Nicolaos Starreveld
FNWI
Universiteit van Amsterdam
n.j.starreveld@uva.nl

Hans Sterk
Faculteit Wiskunde en Informatica
Technische Universiteit Eindhoven
h.j.m.sterk@tue.nl

Interview Edwin van den Heuvel

Data-analyse COVID-19

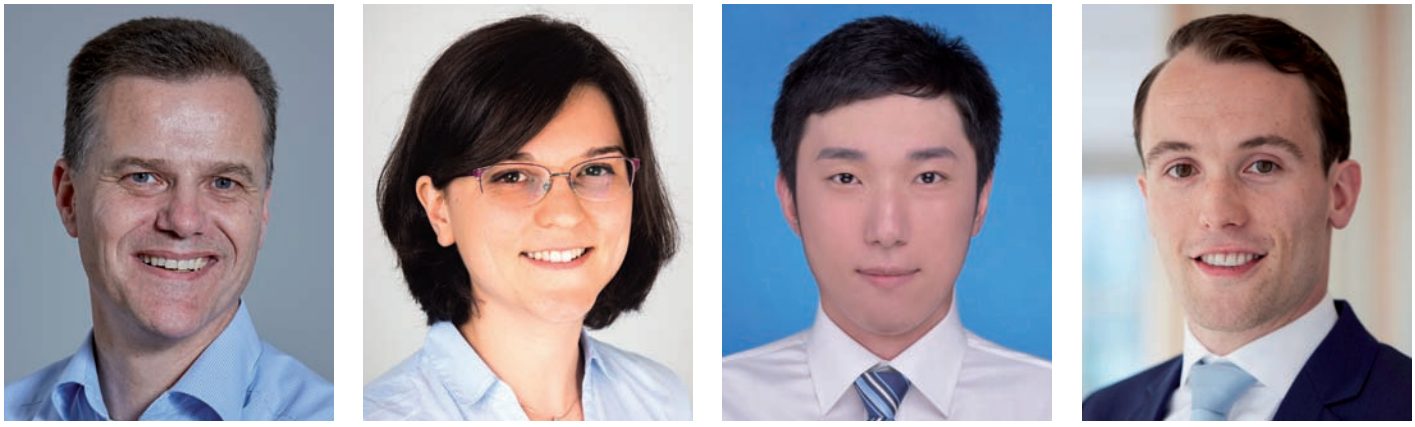
Edwin van den Heuvel, hoogleraar statistiek aan de Technische Universiteit Eindhoven, aarzelde niet toen de corona-epidemie om zich heen greep, en ging met zijn teamleden aan de slag. Ze verzamelden dagelijks data, evalueerden de data, beoordeelden de data op betrouwbaarheid, bouwden en sleutelden aan hun model en gebruikten het om vooruit te kunnen kijken. De dagelijkse voorspellingen die Van den Heuvel en zijn team maakten, bleken zeer accuraat te zijn, en leidden tot samenwerking met meerdere ziekenhuizen die de uitkomsten gebruikten om de benodigde capaciteit beter in te schatten. Op 8 mei spraken onze redacteuren Nicolaos Starreveld en Hans Sterk met Van den Heuvel die in dit interview het verhaal van de afgelopen weken vertelt.

Je bent dus begonnen met het verzamelen van data. Waar haal je die data vandaan?
"Daar hebben we meerdere bronnen voor gebruikt, we begonnen met de data die de Johns Hopkins University publiceert op haar website. Maar al snel zijn we ook naar andere bronnen en allerlei officiële websites gaan kijken. Bijvoorbeeld naar de websites van het RIVM en van vergelijkbare instanties in landen als Frankrijk en Italië. Die bronnen bekeken we allemaal, ook om te zien of er geen tegenstrijdigheden zaten in de cijfers. Die tegenstrijdigheden bleken we inderdaad wel tegen te komen."

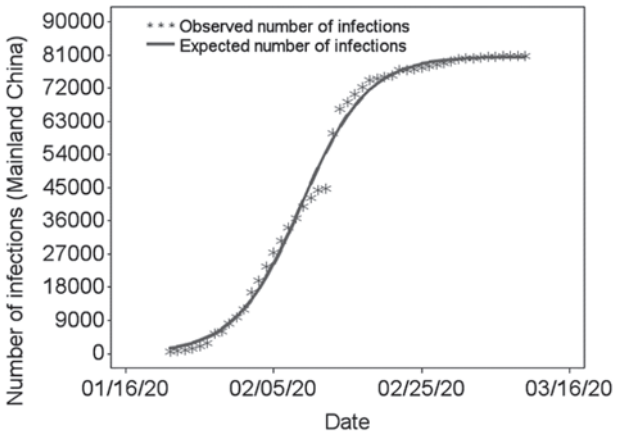
Hoe beoordeel je de kwaliteit van de data?
"Dat is lastig. We hebben een aantal dingen wel een keer gezien. Op een gegeven moment zagen we bij Johns Hopkins opvallende verdubbelingen; die verdubbelingen traden op omdat ze vanaf zeker moment ook alle verzorgingshuizen mee waren gaan tellen. Die stonden dus niet in de officiële cijfers, maar wel in de cijfers van Johns Hopkins. Daar hebben we toen niet voor gekozen. We hebben dus keuzes moeten maken om te zien wat we zouden modelleren."

En wat voor middelen gebruiken jullie om de data te analyseren?
"We gebruiken twee softwarepakketten, R en SAS. En dat komt omdat we in het begin gezien hebben dat als je niet te veel data hebt, de schattingen met onze modellen redelijk gevoelig zijn voor startwaardes. Om uiteindelijk de parameters te kunnen schatten moet je een optimalisatieprobleem oplossen. En de verschillende methodes zijn relatief gevoelig voor keuzes die je daarin maakt. We draaiden altijd met verschillende pakketten en met verschillende numerieke methoden binnen die pakketten om uit te sluiten dat een gevonden optimum op een toevalligheid berustte."

En wat voor modellen gebruiken jullie?
"In het begin zijn we gestart met de 'Verhulst-curve', dit betreft een model met drie parameters. Eén parameter is het maximum, de waarde waarop de curve is afgevlakt, wat we proberen te schatten. Er zit een parameter in die de toename beschrijft, de groeisnelheid. En er zit nog een derde parameter in, het tijdstip waarop je halverwege bent. We zijn ermee begonnen omdat China liet zien dat je eigenlijk geen complexere modellen zou hoeven nemen dan Verhulst. Maar later, toen we bezig waren met Europa, kwamen we erachter dat dat voor Europa toch niet het geval was. In het begin leek Europa redelijk de Verhulst-curve te volgen, maar de manier



Edwin van den Heuvel en zijn teamleden Marta Regis (postdoc), Zhuozhao Zhan (UD) en Richard Post (PhD)



Figuur 1 Model toegepast op data uit China.

waarop de afvlakking in Europa plaatsvond, gebeurde niet volgens Verhulst. Het Verhulst-model is symmetrisch, maar in Europa zie je deze symmetrie niet. Toen we op het eind van die curves kwamen, zagen we dat de afvlakking die Verhulst voorspelde veel sneller gaat dan wat er eigenlijk in werkelijkheid gebeurde. Dus toen hebben we andere logistische curves gepakt die meer parameters hebben, om ook dat fenomeen te kunnen volgen en voorspellen. Voor het geval van Europa hebben we twee parameters toegevoegd in het model. We begonnen met Verhulst, maar toen we tekortkomingen constateerden, voegden we twee parameters toe, en vervingen beide factoren door een macht, deze extra parameters zijn dus een soort power-parameters.

$$\bar{Y}'(t) = k\bar{Y}(t)^{\gamma}\left(1-\frac{\bar{Y}(t)}{M}\right)^{\eta}, \quad \gamma,\eta > 0. \quad (2)$$

Toen we begonnen waren vonden de meeste mensen die ons werk zagen het leuk, en waren ze enthousiast over het inzicht in de groei die het verschafte. Maar er waren ook mensen die zeiden 'hallo, je trekt er een soort S-je doorheen en je vindt het volgende punt, je trekt er dus alleen een lijntje door en dat is het.' Mensen zeiden ook: 'Is dit nou wetenschap?' Maar het ligt dus echt wel iets ingewikkelder. Deels is het waar wat ze zeggen, maar dat 'doortrekken' valt nog niet mee."

Hoe voerde je die discussie?

"We hebben diverse e-mailwisselingen gehad met mensen, maar we hebben op onze website ook een technisch document geplaatst, om onze werkwijze toe te lichten. En toen zeiden ook mensen die eerst wat negatief waren, dat ze zagen dat er meer achter zat dan ze verwacht hadden. En dat heeft de discussie wat getemperd. Maar je kunt er nog steeds een heleboel van vinden. En in het bijzonder ook van de data, hoe goed is de data? Dat blijft een zorg."

Ben je zelf tevreden met wat jullie met de data konden?

"We zijn er eigenlijk zo maar ingestapt, niet wetende waar we in waren gestapt, meer van: we moeten iets doen. Het is sowieso informatief gebleken voor een heleboel mensen. Er kwamen mensen dagelijks terug om te kijken hoe het ervoor stond. Wat ik eigenlijk het belangrijkste vind is dat onze voorspellingen ook zijn gebruikt als input voor ziekenhuisopnames. We hebben met een aantal ziekenhuizen in Nederland samengewerkt, waarbij we probeerden iets verder vooruit te voorspellen, zodat zij konden inschatten of ze moesten opschalen. En dat heeft in ieder geval bij een aantal ziekenhuizen geholpen. Daar heeft het ook echt betekenis gehad. Dat vind ik het belangrijkste, dat we iets in de praktijk hebben kunnen bijdragen."

Hoe zou je verder willen?

"We willen inderdaad verdergaan en zijn ook aan het zoeken naar financieringsmogelijkheden. Een van de dingen die we zouden willen doen, of beter zouden willen doen, is kijken of we misschien een soort dashboard kunnen maken waardoor we precies kunnen zien wat er gebeurt met de verspreiding in het land. Op welke parameters moet je goed letten en kun je gecombineerde parameters gebruiken om te zien of er iets verandert? We hebben intussen ook een analyse gedraaid om te kijken wat de invloed is geweest van beslissingen van de overheid om tot de lockdown over te gaan. Heeft dat invloed gehad op de verspreiding? Nou, dat heeft het. En de verschillen tussen landen zijn ook interessant. We willen de data ook analyseren om de invloed van maatregelen te kunnen duiden. Voor het moment willen we kijken of de verspreiding onder controle blijft als we de overheidsmaatregelen terugdraaien."

Wil je jullie werk ook in het onderwijs gebruiken?

"We zijn een soort capita selecta begonnen in deze onderwijsperiode, we hebben uiteraard geen colleges over dit onderwerp. Er zijn drie studenten bezig en die zijn nu naar literatuur aan het kijken, ze proberen wat zaken diepgaander te begrijpen, en ze onderzoeken ook andere modellen. Dat konden we onmogelijk allemaal zelf doen. En als dit succesvol is, dan moeten we er misschien wel een vak van maken. Ook wel aardig om te zeggen: we hadden in een vroeg

Het Verhulst-model

Genoemd naar de Belgische wiskundige Pierre François Verhulst. Het Verhulst-model of ook wel de logistische functie, is een model dat gebruikt wordt in de populatiegroeidynamica om het verloop van de omvang van een populatie $Y(t)$ in de tijd te beschrijven. In het geval van een epidemie waar een individu vatbaar of besmet kan zijn, beschrijft $Y(t)$ bijvoorbeeld het aantal mensen dat besmet is geraakt. In het Verhulst-model is de verandering van $Y(t)$ zowel evenredig met het aantal individuen dat besmet is, dus $Y(t)$, maar ook met het aantal individuen dat nog vatbaar is, dus $M - Y(t)$. De parameter M beschrijft het maximale aantal individuen dat besmet kan raken in de populatie. Formeel beschreven is het Verhulst-model een differentiaalvergelijking van de vorm

$$Y'(t) = kY(t)\left(1-\frac{Y(t)}{M}\right), \quad (1)$$

waar k de groeisnelheid representeert. Een oplossing van deze vergelijking is gegeven door

$$Y(t) = \frac{M}{1+Me^{-k(t-t_c)}},$$

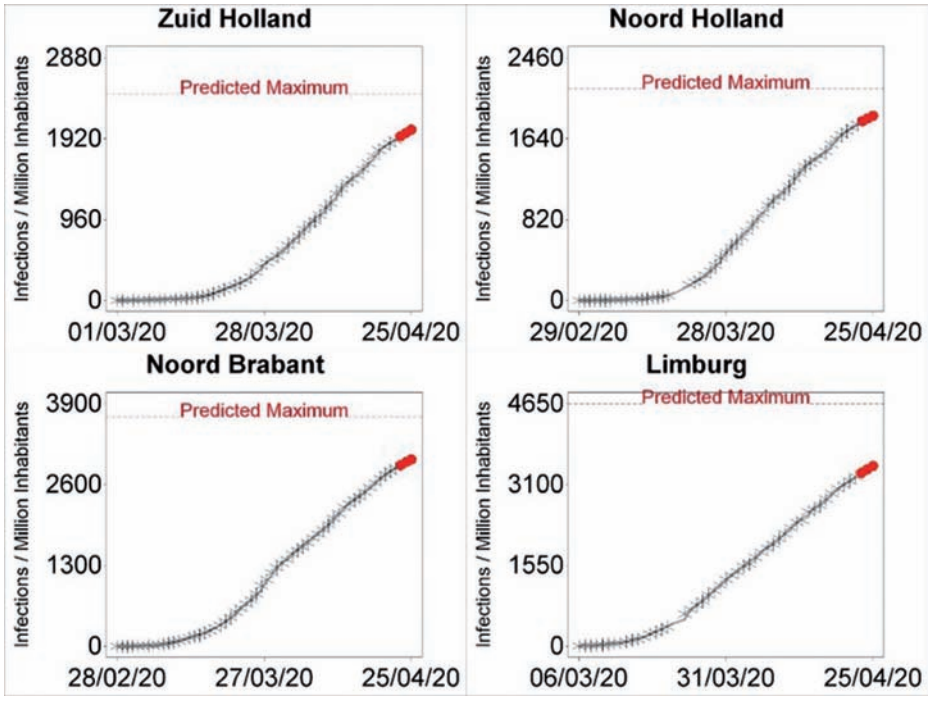
waar t_c het 'turning point' van de curve is. De twee factoren aan de rechterkant van (1) modelleren respectievelijk de fractie vatbare individuen en het aantal besmettelijke individuen, beide factoren moeten hoog zijn voor het virus om zich snel te verspreiden.

stadium een bachelorstudent van wiskunde die een bachelorproject wilde doen over COVID-19; die student hebben we gelijk in het team getrokken. Die voert nu ook simulaties uit en doet wat onderzoekjes."

Je leest veel in de krant en in het nieuws over het model, over het reproductiegetal, over hoe het mis kan gaan. Wat denk je dat het publiek, ook academici, hierover moet weten? Is het nuttig om te weten hoe een model werkt, hoe een model gebouwd wordt, of hoe je statistiek doet, of is dat te ingewikkeld?
"Wie ben ik om daar iets over te zeggen? Het is lastig om de juiste balans te vinden in informatievoorziening. Het RIVM bijvoorbeeld kijkt ook op subgroepniveau en gebruikt ook leeftijdsgebonden informatie, dat hebben wij niet, maar zou ik zeker ook willen gebruiken. Het is denk ik echter heel moeilijk aan het publiek uit te leggen wat voor complexiteiten er allemaal spelen. Het Verhulst-model is nog wel uit te leggen, maar zodra je ingewikkelder modellen en 'epidemic disease models' gebruikt, wordt uitleggen al een stuk lastiger.

Een college op TV zoals Robbert Dijkgraaf wel eens deed voor De Wereld Draait Door?
"Ik heb laatst een webinar gegeven bestemd voor artsen, om uit te leggen wat je met die getallen kunt. Nou denk ik dat het webinar iets te lastig was, omdat je de modellen toch echt wel nodig hebt als je de getallen wilt interpreteren. En een van de dingen die ik van belang vind bij de modellen is dat je je niet op één model moet fixeren, omdat de data niet wordt verzameld op een manier die goed past bij een gegeven model. Dat gaat dan toch iets meer de richting op van data science dan van zeg maar de fundamentelere wiskunde. Je kunt zeker het een en ander uitleggen hoor, in al die modellen zitten parameters die te begrijpen zijn. Als we van *susceptible* naar *infected* gaan, dat heeft te maken met het aantal contacten en de kans dat je besmet raakt tijdens zo'n contact, dat is prima uit te leggen. Maar zo'n heel model? Ik denk dat zo'n college à la Robbert Dijkgraaf een goed idee zou zijn."

Het reproductiegetal R is een redelijk simpel product van drie getallen, Is dat misschien te simpel?
"Ik maak me daar wel wat zorgen over. Het getal heeft een heldere interpretatie, maar de schatting van dat getal is toch niet zo triviaal, want dat heeft ook te maken met het model dat er onder zit en als je daar andere aannames bij maakt, verandert de schatting. En we hebben gezien in een aantal analyses dat sommige aannames behoorlijk invloed hebben op het schatten van bepaalde parameters in dat model, dus je moet goed weten welke aannames goed zijn en welke niet. Dat is iets waar het RIVM mee worstelt, en wij zelf ook. Nou heeft het RIVM wel iets meer informatie, dus dat kan nog helpen. Ik denk dat als je focust op één model of één getal, dat dat toch wel iets te beperkt is met deze data, dus dat je daar heel voorzichtig mee moet zijn."



Figuur 2 Verbeterd model toegepast op data in Nederland.

Het is nu een product van drie getallen die R. Is het denkbaar dat het een ingewikkelder expressie zou moeten zijn?
"Nee, dat denk ik niet. Ik denk wel dat je naar dit soort dingen moet blijven kijken, maar ik denk dat je moet aanvullen. Je moet het aanvullen met andere karakteristieken om te zien of je consistent bezig bent. Zo'n R kan zeggen dat het de goede kant uit gaat, maar andere signalen staven dat misschien niet. Dus het is goed naar verschillende signalen te kijken in die data: Ik zie dit gebeuren en dat lijkt de goede kant uit te gaan, ik zie daar een soort confirmatie, okay dat geeft me al wat meer vertrouwen. We hebben ook gekeken of we vergelijkbare patronen in andere landen zien. Wat we nu zien in Nederland, zien we dat ook in landen die net iets 'voor' zijn. Dan begin je vertrouwen te kweken. Dat is een beetje de aanpak die wij volgen."

In je team heb je statistici, met hun eigen specialisaties. Heb je behoefte aan een breder samengesteld team?
"Ja, we zouden zeker baat hebben gehad bij een breder samengesteld consortium: een epidemioloog bijvoorbeeld, die vanuit zijn expertise zou kunnen kijken, en de medische kant meeneemt. En wat ook nuttig zou zijn geweest, was iemand die handiger is in het verzamelen van de data van de diverse bronnen. Het zou zeker handig zijn geweest met een breder team, maar je doet het met de middelen die je hebt."

Laatste vraag: die gaat over het reproductiegetal. Ik lees wel dat je door die 1,5 m afstand het aantal contacten die iemand heeft beïnvloedt, en een ander artikel dat juist de kans op besmetting benadrukt. Er zijn verschillende meningen over. Is dat erg?
"In die modellen zitten de 'effective contact rates', dat is waar het over gaat. Het product van de kans dat je besmet raakt en het aantal contacten dat je hebt. Het gaat uiteindelijk om die effective contact rates, en of die naar beneden gaan. Voor het effect maakt het niet uit, welke factor daarvoor verantwoordelijk is, want het is toch het product waar je naar kijkt."