Assessing the Quality of Mobile Phone Data as a Source of Statistics

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

[1] *Proximus, Brussels, Belgium; freddy.demeersman@proximus.com*
[2] *Statistics Belgium, Brussels, Belgium, marc.debusschere@economie.fgov.be*
[3] *Eurostat, Luxembourg, Luxembourg; albrecht.wirthmann@ec.europa.eu*

Abstract

Mobile phone data are among the most promising big data sources presently under scrutiny by official statistics, although until now only limited showcase examples have been presented. This paper moves one step further and aims to assess the quality of Belgian mobile phone data (from the major network operator, Proximus) in an analysis conducted jointly by Statistics Belgium, Eurostat and Proximus, focusing on *actual present population*.
The mobile phone data, aggregated for privacy reasons, were tested for internal consistency and compared to results of the 2011 Census**.** They were shown to provide valid and accurate information which may serve as a complement to traditional statistics, but also as an entry point in real time to phenomena inaccessible until now.

Keywords: mobile phone data, big data, official statistics, quality, Belgium.

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[1] 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.

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

2. Data description

1.1 Mobile phone data

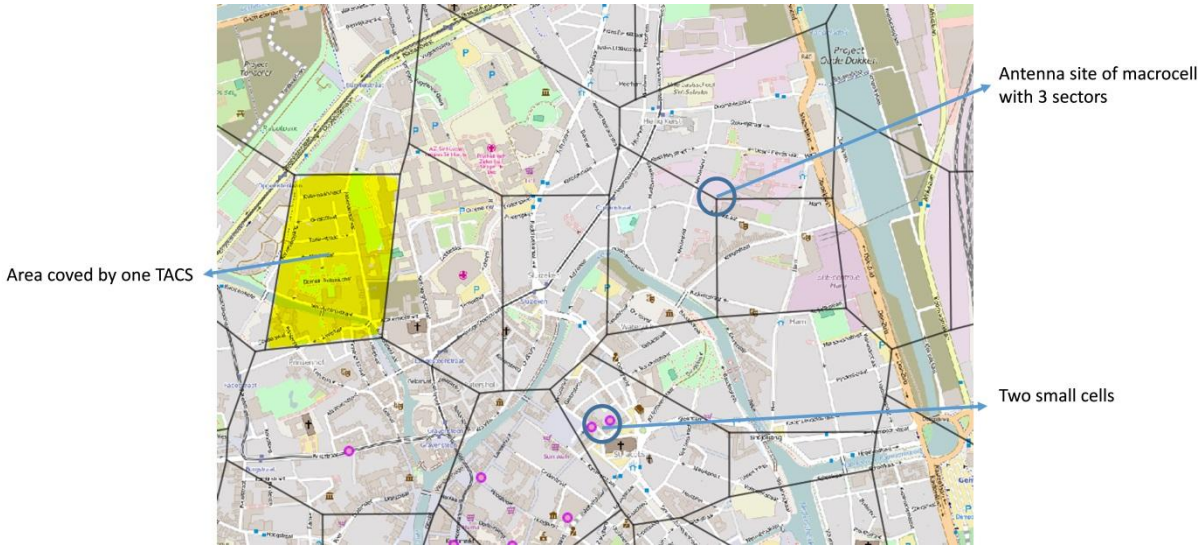
The focus in most studies so far (European Commission, 2014) has been on CDRs, call detail records, used for billing purposes; these reveal the time and location of a mobile phone whenever it is used. Nowadays, however, network probing systems capture all signalling events, including non-billable transactions, 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[2] 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, 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[3] are then added as points and remapped to their "parent"

[2] A location area is a logical grouping of cells (a cell is the area covered by an antenna); a mobile in idle state does not signal to the network when changing from one cell to another, unless the latter is in a new location area.
[3] Microcells (covering a small area, e.g. part of a street), picocells or femtocells (typically for indoor coverage).

Voronoi polygon (see Fig. 1). 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.

Fig.1: *Voronoi diagrams of macro-TACS with small cells mapped to their overlaying TACS.*



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 intrapolation (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 avoid privacy issues, all data used at this stage were aggregates (counts of devices per TACS) which do not contain any information traceable to individuals.

1.2 Census data

The Census[4] 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.

3. Methods

In order to compare the mobile phone and Census datasets, they need to be mapped first onto a common geographical area, either the 1km x 1km Standard European grid or the TACS (Voronoi diagrams representing the mobile phone network).

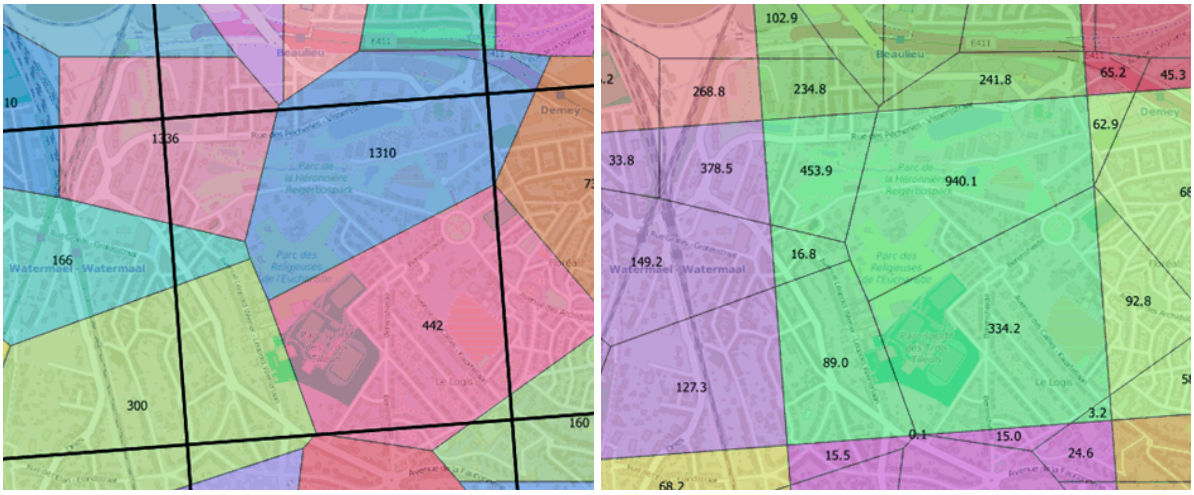
3.1 Mapping mobile phone data to the 1 km² Standard European grid

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. 2 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.

[4] The Census 2011 was compiled entirely from administrative sources (e.g. population, social security, tax, dwellings, enterprises, educational level registers).

Fig. 2: *Counts per TACS (Voronoi polygon) converted by proportional allocation to 1 km² grid counts for a small area in Brussels.*



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. land cover).

3.2 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 TACS.

3.3 Statistical analysis of the mobile phone dataset: cluster analysis

Cluster analysis was performed on the mobile phone datasets for both days separately and taken together, summed by TACs and 1 km² grids, at 15 minute and one hour intervals. Absolute numbers of phone counts were normalised to a mean of 0 and a standard deviation of 1 (procedure “scale” in R, version 3.2.3, base package) before KMEANS clustering (stats

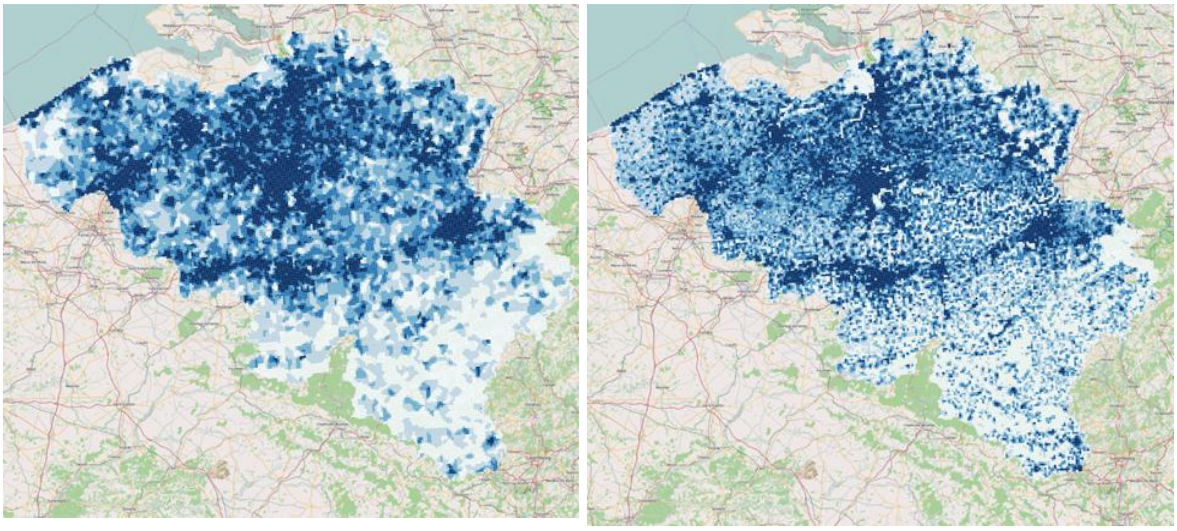
package, Hartigan and Wong algorithm with random centres). Optimal numbers of clusters were determined using the within-groups sum of squares (SSW in R), finally yielding three clusters for the Thursday data and four for the Sunday dataset. In order to verify their plausibility, results were displayed 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 hourly patterns analysed.

4. Results

4.1 Estimating population density from mobile phone versus Census 2011 data

The two maps below visually represent population density per km²; the one on the left is based on Voronoi counts of mobile devices on Thursday 8 October 2015 at 4 a.m. converted to 1 km² grids, the right one on the 2011 Census derived from administrative records in the Population register. The Pearson correlation between these two datasets is 0.85.

Fig. 3: *Population density per km² based on mobile phone data (left) and 2011 Census (right).*

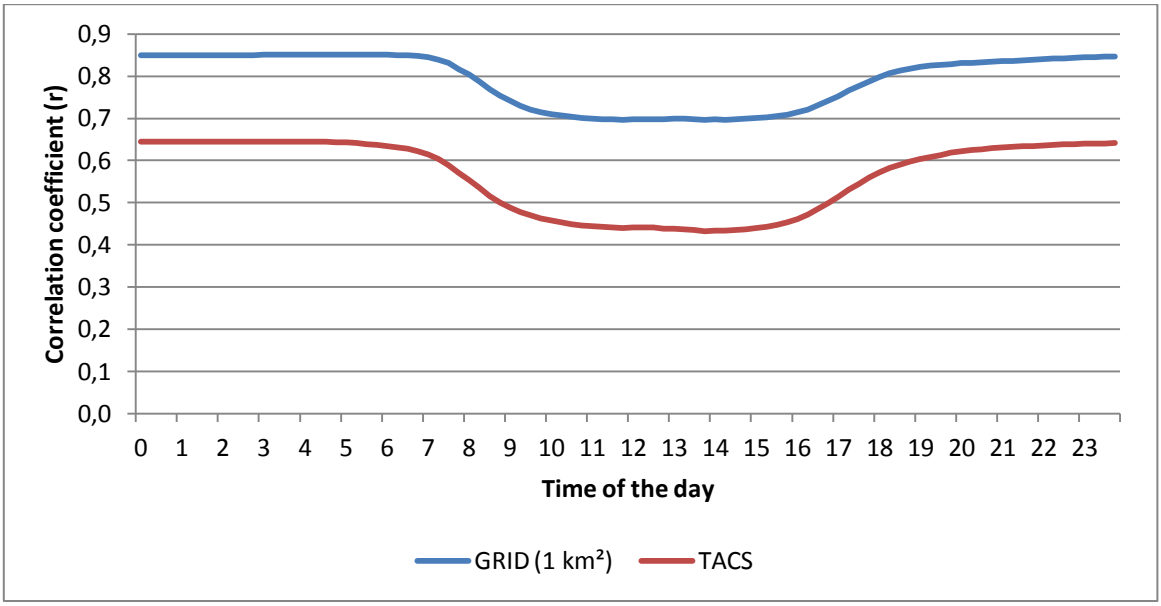


The similarity between both maps above is striking, but one problem already is apparent: in less densely populated areas with large TACS of several square kilometres (e.g. Ardennes in the southeast) mobile phone data seem less precise. Further detailed analysis shows other

mismatches, e.g. in the port of Antwerp, the Zaventem airport, big park areas in Brussels, … These problems can be addressed by more detailed investigation (see 6.).

Fig. 4 shows the correlations between both datasets at 15' intervals on Thursday 8 October 2015 (96 registrations), either based on grids of 1km² (blue line, on top) or on TACS (red line). Their pattern is quite similar, but at different levels. Not unexpectedly, correlations are highest at night and they display a regular 24-hour pattern, with a rather rapid decrease in the morning and a more gradual increase in the evening. What is surprising, however, is that correlations are markedly higher for the 1 km² grids, up to 0.85 at night, while for TACS they are only in the region of 0.65. As it is not clear at this moment what is causing this difference, a prime candidate for further study (see 6.) is testing hypotheses on the optimal area type and size for combining mobile phone and statistical data.

Fig. 4: *Pearson correlation between mobile phone and Census data every 15 minutes on Thursday , for1 km² grids (blue, above) and TACS (red. below).*



4.2 Assessing validity and accuracy of mobile phone data for estimating population density

In order to identify problem TACS in the mobile phone data requiring further analysis, Voronoi counts at night were compared with Census estimates of resident population for each

TACS polygon. In a contingency table (Fig. 5) between density deciles from both sources a strong concentration of frequencies appears around the diagonal of the matrix, and a calculation of the distance between mobile phone and Census deciles shows that for over 35% of TACS both sources perfectly agree, while in nine out of ten cases the distance is 2 deciles or less. From this it can be concluded that both datasets are rather close and valid approximations of population density. However, as Proximus market share is high but far from complete and probably unevenly distributed, large local discrepancies can be observed (which might be resolved through additional information).

Fig. 5: *Number of TACS per decile in the mobile phone (y axis) and Census dataset (x axis).*

# 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

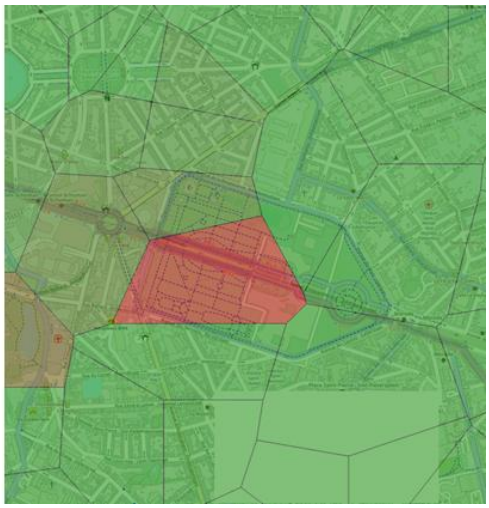
Mapping the distances in detail reveals interesting discrepancies which invite further investigation. Some examples are shown in Fig. 6 (with light green indicating agreement and shifts towards red the opposite):

- on the left map showing the airport at Zaventem, the red polygons correspond with the passenger terminal where no one lives but mobile devices are detected even in the middle of the night;
- in the right map the red polygon mostly but not totally covers the Jubelpark/Parque du Cinquantenaire in Brussels; although obviously uninhabited, mobile phones are nevertheless spotted at night; these could either be ‘spill-over’ from surrounding multi-

storey apartment buildings (Voronoi polygons never exactly coincide with cell footprints) or even from night traffic on a motorway going underground in the park.

These examples suggest that adding data on local conditions may significantly increase the overall validity and accuracy of the mobile phone dataset.

Fig. 6: *Difference in density deciles for Zaventem airport (left) and Jubelpark/Parc du Cinquantenaire, Brussels (right) – green denotes agreement, red difference.*



4.3 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 a characteristic and meaningful temporal pattern.

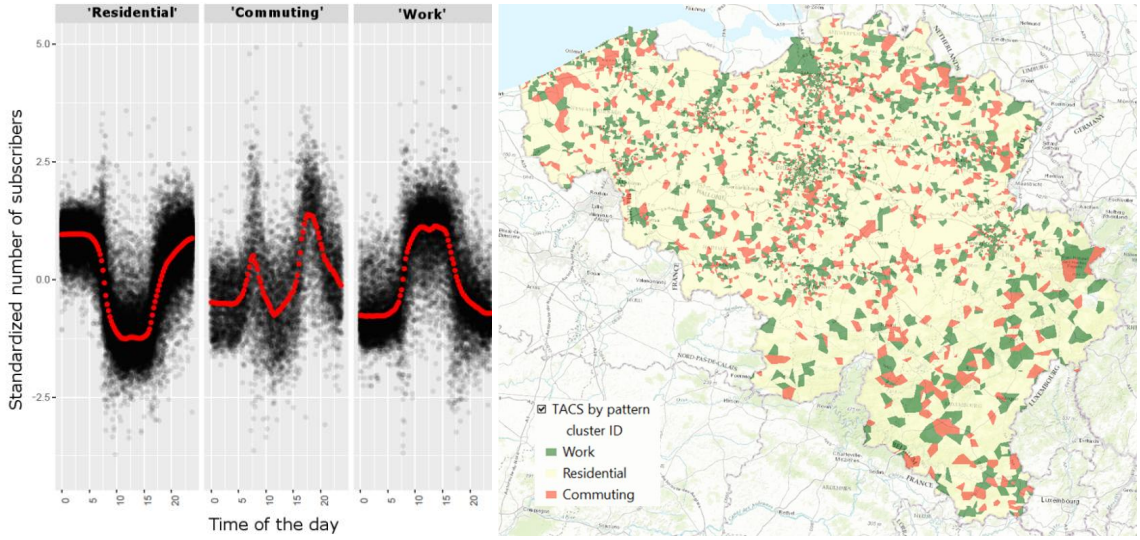
Looking at the means of the normalised number of phones during the day shows three patterns on a Thursday which account for most of the reduction in the within-groups sum of squares (SSW) and which can be interpreted in a meaningful way (see Fig. 7, left):

- 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 (cluster 2);
- 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 (cluster 1);

- two peaks, in the morning (around 7.30 am) and the evening (around 6 pm), which seems to correspond with a *commuting zone* peaking during rush hours.

A geographical representation of this three-way classification of TACS (Fig. 7, right) shows a coherent and unsurprising picture, with most of the territory occupied by residential areas (cluster 2) versus a few working areas (cluster 1) and commuting areas (cluster 3) usually bridging these two.

Fig. 7: W*eekday TACS identified as 'work', 'residential' or 'commuting', with mapping.*



The case of the Sunday is more complex, with a larger number of clusters which are less obvious and more difficult to interpret. Understanding this pattern will require further analysis.

A cluster analysis using 1 km² grid data shows essentially similar results, whereas using hourly instead of 15-minute data does not affect the patterns either.

5. Discussion

5.1 Mobile phone data as a valid source to assess population density (validity)

The theoretical assumption that people's residence is where their mobile phone spends the night is clearly vindicated by the consistently high correlation during the night of about 0.85

between mobile phone counts and Census population density which then markedly decreases during the day. This is confirmed by the striking similarity in geographical mapping of mobile phones at night and register-based population densities, even though both sources suffer from inevitable gaps and inaccuracies. Obvious ones on the side of mobile phone data: not all people have a mobile device and some people have several, data come from only one mobile network operator with a high but still limited and geographically variable market share, people and their phones do not all spend every night at their place of residence (e.g. tourist trip, in hospital, working night shifts, …). Population registers, on the other hand, have a time lag or may be incomplete because some residents are not registered or usually reside at another place than their official domicile.

However, both have also unique advantages: registers are fairly complete and hence representative to a high degree, while mobile phone data are recording actual situations in real time, not affected by non-response or non-registration bias. Combining the unique advantages of both sources should yield statistics at the same time more valid, accurate and timely than either separately. A statistical process could be established where valid and accurate 'flash' estimates of resident population based on mobile phone data and auxiliary datasets, are regularly validated and if necessary corrected with data from the population register.

5.2 Correlation between population density based on mobile phone versus Census data (accuracy)

The high correlation coefficients, around 0.85 for the 1 km² grid data at night (Fig. 4), indicate that both datasets accurately capture the underlying concept of actual present population. Many of the discrepancies which were found can be explained through the use of auxiliary datasets (see 6.). Taking these into account may increase correlations even further.

The cluster analysis shows that small geographical areas can be characterised on the basis of the changing numbers of mobile devices they contain, thus confirming the validity and accuracy of the mobile phone dataset.

5.3 How can the value of mobile phone data be further enhanced?

Any dataset which is spatially and temporally organised in a way which overlaps with the mobile phone dataset, can be used to better understand it and to identify variation which may then be filtered out. Examples are meteorological data, calendars (holidays, events), land use (e.g. roads, railways, train stations) and similar geocoded datasets, information on incidents at a particular location and time, …

A second potential improvement concerns the optimal spatial and temporal granularity of the mobile phone dataset. For the present study devices were counted every 15' for two days (2 x 96 time points) for about 11,000 TACS (covering the Belgian territory of 30.528 km²). More frequent and even continuous recording is possible, for smaller areas and even for individual devices (raising privacy issues which of course would need to be addressed first). More temporal or spatial granularity and longer observation periods will all inflate the size of the dataset, possibly beyond available resources. In a particular statistical context not all detail which is possible might be needed, so optimal sizing should be investigated. For instance, for the estimation of actual population density a limited number of observations at night (e.g. every two hours) would do. To determine fluctuations in actual present people in a given area the interval of 15' used in the present study is probably sufficient. Microstudies at precise locations (e.g. measuring traffic flow patterns) would however require more frequent observations and longer periods.

6. Future research

The present paper has a deliberately limited scope: exploring a totally new type of data and testing validity and accuracy with rather modest research questions. But even at this stage it is obvious there is considerable potential for further research, some of which already underway. Two approaches can be distinguished: optimising the present analysis; and, as a next step, raising new statistical questions and identifying the mobile phone datasets required to answer them.

6.1 Present dataset

Study of the mobile phone dataset has generated numerous new research questions; some are already under study and will be reported in follow-up papers. A non-exhaustive list:

- Optimal temporal resolution (period and frequency) of mobile phone data to assess present population.
- Best geographical unit for connecting mobile phone to statistical data: TACS or regular grids (1 km² or even smaller ones made possible by advances in mobile phone technology).
- Optimal size / resolution of spatial units dependent on the phenomena under investigation or the specific statistical results to be achieved.
- Feasibility to connect mobile datasets at a more basic level, via geocoordinates, i.e. precisely located mobile devices and geocoded statistical data.
- Systematic identification and resolution of problem areas (where datasets fail to match) through detailed local knowledge (cfr. examples in 4.2).
- Adding auxiliary spatiotemporal datasets to reduce the unexplained variation; some examples: land use, degree of urbanisation, limits of built-up area, traffic infrastructure (roads, railways, train stations, airports, etc.).

6.2 New data requests

Many other statistical questions may be answered by similar datasets, extended or modified for a specific purpose:

- Labour mobility patterns may be detected from comparing TACS types with Census data on active versus non-active population.
- Commuting can also be studied by comparing weekdays, and combining this with possibly impacting factors (weather, season, specific incidents or events, …).
- Cross-border commuting, labour migration, international tourism etc. can be investigated by combining counts of foreign mobile devices, roaming data (in and out) and data from

operators of neighbouring countries (e.g. Luxembourg), from which a more comprehensive European picture might emerge.

- If mobile devices are tracked individually and linked with other data via an individual key rather than at an aggregated geographical level as was done here, it becomes possible to measure movements in time and space and to assign persons' most likely living place, workplace and other elements of the 'usual environment'; these are essential, at a later stage, to produce detailed statistics on commuting, transport preferences, labour migration, migration , tourism, and possibly even on time use and lifestyle - but all privacy issues have to be settled first!
- Finally, there is still room for improving the accuracy in positioning mobile devices through triangulation techniques, network densification, use of GPS data, coupling with other sources such as WiFi location data etc …. This will increase precision and make it possible to answer questions previously unanswerable, but at the same time it increases the size and complexity of the datasets.

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 by integrating other detailed 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 pre-processing, 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.

8. References

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