New opportunities for statistics on population and mobility from the use of mobile phone data

Fernando Reis [(1)], Albrecht Wirthmann [(1)], Patrick Lusyne [(2)], Youri Baeyens [(2)], Freddy De Meersman [(3)], Marc Debusschere [(2)], Hannes Reuter [(1)], Gerdy Seynaeve [(3)]

[(1)] Eurostat, Luxembourg, Luxembourg

[(2)] Statistics Belgium, Brussels, Belgium

[(3)] Proximus, Brussels, Belgium

Corresponding author:

Fernando Reis

European Commission (Eurostat),

Joseph Bech building, 5 rue Alphonse Weicker, L-2721 – Luxembourg, Luxembourg.

E-mail: fernando.reis@ec.europa.eu

Abstract

The integration of big data in official statistics is now steadily moving forward in Europe. Several national statistical institutes have been assessing the use of big data sources, analytics and technologies for a few years already and a joint European project involving statistical organisations from 20 countries is currently exploring the potential of various big data sources for official statistics and is preparing the conditions for subsequent production of statistical data. This paper presents results from one such project involving Statistics Belgium, Eurostat and the Belgian mobile network operator Proximus. This project is analysing aggregated geolocation data collected by Proximus on the functioning of its mobile network for producing statistics on the population and its mobility. The initial approach was to attempt to reproduce existing statistics, in particular on the resident population, and to explore new statistics which may be of public interest, i.e. indicators of present de facto population at a high spatiotemporal detail. However, this new data source offers also the possibility to re-think existing statistical concepts of population and their operationalization, in directions probably not fully explored in the past because traditional data sources would not allow it. The current statistical concept of resident population allocates an individual to the place where he/she spends his/her resting period (normally the night) for at least 12 months. In a first step, we assess the sensitivity of resident population statistics to a change of the threshold from 12 months to 6 months or 3 months. We also investigate the possibility of dropping the time period altogether as a threshold and instead simply breaking the population in each spatial unit down by the amount of time they have been living there.

Keywords: Big data, mobile phone data, population, mobility

1. Introduction

Recent studies in several countries have demonstrated the potential of mobile phone data as a viable alternative to more traditional data sources used by national statistical institutes, most notably in the statistical domains of population, migration, tourism and mobility (Altin e.a., 2015, Deville e.a., 2014; European Commission, 2014). Quality issues need to be addressed, however, before they can be integrated into any regular statistical production. This paper focuses on the validity and accuracy of mobile phone data as a measure of resident population density in Belgium, to be compared with results of the Belgian Census 2011, produced by Statistics Belgium on the basis of the Belgian population register. The mobile phone data were obtained from Proximus, the leading mobile network operator in Belgium.

Three research questions are addressed:

(1) do mobile phone data constitute a valid source to assess population density? (validity);

(2) what is the relation between population density based on mobile phone versus Census data? (accuracy);

(3) how can the value of mobile phone data be further enhanced for this purpose? (data integration and replicability)

Both the mobile phone and Census datasets are approximations of reality, with known limitations:

- the Census data show the registered population based on the place of residence as recorded in the Population register, which is not necessarily the actual residence;
- mobile phone data, on the other hand, show the actual present population in an area, which at night should be highly indicative of the actual place of residence, but likely to be biased by incomplete coverage (more than one device per person or none, varying local market shares if not all operators provide data, atypical work or living arrangements, …).

The quality of both data sources can be improved by further analysis, more observations and the use of additional information sources. In the case of mobile phone data, for instance, observation of individual phones over a longer period should make it possible to assign a 'most likely living place' and hence more accurately assess the actual resident population.

The degree to which both sources converge provides a lower limit of validity and accuracy. We postulate that a high correlation between Census data and mobile phone counts at night indicate that both are valid and accurate measures of actual resident population.

2. Roadmap for the adoption of big data in the European Statistical System

The basis of the answer of the European Statistical System to the challenges and opportunities brought by big data is the Scheveningen memorandum agreed by the heads of the statistical authorities at national and European level in 2013.

The Scheveningen Memorandum recognises that the increasing level of digitisation of society, and the consequent digital footprint people leave, offers an opportunity for the compilation of statistics and that it should be incorporated in the conceptual design of official statistics. In particular, it provides alternatives to deal with the current challenges faced by official statistics, such as the decreasing response rates and the need to increase the overall efficiency of statistical production systems. However, the Scheveningen Memorandum also recognises that the use of big data poses challenges to the ESS. Therefore, it calls for an examination of the potential of big data sources and the development of an official statistics big data action plan and roadmap.

The Eurostat task force on big data worked together with the ESS Task force to elaborate several actions in the several horizontal areas shown in figure 1, including pilots of big data usage for producing official statistics. The purpose of the pilots is to gain experience in using big data in the context of official statistics. This includes identifying, analysing and solving issues in horizontal areas as well as investigating and developing future business models for statistical data production related to specific data sources or statistical products. The Big Data Action Plan and Roadmap was then endorsed by the European Statistical System Committee (ESSC) in 2014.

The overall purpose of the roadmap is to enable the ESS to gradually integrate big data sources into the production of European and national statistics. Given the rapid evolution of technology, availability of sources and public perception, it is premature to provide a very detailed description of the final “to be” state of the ESS related to big data usage. Therefore, the roadmap constitutes a high-level description of where we would like to be, in terms of long-term vision (beyond 2020), medium-term aims (by 2020) and short-term objectives (by the end of 2016).

3. Mobile phone data for population statistics

One of the pilots run at Eurostat was the use of mobile phone data for population, mobility and tourism statistics. Mobile phone data provides an excellent test case for the use of big data in official statistics. Firstly, it has an extraordinary potential as a data source for official statistics in a very large spectrum of applications. It provides a relatively high spatial and temporal resolution for the positioning of mobile devices, and ultimately of the persons carrying them, which can be used in the production of detailed population statistics. However, the geo-location of individuals and identification of their mobility profiles together with the network information (the set of links between persons reflected by their communication events) combined with other spatial layers, gives it many further potential applications. Secondly, although mobile phone data suffers from selectivity bias, there are official statistics on the penetration rate of mobile phones which can help in assessing and possibly correct this bias. Thirdly, it is a relatively sustainable source as the chances that it will cease to exist are relatively low compared to other big data sources.

The first most critical step in this project was the establishment of a research partnership between Eurostat, Proximus and Statistics Belgium. The three institutions agreed on the core research questions, the data to be used, the methodological approach, and the way in which the quality of results could be assessed.

Clearly the benefits and lessons from this collaborative work extend much beyond the scope of this project. While assessing the potential of mobile phone data for providing policy relevant insights about population, tourism, migration, mobility, transport and other related fields was our central target in this project. We also had the opportunity to address successfully some horizontal issues such as privacy, confidentiality, algorithmic transparency and quality assurance.

4. Description of the data used

4.1. Description of mobile phone data

The focus in most studies so far (European Commission, 2014) has been on CDRs, call detail records, used for billing purposes. These reveal the time and location of a mobile phone whenever it is used. Nowadays, however, network probing systems capture all signalling events, including non-billable transactions, and therefore offer a much better time granularity. In the Proximus network, the amount of useful signalling events is about 10 times higher than the amount of CDRs. For each device on the network a position is recorded at least every 3 hours; with an active data connection, this interval decreases to approximately once every hour. In practice transactions are recorded even more frequently, especially for smartphones which often connect to the network without the owner being aware of this. Also, with newer technologies such as 4G more location samples become available; and intervals are further reduced when devices perform location updates as they move from one location area to another.

For each transaction on a mobile network, the mobile phone location is known down to the level of a cell identity. A mobile phone network is a cellular system which over time has grown ever more complex; antenna sites nowadays typically contain multiple technologies (2G, 3G, and 4G) and multiple cells. For the purpose of this study, a construct called TACS (Technology-Agnostic Cell Sector) was developed: the area served by all cells on a particular site with the same azimuth (direction of antenna main lobe) and irrespective of the technology used, consisting of all locations closer to the cell site than to surrounding ones (each also TACS). The resulting polygons are represented as Voronoi diagrams, making it possible to build a simplified model of the mobile network. As a first step, only macro-TACS (roof top cells) are considered; smaller cells are then added as points and remapped to their “parent” Voronoi polygon (see Fig. 2). This representation of TACS as Voronoi polygons is an approximation of cell coverage, which disregards the complexity of the different technology layers and thus allows for fast and performant calculations.

A heatmap was then created in the form of Voronoi polygons for each TACS, showing the number of devices in each of them, on the basis of localisations by the mobile network. Discontinuities in time were resolved by interpolation (a device supposedly staying on its last known position, until a new one is known) and additional filtering was performed (e.g. machine to machine). Data were recorded every 15 minutes for one weekday (Thursday 8 October 2015) and one Sunday (11 October 2015).

To protect the confidentiality of individual data, all data used at this stage were aggregates (counts of devices per TACS) which do not contain any information traceable to individuals.

4.2. Description of the census data

The Census data record the Belgian population on 1 January 2011. The variable ‘place of residence’ refers to the registered place of residence as declared in the population register and is used as a proxy for the place where people usually reside during the night, in the early mornings and in the evenings.

These data were aggregated at the level of both 1 km² grids and Voronoi-areas. To produce km² grid data and Voronoi data for the Census, the addresses in the population register were geocoded on the basis of a match between the population register and data from the land registry which is the source for the register of dwellings and buildings.

5. Methods

5.1. Mapping mobile phone data to the 1 km² Standard European grid

While mobile phone data are organised in TACS varying in size (from quite small in cities to several square kilometres in less densely populated areas), the Census uses 1 km² grid areas (see Fig. 3 with an example in Brussels; the numbers in the polygons are mobile phone counts).

The number of devices recorded in each polygon (Voronoi count) is split up proportionally per area and the resulting subtotals are allocated to each of the different 1 km² grids they are part of. These can then be summed for each km² grid.

This method works very well for areas where TACS are relatively small but has its limits when large TACS cover huge areas where very few or no phones are present at night (e.g. the forests in the Ardennes). As population is evenly distributed across a TACS polygon in these cases too, it is obviously impossible to obtain a correct estimation of the population density per km². This problem could be mitigated by additional datasets (e.g. on land use).

5.2. Mapping of population density from Census data to TACS (Voronoi polygons)

This is the opposite of the previous approach. However, instead of assigning 1 km² population proportionally to the TACS or parts of TACS it contains, analogous to the method described above, census observation points (i.e. addresses) were allocated directly and therefore more precisely to a grid cell.

5.3. Analysis of the mobile phone dataset: cluster analysis

The cluster analysis was performed on the average number of devices, over the three days, for each period of 15 minutes in the day (96 points for a whole day) in each TAC. The absolute number of devices during the day were normalised to a mean of 0 and a standard deviation of 1 in each TAC (procedure 'scale' in R, version 3.2.3, 'base' package) and the clustering was performed with the k-means algorithm (function 'kmeans' from 'stats' R package, with Hartigan and Wong algorithm with random centres). The number of clusters was chosen based on the within-groups sum of squares, finally yielding three clusters. In order to verify their plausibility, results were displayed as charts and overlaid with topographic maps to see whether the three workday patterns coincide with respective topographic features. Correlations between number of mobile phones and Census resident population were calculated in R for the different clusters and their time patterns analysed. In addition, the contribution of each of the three clusters (profiles) to an individual Voronoi polygon was estimated using a linear regression model with constraints (using function 'sem' from 'lavaan' R package to estimate a structural equations model).

6. Results

6.1. Estimating population density from mobile phone versus Census 2011 data

The two maps in Fig. 4 visually represent population density per km² based on Voronoi polygons; the one on the left is derived from mobile phone counts per Voronoi polygon and calculated as an average between 3.45h and 4.30h in the morning over the three week days. The map on the right shows the population density based on the 2011 Census and geocoded using the dwellings and buildings register.

The density of the TACS was scaled according to deciles, partially compensating for the fact that the market share of the mobile phone provider is less than 100%. While both maps show similar density patterns, the map derived from the 2011 Census data shows higher densities in the areas that are adjacent to the agglomeration areas.

Fig. 5 shows the correlation between both datasets at 15 min. intervals during the observed days based on the TACS. Correlation between mobile phone data and census data is higher during the night time dropping in the morning, reaching a minimum around 14h and increasing between 16h and 19h. Although there is a striking similarity of the density of resident population and mobile phones visible in Fig. 4, the Pearson correlation only reaches a maximum of 0.65 during the night time. It seems that there are additional factors, such as delimitation of the TACS, market shares of the mobile phone operator or disparities in distribution of geographic locations between the total population and the mobile provider's clients that influence the correlation.

6.2. Cluster analysis of mobile phone data

The purpose of the cluster analysis was to verify whether TACS can be grouped into a limited number of categories with characteristic and meaningful temporal patterns.

Looking at the means of the normalised number of phones during the day shows three patterns during the weekdays which account for most of the reduction in the within-groups sum of squares (SSW) and which can be interpreted in a meaningful way (see Fig. 6):

- densities above average at night and below average during the day, corresponding to a residential area with people leaving in the morning and returning in the evening;

- densities below average at night and above average during the day, suggesting a work area with people entering the TACS in the morning and leaving again in the evening;

- two peaks with above average density in the morning (around 7.30h) and the evening (around 18h), which seems to correspond with a commuting zone peaking during rush hours.

A geographical representation of this three-way classification of TACS within the vicinity of Brussels (Fig. 7) shows a coherent picture, with most of the territory occupied by residential areas versus fewer working areas in and around city centres and commuting areas usually bridging these two.

7. Conclusions

A comparison of mobile phone data with register-based census data, shows them to be a valid and accurate source to approximate actual present population. Mobile phone data are, moreover, extremely timely, easily computable and not dependent on subjective responses. Their quality in this context will be further enhanced increased by confronting them with detailed other spatiotemporal datasets.

However, mobile phone data also present challenges from a statistical perspective. First of all, the data themselves are new and largely unexplored, and likely to be biased in unknown and possibly unknowable ways (e.g., no one-to-one link between persons and devices, networks only partially covering total population, selectively as to age, gender and other important variables). Other issues are guaranteed data access over time, the size of datasets compared to storage and processing capacity of statistical institutes, information about preprocessing, and maybe most importantly, concerns about privacy and other legal aspects such as data ownership or non-disclosure guarantees towards network operators.

The logical next steps, expanding to other types of mobile data, longer time periods, more spatial and temporal granularity and use of relevant auxiliary data, offer great promise not only for population and migration statistics, but also for domains like mobility and transport, labour mobility and migration, and tourism.

Finally, a crucial condition for long-term success in integrating mobile phone data in official statistics is a mutually beneficial partnership between mobile network operators and statistical institutes. Official statistics obviously has a lot to gain, but for operators as well the necessary investment to process the data for statistical purposes needs to be offset by deeper insight in their own data and access to valuable additional datasets, making it possible to successfully and profitably exploit the mobile phone data.

References

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

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

[3] 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

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

Figure captions

Figure 1: Areas of the ESS Big Data Action Plan and Roadmap 1.0

Figure 2: Voronoi diagrams of macro-TACS with small cells mapped to their overlaying TACS

Figure 3: Counts per TACS (Voronoi polygon) converted by proportional allocation to 1 km² grid counts for a small area in Brussels

Figure 4: Density per km² from 2011 Census (left) and mobile phone data (right)

Figure 5: Pearson correlation between mobile phone and census data for TACS

Figure 6: TACS identified as 1: ‘work’, 2: ‘commuting’ or 3: ‘residential’

Figure 7: Mapping of the cluster results in the vicinity of Brussels and Leuven

Figure 1

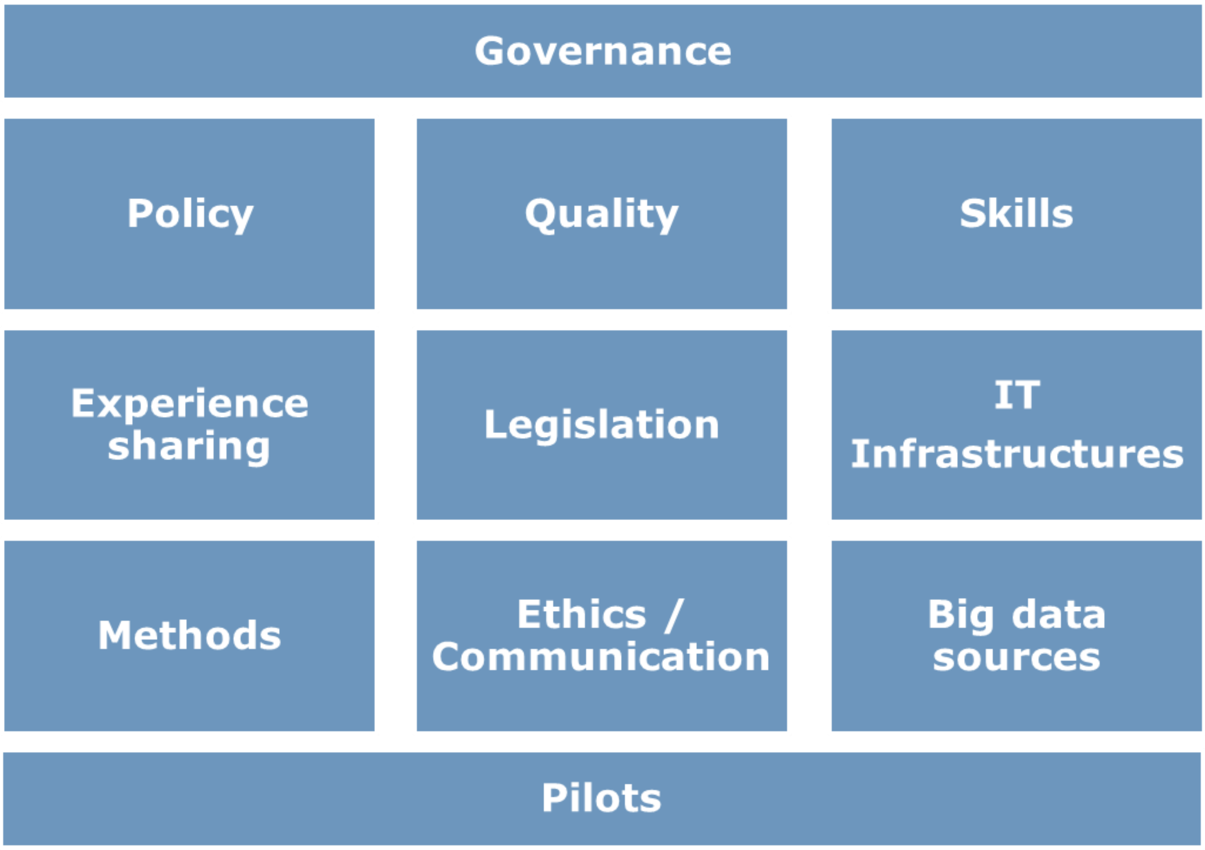

Governance
Policy
Quality
Skills
Experience sharing
Legislation
IT Infrastructures
Methods
Ethics / Communication
Big data sources
Pilots

Figure 2

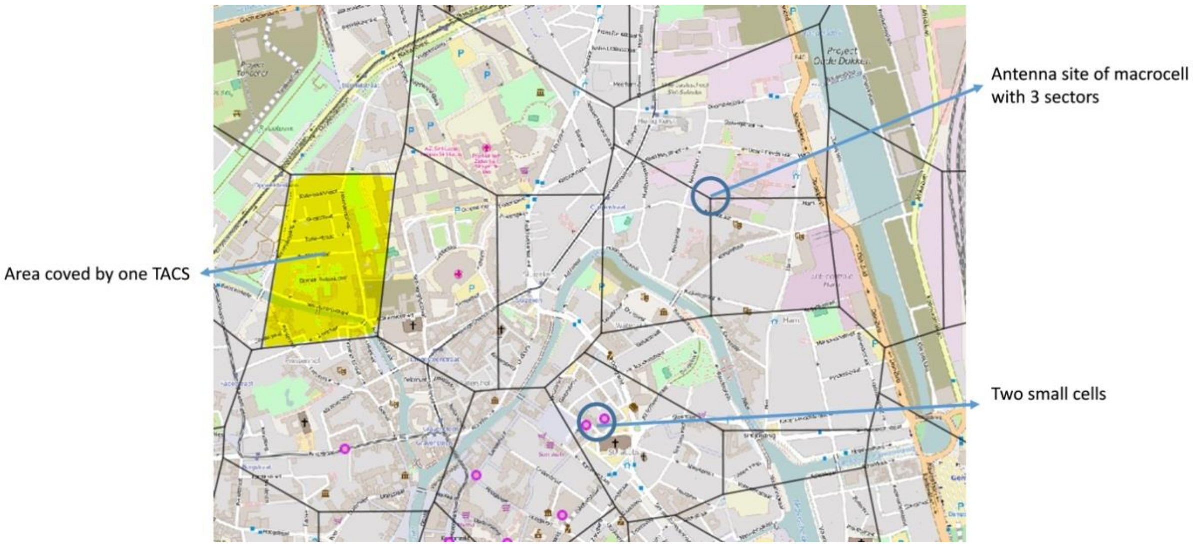

Antenna site of macrocell with 3 sectors
Area coved by one TACS
Two small cells

Figure 3

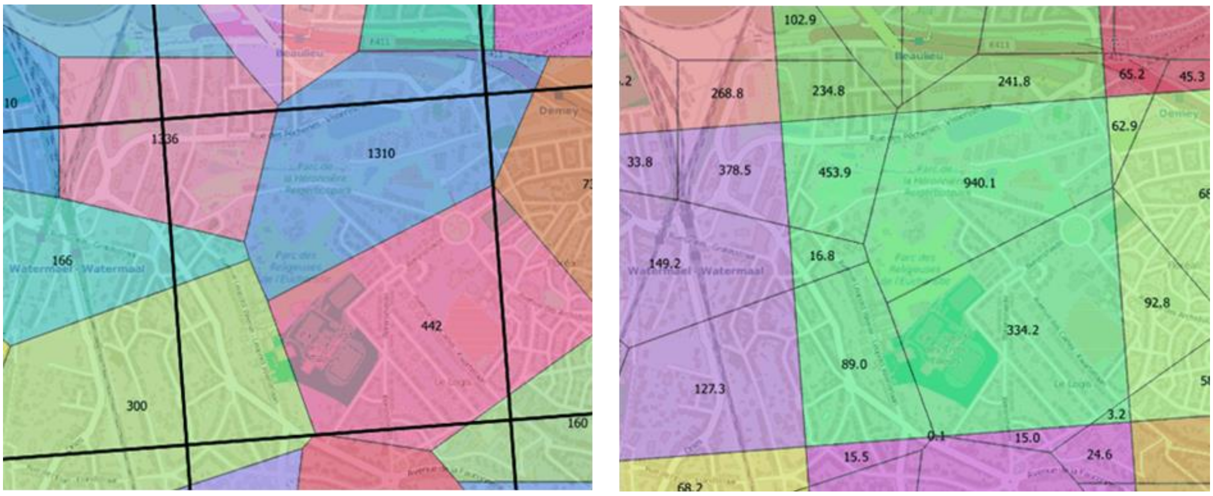

10
1336
1310
Beaulieu
Demey
166
Watermael - Watermael
442
300
160
102.9
.2
268.8
234.8
241.8
65.2
45.3
62.9
33.8
378.5
453.9
940.1
68
149.2
16.8
Watermael - Watermael
92.8
334.2
89.0
127.3
3.2
0.1
15.0
24.6
15.5
68.2

Figure 4

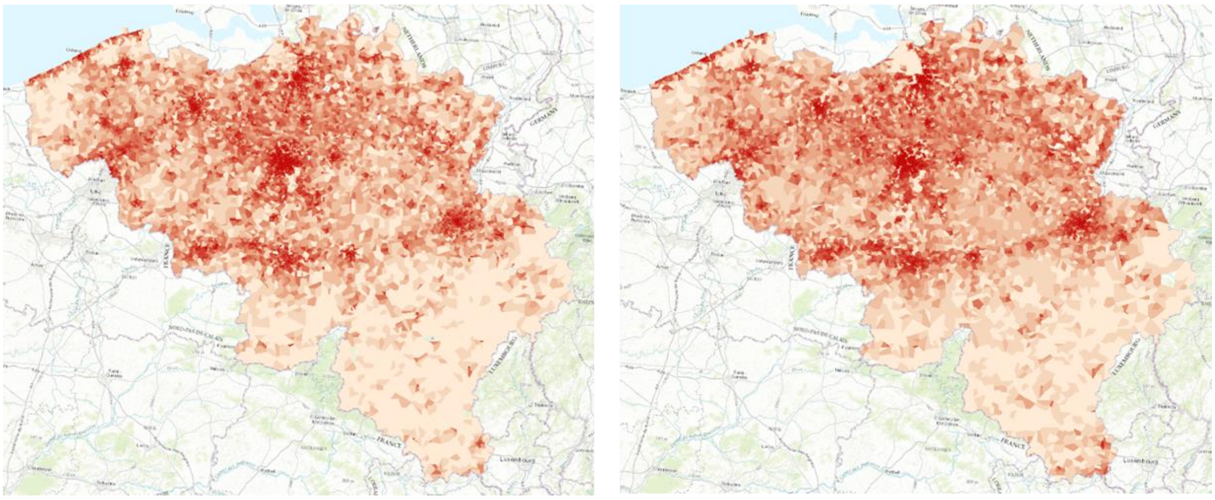

NETHERLANDS
GERMANY
FRANCE
NORD-PAS-DE-CALAIS
LUXEMBOURG
NETHERLANDS
GERMANY
FRANCE
NORD-PAS-DE-CALAIS
LUXEMBOURG

Figure 5

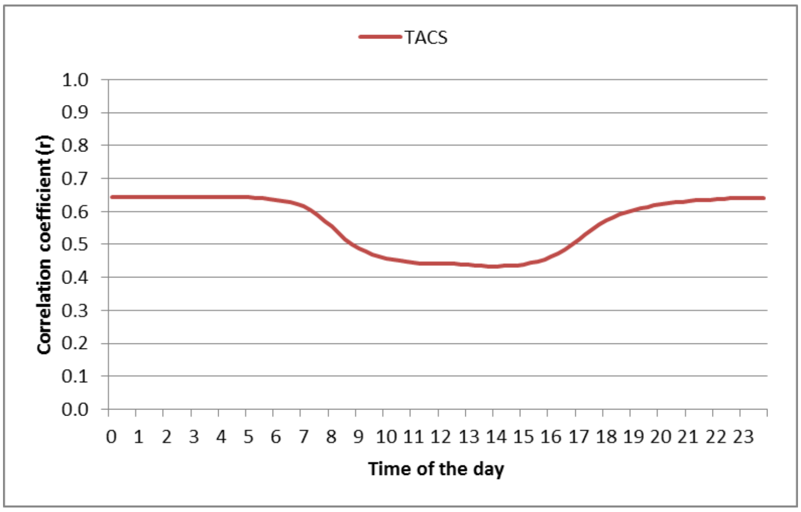

TACS
Correlation coefficient (r)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Time of the day

Figure 6

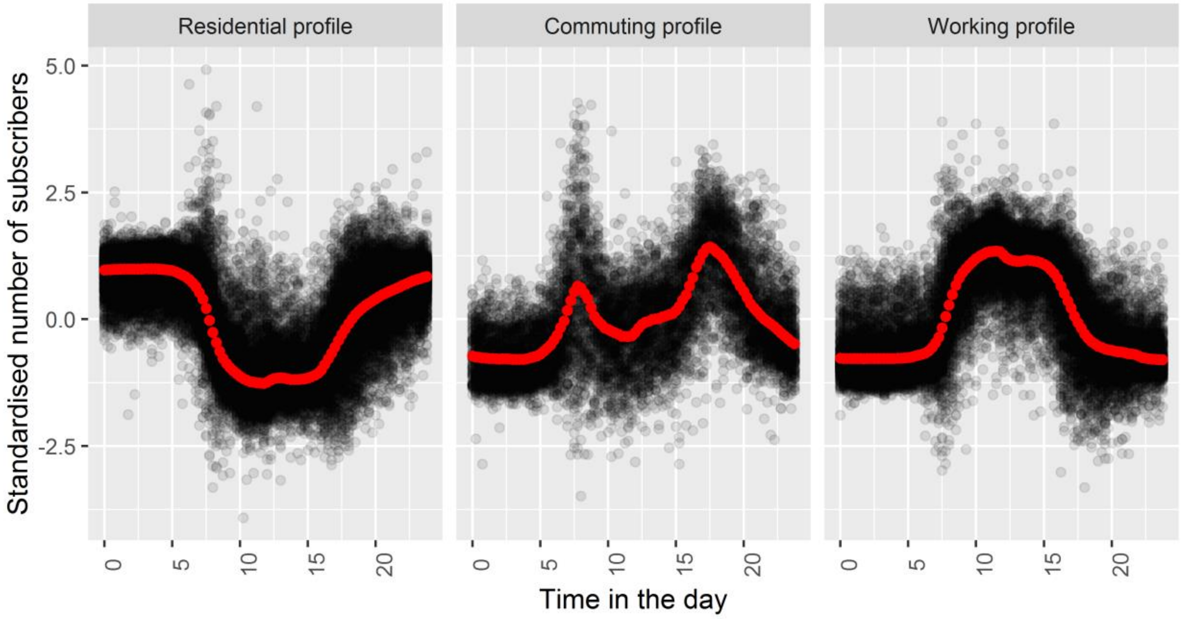
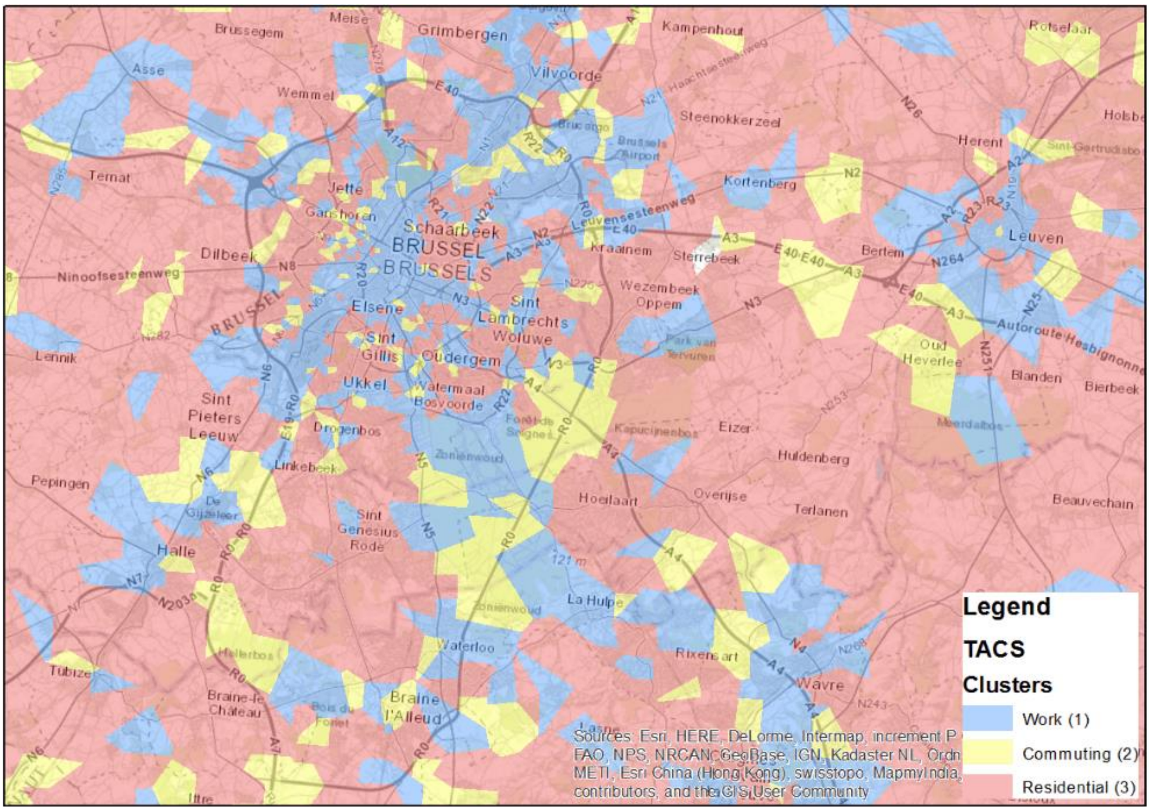

Residential profile
Commuting profile
Working profile
Standardised number of subscribers
5.0
2.5
0.0
-2.5
0
5
10
15
20
Time in the day

Figure 7



Sources: Esri, HERE, DeLorme, Intermap, increment P
FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordn
METI, Esri China (Hong Kong), swisstopo, MapmyIndia,
contributors, and the GIS User Community