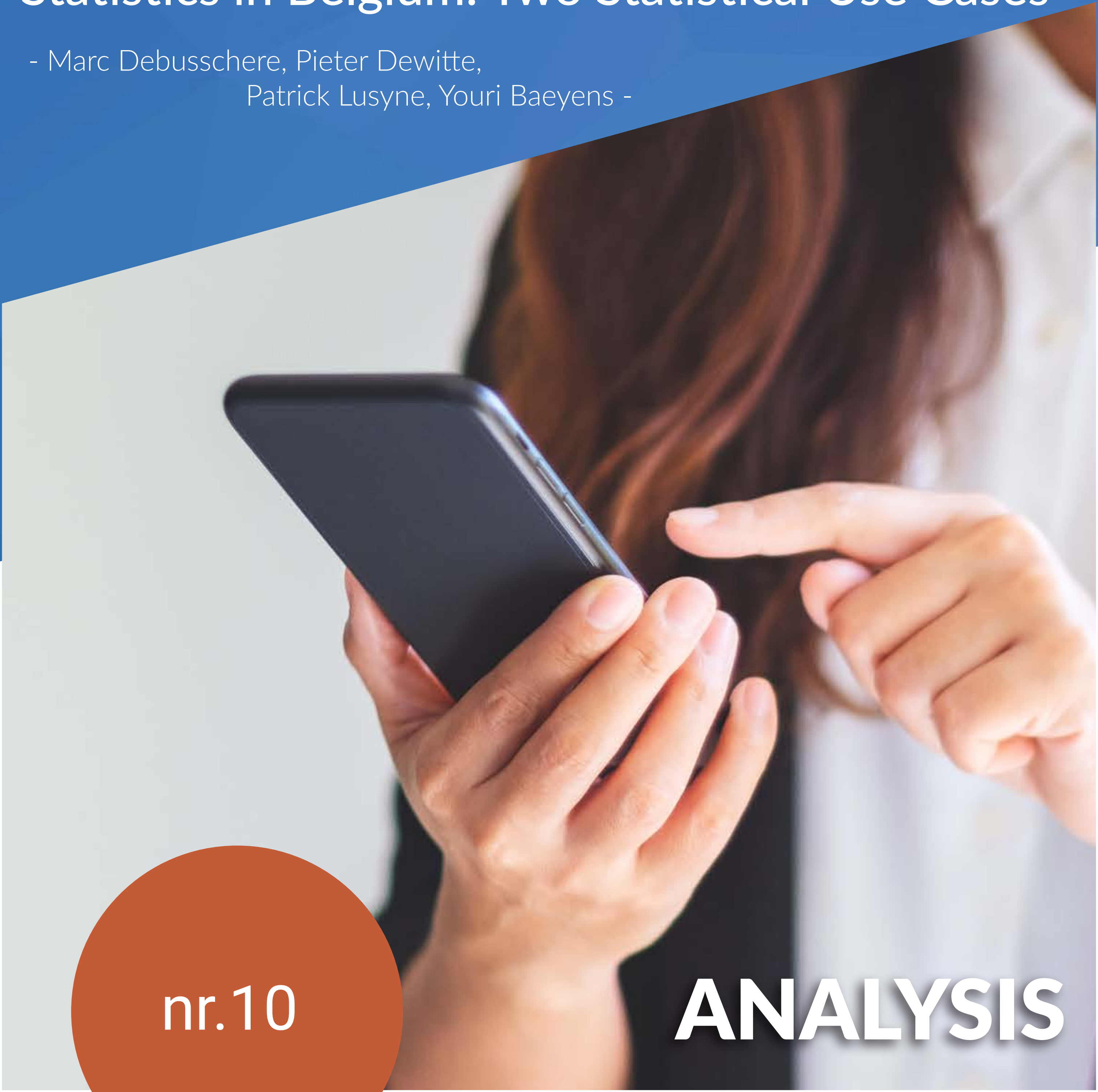
STATBEL
Belgium in figures

# Mobile Phone Data as a Source for Commuting Statistics in Belgium: Two Statistical Use Cases

- Marc Debusschere, Pieter Dewitte,
Patrick Lusyne, Youri Baeyens -



nr.10

ANALYSIS

06.2020

economie

FPS Economy, S.M.E.s, Self-employed and Energy

.be

# Mobile Phone Data as a Source for Commuting Statistics in Belgium: Two Statistical Use Cases

Marc Debusschere, Pieter Dewitte, Patrick Lusyne, Youri Baeyens[1]

[1] Statisticians at Statbel (Directorate-general Statistics – Statistics Belgium)

## ABSTRACT

Statbel, the Belgian statistical bureau, has had various contacts since the end of 2015 with Belgium's three mobile network operators, in its attempts to use mobile phone data for official statistics. One important outcome is a set of concrete potential use cases for compiling commuting statistics based on mobile phone data. These use cases consist of a detailed description of the data needed, the operations to be performed and the expected statistical results. These applications of course fully respect the Statbel charter on big data and privacy[2].

The first statistical use case aims at measuring the patterns of presence and absence at the living place and workplace over a whole year. Complementing this static approach, the second use case seeks to dynamically connect the living place and workplace, in order to create a « Living Place/Workplace Matrix » which will be considerably more granular and timely than the ones produced at present on the basis of administrative data.

Finally, and by way of conclusion, the conditions under which these statistical use cases may give rise to experimental statistics and ultimately to new statistical production lines are discussed.

[2] See https://statbel.fgov.be/en/about-statbel/privacy/statbel-big-data-and-privacy: Statbel, big data and privacy.

# CONTENT

# 1. MOBILE PHONE DATA AND OFFICIAL STATISTICS

From their start in the early 19th century, official statistics were almost exclusively based on surveying citizens and enterprises. The past twenty years have seen an increasing reliance on administrative data, motivated by a concern to reduce the costs and the burden to respondents,. Recently, however, a 'third data revolution' has commenced: exploiting the immense and largely unstructured deluge of 'big data' continuously generated by our society through sensors and cameras, satellites, machine-to-machine communication, e-business, electronic payments and withdrawals, various activities on the internet, social media, ... in order to produce most extant statistics in a better and more rapid way, but also to track phenomena which until now could not be properly measured.

The *Scheveningen Memorandum on Big Data and Official Statistics*[3,], drafted by DGINS[4] and adopted by the ESSC[5] on 27 September 2013, can be considered the official starting point for the integration of big data within the European Statistical System. It took concrete shape in ESSnet Big Data I and II[6,], an extensive collaboration platform to investigate the usability of various types of big data, including mobile phone data, and to test their uses through pilot studies.

Among the various types of big data, mobile phone data look particularly promising as a potential source for statistics, especially in the domains of population, migration, tourism and mobility. Using them is expected to result in faster and even instant statistics, much more granular, with almost perfect coverage, free from response bias, at a lower cost and without the need to inconvenience citizens and businesses. Additionally, they might provide direct access to phenomena defying measurement until now (such as actual present versus registered population or detailed commuting patterns in function of the weekday, weather conditions, etc.). Numerous pilot studies in Europe and in the rest of the world have demonstrated convincingly they are a viable alternative to the more traditional data sources used by national statistical institutes (De Meersman et al., 2016, reference list). The ESSnet Big Data workpackage on mobile network data has considerably expanded our knowledge about the characteristics of mobile phone data and their potential and limitations, and it has laid the foundations for a statistical methodology to exploit them. But real progress is still strongly hampered by the extremely limited access to mobile phone data for statistical production, and even for the testing of methods and approaches.

# 2. MOBILE PHONE DATA AND STATBEL

Belgium has three mobile network operators: Proximus (former incumbent), Telenet/Base and Orange Belgium, with respective 2018 market shares of approximately 41%, 30% and 24.6%[7]. All three have been contacted by Statbel with proposals to jointly explore their data and to combine them with statistical data both for Statbel's statistical use cases and for the operators' commercial ones.

Exploratory talks with Telenet/Base in 2015 and 2018 have not yet resulted in concrete plans or projects in spite of the interest expressed on both occasions.

A first contact with Orange in 2016 ended with a similar lack of direct results, but in 2018 a proposal by MIT (Boston, US) to combine mobile network signalling data and Statbel fiscal income data to study social segregation in Brussels provided a more concrete collaboration opportunity. Because of staff changes in both MIT and Orange this had to be put on hold and there seems to be no prospect on finalising in the short term.

With Proximus and Eurostat, the European statistical office, a common project was started in December 2015 to explore and analyse aggregated Proximus network signalling data for their information value and potential uses, resulting in 10 publications[8] and a joint press conference in September 2016 to present outcomes to the general public. Within the context of the 2018 MIT project mentioned above, Proximus has created an aggregated mobile network signalling dataset to be

[3] See https://ec.europa.eu/eurostat/documents/42577/43315/Scheveningen-memorandum-27-09-13 (PDF download).

[4] Yearly conference of the directors general of the national statistical institutes of the European Statistical System, see https://ec.europa.eu/eurostat/web/ess/about-us/ess-gov-bodies/dgins.

[5] European Statistical System Committee. See https://ec.europa.eu/eurostat/web/ess/about-us/ess-gov-bodies/essc.

[6] See https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/.

[7] See https://www.internetproviders.be/overzicht-mobiel-internet-op-belgische-markt/, retrieved 27 September 2019 (Dutch); totals do not add up to 100% because of minor market shares by operators without own network.

[8] See De Meersman e.a (2016) for an overview of the project.

linked to a similarly aggregated custom-made Statbel fiscal income dataset, using a novel and very detailed method which however does no individual linking, thus excluding any privacy or confidentiality issues.

A major obstacle to using big data such as mobile phone data for official statistics, not just in Belgium but also in the other EU Member States and even globally, is obtaining access to privately held data when owners see possible drawbacks but no obvious advantage in opening them up. Mobile network operators are particularly wary of providing data and even of exploiting the data themselves because of the real or perceived risk of a privacy breach which could cause public outrage. Furthermore, competition issues are an additional concern: operators are very wary, in spite of all guarantees offered, of possible leaking of strategic network signalling data to their competitors. This reluctance can only be overcome if risks can be shown to be minimal or even non-existent, and if commercial or statistical uses are of sufficiently high value.

A paper presented at the October 2018 DGINS Conference in Bucharest (Debusschere, Waeyaert, Van Loon, 2018) identifies four key factors required for access by official statistics to big data such as mobile phone data:

1) a clear and detailed business case;

2) high-level engagement and active support;

3) the fostering of trust by absolutely guaranteeing confidentiality and privacy;

4) specific legislation.

The methodology proposed by Statbel for the MIT Project is highly relevant for the third issue, of strict guarantees for data security, thanks to two innovative approaches:

- individual data, both mobile phone and statistical data, need not leave the respective owner's datawarehouse, ensuring privacy and confidentiality control and ownership;
- mobile phone and statistical data are not linked individually but at a highly detailed yet anonymous geographical level: statistical variables are aggregated according to mobile network antenna areas (using Voronoi shapefiles provided by the mobile network operator) so they can be combined seamlessly with mobile phone data aggregates for the same areas.

It should be noted this methodology can also be applied to commercial use cases of mobile network operators, considerably enriching mobile phone data with made-to-measure statistical context information. This might provide an incentive for operators to collaborate with national statistical institutes, for mutual benefit.

## 3. STATISTICAL USE CASES

The present article focuses on the first of the four critical factors mentioned in the DGINS paper, developing clear and detailed use cases, more particularly in the domain of mobility and labour commuting. At present Statbel compiles statistics on commuting which are either limited in geographical detail or published with a considerable time lag, or both. The 'Living Place/Workplace Matrix' produced in the context of the decennial Census is based on administrative data, at the 'statistical sector' level (corresponding to 'quarters' of municipalities), obtained from Belgium's persons and social security registers. The quarterly Labour Force Survey is a sample survey also containing questions on commuting, but due to limits in the sample size no results at a more detailed level than Belgium's three regions are disseminated.

At the first stages of investigating the potential of mobile phone data for official statistics, numerous pilot studies were conducted with the focus on mobile phone data characteristics, their statistical validity and reliability, quality and methodological issues. The results, amongst others from the joint Statbel-Eurostat-Proximus project, look quite promising: it seems highly probable that mobile phone data can be used to partially or wholly replace, complement, enrich or validate official statistics, in domains such as population and migration, mobility and commuting, transport, tourism, … The logical next step is then the elaboration of statistical use cases, by identifying an existing or devising a new statistical product, selecting the mobile phone data needed as input and specifying the operations to be performed on them to arrive at official statistics based at least partially on mobile phone data.

The key element of a statistical use case is a concrete mobile phone data request which is sufficiently detailed and realistic in terms of the complexity and runtime of the query, the size of the resulting dataset and the handling of privacy issues.

Furthermore, to be useful for official statistics in the long term, it should be sustainable, i.e. repeatable with the agreed frequency without additional effort, and thus able to support regular statistical production.

The two statistical use cases presented below describe in detail which mobile phone data, aggregates and calculations would be needed to serve as a basis for statistics on commuting, the pattern of traveling from a person's living place to the workplace and back. These can of course then serve secondary purposes such as determining the transport modes (for instance by combining these results with land use or building register data), calculating traveling times, assessing the environmental impact, … Another possible use is 'social geography': determining the mainly residential, working or communing areas of a territory, or defining 'influence zones' of urban areas, or urban sprawl from a work commuting point of view.

# 4. IMPROVING THE STATISTICS ON LIVING PLACE AND WORKPLACE

## 4.1. Backgrounds

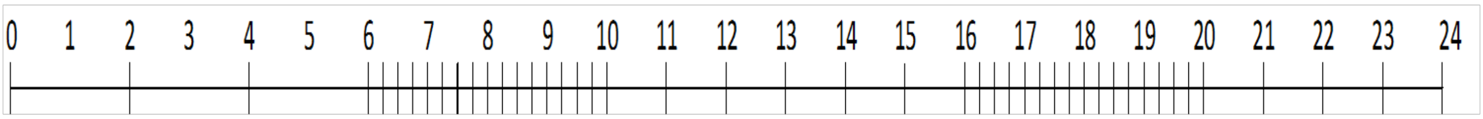
Mobile phone data have been used already, in the Statbel-Eurostat-Proximus project, for replicating and hence implicitly validating the Belgian Census results on population as to residence, based on the Belgian Population Register (De Meersman e.a, 2016).The promising result, and more specifically the high 0.85 correlation between both data sources, makes it now possible to more precisely identify the specific mobile phone data which can make population statistics with regard to living place and workplace more accurate.

The Population Register entry 'domicile', the officially register place of residence, is a good although not perfect approximation of the living place, the actual or habitual dwelling place. Mobile phone data offer an alternative estimate of the latter variable, also with its shortcomings, albeit different ones. Therefore both sources are complementary and flaws in either might be compensated by information derived from the other source. By combining both and, as the case may be, adding still other datasets (e.g., CORINE Land cover[9]), it seems very likely that the living place can be assessed more accurately.

For the workplace a similar approach can be developed, through combining data from the Crossroads Bank for Social Security (CBSS) and the Crossroads Bank for Enterprises (CBE) with mobile phone data.

## 4.2. Data request

The number of mobile devices for each Voronoi-cell (N=11.000), extrapolated from the local Proximus market share to the total, measured at the points in time (N=45) on an 24-hour timeline, for each day of the past 12 months (N=365); accompanied by the topology (shapefile**)** of the Voronoi cells on the first day, and insofar as these are undergoing changes during the 12 months (in order to take into account any change occurring while converting to other geographical mappings such as the km² grid).



This timeline is composed of the following parts, with varying measurement frequencies:

- Between 00:00 and 06:00 (quiet period) every 2 hours
- Between 06:00 and 10:00 (morning rush) every 15 minutes
- Between 10:00 and 16:000 ('normal' office hours) every hour
- Between 16:00 and 20:00 (evening rush) every 15 minutes

[9] See https://www.eea.europa.eu/publications/COR0-landcover for more information.

- Between 20:00 and 24:00 (evening rest) every hour

Apart from the analysis of the division over the day of presence at a certain location, the availability of a whole year also permits a detailed assessment of the differential effects of working day versus Saturday versus Sunday, different working days, bank holidays, holiday periods, one-time events (e.g. disaster, terrorist attack, strike), weather conditions, seasonal influences, ...

The queried dataset, from a total population of about 400 billion mobile phone localisation records, would contain somewhat over 180 million records and for regular statistical production the query would need to be run annually (for instance in January for the past calendar year).

Using 96 points in time, so every 15 minutes, instead of the proposed 45 would double the size of the dataset, but may simplify the query parameters.

# 5. LIVING PLACE/WORKPLACE MATRIX (ORIGIN & DESTINATION OF COMMUTING)

## 5.1. Backgrounds

The dataset mentioned above provides a good global view of on the one hand the usual living place and on the other hand the workplace of the population, but not about the relationship between both and over the journeys from the one to the other and back. Voronoi area totals are not adequate for this purpose, more detail is needed. A first possibility is tracking mobile devices individually over time and location and aggregating the results. There is, however, another solution which avoids any privacy issues raised by individual tracking: assigning the most probable living place and working place for each mobile device, via an algorithm that checks where the device is located most frequently at certain periods during the day, and then aggregating these observations to a dataset.

This mobile phone dataset could serve, in the context of the Census, to validate the « Living Place/Workplace Matrix » compiled at present by Statistics Belgium on the basis of administrative data, but it could also significantly increase its timeliness, accuracy and level of detail in time and space.

The Living Place/Workplace Matrix is being used, amongst others, by the social partners (employers and unions) represented in the Central Economic Council and by the Federal Planning Bureau for their economic models.

## 5.2. Data request

The requested dataset should make it possible to develop an algorithm for determining the living place and workplace. Depending on test results, the size of the dataset may have to be modulated for other combinations of points in time or periods, without however significantly affecting the data volume.

Concretely, the results of a calculation on the basis of the data for the latest month of October are requested; October being considered to be the most ‘normal’ ‘typical’ month as to length and absence of bank holidays and, as a rule, school holidays. Below two possible algorithms are presented, with a similar output: a cross table for the number of mobile devices of approximately 11,000 by 11,000 Proximus Voronoi areas. This corresponds to a theoretical maximum of 121,000,000 records, but most of these will probably have as value 0. A record of the proposed dataset consists then of a living place cell, a workplace cell, the  number of mobile devices and the topology (shapefile) of both Voronoi areas.

### 5.2.1. Scenario 1

- Step 1: determine, for each mobile device, the Voronoi area which can be considered the most probable living place of the person using the device. For each device and for each day of October the Voronoi area in which the device is located at 04:00 is identified. The most probable living place then is the Voronoi area in which a device most often was observed at 04:00 during the month of October (or, in other words, the modus of the distribution of Voronoi areas measured at 04:00 every day in October, for a given device).
- Step 2: determine, for each mobile device, the Voronoi area which can be considered the most probable workplace of the person using the device. For every device the Voronoi area is determined in which it was observed at 10:00,

11:00, 14:00 and 15:00 on each working day (Monday to Friday, both included) in October. The most probable workplace is then the Voronoi area in which the device was observed most frequently on these points in time (or, in other words, the modus of the distribution of the Voronoi areas measured at 10, 11,14 and 15 o'clock of each working day in October for a given device).

- Step 3: After having defined the most probable living place and workplace for each mobile device, these are aggregated into a cross table living place X workplace of approximately 11.000 X 11.000 cells, or a total of about 121,000,000 values (of which a majority is probably zero, because no single mobile device 'living' in cell x is 'working' in cell y).

| | | Most probable living place October YYYY | | | | | |
|---|---|---|---|---|---|---|---|
| | | Cell 1 | Cell 2 | Cell 3 | Cell 3 | Cell 4 | Cell 5 |
| Most probable living place October YYYY | Cell 1 | | | | | | |
| | Cell 2 | | | | | | |
| | Cell 3 | | | | | | |
| | Cell 4 | | | | | | |
| | Cell 5 | | | | | | |

### 5.2.2. Scenario 2

In scenario 2 the living place is determined in the same way as in scenario 1, but whereas the most probable workplace is always determined in scenario 1 for a given mobile device, scenario 2 excludes those cases where the mobile device is found too infrequently at one location.

Suppose we have *n* mobile devices and *m* Voronoi areas. Let us note the number of times (the *frequency*) that the *i*-th device is found in the *j*-th Voronoi area (measured at 10, 11, 14 and 15 o'clock on each weekday of October) as $f_{ij}$ . Let us note the index number of the Voronoi area where the *i*-th mobile device appeared most frequently as $Mod_i$ (the modus of the Voronoi cells where the *i*-th mobile device appeared). Or, expressed as a formula:

$$f_{iMod_i} = \max_{j \in \{1,\ldots,m\}} \{f_{ij}\}$$

We calculate the probability *p* that the *i*-th mobile device is located in Voronoi area $Mod_i$ at a random 'working moment in time'

$$p_i = \frac{f_{iMod_i}}{\sum_{j=1}^{m} f_{ij}}$$

$p_i$ gives an idea about how often a person is to be found at the same location during the working day. We define the 'most probable workplace' as the $Mod_i$ -the Voronoi area if $p_i$ is larger than or equal to a certain minimum value $\alpha$ . If $p_i < \alpha$ , the 'most probable workplace' is considered as non-defined.

$\alpha$ could be set as the lowest decile of *p*. Or differently expressed

$$\alpha = \inf\left\{ p \left| \frac{\#\{p_i | p_i \le p\}}{n} \ge 0.1 \right.\right\}$$

Once the ‘most probable living place’ and the ‘most probable workplace’ have been determined for each mobile device, we aggregate the dataset. If, for instance, we want the number of mobile devices by most probable living place and most probable workplace for a given month, we will get a cross table.

| | | Most probable living place October YYYY | | | | | |
|---|---|---|---|---|---|---|---|
| | | Cell 1 | Cell 2 | Cell 3 | Cell 3 | Cell 4 | Cell 5 |
| Most probable living place October YYYY | Cell 1 | | | | | | |
| | Cell 2 | | | | | | |
| | Cell 3 | | | | | | |
| | Cell 4 | | | | | | |
| | Cell 5 | | | | | | |

## 6. CONCLUSIONS

The ultimate aim of exploring the potential of mobile phone and other privately held big data as a source for official statistics is their integration into regular statistical outputs produced and disseminated with a predetermined methodology, frequency, timeliness, level of detail and quality standards. Developing concrete use cases is the logical second step in this process, after the exploratory stage and before the development of experimental statistics and, eventually, full-fledged official statistics.

But developing use cases is also crucial in overcoming the major obstacle to achieving this end: the lack of data access. Policy makers and other potential users, including the mobile network operators themselves, will only be persuaded this is essential if it can be shown that costs are low, risks perfectly manageable and potential benefits very high indeed. Use cases are a concrete operationalisation of these benefits, demonstrating the essential knowledge to be gained from mobile phone data.

The two use cases presented in this article outline concretely and practically how very detailed and timely information about commuting can be extracted from aggregated mobile phone data, without any threat to privacy or data confidentiality, for a marginal cost once the initial database query and data treatment has been developed. Similar statistical use cases can be developed for other high-interest domains, some concrete examples:

- actual present population, both its size and composition (breakdown in inhabitants, tourists, people at work, visitors of hospital or school, recreating, … via combination with spatial datasets such as buildings register, land use);
- migration: labour migration and cross-border commuting via roaming data;
- tourism: identifying visitor numbers of (parts of) municipalities and tourism spots, events with origin, length of stay, on the basis of localisation network data and roaming data;
- mobility and transport of persons: traffic flows with origin and destination, transport mode by a combination of spatial datasets such as buildings register or land use, granular measurement of intensiveness of use of road, rail or air traffic nodes (road interchanges, train stations, airports).

A possible further step, which also might interest mobile network operators, is the combination of aggregated mobile phone datasets with similarly aggregated statistical and spatiotemporal datasets, resulting in wholly new statistical products being capable of replying to statistical or commercial queries which could not be answered before. For instance: what is the impact of meteorological conditions on rush hours? Where do the people in a particular location come from and what are the average demographics, income and education level of those areas of origin?

# 7. REFERENCES

F. De Meersman, G. Seynaeve, M. Debusschere, P. Lusyne, P. Dewitte, Y. Baeyens, A. Wirthmann, C. Demunter, F. Reis, H.I. Reuter (2016): *Assessing the Quality of Mobile Phone Data as a Source of Statistics* (mirror site), Q2016 Conference paper, June 2016 (pdf download)

**M. Debusschere, N. Waeyaert, K. van Loon (2018)**: *Key Factors for Obtaining Access to Big Data,* DGINS Conference paper, October 2018 ( not published, available at request)

# ABOUT STATBEL

Statbel, the Belgian statistical office, collects, produces and publishes objective and relevant figures on the Belgian economy, society and territory.

Statbel produces scientific statistics based on administrative data sources and surveys. The statistical results are published in a user-friendly way, and are available to everyone at the same time.

The data collected are used by Statbel for statistical purposes only. As statistical office, we guarantee at all times the privacy and the protection of confidential data.

Visit our **website**

www.statbel.fgov.be

Or **contact** us

Email: statbel@economie.fgov.be

Statbel (Directorate-General Statistics - Statistics Belgium)
North Gate - Boulevard du Roi Albert II, 16, 1000 Brussels
Email: statbel@economie.fgov.be

Enterprise number
0314.595.348

Responsible publisher
Nico Waeyaert

North Gate
Boulevard du Roi Albert II, 16
1000 Brussels