Statistics Belgium

ANALYSE

Big data et statistiques

un recensement tous les quarts d'heure…

Big data et statistiques : un recensement tous les quarts d'heure…

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 DG Statistique – Statistics Belgium, Bruxelles, Belgique, marc.debusschere@economie.fgov.be

2 Proximus, Bruxelles, Belgique, freddy.demeersman@proximus.com

3 Eurostat, Luxembourg, Luxembourg, albrecht.wirthmann@ec.europa.eu

Résumé

Dans le cadre d'un projet conjoint, la DG Statistique, Proximus et Eurostat ont examiné la possibilité d'exploiter les données de la téléphonie mobile (provenant de Proximus) pour la production de statistiques officielles. Une étude préliminaire pose les premiers jalons en examinant s'il est possible de mesurer la population réellement présente de manière valable et précise, et ce, en comparaison avec le Census 2011.

Cet article est paru précédemment dans le numéro 8 du Carrefour de l'Economie du SPF Economie.

1. Introduction : la troisième révolution des données

Dès leur apparition au début du 19e siècle, les statistiques officielles s'appuyaient sur des enquêtes menées auprès des citoyens et des entreprises. Afin de réduire les coûts et la charge de réponse, le recours aux fichiers administratifs n'a cessé de prendre de l'ampleur ces 20 dernières années. Le Census 2011 a ainsi marqué un tournant important en Belgique : contrairement au recensement précédent de 2001, pour lequel il a fallu encore compléter et traiter des millions de formulaires, le Census 2011 reposait entièrement sur des bases de données administratives.

L'heure de la troisième révolution en statistique est toutefois venue, celle du big data. On laisse tous en continu des traces électroniques derrière nous, que ce soit au moyen de capteurs, de caméras, de paiements électroniques ou de retraits, d'activités en ligne en tout genre, etc. En principe, il est possible d'exploiter cette masse immense de données, en grande partie déstructurée, afin de produire les statistiques actuelles plus rapidement et plus efficacement, voire même de décrire des phénomènes inexplorés jusqu'à présent.

Les données de la téléphonie mobile constituent une source potentielle particulièrement prometteuse, et ce, dans nombre de domaines statistiques : population, migration, mobilité et migration de la main-d'œuvre, transport, mouvements transfrontaliers, tourisme, etc. De récentes études pilotes menées dans différents pays ont montré que ces données massives constituent une alternative viable aux sources plus traditionnelles utilisées par les instituts nationaux de statistique (Altin et al., 2015 ; Deville et al., 2014 ; Commission européenne, 2014). Toutefois, avant de pouvoir intégrer cette nouvelle source d'information dans la production statistique régulière, il convient d'en évaluer minutieusement la qualité. Cet article se penche sur la validité et la précision des données de la téléphonie mobile en tant qu'instrument de mesure de la densité de population en Belgique, en dressant une comparaison avec les résultats du Census belge de 2011. Les données de téléphonie mobile proviennent de Proximus, le plus grand[1] opérateur de Belgique.

[1] 40,3 % de part de marché en 2012 (http://economie.fgov.be/fr/consommateurs/Internet/Telecommunications/teledistribution/).

Trois questions de recherche ont été traitées dans cet article :

(1) les données de la téléphonie mobile constituent-elles une source de données valable pour estimer la densité de population ? (validité) ;

(2) quel est le rapport entre la densité de population basée sur les données de la téléphonie mobile et celle basée sur les données du Census ? (précision) ;

(3) comment est-il possible d'accroître la valeur des données de la téléphonie mobile pour de telles applications ? (intégration des données et reproductibilité).

Tant les données de la téléphonie mobile que les données du Census constituent une approximation de la réalité et présentent quelques imperfections bien connues :

- les données du Census reflètent la population enregistrée sur la base du lieu de résidence mentionné dans le Registre national, lequel ne correspond pas nécessairement au domicile effectif ;

- les données de la téléphonie mobile, quant à elles, constituent une approximation de la population réellement présente dans une zone déterminée ; ces données fournissent, pendant la nuit, une bonne indication du lieu de résidence effectif, mais peuvent introduire un certain biais en raison d'une couverture incomplète (plus d'un appareil par personne ou aucun, parts de marché variables si tous les opérateurs ne fournissent pas des données, situations atypiques en termes de logement ou d'emploi, etc.).

La qualité des deux sources de données peut être améliorée par de plus amples analyses, davantage d'observations et l'utilisation de sources d'information complémentaires. Dans le cas des données de la téléphonie mobile, par exemple, l'observation des appareils individuels sur une plus longue période permettrait de déterminer un « lieu de résidence probable ».

La mesure dans laquelle les deux sources convergent constitue la limite inférieure de leur validité et de leur précision. Nous présupposons qu'une forte corrélation entre les données du Census et les comptages des téléphones portables pendant la nuit

démontre que ces deux sources de données constituent un instrument de mesure valable et précis de la population résidente effective.

2. Données

2.1 Données de la téléphonie mobile

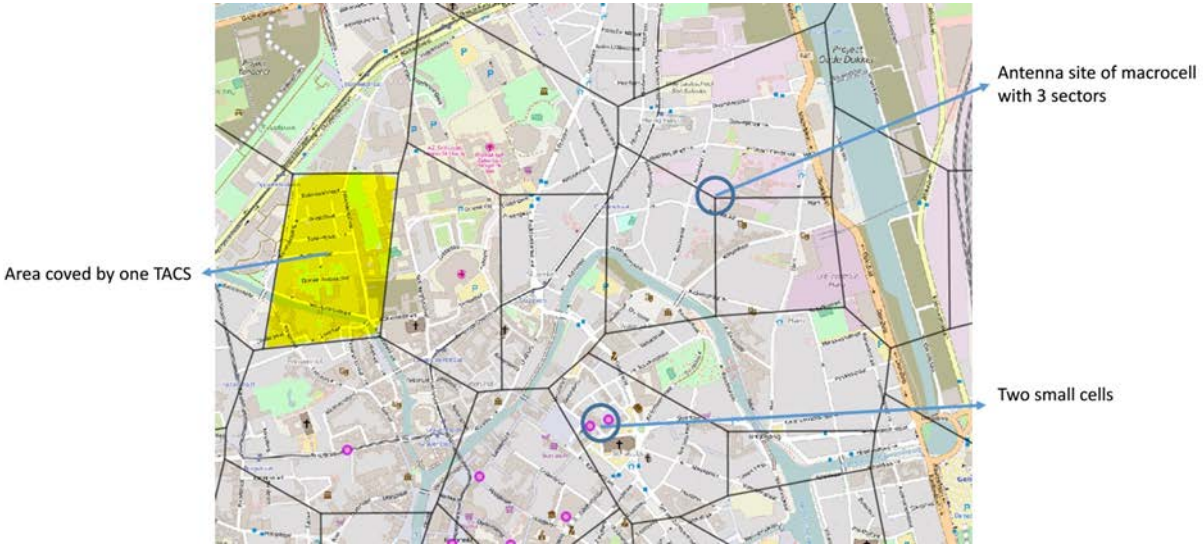
La plupart des études (Commission européenne, 2014) ont jusqu'à présent utilisé les CDR (call detail records ou enregistrements détaillés des appels), qui sont nécessaires pour pouvoir facturer. Ces enregistrements indiquent le lieu et le moment de chaque utilisation d'un téléphone portable. Il existe pourtant une source de données beaucoup plus complète : le système d'exploration du réseau (network probing system) capte chaque activité de signal, en ce compris les transactions non facturables. Il permet ainsi de fournir des détails spatio-temporels beaucoup plus précis. Au sein du réseau Proximus, le nombre de signaux exploitables est environ dix fois plus élevé que le nombre de CDRs. Pour chaque appareil présent sur le réseau, une position est déterminée au minimum toutes les trois heures. Dans le cas d'une connexion de données active, cet intervalle passe à environ une fois par heure. En pratique, la fréquence d'enregistrement des transactions est même encore plus élevée, surtout pour les smartphones qui se connectent fréquemment au réseau sans que le propriétaire en soit conscient. Pour les technologies plus modernes comme la 4G, davantage de localisations sont disponibles. Les intervalles diminuent encore lorsque les appareils sont en mouvement, car ils effectuent une mise à jour de la position à chaque fois qu'ils passent d'une zone de localisation[2] à une autre.

Lors de chaque transaction d'un téléphone portable sur un réseau mobile, la position de ce téléphone est connue jusqu'au niveau de la cellule. Un réseau de téléphonie mobile est un système cellulaire qui est devenu plus complexe au fil du temps : en règle générale, les sites d'antennes couvrent actuellement des cellules et des technologies multiples (2G, 3G et 4G). Dans le cadre de cette étude, le concept de TACS (Technology-Agnostic Cell Sector ou secteur cellulaire indépendant de la technologie sous-jacente) a été développé. Il s'agit de la zone desservie par toutes les cellules ayant le même azimut (direction du lobe principal de l'antenne), quelle que soit la technologie utilisée ; cette zone comprend toutes les positions qui se situent plus près de la station cellulaire que des cellules environnantes (chacune étant un TACS également). Les polygones qui en résultent sont présentés sous forme de

[2] Une « zone de localisation » est un regroupement logique de cellules (une « cellule » est la zone couverte par une antenne) ; un téléphone portable en mode inactif n'émet pas de signaux vers le réseau lorsqu'il change d'une cellule à l'autre, sauf s'il passe dans une nouvelle zone de localisation.

diagrammes de Voronoï[3], ce qui permet d'établir un modèle simplifié du réseau mobile. En faisant ainsi abstraction de la complexité des différentes couches technologiques et de la relation réelle et complexe entre les grandes et les petites cellules[4], il est possible de procéder à des calculs rapides et efficaces.

Fig. 1 : macro-TACS et regroupement des petites cellules avec leurs TACS « mères »



Une carte (*heatmap*) a ensuite été créée sous forme de polygones représentant chacun un TACS. Cette carte montre le nombre d'appareils détectés dans chacun d'eux. Les discontinuités dans le temps ont été résolues par interpolation (un appareil est supposé se trouver à l'endroit de sa dernière position jusqu'au moment où une nouvelle position est connue) et des filtres supplémentaires

[3] Il s'agit de la subdivision d'une surface plane en zones (polygones de Voronoï) sur la base de la distance jusqu'aux points centraux, soit les antennes dans le cas présent ; chaque point à l'intérieur d'un polygone se trouve plus près de cette antenne que des autres antennes.

[4] Les microcellules (couvrant une petite zone, p.ex. une partie d'une rue), picocellules ou femtocellules (généralement pour la couverture intérieure) sont regroupées avec leurs macro-TACS (cellules « mères »).

ont été appliqués (p.ex. suppression des communications de machine à machine). Les données ont été enregistrées toutes les 15 minutes lors d'un jour de semaine (jeudi 8 octobre 2015) et un dimanche (11 octobre 2015).

Pour éviter tout problème en matière de vie privée, toutes les données utilisées à ce stade étaient des agrégats (comptages des appareils par TACS), qui ne contenaient aucune information pouvant être associée à une personne en particulier.

2.2 Données du Census

Les données du Census[5] recensent la population belge au 1er janvier 2011. La variable « lieu de résidence » correspond au domicile repris dans le registre de population. Celle-ci est utilisée comme variable substitutive du lieu où séjournent habituellement les individus pendant la nuit, au petit matin et durant la soirée.

Ces données ont été agrégées tant au niveau d'une grille de cellules de 1 km² que des TACS. Pour produire les données relatives aux cellules de 1 km² et aux TACS pour le Census, les adresses dans le registre de population ont été géocodées à l'aide d'informations provenant du cadastre (la base du registre des logements et des bâtiments).

3. Méthodologie

Pour pouvoir comparer les données de la téléphonie mobile et les données du Census, elles doivent d'abord être insérées dans un espace géographique commun, soit la grille européenne standard (1 km x 1 km), soit les TACS (diagrammes de Voronoï représentant le réseau de téléphonie mobile).

3.1 Conversion des données de la téléphonie mobile en grille européenne standard de 1 km²

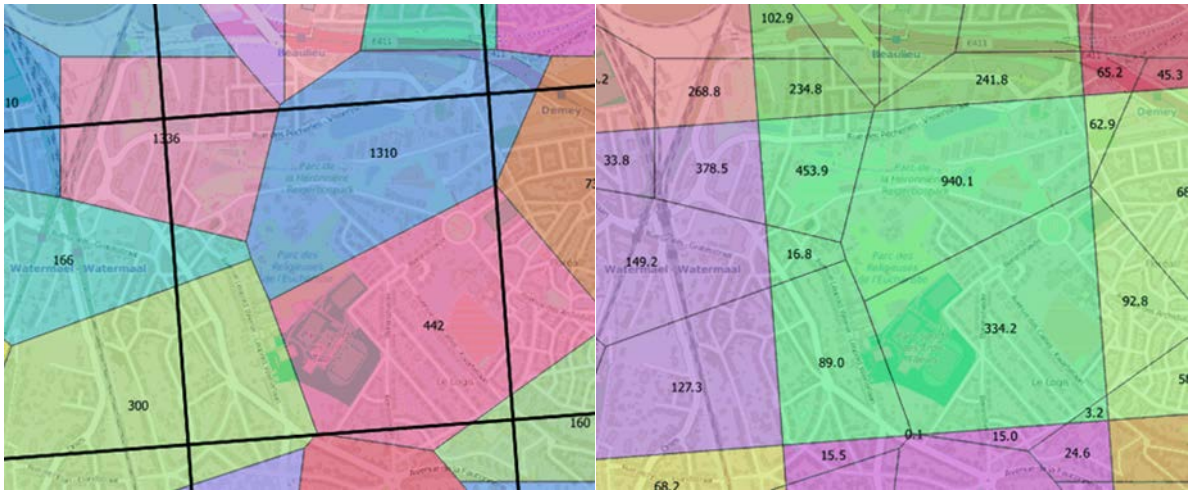
Alors que les données de la téléphonie mobile sont organisées en TACS de tailles différentes (de relativement petite taille dans les villes à plusieurs kilomètres carrés dans les zones moins densément peuplées), le Census se base quant à lui sur une grille dont les

[5] Le Census 2011 reposait entièrement sur des sources administratives (p.ex. registres de population, sécurité sociale, impôts, logements, entreprises, niveau d'enseignement).

cellules ont une taille constante de 1 km² (voir Fig. 2 avec un exemple pour Bruxelles ; les nombres dans les polygones sont les comptages des téléphones portables).

Les comptages réalisés au niveau des TACS sont répartis, par attribution proportionnelle, sur les km² de la grille de cellule.

Fig. 2 : Comptages par TACS convertis, par attribution proportionnelle, en km² pour une petite zone à Bruxelles



Cette méthode fonctionne très bien pour les zones où les TACS sont relativement petits, mais pose problème lorsqu'un TACS couvre une surface importante qui compte des zones inhabitées (p.ex. les forêts ardennaises). La population est supposée être distribuée de manière uniforme sur l'ensemble du TACS. Elle sera répartie sur les km² couverts totalement ou partiellement par le TACS. Une zone non habitée pourra ainsi se voir attribuer une partie du comptage. Ce problème pourrait être surmonté en utilisant des fichiers de données supplémentaires (p.ex. sur contour des zones habitées).

3.2 Conversion de la densité de population sur la base des données du Census en TACS

C'est l'inverse de l'approche précédente. Toutefois, au lieu d'attribuer proportionnellement la population des km² aux TACS (en suivant la méthodologie décrite ci-dessus), les points d'observation du Census (à savoir les adresses) ont directement été attribués aux polygones. Nous obtenons donc un comptage précis du nombre de personnes résidant dans un TACS.

3.3 Analyse statistique du fichier des données de la téléphonie mobile : clustering

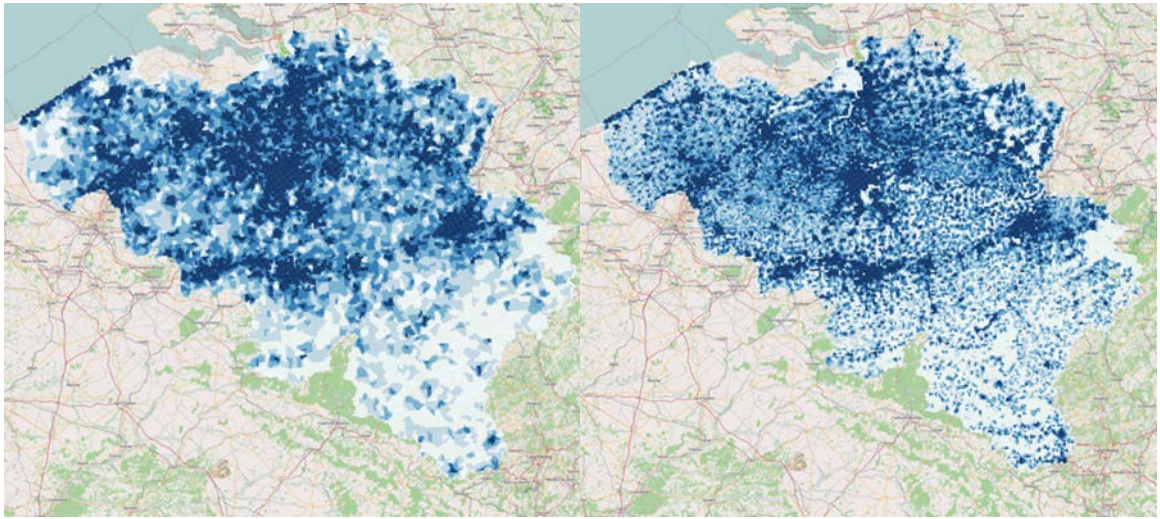
Un clustering a été effectué sur les mesures des TACS et des cellules de grille de 1 km² pour les deux jours pris séparément, d'une part, et pour les deux jours pris ensemble. Les valeurs absolues des comptages des appareils ont été standardisées (fixation de la moyenne à 0 et de l'écart-type à 1 - en utilisant la fonction « scale » en R, version 3.2.3.) avant de procéder à un clustering K-means (package stats, algorithme de Hartigan et Wong avec centres aléatoires). Le nombre optimal de clusters a été déterminé par la somme des carrés intra-groupes (SSW dans R), ce qui a abouti finalement à trois clusters pour les données du jeudi et à quatre pour celles du dimanche. Afin d'en tester la plausibilité, les résultats ont été représentés graphiquement et associés à des cartes topographiques afin de vérifier si les trois clusters pour le jour de semaine cadraient bien avec certaines caractéristiques topographiques. Des corrélations entre le nombre de téléphones portables et la population présente sur la base du Census ont été calculées au moyen de R pour chaque cluster et leur structure au fil des heures a été analysée.

4. Résultats

4.1 Estimation de la densité de population : données de la téléphonie mobile versus Census 2011

Les deux cartes ci-dessous montrent la densité de population par km². La carte de gauche est basée sur les comptages TACS des téléphones portables le jeudi 8 octobre 2015 à 4h00, convertis en cellules de grille de 1 km², tandis que la carte de droite est basée sur le Census 2011, dérivé des fichiers administratifs du Registre national. Le coefficient de corrélation de Pearson entre les deux séries de données s'élève à 0,85.

Fig. 3 : densité de population par km² : données de la téléphonie mobile (à gauche) et Census 2011 (à droite)

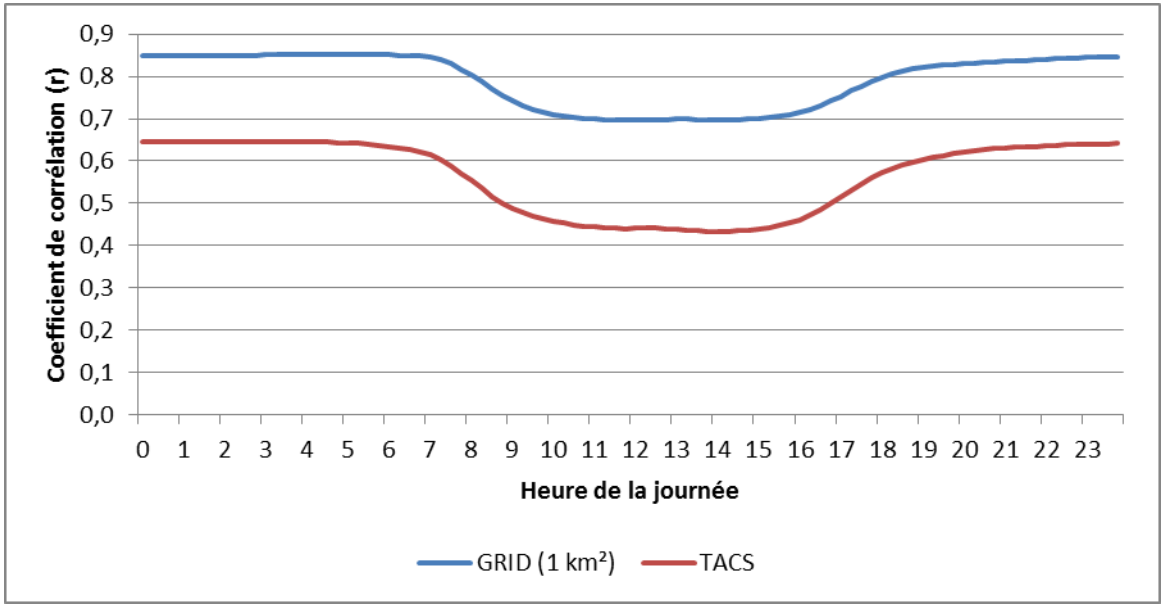


La ressemblance entre les deux cartes ci-dessus est frappante, mais un problème est déjà apparent : dans les zones moins densément peuplées avec des TACS étendus de plusieurs kilomètres carrés (p.ex. dans les Ardennes au sud-est), les données de la téléphonie mobile semblent moins précises (explication : voir point 3.1). Une analyse plus en profondeur met en évidence d'autres zones où la concordance fait défaut, p.ex. le port d'Anvers, l'aéroport de Zaventem, les grands parcs à Bruxelles, etc. Il sera peut-être possible de remédier à ces problèmes par des recherches plus approfondies (voir 6).

La figure 4 montre la corrélation entre les deux fichiers de données avec un intervalle de temps de 15 minutes pour le jeudi 8 octobre 2015 (96 observations), soit sur la base des km² (ligne bleue, en haut), soit des TACS (ligne rouge). Leur structure est très semblable, mais à un niveau différent. Il n'est guère étonnant de constater que les corrélations sont les plus élevées pendant la nuit et qu'elles présentent une structure régulière sur 24 heures avec une baisse relativement rapide dans la matinée et une hausse plus progressive dans la soirée. Par contre, il est surprenant de constater que les corrélations se révèlent nettement supérieures pour les

km² avec un coefficient allant jusqu'à 0,85 la nuit, contre un coefficient d'environ 0,65 pour les TACS. Étant donné que la cause de cet écart n'est pas claire à l'heure actuelle, la première piste de recherche (voir 6) est de tester des hypothèses sur la taille et le type de zones les plus appropriés pour pouvoir combiner les données de la téléphonie mobile et les données statistiques.

Fig. 4 : corrélation de Pearson entre les données de la téléphonie mobile et les données du Census avec un intervalle de 15 minutes le jeudi, en cellules de 1 km² (bleu, en haut) et en TACS (rouge, en bas)



4.2 Évaluation de la validité et de la précision des données de la téléphonie mobile pour l'estimation de la densité de population

Afin d'identifier dans les données de la téléphonie mobile les TACS problématiques qui nécessitent une analyse plus approfondie, les comptages TACS effectués pendant la nuit ont été comparés avec les estimations de la population résidente dans le TACS calculées à partir du Census. Un tableau de contingence (Fig. 5) entre les déciles de densité des deux sources montre une forte concentration de fréquences autour de la diagonale de la matrice. Un calcul de la distance entre les déciles de la téléphonie mobile et du Census montre une concordance parfaite pour plus de 35 % des TACS et une distance de deux déciles, au plus, dans neuf cas sur dix. On peut en conclure que les deux fichiers de données permettent une approximation assez proche et valable de la densité de population. Le décalage constaté s'explique probablement par le fait que la part de marché de Proximus n'est pas constante. Elle varie d'un TACS à l'autre.

Fig. 5 : nombre de TACS par décile dans les fichiers de données de la téléphonie mobile (axe des y) et du Census (axe des x)

# 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

Une cartographie détaillée des distances révèle des écarts intéressants, qui méritent une analyse plus en profondeur. La figure 6 présente quelques exemples (le vert clair y indique une concordance et les différentes nuances de rouge l'inverse) :

- la carte de gauche montre l'aéroport de Zaventem. Les polygones rouges correspondent au terminal de passagers où personne n'habite mais où des appareils portables sont détectés même au milieu de la nuit ;
- le polygone rouge sur la carte de droite couvre en grande partie, mais pas totalement, le Parc du Cinquantenaire à Bruxelles. Bien qu'il ne soit effectivement pas habité, on y détecte quand même des téléphones portables pendant la nuit. Les TACS ne correspondent jamais tout à fait à la couverture effective de l'antenne ou cell footprint. Les comptages se rapportent peut-être, en réalité, à une autre zone habitée, comme les immeubles à appartements qui se trouvent en bordure du parc ou à la circulation nocturne du tunnel routier du Cinquantenaire.

Ces exemples suggèrent que l'ajout de données sur les conditions locales peut améliorer significativement la validité et la précision globales des fichiers de données de la téléphonie mobile.

Fig. 6 : différence en déciles de densité pour l'aéroport de Zaventem (à gauche) et le Parc du Cinquantenaire à Bruxelles (à droite) – la couleur verte indique une concordance, la couleur rouge une différence

4.3 Clustering des données de la téléphonie mobile

L'objectif du clustering était de vérifier si les TACS pouvaient être regroupés en un nombre limité de catégories présentant une structure temporelle caractéristique et significative.

Si l'on examine la moyenne des nombres standardisés de téléphones portables pendant la journée du jeudi, trois structures expliquent la majeure partie de la diminution de la somme des carrés intra-groupes (SSW) et peuvent être interprétées de manière rationnelle (voir Fig. 7 à gauche) :

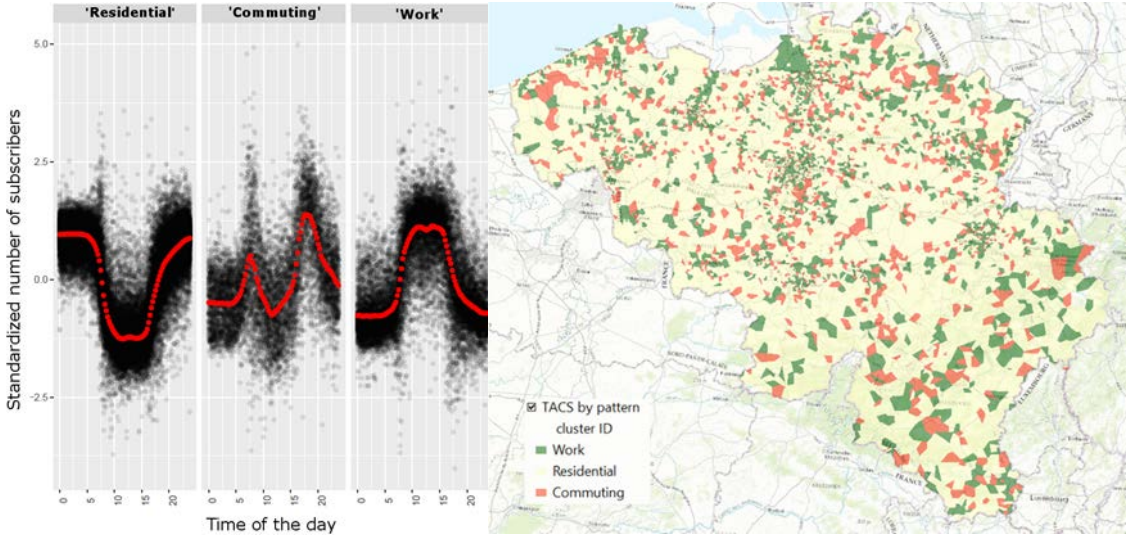
- supérieur à la moyenne la nuit et inférieur à la moyenne le jour, ce qui correspond à des zones résidentielles où les habitants partent le matin et rentrent le soir (cluster 2) ;

- deux pics, un le matin (vers 7h30) et un le soir (vers 18h00), ce qui semble correspondre à une zone de navette, où l'on observe des pics pendant les heures de pointe (cluster 3).

- inférieur à la moyenne la nuit et supérieur à la moyenne le jour, ce qui suggère une zone de travail où des personnes entrent dans les TACS le matin et les quittent le soir (cluster 1) ;

Une représentation géographique de cette classification des TACS en trois clusters (Fig. 7 à droite) montre une image cohérente et peu surprenante, avec une majorité de zones d'habitation (cluster 2) et quelques zones de travail (cluster 1) ainsi que des zones de navette (cluster 3), qui relient généralement ces deux premiers clusters.

Fig. 7 : TACS de semaine identifiés comme zones de travail, d'habitation ou de navette, avec représentation géographique



Le cas du dimanche est plus complexe, avec un plus grand nombre de clusters qui sont plus difficiles à interpréter. Une étude plus approfondie sera nécessaire pour comprendre cette structure.

Un clustering des données par cellule de grille de 1 km² montre grosso modo des résultats similaires. Travailler « par heure » plutôt que par « tranche de 15 minutes » n'influence pas les résultats.

5. Discussion

5.1 Les données de la téléphonie mobile en tant que source fiable pour l'estimation de la densité de population (validité)

L'hypothèse théorique selon laquelle le lieu de résidence des personnes correspond à l'endroit où leur téléphone portable passe la nuit, est clairement soutenue par la corrélation invariablement élevée de 0,85, observée pendant la nuit entre les comptages des téléphones portables et la densité de population tirée du Census, qui diminue toutefois sensiblement pendant la journée. Cette hypothèse est également confirmée par la ressemblance frappante entre des cartes de densité, même si les deux sources souffrent d'inévitables lacunes et imprécisions. En ce qui concerne les données de la téléphonie mobile, quelques exemples évidents sont : le fait que tout le monde ne possède pas un téléphone portable et que certaines personnes en possèdent plusieurs - le fait que les données ne proviennent que d'un seul opérateur disposant d'une part de marché importante mais tout de même limitée et variant d'un endroit à l'autre, le fait que toutes les personnes et leur téléphone ne passent pas chaque nuit à leur domicile (p.ex. en raison d'un voyage touristique, d'un séjour à l'hôpital, d'un travail de nuit, etc.). Quant aux registres de population, ils ne sont disponibles que tardivement ou peuvent être incomplets, étant donné que certains habitants ne sont pas enregistrés ou résident habituellement à un autre endroit que leur domicile officiel.

Toutefois, ces deux sources présentent aussi des avantages uniques : les registres sont relativement complets et donc largement représentatifs, tandis que les données de la téléphonie mobile reflètent la situation réelle actuelle, sans influence des effets liés à la non-réponse ou au non-enregistrement. La combinaison des avantages uniques des deux sources devrait permettre d'obtenir des statistiques à la fois plus valables, plus précises et disponibles plus rapidement qu'en utilisant l'une des deux sources isolément. Une procédure statistique pourrait être élaborée, dans laquelle des estimations « flash » valables et précises de la population basées sur des données de la téléphonie mobile et des fichiers de données complémentaires sont validées à intervalles réguliers et éventuellement corrigées à l'aide du registre de population.

5.2 Corrélation entre la densité de population sur la base des données de la téléphonie mobile et des données du Census (précision)

Les coefficients de corrélation élevés, aux alentours de 0,85 pour les données de la grille de cellules de 1 km² pendant la nuit (Fig. 4), montrent que les deux fichiers de données saisissent avec précision le concept sous-jacent de population réellement présente. De nombreuses disparités constatées peuvent s'expliquer à l'aide de fichiers de données complémentaires (voir 6). En les prenant en compte, les corrélations seraient probablement encore plus élevées.

Le clustering démontre que de petites zones géographiques peuvent être caractérisées sur la base du nombre variable de téléphones portables qu'elles abritent et confirme donc la validité et la précision du fichier de données de la téléphonie mobile.

5.3 Comment est-il possible d'accroître la valeur des données de la téléphonie mobile ?

Tout fichier de données, qui est organisé de manière spatiale ou temporelle de telle manière à chevaucher le fichier de données de la téléphonie mobile, peut être utilisé pour mieux l'interpréter et identifier des variations qui peuvent ensuite être filtrées. Il peut s'agir, par exemple, de données météorologiques, de calendriers (jours fériés, événements), de fichiers de données sur l'utilisation du sol (p.ex. les routes, les voies ferrées, les gares), de données géocodées similaires, d'informations sur les incidents survenus à un endroit et à un moment donnés, etc.

Une deuxième amélioration potentielle concerne la granularité (le niveau de détail) spatiale et temporelle optimale des données de la téléphonie mobile. Pour la présente étude, les appareils ont été comptabilisés toutes les 15 minutes pendant deux jours (2 X 96 moments) pour environ 11.000 TACS (qui couvrent ensemble le territoire belge de 30.528 km²). Un enregistrement plus fréquent et même continu est possible, pour des zones plus restreintes voire même des appareils individuels (ce qui soulève des problèmes en matière de protection de la vie privée qu'il convient bien entendu de résoudre au préalable). Une granularité temporelle ou spatiale plus fine et des périodes d'observation plus longues contribueront à augmenter la taille du fichier de données, peut-être même au-delà des capacités disponibles. Dans un contexte statistique spécifique, tous les détails qu'il est possible d'obtenir ne sont peut-être pas nécessaires. Il est donc préférable d'analyser la taille et le niveau de détail optimaux. Pour l'estimation de la densité de population réelle, par exemple, un nombre limité d'enregistrements pendant la nuit (p.ex. toutes les deux heures) pourrait suffire. Pour déterminer les fluctuations de la population réellement présente dans une zone donnée, l'intervalle de 15 minutes utilisé dans

la présente étude est sans doute suffisant. Des micro-études relatives à une localisation précise (p.ex. pour mesurer les flux de trafic) nécessitent toutefois des observations plus fréquentes sur de plus longues périodes.

6. Recherches complémentaires

La présente étude s'est volontairement limitée à l'exploration d'un tout nouveau type de données et à l'évaluation de la validité et de la précision sur la base de questions de recherche plutôt restreintes. Toutefois, à ce stade déjà, il est évident qu'il existe d'innombrables possibilités d'analyses complémentaires, dont certaines sont déjà en cours. On peut distinguer deux approches à cet égard : l'optimisation de l'analyse actuelle et, dans un second temps, la formulation de nouvelles questions statistiques et l'identification des fichiers de données de la téléphonie mobile nécessaires pour y répondre.

6.1 Fichier de données actuel

L'étude du fichier de données de la téléphonie mobile a généré d'innombrables nouvelles questions de recherche. Certaines sont déjà à l'étude et seront traitées dans de prochaines publications. En voici une liste non exhaustive :

- la résolution temporelle (période et fréquence) optimale des données de la téléphonie mobile pour estimer la population réellement présente ;
- la meilleure subdivision géographique pour coupler les données de la téléphonie mobile aux fichiers de données statistiques : TACS ou cellules de grille (de 1 km², voire moins, ce qui est désormais possible grâce aux progrès de la technologie mobile) ;
- la taille/résolution optimale des unités spatiales, en fonction des phénomènes étudiés ou des résultats statistiques spécifiques qui doivent être obtenus ;
- la faisabilité de coupler des fichiers de données mobiles à un niveau plus élémentaire au moyen des coordonnées géographiques, p.ex. appareils mobiles localisés précisément et données statistiques géocodées ;
- identifier systématiquement les zones problématiques (où les fichiers de données ne correspondent pas) et les résoudre grâce à des connaissances locales détaillées (voir exemples au point 4.2) ;

- ajouter des fichiers de données spatio-temporels complémentaires afin de réduire les variations inexpliquées. Quelques exemples : l'utilisation du sol, le degré d'urbanisation, les frontières des zones bâties, les infrastructures de transport (routes, voies ferrées, gares, aéroports, etc.).

6.2 Nouvelles demandes de données

Il est possible de répondre à de nombreuses autres questions statistiques au moyen de fichiers de données similaires, étendus ou modifiés pour un objectif spécifique :

- les schémas de mobilité de la main-d'œuvre peuvent être détectés en comparant les types de TACS aux données du Census sur la population active et non active qui y réside ;

- le comportement des navetteurs peut être étudié en détail en comparant les jours de semaine, et ce, en combinaison avec des facteurs susceptibles d'influencer ce comportement (temps, saison, incidents ou événements spécifiques, etc.) ;

- les navetteurs transfrontaliers, la migration de la main-d'œuvre, le tourisme international, etc. peuvent être étudiés en combinant les comptages des appareils portables étrangers, les données sur le roaming (entrant et sortant) et les données provenant des opérateurs de pays voisins (p.ex. le Luxembourg), permettant ainsi de dresser un bilan européen plus complet ;

- si les téléphones portables sont suivis individuellement et sont couplés à d'autres données au moyen d'une clé individuelle plutôt qu'à un niveau géographique agrégé, comme nous l'avons fait ici, il devient possible de mesurer les mouvements dans le temps et l'espace et de déterminer le lieu de résidence, le lieu de travail et l'environnement habituel les plus plausibles. Ces données sont essentielles pour pouvoir produire à un stade ultérieur des statistiques détaillées sur les navetteurs, les préférences relatives au mode de transport, la migration de la main-d'œuvre, la migration, le tourisme, voire peut-être même l'emploi du temps et le mode de vie. Toutes les questions relatives au respect de la vie privée doivent toutefois être réglées au préalable.

- enfin, il reste encore une marge d'amélioration de la précision en matière de localisation des appareils portables grâce aux techniques de triangulation, à la densification du réseau, à l'utilisation des données GPS, au couplage à d'autres sources comme les données de localisation par WiFi, etc. Cela permettra d'augmenter la précision et de répondre à certaines questions restées sans réponse précédemment, mais cela accroîtra simultanément la taille et la complexité des fichiers de données.

7. Conclusions

Il ressort d'une comparaison entre les données de la téléphonie mobile et les données du Census basées sur des registres qu'elles constituent une source valable et précise pour l'estimation de la population réellement présente. En outre, les données de la téléphonie mobile sont actuelles, faciles à calculer et non tributaires de réponses subjectives. Il est possible d'en accroître davantage la qualité par l'intégration d'autres séries de données spatio-temporelles détaillées.

Les données de la téléphonie mobile constituent également un défi d'un point de vue statistique. Tout d'abord, ces données sont nouvelles, en grande partie inexploitées, et probablement affectées par des biais inconnus et peut-être impossibles à estimer (p.ex. pas de relation univoque entre les personnes et les appareils, réseau qui ne couvre que partiellement le territoire, sélectivité par rapport à l'âge, le sexe et d'autres variables clés). D'autres préoccupations sont l'accès garanti aux données au fil du temps, le volume des fichiers de données par rapport à la capacité de stockage et de traitement des instituts de statistique, les informations relatives au pré-traitement, et peut-être le point le plus important : les questions relatives au respect de la vie privée et aux autres aspects juridiques tels que les droits de propriété des données ou les garanties de confidentialité envers les opérateurs des réseaux mobiles.

Les prochaines étapes logiques, à savoir : l'extension à d'autres types de données mobiles, le recours à des périodes de temps plus longues, une plus grande granularité spatiale et temporelle ainsi que l'utilisation de données complémentaires pertinentes, se révèlent très prometteuses, pas seulement pour les statistiques démographiques et migratoires, mais aussi pour des domaines tels que la mobilité et le transport, la migration et la mobilité de la main-d'œuvre ainsi que le tourisme.

Pour garantir le succès à long terme de l'intégration des données de la téléphonie mobile dans les statistiques officielles, l'une des conditions sine qua non est un partenariat mutuellement bénéfique entre les opérateurs des réseaux mobiles et les instituts de statistique. Il est évident que les statistiques officielles y ont beaucoup à gagner. Pour les opérateurs également, l'investissement

nécessaire pour pouvoir traiter leurs données à des fins statistiques, lequel n'est pas à sous-estimer, doit être compensé par une plus grande compréhension de leurs propres données et un accès à des séries de données complémentaires de qualité, de sorte qu'ils puissent exploiter les données de la téléphonie mobile de manière fructueuse et rentable.

Les efforts consentis en vue d'utiliser les données de la téléphonie mobile pour la production de statistiques officielles constituent les prémices d'une révolution imminente dans le domaine de la statistique. À l'avenir, les statistiques seront, sans doute, en partie produites quasi instantanément et sans la contribution des citoyens ou des entreprises sur la base de big data de toutes sortes, en les complétant si nécessaire par des fichiers administratifs et en les validant à intervalles réguliers par des enquêtes limitées.

8. Bibliographie

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

F. De Meersman, G. Seynaeve, M. Debusschere, P. Lusyne, P. Dewitte, Y. Baeyens, A. Wirthmann, C. Demunter, F. Reis F., H.I. Reuter (2016, in press), Assessing the Quality of Mobile Phone Data as a Source of Statistics

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

Consultez notre site web
www.statbel.fgov.be

SPF Économie, P.M.E., Classes moyennes et Énergie
Direction générale Statistique - Statistics Belgium

Responsable Communication Stephan Moens
statpress@economie.fgov.be
North Gate - Bd. du Roi Albert II, 16 - 1000 Bruxelles
E-mail : statbel@economie.fgov.be

Numéro d'entreprise : 0314.595.348
Editeur responsable : Nicolas Waeyaert
North Gate - Bd. du Roi Albert II, 16 - 1000 Bruxelles