FITTING THE THEORY TO THE DATA:
HET VOORSPELLEN VAN OVERLAST

SELMAR SMIT, BOB VAN DER VECHT & LAYLA LEBESQUE

Theorie en praktijk lijken vaak ver uit elkaar te liggen. Toonaangevende theorieën zijn vaak beschrijvend, generiek en kwalitatief, waar de praktijk vraagt om specifieke, kwantitatieve uitspraken. Een voorbeeld hiervan vinden we in de sociale wetenschappen. Gedragstheorieën als de *rational choice*, *planned behaviour* en *environmental criminology* leveren algemene beschrijvingen over welke aspecten mogelijk het gedrag van een individu beïnvloeden. In praktijk blijkt dat dergelijke theorieën wel handvatten bieden, maar moeilijk gebruikt kunnen worden om gedrag van een individu, of zelfs een groep in kaart te brengen en te voorspellen. En juist dat laatste is in praktijk meestal het interessants. Jongerenwerkers zouden graag willen weten wie er de grootste kans loopt om op het slechte spoor te geraken. De gemeente en politie zouden graag willen weten wat zij kunnen doen om slecht gedrag te ontmoedigen. En menig bedrijf zou een grote pot geld over hebben om de adoptie van hun product te kunnen voorspellen. Dergelijke voorspellingen worden nu vooral gedaan op basis van datamining , statistiek en onderbuikgevoel en negeren op die manier de grote schat aan kennis die aanwezig is vanuit de sociale wetenschappen. Met de opkomst van krachtige computers kan dit gat tussen praktijk en theorie mogelijk gedicht worden. Onder de noemer *fitting the theory to the data* beschrijven we in dit artikel een specifiek voorbeeld waarin gedragstheorieën uit de *environmental criminology* worden omgevormd tot een voorspellend model van overlast voor de regio Haaglanden.

Praktijk: het voorspellen van overlast

Over criminaliteit en overlast bestaan veel theorieën (Lochner, 2004) maar niet elke theorie is even geschikt om omgevormd te worden tot een voorspellend model. Soms is het dat de theorie er niet geschikt voor is, maar het is ook mogelijk dat de empirische data niet voorhanden zijn, of dat het voorspellend model zelf niet van nut is. Zo gaan veel voorspellende modellen alleen uit van de sociale en economische factoren in een wijk. Maar omdat dergelijke factoren voor beleidsmakers niet makkelijk te beïnvloeden zijn, bieden ze weinig handvatten voor de ontwikkeling van beleidsinterventies. Wat ze wel kunnen doen is bepalen of ergens een buurthuis moet worden gebouwd, een uitgaansdistrict moet worden verplaatst of een park moet worden aangelegd. Dit zijn relevante beslissingen, want het effect van een dergelijke ingreep is zeer afhankelijk van de omgeving. Wat in de ene buurt tot overlast leidt, hoeft niet noodzakelijk hetzelfde effect te hebben in een andere buurt. Zo zijn er nauwelijks meldingen van problemen bij het café De Uylenburg aan de rand van Delft, terwijl 2.5 kilometer verderop bij de cafés in het centrum er een *hotspot* ligt van overlast.

Het ligt dus, logischerwijs, niet enkel aan het type gebouwen dat er staat, maar ook aan de omgeving waarin ze staan. Het bepalen van het effect van gebouwen op de hoeveelheid overlast in een buurt is dus meer dan enkel een simpele optelsom van de individuele effecten.

Theorie: precipitators en attractors

Op het gebied van omgevingsfactoren zijn er twee theorieën die verklaringen aandragen waarom op de ene locatie wel, en op de andere locatie geen overlast plaatst vindt. De eerste komt van Brantingam & Brantingam (1995) waarin zogenaamde *crime attractors* worden geïntroduceerd. *Attractors* zijn plaatsen die potentiele overlastveroorzakers aantrekken, maar niet noodzakelijk zelf overlast veroorzaken. Een voorbeeld hiervan is een bankje in het park. Hoewel deze op zichzelf geen overlast veroorzaakt, kan het wel overlastveroorzakers aantrekken. Wortley (2008) beschrijft juist een verklaring voor de hoeveelheid criminaliteit in een gebied. Hij introduceert *crime precipitators;* omgevingsfactoren die aanmoedigend werken op personen om overlast te veroorzaken. Een café en discotheek zijn logische voorbeelden van een *precipitator*.

Van theorie naar model

De theorieën van Brantingham & Brantingham en Wortley kunnen relatief eenvoudig worden omgezet naar een (wiskundig) model. Elk object in de omgeving is van een bepaald type, en van elk type wordt met behulp van vier verschillende parameters gedefinieerd wat de invloed is op de totale hoeveelheid overlast. De eerste twee parameters (a en b) bepalen de hoogte en uitstoot-afstand voor het precipitator gedeelte. De laatste twee (c en d) bepalen de mate van aantrekking en het bereik van de attractors.

Met de behulp van de formules uit figuur 1 is daarmee zowel de totaal aangetrokken hoeveelheid overlast te berekenen voor een bepaald object (A_j), als de hoeveelheid overlast die uiteindelijk terecht komt op een specifieke x,y locatie (R_{xy}). Hierbij gebruiken we de (journey to crime) distance decay function uit (Wilson, 1970) om de afstand tussen twee punten (D) om te zetten naar uitstoot.

Fitting the theory to the data

Hoewel het model nu een goede representatie is van de theorieën van Brantingham & Brantingham en Wortley, is het nog niet direct bruikbaar als voorspellend model. Daartoe gaan we het model kalibreren met empirische data van omgevingsobjecten en overlastcijfers uit de regio Haaglanden. De dataset van objecten halen we uit OpenStreetMap en bestaat uit 128 verschillende objecttypes. Daarom moeten de bijbehorende 512 parameters nog gedefinieerd worden om tot voorspellingen te kunnen komen; voor elke objecttype 4 parameters. Dit is wat wij *fitting the theory to the data* noemen; het kalibreren van een kwantitatief model (gebaseerd op bestaande theorieën) met parameterwaarden die passen bij de gegevens van een bepaald gebied. Gezien de complexiteit van het model, hebben we hierbij gekozen om gebruik te maken van de machine-learning techniek *backpropagation*. Backpropagation is een vorm van *supervised learning*, die in staat is om voor een (set van) geparameteriseerde formules de waardes af te leiden die zo goed mogelijk passen bij een database van trainingsgegevens (Mehryar Mohri, 2012). Met trainingsgegevens bedoelen we hier een combinatie van input en gewenste output, zoals de (x,y) coördinaten van een bepaald punt en de bijbehorende gemeten hoeveelheid overlast rond dezelfde locatie.

Het algoritme start met het willekeurig initialiseren van alle *a*, *b*, *c* en *d* waarden voor alle objecttypen. Gegeven de set met trainingsgegevens en alle objectlocaties, kan nu voor elke (x,y) coördinaat berekend worden wat dit (geheel willekeurige) model voor voorspelling doet qua hoeveelheid overlast (R_{xy}) en in hoeverre deze afwijkt van de gemeten waarde, de zogenaamde *fout*.

Voor elk van de coördinaten is tevens te bepalen wat hun afstand (D_{xyj}) is tot elk van de attractors en wat daarvan de attractionwaarde was (A_j). Hierdoor is het voor elke coördinaat en elk objecttype in de trainingsset mogelijk om te bepalen welke kant d_j op zou moeten bewegen (hoger of lager) om de fout voor deze coördinaat te verkleinen. Een andere mogelijkheid om de fout te verkleinen is door juist de waarden van A_j aan te passen. Logischerwijs kan dat enkel door a_i, b_i of c_i aan te passen. Wederom kun je voor elk van deze waarden vaststellen welke kant deze op zouden moeten bewegen om de fout van R_{xy} voor deze coördinaat te verkleinen. Als we daarna al deze richtingen optellen voor alle coördinaten in de trainingsset, hebben we voor elk van de 512 parameters een indicatie naar welke kant deze aangepast zou moeten worden om de totale fout te verkleinen.

De volgende stap in het algoritme is om al die 512 waarden een heel klein beetje aan te passen in de berekende richting. Nu er dus 512 nieuwe waarden zijn, die waarschijnlijk beter zijn dan de oorspronkelijke 512 kan hetzelfde proces herhaald worden. Opnieuw worden alle fouten berekend, opnieuw de richtingen bepaald, en opnieuw de waarden aangepast, totdat verdere verbetering niet mogelijk is. Een te grote aanpassing van de parameters leidt tot 'heen en weer schieten' (REF), een te kleine aanpassing zorgt voor langzame convergentie, en kan leiden tot het blijven hangen in een lokaal optimum.

Als de parameterwaarden niet meer veranderen, is het algoritme klaar, en is het model zo goed mogelijk gefit op de bestaande gegevens.

Resultaten

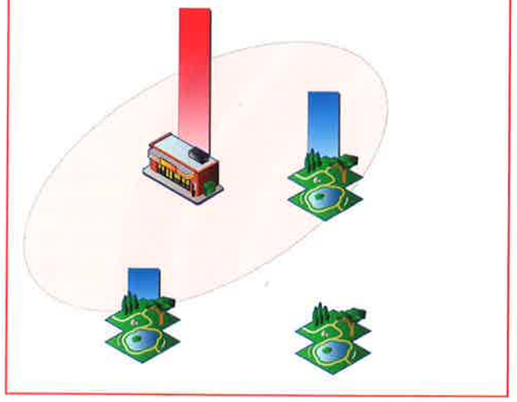
Interessante vraag is nu: 'Hoe goed representeert een dergelijk model de werkelijkheid?' of om de vraag anders te formuleren: 'Hoe goed is de theorie op de data gefit?'. Hierbij is het van belang te realiseren dat een model met een grote hoeveelheid vrijheidsgraden altijd bijna perfect gefit kan worden op een set gegevens. Het is dus van belang om niet te kijken naar de fit tussen trainingsdata en de bijbehorende voorspellingen (figuur 2) maar naar de voorspellingen voor een gebied dat niet is meegenomen in de trainingsdata. Specifiek voor dit doel is de stad Delft buiten de trainingsdata gehouden.

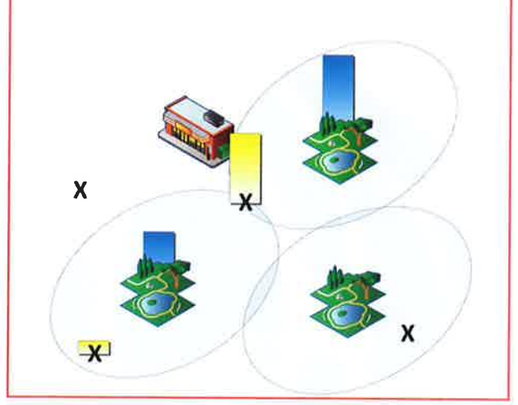
Figuur 2. De daadwerkelijke overlast in de trainingsdata (links) en de voorspelde waarde (rechts). Een bijna perfecte fit (een correlatie van 0,92).

Figuur 3. De daadwerkelijke overlast in Delft (links) en de voorspelde overlast (rechts)

$$A_j = \sum_{i \in Objecten} c_j . a_i \cdot e^{-\frac{1}{D_{ji} \cdot b_i}}$$

$$R_{xy} = \sum_{j \in Objecten} A_j \cdot e^{-D_{xyj} \cdot d_j}$$





Figuur 1. Het Precipitator & Attractor Model. De discotheek heeft een hoogte (rood) en een uitstootbereik (cirkel) van overlast. De overlast wordt aangetrokken (blauw) door de parken, afhankelijk van hun afstand tot de discotheek. Voor elke locatie (X) kan de overlast (geel) worden berekend op basis van de afstand tot de parken.

Als we de voorspellingen en daadwerkelijke cijfers van Delft naast elkaar leggen (figuur 3) blijkt dat de verhoudingen tussen de delen van de stad Delft redelijk goed zijn geschat (een correlatie van 0,79), maar de ordergrootte verkeerd is.

Dit kan veroorzaakt worden door een veel hogere concentratie van objecten in de stad Delft dan in de regio Haaglanden, of zelfs in Den Haag zelf. Logischerwijs zorgt dit direct ook voor hogere voorspellingen, aangezien zowel de afstanden als de hoeveelheid objecten heel anders is, dan in de trainingsset. Dit is mogelijk een gevolg van de *crowdsourcing* aanpak van OpenStreetMaps, welke de bron was van de objecten database, waarbij de detaillering van een gebied afhangt van de gebruikers en daarom niet uniform is.

Conclusies

Gezien de resultaten kunnen we concluderen dat de *fitting the theory to the data* aanpak succesvol is geweest. Het was niet alleen mogelijk om de bestaande theorieën uit de environmental criminology om te vormen tot een kwantitatief voorspelmodel op basis van data uit de regio Haaglanden, maar deze lijkt ook goed te generaliseren naar een ander gebied als Delft. Om het daadwerkelijk in praktijk te kunnen inzetten, zou het model nog verder verrijkt moeten worden met additionele informatiebronnen. Maar zelfs in de huidige vorm biedt het al handvatten aan de 'praktijk', zoals beleidsmakers. Naast deze praktische toepassing is het tevens niet ondenkbaar dat deze aanpak ook gebruikt kan worden door 'theoretici' om bestaande theorieën aan te scherpen, of uit te breiden door te kijken in hoeverre de data past op de theorie.

Literatuur

Brantingham, P., & Brantingham, P. (1995). Criminality of place. *European Journal on Criminal Policy and Research*, 3(3), 5-26.

Lochner, L. (2004). Education, Work, and Crime: A Human Capital Approach. *International Economic Review*, 45(3), 811-843.

Mehryar Mohri, A. R. (2012). Foundations of Machine Learning. Cambridge, MA: The MIT Press.

Wilson, A. G. (1970). *Entropy in Urban and Regional Planning*. Buckinghamshire: Leonard Hill Books.

Wortley, R. (2008). Situational Crime Precipitators. In R. Wortley, *Environmental Criminology and Crime Analysis* (pp. 48-69). Cullompton, UK: Willan Publishing.

Selmar Smit is aan de Vrije Universiteit gepromoveerd op het onderwerp machine learning, en sindsdien werkzaam als data scientist bij TNO.
E-mail: <selmar.smit@tno.nl>
Bob van der Vecht studeerde kunstmatige intelligentie aan de Rijksuniversiteit Groningen en is hierin in 2009 gepromoveerd aan de Universiteit Utrecht. Hij werkt sindsdien als onderzoeker bij TNO op het gebied van operations research.
E-mail: <bob.vandervecht@tno.nl>
Layla Lebesque heeft Econometrics and Operations Research gestudeerd aan de Universiteit Maastricht en is werkzaam bij TNO als technisch consultant op het gebied van data modeling & operations research.
E-mail: <layla.lebesque@tno.nl>