Big data en officiële statistieken: een gesprek met Marc Debusschere

Crompvoets, J; van der Voort, Haiko

DOI
10.5553/Bk/092733872016025001004

Publication date
2016

Document Version
Final published version

Published in
Bestuurskunde

Big data en officiële statistieken: een gesprek met Marc Debussschere*

Joep Crompvoets & Haiko van der Voort

Vlaamse en Nederlandse overheden en semi-overheden nemen steeds vaker initiatieven om met big data te werken. We spraken een van de pioniers. Marc Debusschere is coördinator administratieve en big data bij de Algemene Directie Statistiek – ook genoemd *Statistics Belgium*. Als zodanig is hij betrokken bij het project Exploitatie mobiele telefoondata voor officiële statistieken. Dit project is een samenwerking tussen *Statistics Belgium*, telecommunicatiebedrijf Proximus en Eurostat. Het project omvat een eerste exploratie van de mogelijkheid om mobiele-telefoondata te gebruiken voor officiële statistieken. De mogelijkheden moesten geconcretiseerd en geoperationaliseerd worden om zo snel mogelijk tot exploitatie te komen.

Variëteit en omvang

Dit bleek geen sinecure. Ten eerste is er een grote variëteit aan datatypes die mogelijk in aanmerking komen: telecomoperatoren beschikken over meer dan twintig verschillende soorten gegevens, zoals CDR (*Call Detail Records*), elektromagnetische signaalgegevens, facturatiegegevens, *roaming*-gegevens van buitenlandse operatoren, et cetera. In eerste instantie werd geopteerd om te werken met de volstrekt anonieme elektromagnetische signaalgegevens. Ten tweede zijn de databestanden zo omvangrijk dat ze pas na selectie en verwerking hanteerbaar worden voor Statistics Belgium. Over de selectie- en verwerkingsmethoden moeten afspraken gemaakt worden tussen de partijen. Ten derde moeten de telecomdata op een specifieke manier gegeocodeerd worden, zodat ze zijn te vergelijken en valideren met gegevens van Statistics Belgium.

Gefaseerd

Omwille van deze technische uitdagingen werd het project gefaseerd en cumulatief georganiseerd. Masten pikken de signalen op binnen hun 'cel' (11.000 voor België). Eerst zal het aantal aanwezige actieve mobiele telefoons per cel geregistreerd worden (de *present population*) om dit te koppelen aan gegeocodeerde datasets van Statistics Belgium. Via te ontwikkelen en te testen algoritmen zal dan gepoogd worden de *resident population* in een gebied te schatten. Daarna wordt de scope uitgebreid naar de *usual environment*, inclusief bijvoorbeeld woon-werkmobiliteit. En uiteindelijk wordt alles beschouwd, dus ook bewegingen buiten de

* Prof. dr. ir. J. Crompvoets is senior researcher/consultant/lecturer aan de KU Leuven. Dr. H.G. van der Voort is universitair docent aan de Technische Universiteit Delft.

usual environment, ofwel: toeristische reizen voor werk of ontspanning. Deze telecomdata zullen dus hopelijk relevant blijken voor demografische, mobiliteits-, sociale en toerismestatistieken!

Motieven

De drie partijen hebben elk hun eigen motief om dit project aan te vatten. Statistics Belgium wil graag nieuwe databronnen exploreren en exploiteren, vanuit drie motieven. Een nieuwe manier van werken met behulp van big data belooft kostenefficiëntie en minder administratieve lasten voor burgers en ondernemingen, wat ook een politieke prioriteit is. Er is ook een statistisch-inhoudelijk motief: door technologische en maatschappelijke evoluties laat iedereen tegenwoordig een groeiende 'elektronische voetafdruk' na (bijvoorbeeld via de mobiele telefoon, elektronisch betalen, surfgedrag, sociale media, camera's, *internet of things*, et cetera). Dit is een nieuwe datastroom die geëxploiteerd kan worden. Bovendien is er de belofte van snellere verspreiding van relevante informatie (bijna *real-time*) en volledigere datasets van de gehele populatie. Big data zullen echter nooit het volledige plaatje kunnen geven, en altijd aangevuld moeten worden met traditionele enquêtes en administratieve data. Debusschere veronderstelt dat Proximus geïnteresseerd is in het vinden van commerciële toepassingen voor hun big data en daarom in projecten stapt die hun knowhow vergroten en partners opleveren voor gezamenlijke ontwikkeling, ook in een Europese en internationale context via Eurostat. Eurostat ten slotte wil het gebruik van big data voor Europese statistieken stimuleren via het faciliteren van concrete samenwerking in pilots, zoals deze.

'Geef ons data'

Cruciaal voor een pilotproject als dit is het juiste businessmodel. In dit geval is het model gebaseerd op onderlinge afstemming en het zoeken naar win-winsituaties. 'Geef ons data voor officiële statistieken' werkt niet. Het is onmogelijk om data-eigenaars te dwingen, ook al omdat de initiatiefnemer de data niet kent en de infrastructuur niet heeft. Een wetgevend kader ontbreekt voorlopig, al zal dat er ooit wel komen. Statistics Belgium speelt dan ook de rol van makelaar en datakoppelaar. Uiteraard zijn zij zeer geïnteresseerd in het ontvangen van data, maar daarnaast zijn zij ook dataleverancier en leverancier van methodologische en statistische expertise.

Schaarste

Bestuurlijke succescriteria zijn dan ook de vertrouwensbasis tussen partijen op basis van een wederzijds belang, en het sporen van de eigen doelen met die van de data-eigenaar. Maar er zijn ook technologische succescriteria. Hoe om te gaan met schaarse infrastructuur die met grote datasets om kan gaan? Welke data moeten

Bestuurskunde 2016 (25) 1
doi: 10.5553/Bk/092733872016025001004

bewaard en verwerkt worden (en welke niet)? Hiervoor is er schaarste aan competent personeel, vooral *data scientists*. In deze context zijn ook de universiteiten als partners potentieel belangrijk. De schaarste aan competent personeel kan een project als deze in de kiem smoren. Op termijn zal er dus moeten geïnvesteerd worden in personeel met specifieke skills. Daartoe is een nieuwe prioritering nodig, waarvoor politiek draagvlak moet worden gezocht. Ten slotte is er tolerantie voor *trial and error* nodig. De stap van exploratie naar exploitatie moet namelijk zo snel mogelijk worden genomen. De enige manier om kennis te maken met de mogelijkheden van big data is deze te ervaren.

Pril

Het project is nog pril. De respondent heeft de avond voor het interview de eerste databeschrijving ontvangen (10-2-2016)! Een eerste rapport wordt midden april verwacht, met tentatieve resultaten voor hopelijk de *present population* en de *resident population*. Big data is dan ook grotendeels nog onbekend terrein. Daarom moeten verwachtingen niet te hoog gespannen zijn. Er is geen zekerheid of garantie dat de uiteindelijke methoden werken en dat er resultaten zoals reguliere officiële statistieken geleverd kunnen worden. Daarenboven zijn privacy en databescherming van het allergrootste belang. In dit stadium is er geen gevaar voor *big brother*-toestanden, omdat enkel anonieme statistische aggregaten worden gebruikt. In een (veel) later stadium – als persoonlijke gegevens gebruikt zouden worden – moeten de nodige *safeguards* worden ingebouwd, onder toezicht van de Belgische Privacycommissie.