna elk ontwerp een combinatie vergt van verschillende energieconversiesystemen. Dit benadrukt wederom het voordeel en belang van een multi-energy perspectief. De aggregatie per locatie laat zien dat de meeste investeringen zich concentreren op locaties met de hoogste vraag naar energie. Echter, er is op deze locaties niet altijd ruimte om benodigde assets neer te zetten. Dit toont hoe relevant het is om geografische aspecten mee te nemen om hierop in te kunnen spelen. Tot slot laat de aggregatie op tijdsperiode zien dat de meeste investeringen gedaan worden in de tweede helft van het tijdspad, na 2030. Op dat moment zijn de uitdagingen van de energietransitie het grootst en worden bovendien snel ontwikkelende technologieën economisch steeds interessanter. Dit toont de urgentie om nu structureel beleid te implementeren om straks de beste investeringen te kunnen doen.

Conclusie

Om stedelijke beleidsmakers te ondersteunen in het ontwikkelen van energietransitieplannen van vandaag naar een duurzame toekomst, stelt deze thesis een multi-energy raamwerk voor.



Figuur 9: Exploratory modeling methode

Dit vertaalt zich naar een optimalisatiemodel voor investeringsplannen op lange termijn met meerdere tijdstappen, inclusief een exploratory modeling methode om de onzekerheid op zulke termijnen mee te nemen. Het raamwerk kan effectief ontwerpen maken van energietransitieplannen, ondanks de enorme uitdagingen en diepe onzekerheid die hierbij komen kijken. De resultaten zijn consistent en reageren als verwacht op veranderingen in onzekere parameters. Naast duidelijke verschillen bij de verschillende tests op de stedelijke case, zijn er ook vele overeenkomsten in de investeringspatronen. Beide soorten resultaten zijn nuttig voor een stedelijke beleidsmaker en bevestigen de toegevoegde waarde van het raamwerk voor het ontwerpen van energietransitieplannen in stedelijke gebieden.

Literatuur

I. van Beuzekom, B. Hodge en J. Slootweg. „Exploring uncertainty in long-term, multi-stage investment planning of integrated urban energy systems". In: *Energy Reports - under review* (2022).

I. van Beuzekom. „Optimizing Investment Planning of Integrated Multi-Energy Systems to support urban decision makers design an energy transition pathway". Proefschrift. Eindhoven University of Technology, 2022. ISBN: 978-90-386-5600-7.

I. van Beuzekom, B.-M. Hodge en H. Slootweg. „Framework for optimization of long-term, multi-period investment planning of integrated urban energy systems". In: *Applied Energy* 292 (jun 2021), p. 116880. DOI: 10.1016/j.apenergy.2021.116880. URL: https://doi.org/10.1016%5C%2Fj.apenergy.2021.116880.

N. Bonenkamp. „Designing Multi-Energy Systems. Solution methods for a multi-period network design problem". Masterscriptie. Delft University of Technology, 2020.

Centraal Planbureau. *Energietransitiescenario's*. Available: https://www.pbl.nl/publicaties/nationale-energieverkenning-2016. 2016.

ETSAP. *Technology data*. Energy Technology Systems Analysis Program (ETSAP) of the International Energy Agency (IEA). Data retrieved from: https://iea-etsap.org/index.php/energy-technology-data. Accessed 02.2021.

Gemeente Eindhoven. *Eindhoven Klimaatplan 2016-2020*. 2016.

S. Pfenninger en I. Staffell. „Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data". In: *Energy* 114 (nov 2016), p. 1251–1265. DOI: 10.1016/j.energy.2016.08.060. URL: https://doi.org/10.1016%5C%2Fj.energy.2016.08.060.

I. Staffell en S. Pfenninger. „Using bias-corrected reanalysis to simulate current and future wind power output". In: *Energy* 114 (nov 2016), p. 1224–1239. DOI: 10.1016/j.energy.2016.08.068. URL: https://doi.org/10.1016%5C%2Fj.energy.2016.08.068.

Iris van Beuzekom rondde haar PhD onderzoek in 2022 af en werkt als consultant bij ORTEC.
Iris.vanBeuzekom@ortec.com, LinkedIn: irisvanbeuzekom

Peilen in Rusland is nog niet zo eenvoudig

Jelke Bethlehem

Kun je in Rusland onder de huidige omstandigheden een methodologisch verantwoorde peiling houden? Kun je een representatieve steekproef van Russen trekken? En geven de personen in de steekproef dan eerlijk antwoord op de gestelde vragen? Vrije meningsuiting is immers amper mogelijk, de pers is aan banden gelegd en demonstranten worden onmiddellijk gearresteerd. Twee onderzoekers (Philipp Chapkovski en Max Schaub) deden toch een poging tot een opiniepeiling. Ze probeerden de mening van de Russen te meten over de acties van het Russische leger in Oekraïne. Dat blijkt toch niet zo eenvoudig te zijn.

Philipp Chapkovski is van de HSE (Higher School of Economics in Moskou) en Max Schaub is van de Universiteit van Hamburg. Hun onderzoeksproject staat beschreven in een artikel op de academische blog van EUROPP (European Politics and Policy) en LSE (de London School of Economics). Zie ook (Chapkovski en Schaub,

2022).

Voor het trekken van de steekproef zou je de beschikking willen hebben over een bevolkingsregister met daarin alle Russen. Daaruit zou je dan een aselecte steekproef kunnen trekken. De twee onderzoekers hadden dat niet. Daarom zochten ze een ander steekproefkader. Ze besloten gebruik te maken van een online panel. Dat was *Toloka*. Toloka is de Russische versie van *MTurk*. MTurk (*Amazon's Mechanical Turk*) is een online crowdsourcing platform dat onderzoekers kan assisteren bij het rekruteren van mensen voor allerlei betaalde (online) taken. Volgens Amazon geeft MTurk toegang tot meer dan 500.000 panelleden in meer dan 190 landen. Hierbij moet je wel bedenken dat 75% van de panelleden in de VS zit. Dus de panelleden zijn niet netjes verdeeld over de landen in de wereld.

De beide panels MTurk en Toloka zijn gevuld met personen die zichzelf spontaan hebben opgegeven voor het regelmatig uitvoeren van online taken. Er is dus sprake van zelfselectie. En het probleem van zelfselectie is dat het de representativiteit van de peilingen ernstig kan aantasten. Analyses hebben uitgewezen dat MTurk te veel jongeren bevat, te veel hoog opgeleiden, te weinig gelovigen en te veel liberalen. En in de VS zijn Afro-Amerikanen, Latijns-Amerikanen en Aziaten oververtegenwoordigd. De conclusie is dat steekproeven uit MTurk niet representatief zijn.

De problemen met de representativiteit deden zich ook voor bij Toloka. Chapkovski en Schaub trokken een steekproef van 2.998 personen. Tabel 1 vergelijkt voor een aantal variabelen de verdeling in de steekproef met die in de populatie. De cijfers over de bevolking zijn afkomstig uit de volkstelling die in 2010 in Rusland is gehouden.

Variabele	Steekproef	Populatie	Verschil
Man	49 %	45 %	4 %
Ouder dan 40 jaar	36 %	58 %	-22 %
Universitaire opleiding	60 %	29 %	31 %
Stedelijk gebied	46 %	31 %	15 %
Aantal personen	2.998	112.557.618	

Tabel 1: De verdeling van de variabelen geslacht, leeftijd, opleiding en mate van verstedelijking in de steekproef en in de populatie

Mannen zijn wat oververtegenwoordigd in de steekproef en ouderen zijn zwaar ondervertegenwoordigd. Verder zitten er veel te veel hoger opgeleiden in de steekproef. En ook zijn de stedelijke gebieden oververtegenwoordigd. Kortom, de steekproef is niet representatief. Je zou kunnen proberen de peiling te corrigeren voor het gebrek aan representativiteit. Dat kan met een wegingsprocedure. Daarvoor heb je weegvariabelen nodig. Dat zijn variabelen

Figuur 1: De vraag in deelsteekproef 1

Figuur 2: De vraag in deelsteekproef 2

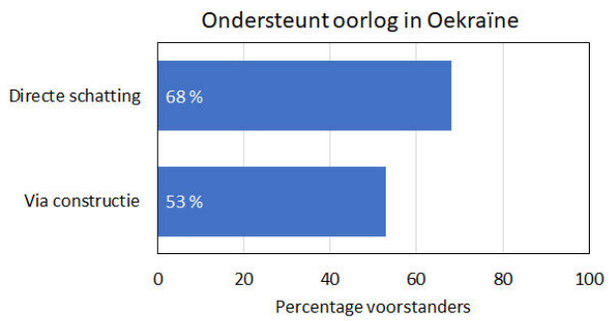
waarvoor de verdeling in de steekproef en in de populatie beschikbaar zijn. Tabel 1 laat zien dat zulke variabelen beschikbaar waren. Helaas hebben de onderzoekers toch geen weging uitgevoerd.

En dan de vraagstelling. Hoe vraag je de mening van de respondenten over de oorlog in Oekraïne als ze die mening niet durven geven, en dus misschien wel een sociaalwenselijk antwoord geven? De onderzoekers hebben daarvoor een speciale constructie bedacht. Ze verdeelden de steekproef via loting in twee deelsteekproeven. De personen in de eerste deelsteekproef kregen de vraag in Figuur 1 voorgelegd.

De lijst bevat vier mogelijke maatregelen: acties van het Russische leger in Oekraïne, legalisatie van het homohuwelijk, verhoging van de uitkering voor arme Russische gezinnen en overheidsmaatregelen om abortus te voorkomen. De respondenten moesten aangeven achter hoeveel van deze maatregelen ze staan. Ze zeggen dus niet achter welke maatregelen ze staan. Het gaat alleen om het aantal maatregelen waar ze achter staan. Ze hoeven dus bijvoorbeeld niet te zeggen of ze de oorlog in Oekraïne steunen.

De personen in de tweede deelsteekproef kregen een iets andere vraag voorgelegd, zie Figuur 2. In deze vraag gaat het om drie maatregelen. Eén maatregel minder dus. De maatregel over de oorlog in Oekraïne is weggelaten. Door het aantal voorstanders van drie maatregelen in de tweede vraag af te trekken van het aantal voorstanders van vier maatregelen in de eerste vraag, krijg je een schatting voor het aantal voorstanders van de oorlog in Oekraïne.

In de tweede steekproef werd ook nog direct (zonder speciale constructie) gevraagd naar de mening over de oorlog in Oekraïne. Dat maakt het mogelijk om het antwoord op deze directe vraag te vergelijken met het antwoord dat met de speciale constructie is verkregen. De resultaten staan in de Figuur 3.



Figuur 3: Schattingen voor het percentage Russen dat de oorlog in Oekraïne ondersteunt

Er is een duidelijk verschil tussen beide aanpakken. Als je de vraag direct stelt, zonder speciale constructie, dan steunt twee derde (68%) van de Russen de oorlog in Oekraïne. Corrigeer je voor het effect van sociaalwenselijke antwoorden, dan daalt het percentage naar nog maar net boven de helft (53%).

Je kunt hieruit in ieder geval de conclusie trekken dat een simpele peiling met gevoelige directe vragen niet de correcte antwoorden oplevert. Je moet iets doen om sociaalwenselijke antwoorden te vermijden of te repareren.

Je moet je ook afvragen of de gecorrigeerde peiling wel het juiste antwoord oplevert. Misschien is het wel zo dat een aantal respondenten het nog steeds niet vertrouwt en ten onrechte aangeeft de invasie in Oekraïne steunen. Dus dan zou het juiste antwoord een nog lager percentage kunnen zijn. Verder onderzoek lijkt nodig. Voorzichtigheid is geboden met Russische peilingen.

Literatuur

P. Chapkovski en M. Schaub. „Do Russians tell the truth when they say they support the war in Ukraine? Evidence from a list experiment". In: (2022). URL: https://blogs.lse.ac.uk/europpblog/2022/04/06/do-russians-tell-the-truth-when-they-say-they-support-the-war-in-ukraine-evidence-from-a-list-experiment/.

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

Letter from the president

It's my honour and pleasure to invite you all to our society yearly highlight: the Annual Meeting! It's possible to have in-person meetings and we make full use of this to also host a conference dinner after the meeting. Just as last year, the meeting will be hybrid such that those who can't attend in person can still participate.

The Annual Meeting takes place on Thursday 23 March. This year the venue is In de Driehoek in Utrecht. Members can register for the day for 65 euro, which includes lunch and drinks. The conference dinner (€ 55) takes place at De Rechtbank, a few minutes walk from the conference venue. Thanks to a generous sponsorship from LUXs, student members can join both the meeting itself and the conference dinner for a reduced price. Online participation is free (but less fun than in-person participation).

Once again, the Annual Meeting committee succeeded in setting up an interesting programme with a diverse range of speakers. The day's topic is *Statistics and Operations Research for Robust Decision Making*.

In its essence, statistics and OR are about making informed decisions based on incomplete information and uncertainty. How can these decisions be made in a robust way, especially when these are high-stake decisions. This requires the best data sources, the best methods and optimal algorithms.

The four speakers will highlight this topic from different viewpoints. At the end of the day, before the drinks and dinner, there will be a round table discussion with the four speakers.

Furthermore, we will have the regular items on the agenda: the (members only) General Assembly, for which documents will be e-mailed two weeks prior to the day, and the ceremony for the Willem R. van Zwet Award and the Jan Hemelrijk Award.

I hope you're looking forward as much to this day as I am, and I hope to see many of you there on 23 March!

Casper Albers
President VVSOR

VVSOR Annual meeting March 23, 2023

ROBUST decision making

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

Thursday March 23, 2023

10:45 – 17:15

hybrid: online & at In de Driehoek

Willemsplantsoen 1 C, 3511 LA Utrecht

What is the role of statistics and operations research in robust decision making? Four speakers will discuss how their research contributes to robust decision making, all of them with a different perspective on the topic and coming from an other field of study.

- Prof. dr. Peter Grünwald
- Prof. dr. Frank Pijpers
- Prof. dr. Gianluca Baio
- Dr. Julie Rozenberg

This year the Annual Meeting will be a hybrid event at In de Driehoek in Utrecht and broadcasted live online. During the Q&A questions can be asked via the chat function. We will have a general assembly for members (in Zoom), followed by the actual event with four talks and two award presentations. The AM 2023 will be in English.

Attending this year's annual meeting online is free of charge. Attending the meeting at In de Driehoek (including drinks and lunch) costs 65 euro. Reduced price for students: 30 euro. Additional registration is required for dinner and pubquiz at "De Rechtbank"

Please register on the vvsor-website https://www.vvsor.nl/articles/vvsor-annual-meeting-2023.

DATE

Thursday, March 23, 2023

VENUE

Online via Livestream and at In de Driehoek, Willemsplantsoen 1C, 3511 LA Utrecht

REGISTRATION

Registration for the conference is mandatory at https://www.vvsor.nl/articles/vvsor-annual-meeting-2023. Detailed information can be found on our website.

LANGUAGE

The talks at the annual meeting will be in English.

ALGEMENE LEDENVERGADERING (ALV)

The Annual General Meeting of members (ALV) takes place on March 23, 10:45 – 11:45, via Zoom. The relevant documents will be e-mailed two weeks before the meeting.

SNACKS AND DRINKS

Lunch and drinks during the breaks will be provided.

DINNER WITH PUBQUIZ

Dinner at De Rechtbank, Korte Nieuwstraat 14, 3512 NM Utrecht. The pubquiz will be organized by the Young Statisticians.

ORGANIZING COMMITTEE

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

PLEASE REGISTER BEFORE MARCH 16

10:15 - 10:45	**Registration + coffee & tea**
10:45 - 11:45	**ALV, General Assembly (members only)**
11:45 - 12:25	**Lunch at the 'In de Driehoek'**
12:25 - 12:30	**Prof. dr. Casper Albers | Welcome & Opening of the AM 2023**
12:30 - 13:00	**A new type of robustness: valid tests and confidence intervals without setting alpha in advance** Prof. dr. Peter Grünwald
13:00 - 13:30	**Official statistics as support for robust public governance** Prof. dr. Frank Pijpers
13:30 - 13:50	**Break 1 with coffee & tea**
13:50 - 14:20	**Ceremony of the Willem R. van Zwet Award and the Jan Hemelrijk Award** Prize winners will be presented by the juries, followed by a short presentation by the laureates
14:20 - 14:50	**Making robust decisions in health technology assessment: the value of value of information** Prof. dr. Gianluca Baio
14:50 - 15:10	**Break 2 with coffee & tea**
15:10 - 15:40	**Decision making under deep uncertainty applied to economic development policies** Dr. Julie Rozenberg
15:40 - 16:10	**Final panel discussion with speakers (Q&A) & Closure**
16:10 - 16:15	**Wrap up & Finish**
16:15 - 17:15	**Drinks at "In de Driehoek"**
17:30 - 21:30	**Dinner + Pubquiz at "De Rechtbank" (extra registration requiered)**

12:30 – 13:00

A new type of robustness: valid tests and confidence intervals without setting alpha in advance

Prof. dr. Peter Grünwald
Research institute for mathematics and computer science in the Netherlands (CWI)

A standard practice in hypothesis testing is to mention the p-value alongside the accept/reject decision. We show a major advantage of mentioning an e-value instead.

With p-values, we simply cannot use an extreme observation (e.g. p « alpha) for getting better frequentist decisions. With e-values we can, since they provide Type-I risk control in a generalized Neyman-Pearson setting with the decision task (a general loss function) determined post-hoc, after observation of the data — thereby providing a handle on the age-old "roving alpha" problem in statistics: we obtain risk (expected loss) bounds which hold independently of the loss, or any alpha, being set in advance.

The reasoning can be extended to confidence intervals. E-values were originally (in 2019) introduced because of their ability to deal with optional continuation, i.e. gathering additional data whenever one sees fit. Their ability to deal with post-hoc decision tasks provides a second, independent argument for embracing them. This work is based on P. Grünwald. Beyond Neyman-Pearson. arXiv 2205.00901, 2022; and P. Grünwald. The E-Posterior. Phil. Trans. Soc. London Ser. A, 2023.

Prof. dr. Peter Grünwald heads the machine learning group at CWI in Amsterdam, the Netherlands. He is also full professor of statistics at the mathematical institute of Leiden University. A recipient of NWO VIDI and VICI grants, in 2010 he was co-awarded VVSOR's Van Dantzig Award, the highest Dutch award in statistics and operations research. From 2018-2022 he served as President of the Association for Computational Learning, the organization running COLT, the world's prime annual conference on machine learning theory, which he chaired himself in 2015. Since about 2018, his research group has focused almost exclusively on safe anytime-valid inference and e-values.

13:00 – 13:30

Official statistics as support for robust public governance

Prof. dr. Frank Pijpers
Korteweg-de Vries Institute for Mathematics (UvA)/Statistics Netherlands (CBS)

In this talk, I will argue that for effective evidence-based public governance it is imperative that supplementary analysis and interpretation be provided by national statistical institutes (NSIs). NSIs must disseminate not only data, but also the possibilities and limitations of interpretation of the data being published, or the micro-level data that underpin them.

Public governance decisions often implicitly assume causalities, which have not necessarily been demonstrated or properly tested. Even outside of the controlled and isolated settings of experiments, in some cases it is possible to explore and test hypotheses of causality, which the work of renowned researchers such as Imbens and Angrist have demonstrated.

Causality testing brings together some traditional statistics with the relatively new field of complexity science. The traditional part is quantifying margins of uncertainty, that ought to be available for everything NSIs publish. The non-traditional part comes from the realisation that what we designate as trends or properties of society or the economy, are emergent from the myriad of individual interactions between people, which implies that "causes" are always through multiple pathways of mechanisms. I hope to illustrate this point using some examples from recent 'case work'.

Prof. dr. Frank Pijpers is senior methodologist at Statistics Netherlands and professor by special appointment at the Korteweg-de Vries Institute for Mathematics of the University of Amsterdam. The focus of his chair is on complexity for official statistics. Before joining Statistics Netherlands in 2010, Frank carried out fundamental research in astrophysics at various universities in Europe and worked for the UK Government Operational Research Service .

14:20 – 14:50

Making robust decisions in health technology assessment: the value of value of information

Prof. dr. Gianluca Baio
University College London (UCL)

Health technology assessment (HTA) is the final stage of clinical development, where interventions are assessed for their "value-for-money". Bodies such as the National Institute for Health and Care Excellence (NICE) in the UK or Zorginstituut Nederland (ZIN) in the Netherlands act as conduits for the relevant (public) healthcare provider to suggest whether a given intervention should be funded, once it's put on the market. Often, decisions are based on limited evidence, which generates large uncertainty over the decision-making process and the possible consequences of making the "wrong" choice, including sunk costs associated with switching from one technology to another.

Value of information (VoI) is a principled set of techniques that can be used to assess the impact of uncertainty in model inputs over the decision-making, as well as to prioritise research into specific components of a model, in order to reduce the underlying, key drivers of the uncertainty.

In this talk, I will briefly introduce the main concepts around HTA and VoI and present some recent computational and substantive development in the VoI methodology.

Prof. dr. Gianluca Baio is a professor of Statistics and Health Economics in the Department of Statistical Science at University College London (UK). Gianluca's main interests are in Bayesian statistical modelling for cost effectiveness analysis and decision-making problems in the health systems, hierarchical/multilevel models and causal inference using the decision-theoretic approach. Gianluca leads the Statistics for Health Economic Evaluation research group within the department of Statistical Science. He has been the 18th Armitage Lecturer in November 2021.

15:10 – 15:40

Decision making under deep uncertainty applied to economic development policies

Dr. Julie Rozenberg
World Bank

Climate change, pandemics, financial crises, are examples of deep uncertainties that challenge decision making for public policy. Deep uncertainty occurs when decision makers and stakeholders do not know or cannot agree on how likely different future scenarios are. For economists assessing the consequences of policy options or infrastructure investments, the presence of deep uncertainty requires using new techniques that look for robust decisions—performing well under multiple future conditions—rather than an optimal solution under a single prediction of the future. This talk will give examples of how robust decision making methods and tools are used to support decisions for infrastructure investments or to analyze future climate change impacts on poverty reduction goals.

Dr. Julie Rozenberg is a widely published economist with 15 years of experience working on the link between development policy and climate change adaptation and mitigation. She works as a Senior Economist at the World Bank where she leads applied research to support decision making for investments and public policies in developing countries. Julie is also an editor for Wires Climate Change, and the Vice President of the Society for Decision Making Under Deep Uncertainty.

Gerrit Stemerdink

Column

Are you from the ladies program?

In het verleden hadden grote conferenties vaak een speciaal programma voor de partners van de deelnemers. Zo'n programma stond aangekondigd als het 'Ladies Program'. Men ging er namelijk gemakshalve van uit dat de deelnemer een man was, en vrijwel altijd was dat ook zo. Terwijl de mannen moeilijke lezingen bijwoonden, werden hun echtgenotes onthaald op attracties als bezoeken aan parfumfabrieken, musea, exclusieve winkels en exotische tuinen.

Ook de wereld van statistische software was in de jaren '70 en '80 nog zo'n mannenbolwerk, slechts sporadisch kwam men daar een vrouw tegen. En dan gaan mannen wel eens in de fout als ze zo'n zeldzame verschijning ontmoeten.

De International Association for Statistical Computing, een ISI-afdeling, organiseert al sinds 1974 tweejaarlijks de Compstat conferenties. Compstat 1978 vond plaats in Leiden en dat had in de jaren daarna geleid tot een relatief groot aantal Nederlandse leden van de IASC en deelname aan de Compstat conferenties.

Compstat 1984 werd in Praag georganiseerd. Het was een geweldige ervaring, een prachtige stad, heel vriendelijke mensen en fantastisch bier. Zoals gebruikelijk was er ook hier een relatief grote Nederlandse deelname, ruim 20

waarvan het merendeel elkaar goed kende vanuit de Sectie Statische Programmatuur van de VVS. Onze delegatie telde zowaar drie vrouwen die redelijk bekijks trokken in deze masculiene omgeving, ook al doordat een van hen lang blond haar had. Zelf vonden wij hun deelname eigenlijk niets bijzonders, het waren gewoon zeer actieve SSP-leden die we goed kenden.

Ook hier was er een Ladies Program, bijna dagelijks konden de dames kiezen voor een interessante activiteit. Daarnaast werd traditiegetrouw voor álle deelnemers en hun partners op de woensdagmiddag een toeristisch uitstapje georganiseerd. In Leiden 1974 bezocht men uiteraard de Deltawerken en in Praag 1984 ging het naar slot Konopiště. Dat was het jachtslot van de in 1914 in Sarajevo neergeschoten aartshertog Franz-Ferdinand. Het lag in een mooie dichtbeboste omgeving en was omringd door een groot park. Na een tour door het slot, met honderden harnassen, opgezette dieren en geweien van neergeschoten herten, ging de rondleiding verder door dat park.

Een van de Nederlandse deelnemers kwam niet uit de SSP-kring, hij kwam uit een heel andere omgeving. Het was een erg aardige man, een beetje een verlegen charmeur, die kennelijk behoefte had aan een praatje. Mij kende hij wel redelijk goed en toen ik met die blonde Nederlandse vrouw en twee Tsjechen in het Engels in gesprek was, durfde hij het aan de betreffende dame aan te spreken. Zijn openingszin was ”Are you from the ladies program?” Zij had hem kennelijk als Nederlander herkend en antwoordde in het Nederlands ”Nee, ik heb vanochtend een presentatie gehouden, was u daar niet bij?” Zelden heb ik iemand zo totaal verbluft en zoekend naar woorden zien staan.

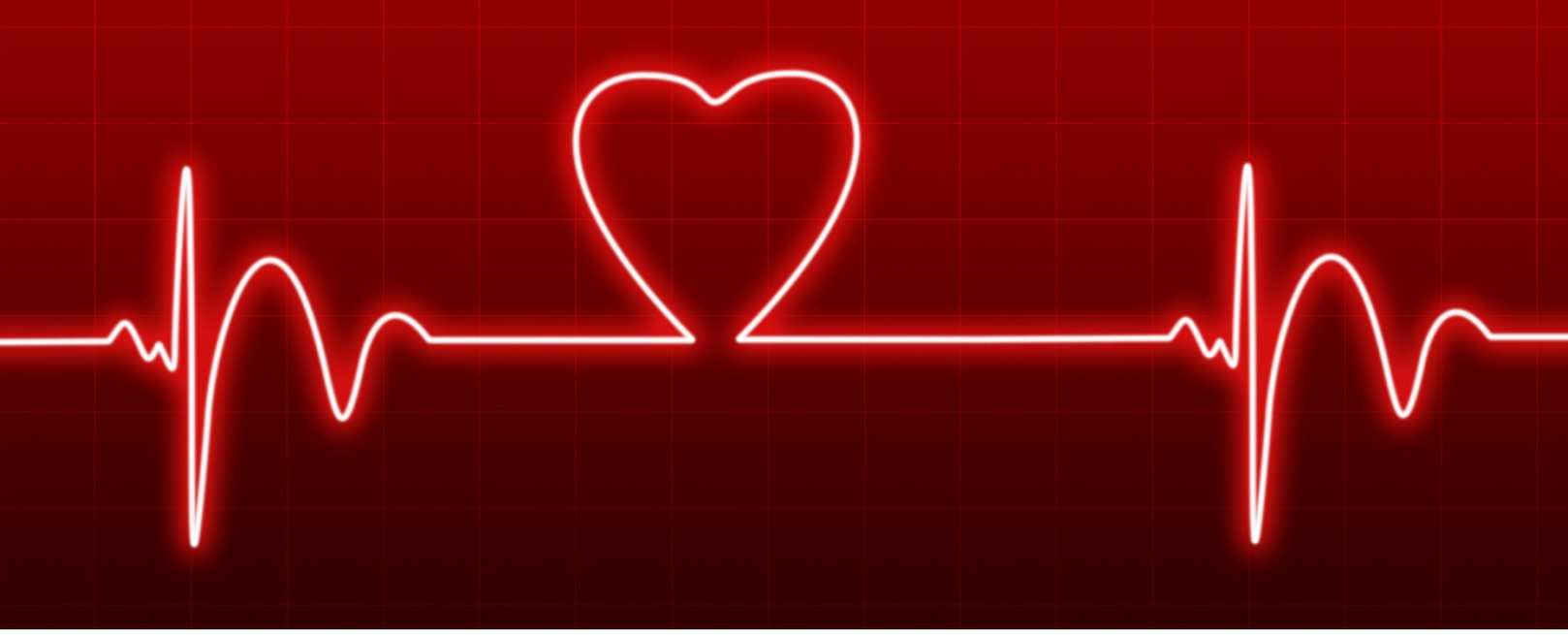
In de jaren ’90 verdween het Ladies Program geleidelijk aan bij de meeste conferenties, het raakte achterhaald, maar ik denk niet dat dit voorval daar mede de oorzaak van was.

Young statisticians

The Young Statisticians held a Statistics Café on the 8th of December on the topic of “Becoming an expert in Statistics & Data Science”. Laura Zwep, a PhD candidate at Leiden University, talked about her journey in academia and Vladimir Hazeleger, a Data Engineer and Consultant at Riviq, gave his perspective on what it’s like in the industry.

Next event

Our next event is going to be a company visit to NS (Nederlandse Spoorwegen). For more information sign up for our newsletter, follow us on LinkedIn or Instagram or keep an eye on our website: vvsor.nl/young-statisticians.



Een optimale meldingsstrategie voor burgerhulpverleners

Caroline Jagtenberg, Pieter van den Berg en Océane Fourmentraux

Elke dag krijgen 40 mensen in Nederland buiten het ziekenhuis een hartstilstand. Zij hebben over het algemeen een kleine overlevingskans, maar snelle reanimatie verbetert hun kansen aanzienlijk. Een manier om de tijd tot reanimatie te verkorten is door burgerhulpverlening.

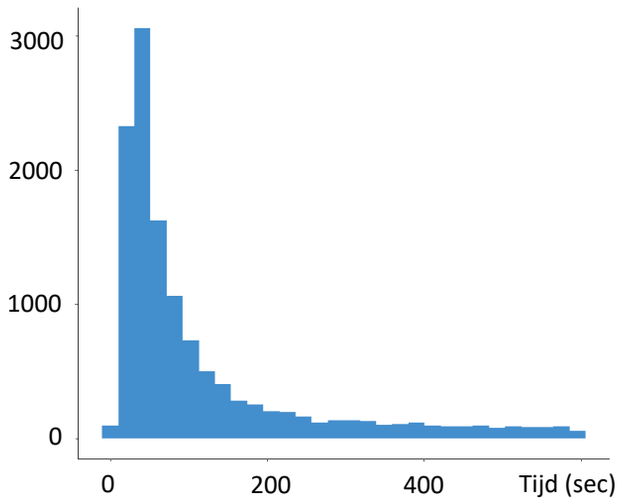
Burgerhulpverlening is het oproepen van getrainde vrijwilligers via een app. In Nederland bestaat de app HartslagNu, en veel andere landen hebben een vergelijkbaar systeem. Een vraag bij dit soort systemen is hoeveel vrijwilligers gealarmeerd moeten worden om de overlevingskans van de patiënt te bevorderen, zonder dat vrijwilligers overbelast raken. Kan het hierbij zinnig zijn om niet alle meldingen onmiddellijk te versturen, maar om dit met enige tussenpozen te doen?

Hoe werkt burgerhulpverlening?

Doorgaans is het de medewerker van de alarmcentrale die het systeem activeert, waarna meerdere vrijwilligers - automatisch gekozen op basis van hun GPS-locatie - een melding krijgen met instructies. Vrijwilligers kunnen zo'n melding accepteren of afwijzen, of het simpelweg niet opmerken en daarom niet reageren. Voor dit onderzoek hebben wij data gekregen van de GoodSAM app in Nieuw-Zeeland. Daaruit blijkt dat de tijd die verstrijkt tussen een melding en een reactie sterk uiteen loopt, met een piek rond de 30 seconden (Figuur 1).

Vanuit het perspectief van één patiënt levert het alarmeren van zoveel mogelijk vrijwilligers de hoogste overlevingskans op; frequente meldingen kunnen echter ook tot nadelen leiden. Als een vrijwilliger zich bijvoorbeeld ter plaatse haast en merkt dat er al meerdere andere burgerhulpverleners aanwezig zijn, is de kans kleiner dat hij of zij de volgende keer nog komt opdagen. Daarnaast kan een vrijwilliger die veel meldingen krijgt de app moe worden en deze deïnstalleren, of op zijn minst het geluid van de meldingen uitzetten. Daarom moet een 'optimale' strategie niet zomaar iedereen die in de buurt is oproepen, maar moet het overleven van de huidige patiënt zorgvuldig worden af-

gewogen tegen het overleven van toekomstige patiënten.

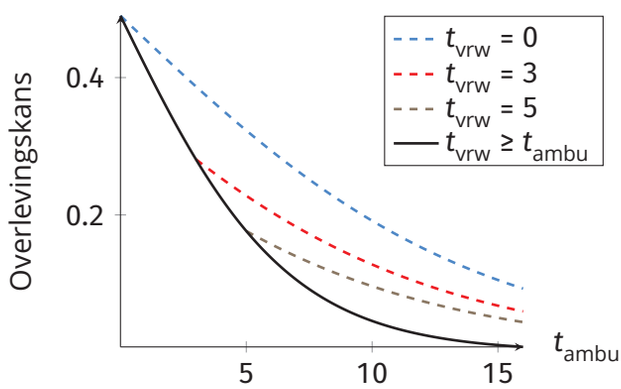


Figuur 1: Histogram van de verstreken tijd tussen melding en reactie van een burgerhulpverlener, gemeten voor alle meldingen tussen 2017 en 2020 in de GoodSam app in Nieuw-Zeeland

Model

We zoeken naar een zogenaamde meldingsstrategie: een strategie die aangeeft op welk moment we een melding moeten sturen naar welke vrijwilliger. Definieer tijdstip 0 als het moment waarop we weten dat het om een hartstilstand gaat en het systeem geactiveerd wordt. Hierbij veronderstellen we dat op dat moment de looptijd van elke vrijwilliger in de buurt bekend is. Dit is enigszins realistisch omdat in veel meldkamers de software nu ook al automatisch de looptijd van burgerhulpverleners berekent via een routeplanner. We gaan ervan uit dat een vrijwilliger een bepaalde tijd nodig heeft voordat hij of zij de melding ziet. Van deze tijd achten we alleen de verdeling bekend; neem bijvoorbeeld de histogram uit Figuur 1. De vrijwilliger accepteert vervolgens met een bepaalde kans: deze hang niet af van de geschiedenis van de vrijwilliger, noch van diens afstand tot de patiënt. Mogelijk hangt het wel af van hoe lang de melding al actief is op het moment dat deze gezien wordt (dit noemen we de reactietijd).

De overlevingskans van een patiënt hangt sterk af van de tijd totdat begonnen wordt met reanimatie. In de medische literatuur zijn er verschillende overlevingsfuncties beschikbaar die de tijd tot reanimatie vertalen naar een overlevingskans van de patiënt. De reanimatie start op het moment dat de eerste vrijwilliger arriveert bij de patiënt. De responstijd van een vrijwilliger bestaat uit: (1) de tijd die het het systeem kostte om de melding uit te sturen, plus (2) de reactietijd, plus (3) de looptijd. Voor het bepalen van de overlevingskans is het minimum van de responstijden over alle vrijwilligers van belang. Omdat naast de vrijwilligers ook een ambulance uitrukt die betere zorg kan bieden, hangt de overlevingskans ook af van de responstijd van de ambulance. Dit is te zien in Figuur 2.



Figuur 2: Overlevingskans van een patiënt met hartstilstand waarbij een vrijwilliger op tijd t_{vrw} begint te reanimeren en de ambulance arriveert op t_{ambu}. Alle tijden zijn in minuten

Dan is er nog de vraag hoe je het nadeel van te veel meldingen moet modelleren. Om te beginnen kun je simpelweg het *aantal* meldingen meten dat je in totaal verstuurt. Wij rekenen daarnaast ook het verwachte *aantal overtollige* vrijwilligers uit dat bij de patiënt zal aankomen. Voor elke mogelijke melding bestuderen we hoe deze getallen zich verhouden tot de marginale contributie van die vrijwilliger. Om deze contributie te berekenen moet je de verdeling van diens responstijd afzetten tegen de verdeling van de responstijd van alle andere vrijwilligers samen, en dat verschil vertalen naar een overlevingskans. Simpel gezegd: wat is de kans dat juist dié vrijwilliger het verschil gaat maken tussen leven en dood?

Mogelijke oplossing

Dit probleem hebben we voor het eerst geïntroduceerd in (Henderson e.a., 2022), waar we suggereerden dat het opgelost kan worden met dynamisch programmeren. Naast deze oplossingsrichting hebben we ook de mogelijkheid verkend dit probleem op te lossen met behulp van machine learning (Fourmentraux, 2022). Het algemene idee hierbij is om vooraf een flink aantal meldingsstrategieën te definiëren en het algoritme te trainen om op basis van de locaties van de vrijwilligers de juiste strategie te selecteren. Het selecteren van de juiste strategie doen we op basis van een beslisboom, of random forest. Het voordeel van het gebruik van een beslisboom is dat al het rekenwerk vooraf gebeurt en dat in real-time alleen de beslisboom afgelezen hoeft te worden.

Om het algoritme te trainen creëren we een groot aantal scenario's op basis van een vrijwilligersdichtheid. Daarnaast bepalen we de verschillende meldingsstrategieën die we willen beschouwen. Hoeveel en welke strategieën dat zijn, hangt af van de maximale complexiteit van de uiteindelijke beslisboom. Omdat de vrijwilligers als een homogene groep worden beschouwd is het nooit optimaal om een vrijwilliger te waarschuwen die ver weg is als er nog niet gealarmeerde vrijwilligers dichterbij zijn. We hoeven dus alleen te beslissen over het *aantal* gewaarschuwde vrijwilligers en de timing van de waarschuwingen.

Alvorens we aan machine learning beginnen, bepalen we voor elk scenario de optimale strategie. Dit doen we door voor elke strategie enkele prestatie-indicatoren (overlevingskans, het verwachte aantal meldingen, het verwachte aantal overtollige vrijwilligers dat arriveert) te simuleren. De reactietijden en het antwoord (acceptatie of afwijzing) simuleren we met behulp van de GoodSAM data. De optimale strategie voor elk scenario balanceert de verschillende prestatie-indicatoren zo goed mogelijk. Ten slotte trainen we zoals gezegd een beslisboom of random forest, die het classificatieprobleem oplost: deze voorspelt welke van de beschouwde strategieën optimaal gaat zijn voor een scenario dat je nog niet eerder gezien hebt.

Resultaten

Ter illustratie presenteren we hier een kleine casus waarin 6 vrijwilligers op minder dan 10 minuten lopen van de patiënt aanwezig zijn en waarbij elke vrijwilliger met 50% kans een inkomende melding accepteert. We beschouwen twee typen strategieën:

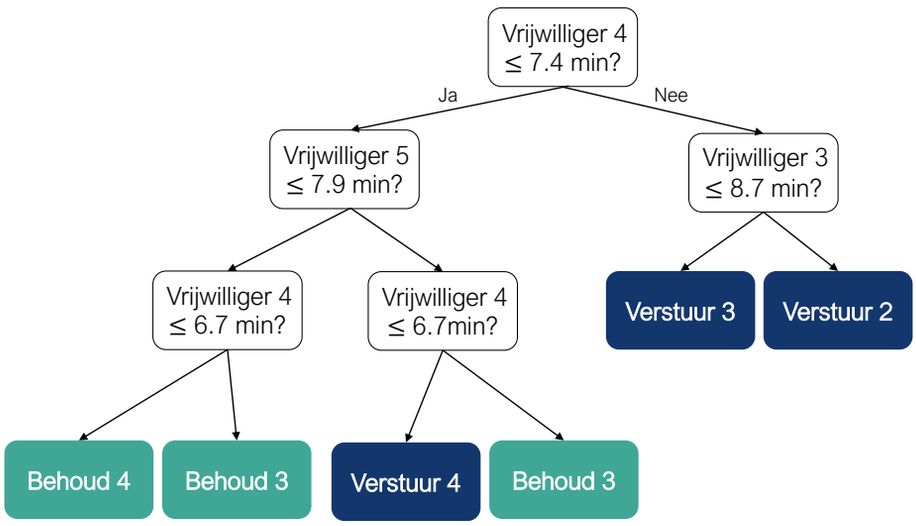
1. Verstuur X: alarmeer X vrijwilligers op tijdstip 0
2. Behoud X: alarmeer X vrijwilligers op tijdstip 0 en alarmeer een extra vrijwilliger zodra iemand een melding afwijst

In totaal geeft dit 11 strategieën, omdat X loopt van 1 tot en met 6 en voor X = 6 twee strategieën hetzelfde zijn.

Hoewel het random forest betere voorspellingen geeft, is een beslisboom beter interpreteerbaar en daarom laten we er zo een zien in Figuur 3. We zien hoe de optimale strategie afhangt van de looptijd van de verschillende vrijwilligers.

Figuur 3 geeft een voorbeeld van hoe een oplossing eruit zou kunnen zien. In deze boom zien we bij de knopen staan: "Vrijwilliger X", dit geeft de looptijd van de X^e dichtstbijzijnde vrijwilliger naar de patiënt. De eerste vertakking hangt dus af van of het systeem vier vrijwilligers binnen een straal van 7,4 minuten lopen gevonden heeft. We zien dat in de drie blauwe leafs de strategie 'Verstuur X' wordt aanbevolen. Dit gebeurt bijvoorbeeld aan de gehele rechterkant van de boom waarbij de vierde vrijwilliger van vrij ver moet komen. Intuïtief is dit als volgt te verklaren: omdat de extra vrijwilliger van ver moet komen en er ook al tijd verstreken is, weegt de marginale contributie in overlevingskans niet op tegen het versturen van een extra melding. In de drie groene leafs is de strategie 'Behoud X' optimaal, de adaptieve strategieën dus.

Het voorbeeld laat zien dat een strategie op basis van een beslisboom relatief eenvoudig is



Figuur 3: Beslisboom om te bepalen welke vrijwilliger(s) een melding moeten krijgen, voor een context waarin het acceptatiepercentage 50% is en er 6 vrijwilligers zijn in een straal van 10 minuten

uit te leggen aan app-beheerders en de beslisregels zouden kunnen worden opgenomen in de software die bepaalt wie een melding krijgt. Ook wanneer veel meer strategieën als input beschouwd worden, denken wij dat het een goed idee is om de uiteindelijke beslisboom beperkt te houden.

Je zou dit onderzoek op verschillende manier kunnen uitbreiden. Denk bijvoorbeeld aan vrijwilligers die verschillende vervoersmiddelen gebruiken. Of het alarmeren van extra vrijwilligers die niet rechtstreeks naar de patiënt gaan, maar ergens een AED ophalen.

Naast het optimaliseren van een meldingsstrategie zijn er meer manieren om burgerhulpsystemen te verbeteren. Zo publiceerde de Volkskrant laatst een idee voor HartslagNu: om te voorkomen dat vrijwilligers op moment suprême moeten werken met een interface waar ze nog helemaal niet bekend mee zijn, kan je ze laten wennen aan de app door middel van oefenmeldingen (Kuijk, 22-10-2022). Daarnaast is er in de medische literatuur ook al zeer veel aandacht voor dergelijke systemen. Die studies zijn vrijwel altijd retrospectief: ze volgen een groep patiënten en proberen patronen te ontdekken in het al dan niet overleven. Met dit artikel hopen wij te hebben laten zien dat het ook mogelijk is om deze systemen te bestuderen op een manier die zowel data-gedreven als proactief is.

Literatuur

O. Fourmentraux. „Volunteer dispatch strategies for cardiac arrest cases". Masterscriptie. Vrije Universiteit, 2022.

S. G. Henderson e.a. „How should volunteers be dispatched to out-of-hospital cardiac arrest cases?" In: *Queueing Systems* (2022), p. 1–3.

J. van Kuijk. „Volgens reanimatie-app Hartslagnu telt elke minuut, maar de app lijkt niet ontworpen om snel te kunnen helpen". In: *Volkskrant* (22-10-2022).

Caroline Jagtenberg werkt als universitair docent bij de afdeling Operations Analytics op de Vrije Universiteit Amsterdam. Ze zit ook in de redactie van STAtOR. C.j.jagtenberg@vu.nl

Pieter van den Berg is als universitair hoofddocent verbonden aan de Technology and Operations Management afdeling van de Rotterdam School of Management, Erasmus Universiteit Rotterdam.

Océane Fourmentraux behaalde haar master Econometrics & OR aan de Vrije Universiteit Amsterdam. In september 2022 begon ze als research assistant bij de afdeling Technology and Operations Management van Erasmus Universiteit Rotterdam.

De auteurs bedanken GoodSam New Zealand voor het delen van hun data.

Bernard Zweers

Column

Illustraties: Pixabay

Ode aan de efficiëntie

Rotterdam, het was al vrij snel duidelijk voor mij dat ik daar niet wilde studeren. De hoge gebouwen, de wegen met veel ruimte voor de auto, een koopgoot. Het kwam op mij over als een stad die was gebouwd om zo veel mogelijk mensen te kunnen huisvesten en je zo snel mogelijk te verplaatsen. Dat leek me nou niet echt gezellig. Nee, het werd Amsterdam waar de autoluwe grachten, de scheve huizen en de oude kroegen voor de sfeer zorgen. Een keus waar ik tot de dag van vandaag geen spijt van heb.

Ik was dan ook verheugd toen in het najaar van 2022 het boek *Rotterdam: een ode aan de ineffiëntie* van Arjen van Veelen (2022) uitkwam. In dit boek beschrijft hij hoe de stad Rotterdam te veel draait om efficiëntie. Het is toch prettig om iets te lezen wat je bevestigt in je mening. Dat gevoel draaide honderdtachtig graden toen ik las dat van Veelen de container als symbool gebruikte voor de focus op efficiëntie. Tijdens mijn promotie heb ik me bezig gehouden met het optimaliseren van container vervoer. Ik was in twijfel: was ik het nu wel of niet eens met de strekking van het boek?

Het werd tijd om het boek te lezen en te kijken wat van Veelen tegen efficientië had. De vooroorlogse spreuk van Rotterdam, *Navigare necesse est*, komt geregeld terug in het boek. Van Veelen vertaalt haar vrijelijk als *schip moet varen*. Alles draait in de Rotterdamse haven om een schip zo veel mogelijk te laten varen. De bemanning van een schip brengt soms maanden achter elkaar doo op zee. Opgesloten in hun hut raken veel zeelieden depressief. Aangezien het laden en lossen van een schip met containers zo snel gaat en een schip moet varen, is de tijd aan land voor de meesten zeer beperkt.

Een ander aspect dat van Veelen vaak aanhaalt is just-in-time management. Dit concept dat ooit door Toyota is bedacht, en zorgt dat de fabriek geen voorraden meer nodig heeft omdat die precies op tijd door de leveranciers worden aangeleverd. Deze methode lijkt ervoor te zorgen dat er minder producten nodig zijn, maar van Veelen claimt dat het juist voor een toename van verspilling zorgt. De macht van Toyota zorgt ervoor dat toeleveranciers er alles

aan doen om genoeg en op tijd te kunnen leveren. Hierdoor onstaat een heel pyramidespel van allerlei onderleveranciers waarin verspilling niet wordt geminimaliseerd, maar wordt afgewendeld op anderen.

Het eigen leven van Veelen wordt ook gedomineerd door efficientië met een oppas, een schoonmaakster en een boodschappenbezorger. Allemaal simpel te regelen met een app en ze zorgen ervoor dat het gezin van Veelen de tijd zo goed mogelijk kon besteden. Gaat er echter iets mis in één van deze schakels, omdat bijvoorbeeld de oppas ziek is of een tijdslot van een bezorger al vol is, dan loopt alles in de soep.

Overdonderd door al deze voorbeelden, ging ik eens peilen hoe mijn omgeving dacht over het woord 'efficiëntie'. Vrijwel iedereen had een negatieve connotatie bij dat woord. Wat toch eigenlijk wel vreemd is, want als je iets efficiënt organiseert dan heb je minder middelen nodig om een bepaald doel te halen dan als je iets minder efficiënt hebt geregeld. Hoe kon je daar nu iets tegen hebben? Het hele vakgebied Operations Research gaat eigenlijk over efficiëntie. Hadden mijn vrienden en familie nu slechte gevoelens bij het werk dat ik deed?

Als je de voorbeelden van van Veelen leest of in het nieuws hoort over doorgeslagen efficiëntie in het onderwijs, de zorg of de rechtspraak, is het niet gek dat je een negatieve associate krijgt met het woord efficiëntie. Alleen zijn dit allemaal voobeelden van slechte optimalisatie, namelijk van *ééndimensionale, deterministische, indivuele efficiëntie*. Wat ik hiermee bedoel zal ik kort uitleggen.

Eéndimensionale efficiëntie focust zich alleen maar op het optimaliseren van een enkel criterium. Eigenlijk is dit altijd het minimaliseren van de kosten of het maximaliseren van de winst. Dit gaat ten koste van bijvoorbeeld de arbeidsomstandigheden van de zeelieden. Je zou natuurlijk ook meerdere dimensies kunnen optimaliseren waardoor de extreme uitkomsten niet meer voorkomen.

Met deterministische efficiëntie wordt niet of onvoldoende rekening gehouden met onzekerheid. Het schema van de familie van Veelen houdt te weinig rekening met een verstoring in één van de schakels. Van Veelen zegt zelf in het boek dat zijn agenda voelt als blokken die zo efficiënt mogelijk op elkaar gestapeld moeten worden, maar daardoor is er geen ruimte om onverwachte gebeurtenissen op te vangen.

Ten slotte is het voorbeeld van just-in-time management, een goede illustratie van wat er mis gaat als je te individueel optimaliseert. Bij het alleen optimaliseren van een enkele schakel in de keten, krijg je oplossingen die hoogstwaarschijnlijk niet optimaal zijn voor het gehele netwerk.

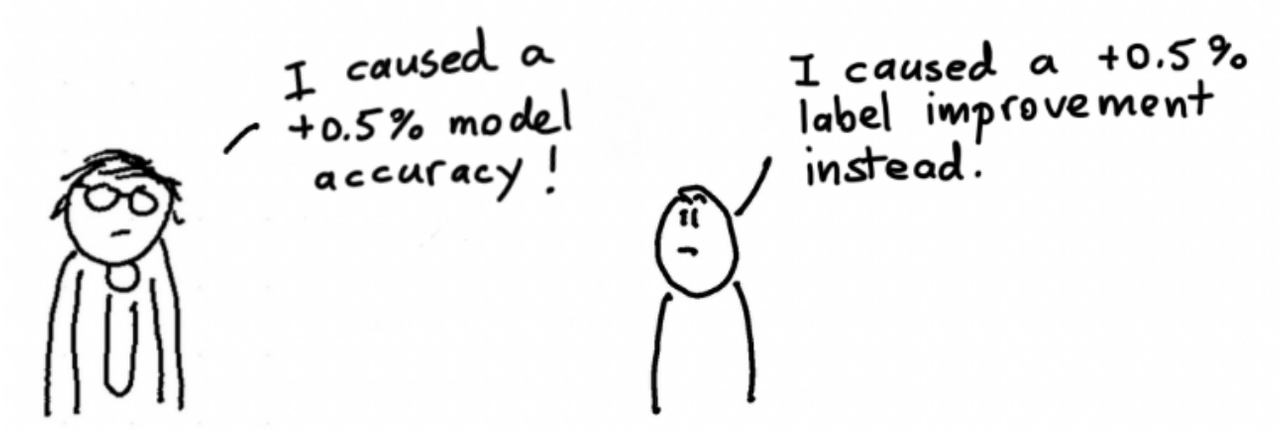
Voor de OR-professional voelen de bovenstaande zinnen hopelijk als een boel open deuren die worden ingetrapt. Met allerlei technieken, zoals stochastische, robuste of multicriteria optimalisatie kunnen we meerdimensionale, stochastische en holistische efficiëntie nastreven. Toch kunnen we ons afvragen of dat ook daadwerkelijk gebeurt in het dagelijks leven. Hoeveel van de artikelen die je recent hebt gelezen behandelen ten minste twee van de onderdelen van goede efficiëntie? In hoeveel projecten worden meerdere onderdelen in de praktijk gebracht?

Minstens zo belangrijk als bewustzijn binnen ons vakgebied, is dat het algemene publiek ook bekend is met het feit dat wiskundige modellen niet per se tot slechte efficiëntie leiden. Sterker nog, misschien kunnen de huidige voorbeelden van slechte efficiëntie wel het beste worden opgelost met meer wiskundige optimalisatie. Als we dat als gemeenschap voor elkaar krijgen, heeft het volgende boek dat over Rotterdam (of een andere Nederlandse stad) gaat, hopelijk als ondertitel: ode aan de efficiëntie.

Literatuur

A. van Veelen. *Rotterdam. Een ode aan inefficiëntie*. De Correspondent, 2022.

Bernard Zweers Doing the Math
email: bernard@doingthemath.nl



https://koaning.io/posts/labels/

Bad labels

Vincent Warmerdam

Supervised machine learning is een krachtig gereedschap om automatisch verbanden tussen kenmerken en labels te vinden. De aandacht van de onderzoeker is verschoven van het specificeren van modellen en controleren van modelaannames naar het finetunen van hyperparameters. Daarbij wordt wel eens vergeten hoe groot de invloed van foutief gelabelde data is.

Onderstaande blogpost over foutieve labels is overgenomen van koaning.io met Vincent's toestemming.

Bit of background

It turns out that bad labels are a huge problem in many popular benchmark datasets. To get an impression of the scale of the issue, just go to labelerrors.com. It's an impressive project that shows problems with many popular datasets; CIFAR, MNIST, Amazon Reviews, IMDB, Quickdraw and Newsgroups just to name a few. It's part of a research paper (Northcutt, Athalye en Mueller, 2021) that tries to quantify how big of a problem these bad labels are. The table from the paper gives a nice summary. It's a huge problem.

The issue here isn't just that we might have bad labels in our training set, the issue is that they appear in the validation set. If a machine

learning model can become state of the art by squeezing another 0.5% out of a validation set one has to wonder: are we really making a better model? Or are we creating a model that is better able to overfit on the bad labels?

Another dataset

The results from the paper didn't surprise me much, but it did get me wondering how easy it might be for me to find bad labels in a dataset myself. After a bit of searching I discovered the Google Emotions dataset (Demszky e.a., 2020). This dataset contains text from Reddit (so expect profanity) with emotion tags attached. There are 28 different tags and a single text can belong to more than one emotion

The dataset also has a paper about it which explains how the dataset came to be (Demszky e.a., 2020). It explains what steps have been taken to make the dataset robust.

- There are 82 raters involved in labeling this dataset. Each example should have at least 3 people checking it. The paper mentions that all the folks who rated were from India but spoke English natively.
- An effort was made to remove subreddits that were not safe for work or that contained too much vulgar tokens (according to a predefined word list).
- An effort was made to balance different subreddits such that larger subreddits wouldn't bias the dataset.
- An effort was made to remove subreddits that didn't offer a variety of emotions.
- An effort was made to mask names of people as well as references to religions.
- An effort was made to, in hindsight, confirm that there is sufficient interrated correlation.

All of this amounts to quite a lot of effort indeed. So how hard would it be to find bad examples here?

Quick trick

Here's a quick trick that seems worthwhile. Let's say that we train a model that is very general. That means high bias, low variance. You may have a lower capacity model this way, but it will be less prone to overfit on details.

After training such a model, it'd be interesting to see where the model disagrees with the training data. These would be valid candidates to check, but it might result in a list that's a bit too long for comfort. So to save time you can sort the data based on the `predict_proba()`-value. When the model gets it wrong, that's interesting, but when it *also* associates a very low confidence to the correct class, that's an example worth double checking.

So I figured I would try this trick on the Google emotions dataset to see what would happen. I tried predicting a few tags chosen at random and tried using this sorting trick to see how easy it was to find bad labels. For each tag, I would apply my sorting to see if I could find bad labels in the top 20 results. Figure 1 shows some of the results. I don't know about you, but many of these examples seem wrong.

Friggin' strange

Before pointing a finger, it'd be good to admit that interpreting emotion isn't a straightforward task. At all. There's context and all sorts of cultural interpretation to consider. It's a tricky task to define well.

The paper also added a disclaimer to the paper to make people aware of potential flaws in the dataset. Here's a part of it:

> *"We are aware that the dataset contains biases and is not representative of global diversity. We are aware that the dataset contains potentially problematic content. Potential biases in the data include: Inherent biases in Reddit and user base biases, the offensive/vulgar word lists used for data filtering, inherent or unconscious bias in assessment of offensive identity labels, annotators were all native English speakers from India. All these likely affect labeling, precision, and recall for a trained model."*

Adding this disclaimer is fair. That said. It really feels just a bit too weird that it was that easy for me to find examples that really seem

Label = 'love'

- Weird game lol
- Looks like it. I didn't make it, I just found it.
- Wow, you people...

Label = 'not love'

- Very nice!! I love your art! What journal is this? I love the texture on the pages.
- love love love this. so happy for the both of you.
- I LOVE IT, I would love if they will make season 2... I really enjoyed it
- I love this, my wife told me about something she read on reddit yesterday and I was like.... well just like ol [NAME] here!!!

Label = 'curiosity'

- I actually enjoy doing this on my own. Am I weird?
- She probably has a kid by now.
- So much time saved. Not.
- Didn't you just post this and people told you it was dumb and not meant for this sub?

Label = 'not curiosity'

- I cant wait. I'm curious if it will give us any more insight into the incident other than what we already know.
- Why do you guys hate [NAME]? I'm neutral leaning slightly positive on him. Just curious why the strong negative opinion?
- What does that even mean? How does one decide right or wrong with something so vague?
- Wait, this is actually a really interesting point. That could/should play a factor if he's a legitimate candidate.
- Is it weed? I'm curious to ask if you know what weed smells like?

Label = 'not excitement'

- I am inexplicably excited by [NAME]. I get so excited by how he curls passes
- Omg this is so amazing ! Keep up the awesome work and have a fantastic New Year !
- I just read your list and now I can't wait, either!! Hurry up with the happy, relieved and peaceful onward and upward!! Congratulations 😍
- I absolutely love that idea. I went on an anniversary trip with a couple once and it was amazing! We had so much fun.
- Happy New Year! Looks like you had a great time there! Cheers! Here's to a great 2019 hopefully in both baseball and life!

Label = 'not joy'

- Happy cake day! Have a great day and year, cheers.
- It's wonderful and gives me happy happy feels
- Happy to hear this exciting news. Congratulations on your fun-filled morning.
- It's good, good, good, good - good good good!
- My son and I both enjoy taking pictures. It gives us pleasure. Part of the fun for us on vacation is taking pictures of new things.

Label = 'not gratitude'

- Thanks. Nice input as always.
- Thanks. I didn't quite get it from the original. Appreciate the time.
- This made my hump day. Thank you good sir
- Excellent work thank you for this. This is why I love Reddit.
- You're amazing thank you so much!! :)

Figuur 1

so clearly wrongly labeled. I didn't run through the whole dataset, so I don't have a number on the amount of bad labels but I'm certainly worried now. Given the kind of label errors, I can certainly imagine that my grid search results are skewed.

What does this mean?

The abstract of the paper certainly paints a clear picture of what this exercise means for state-of-the-art models:

> *"We find that lower capacity models may be practically more useful than higher capacity models in real-world datasets with high proportions of erroneously labeled data. For example, on ImageNet with corrected labels: ResNet-18 outperforms ResNet-50 if the prevalence of originally mislabeled test examples increases by just 6%. On CIFAR-10 with corrected labels: VGG-11 outperforms VGG-19 if the prevalence of originally mislabeled test examples increases by 5%. Traditionally, ML practitioners choose which model to deploy based on test accuracy – our findings advise caution here, proposing that judging models over correctly labeled test sets may be more useful, especially for noisy real-world datasets."*

So what now?

More people should check their labels more frequently. Anybody is free to try out any trick that they like, but if you're looking for a simple place to start, check out the cleanlab (https://github.com/cgnorthcutt/cleanlab) project. It's made by the same authors of the labelerrors-paper and is meant to help you find bad labels. I've used it a bunch of times and I can confirm that it's able to return relevant examples to double-check.

Here's the standard snippet that you'd need:

```
from cleanlab.pruning import get_noise_indices

# Find label indices
ordered_label_errors = get_noise_indices(
    s = numpy_array_of_noisy_labels,
    psx = numpy_array_of_predicted_probabilities,
    # Order label errors
    sorted_index_method = 'normalized_margin',
)

# Use indices to subset dataframe
examples_df.iloc[ordered_label_errors]
```

It's not a lot of effort and it feels like such an obvious thing to check going forward. The disclaimer on the Google Emotions paper checks a lot of boxes, but imagine that in the future they'd add 'we checked out labels with cleanlab before releasing it'. For a dataset that's meant to become a public benchmark, it'd sure be a step worth adding.

For everyone; maybe we should spend less time tuning parameters and instead spend it trying to get a more meaningful dataset. If working at Rasa is teaching me anything, it's that this would be time well spent.

Literatuur

D. Demszky e.a. „GoEmotions: A Dataset of Fine-Grained Emotions". In: *CoRR* abs/2005.00547 (2020). arXiv: 2005.00547. URL: https://arxiv.org/abs/2005.00547.

C. G. Northcutt, A. Athalye en J. Mueller. „Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks". In: *arXiv* (2021). DOI: 10.48550/ARXIV.2103.14749. URL: https://arxiv.org/abs/2103.14749.

Vincent Warmerdam is a senior data professional who worked as an engineer, researcher, team lead, and educator in the past. He is especially interested in understanding algorithmic systems so that one may prevent failure. As such, he has had a preference for simpler solutions that scale, as opposed to the latest and greatest from the hype cycle. He was an invited speaker at the 18th January NGB/LNMB Seminar "Fairness in Machine Learning and Operations Research".

Jelke Bethlehem

Column

Over de gevaren van zelfselectie-peilingen

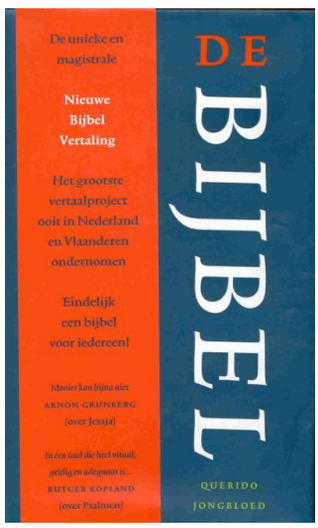
Als je een peiling wilt uitvoeren, dan moet je een steekproef trekken uit de te onderzoeken groep (de doelpopulatie). Dat kun je op verschillende manieren doen. Voor een goede, representatieve peiling moet je een kanssteekproef *trekken. Je moet de steekproef dus* loten *uit die doelpopulatie. We noemen dit een* aselecte steekproef*. Helaas gebeurt het maar al te vaak dat de steekproef niet aselect is, maar dat hij tot stand is gekomen via* zelfselectie*.*

Dan kan iedereen die dat wil, de vragenlijst invullen. Dus ook personen die niet behoren tot de doelpopulatie. Soms kun je de vragenlijst zelfs meer dan één keer invullen. Dit leidt allemaal tot steekproeven die meestal verre van representatief zijn. Maar misschien nog belangrijker is het dat groepjes personen de uitkomsten van de peiling kunnen manipuleren. Zo kunnen ze de peiling naar hun hand zetten met gericht gekozen antwoorden op de vragen. In deze bijdrage geven we een paar voorbeelden van gemanipuleerde zelfselectie-peilingen.

De nieuwe Bijbelvertaling

We beginnen met een al wat ouder voorbeeld. Dat is de *NS Publieksprijs*. In 2005 werd de winnaar via een online zelfselectie-peiling bepaald. Je kon niet alleen stemmen op één van de zes genomineerde boeken, maar je kon ook zelf een ander boek opgeven. In totaal brachten 92.000 mensen hun stem uit. Tot verbazing van iedereen werd niet een van de genomineerde boeken tot winnaar gekozen. 72% van de stemmers koos voor de nieuwe Bijbelvertaling. Zie figuur 1. Deze uitslag was het resultaat van een campagne gevoerd door onder anderen het dagblad Trouw, de Evangelische Omroep,

het Nederlands Bijbelgenootschap, de Katholieke Bijbelstichting en de Protestantse Kerk om te stemmen op de nieuwe Bijbelvertaling. De gang van zaken was niet tegen de regels. Het is echter wel duidelijk dat de stemmers geen goede afspiegeling vormden van de Nederlandse bevolking. Er was dus geen sprake van een aselecte steekproef. Het was een geval van zelfselectie.



Figuur 1: De nieuwe Bijbelvertaling won in 2005 de NS-Publieksprijs

Was 'moestuinsocialisme' toch woord van het jaar?

Op 16 december 2014 maakte de redactie van de *Dikke Van Dale* bekend dat 'dagobertducktaks' was gekozen tot *Woord van het Jaar 2014*. In een online peiling met zelfselectie kon je kiezen uit 10 genomineerde woorden. 60.000 mensen deden mee aan de verkiezing. Het woord 'dagobertducktaks' kwam op de eerste plaats met 18% van de stemmen. Op de tweede plaats eindigde 'moestuinsocialisme' met 14% van de stemmen. Klopte deze uitslag wel? Was hij wel representatief voor de Nederlandse bevolking.

Auteur en taalkundige *Wim Daniëls* (zie figuur 2) vond het maar niks dat 'dagobertducktaks' had gewonnen. In het journaal van de NOS (het achtuurjournaal van 16 december 2014) vertelde hij dat de peiling in zijn ogen niet eerlijk was geweest omdat het FNV haar leden had opgeroepen om op dit woord te stemmen. Dat hadden ze inderdaad in grote getale gedaan. En zo won 'dagobertducktaks'. Wim Daniëls had gehoopt dat 'moestuinsocialisme' zou winnen. Hij had op dat woord gestemd. Dat woord was in zijn ogen de echte winnaar.

Figuur 2: Wim Daniels hoopte dat 'moestuinsocialisme' woord van het jaar zou worden en niet 'dagobertducktaks'

Het item in het journaal wekte de suggestie op dat de verkiezing niet eerlijk was verlopen. En de schuld daarvan werd gelegd bij het FNV die haar leden had opgeroepen op 'dagobertducktaks' te stemmen en niet op 'moestuinsocialisme'. Die oproep was echter niet tegen de regels. Wel kun je concluderen dat je zo geen representatieve peiling krijgt. Dit was een gemanipuleerde zelfselectie-peiling.

Willen natuurliefhebbers vuurwerk verbieden?

Een derde voorbeeld van een online peiling met zelfselectie troffen we aan bij het programma *Vroege Vogels* van de VARA. Tegen het einde van 2014 werd in dit programma uitgebreid aandacht besteed aan vuurwerkoverlast. Het programma leek er vanuit te gaan dat natuurliefhebbers wel tegen vuurwerk zouden zijn. Dat bleek uit de behoorlijk sturende vraagstelling in een peiling. Zie figuur 3.

Op de website van *Vroege Vogels* werd een zelfselectie-peiling geplaatst waarin de bezoekers zich konden uitspreken over een verbod op particulier vuurwerk. Aanvankelijk verliep de peiling naar verwachting. Zo'n 90% van de deelnemers wilde een verbod op vuurwerk. Maar toen kwam er een kentering. Ineens waren er

duizenden tegenstanders van een vuurwerkverbod. Op zondagmorgen 28 december 2014 hadden bijna 5.000 mensen hun stem uitgebracht. Nog maar 46% was voor een vuurwerkverbod en een meerderheid van 53% was tegen. Dat was een behoorlijke omwenteling. Hoe kon dat? Enig onderzoek leerde dat ook de voorstanders van vuurwerk de website hadden ontdekt. Een voorbeeld daarvan was de website *Freakpyromaniacs.com*. In het forum op deze website werden vuurwerkliefhebbers opgeroepen naar de website van *Vroege Vogels* te gaan en te stemmen tegen een vuurwerkverbod. En dat deden ze dus.

Figuur 3: De sturende vraagstelling in de anti-vuurwerk-peiling

Je kunt je afvragen wat nu precies de doelpopulatie van deze peiling was. Waren het alleen de kijkers/luisteraars naar *Vroege Vogels*? Of waren het alle Nederlanders? In ieder geval kon iedereen meedoen aan de peiling. Hoe dan ook, het lijkt erop dat vuurwerkliefhebbers behoorlijk oververtegenwoordigd waren. Dit is dus ook weer een voorbeeld van een manipuleerde peiling.

Is 'prikspijt' echt het woord van het jaar in 2021?

Op 21 december 2021 werd 'prikspijt' door woordenboekmaker *Van Dale* uitgeroepen tot het *Woord van het Jaar 2021*. Het woord kreeg in een online peiling ruim 82% van de in totaal 49.000 uitgebrachte stemmen. Dat is een bijzonder hoog percentage. Het was zo hoog dat het argwaan opriep. Was het echt zo dat 82% van alle Nederlanders vond dat het relatief onbekende woord 'prikspijt' het Woord van het Jaar was? Kortom was de peiling wel representatief?

Nadere beschouwing leert dat de peiling verkeerd was opgezet. Voor een goede, representatieve peiling had je de steekproef netjes moeten loten uit de hele bevolking. Deze peiling was echter gebaseerd op zelfselectie. Iedereen die dat wilde, kon meedoen aan de peiling. En er was een groot gevaar op manipulatie van de peiling.

Winnaar van deze verkiezing werd, met een overweldigende meerderheid van 82,2%, het woord 'prikspijt'. Op de tweede plaats, met veel minder stemmen, eindigde 'woonprotest' met 3,7% van de stemmen. Op de derde plaats kwam 'wappiegeluid' met 3,6%.

Waar kwam dat overweldigende aantal stemmen voor 'prikspijt' vandaan? Nader onderzoek leert dat in de sociale media door veel mensen was opgeroepen om toch vooral op dit woord te stemmen. Het leek erop dat de stemmers op 'prikspijt' vooral uit de hoek van de *antivaxers* kwamen. Dus deze groep was zwaar oververtegenwoordigd in de peiling. Daarom moeten we concluderen dat ook deze peiling niet representatief was. Hij was weer een gemanipuleerde zelfselectie-peiling.

Figuur 4: In 2021 werd 'prikspijt' gekozen tot Woord van het Jaar

Hoe de peiling voor de NS-Publieksprijs in 2022 ontspoorde

Ook in 2022 organiseerde de NS weer een online peiling met zelfselectie voor de NS Publieksprijs. Op de website van de prijs kon je een keuze maken uit zes genomineerde boeken. En zat je favoriete boek er niet bij, dan kon je de titel en/of auteur van dat boek alsnog intypen.

Om de stemmers te kunnen melden dat ze hun stem hadden uitgebracht, moesten ze hun e-mailadres opgeven. Daarna kregen ze een e-mail met daarin de bevestiging van hun keuze. Het ging helemaal mis met deze stemprocedure. In plaats van je eigen e-mailadres, kon je ook het e-mailadres van iemand anders opgeven. En dus kreeg die andere persoon dan bericht. Zo meldde Caroline van der Plas (Tweede Kamerlid van BBB) op Twitter dat ze niet had gestemd, maar dat wel iemand anders haar e-mailadres had gebruikt. Ze kreeg de bevestiging dat ze op een boek van Thierry Baudet ('Het Corona Bedrog') had gestemd. En dat terwijl ze helemaal niet had gestemd. Ook andere Kamerleden melden misbruik van hun e-mailadres.

Kennelijk was er sprake van een groep stemmers die probeerde de uitslag te manipuleren. Ze probeerden zoveel mogelijk stemmen te krijgen voor het (niet-genomineerde) boek van Thierry Baudet. Na veel klachten werd de peiling door de NS stopgezet: "we begrijpen de commotie en zoeken grondig uit wat er aan de hand is. We komen spoedig met een oplossing en tot die tijd sluiten we tijdelijk de stemmodule".

Na verloop van enige tijd werd de stemmodule voor de NS Publieksprijs weer geopend: "stemmen die vanaf dit moment worden uitgebracht zijn rechtmatig. Bij reeds uitgebrachte stemmen wordt om verificatie van de stem gevraagd: zolang die verificatie niet wordt gegeven zijn deze reeds uitgebrachte stemmen ongeldig".

Na deze reparatie was het niet meer mogelijk om e-mailadressen van anderen te misbruiken. Wat bleef was dat hier sprake was van een peiling op basis van zelfselectie en niet van een aselecte steekproef. Dit liet nog steeds de mogelijkheid open om meer dan één keer te stemmen of met groepjes mensen proberen de uitslag te manipuleren. En je kunt je ook afvragen wat precies de doelgroep was. Er konden ook buitenlanders (bijvoorbeeld Vlamingen) mee doen. De stemmers vormen daarom geen representatieve selectie uit de Nederlandse bevolking.

Uiteindelijk besloot de NS helemaal te stoppen met de peiling. Er werd (terecht) geen winnaar bekend gemaakt. Niemand won de NS-Publieksprijs.

Figuur 5: Peiling voor de NS-Publieksprijs 2022

Conclusie: vermijd zelfselectie-peilingen!

Deze voorbeelden tonen de gevaren van online peilingen die zijn gebaseerd op zelfselectie. De boodschap is duidelijk. Daarom moet je heel voorzichtig zijn als je de uitkomsten van zulke peilingen wilt gebruiken. Het zou heel goed kunnen dat de uitkomsten gemanipuleerd en dus ernstig vertekend zijn. Als het bij een online peiling niet duidelijk is hoe de steekproef is getrokken, kun je dit beter eerst uitzoeken. En als er geen informatie beschikbaar is, dan kun je maar beter niet al te veel waarde hechten aan de uitkomsten van een online peiling.

Jelke Bethlehem is expert op het gebied van steekproeven, vragenlijsten en weergave van onderzoeksresultaten.
email: mail@jelkebethlehem.nl