Statistics Belgium

ANALYSE

Big data en statistiek

om het kwartier een volkstelling …

Big data en statistiek: om het kwartier een volkstelling …

Marc Debusschere[1], Patrick Lusyne[1], Pieter Dewitte[1], Youri Baeyens[1], Freddy De Meersman[2], Gerdy Seynaeve[2], Albrecht Wirthmann[3], Christophe Demunter[3], Fernando Reis[3], Hannes I. Reuter[3]

1 Statistics Belgium (AD Statistiek), Brussel, België, marc.debusschere@economie.fgov.be

2 Proximus, Brussel, België; freddy.demeersman@proximus.com

3 Eurostat, Luxemburg, Luxemburg; albrecht.wirthmann@ec.europa.eu

Samenvatting

In een gezamenlijk project onderzoeken Statistics Belgium, Proximus en Eurostat de mogelijke bruikbaarheid van mobiele-telefoondata (afkomstig van Proximus) voor officiële statistieken. Een eerste studie legt de grondslag door na te gaan of de werkelijk aanwezige bevolking geldig en accuraat gemeten kan worden, vergeleken met de Census 2011.

Dit artikel is eerder verschenen in Trefpunt Economie nr. 8 van de Fod Economie.

1. Inleiding: de derde datarevolutie

Officiële statistieken waren sinds hun begin in de vroege negentiende eeuw gebaseerd op enquêtes bij burgers en ondernemingen. Om de kosten te beperken en de belasting voor wie moet antwoorden te beperken, werd de laatste twintig jaar in toenemende mate een beroep gedaan op administratieve bestanden; daarbij vormde in België de Census 2011 een belangrijk keerpunt: in tegenstelling tot de vorige volkstelling, van 2001, toen nog miljoenen formulieren ingevuld en verwerkt moesten worden, werd de Census 2011 volledig vanuit administratieve bestanden opgesteld.

Maar nu is al de tijd aangebroken voor de derde datarevolutie in de statistiek: big data! Iedereen laat continu elektronische sporen na, via sensoren, camera's, elektronische betalingen of afhalingen, online activiteit allerhande, … Het is in principe mogelijk om deze immense en grotendeels ongestructureerde vloedgolf van 'big data' te gebruiken om de meeste bestaande statistieken sneller en beter te produceren en zelfs om fenomenen te beschrijven die tot nu toe volledig buiten beeld bleven.

Binnen het geheel van big data vormen mobiele-telefoondata een bijzonder veelbelovende potentiële bron, voor tal van statistische domeinen: bevolking, migratie, arbeidsmigratie en -mobiliteit, transport, bewegingen over de grenzen, toerisme, enz. Recente pilootstudies in diverse landen hebben aangetoond dat ze een haalbaar alternatief vormen voor de meer traditionele gegevensbronnen die de nationale statistische instituten gebruiken (Altin ea, 2015, Deville ea, 2014; European Commission, 2014). Maar vooraleer mobiele-telefoondata in de reguliere statistiekproductie geïntegreerd kunnen worden, moet hun kwaliteit zorgvuldig geëvalueerd worden. Dit artikel focust op de geldigheid en nauwkeurigheid van mobiele-telefoondata als een meting van de bevolkingsdichtheid in België, in vergelijking met de resultaten van de Belgische Census 2011, door Statistics Belgium opgesteld op basis van het Belgische bevolkingsregister (Rijksregister). De mobiele-telefoondata zijn afkomstig van Proximus, de grootste[1] mobiel-netwerkoperator in België.

[1] 40.3% marktaandeel in 2012 (http://economie.fgov.be/nl/consument/Internet/telecommunicatie/teledistributie/).

Drie onderzoeksvragen worden behandeld:

(1) vormen mobiele-telefoondata een geldige gegevensbron om de bevolkingsdichtheid te ramen? (geldigheid);

(2) wat is het verband tussen bevolkingsdichtheid gebaseerd op mobiele-telefoondata versus Censusgegevens? (nauwkeurigheid);

(3) hoe kan de waarde van mobiele-telefoondata voor deze toepassingen verder verhoogd worden? (data-integratie en reproduceerbaarheid)

Zowel de mobiele-telefoondata als de Censusgegevens zijn een benadering van de werkelijkheid; met welbekende beperkingen:

- de Censusgegevens tonen de geregistreerde bevolking op basis van de woonplaats vermeld in het Rijksregister, die niet noodzakelijk de werkelijke woonplaats is;
- de mobiele-telefoondata, anderzijds, benaderen de werkelijk aanwezige bevolking via de telefoons aanwezig in een gebied, wat gedurende de nacht een sterke aanwijzing biedt voor de werkelijke woonplaats, maar met vertekeningen als gevolg van onvolledige dekking (meer dan één apparaat per persoon of geen, wisselende marktaandelen als niet alle operatoren data leveren, atypische werk- of woonsituaties, …).

De kwaliteit van beide databronnen kan verbeterd worden door verdere analyse, meer waarnemingen en het gebruik van bijkomende informatiebronnen. In het geval van mobiele-telefoondata, bijvoorbeeld, maakt het observeren van individuele apparaten over een langere periode het mogelijk om een ‘meest waarschijnlijke woonplaats’ toe te kennen en zo tot een betere raming te komen van de werkelijk aanwezige bevolking.

De mate waarin beide bronnen convergeren vormt een ondergrens voor hun geldigheid en nauwkeurigheid. We postuleren dat een hoge correlatie tussen Censusgegevens en tellingen van mobiele telefoons tijdens de nacht aantoont dat beide een geldige en nauwkeurige meting bieden van de werkelijke bevolking.

2. De gegevens

2.1 Mobiele-telefoondata

De meeste studies tot nu toe (European Commission, 2014) gebruikten CDR's ('call detail records', registraties van de oproepdetails), nodig om te kunnen factureren; ze tonen de locatie en het tijdstip telkens als een mobiele telefoon gebruikt wordt. Tegenwoordig capteren netwerkpeilsystemen echter elke signaalactiviteit, met inbegrip van niet-factureerbare transacties, waardoor ze een veel fijner tijdruimtelijk detail bieden. In het Proximus-netwerk is daardoor het aantal bruikbare signaalgebeurtenissen ongeveer tienmaal groter dan het aantal CDR's. Voor elk apparaat op het netwerk wordt minstens om de 3 uren een positie bepaald; bij een actieve dataconnectie vermindert dit interval tot ongeveer eenmaal per uur. In de praktijk worden transacties zelfs nog frequenter geregistreerd, vooral voor smartphones die vaak met het netwerk connecteren zonder dat de eigenaar zich daarvan bewust is. Verder komen met nieuwere technologieën zoals 4G meer lokalisaties beschikbaar; en de intervallen verminderen nog meer wanneer apparaten locatie-updates uitvoeren telkens ze van een locatiegebied[2] naar een ander overgaan.

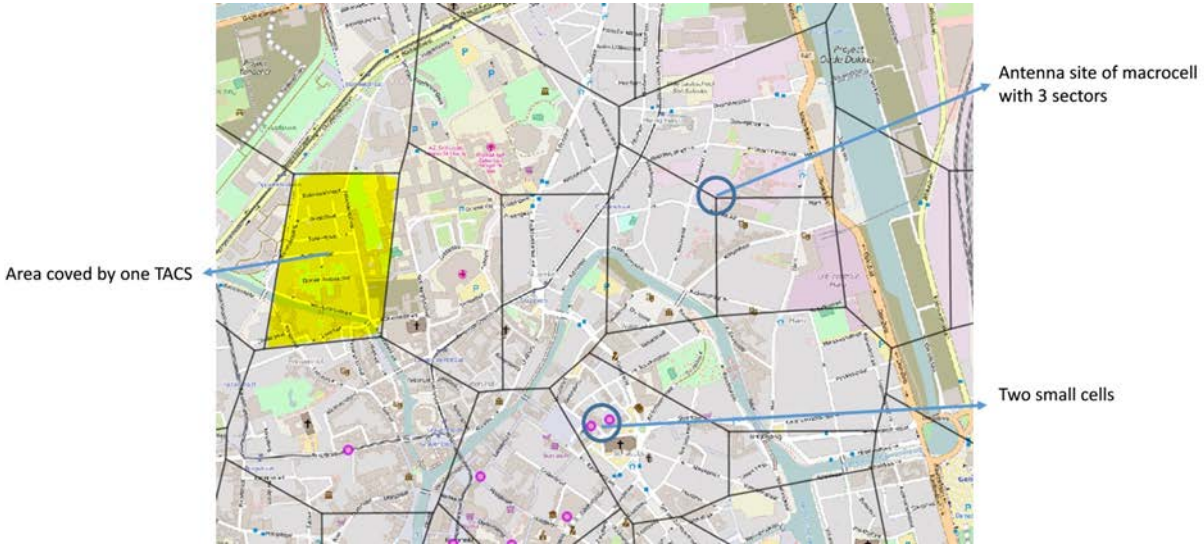
Bij elke transactie van een mobiele telefoon op een mobiel netwerk is de locatie van die telefoon gekend tot op het niveau van de celidentiteit. Een mobiele-telefoonnetwerk is een cellulair systeem dat geleidelijk complexer geworden is: antenne-opstelpunten omvatten tegenwoordig typisch meervoudige technologieën (2G, 3G, 4G) en meervoudige cellen. Met het oog op deze studie werd een construct genaamd TACS (Technology-Agnostic Cell Sector) ontwikkeld: het gebied bediend door alle cellen met dezelfde azimuth (richting van de hoofdbundel van de antenne) en ongeacht de gebruikte technologie, dat bestaat uit alle locaties die zich dichter bij de cel bevinden dan bij de omringende (ook elk een TACS). De resulterende veelhoeken worden voorgesteld als een Voronoi-diagram[3], wat het mogelijk maakt om een vereenvoudigd model van het mobiel netwerk te bouwen. Door op deze manier

[2] Een locatiegebied is een logische groepering van cellen (een cel is het gebied bestreken door een antenne); een mobiele telefoon in rusttoestand seint niet naar het netwerk als hij verandert van de ene cel naar de andere, tenzij deze laatste zich in een nieuw locatiegebied bevindt.

[3] De opdeling van een plat vlak in gebieden (Voronoi-veelhoeken) op basis van de afstand tot centrale punten, in dit geval antennes; elk punt in een veelhoek bevindt zich dichter bij diens antenne dan bij andere antennes.

abstractie te maken van de complexiteit van de verschillende technologielagen en van de ingewikkelde reële relatie tussen grote en kleine cellen[4], worden snelle en performante berekeningen mogelijk gemaakt.

Fig.1: Macro-TACS met kleine cellen ingedeeld bij hun overkoepelende TACS.



Vervolgens werd een heatmap aangemaakt, als een geheel van veelhoeken die elk een TACS voorstellen; die kaart toont het aantal aanwezige apparaten in elke TACS op basis van lokalisaties door het mobiel netwerk. Tijdsdiscontinuïteiten werden opgelost via interpolatie (een apparaat wordt verondersteld zich op zijn laatste positie te bevinden tot een nieuwe gekend is) en bijkomende filters werden toegepast (bv. verwijderen van 'machine naar machine'). Gegevens werden om de 15 minuten geregistreerd voor één weekdag (donderdag 8 oktober 2015) en één zondag (11 oktober 2015).

[4] Microcellen (die een klein gebied afdekken, bv. een deel van een straat), picocellen of femtocellen (typisch voor dekking binnenshuis) worden ingedeeld bij hun overkoepelende macro-TACS (dakcellen).

Om privacyproblemen te voorkomen, zijn alle in dit stadium gebruikte gegevens aggregaten (tellingen van apparaten per TACS), die geen individueel toewijsbare informatie bevatten.

2.2 Censusdata

De gegevens van de Census[5] registreren de Belgische bevolking op 1 januari 2011. De variabele 'woonplaats' refereert naar de geregistreerde verblijfplaats zoals opgegeven in het bevolkingsregister en wordt gebruikt als proxy voor de plaats waar mensen gewoonlijk verblijven gedurende de nacht, in de vroege morgen en de avond.

Deze gegevens werden geaggregeerd zowel voor roostervierkanten van 1 km² als voor TACS. Om km²-roostergegevens en TACS-gegevens aan te maken voor de Census, werden de adressen in het bevolkingsregister gegeocodeerd op basis van de overeenstemming tussen het bevolkingsregister en kadastergegevens, die de basis vormen voor het register van woningen en gebouwen.

3. Methoden

Om de mobiele-telefoondata en de Census-gegevens te kunnen vergelijken, moeten ze eerst voorgesteld worden in een gemeenschappelijke geografische ruimte, hetzij het 1km x 1km Standaard Europees Rooster of TACS (Voronoi-diagrammen die het mobiele-telefoonnetwerk repliceren).

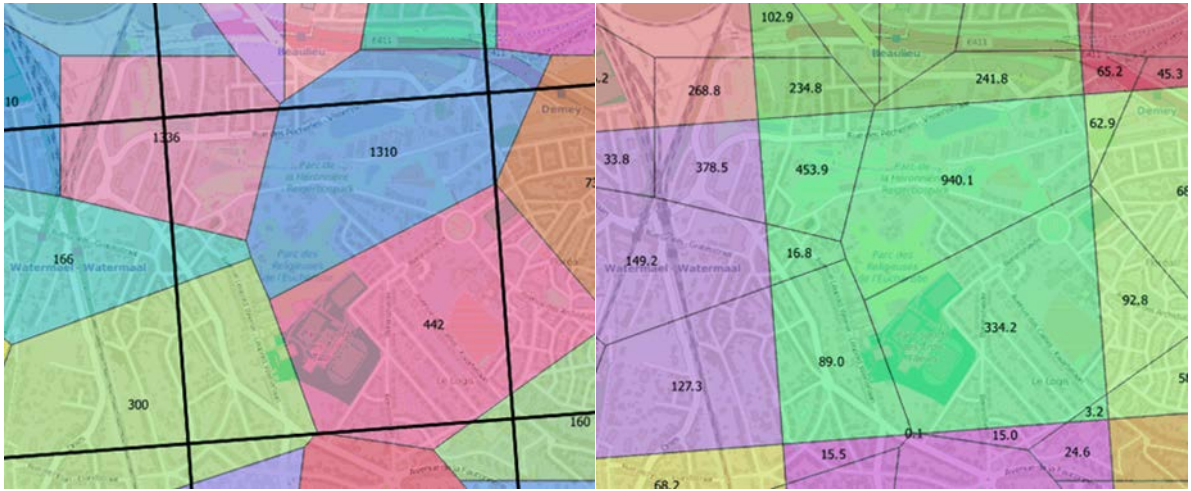
3.1 Mobiele-telefoondata omzetten naar het 1 km² Standaard Europees Rooster

Terwijl mobiele-telefoondata georganiseerd zijn in TACS die in omvang variëren (van vrij klein in steden tot ettelijke vierkante kilometers in dunner bevolkte gebieden), gebruikt de Census roostervierkanten van 1 km² (zie Fig. 2 met een voorbeeld in Brussel; de getallen in de veelhoeken zijn tellingen van mobiele telefoons).

[5] De Census 2011 werd volledig opgesteld vanuit administratieve bronnen (bv. registers van bevolking, sociale zekerheid, belastingen, woningen, ondernemingen, onderwijsniveau).

Het aantal getelde apparaten in elke veelhoek (TACS-tellingen) wordt proportioneel per oppervlakte opgesplitst en de resulterende subtotalen worden dan toegewezen aan elk van de roostervierkanten van 1 km² waartoe ze behoren; Die kunnen vervolgens opgeteld worden voor elk km²-roostervierkant.

Fig. 2: Tellingen per TACS omgezet via proportionele toewijzing in totalen per roostervierkant van 1 km² voor een klein gebied in Brussel.



Deze methode werkt erg goed voor gebieden waar TACS relatief klein zijn, maar ze ondervindt problemen wanneer een uitgebreide TACS grote gebieden omvat waarin weinig of geen telefoons tijdens de nacht aanwezig zijn (bv. in de Ardense wouden). Vermits ook in deze gevallen de bevolking evenredig over een TACS verdeeld wordt, is een correcte raming van de bevolkingsdichtheid per km² uiteraard niet mogelijk. Dit probleem kan misschien ondervangen worden door bijkomende databestanden (bv. over bodemgebruik).

3.2 Omzetten van bevolkingsdichtheid op basis van Census-gegevens naar TACS

Dit is het omgekeerde van de vorige benadering. Maar in plaats van de bevolking van elk vierkant van 1 km² proportioneel toe te wijzen aan de TACS of delen van TACS die het omvat, analoog aan de hierboven beschreven methode, worden censusobservatiepunten (d.w.z. adressen) direct en daardoor ook nauwkeuriger ondergebracht in een bepaalde TACS.

3.3 Statistische analyse van het mobiele-telefoondatabestand: clusteranalyse

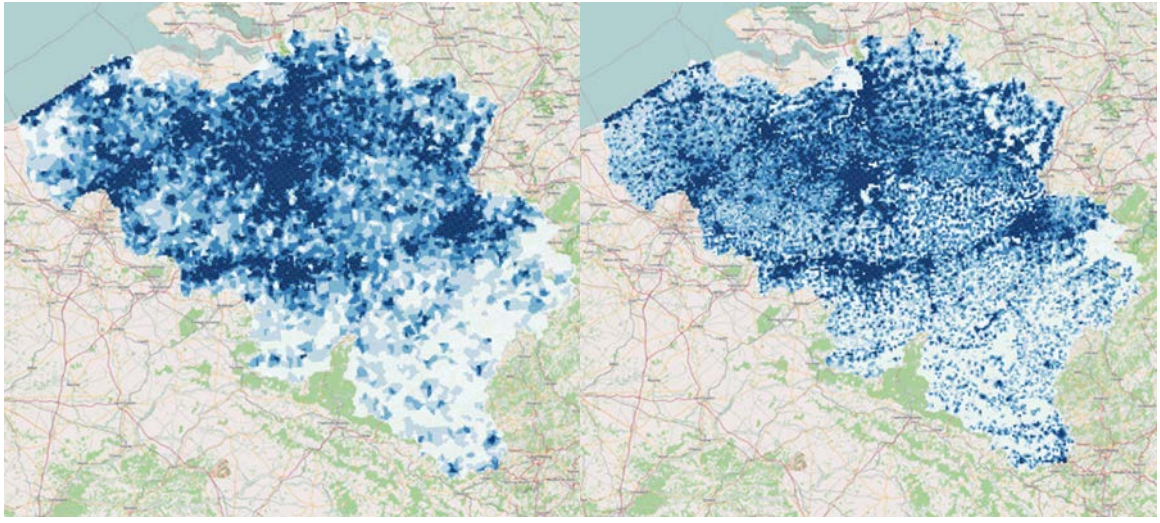
Een clusteranalyse werd uitgevoerd op de bestanden van de mobiele-telefoondata, zowel voor de twee dagen apart als voor beide gezamenlijk, gesommeerd per TACs en per roostervierkant van 1 km², voor een interval van 15 minuten en van één uur. De absolute cijfers van de tellingen van de apparaten werden genormaliseerd naar een gemiddelde van 0 en een standaardafwijking van 1 (procedure “scale” in R, versie 3.2.3, basispakket) vóór KMEANS-clustering (stats package, algoritme van Hartigan en Wong met willekeurige middelpunten). Optimale clusteraantallen werden bepaald via de som van kwadraten binnen groepen (SSW in R), uiteindelijk resulterend in drie clusters voor de dataset van de donderdag en vier voor die van de zondag. Om hun geloofwaardigheid te testen, werden de resultaten grafisch voorgesteld en gecombineerd met topografische kaarten om na te gaan of de drie patronen voor de werkdag overeenstemmen met bepaalde topografische kenmerken. Correlaties tussen tellingen van mobiele telefoons en aanwezige bevolking op basis van de Census werden berekend via R voor elk van de clusters, en hun patroon over de uren heen werd geanalyseerd.

4. Resultaten

4.1 Raming van de bevolkingsdichtheid: mobiele-telefoondata versus Census 2011.

De twee kaarten hieronder tonen de bevolkingsdichtheid per km²; de linkse is gebaseerd op TACS-tellingen van mobiele apparaten op donderdag 8 oktober 2015 om 04:00, omgezet naar 1 km² roostervierkanten, de rechtse op de 2011 Census afgeleid van administratieve records in het Rijksregister. De Pearson-correlatie tussen beide datasets bedraagt 0.85.

Fig. 3: Bevolkingsdichtheid per km²: mobiele-telefoondata (links) en Census 2011 (rechts).

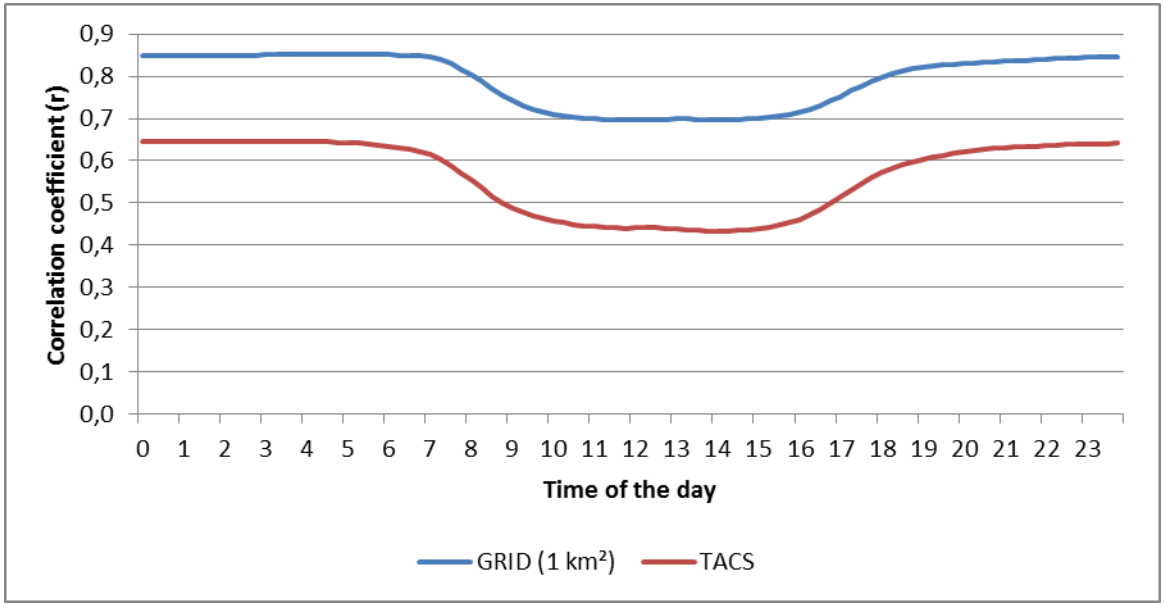


De gelijkenis tussen beide kaarten hierboven is treffend, maar één probleem is al duidelijk: in dunner bevolkte gebieden met grote TACS van meerder vierkante kilometers (bv. in de Ardennen in het zuidoosten) lijken de mobiele-telefoondata minder nauwkeurig. Een dieper gravende analyse toont andere locaties met een gebrekkige overeenstemming, bv. de haven van Antwerpen, de luchthaven van Zaventem, grote parken in Brussel, … Deze problemen kunnen aangepakt worden door meer gedetailleerd onderzoek (zie 6.).

Fig. 4 toont de correlatie tussen beide gegevensbestanden met een tijdsinterval van 15 minuten op donderdag 8 oktober 2015 (96 observaties), ofwel gebaseerd op roostervierkanten van 1km² (blauwe lijn, bovenaan) ofwel op TACS (rode lijn). Hun patroon is erg gelijkaardig, maar op een verschillend niveau. Het wekt geen verbazing dat de correlaties het hoogst zijn tijdens de nacht en dat ze een regelmatig 24-uurpatroon vertonen, met een vrij snelle daling in de morgen en een meer geleidelijke toename ‘s avonds. Wat echter wel verwondert, is dat de correlaties merkelijk hoger liggen voor de vierkanten van 1 km², tot 0.85 ’s nachts, terwijl ze voor

de TACS dan maar in de buurt van 0.65 liggen. Vermits op dit moment niet duidelijk is wat dit verschil veroorzaakt, is een eerste kandidaat voor verder onderzoek (zie 6) het testen van hypothesen over het optimale type en omvang van de gebieden om mobiele-telefoondata en statistische gegevens te combineren.

Fig. 4: Pearson-correlatie tussen mobiele-telefoondata en Census-gegevens elke 15 minuten op donderdag, voor 1 km² vierkanten (blauw, boven) en voor TACS (rood, onder).



4.2 Beoordeling van de geldigheid en nauwkeurigheid van mobiele-telefoondata voor het ramen van de bevolkingsdichtheid

Om problematische TACS in de mobiele-telefoondata die verdere analyse vereisen te identificeren, werden TACS-tellingen tijdens de nacht voor elke TACS-veelhoek vergeleken met schattingen over de daar woonachtige bevolking afkomstig van de Census. In een afhankelijkheidstabel (Fig. 5) tussen de dichtheidsdecielen van beide bronnen, blijkt een sterke concentratie van frequenties rond de diagonaal van de matrix; een berekening van de afstand tussen mobiele-telefoon- en censusdecielen toont een perfecte overeenstemming voor meer dan 35% van de TACS, terwijl in negen van de tien gevallen de afstand 2 decielen of minder bedraagt. Hieruit kan besloten worden dat beide gegevensbestanden een nauw verwante en geldige benadering van de bevolkingsdichtheid opleveren. Maar omdat het marktaandeel van Proximus wel hoog maar ver van volledig is en bovendien wellicht ongelijk verdeeld, worden grote lokale discrepanties vastgesteld (mogelijk op te lossen mits bijkomende informatie).

Fig. 5: Aantal TACS per deciel in de mobiele-telefoon- (y as)en Census-dataset (x-as).

# of ID_VORONOI	RK_STATBEL_POP_DENSITY										
RK_PROXIMUS_POP_DENSITY	0	1	2	3	4	5	6	7	8	9	Grand Total
0	515	328	102	52	16	11	6	4		1	1035
1	184	316	276	128	68	37	23	4			1036
2	92	160	295	267	127	54	31	9	1		1036
3	64	99	158	279	257	111	45	20	3		1036
4	43	50	94	161	288	241	117	30	11	1	1036
5	39	32	58	71	154	303	233	108	33	5	1036
6	43	21	27	47	80	171	318	211	101	17	1036
7	24	19	17	22	27	74	179	381	214	79	1036
8	16	9	6	7	17	27	73	219	418	244	1036
9	15	2	3	2	2	7	11	50	255	688	1035
Grand Total	1035	1036	1036	1036	1036	1036	1036	1036	1036	1035	10358

Het gedetailleerd in kaart brengen van de afstanden onthult interessante afwijking die uitnodigen tot verder onderzoek. Fig. 6 toon enkele voorbeelden (waarbij lichtgroen wijst op overeenstemming en graduele verschuivingen naar rood het omgekeerde):

- de linkse kaart toont de luchthaven van Zaventem, de rode veelhoeken komen overeen met de passagiersterminal waar niemand woont maar waar mobiele apparaten zelfs midden in de nacht gedetecteerd worden;

- de rode veelhoek op de rechterkaart dekt grotendeels maar niet volledig het Jubelpark in Brussel; alhoewel uiteraard onbewoond, worden er toch ’s nachts mobiele telefoons opgepikt; die kunnen ofwel een ‘overvloei’ zijn van de omringende appartementsgebouwen met meerdere verdiepingen (TACS vallen nooit exact samen met de ‘cel-footprint’, de effectieve antenneradius) of zelfs van nachtelijk verkeer op de autoweg die in het park ondergronds gaat.

Deze voorbeelden suggereren dat het toevoegen van gegevens over lokale omstandigheden de globale geldigheid en nauwkeurigheid van mobiele-telefoondataset significant kunnen verbeteren.

Fig. 6: Verschil in dichtheidsdecielen voor de luchthaven van Zaventem (links) en het Jubelpark in Brussel (rechts) – groen wijst op overeenstemming, rood op verschil.

4.3 Clusteranalyse van mobiele-telefoondata

De bedoeling van de clusteranalyse was het verifiëren of TACS gegroepeerd kunnen worden in een beperkt aantal categorieën met een karakteristiek en betekenisvol tijdspatroon.

Als we het gemiddelde van de genormaliseerde aantallen mobiele telefoon gedurende de dag bekijken, dan verklaren drie zinvol te interpreteren patronen op donderdag het grootste deel van de afname van de som van kwadraten binnen groepen (SSW) (zie Fig. 7, links):

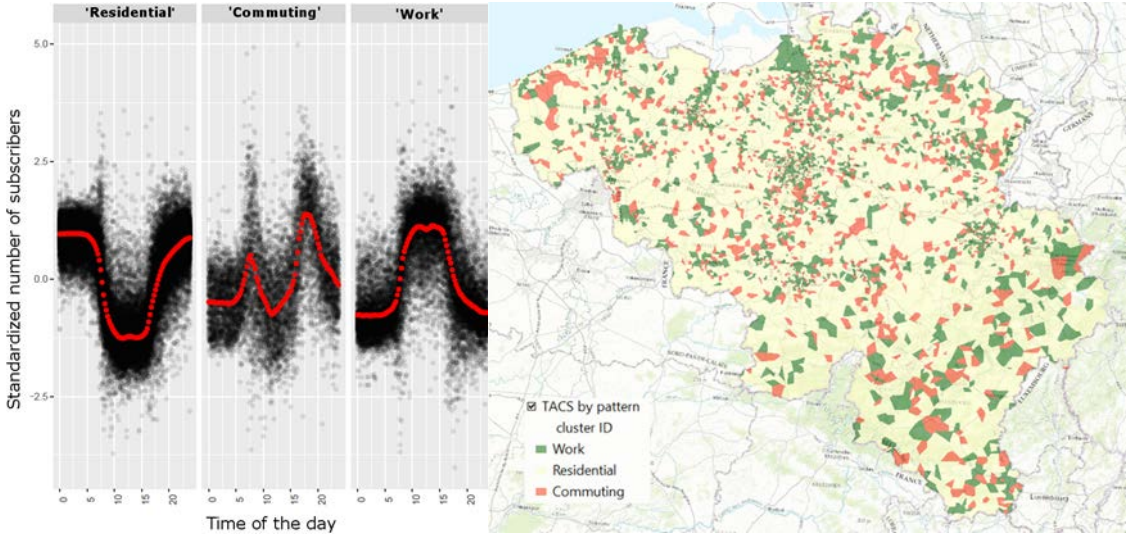
- Bovengemiddeld ’s nachts en ondergemiddeld tijdens de dag, overeenstemmend met woongebieden waar mensen ’s morgens vertrekken en ’s avonds terugkeren (cluster 2);

- Ondergemiddeld ’s nachts en bovengemiddeld tijdens de dag, wat een werkgebied suggereert met mensen die ’s morgens de TACS binnenkomen en deze ’s avonds weer verlaten (cluster 1);

- Twee pieken, één ‘s morgens (rond 7u30) en één ’s avonds (rond 18u), wat lijkt overeen te stemmen met een pendelgebied dat piekt tijdens de spitsuren.

Een geografische voorstelling van deze drievoudige klassering van TACS (Fig. 7, rechts) toont een coherent en weinig verrassend beeld, met een meerderheid van woongebieden (cluster 2) en enkele werkgebieden (cluster 1) en pendelzones (cluster 3) die deze eerste twee doorgaans verbinden.

Fig. 7: Weekdag-TACS geïdentificeerd als ‘werk’, ‘woon’ of ‘pendel’, met geografische voorstelling.



Het geval van de zondag is complexer, met een groter aantal minder voor de hand liggende clusters die moeilijker te interpreteren zijn. Verdere studie zal vereist zijn om dit patroon te begrijpen.

Een clusteranalyse van de gegevens per 1 km² roostervierkant toont in wezen gelijkaardige resultaten, en ook gegevens per uur in plaats van per 15 minuten beïnvloeden het patroon niet.

5. Bespreking

5.1 Mobiele-telefoondata als een geldige bron om bevolkingsdichtheid te ramen (validiteit)

De theoretische veronderstelling dat de woonplaats van mensen gelijk is aan de plek waar hun mobiele telefoon de nacht doorbrengt, wordt duidelijk ondersteund door de consistent hoge correlatie tijdens de nacht van 0.85 tussen tellingen van mobiele telefoons en de bevolkingsdichtheid afgeleid uit de Census die echter merkbaar afneemt gedurende de dag. Dit wordt verder bevestigd door de opvallende gelijkenis tussen het geografisch in kaart brengen van de mobiele telefoons 's nachts en de bevolkingsdichtheid op basis van het register, ook al hebben beide bronnen last van onvermijdelijke leemten en onnauwkeurigheden. Evidente voorbeelden wat de mobiele-telefoondata betreft: niet iedereen heeft een mobiele telefoon en sommige mensen hebben er meer dan één, de gegevens zijn afkomstig van slechts één netwerkoperator met een hoog maar toch beperkt en geografisch variabel marktaandeel, mensen en hun telefoons brengen niet allen de nacht door in hun woonplaats (bv. omwille van een toeristische trip, hospitaalverblijf, nachtwerk, ...). Anderzijds lijden bevolkingsregisters onder late beschikbaarheid of kunnen ze onvolledig zijn omdat sommige inwoners niet geregistreerd zijn of gewoonlijk verblijven op een andere plaats dan hun officiële domicilie.

Toch hebben ze allebei ook unieke voordelen: registers zijn vrij volledig en dus in grote mate representatief, terwijl mobiele-telefoondata de werkelijke actuele toestand weergeven, niet beïnvloed door effecten van non-respons of non-registratie. De combinatie van de unieke voordelen van beide zou moeten resulteren in statistieken die tegelijk geldiger, nauwkeuriger en meer tijdig zijn dan wanneer men maar één van beide gebruikt. Een statistische procedure zou uitgewerkt kunnen worden, waarbij geldige en nauwkeurige 'flash'-ramingen van de bevolking gebaseerd op mobiele-telefoondata en aanvullende datasets op geregelde tijdstippen gevalideerd en eventueel gecorrigeerd worden met behulp van het bevolkingsregister.

5.2 Correlatie tussen bevolkingsdichtheid op basis van mobiele-telefoondata versus Census-gegevens (nauwkeurigheid)

De hoge correlatiecoëfficiënten, rond 0.85 voor de gegevens van het rooster van vierkanten van 1 km² 's nachts (Fig. 4), tonen aan dat beide datasets het onderliggende concept van werkelijk aanwezige bevolking nauwkeurig capteren. Veel vastgestelde discrepanties kunnen verklaard worden met behulp van aanvullende datasets (zie 6.). Als deze mee in rekening gebracht worden, zullen de correlaties wellicht nog hoger uitkomen.

De clusteranalyse demonstreert dat kleine geografische gebieden gekarakteriseerd kunnen worden op basis van het wisselend aantal mobiele telefoons die ze bevatten en bevestigt dus de geldigheid en nauwkeurigheid van de mobiele-telefoondataset.

5.3 Hoe kan de waarde van mobiele-telefoondata verder opgedreven worden?

Elke dataset die spatiaal of temporeel georganiseerd is op een manier die overlapt met de mobiele-telefoondataset, kan gebruikt worden om deze laatste beter te interpreteren en om variatie te identificeren die vervolgens uitgefilterd kan worden. Voorbeelden zijn meteorologische gegevens, agenda's (feestdagen, evenementen), bodemgebruik (bv. wegen, spoorwegen, treinstations) en gelijkaardige gegeocodeerde datasets, informatie over incidenten op een bepaalde tijd en plaats, …

Een tweede potentiële verbetering betreft de optimale spatiale en temporele granulariteit (mate van detail) van de mobiele-telefoon. In de huidige studie werden de toestellen elke 15 minuten gedurende twee dagen geteld (2 x 96 tijdstippen) voor ongeveer 11,000 TACS (die samen het Belgisch grondgebied van 30.528 km² afdekken). Een meer frequente en zelfs continue registratie is mogelijk, voor kleinere gebieden en zelfs voor individuele toestellen (wat privacyproblemen oproept die uiteraard eerst aangepakt zouden moeten worden). Een grotere temporele of spatiale granulariteit en langere observatieperiodes zullen de omvang van de dataset doen toenemen, mogelijk tot buiten het bereik van de beschikbare capaciteit. In een specifieke statistische context is misschien niet al het detail dat mogelijk is ook nodig, het is dus aangewezen om de optimale omvang en detail te onderzoeken. Voor de raming van de werkelijke bevolkingsdichtheid, bijvoorbeeld, kan wellicht een beperkt aantal registraties tijdens de nacht (bv. om de twee uur) volstaan. Om schommelingen in de werkelijk aanwezige bevolking in een bepaald gebied te bepalen, is het interval van 15 minuten gebruikt in de huidige studie, wellicht adequaat. Microstudies op een precieze locatie (bv. om verkeersstromen te meten) vereisen echter meer frequente waarnemingen over langere perioden.

6. Verder onderzoek

De huidige studie beperkte zich bewust tot het exploreren van een totaal nieuw type van data en het evalueren van de geldigheid en nauwkeurigheid met eerder bescheiden onderzoeksvragen. Maar zelfs in dit stadium werd het al duidelijk dat er een groot potentieel is voor verdere analyses, waarvan sommige al gestart zijn. Daarbij kunnen twee benaderingen onderscheiden worden: het optimaliseren van de huidige analyse; en, als een volgende stap, het stellen van nieuwe statistische vragen en het identificeren van de mobiele-telefoondatasets die nodig zijn om deze te beantwoorden.

6.1 Huidige dataset

De studie van de mobiele-telefoondataset heeft talloze nieuwe onderzoeksvragen gegenereerd; sommige daarvan zijn al in onderzoek en zullen in vervolgpublicaties gerapporteerd worden. Een niet-exhaustieve lijst:

- Optimale temporele resolutie (periode en frequentie) van mobiele-telefoondata om de werkelijk aanwezige bevolking te ramen.
- Beste geografische onderverdeling om mobiele-telefoondata te koppelen aan statistische datasets: TACS of roostervierkanten (van 1 km² of zelfs kleiner, nu mogelijk dankzij vooruitgang in mobiele technologie).
- Optimale omvang / resolutie van spatiale eenheden, afhankelijk van de bestudeerde fenomenen of welke specifieke statistische resultaten dit moet opleveren.
- Haalbaarheid om mobiele datasets op een meer elementair niveau te koppelen, via geocoördinaten, en met name precies gelokaliseerde mobiele toestellen en gegeocodeerde statistische data.
- Systematisch identificeren van probleemgebieden (waar de datasets niet overeenstemmen) en deze oplossen via gedetailleerde lokale kennis (cfr. de voorbeelden in 4.2).
- Toevoegen van aanvullende spatiaal-temporele datasets om de onverklaarde variatie te reduceren; enkele voorbeelden: bodemgebruik, urbanisatiegraad, grenzen van de bebouwde kom, verkeersinfrastructuur (wegen, spoorwegen, treinstations, luchthavens, enz.).

6.2 Nieuwe dataverzoeken

Veel andere statistische vragen kunnen beantwoord worden met behulp van gelijkaardige datasets, uitgebreid of gewijzigd voor een specifieke doelstelling:

- Patronen van arbeidsmobiliteit kunnen gedetecteerd worden door het vergelijken van TACS-types met Censusdata over daar residerende actieve versus niet-actieve bevolking.

- Pendelgedrag kan gedetailleerd bestudeerd worden door het vergelijken van de weekdagen, en dit te combineren met mogelijk beïnvloedende factoren (weer, seizoen, specifieke incidenten of gebeurtenissen, …).

- Pendelen over de grenzen heen, arbeidsmigratie, internationaal toerisme, enz. kan onderzocht worden door het combineren van tellingen van buitenlandse mobiele apparaten, gegevens over roaming (inkomend en uitgaand) en gegevens afkomstig van operatoren van omliggende landen (bv. Luxemburg), waaruit een meer omvattend Europees beeld geschetst kan worden.

- Als mobiele telefoons individueel gevolgd worden en aan andere data gekoppeld via een individuele sleutel in plaats van op een geaggregeerd geografisch niveau zoals hier gedaan werd, wordt het mogelijk om bewegingen in tijd en ruimte te meten en om de meest waarschijnlijke woonplaats, werkplaats en ‘gebruikelijke omgeving’ te bepalen; die zijn essentieel om in een later stadium gedetailleerde statistieken te kunnen produceren over pendelen, voorkeuren inzake vervoersmodus, arbeidsmigratie, migratie, toerisme, en misschien zelfs over tijdgebruik en levensstijl – maar alle privacykwesties moeten eerst geregeld worden!

- Tenslotte is er nog ruimte voor verbetering van de nauwkeurigheid in het lokaliseren van mobiele apparaten via triangulatietechnieken, netwerkverdichting, het gebruik van GPS-gegevens, het koppelen met andere bronnen zoals wifi-lokalisatiegegevens, enz. Dit zal de precisie verhogen en het mogelijk maken om vragen te beantwoorden die voorheen onbeantwoordbaar waren, maar tegelijk vergroot dit de omvang en de complexiteit van de datasets.

7. Besluiten

Een vergelijking van mobiele-telefoondata met op registers gebaseerde censusgegevens toont aan dat ze een geldige en nauwkeurige bron zijn om de werkelijk aanwezige bevolking mee te ramen. Daarenboven zijn mobiele-telefoondata uitermate recent, gemakkelijk te berekenen en niet afhankelijk van subjectieve antwoorden. Hun kwaliteit in deze context kan nog verder verhoogd worden door de integratie van andere gedetailleerde spatiaal-temporele datasets.

Maar mobiele-telefoondata vormen ook een uitdaging vanuit statistisch perspectief. De data zijn vooreerst nieuw en grotendeels onontgonnen, en wellicht vertekend op onbekende en misschien onkenbare manieren (bv. geen één-op-één verband tussen personen en toestellen, netwerken die maar deels de totale populatie afdekken, selectiviteit met betrekking tot leeftijd, geslacht en andere belangrijke variabelen). Andere kwesties zijn gegarandeerde toegang tot de data doorheen de tijd, de omvang van de datasets vergeleken met de opslag- en verwerkingscapaciteit van statistische instituten, informatie over de gedane voorbewerkingen, en misschien het belangrijkste: bezorgdheid over privacy en andere juridische aspecten zoals eigendomsrechten over de data of vertrouwelijkheidsgaranties naar de netwerkoperatoren toe.

De logisch volgende stappen, het uitbreiden naar andere types van mobiele data, langere tijdsperiodes, grotere spatiale en temporele granulariteit en het gebruik van relevante aanvullende data, zijn uiterst veelbelovend, niet alleen voor bevolkings- en migratiestatistieken, maar ook voor domeinen als mobiliteit en transport, arbeidsmobiliteit en –migratie, en toerisme.

Een cruciale voorwaarde voor langetermijnsucces bij het integreren van mobiele-telefoondata in officiële statistieken is een wederzijds voordelig partnerschap tussen mobiel-netwerkoperatoren en statistische instituten. Het is duidelijk dat officiële statistieken hier erg veel bij te winnen hebben, maar ook voor de operatoren moet de noodzakelijke en niet te onderschatten investering om hun data voor statistische doeleinden te verwerken, gecompenseerd worden door dieper inzicht in hun eigen data en toegang tot waardevolle aanvullende datasets, zodat ze hun mobiele-telefoondata op een succesvolle en winstgevende manier kunnen exploiteren.

De poging om mobiele-telefoondata te gebruiken voor het maken van officiële statistieken is het eerste teken van een nakende paradigmaverschuiving in de statistiek. In de toekomst zou het best kunnen dat statistieken bijna onmiddellijk en zonder ondervraging van burgers of ondernemingen geproduceerd worden op basis van alomtegenwoordige big data van allerlei aard, waar nodig aangevuld met administratieve bestanden en op regelmatige tijdstippen gevalideerd door beperkte enquêtes.

8. Literatuurlijst

European Commission (2014): Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics, Eurostat

L. Altin, M. Tiru, E. Saluveer & A. Puura (2015): Using Passive Mobile Positioning Data in Tourism and Population Statistics, NTTS 2015 Conference abstract

P. Deville, C. Linarde, S. Martine, M. Gilbert, F.R. Stevens, A.E. Gaughan, V.D. Blondela & A.J. Tatem (2014): Dynamic population mapping using mobile phone data, PNAS 2014 111 (45) 15888-15893

F. Ricciato, P. Widhalm, M. Craglia & F. Pantisano (2015): Estimating Population Density Distribution from Network-based Mobile Phone Data, JRC Technical Report

Bezoek onze website
www.statbel.fgov.be

FOD Economie, K.M.O., Middenstand en Energie
Algemene Directie Statistiek - Statistics Belgium

Communicatieverantwoordelijke Stephan Moens
statpress@economie.fgov.be
North Gate - Koning Albert II-laan 16 - 1000 Brussel
E-mail: statbel@economie.fgov.be

Ondernemingsnummer: 0314.595.348
Verantwoordelijke uitgever: Nicolas Waeyaert
Norht Gate - Koning Albert II-laan 16 -1000 Brussel