Casper Albers
Psychometrie & Statistiek
Rijksuniversiteit Groningen
c.j.albers@rug.nl

Column Casper grijpt een kans

Coronastatistiek

Sinds het begin van de coronacrisis, vliegen de statistieken en grafieken ons om de oren. Casper Albers bekijkt enkele epidemiologische modellen en benoemt de mogelijkheden én limitaties die statistiek in deze crisis kan bieden.

Als een vlinder met haar vleugels klapt in Brazilië, veroorzaakt zij een tornado in Texas. Deze zin — en vele variaties daarop — beschrijven het vlindereffect. Dit effect beschrijft hoe bij (deterministische) niet-lineaire systemen microscopisch kleine verschillen in de beginvoorwaarden kunnen leiden tot gigantische verschillen in een later stadium. Hetzelfde geldt, in grote lijnen, ook voor complexe stochastische systemen waar een klein verschil in parameterwaarden uiteindelijk tot gigantische verschillen kunnen leiden.

Dit is natuurlijk enorm relevant bij de huidige coronacrisis. Op 15 april legde de Duitse bondskanselier Merkel dit duidelijk uit: als de reproductiefactor R (het gemiddeld aantal personen dat een ziek iemand besmet) $1{,}1$ is, dan zou Duitsland pas in oktober de grenzen van de zorgcapaciteit bereiken. Bij een reproductiefactor die $1{,}2$ is (dus slechts $0{,}1$ hoger) is die grens al drie maanden eerder in zicht.

Die redenatie gaat uit van exponentiële groei: de eerste geïnfecteerde besmet R personen, die besmetten elk ook weer R personen zodat je met R^2 besmettingen zit en na t stappen heb je R^t besmettingen. In een blad voor wiskundigen hoef ik verder niet uit te leggen hoe dat explodeert; en ook hoe dat snel tot onrealistische voorspellingen leidt: bij een populatiegrootte N heb je na $\log(N)/\log(R)$ stappen meer voorspelde besmettingen dan dat er personen zijn. Voor een Nederlandse bevolking van een kleine 18 miljoen inwoners en $R \approx 2{,}5$, zit je daar binnen 16 stappen al. Zo'n exponentieel model is dus enkel geschikt om de beginfase van de pandemie te modelleren. Maar zelfs daar gaat exponentiële groei te snel. Een besmet iemand kan zijn/haar partner infecteren, maar deze kan vervolgens niet zijn/haar partner infecteren: die is al besmet. Omdat de meeste mensen doorgaans toch met betrekkelijk weinig anderen in aanraking komen, en voornamelijk steeds dezelfde — gezin, collega's, buren, enzovoort — vlakt de groei al snel af.

Voor die beginfase van de groei zijn er honderden verschillende wiskundige modellen. Deze modellen zijn ontwikkeld naar aanleiding van andere epidemieën, zoals ebola en hiv/aids, en de onderliggende ideeën zijn ook bruikbaar voor COVID-19. De manier hoe er in de modellen wiskunde wordt bedreven, de ene keer via deterministische differentiaalmodellen en de andere keer juist via stochastische modellen, laat goed zien hoe multidisciplinair dit vakgebied is. Voor wie meer wilt weten: Chowell en collega's [2] hebben recent een toegankelijk overzichtsartikel over epidemiologische groeimodellen geschreven. Een algemener achtergrondswerk, met ook de nodige informatie over de wiskundige modellen, is het boek van Anderson en May [1].

Gegeneraliseerd groeimodel

Veel modellen zijn een submodel of variant van het *generalized growth model* [4]:

$$\frac{\mathrm{d}C(t)}{\mathrm{d}t} = rC(t)^p.$$

Hier is $C(t)$ het cumulatief aantal besmettingen tot aan tijd t en de afgeleide $\mathrm{d}C(t)/\mathrm{d}t$ de incidentie; $r > 0$ beschrijft de groeisnelheid en $p \in [0,1]$ beschrijft de zogenaamde groeivertraging. Als $p = 0$ hebben we lineaire groei en als $p = 1$ het exponentiële groeimodel dat al in 1798 door Thomas Maltus beschreven werd. Als $0 < p < 1$ dan is er sprake van sub-exponentiële groei en kan het model herschreven worden tot een polynomiaal groeimodel met groeifactor $1/(1-p)$ [2,4].

Dat de groeifactor langzaam afremt is iets dat ook bij de huidige corona-ontwikkelingen te zien is — en ook al te zien was voor-

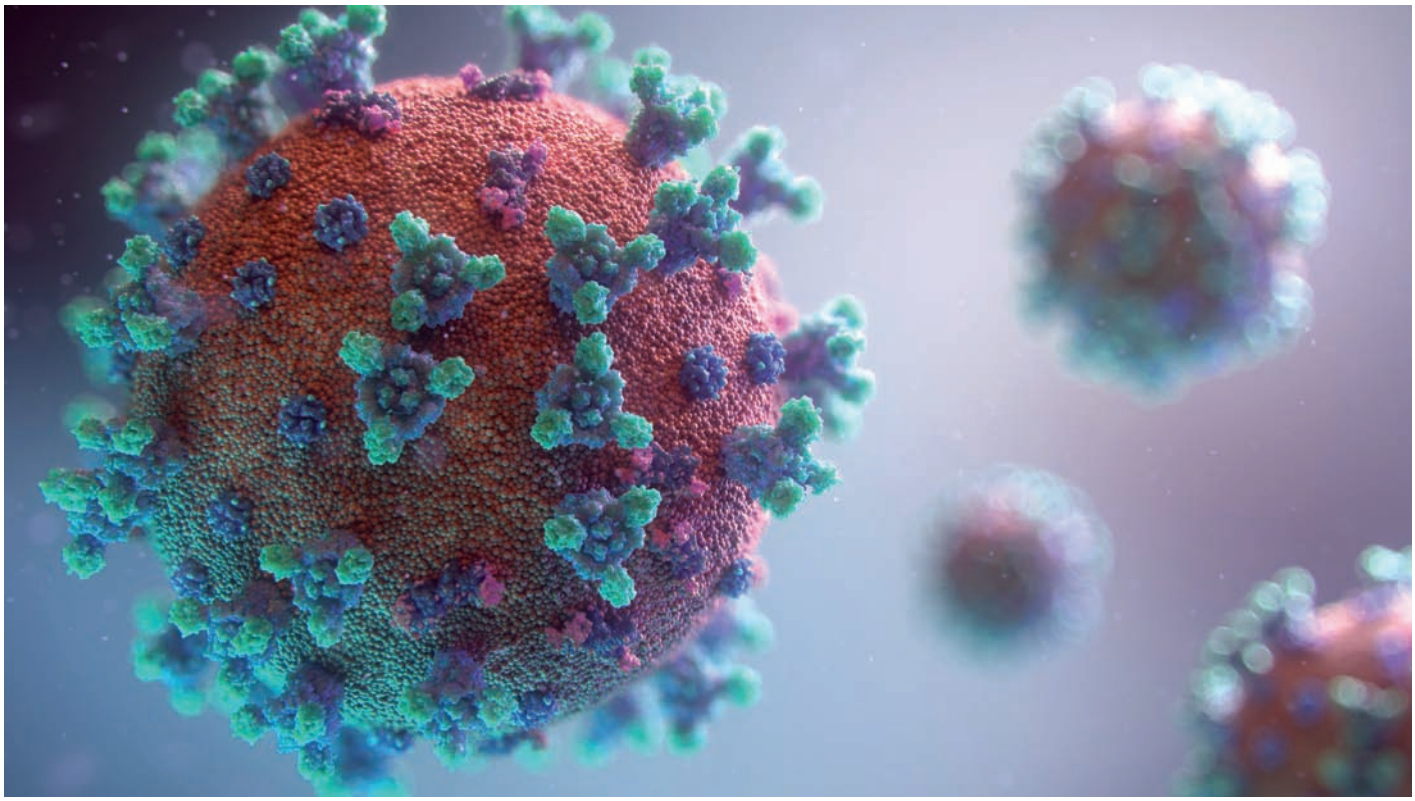

Foto: Fusion Medical Animation / Unsplash

dat de afname toe te schrijven kon zijn aan de lockdown(achtige) maatregelen die in verschillende landen genomen zijn: zelfs in de beginfase is de groei lager dan dat van een puur exponentieel groeimodel.

Het elegante van het gegeneraliseerde groeimodel is dat het een (relatief) eenvoudig en elegant model is. Die eenvoud is tegelijkertijd haar achilleshiel: als je de complexe werkelijkheid wilt beschrijven met een te eenvoudig model, sla je de plank al snel mis.

Het standaard SIR-model

Een stap ingewikkelder dan het gegeneraliseerde groeimodel is het SIR-model, wat de afkorting is van *S*usceptible ('vatbaar'), *I*nfected ('geïnfecteerd'), *R*ecovered ('hersteld'). In dit model, wordt de gehele populatie opgedeeld in deze drie categorieën (de 'recovered' categorie bevat zowel degenen die genezen zijn als de overledenen; beiden kunnen geen anderen meer besmetten), middels notatie $S(t)$, $I(t)$ en $R(t)$ voor het aantal personen in deze categorieën op tijdstip t. Het meest eenvoudige SIR-model neemt aan dat de totale populatiegrootte, $N = S(t) + I(t) + R(t)$, constant is en dat de overgangskansen van $S \to I$ en $I \to R$ dat ook zijn. Dan is het model te schrijven in drie differentiaalvergelijkingen:

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = -\beta \frac{S(t)I(t)}{N},$$
$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = \beta \frac{S(t)I(t)}{N} - \gamma I(t),$$
$$\frac{\mathrm{d}R(t)}{\mathrm{d}t} = \gamma I(t).$$

Met β wordt de overgang van vatbaar naar geïnfecteerd gemodelleerd; β is het product van het gemiddeld aantal contacten dat iemand op een bepaald tijdspunt heeft en de kans van overdracht van de ziekte bij contact. Met γ wordt de overgang van geïnfecteerd naar niet meer geïnfecteerd (= weer gezond + overleden) gemeten. Deze β en γ samen leiden tot het beroemde reproductiegetal R_0 dat de mediaberichten domineert:

$$R_0 = \beta/\gamma.$$

Als $R > 1$, dan is er sprake van een epidemie.

Ook het SIR-model is natuurlijk te eenvoudig. Zo is N niet constant in de tijd — er worden mensen geboren en er sterven mensen aan andere oorzaken; en zijn β en γ ook niet constant (zeker wanneer er lockdown-maatregelen zijn zal β dalen, en bij betere medische behandeling zal γ stijgen). Daarnaast sluit dit model uit dat herstelde personen nogmaals ziek worden ($R \to S$ is uitgesloten).

Wellicht het grootste gebrek van dit model is dat het aanneemt dat het ziekteverloop bij iedereen identiek is. R_0 wordt bij corona geschat op ongeveer $2{,}5$, wat betekent dat elke besmette persoon *gemiddeld* $2{,}5$ anderen besmet. Echter, helemaal niemand besmet $2{,}5$ anderen, simpelweg omdat aantallen integers moeten zijn. De ene zit het grootste deel van de week thuis en gaat alleen soms naar buiten voor een boodschap; terwijl de ander dagelijks een uur in de trein zit en vaak op feestjes komt (dit wordt contactheterogeniteit genoemd). Ook de incubatieduur is niet voor iedereen gelijk en een aantal andere factoren ook niet. Er is dus een model nodig dat met individuele verschillen om kan gaan. En dan biedt de statistiek uitstekende mogelijkheden.

Het stochastische SEIR-model

Een interessante uitbreiding van het SIR-model is het SEIR-model: de categorie *E*xposed ('blootgesteld') wordt toegevoegd. Er is bij veel infecties, waaronder corona, namelijk een incubatieperiode waarin men al wel is blootgesteld maar nog niet besmettelijk is.

Middels een extra differentiaalvergelijking kan je het $S \to E \to I \to R$-proces beschrijven en dan heb je een model dat zeer vergelijkbaar is met het SIR-model. Het kan echter ook via stochastische vergelijkingen, zoals Lekone en Finkenstädt [3] laten zien.

Het aantal vatbare personen op dag $t+1$ is het aantal vatbare personen op dag t, minus het aantal nieuwe besmettingen op dag t: $S(t+1) = S(t) - B(t)$. Het aantal blootgestelde personen stijgt dagelijks met die $B(t)$ en daalt met $C(t)$, het aantal besmette personen dat overgaat naar de geïnfecteerde categorie. Het aantal geïnfecteerden stijgt dagelijks met die $C(t)$ en daalt met $D(t)$, het aantal overgangen naar de restcategorie $R(t)$. Ook dit model neemt aan dat N constant is in de tijd.

De dagelijkse aantallen $B(t)$, $C(t)$ en $D(t)$ worden gemodelleerd via binomiale verdelingen:

$$\begin{aligned} B(t) &\sim \text{Bin}(S(t), P(t)), \\ C(t) &\sim \text{Bin}(E(t), p_C)), \\ D(t) &\sim \text{Bin}(I(t), p_R)). \end{aligned}$$

Hierbij is de kans $P(t)$ tijdsafhankelijk,

$$P(t) = 1 - \exp\left(-\beta(t)\frac{I(t)}{N}\right)$$

en de andere twee kansen niet,

$$p_C = 1 - \exp(-\rho), \qquad p_R = 1 - \exp(-\gamma).$$

Oftewel, $\beta(t)$ kan gezien worden als de tijdsafhankelijke overdrachtssnelheid, $1/\rho$ de *gemiddelde* incubatieduur en $1/\gamma$ de *gemiddelde* lengte van infectie. Deze kansen zijn een logisch gevolg van de aanname dat de duur van verblijf in een van de compartimenten exponentieel verdeeld is. Het fijne van exponentiële verdelingen is dat deze compleet worden vastgelegd door hun gemiddelde: hoewel je dus maar één parameter gebruikt, bied je toch de mogelijkheid om individuele verschillen te erkennen in je model.

Door $\beta(t)$ tijdsafhankelijk te maken, kan je het in het model verwerken als er maatregelen om overdracht tegen te gaan genomen worden. De beroemde R_0 is daardoor ook tijdsafhankelijk:

$$R_0(t) = \frac{\beta(t)}{\gamma}\frac{S(t)}{N} \approx \frac{\beta(t)}{\gamma}$$

(de stap met $\approx$ volgt uit dat $S(t) \approx N$ in de beginfase van elke epidemie). Middels simulatiestudies laten [3] zien dat dit model vrij accuraat de parameters van het model, inclusief R_0 kan schatten, zelfs wanneer er betrekkelijk weinig data beschikbaar is. Tevens laten de auteurs hun model los op een aantal ebola-datasets. Omdat statistische methodologie wordt gebruikt, worden alle schattingen automatisch vergezeld van onzekerheidsmarges. Dergelijke modellen kunnen dus zeer bruikbaar zijn om zowel het verloop van de ziekte en het (vermoede) effect van maatregelen te schatten, alsmede om aan te geven hoe nauwkeurig de schattingen zijn.

Statistische beschouwingen

Zoals George Box al zei: "All models are wrong (but some models are useful)." Het grote probleem dat coronamodellen onderscheidt van andere takken van statistisch modelleren is dat als er nu fouten gemaakt worden er direct doden vallen. Maar eerst afwachten totdat we een beter model en meer en betere data hebben is nóg dodelijker.

Vanzelfsprekend zijn er nog meer complicaties te bedenken die om een ingewikkelder model schreeuwen. Iemand uit Uden zal niet zo snel iemand uit Appingedam infecteren, simpelweg omdat er maar weinig mensen vanuit Uden naar Appingedam, en vice versa, gaan; en modellen die toegerust zijn met een spatiale component kunnen hiermee omgaan. Daarnaast: het is onmogelijk om voor iedereen te meten of zij in de S-, E-, I- of R-groep behoren (zouden we dit wel makkelijk kunnen meten, dan konden we korte metten met het virus maken). Zelfs het aantal overledenen door corona is niet exact, en zelfs niet met een marge van kleiner dan circa 25%, vast te stellen. Dit, gecombineerd met het feit dat we nu nog steeds slechts enkele maanden aan dagelijkse metingen hebben, zorgt er voor dat het gewoon onmogelijk is om bepaalde zaken nauwkeurig te schatten.

Hier komt bovenop dat we het model loslaten op de werkelijkheid en niet op een mooi opgezette *randomised controlled trial*. Mocht straks het aantal infecties stijgen, dan is het — op basis van statistiek — onmogelijk om te stellen dat dit komt door de heropening van de basisscholen of die van de kappers: beiden vonden ongeveer tegelijk plaats en het is niet mogelijk hun effecten los te meten.

Het is dus zaak om de statistiek en wiskunde in deze geen grotere rol te geven dan welke zij verdiend: die van hulpwetenschap. Daadwerkelijk zinnige interpretaties zijn niet puur *data driven* te trekken; je ontkomt er niet aan om de cijfers in de juiste virologische, epidemiologische en sociale context te plaatsen. De coronacrisis is één grote schreeuw om meer interdisciplinaire wetenschap.

Referenties

1 R.M. Anderson en R.M. May, *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1991.

2 G. Chowell, L. Sattenspiel, S. Bansal en C. Viboud, Mathematical models to characterize early epidemic growth: A review, *Physics of Life Reviews* 18 (2016), 66–97.

3 P.E. Lekone en B.F. Finkenstädt, Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study, *Biometrics* 62 (2006), 1170–1177.

4 C. Viboud, L. Simonsen en G. Chowell A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks, *Epidemics* 15 (2016), 27–37.