Barend Köbben

Data in overvloed

In de eerste aflevering van deze rubriek (*Geografie* september 2008) beloofden we dat we onze pijlen niet zouden richten op de zware overtreders, kaarten waarvan iedereen wel ziet dat ze lelijk, slecht gemaakt of overduidelijk fout zijn, maar dat we ons juist zouden bezighouden met de subtielere valkuilen waarin iedere kaartenmakende geograaf kan belanden.

Een valkuil die waarschijnlijk de meeste slachtoffers maakt, betreft de data. Nog niet zo lang geleden was het vaak lastig aan data te komen om kaarten op te baseren. In deze tijd van Wikipedia en Web 2.0 lijkt het wel of op elke digitale straathoek data in overvloed te krijgen zijn. Maar als er iets duidelijk geworden is uit de recente ophef over de IPCC klimaatrapporten, is het wel dat *veel* data nog niet hetzelfde is als de *goede* data. Welke data moet je kiezen en wat betekenen de data eigenlijk?

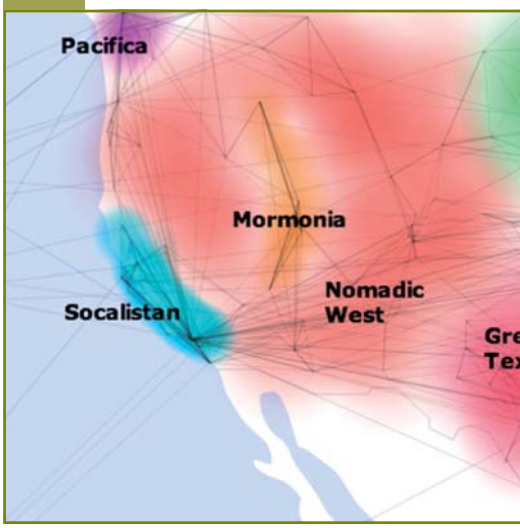
Juiste data

Het eerste dilemma is de keuze van de juiste data. In basiscursussen leer je dat de keuze voor een bepaald kaarttype afhangt van het doel van de kaart en de beoogde gebruikers. Bij verschillende kaarttypen passen verschillende datatypen, dus moet je zorgen dat de data passend zijn of passend gemaakt worden. Maar dit alles vormt nog een behoorlijk grove zeef, waardoor in veel gevallen nog een heleboel data in het mandje 'geschikt' vallen. Neem de grafiek op pagina 9 van deze *Geografie*, die de crisis van de euro moet verduidelijken. De grafiek geeft de ontwikkeling van de overheidsschuld en de begrotingstekorten in de periode 2002-2009 weer. De data komen grotendeels uit Eurostat-tabellen, en de factor 'begrotingstekort' wordt hierin omschreven als *general government deficit and surplus as percentage of GDP*. Maar de Eurostat-tabellen gaan niet verder dan 2008, dus voor 2009 hebben we gezocht naar vergelijkbare data en daarin voorzag een rapport van het EC Directoraat-Generaal Economische en Financiële Zaken. Maar hier kun je kiezen tussen maar liefst vijf tabellen die allemaal iets over begrotingstekorten lijken te zeggen, alle vijf in *% of GDP* zijn, maar waarvan de getallen behoorlijk uiteenlopen. Welke is vergelijkbaar met de Eurostat-data? De beschrijvingen helpen niet veel verder; ze variëren van *general government balance: net lending or borrowing of general government* voor de eerste tabel, tot *cyclically adjusted primary balance: corrected for the influence of the business cycle* voor de vijfde. En het maakt nogal uit welke je kiest: figuur 1a toont een fragment van de grafiek op pagina 9, gemaakt met de cijfers uit de eerste tabel. De vijfde tabel levert figuur 1b op, en maakt het verhaal over de europroblemen van de hier getoonde landen behoorlijk positiever! Wij hebben uiteindelijk de keuze gemaakt door van hetzelfde EC DG-rapport de versie uit 2008 te nemen en de cijfers daaruit te vergelijken met de Eurostat-tabellen. Daarbij bleek dat alleen de cijfers voor de eerste tabel vrijwel overeenkwamen en dus klaarblijkelijk hetzelfde verschijnsel weergaven...

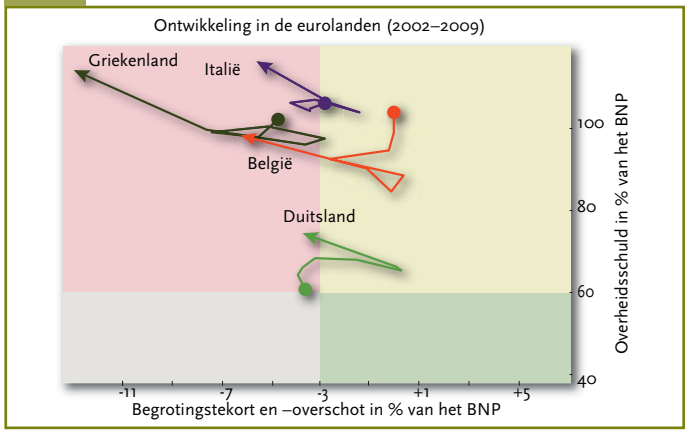
Heel veel data

Een heel ander keuzeprobleem is dat we soms de beschikking hebben over veel data. Heel veel data. Echt heeeel veel data... Neem de 210 miljoen publiek beschikbare *Facebook*-profielen waarnaar Pete Warden onderzoek doet en waarover hij regelmatig bericht op zijn blog. *Facebook* is een *social network* waarin mensen over de hele wereld vrijwillig hun

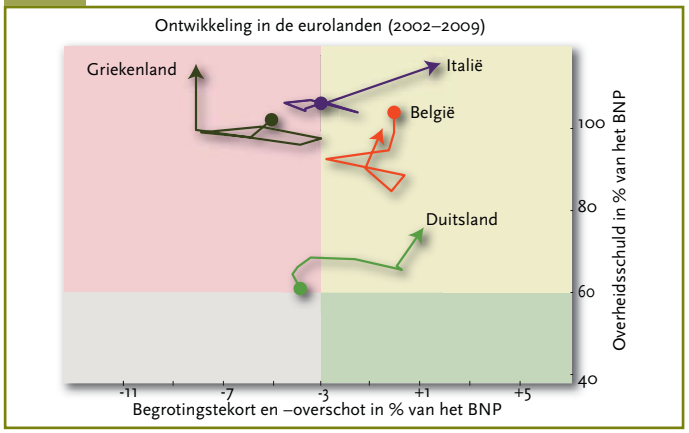
Figuur 2: *Facebook*-links

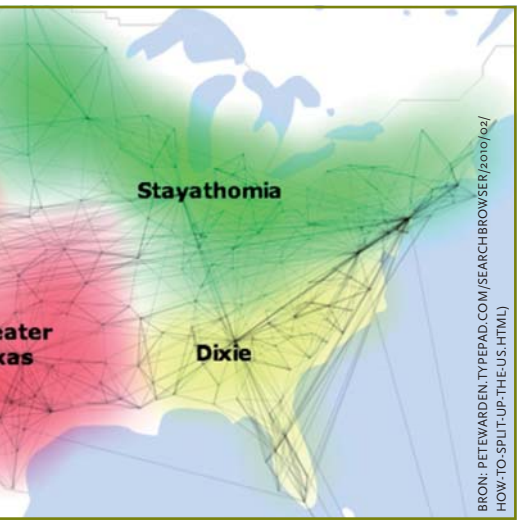


Figuur 1a: grafiek met de originele data



Figuur 1b: grafiek met de alternatieve data voor 2009





gegevens hebben gestopt over waar ze wonen, wat en wie ze leuk vinden en wie hun vrienden zijn. Omdat de gebruikers en al hun contacten in kaart te brengen zijn (via de in het profiel ingevoerde woonplaatsen), kun je uit deze enorme berg data geografische netwerken en patronen destilleren. Warden heeft een aardige webapplicatie gemaakt (petewarden.typepad.com/searchbrowser) waarin je een locatie kunt kiezen en dan direct ziet met welke andere locaties de meeste *Facebook*-links bestaan. Die informatie kun je dan weer combineren met de lijstjes van meest genoemde dingen in de *Facebook*-rubrieken als Likes en Friends. Op basis daarvan heeft Warren een interessante indeling van de VS gemaakt (figuur 2).

Een aantal van de onderscheiden gebieden zijn klassiek, zoals Dixie, het oude Zuiden, waar 'God' hoog scoort op de *fan pages*. Andere zijn verrassender, zoals Greater Texas, dat een groot deel van wat traditioneel als het mid-westen en het oude zuiden wordt beschouwd tot zijn invloedssfeer mag rekenen. En hier zijn de fan pages ook echt anders: 'God shows up, but always comes in below the Dallas Cowboys for Texas proper, and other local sports teams outside the state'.

Leuk toch, zulke geografische bevindingen uit zo'n berg data? Jazeker, maar hoe kun je nou weten wat zijn conclusies waard zijn? Nog minder dan bij de IPCC-rapporten kun je als gebruiker zelf controleren hoe en wat hij nou uit die miljoenen webpagina's heeft gehaald en wat die data echt vertellen. De auteur weet dat ook wel, hij schrijft ergens dat iemand over hem zei: 'Pete, you think by coding'. Hij is in de data geïnteresseerd vanwege de technologische uitdagingen die ze bieden om uit die databrij interessante of leuke informatie te halen *(data mining)*. Maar hij realiseert zich ook wel dat 'it summed up a lot of both my strengths and weaknesses'.

Foute keus

Eén voorbeeld van waar het fout kan gaan geeft hij zelf aan in zijn beschrijving van de gebieden. Bij Greater Texas noemt hij 'a few interesting name hotspots, like Alexandria, LA boasting Ahmed and Mohamed as #2 and #3 on their top 10 names'. Maar zoals bij veel van dit soort websites krijg je in *Facebook* bij het invullen van je woonplaats al tijdens het intypen van de eerste paar letters een rijtje suggesties, waarvan je bij het indrukken van de enter-toets automatisch de bovenste kiest. En als je als Egyptenaar uit Alexandrië even niet oplet, woon je dus zo in 'Alexandria, Louisiana, USA', en komt de toppositie van de naam Ahmed ineens in een ander licht te staan...

Maar het wordt bij dit soort data vooral lastig als je niet echt diepgaande kennis hebt van de data. Kijk bijvoorbeeld eens naar de kaart die je op deze site kunt maken van Nederland (figuur 3). Hij lijkt niet zo vreemd, met bijvoorbeeld veel relaties tussen Nederlandse Facebookers en die in de VS, Indonesië en het Verenigd Koninkrijk. Maar wat doet Starbucks in de lijst van 'Likes', terwijl die keten in heel Nederland maar een paar vestigingen heeft? Zijn de *Facebook*-gebruikers zo'n internationaal georiënteerd stel wereldreizigers? Wat veel buitenlanders, en wellicht ook Pete Warden, zich niet realiseren, is dat *Facebook* in Nederland in tegenstelling tot de rest van de wereld maar een weinig bezochte site is. Hier gebruikt de overgrote meerderheid het van oorsprong Nederlandse Hyves, en als je dus wilt weten wat er in de Nederlandse *social networks* speelt, moet je Hyves-profielen analyseren. Maar dat de Nederlandse *Facebook*-profielen als het ware een niet-representatieve steekproef zijn, is op geen enkele manier uit de data, en al helemaal niet uit de kaarten op te maken! •

Figuur 3: Geografische relaties in Nederlandse *Facebook*-profielen.