City data from LFS and big data

Project 2016CE16BAT107

Final report

February 2019

EUROPEAN COMMISSION

Directorate-General for Regional and Urban Policy
Directorate Policy
Unit Policy Development and Economic Analysis.

Contact: Moray Gilland

E-mail: REGIO-B1-PAPERS@ec.europa.eu

European Commission
B-1049 Brussels

Manuscript completed in February 2019.

Luxembourg: Publications Office of the European Union, 2019

PDF ISBN 978-92-76-08678-9 doi:10.2776/120066 KN-02-19-484-EN-N

City data from LFS and Big Data

Project 2016CE16BAT107

Final report

Prepared by
Johan van der Valk, Martijn Souren, Martijn Tennekes, Shan Shah, May Offermans, Edwin de Jonge, Jan van der Laan, Yvonne Gootzen, Sander Scholtus, Anna Mitriaieva (Statistics Netherlands)
Benjamin Sakarovitch (INSEE)
Sandra Hadam, Markus Zwick, Martina Rengers (Destatis)
Alex Kowarik, Marlene Weinauer, Johannes Gussenbauer (Statistics Austria)
Marc Debusschere, Anja Termote (Statbel)

Directorate-General for Regional and Urban Policy

2019 EN

Contents

SUMMARY

This document is a result of a project to explore how data collection for cities and functional urban areas can be improved. Cities are getting growing policy attention in Europe. All these activities underline the interest in city level indicators. Currently, a number of indicators are collected at city and functional urban area level, but the indicators typically collected by the Labour Force Survey can rather be found at the regional level than at the city level. The project consisted of two sets of activities. The first part was dedicated to the Labour Force Survey (LFS) to investigate if this source can provide a limited set of annual indicators for functional urban areas (FUA's) and whether gaps in the regional LFS time series could be filled by reattributing previous LFS respondents to the new NUTS regions. The second part dealt with the question if anonymised aggregated mobile phone data can be used to estimate commuting flows or population flows between municipalities or between statistical grid cells. The project was carried out by a consortium of 5 statistical institutes : INSEE (FR), Destatis (DE), Statbel (BE), Statistik Austria (AT) and Statistics Netherlands (NL) as project leader.

The work on the **LFS** consistent of an analysis of the countries covered in the consortium to look at the possibilities of using it for information on FUA's. Subsequently, the results of were discussed at a workshop on using the LFS for low regional for which all EU Member States were invited. The following conclusions were drawn:

1. Long consistent time series of LFS regional data of a limited set of indicators is possible and he statistical institutes are willing to provide historic time series;
2. Using the LFS for information on FUA's is not without issues since because sample and weighting schemes were not designed to produce results for FUA's which impacts on the reliability.
3. Most countries are okay with using NUTS 3 as building blocks as long as users are made aware that the results could be inconsistent compared with other regional results from the LFS.

One should be very careful in producing outputs in a creative way for varying geographies. Sample sizes should be larger enough. And even when they are large enough quality issues remain. The experts recommended that NSI's are involved in producing such outputs using the best methods available and checking if the results are of acceptable quality. Since only a limited set of indicators is required it would be feasible to define a set of tables that can be produced and verified by the NSI's. It would not involve a lot of work.

The members of the consortium worked on analysing and developing methodology on anonymised aggregated **mobile phone network data** given the specific circumstance in their country. Mobile phone network data is log data of cell phone connection events between mobile network cell towers and mobile devices. In this document we refer to this data by mobile phone data. In this project we learned several lessons attempting to use anonymised aggregated mobile phone data (not traceable/relatable to specific persons or personal information) to produce statistics on low regional level in general and city data in particular. It makes sense to present those lessons following the three process steps involved in producing statistics: acquiring, processing and disseminating data. We present them below.

First of all, statistical institutes will never be able to acquire mobile phone micro data because of legal, privacy and ethical reasons. They will have to settle with receiving anonymised aggregated information from Mobile Network Operators (MNO's) as input to produce statistics. Furthermore, purchasing such an information is not a valid option for producing official statistics. An alternative is that this information is provided within the context of a cooperation agreement. The important question here is of course: what is in it for the MNO? It is still early to answer this question. The future will tell. At this stage one can mention a few points and elements that seem to have potential. The community of statistical institutes are able to develop and maintain a methodology as open standard how to produce statistics from anonymised aggregated mobile phone data. This means that the MNOs do not have to do this work but can benefit from it. This methodology

could be applied by the MNOs to produce customised statistical information. In addition, they could build new modules on this general software and methodology to produce more sophisticated products. Finally, the open standard methodology is internationally and universally usable which is beneficiary to all parties concerned. This approach has to explored and needs to be tested in practice but it seems to have potential for a bright future.

The second part of the process is the methodology how to produce statistical output starting with the raw signalling or CDR data. The section on anonymised aggregated mobile phone data deals with that part of the process. The bottom line is that quite some methodology has been developed but is still not yet at a mature level. A lot needs to be done. However it is do-able. A basic method is ready. It can already be used to produce outputs. The method must be improved obviously in the next phase. An important step to take is to test it in other countries using data from other MNOs. This will all helps to improve the methodology and define new and better outputs. It is important that this work is carried out by several institutes in a coordinated way. The resources has to be pooled, innovation has to be done jointly and the methods must be tested in several countries. This is required in order come to an international standard in an efficient way.

The final element of the process concerns the output mobile phone data should produce. The specific features and possibilities of anonymised aggregated mobile phone data forces us to define products that go beyond the standard and traditional products. Methods, formats, standards and platforms have to be defined, developed, maintained and improved. In addition, we need interactive visualisation tools that enables a range of users to produce tailored outputs or carry out all sorts of analysis. These tools have to developed making use of state of the art technologies. It requires an longstanding and never ending innovation process that has to be organised. This task should not be underestimated. The statistical community will have to organise the involvement of a broad range of stakeholders. Varying from national and international organisations and authorities to local players like cities or regional governments.

We argue that using anonymised aggregated mobile phone data is a game changer for the way statistics are produced. We will have to adapt the way we work in all stages of the statistical production process. In the past, NSI's could work quite independently to produce statistics. They collected data themselves and defined to a large extent the products they supply. Circumstances of data providers of users could largely be ignored. With many big data sources in general and mobile phone data in particular this is not the case anymore. The methods, products the products they create as the organisation of the work is depending on other actors forcing NSI's to cooperate with the actors involved. Acquiring data requires collaboration with MNO's in dealing privacy sensitive data and find creative solutions for a partnership that is beneficiary to all parties without disturbing their main processes. It could even lead to a situation that NSI's have to accept, at least for the time, that a part of the process will be a black box where they have no saying on. Developing methodologies will require international cooperation between statistical institutes, collaboration with scientists and international organisations since setting open standards is a common goal concerning all. Finally, new kind of products will have to be development in close cooperation with all kinds of users.

1. INTRODUCTION

In 2016, many actions highlighted the growing policy attention to cities. The Urban Agenda for the EU was adopted, which includes a specific pillar on better knowledge and data. The State of European Cities report was published jointly with UN-Habitat. The new global Urban Agenda was adopted at the UN-Habitat III conference. The UN Sustainable Development Goals, including a specific urban goal, were approved. Last but not least, Eurostat produced its first publication dedicated entirely to cities: Urban Europe. All these activities underline the interest in city level indicators. Both the current programmes and the preparation of Cohesion Policy post 2020 would benefit from more information and indicators for cities. Currently, a number of indicators are collected at city and functional urban area level, but the indicators typically collected by the Labour Force Survey can rather be found at the regional level than at the city level. Commuting flows are often only measured once every ten years with the census, but some countries do not include this indicator in their census and struggle to find alternative sources for this information. New developments, such as the use of big data, may allow a regular, cost-effective and harmonised data collection. The overall goal of this action is to explore how data collection for cities and functional urban areas can be improved.

The project aimed to understand whether 1) the Labour Force Survey (LFS) can provide a limited set of annual indicators for functional urban areas, 2) whether big data can estimate commuting flows or population flows between municipalities or between statistical grid cells and 3) whether gaps in the regional LFS time series could be filled by reattributing previous LFS respondents to the new NUTS regions.

The project was carried out by a consortium of 5 statistical institutes : INSEE (FR), Destatis (DE), Statbel (BE), Statistik Austria (AT) and Statistics Netherlands (NL) as project leader. The project started in January 2018 with a duration of 12 months. This document is the final report presenting the results of the project.

2. LABOUR FORCE SURVEY

2.1. Introduction

The Labour Force Survey (LFS) is the main source in the EU for harmonised labour market statistics. This survey is based on a sample and therefore allow a limited degree of regional breakdown. Information on NUTS 2 in generally available. The European Commission and OECD developed Functional Urban Areas (FUA's) which play an important role in current policies dedicated to urban regions. The following classification of functional urban areas into four types according to population size is distinguished:

- Small urban areas, with a population below 200 000 people;
- Medium-sized urban areas, with a population between 200 000 and 500 000;
- Metropolitan areas, with a population between 500 000 and 1.5 million;
- Large metropolitan areas, with a population of 1.5 million or more.

One would expect expected that the populations of in particular Metropolitan urban areas, with population between 500,000 and more, are of the same order as many NUTS 2 regions. Therefore, in principle, the sample size should large enough to provide LFS statistics. Small urban areas can be excluded from the analysis, since the corresponding sample sizes are too small. For Functional Urban Areas was investigated the possibilities to generate information that involves the following indicators: employment rate, unemployment rate and educational attainment.

2.2. Analysis on using LFS for information on Functional Urban areas

Statistics Austria, Statistics Belgium, Destatis, INSEE and Statistics Netherlands reported on the use of the LFS for information on Functional Urban Areas. They were all asked to try to produce the following indicators: employment rate, unemployment rate and educational attainment of persons 25-64 years of age.

Austria

In Austria the Functional Urban Areas consist of the capitals of six federal states and their commuting area, which are Vienna, Linz of Upper Austria, Graz of Styria, Salzburg of Salzburg, Innsbruck of Tyrol and Klagenfurt of Carinthia.

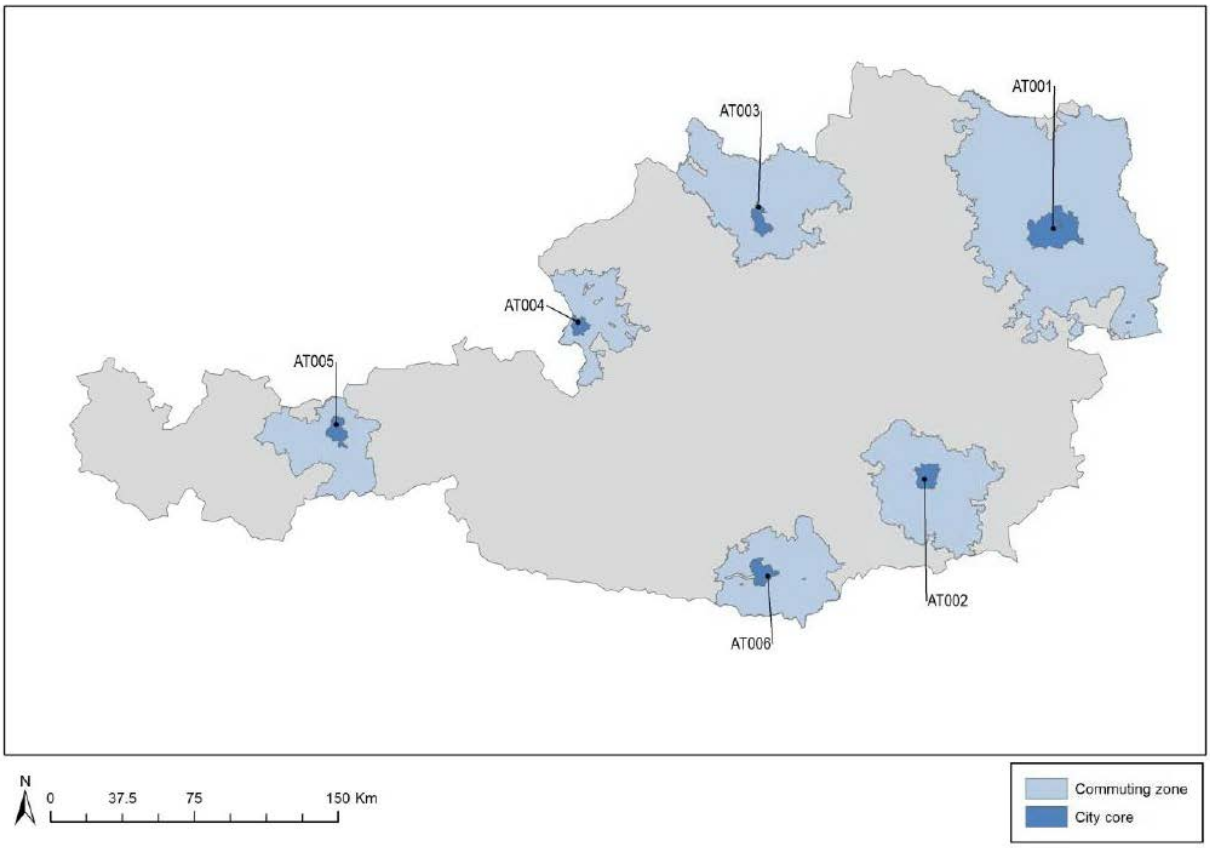


Figure 1: Functional Urban Areas in Austria

It was possible to produce the three indicators for the FUA's in Austria. The shares were compared to that of the corresponding Federal states Employment rates were similar. For unemployment shares of the FUA's were higher than that of the federal state except for Vienna. For more detail see Annex 1.1.

In case of Austria in order to produce reliable using LFS estimates for FUAs some improvement needs to be done. For instance, using variables from other administrative sources which highly correlate with the education and employment rate. Taking the corresponding population margins from these sources and specifically calibrating the sampling weights can further improve the estimates for this specific problem. Statistics Austria tested multiple approaches for calibrating the sampling weights for which the results are presented for 2 of those approaches in the following. The approaches correspond to

- Calibrating on sex and age;
- Calibrating on sex, age and employment status (taken from administrative sources).

But to insure good quality of data by using these approaches more tests and analysis need to be done. Important to note is the issue of recoding administrative data to correspond with the variables of interest in the LFS. For Statistics Austria this did not work in all cases and is still an open issue.

Belgium

Belgium has 11 Functional Urban Areas. Brussels is the capital of Belgium and the only large functional urban area in Belgium, with a total population above 1.5 million. Antwerpen, Liège and Gent are functional urban areas with population between 500,000 and 1.5 million. Charleroi, Brugge, Namur and Leuven are medium-sized urban areas with population between 200,000 and 500,000. Mons, Kortrijk and Oostende are small urban areas with population between 50,000 and 200,000. Antwerpen, Brugge, Gent,

Kortrijk, Leuven and Oostende are situated in the Dutch-speaking part of Belgium. Charleroi, Liège, Mons and Namur are French-speaking. Brussels has two official languages: both Dutch and French1. Figure 2 presents the FUA's on a map.

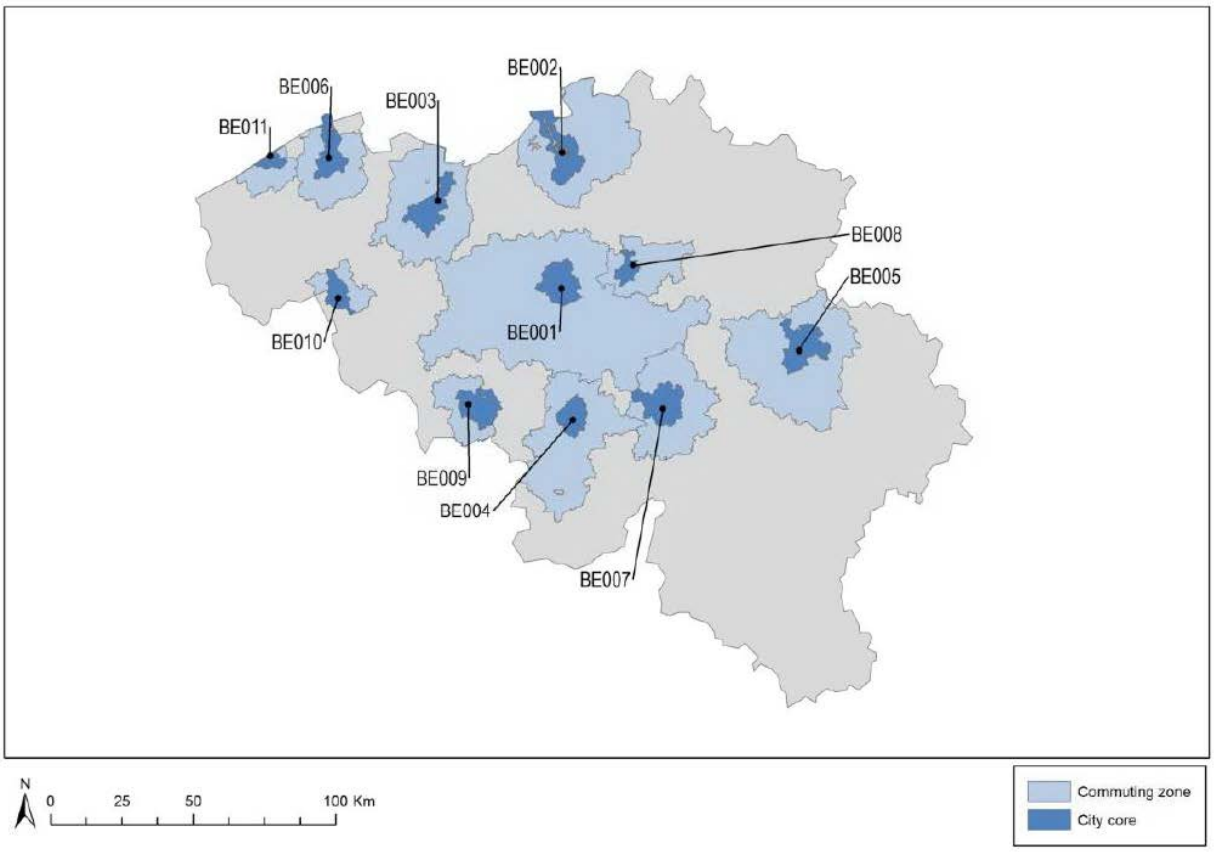


Figure 2: Functional Urban Areas in Belgium

A first analysis of LFS-indicators for the Belgian FUA's shows that the FUA's largely differ.

There are FUA's with high employment, low unemployment rates and high percentages of high educated persons. Other FUA's perform less well for one or more of these indicators. The FUA's Brussels and Antwerp, the biggest Belgian FUA's, score average. Brussels and Antwerp have average levels of employment, unemployment and higher education diploma's. For more details see Annex 1.2.

Despite the fact that the Belgian LFS is not designed to generate statistics on the level of FUA's, first analyses show that LFS can be used to a certain extent for large and medium sized FUA's, with the exception of unemployment. But it is difficult to interpret evolutions from year to year because of the sometimes large confidence intervals. The presented results should be interpreted very carefully and further analysis is needed.

In case of Statistics Belgium there are general issues about the Belgian LFS: sampling method, stratification, extrapolation, publication thresholds and issue related consequences of the Belgian LFS methodology at the level of FUA's.

The Netherlands

For the Netherlands in total 35 FUA's can be identified (figure 3). These cover 82.3 percent of the total population of 15-64 years old. More than half of these areas are so-called small urban areas which will not be discussed. The remaining 16 metropolitan or

medium-sized areas cover 65.7 percent of the population in the Netherlands. The metropolitan FUA's can be found mostly in the Western part of the country: Amsterdam, Rotterdam, The Hague and Utrecht. One area is located in the Southeast: Eindhoven. The medium-sized areas can be found all across the country.

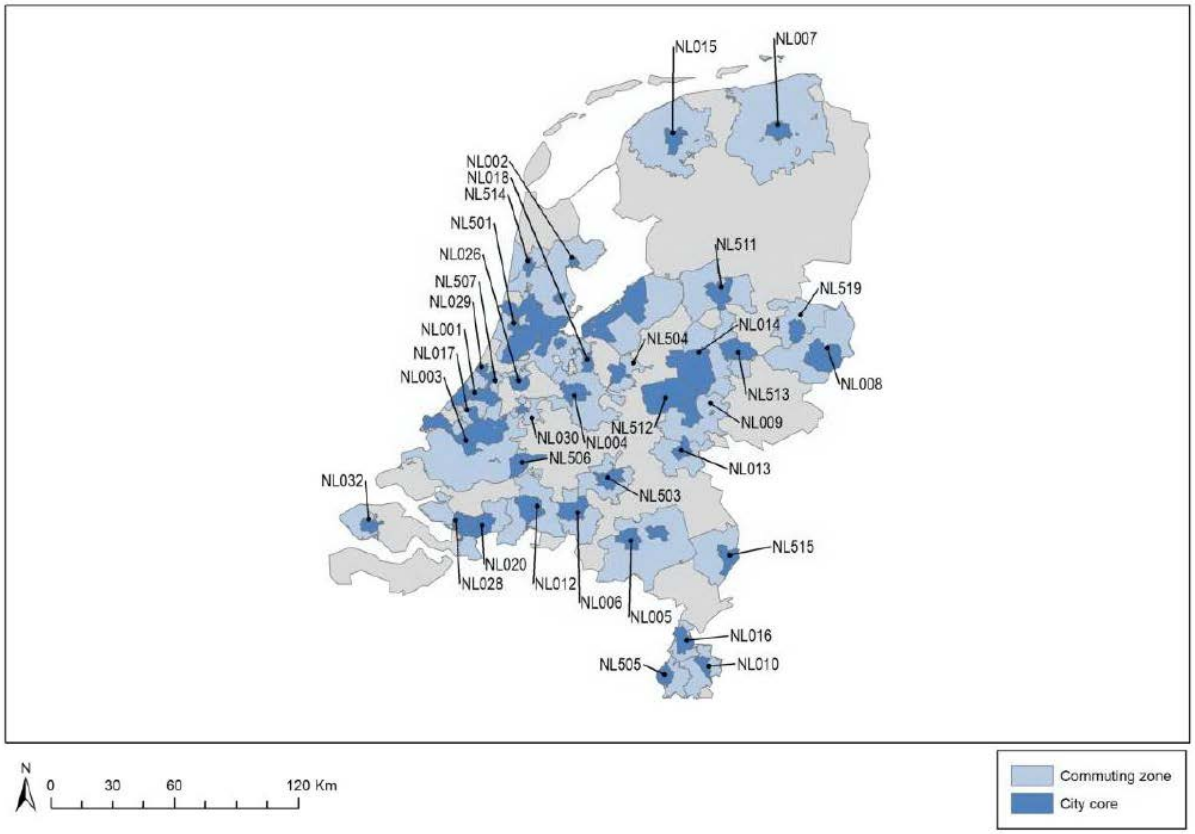


Figure 3: Functional Urban Areas in the Netherlands

For the Netherlands it is possible to provide indicators for FUA's based on the LFS. They see show quite some variation in shares for all three indicators . For more details see Annex 1.3.

The results show that the labour market can be very different between areas within a rather small country as the Netherlands. It is therefore very useful to have precise and unbiased statistics. GREG estimates can be used but should be handled with care. Small Area Estimation (SAE) can help to improve precision and reduce the bias, especially for employment and education statistics. The biggest issue for regional statistics in The Netherlands is the precision of the estimates. This was the main reason why SAE were introduced. When analysing the results it showed that some areas also have more selection biases than others. The bias can be tackled by introducing auxiliary information on a regional level. At the moment this is incorporated in the SAE but it could also be tackled by improving the weighting on regional level. Producing the GREG estimates is a fairly straightforward. For SAE this is more complicated and time-consuming. But since only a very limited set of indicators have to be produced it seems a sensible approach.

Germany

In Germany, there are in total 208 units which are relevant for determining FUAs. The FUAs are composed of 125 city cores called category 'C', and 83 commuting areas of these cores, known as category 'F'. Especially through the agglomeration of the cities in North-Rhine Westphalia, e.g. the Ruhr area, there exist more city cores than commuting zones. Furthermore, 11 commuting areas are indicated as city cores at the same time –

as mentioned above (see red highlighted areas). Therefore it should be decided for a clear assignment of these areas in order to avoid double counting.

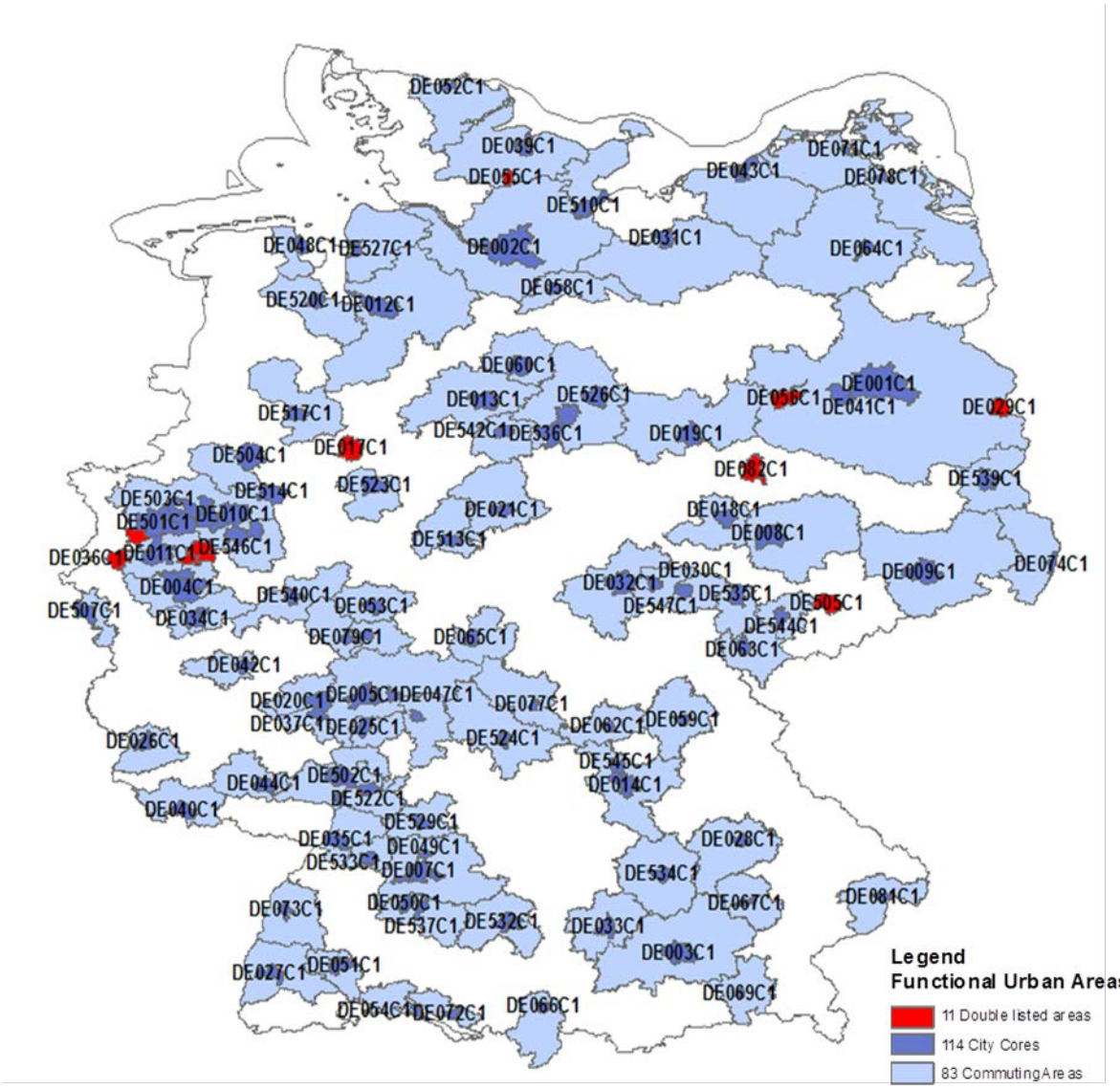


Figure 4: Functional Urban Areas in Germany

For the German FUA's no universal conclusions can be drawn regarding indicators based on LFS data. Uncertainty and deviation in the data and indicators depend on the number of observations. The areas with the smallest sample sizes are in general small urban areas. In comparison, metropolitan areas or large metropolitan areas are accompanied by higher sample sizes, lower CVs and unbiased estimates. Due to the high number of observations for estimating the employment rate, the direct estimation works quite well. The relative standard errors are comparatively low in comparison to those of the unemployment rate for FUAs. The results of direct estimation for the unemployment rate go together with larger standard errors because of lower numbers of observations and biased estimators. Indicators with breakdowns increase the likelihood that out-of-sample domains arise because of missing observations in the sample.

Overall, the methodological approach of the Urban Audit provides reliable indicators for the FUAs. Next to this approach alternative estimation methods can be used like small area estimation in combination with register data or administrative data. A direct estimation by using solely LFS information is not recommended for medium and smaller urban areas. But it is useful to get an overview of indicators for metropolitan and large metropolitan areas.

France

INSEE did not carry out a full-fledged analysis on FUA's. Only sample sizes were determined and a comparison with Census data was carried out. Their conclusion based on this research is that the LFS should be used for analysis on national level. The French LFS cannot be used for analysis at FUA level.

2.3. Workshop on city data from LFS

On 24-25 of September 2018 a LFS workshop took place on Malta at the premises of the Maltese statistical office. Participating institutions were the statistical offices of BE, AT, DE, EE, SI, UK, ES, IT, HU, HR, FI, PL, GR, MT, NL, BA, ME, KO and TR, Eurostat and the ILO. The topic of the workshop was to identify the potential of using the LFS to provide information on cities and urban areas. Increasingly, international and national policies are targeted to urban areas. For identifying opportunities and weaknesses and to monitor the effectiveness of policy measures, labour market information for these areas is needed. Using the LFS as a data source has in principle high potential. It could provide internationally comparable labour market data which allows comparing urban areas in the EU. The objective of the workshop was to explore the possibilities of using the LFS for regional information and to provide recommendations for the future.

Using LFS for information on Functional Urban Areas

The European Commission and the OECD defined so-called Functional Urban Areas (FUA's) composed by the cities and their commuting zones . They play an important role in current policies dedicated to urban regions. For the purposes of this project, the following classification of functional urban areas into four types according to population size are distinguished:

- Small urban areas, with a population below 200 000 people;
- Medium-sized urban areas, with a population between 200 000 and 500 000;
- Metropolitan areas, with a population between 500 000 and 1.5 million;
- Large metropolitan areas, with a population of 1.5 million or more.

One would expect that the populations of in particular functional urban areas, with population over 500,000 , are of the same order of size as many NUTS 2 regions. Therefore, in principle, the sample size should be large enough to provide LFS statistics for such regions. Small urban areas can be excluded from the analysis, since the corresponding sample sizes are too small.

Members of the project consortium investigated the possibilities to generate information that involves the following indicators: employment rate, unemployment rate and educational attainment. At the workshop colleagues from Statistics Austria, Statistics Belgium, and Statistics Netherlands presented their findings. The results showed that all three indicators can be produced for metropolitan and medium-sized urban areas. It also gave some interesting insights showing the power of FUA's as a tool for analysis. An illustrative example was the situation of the Belgian capital Brussels. The employment rate in the FUA Brussel is considerably higher compared to that of the Brussels capital region. The latter is generally used for labour market analysis since this is the most relevant administrative unit. A general finding was that unemployment rates were quite unreliable for medium-sized urban areas due to too small sample sizes. Therefore it does not seem sensible to produce this particular indicator for relatively small FUA's.

All presenters pointed out that producing estimates for FUA's directly based on the original microdata involve some serious quality issues. First of all, the samples are not designed to produce indicators for such regions. Therefore, without reweighting to population totals representativity cannot be guaranteed and consistency with demographic information is lacking. But even more importantly, confidence intervals are generally large for the direct estimates for such regions. The Dutch colleague argued that using Small Area Estimation methods are a good way to deal with this issue.

Incorporating auxiliary information from registers makes it possible to increase the reliability substantially.

In the discussion at the workshop the participants supported the concerns on the quality of the statistics that were raised by the presenters. The opinions of the participants varied concerning the question to what extent issues of representativity or coherence with other statistics are a major problem. Low levels of reliability of the LFS estimates was generally considered as being a more serious issue. Using SAE techniques with auxiliary administrative information is in principle a good statistical method to deal with this issue. The number of indicators and breakdowns that need to be produced are limited. That means that the relatively technically challenging work and resources involved will be limited too. This points in favouring such an approach. However, is still an open question to what extent these methods can be applied throughout the EU in a harmonised approach. Countries have different situations regarding the availability of administrative data and the possibility to link this with survey data at micro level. So it is not possible to apply SAE techniques in all countries. It seems that a valid approach could be that countries use the best method possible to produce the information for FUA's. This could be a SAE method for one country and direct estimate for another country. However, Eurostat is working together with the research community on a project which main deliverable is guidelines on SAE to be used on voluntary basis by the countries without established national methodologies.

The need for longer time series from the LFS

Eurostat explained the need for longer time series from the LFS. There is increasing demand for sub-national data and analysis. Firstly, there is need for consistent time series versus changing (regional and local) boundaries. And secondly, various policy fields require data at various territorial levels. Territorial focus is essential in Cohesion Policy. The instruments and focus, vary. They could be for instance on cross-border cooperation programme areas or actions in urban areas or urban-rural relationships. Consequently, reporting on the implementation of the programmes requires various territorial data, timely updated with longer time series. How to make the data collection process flexible enough to cope with boundary changes and to facilitate the provision of data at different territorial levels? NUTS classification changes every 3 years. Member States are obliged to send a limited historical time series just 4 years after the adoption of the NUTS regulation. How to keep NUTS codes up-to-date in LFS data output and to maintain the longest possible time series?

Eurostat presented a possible solution to deal with both the issue of dealing boundary changes to provide consistent time series and the issue of growing needs for different geographical areas. If the LFS microdata includes a kind of geocoding that allows a distinction in (1 km2) grids the backward calculations could be easily done. This would be very flexible in producing data for all kinds of geographies of interest. It would also enable production of data by urban centres, urban clusters and rural grid cells which would be much more precise than the 3 DEGURBA classes. DEGURBA needs to be updated every time municipality boundaries change which results in break in time series. The first question that Eurostat put to the participants regarding this idea was: are countries technically able to provide geocoded LFS data? The second question would be if the answer is yes: is this a good idea to incorporate geocoding in the LFS micro data file?

Next steps that were mentioned in the presentation of Eurostat are that DG REGIO will prepare a document on the policy needs regarding longer time series. In addition, Eurostat proposed to initiate a test case with few countries that would volunteer to prove the feasibility of geo-coding of the LFS. Subsequently Eurostat will report on the results to the relevant WGs on the progress. A request to the participants of the workshop was to express their interest in testing the feasibility of geocoding the LFS.

Most of the participants of the workshop clearly stated that they were in principle able and willing to provide long historical time series if required. The extra work involved was

considered limited. At national level they receive similar requests from users. So most of that work is carried out anyway. This is considered as being part of their regular work.

Regarding the geocoding most of the countries said that it would be technically possible or will be possible in the near future to geocode surveys. However, this does not necessarily mean that it is a good idea to incorporate this in surveys in general and the LFS in particular. Many grid cells will be empty anyway and most of them will only contain only very few elements. Confidentiality is also an important issue. Some countries could even face legal issues that prevent them from providing such information. On the other hand, one has to realise that the 1 sq. km grid data from the 2021 Census is included in the implementing Regulation with the confidentiality rules already established. Another issue is that although aggregating survey data using grids as building block is technically possible it is not necessarily a valid approach from a statistical point of view. The LFS is weighted using NUTS 2 regions. Information that produced for other geographical areas will suffer from imperfect representativity and coherence issues with other statistics. These statistics will therefore have lower quality. For these reasons, it was concluded that the issues related to incorporating geocoding in EU surveys should be further investigated in the light of the reservations that were expressed.

Using NUTS3 regions as building blocks

In a recent project on exploring the possibilities of using the LFS to produce labour market information for border regions NUTS 3 regions were used as building blocks to define macro border regions. Such a combination of several NUTS 3 regions amounted to regions of the size of NUTS 2 regions. This would allow the production of indicators based on the LFS of reasonable reliability because of sufficiently large sample sizes. This method was seen by the researchers as a good approach to make the best use of the LFS given the limitations due to the sample sizes. Eurostat applies the same approach for publishing data on coastal regions, mountain regions, border regions or island regions. The participants were asked to comment on this practices.

The opinions on the use of NUTS 3 regions as building blocks to create larger regions differed. A large number of participants found it a logical an acceptable approach. Others expressed their concerns that representativity and coherence with other statistics are not guaranteed since the constructed regions does not fit the regions that are used in the sampling and weighting procedures (metropolitan regions composed by a single NUTS 3). One should therefore handle the results with care. Dissemination thresholds should be applied and users should be informed about quality issues like imperfect representativity and possible incoherence with other statistics.

Conclusions of the workshop

It was good to have had an opportunity to discuss the issues in using the LFS to produce regional information with a large number of LFS-experts. All participants acknowledged the demand for more and better regional information. The LFS can be used to satisfy the needs for that kind of information, albeit to a certain extent.

All countries apply sampling designs and weighting methods for the LFS incorporating regions at NUTS 2 level. As a consequence, to produce a broad spectrum of indicators for such regions is in order as long as the dissemination thresholds are respected. For indicators for other geographical areas the quality is not guaranteed due to risks of lack of representativity, inconsistency with other statistics and lower reliability of the estimates. It is important that the statistics are of acceptable quality. One can conclude from the discussions that the statistical offices would like to be involved in this process in two ways. Firstly, the member states would like to be involved in assessing if an indicator for a certain area is of acceptable quality. Secondly, they would like to be given the possibility to provide better statistics for these regions if they are able to do so. This could mean either reweighting of the sample or using more sophisticated methods like

Small Area Estimation to improve the reliability and representativity of the results. Following this line of reasoning, one can conclude that the issues related to the incorporating geocoding in EU surveys should be further investigated before deciding in favour of such an approach.

Producing information on Functional Urban Areas (FUA's) that are at least of medium size based on the LFS is possible, with the constraint that production of unemployment rates for medium-sizes FUA's should be avoided because of low reliability. Furthermore, it is recommended to improve the results by using reweighting the sample or applying Small Area Estimation. This approach would imply that such tables are provided by the member states rather than by Eurostat.

Backward calculations of regional classifications in case of boundary changes is no real problem for statistical office to carry out. This is considered to a part of their regular work. For that reason, they are prepared to do this when necessary.

Using NUTS 3 as building blocks to create larger regions is considered by many of the participants as a valid approach. Dissemination thresholds should be applied and users should be informed about quality issues like imperfect representativity and possible inconsistencies with other statistics.

2.4. Conclusions on using LFS for city data

Long consistent time series of LFS regional data of a limited set of indicators is possible. The statistical institutes are willing to provide historic time series. This is part of their regular work.

Both the analyses carried out within the project as the opinions of the LFS-experts as expressed at the workshop showed that using the LFS for information on FUA's is not without issues. Generally, estimates on unemployment involves too small numbers to produce reliable results. For other indicators reliability is limited because sample and weighting schemes were not designed to produce results for FUA's. Small area estimation and reweighting could improve the results. But this has to be carried out by the Member States and cannot be done by Eurostat using LFS data files.

INSEE explicitly stated the opinion that all applications using LFS data for regional data that deviates from NUTS 2 is unacceptable. Sampling and weighting methods are designed to produce data on NUTS 2 as most detailed regional level. As a consequence delimiting other regions will have quality issues. Because of the high political sensitivity of the LFS information they are of the opinion that it should not be done. They are even opposed using NUTS 3 areas as building block to create larger regions. Other countries are less extreme in their opinion when it involves other indicators than unemployment. Most think that after using reweighting and if possible and relevant using small area estimation techniques acceptable results can be achieved. Most countries are okay with using NUTS 3 as building blocks as long as users are made aware that the results could be inconsistent compared with other regional results from the LFS.

For large regions like large FUA's and regions that are created by merging a number of NUTS 3 regions is should be possible to use LFS to produce information. Ideally these regions should be an element in the sampling and weighting procedures. This is technically possible. It is not so complicated to carry out. And will not jeopardise the other results. The best way to ensure this is to include those regions in the regular LFS dataset. It will however take some time before this is implemented in all countries. An alternative is that NSI's provide tables on these regions. To produce these table they can use LFS data that they reweight, or they can use small area estimation or use other source that they have available at national level. This is also a feasible approach. A major advantage of this method is that it can be applied immediately. It will of course involve a limited amount of work for the NSI"s.

3. MOBILE PHONE DATA

3.1. Introduction

With the trend of disaggregation there is a growing attention to cities which includes an important data component. Currently, a number of indicators are collected at city and functional urban area level, but the indicators typically collected by the Labour Force Survey can rather be found at the regional level than at the city level. Commuting flows are often only measured once every ten years with the census, but some countries do not include this indicator in their census and struggle to find alternative sources for this information. New developments, such as the use of anonymised aggregated mobile phone data, may allow a regular, cost-effective and harmonised data collection on inter- and national levels. The overall goal of this work package is to explore whether mobile phone data can estimate population flows.

Within the project the following countries have worked with or on anonymised aggregated mobile phone data to produce low regional information on day-time or night-time populations: FR, DE, NL, BE and AT. The results are presented below in separate sections.

3.2. Experiences of using mobile phone data in France

Research problem

Mobile phone data (MPD) are the tracks that are left by connections on a mobile network. As such they represent an important opportunity to get a very fine view of spatial footprints of mobile phone users.

Yet the geographical precision of this data is quite heterogeneous. Basically the antenna that connected to a SIM card is the only information that may be recorded. So depending on the local density of antennae the precision of the location may vary a lot. Furthermore the location may be available on a specific grid, depending of the shape of the network. So to compute aggregates on a given grid, as INSPIRE or municipalities, we need to map the events directly on that grid.

Moreover the location of observed events depends on the model of antennae coverage areas. Models of the extension of the zone receiving a signal from a given antenna may be based on very different information. That information which is provided by the MNO may vary from the output of the propagation models used for radio planning[2] to the coordinates of cell towers.

Thus the research problem is to map events observed on the mobile network to a given grid (as INSPIRE) tanking as input a model of antenna coverage areas. Of course additional data, available to the NSI, is of much help in that task. Figure 5 sums up the procedure.

[2] https://www.forsk.com/crosswave-propagation-model

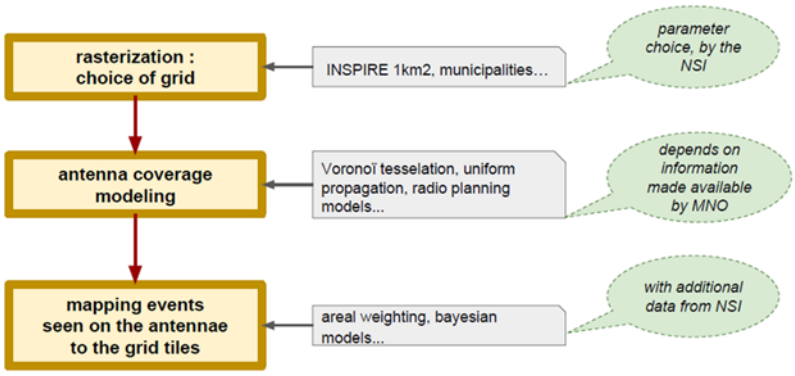


Figure 5: Mapping procedure

Data

There are two sources of data that were used in this work. The first one is mobile phone data, the second is land use data and tax data that are available in the NSI.

The mobile phone dataset we had access to is 5 month of Call Detail Records (CDR) from 2007 from Orange, the French main MNO. It is a pseudonymised dataset that the Data Protection Agency (CNIL) allowed to keep for research purposes. CDR contain information on the MNO customers and keep tracks of their activities as sending or receiving a call or a SMS. Table 1 shows the type of information available in such a dataset. Around 18 million users are recorded in the dataset, representing around 15 billion events.

Table 1: Call detail records structure

SIM card transmitting	SIM card receiving	type of event	transmitting antenna	receiving antenna	timestamp	length
SIM-1	SIM-2	call	A-1	A-2	13/06/2007-14:26:03	7m32s
SIM-1	SIM-3	SMS	A-3	-	25/08/2007-12:04:58	-

The only information made available by the MNO on the antennae is the coordinates of the cell towers. Without further information we model the coverage areas with Voronoï tessellation. It is a common method from the literature[3]. The approximation in using Voronoï tessellation is that the signal is always transmitted to a mobile phone by the closest antenna. It is a strong approximation as in reality there is an important variety of antennae and their coverage areas overlap. Mobile phone connections switch from one antenna to another even when standing still.

There were around 18000 cell towers at the time in France. Figure 6 displays a map of the Voronoï tessellation. One may observe that the polygons are very heterogeneous in size and much denser in urban areas. Figure 7 shows the distribution of the sizes of Voronoï cells.

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

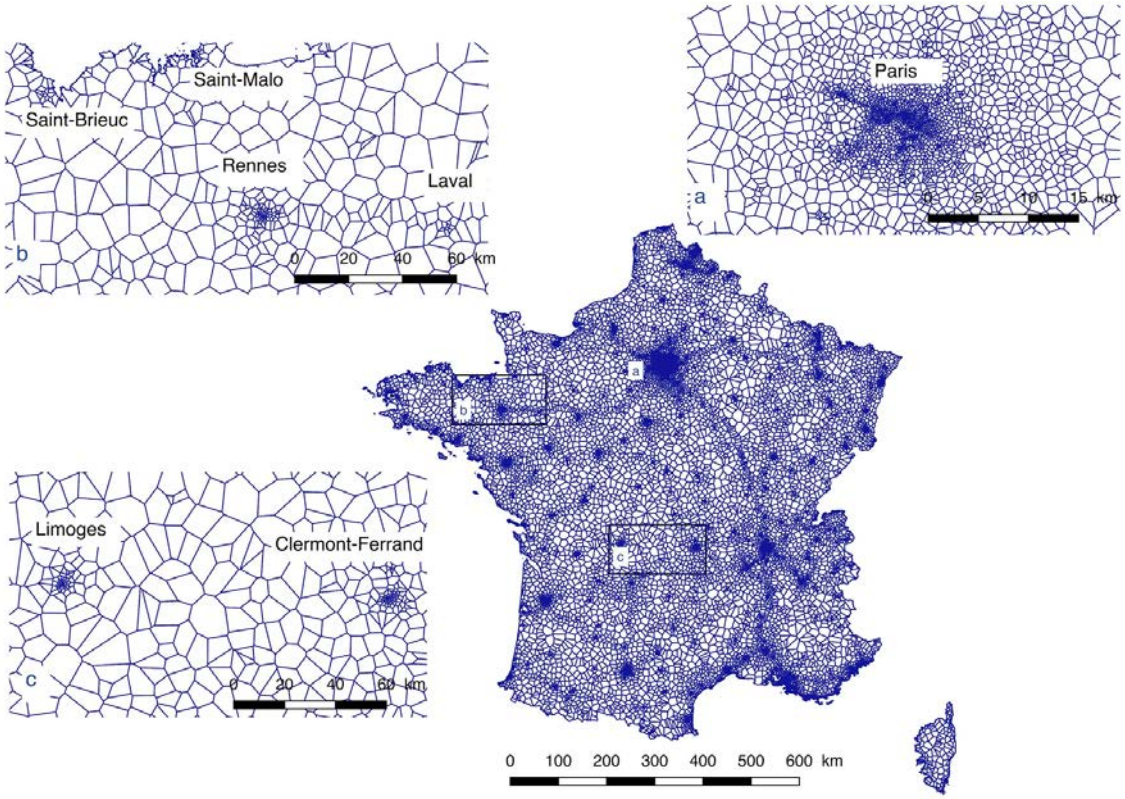


Figure 6: Voronoï tessellation of France from the location of the cell towers

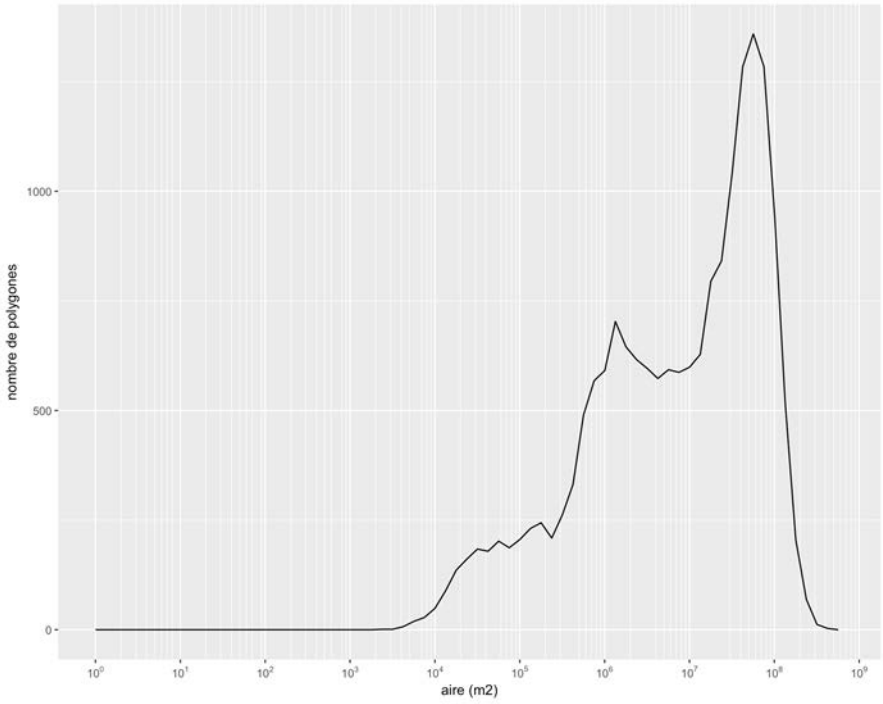


Figure 7: distribution of the Voronoï cells size

Yet the method described in the following section should adapt easily to a better level of information, and a finer modelling of the antenna coverage.

We have access to this dataset in the framework of an agreement signed by Eurostat, Orange labs and INSEE for research purposes. The access to the data is only possible at the MNO premises[4].

Two other data sources have been used. The BD Topo is a dataset on land cover and land use maintained by the French mapping agency (Institut national de l'information géographique et forestière,IGN). It contains four main types of land cover :

- natural zone
- transport infrastructure
- industrial building
- other building

The other data source is transmitted by the tax administration to the NSI. It is a geocoded tax register and thus provides us with the number of people in each household with geographic coordinates. Hence the concept lying behind this dataset is the one of fiscal residency.

Methods

(a) Living in a grid tile

The use case developed is to infer the place of residence for the mobile phone users. We use a 500m*500m grid and try to detect home as one of the tiles in this grid. In fact we will reason with probabilities and describe an individual home as a distribution of probabilities over those tiles. Summing the individual distributions of home we get densities of resident population. The reason for working with nighttime population and place of residence estimates is to compare with a reference. We compare those estimated densities to the tax data so as to be able to validate our methods in handling mobile phone date.

(b) The Bayesian model to locate events

The method to locate each events is inspired from the mobloc package developed by Statistics Netherlands. We make use of the Bayesian model developed in the Netherlands, which is proceeded as

$$\mathbb{P}(tile_i|cell_j) \propto \mathbb{P}(tile_i).\mathbb{P}(cell_j|tile_i)$$

Each event is observed at a cell level, meaning the antenna transmitting the signal. And the problem is to infer in which tile of the grid the mobile phone that generated this record was. Indeed the problem is to evaluate for each grid tile the probability of location knowing what was the cell. This makes the Bayesian description very relevant. Thus this probability is proportional to a prior that we have on each tile time the probability of the signal being transmitted through a given cell knowing that the phone was on such grid tile. This general model applies to different concrete cases with different information on the antenna.

In our case we model the cells through a Voronoï tessellation, so each cell is a Voronoï polygon. Besides the probability of being in the cell j knowing that the phone is in the tile i is computed very simply with areal weighting.

We have an a priori information on each tile through the land cover register. As we locate nighttime events and try to detect home, our prior is that events during the night are located in buildings (non industrial). So we compute for each tile the share of the volume of building in it to weight the events in the grid.

[4]For a description of the Big Data infrastructure set up by the MNO to process this type of data see https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/c/ce/WP5_Deliverable_5.4_Final.pdf

(c) From single events to home detection

There are different heuristics in the literature for inferring home. They have been assessed by Vanhoof[5] on the same dataset. Here we work with a different view as we locate each event as a probability distribution over a grid. So the estimated home will also be a probability distribution. For each user, per month, we sum up the probabilities from the different events recorded between 9 pm and 7 am. Then we normalize the results over the grid. Finally individual homes are considered as these distributions.

(d) Validation

The last step is the confrontation to our validation dataset, the geocoded tax register. We do not adjust our estimates to fit the overall population so there is one side the population of one MNO clients and on the other the global population. Yet we compare visually the densities over the countries. In urban areas we try to assess if we can reproduce the intra urban variations of densities. Hence we concentrate on the last two deciles of the grid tiles in densities and compare the shape of that urban core.

Results

On the national level

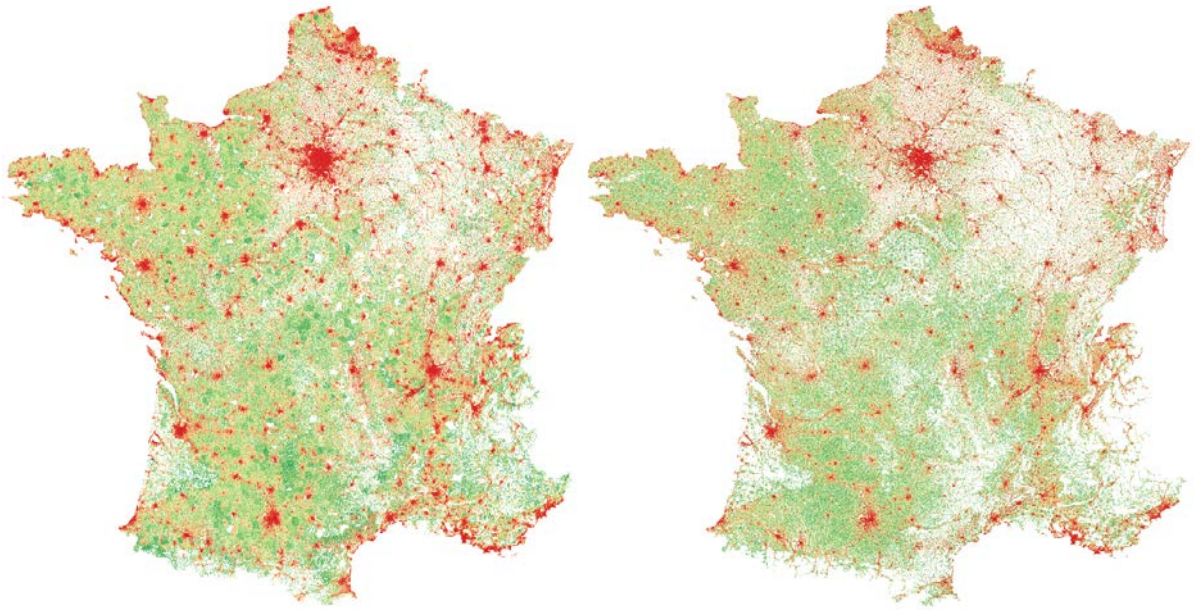


Figure 8: National densities of resident population from mobile phone data (left) and tax data (right) scale : decile of the number of resident per tile of 500m*500m

We compare the results on a national scale on figure 8. It makes clear that the global structure of the most densely populated areas is well reproduced. Yet it also appears that the estimations tend to produce larger urban areas and under evaluate the sparsely populated regions over the territory like in the North-Eastern part of France.

[5]Vanhoof M., Reis F., Ploetz T., Smoreda Z. (2018), Assessing the Quality of Home Detection from Mobile Phone Data for Official Statistics," Journal of Official Statistics In Press . Preprint at: http://eprint.ncl.ac.uk/author_pubs.aspx?author_id=183527

Infra urban variations

Figure 9: Densities of resident population in Marseille urban area from mobile phone data (left) and tax data (right) scale : decile of the number of resident per tile of 500m*500m

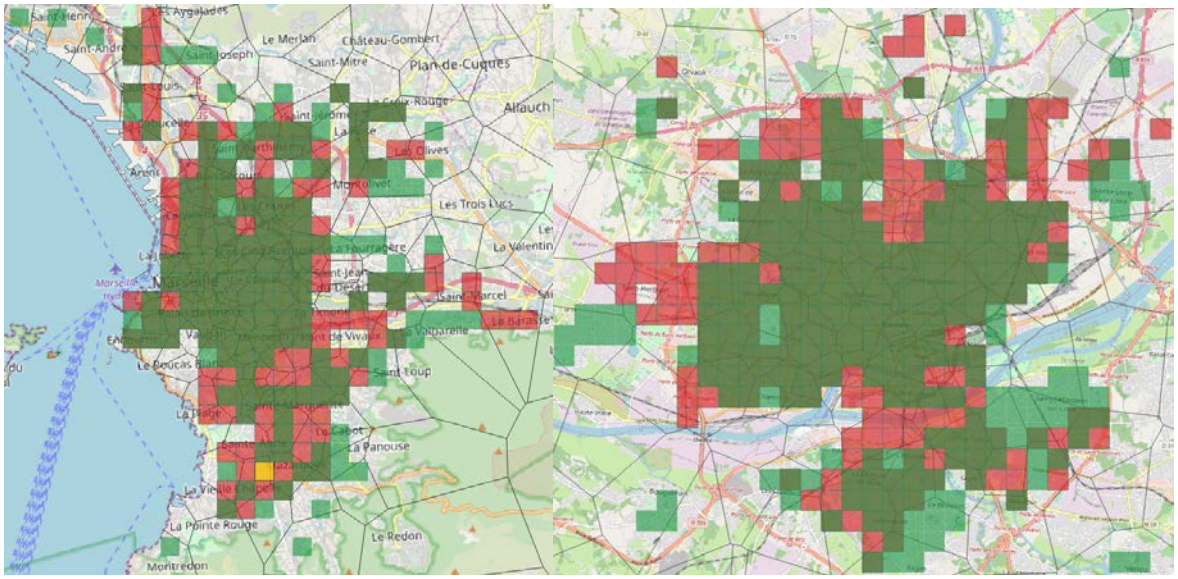


Figure 10: Comparison of the most populated tiles in Marseille (left) and Nants (right) from mobile phone data (red) and tax data (green)

We need to zoom in dense areas to evaluate at a smaller scale the accuracy of our home detection. This is why we densities in figure 9 in the urban area of Marseille. It appears that we miss some of local variations at that scale. Especially inside the larger Voronoï polygons in the periphery we eventually fail to observe some abrupt disparities. Besides the densest part is larger in our estimations than in the reference data.

To compare shape of the densest neighbourhoods we display on figure 10 the last two deciles of the distribution, in green for the tax data and in red for MPD. So when the estimation matches the reference data the tile is coloured in dark green. That makes more visible where the discrepancies are. For example we detect as a dense tile the train station probably because there is more people passing by and communicating in that place although there is very little residences built up. This is likely to be a bias from our method and CDR. Yet we also note that the hospital is considered as a dense tile in the estimation and not in the tax data, and that is because with MPD we measure were people really sleep and not where they declare their taxes. Common features appear also in Nantes urban area.

The prior effect

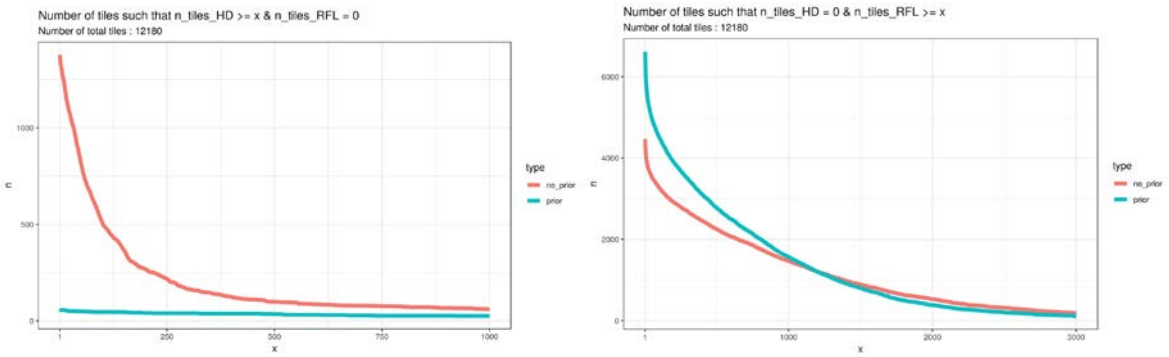


Figure 11: Distribution of the estimated population in each tile from mobile phone data with and without prior on land cover

Moreover we evaluate the effect of adding an a priori information on the places of residence from the land cover register. To describe what this information adds we display some distributions on figure 11.

The one on the left is the distribution of the number of people estimated to live in tiles where we know from the tax data that nobody lives in. It makes clear that with a prior we allocate almost no mobile phone user in tiles that are known to be empty. That comes from the fact that with information on land cover we weight to 0 all the tiles covered by forest or water for instance. The a priori information is a good tool to avoid this kind of obvious error.

The graph on the right hand side shows the distribution of the population of the tiles where we do not detect any home with mobile phone data although we know those tiles to contain some residence from the tax data. Those are populated tiles that we miss. In that case the advantage of the prior is less clear than in the former. Yet it still somehow mitigates that type of error as we observe that me miss less very populated tiles with the prior than without. Although we miss more of the sparsely populated tiles, it is less important mistake than missing out very dense tiles.

Limits and perspectives

Finally we conclude that from relatively sparse mobile phone data (CDR from 2007), using a prior on land cover is quite relevant and avoids making important mistakes in the home detection problem. Still we rely heavily on a poor model (Voronoï tessellation) on the coverage of antennas in the absence of additional information.
The methodology we used fits in the Referential Methodological Framework[6] proposed by Eurostat, which is an important tool to harmonize our procedures at the continental scale.Besides the modular description of the methodology leaves room for improvement if we have access to more data. The next steps should consist in locating the events on

[6] http://www.15th-tourism-stats-forum.com/pdf/Papers/S3/3_1_A_Reference_Methodological_Framework_for_processing_mobile_network_operatordata_for_official_statistics.pdf

an adaptive grid, with larger tiles in the areas where we have less accuracy in the location from the network and less people living. Another piece of work that should be conducted is using tax data as a prior and not only for validation, keeping in mind though that in this case it would be less easy to validate the results.
We hope to access signalling data in a near future, many more applications should therefore be possible, including day time population. We hope that this project may contribute to open some perspectives in using mobile phone data for official statistics in demonstrating how NSI can bring some relevant data in the process. A framework to fuse many data sources is a rich potential for defining FUA encompassing many dimensions of cities.

3.3. Experiences of using mobile phone data in Germany

Introduction

The German mobile communications market currently consists of the three providers Deutsche Telekom, Vodafone and Telefónica with a respective market share of one-third each. In order to research the use of mobile phone data for official statistics, the Federal Statistical Office of Germany (Destatis) entered into a cooperation with T-Systems International GmbH and Motionlogic GmbH (both wholly-owned subsidiaries of Deutsche Telekom AG) in September 2017. The conceptual designs of the planned feasibility studies were developed in consultation with the Federal Network Agency, the Federal Commissioner for Data Protection and Freedom of Information (BfDI), and T-Systems[7].
This report presents the analysis and the results of the project on the use of mobile phone data in official statistics. The project is divided in two parts. The first part examines to what extent mobile phone data could be used to represent the day time and night time population. This part of the project is mainly limited to the federal state of North Rhine-Westphalia (NRW). In the second part the unemployment rate of the Labour Force Survey (LFS) is estimated with static, aggregated and anonymised mobile phone data for Functional Urban Areas (FUAs) by using small area estimation. This Project deals with mobile phone data for the whole territory of Germany.

Data

The data record currently available to Destatis contains mobile phone activities of Deutsche Telekom customers. Due to data protection rules, Destatis only receives anonymised, aggregated data on mobile phone activities from T-Systems, which correspond to signalling data. A mobile phone activity is defined as a length of stay at a location without movement, with all signalling data being evaluated, i.e. phone calls, text messages and data connections[8]. Furthermore, signalling data is produced automatically and only registers the location of the cell tower to which a mobile device is connected at a specific time.

The first set of data contains mobile phone activities in North Rhine-Westphalia for a statistical week that consists of 24-hour days which were selected from the months of April, May and September 2017[9]. The mobile phone activities comprise the average activities on the weekdays selected. The weekdays are categorized in five types of days, Monday, Friday, Saturday, Sunday and the days from Tuesday to Thursday being grouped together. This data is also associated with some additional information such as socio-demographic characteristics of mobile phone users, like age group, gender and nationality proportions of the SIM cards per grid cell. This includes contract and prepay customers. The characteristics, however, are only available for contract customers. Only

[7] See Hadam, S. (2018)
[8] These events are known as Call Detail Records (CDR).
[9] School holidays of NRW and public holidays are excluded here.

values based on a minimum of 30 activities per grid cell[10] were transmitted to Destatis[11]. The grid cells conform with INSPIRE[12] and correspond to the grid cells of the 2011 Census Atlas[13].

The number of mobile phone activities depends on the location and number of cell towers in the various grid cells. Depending on the cell towers' location (rural or urban), their frequencies differ and, as a result, they are sometimes distributed unevenly across the regions. Consequently, an existing geometry may contain 5 to 20 cell towers. For that reason, some geometries are combined to ensure a minimum of 30 activities per grid cell[14]. Furthermore, the length of stay at a location without movement determines the number of mobile phone activities, whereby long mobile phone activities corresponding to the length of the dwell time are counted and included in the data record, while short mobile phone activities are left out of account. The dwell time is defined as the duration of stay of a mobile device at a location or in a grid cell without any movement. The dwell time in the available data record is two hours in order to filter out short mobile phone activities, which result, for instance, from quick movements between the grid cells.

The second set of data contains mobile phone activities of Deutsche Telekom customers from all over Germany for a statistical Sunday evening that is from 8 p.m. to 11 pm. There were eight Sunday evenings selected from the months of April, June and July 2018. [15,16] The mobile phone activities comprise the average activities on Sunday evening. This set of data also contains supplementary information on socio-demographic characteristics of mobile phone users.[17] In compliance with data protection rules, the mobile phone activities were again anonymised[18] and aggregated. Only values based on a minimum of 30 activities per geometry were transmitted to Destatis. The geometries used here correspond to the municipal and district level for Germany.[19]

Mapping nighttime (residential) and daytime population[20]

The aim of the first analysis was to provide a valid picture of the resident and daytime population. To this end, the total number of active SIM cards per grid cell was identified in the aggregated data. The population figures of the 2011 census are used as a benchmark to check the representativity of the data. First, the figures were used to determine the correlation between mobile phone activities and census values by type and time of day for NRW, as shown below in Figure 12. Overall, the values reveal a high correlation of 0.8 between mobile phone activities and census values throughout Saturday and Sunday. On weekdays, the correlation declines to less than 0.7 between 5 a.m. and 4 p.m., which indicates significant differences in the resident population according to the 2011 census and according to the location of mobile phone activities within the given period, which refers to the daytime population. From this figure it can be seen that aggregated mobile phone data enables a distinction to be made between the daytime and the resident population.

---

[10] A grid cell is a square-shaped geographical unit of varying or uniform grid width, with clear cell and spatial reference. Grid cells do not follow national administrative boundaries but represent a suitable territorial delineation. Several cells together form a grid as a large-scale reference system.
[11] The data provider Motionlogic has no access to individual or raw data, either.
[12] Infrastructure for Spatial Information in Europe
[13] For more information on the Census Atlas please see https://atlas.zensus2011.de/, accessed on 29 January 2019.
[14] See **Figure 29** for the original geometry. The grid cells vary from 500 x 500 to 8000 x 8000 meters.
[15] This period was chosen because of the high correlation between mobile phone activities and population figures of the census. See Figure 12.
[16] School holidays and public holidays are excluded here, too.
[17] The characteristics, however, are only available for contract customers, too.
[18] Telekom AG uses a procedure agreed with the Federal Commissioner for Data Protection and Freedom of Information (BfDI) to anonymise the data.
[19] Since this dataset was created mainly for the small area application, the characteristics in the data and geometries were changed or adapted accordingly compared to the first dataset described above.
[20] See also Hadam S. (2018).

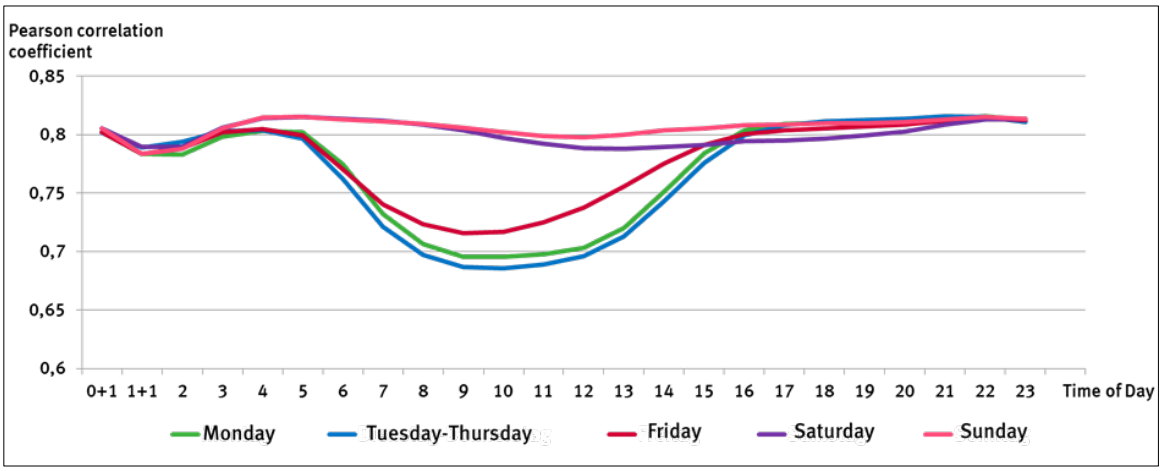


Figure 12: Pearson correlation coefficients referring to the census values and the mobile phone data, by statistical day and period.

Since the mobile phone activities are available on grid cells of different sizes[21], they are converted to the preferred geometry by using a kernel density estimation to enable further comparisons with the population figures of the census. As only aggregated and classified data in a specific geometry are available, a kernel density estimation for heaped and rounded data proposed by Gross et al. (2017) is used. For practical use, the R package Kernelheaping for the statistical programming language R was developed[22]. As Gross (2018) indicates "this package implements a partly Bayesian algorithm treating the true unknown values as additional parameters and estimates the rounding parameters to give a corrected kernel density estimate[…]" using the Stochastic Expectation-Maximization (SEM) algorithm proposed by Gross et al. (2017)[23]. By using the command dshapebivr we can estimate bivariate kernel density estimation for data classified in polygons or shapes. This methodology estimates the kernel density of the mobile activities based on the on the underlying grid cells and creates a uniform map with the hotspots of the mobile activities. Figure 13 clearly shows that on the one hand, large cities and urban areas in NRW have a high kernel density and on the other hand, a distribution of the density and thus of the mobile activities in the course of the day indicates commuter flows.

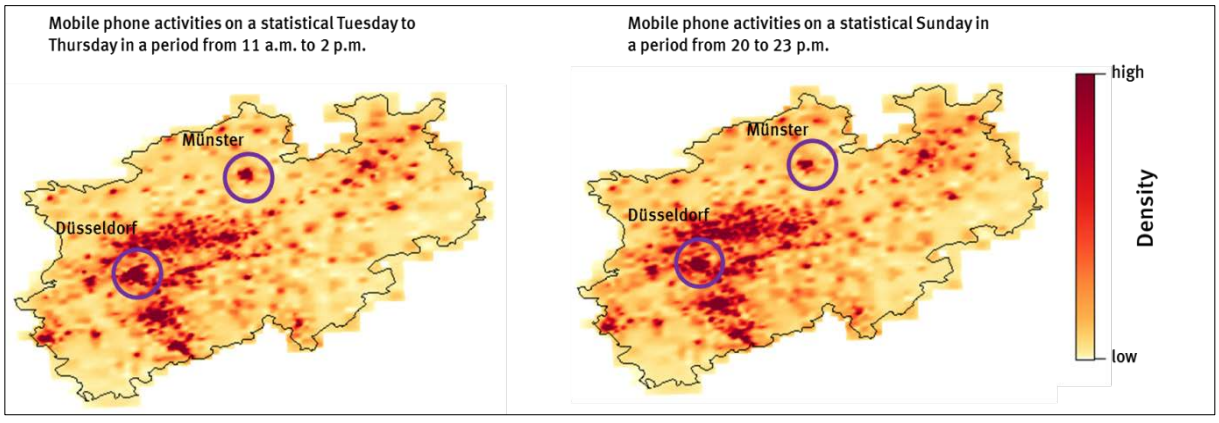


Figure 13: Kernel density estimates referring to mobile phone data, by a statistical Tuesday to Thursday (daytime population) and a statistical Sunday (nighttime, residential population).

[21] See again Figure 22 in the Annex 2.
[22] See Gross, M. (2018).
[23] See Gross, M. (2018).

By using the R command toOtherShape, the observations can be transferred into the desired form, since the function returns the count of the observations in a different shapefile. The advantage of using kernel density estimation is the redistribution of counts to any desired geometry, but the redisribution must be treated with caution since it becomes less accurate, the smaller the geometry becomes. In this case, a redistribution of the activities on the municipality and district level is recommended.

On the basis of the kernel density estimators from Figure 2, the kernel density estimates are converted into mobile phone counts to the district and municipality level and compared with the population figures of the 2011 census. Since a difference of 9 million counts between census values and mobile activities is observed, the latter is extrapolated with a correction factor, which is calculated from the sum of the population divided by the sum of the mobile phone data in the period under consideration.[24] This allows considering the current distribution of the entire German population on the basis of mobile phone data. The differences between the extrapolated mobile phone counts and population figures are presented in Figure 3 and Figure 14. If it is assumed that a good estimation of the resident population by using mobile phone data results from a distortion of maximum ±10% compared to the population figures from census, then a good estimation at the municipality level of 30% on Sunday evening and 32% on Tuesday to Thursday noon is observed.[25] In comparison, the estimation at district level is better. Here a good estimation of 49% on Sunday evening and 34% on Tuesday to Thursday noon is obtained. Since the activities are aggregated and have already been distributed to specific geometries, it is not advisable to include other data sources to distribute the aggregated activities.

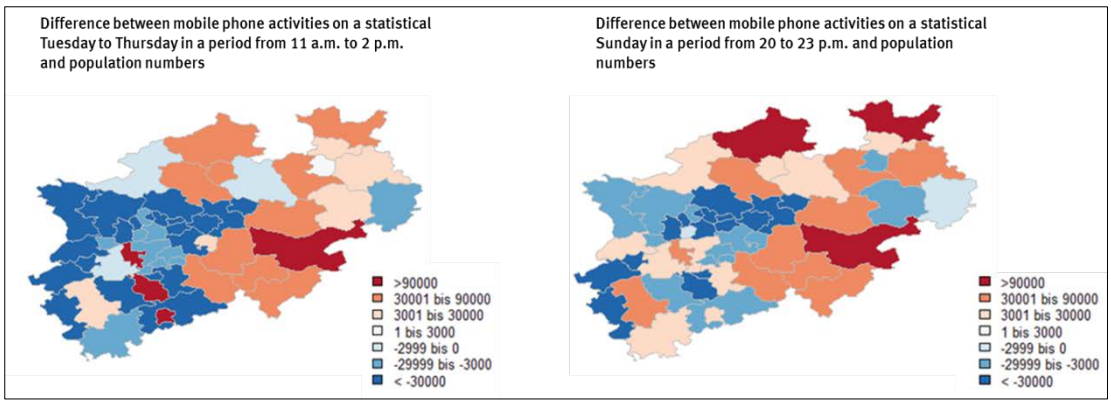
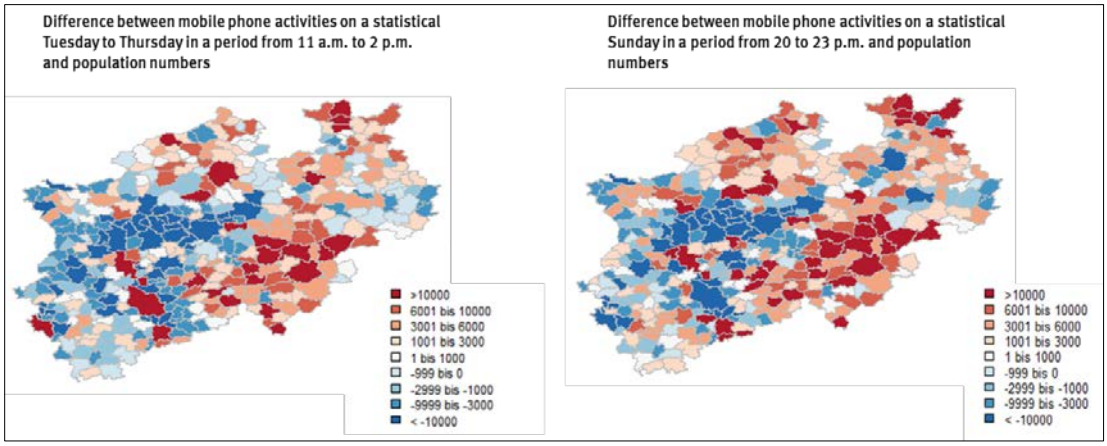


Figure 14: Difference between mobile phone activities and population at district level[26].



[24] This data has not been further modified or extrapolated and has been used in the original form as provided by Motionlogic.
[25] Noon is used here for the time span of 11 a.m. to 2 p.m.
[26] Blue areas: less mobile activity counts than population figures from the 2011 census
Red areas: more mobile activity counts than population figures from the 2011 census

Figure 15: Difference between mobile phone activities and population at municipal level [27].

As mentioned before, Figure 14 and Figure 15 show that the distortions increase as the geometries become smaller. The distortion at the municipality level is greater in NRW as well as in Germany as a whole compared to the district level[28,29]. Therefore, it is not advisable to make a comparison at grid level of, for example, 1 x 1 kilometre. It is also point out that the mobile phone data of the Deutsche Telekom show spatial distortions in terms of their market share. It is known that Deutsche Telekom is more distorted in rural areas than in urban ones. This can be seen in Figure 3, Figure 4 and Figure 24, as the population numbers in agglomerations or heavily urbanized areas are clearly underestimated by mobile activities.

The analysed data allows the distinction between daytime and nighttime population. Using the current available data, it is however not yet possible to describe the commuting patterns as such, i.e. the movement in space. Nevertheless, the results allow deducing commuter regions[30].

The results show that, to some extent, the data could provide a good picture of the population. The differences observed between the population figures based on mobile phone data and those based on census values, may partly be explained by the time difference between the mobile phone data from summer/autumn 2017 and the census data from 2011, but they may also be a result of the extrapolation method used by Motionlogic. The extrapolation is based on Deutsche Telekom's regional market shares across the mobile communications market in the whole territory of Germany. The regional market shares are determined using postal codes. The mobile phone activities are weighted based on the location or postal code of a mobile device's first signal at the very beginning of its activity chain. This means that the extrapolation of all the activities of a mobile device counted throughout a day depends only on the market share of the postal code of the first activity counted. In addition, the figures are only extrapolated to the total number of mobile phone users. At present, roughly 80 percent of the German population owns a mobile phone.[31] Consequently, 20 percent of the population is not considered in the extrapolation. The issue of biases and selectivity will be discussed in future papers.

Estimating unemployment rate from LFS using mobile phone data

The second part of this report will examine whether and to what extent various Labour Force Survey (LFS) indicators can be estimated at the level of Functional Urban Areas (FUAs)[32] using solely mobile phone data. Therefore, an equivalent dataset as explained in chapter 2, page 3 is used. Mobile phone data are used since they have a lot of

---

[27] Blue areas: less mobile phone data than population figures from the 2011 census

Red areas: more mobile phone data than population figures from the census

[28] See Figure 30, Figure 31, Figure 32 and Figure 33 in the Annex 2. Figure 32 and Figure 33 show the relative difference between activity counts and population figures.

[29] In the case of mobile activities throughout Germany, no kernel density estimation was carried out because the data were already available in the desired geometry.

[30] They can be identified based on the territorial units with above and below average population densities during the day, especially in the middle of the day compared with the morning and evening hours.

[31] Cf. https://de.statista.com/statistik/daten/studie/585883/umfrage/anteil-der-smartphone-nutzer-in deutschland/, accessed on 8 June 2017.

[32] The FUAs are composed of city cores and their commuting zones. City cores are urban centres with at least 50,000 inhabitants. The commuting zones contain the surrounding travel-to-work areas of the city cores where at least 15 percent of their employed residents are working in this city. Germany has in total 208 units, which are relevant for determining FUAs (see Figure 34 in the Annex 2). The FUAs are composed of 125 city cores, and 83 commuting areas of these cores. Especially through the agglomeration of the cities in North-Rhine Westphalia, e.g. the Ruhr area, there exist are more city cores than commuting zones.

advantages compared to survey data. They are real-time data and with finer spatial resolution, which means that the latest mobile phone data can be obtained, or more exactly activities, at the spatial resolution of cities or municipalities. According to the model described by Schmid et al. (2017) and as part of collaboration between Destatis and Freie Universität Berlin, the intention is to link mobile phone data to the unemployment rate, which will be estimated for small areas using a small area estimation method. The basic question is where and to what extent small area estimation could be used in combination of mobile phone and official statistical data.

In this part a dataset with aggregated mobile phone activities for whole Germany based on the municipality level is used, since the smallest units in the FUAs are municipalities (see Figure 34). Thus, the mobile phone activities can be aggregated up to the FUA level. Because of the high correlation between population figures of the 2011 census and mobile phone counts (see Figure 12) on the weekend and especially in the evening the time period from 8 p.m. to 11 p.m. of the average of eight Sundays of the months April, June and July 2018 without school holidays or public holidays is focused. This time slot is used since the LFS surveys questioned the resident population, which means that individual information of the home area and the working situation of each observation in the LFS is used.

Basically, LFS indicators are published on NUTS 2 level[33], which contains around 800,000 to 3 million inhabitants. The survey was not designed to produce reliable estimates on smaller level like NUTS 3 or municipalities due to small sample sizes. Hence, traditional design-based estimators are not appropriate and would lead to high sampling variability. The reference year 2016 with an overall sample size of 725,829 observations is considered. Since the FUA level is evaluated the sample size decreases to 533,356 observations of individuals.

At the beginning, a direct estimation is applied. This includes a direct estimation of weighted means of the variable of interest, which is solely based on the sampling and survey design of the LFS. Therefore, a Horvitz-Thompson estimator is used. The direct estimator corresponds to a weighted mean of the variable of interest for each FUA i and is defined as follows

$$\hat{\theta}_i^{direct} = \frac{\sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{j=1}^{n_i} w_{ij}}$$

where y_{ij} is the dependent variable and w_{ij} is the sampling weight of unit j in area i. Considering the one-stage clustered sample or area sample, one gets an unbiased estimator for the LFS design for any breakdown. The information of the direct estimator to apply small area estimation is needed.

Secondly, the results in relation to the coefficients of variation (CVs) are evaluated[34]. The CV is the ratio of the standard deviation to the mean. The higher the coefficient of variation is, the greater is the level of dispersion around the mean. The values of the coefficients in Figure 16 increase drastically, and achieve in some cases even a CV of about 100 percent. For one area in the female's case one also get an out-of-sample domain, which means that there are no observations from LFS to directly estimate the unemployment rate for this area.

[33] The NUTS levels are harmonized geometries in Europe to compare specific areas with each other. In total, Germany has 38 NUTS 2 regions and 402 NUTS 3 regions. https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:02003R1059-20180118&from=DE, accessed on 25 November 2018.
[34] An acceptable CV is about 20%.

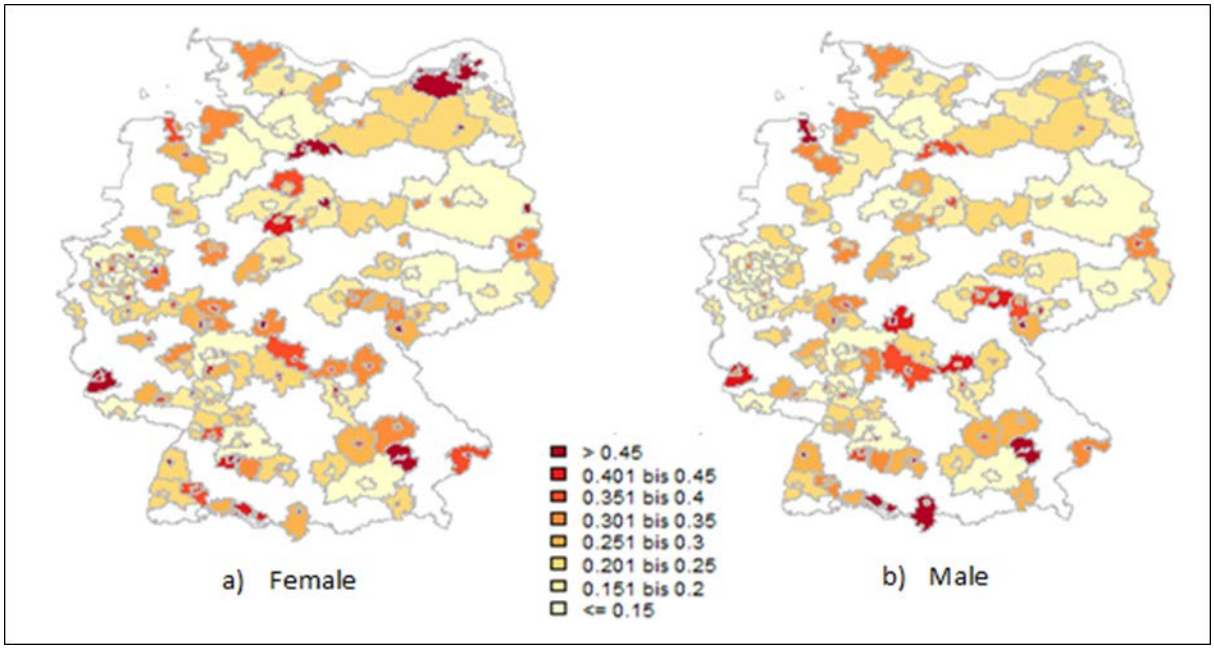


Figure 16: Coefficient of variations (CVs) of the unemployment rate in the females and males' model by gender on FUAs.

To obtain reliable estimates for the FUAs, a small area estimation based on the model of Schmid et al. (2017) is used. Therefore, in addition to the use of LFS information, alternative sources of passively collected mobile phone data are used for small area estimation. The main idea for this application is to use anonymised aggregated mobile phone data of the Deutsche Telekom as auxiliary variables to estimate LFS indicators for FUAs. The applied model corresponds to the Fay-Herriot estimator proposed by Fay and Herriot (1979), which is an area level model that links direct estimates with area-level covariates. The modification is an integrated *inverse sine transformation* to produce estimates in a particular range. Furthermore, area level models require auxiliary information at area level, which should be available for the sampled and out-of-sample domains.

At the first stage a direct estimator is estimated. At the second stage, the so-called linkage model links the direct estimator with the auxiliary variables. This model links the known indicator θ_i with covariate information x_i in linear relation. By combining both models the following area level linear mixed model is obtained

$$\hat{\theta}_i^{direct} = \theta_i + \varepsilon_i = x_i^T\beta + u_i + \varepsilon_i,$$

where the random effects $u_i \sim N(0, \sigma_u^2)$ are identically independently normally distributed and the sampling errors $\varepsilon_i \sim N(0, \sigma_{\varepsilon_i}^2)$ are independently normally distributed. The variables x_i are the mobile phone covariates. The EBLUP under the Fay-Herriot (FH) model is obtained by

$$\hat{\theta}_i^{FH} = x_i^T\hat{\beta} + \hat{u}_i + \varepsilon_i = \gamma_i\hat{\theta}_i^{direct} + (1-\gamma_i)x_i^T\hat{\beta},$$

where $\gamma_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2+\sigma_{\varepsilon_i}^2}$ denotes the shrinkage factor for area i. Since the unemployment rate is a percentage variable and therefore located in an interval between 0 and 1, the dependent variable must be transformed to obtain FH estimators in that particular range.

Therefore, an inverse sine transformation is used like in Schmid et al. (2017)[35]. By doing this, the following transformed FH estimator is obtained

$$\hat{\theta}_i^{FH,trans} = f^{-1}\left(\hat{\theta}_i^{FH}\right) = \sin^2(\hat{\theta}_i^{FH})$$

In order to ensure the internal consistency of the estimators, a benchmark method is applied. The benchmark approach of Datta et al. (2011) is used, which assumes the following

$$\sum_{i=1}^{m} \xi_i \hat{\theta}_i^{FH,bench} = \tau$$

where $\xi_i = \frac{N_i}{N}$ is the ratio of the population size in each region divided by the total population size. The benchmarked transformed FH estimator can be expressed as

$$\hat{\theta}_i^{FH,trans,bench} = \hat{\theta}_i^{FH,trans} + \left(\sum_{i=1}^{I} \frac{\xi_i^2}{\Phi_i}\right)^{-1} \left(\tau - \sum_{i=1}^{I} \xi_i \hat{\theta}_i^{FH,bench}\right) \frac{\xi_i}{\Phi_i}$$

After applying the model described above, the following results in Figure 17 are obtained. Figure 17 shows the regional distribution of benchmarked transformed FH estimates for all FUAs by gender. It can be seen that the unemployment rate of men is in general higher than that of women. One also obtains an estimate of the unemployment rate for the original out-of-sample domain in the females model. In addition, one receives on average higher unemployment rates in the city cores than in the surrounding travel to work areas.

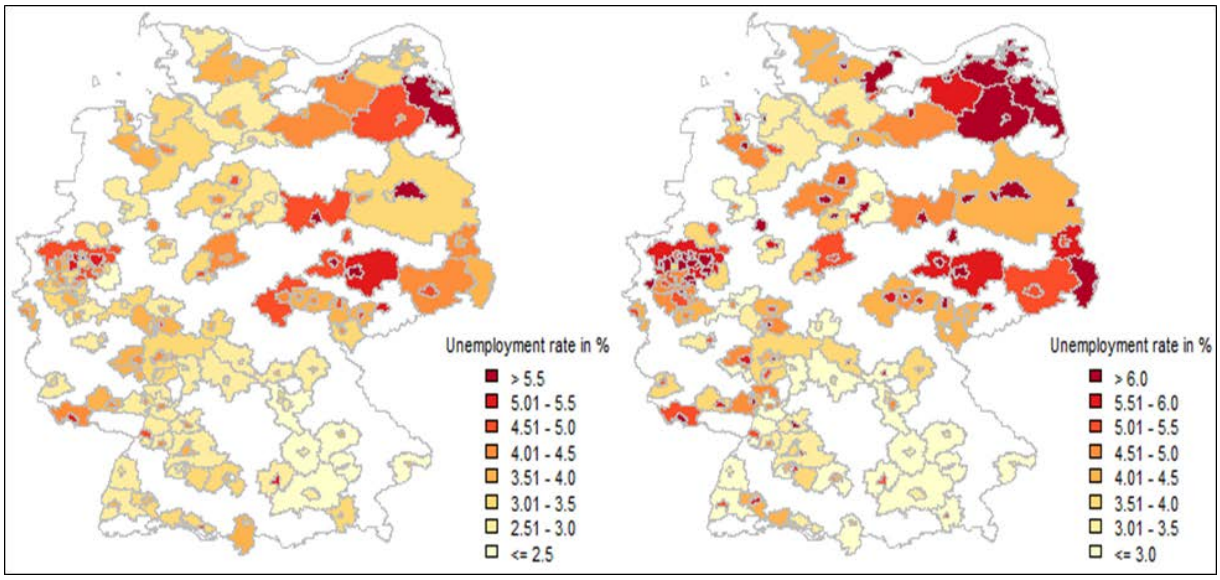


Figure 17: Unemployment rate by gender on FUAs. Benchmarked transformed FH estimator by gender: Females model (left) and males model (right).

Table 2 reports the distribution of the estimated unemployment rates by the direct, the transformed FH and the benchmarked transformed FH estimates by gender for the FUAs. Furthermore, the direct estimators by gender on the national level as benchmarks are used, which are similar to the official national values. Thus, the benchmarking ensures the consistency with the national estimates.

[35] The steps of the transformation are defined in Schmid et al. (2017).

Table 2: Distribution of the females and males unemployment rate.

Gender	Estimator	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
Female	Direct	0.004447	0.025640	0.035640	0.039110	0.047700	0.171500
	FH Trans	0.02034	0.03103	0.03595	0.03695	0.04165	0.07021
	FH Bench	0.02059	0.03140	0.03639	0.03740	0.04215	0.07106
Male	Direct	0.005276	0.031950	0.047930	0.049800	0.063980	0.112600
	FH Trans	0.01265	0.03773	0.04775	0.04766	0.05708	0.10760
	FH Bench	0.01290	0.03847	0.04869	0.04860	0.05820	0.10971

The benchmarked results are somewhat higher in both models than the transformed FH estimates. This is because adjusting the underestimation of the aggregated estimates in the males and females model has a visible effect on the benchmarked estimates to ensure internal consistency. This adjustment results from the fact that the mean transformed FH estimators are in both models lower than the estimated national mean of 3.92% in the females and 4.85% in the males model.

To assess the described results with regard to a possible gain in accuracy, the MSE is used. For the MSE estimation an approximation by Jack knife method is used. By obtaining an MSE it is possible to estimate the CV of the transformed Fay-Herriot model and compare it with the CV of the direct estimation, to determine whether the FH method generates a gain in accuracy. Figure 18 shows that the transformed FH estimator in the females model is more accurate than the direct estimator for the mean of unemployment rate for FUAs, where the red line represents the acceptable CV of 20%. The gain in accuracy is especially larger in the females model than in the males model. Nonetheless, the CVs received are still higher than the limit of 20%. This may be due to the low explanation of the mobile phone data in the model. But from these results it can be conclude that the FH approach helps to receive more reliable results.

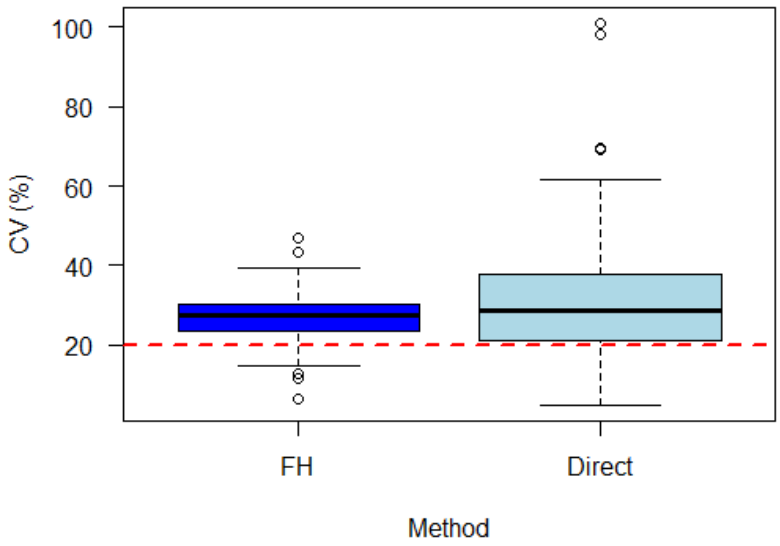
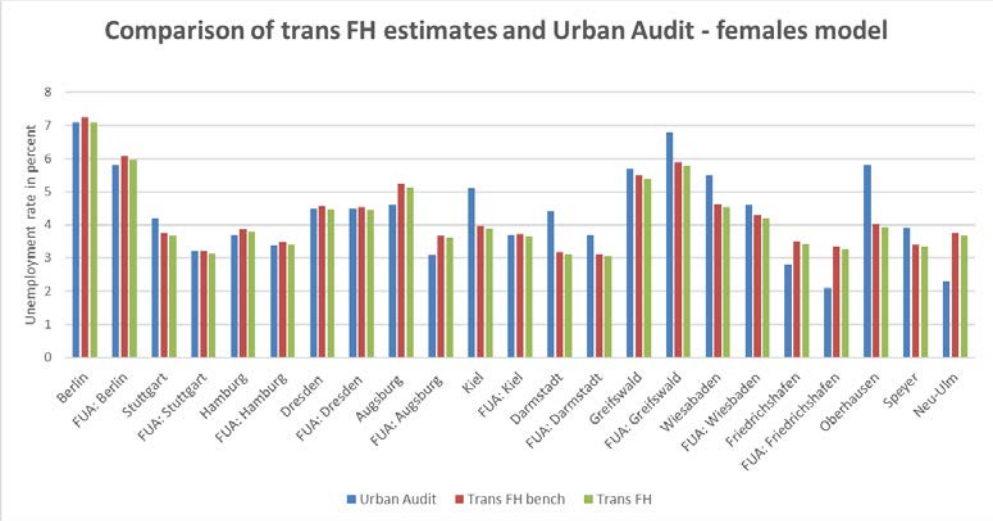


Figure 18: Distribution of CVs in the female model for the transformed FH and the direct estimator.

For the validation of the results, the small area estimators will be compared with the published indicators of the Urban Audit[36]. The Urban Audit is described as a "modelling based on administrative register and sample survey", which uses several register data sources as benchmark and for extrapolation[37,38].

Figure 8 shows the estimated unemployment rates by using mobile phone covariates and the published ones from Urban Audit for selected city cores and FUAs. The urban areas include a set of metropolitan areas, medium sized urban areas and small urban areas as well as their FUA if available.



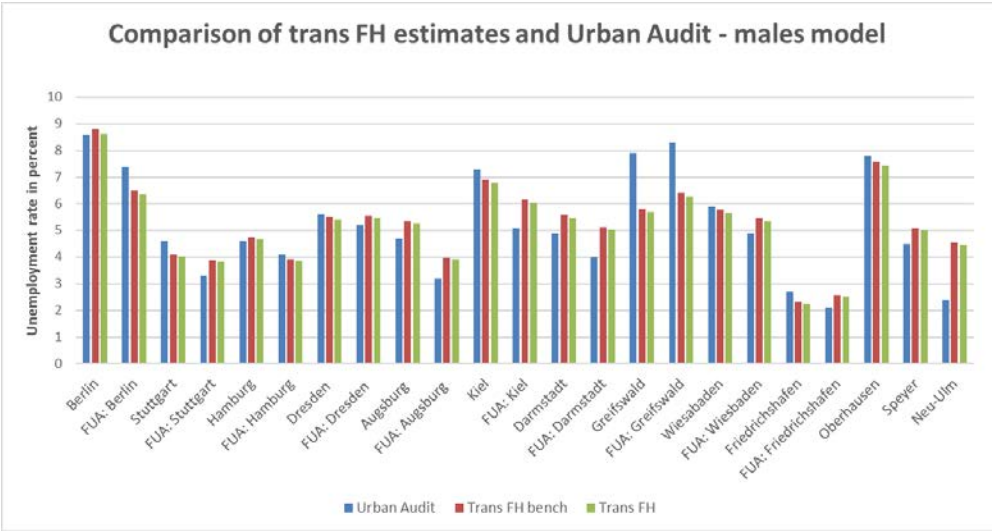


Figure 19: Comparison of estimated transformed FH and published Urban Audit indicators.

On Figure 19 the FH estimates are similar compared to the Urban Audit indicators. Furthermore, the benchmarked FH estimates are always slightly higher than those without benchmark approach. This is due to adjusting the underestimation of the

[36] See https://ec.europa.eu/eurostat/cache/metadata/EN/urb_esms_de.htm, accessed on 09 November 2018.
[37] For comparison purposes, the age groups considered in the LFS observation must also be changed. The LFS reports unemployment rates by the ILO definition for a population aged 15 to 74. However, the rate by the Urban Audit actually refers to a population aged 15 to 64.
[38] To assess the quality of the model-based estimates with the direct estimates, a goodness-of-t test proposed by Brown et al. (2001) can be applied. The null hypothesis of the test assumes that the model-based estimates do not differ significantly from the direct estimates. In this case the results of the test reject the null hypothesis in both models.

aggregated transformed FH estimates to ensure internal consistency. In average, the benchmarked transformed FH estimators work in both models better than the estimators without benchmark. Furthermore, the males model is somewhat better than the females, where the unemployment rates compared to the Urban Audit with the transformed FH bench estimators are underestimated only by 0.02 percentage points and by 0.12 percentage points with the transformed FH estimators. In the females model the benchmarked transformed FH estimators underestimate the rate compared to the Urban Audit indicators by 0.42 percentage points, whereas the transformed FH estimators underestimate the Urban Audit rates by 0.50 percentage points in average.

Nevertheless, this comparison should be treated with caution. In addition to the small area estimators, the Urban Audit indicators were also estimated using additional data sources and do not result from a full survey.

Conclusions on experiences of using mobile phone data in Germany

The results in the first part of this report show that, to some extent, the mobile phone data available could provide a good picture of the population. The data studied allows distinguishing between daytime and nighttime population. Using the data currently available, it is however not yet possible to describe the commuting patterns as such, i.e. the movement in space. The differences observed between the population figures based on mobile phone data and those based on census values may partly be explained by the time difference between the mobile phone data from summer/autumn 2017 and the census data from 2011, but they may also result from the extrapolation method used by Motionlogic. Nevertheless, with the data it is possible to determine the distribution of the population at a given time. For more accurate information and representativity, mobile phone data from all mobile network providers in Germany is required.

The second part of the report showed that it is possible to estimate unemployment rate by using aggregated and anonymised mobile phone data at spatially disaggregated level and obtain reliable results for FUAs. The results show a gain in accuracy compared to the direct estimators. By using the transformed Fay-Herriot model the confidence intervals and the CVs decrease and indicate a smaller uncertainty in the females and in the males model.

These projects indicate a lot of potential by using mobile phone data solely or in combination with other data sources. Therefore, unrestricted access for NSIs to this data source is highly relevant and of great added value.

3.4. Experiences of using anonymised aggregated mobile phone data in The Netherlands

Data access circumstances in the Netherlands

Statistics Netherlands (SN) has had previous experiences in using anonymised aggregated mobile phone data for statistical purposes. Back in 2013 SN worked with a mediating company who produced aggregated anonymised datasets based on data obtained from the mobile network operator (MNO) Vodafone Netherlands. In 2017 a new pilot based on aggregated anonymised datasets was initiated in collaboration with the MNO T-Mobile Netherlands. A collaboration was started on the basis of a contract for half a year with an option for extension under agreement of both parties, and this has been successfully continued. It allowed SN a unique opportunity to study, but not export the raw data, understand its structure and get acquainted with the technical infrastructure surrounding it. One of the aims of SN is to develop an open methodology to generate anonymised aggregates based on signalling data for statistical purposes which can be applied at another MNO's nationally and internationally.

Constructing the flow cube of devices

To estimate statistics on population flows, as well as day- and night-time population figures, a table with anonymised aggregated numbers of devices was produced. Transforming this table to one containing numbers of persons is done at a later stage (see Section 4.4.3). The initial table, which we call a flow cube of devices, has three axes called place of residence, place of presence and time. The element of this cube is a single number – estimated number of devices of which the owner resides in a given place of residence, but is present at a given point in time at a possibly different place.

For example, if the smallest resolution of locations used is districts and the smallest resolution of time used is hours, then a selection of rows from the flow cube could look as follows:

day	hour	place of residence	place of presence	n
20180604	12:00	district A	district A	439
20180604	12:00	district A	district B	18,57897
20180605	17:00	district B	district C	670,798

That is, on 4 of June 2018 at 12:00 approximately 439 devices which have district A as their place of residence are still in district A. At the same time 18,57897 devices which have district A as their place of residence are estimated to be located in another district B. The third row should be interpreted similarly. These device counts are typically decimal numbers due to the estimation methods explained below.

The primary data source used to construct this flow cube of devices was a table of 4G signalling data in the big data environment of the MNO. We would like to stress the fact that none of these data are shared with SN. Only the anonymised aggregated outcome data were delivered to SN to produce statistics. For every interaction, which relies on 4G technology between a device and an antenna in the MNO's network a record is generated in this table. This holds both for Dutch devices, and for ones roaming on the MNO's network. Some of these records are the result of deliberate activity by the device's user, such as calls, SMS messages and mobile data usage. Other records are generated passively by the device, for example, when it switches from one antenna to another. Neither type of records contains the actual contents of the communication. Records of the first type are converted by the MNO into so called CDR (Call Detail Records) data which is then used for billing. For statistical purposes, we treated the two types of records on equal footing. Having a larger supply of them, regardless of their type, is likely to lead to more reliable statistics. One can expect a smart phone to generate hundreds records in this table per day, of which fewer are generated at night. However, the number of records depends on many factors, such as brand of the device and operating system.

Each record in the MNO's 4G signalling data table consists of 321 variables. However, for our purpose, we only needed the following three variables:

- *imsi*: standing for International Mobile Subscriber Identity, which is an anonymous unique identifier of the device that generated the record. As a privacy measure, neither researchers from the MNO nor from SN have direct access to this identifier, as the actual IMSI is partially hashed to the variable imsi.
- *start_time*: a time-stamp signifying the start of the interaction of the device with the antenna which underlies the record,
- *e_cgi*: a unique identifier of the antenna on the MNO's network with which the device interacted.

The other variables may be useful in future research, but this requires substantial knowledge on the engineering aspects of the data.

The country of origin of a device can be extracted from the variable imsi, since the first three digits of the value of this variable are the country's mobile country code (MCC). This allows one to distinguish in particular records generated by Dutch devices from those generated by roaming devices.

The second data source provided by the MNO and used to construct the flow cube of devices is an up-to-date cell plan of the network, which contains various physical properties and settings of the antennas, including their geographical coordinates. This data is used as follows to approximate the locations of the devices that correspond to the imsi's.

Recall that two of the dimensions of the flow cube of devices are spatial: place of residence and place of presence. To fill this cube with numbers of devices therefore suggests that the following two methods are needed: given a record in the signalling data table

- determine the device's (that is, its owner's) place of residence,
- determine the device's place of presence.

Unfortunately, signalling data contains neither of these two pieces of information. Therefore, it is necessary to approximate these instead. A Bayesian model was proposed which estimates the probability $\mathbb{P}(x|a)$ that a device is present at a specific location x given its connection to some antenna a. We have in particular that summing over all locations x gives $\sum_x \mathbb{P}(x|a) = 1$, which reflects that a device (when connected to an antenna) is located with probability 1 somewhere within the MNO's network range. The model considers each device's connection individually and it is independent of the specific device and the moment of the connection. A detailed description of the construction of this probability distribution $\mathbb{P}(x|a)$ can be found in Tennekes et al. (2019).

However, a device can make more than one connection per hour, even with the same antenna. The number of connections per hour can moreover vary strongly between devices and different hours. Since we want to produce day- and night-time population and population flows statistics on an hourly basis it is necessary to account in various computations for this phenomenon. More precisely: given a device i and an hour h, the estimated fraction of h that i spent connected to a is an important quantity. We denote this fraction by $\mathbb{P}_{ih}(a)$ to suggest an alternative interpretation: it is the probability that i connected to a during h. It is calculated from the signalling data by

$$\mathbb{P}_{ih}(a) := \frac{\#\{\text{connections made by } i \text{ at } a \text{ during } h\}}{\#\{\text{connections made by } i \text{ during } h\}}.$$

If i made no connections during h then we define $\mathbb{P}_{ih}(a)$ to be 0. If i made at least one connection during h, then obviously, $\sum_a \mathbb{P}_{ih}(a) = 1$. Note that in this estimation step decimal numbers make their first appearance.

To estimate the place of presence the probability distribution of the device i at hour h is defined as

$$P_{ih}(x) := \sum_a \mathbb{P}_{ih}(a) \cdot \mathbb{P}(x|a),$$

where x stands for an arbitrary location. If i made no connections during h then $\sum_x P_{ih}(x) = 0$. If i made at least one connection during h, then obviously $\sum_x P_{ih}(x) = 1$.

To estimate the place of residence of a device i its *home antenna* a_i^{home} is determined first, meaning (roughly) the antenna to which i was connected to the longest during the observation period of five weeks. One might try to determine this starting by calculating for each antenna the number of hours connected to it by i over the entire period, then sorting the antennas by these hours and finally selecting the top ranked. Due to some computational barriers, this process is implemented slightly differently, as explained next.

First, for each antenna a the number of hours $h_{iw}^{\text{tot}}(a)$ connected to it by the device i was calculated separately per week w. This was done by summing the probabilities $\mathbb{P}_{ih}(a)$ associated to $24 \cdot 7$ hours in that week:

$$h_{iw}^{\text{tot}}(a) \coloneqq \sum_{h \in w} \mathbb{P}_{ih}(a).$$

To reduce the number of hours needed to store all these hour counts, for each device only the top ten antennas per week (in terms of connected hours) were preserved. The datasets per week were then combined into one dataset by summing the number of hours per device and antenna over the period of five weeks:

$$h_i^{\text{tot}}(a) \coloneqq \sum_{w} h_{iw}^{\text{tot}}(a).$$

It is assumed that the top ten per week will always include the antenna the device connected to the largest number of hours during the five week period. Finally, the home antenna a_i^{home} of device i is set to be the antenna which maximises the number of hours $h_i^{\text{tot}}(a)$:

$$a_i^{\text{home}} \coloneqq \arg\max_a h_i^{\text{tot}}(a).$$

Recall that we have already derived the probability distributions $\mathbb{P}(x|a)$. Then for each specific location x we derived the probability that the device i has x as its place of residence:

$$R_i(x) \coloneqq \mathbb{P}\left(x \middle| a_i^{\text{home}}\right).$$

Note that R_i is a probability mass function because $\mathbb{P}\left(\cdot \middle| a_i^{\text{home}}\right)$ is: summing over all locations x gives $\sum_x R_i(x) = 1$, which reflects that a device which has a home antenna has its place of residence somewhere within the MNO's network range.

In this way, every device is assigned a home antenna and a mass function of probable places of residence. However, it was found that about 25% of all devices present in the signalling data connected to their home antenna for less than 60 hours during the observation period of five weeks. The probable places of residence derived from such an antenna are considered unlikely to be near the true place of residence of the device's owner. Within this project we did not study these *wandering devices* closer and decided to discard them before further processing. From this point onwards in this document 'all devices' will stand for only those which crossed this threshold of 60 hours.

Having estimates of a device's place of residence and, at every hour, place of presence now gives all the components to build the following 2-dimensional matrix of device i for hour h:

$$RP_{ih}(x^{\text{r}}, x^{\text{p}}) \coloneqq R_i(x^{\text{r}}) \cdot P_{ih}(x^{\text{p}}).$$

It stores the probability that i has place of residence x^{r} (which is independent of h) and place of presence x^{p} during hour h. It is assumed that the probability distributions of place of residence R_i and place of presence P_{ih} are independent from each other. If i made at least one connection during h, then RP_{ih} is a joint probability mass function: we have

$$\sum_{x^{\mathrm{r}}} \sum_{x^{\mathrm{p}}} RP_{ih}(x^{\mathrm{r}}, x^{\mathrm{p}}) = 1.$$

If i made no connections during h, then the above double sum equals 0.

The total flow cube of devices RPT^{dev} is obtained by summing all matrices $RP_{ih}(x^{\mathrm{r}}, x^{\mathrm{p}})$ over all devices i for each hour h:

$$RPT^{\mathrm{dev}}(x^{\mathrm{r}}, x^{\mathrm{p}}, h) := \sum_{i} RP_{ih}(x^{\mathrm{r}}, x^{\mathrm{p}}).$$

All processing steps explained so far took place at the MNO. Even though the signalling data was already anonymous, in the sense that records do not contain the name, address or demographic data of the owners of the devices which generated them, to further reduce a possible risk of disclosure the elements of the flow cube of devices which are estimated numbers $RPT^{\mathrm{dev}}(x^{\mathrm{r}}, x^{\mathrm{p}}, h)$ of devices strictly lower than 15 were removed. The resulting filtered cube was then exported to Statistics Netherlands.

In our discussion up to now we have not specified the precise meaning of locations x. They can refer to grid cells, or administrative regions such as provinces, municipalities, their districts or neighbourhoods, depending on the level of spatial detail for which one wants to produce population flow statistics. The choice of size of grid cells and the level of administrative regions' detail is however hindered by:

- The spatial density of antennas belonging to the MNO's network. This density differs according to, among other factors, the level of urbanicity. Since a mobile network is optimised for the needs of its users, densely populated regions contain more antennas to ensure optimal service.
- The accuracy of the methods used to estimate the probabilities $\mathbb{P}(x|a)$.
- The mass lost from the cube by the threshold of 15 devices. A greater level of spatial detail namely implies that the cube will contain more cells, each of which is more likely to contain a lower number of devices.
- The market share of the MNO from which one obtained signalling data.

After considering the factors mentioned above it was decided to make aggregates at the MNO at the municipal level (specifically, the Dutch municipality definitions of 2017) to produce statistics for day- and night-time population as well as population flows. However, for the purpose of data analysis municipalities' districts were used.

This flow cube of devices can now be used to estimate the *incoming* and *outgoing flow* of devices for a given location x and hour h. By this we mean the number of devices entering or leaving x from other locations, possibly including x itself. They are computed respectively as follows:

$$F^{\mathrm{in,dev}}(x, h) := \sum_{x^{\mathrm{r}}} RPT^{\mathrm{dev}}(x^{\mathrm{r}}, x, h),$$

$$F^{\mathrm{out,dev}}(x, h) := \sum_{x^{\mathrm{p}}} RPT^{\mathrm{dev}}(x, x^{\mathrm{p}}, h).$$

In other words, the inflow is calculated by summing over all places of presence, while the outflow is calculated by summing over all places of residence.

Constructing the flow cube of persons

The elements of the flow cube RPT^{dev} of devices are estimates of numbers of devices. These figures differ from the corresponding numbers of persons for at least the following reasons:

- They represent merely counts of devices of the MNO(s) from which the signalling data is obtained. The people who communicate via different MNO's are therefore excluded.
- Not everyone owns a mobile phone, or if they do, carry their devices everywhere with them. This holds especially for young children and the elderly.
- Some people might carry multiple devices.

The cube of devices hence needs to be transformed or *calibrated* to a flow cube RPT^{pop} of persons (with axes and dimensions equal to those of RPT^{dev}). Before proceeding to the description of the calibration method, note that from such a cube incoming and outgoing flows of persons can be calculated in the exact same way as from the flow cube of devices. They are denoted by $F^{in,\text{pop}}(x,h)$ and $F^{\text{out,pop}}(x,h)$, respectively, for every location x and hour h.

For the calibration method applied by us an additional data source was needed. Every municipality in the Netherlands registers their residents and residents who moved abroad in the Dutch national *Personal Records Database (PRD)*. Additionally, people who are staying in the Netherlands legally for less than four months are able to have themselves be registered. The PRD is not completely accurate, but is the most reliable data source on Dutch residential population counts. Based on the PRD, Statistics Netherlands publishes periodically figures on the number of residents at several administrative levels, such as municipalities, districts and neighbourhoods. To reduce the risk of disclosure these public figures have rounding methods applied to them. We write $\text{Pop}(x)$ for this publically available number of residents of location x on 1 January 2017.

Our calibration method was based on the assumption that the flow cube RPT^{pop} of persons ought to satisfy the following combination of two equations for all locations x^{r} and x^{p} and hours h:

$$\frac{RPT^{\text{pop}}(x^{\text{r}},x^{\text{p}},h)}{F^{\text{out,pop}}(x^{\text{r}},h)} = \frac{RPT^{\text{dev}}(x^{\text{r}},x^{\text{p}},h)}{F^{\text{out,dev}}(x^{\text{r}},h)}$$

$$F^{\text{out,pop}}(x^{\text{r}},h) = \text{Pop}(x^{\text{r}}).$$

If we momentarily leave out the reference to the hour h for brevity, the first equation can be understood as the following example of a claim:

> Suppose that one tenth of the residents of x^{r} are present in x^{p}. Then also one tenth of the MNO's devices with place of residence x^{r} are present in x^{p}. The converse implication holds as well.

In other words, the first equation assumes a uniform presence of the MNO's devices in the flow of persons from location x^{r}. This homogeneity is not obvious, because, for example, the market share of the MNO in the flow from x^{r} to a location x_1^{p} might differ from that in the flow to another location x_2^{p}. Further research is needed to quantify the bias resulting from this.

The second equation results from the assumption

> The number of residents of x^{r} who are present in any of the regions x considered together equals the number of residents of x^{r}.

This assumption of course introduces a small error since the date (in our case 1 January 2017) for which the figure $\mathrm{Pop}(x^{\mathrm{r}})$ was determined is somewhat different from the observation period for which the data was obtained. A larger error is introduced if the set of regions $\{x\}$ does not also include locations abroad. Residents of x^{r} might be abroad during (part of) the observation period. Correcting for this misestimating would involve additional tourism or holiday statistics, which at this stage we did not attempt in this project.

The two equations above are easily seen to be equivalent to the single equation

$$RPT^{\mathrm{pop}}(x^{\mathrm{r}}, x^{\mathrm{p}}, h) = \frac{RPT^{\mathrm{dev}}(x^{\mathrm{r}}, x^{\mathrm{p}}, h)}{F^{\mathrm{out,dev}}(x^{\mathrm{r}}, h)} \cdot \mathrm{Pop}(x^{\mathrm{r}}).$$

Written in this way all known variables are present on the right hand side, while the variable on the left hand side is the one we wish to compute. The factor calibrating the estimate $RPT^{\mathrm{dev}}(x^{\mathrm{r}}, x^{\mathrm{p}}, h)$ of a number of devices to the estimate $RPT^{\mathrm{pop}}(x^{\mathrm{r}}, x^{\mathrm{p}}, h)$ of a number of persons is hence defined to be the fraction

$$\frac{\mathrm{Pop}(x^{\mathrm{r}})}{F^{\mathrm{out,dev}}(x^{\mathrm{r}}, h)}$$

and it is independent of the place of presence x^{prc}. This calibration method is best illustrated via an example.

Suppose the Netherlands is partitioned into three regions A, B and C, having residential population figures according to the PRD $\mathrm{Pop}(\mathrm{A}) = 5000$, $\mathrm{Pop}(\mathrm{B}) = 750$ and $\mathrm{Pop}(\mathrm{C}) = 1000$, respectively. Fix an hour h and suppose that the corresponding 2-dimensional slice $RPT^{\mathrm{dev}}(\cdot, \cdot, h)$ of the flow cube of devices looks as follows:

		Place of presence			
		A	B	C	Total
Place of residence	A	900	20	80	1000
	B	80	120	50	250
	C	70	40	140	250
	Total	1050	180	270	

Figure 20: The slice $RPT^{\mathrm{dev}}(\cdot, \cdot, h)$ at hour h of the flow cube of devices.

This table for example tells us that $RPT^{\mathrm{dev}}(A, B, h) = 20$. We also added the column totals to this table, that is, the incoming flow $F^{\mathrm{in,dev}}(\cdot, h)$, and the row totals $F^{\mathrm{out,dev}}(\cdot, h)$ for each of the regions A, B and C. The residential population figures $\mathrm{Pop}(\cdot)$ for the regions are higher than the row totals, by factors 5, 3 and 4, respectively. Correcting for this discrepancy via our method implies that the rows of the table above should be multiplied by these calibration factors. We then obtain the slice $RPT^{\mathrm{pop}}(\cdot, \cdot, h)$ at hour h of the flow cube of persons:

		Place of presence			
		A	B	C	Total
Place of residence	A	4500	100	400	5000
	B	240	360	150	750
	C	350	120	560	1000
	Total	5090	580	1110	

Figure 21: The slice $RPT^{\mathrm{pop}}(\cdot, \cdot, h)$ at hour h of the flow cube of persons.

The row totals now equal the residential population figures $\mathrm{Pop}(\cdot)$ and the column totals are the incoming flows of persons $F^{\mathrm{in,pop}}(\cdot, h)$.

Mapping day- time and night-time Dutch population and its flows

The flow cube of persons was used for mapping day- and night-time residential population as well as its flows. For this project, the cube was calibrated at a municipal level, as, at the moment, SN works with data from only one out of three MNOs and methodology still needs some improvement. The output is presented as dashboard (Figure 22) showing hourly Dutch population and its flows during a time period from May 28th until July 7th.

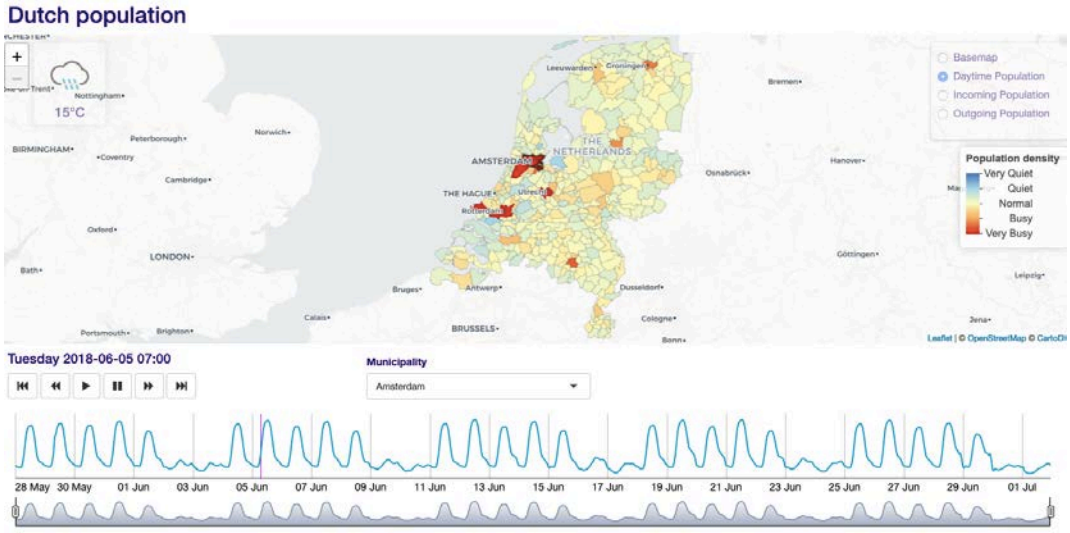


Figure 22: Dutch population in Amsterdam Tuesday 5 of June at 07:00

On Figure 22 graph shows a clear difference between Dutch population during the working days and the weekend in Amsterdam. On working days starting from around 07:00 on the city becomes busy with a pick around 12:00 and at about 23:00 it gets quiet (Figure 23). As we use data based on the Dutch devices only, we suspect that this has to do for the larger part with labour migration for other municipalities surrounding Amsterdam.

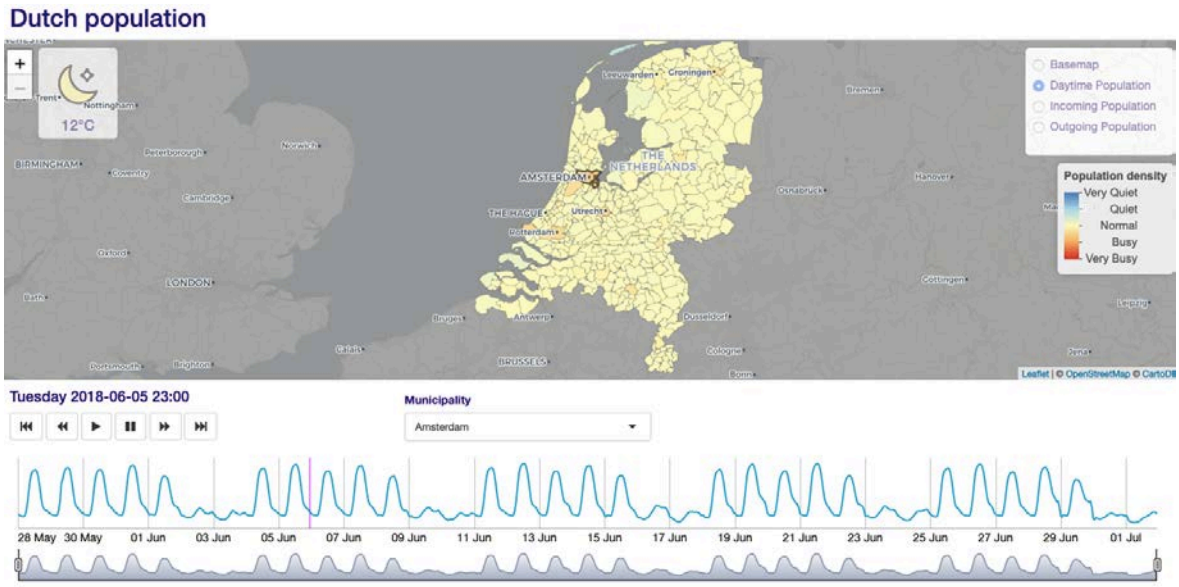


Figure 23: Dutch population in Amsterdam Tuesday 5 of June at 23:00

Analysing the data of other large municipalities Eindhoven, Rotterdam, Utrecht, which are comparable to Amsterdam for being known by its attractiveness for labour migrants familiar pattern can be seen (Figure 24).

Figure 24: Dutch population in Amsterdam, Eindhoven, Rotterdam and Utrecht

Analysing Dutch population numbers based on mobile phone data even from one MNO we could recognize some events and festivals within the county. One example can be “Pinkpop” which took place 15-17 of June in municipality of Landgraaf. On Figure 25 graph shows that starting from 15 of June it gets busy in Landgraaf with certain dynamics and after June 17th the population pattern gets back to normal.

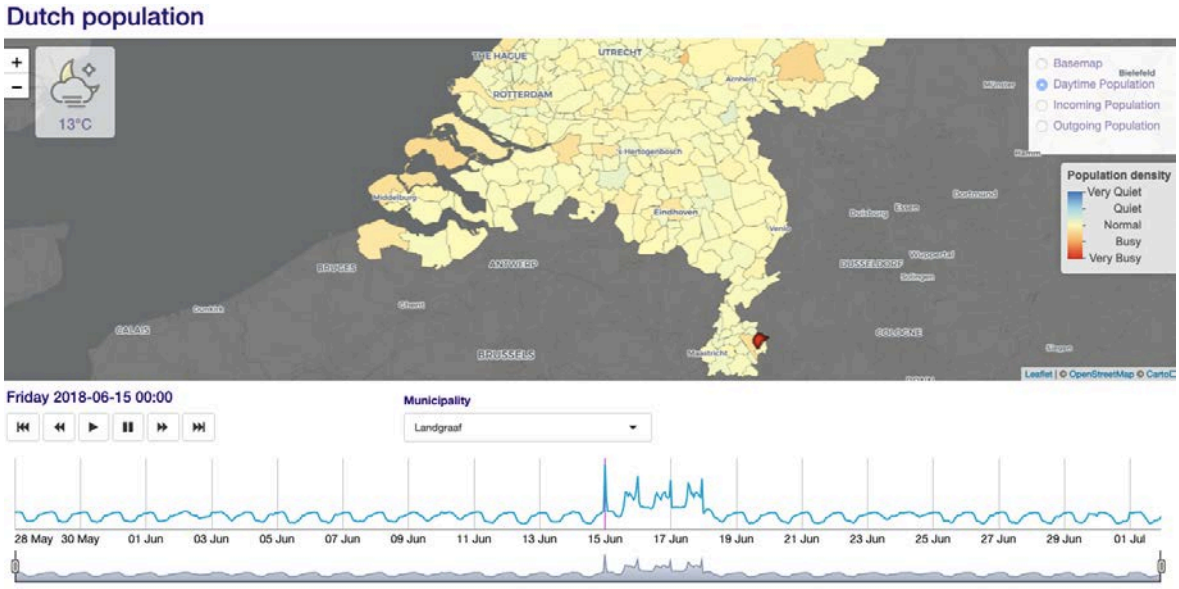


Figure 25: Dutch population in municipality Landgraaf

Another example of the event what can be recognized based on the information received from anonymised aggregated mobile phone data is TT Festival in Assen which took place

27-29 of June in municipality of Assen. As can be seen on Figure 26 starting from 27 of June population pattern in municipality changes. It becomes busy in Assen.

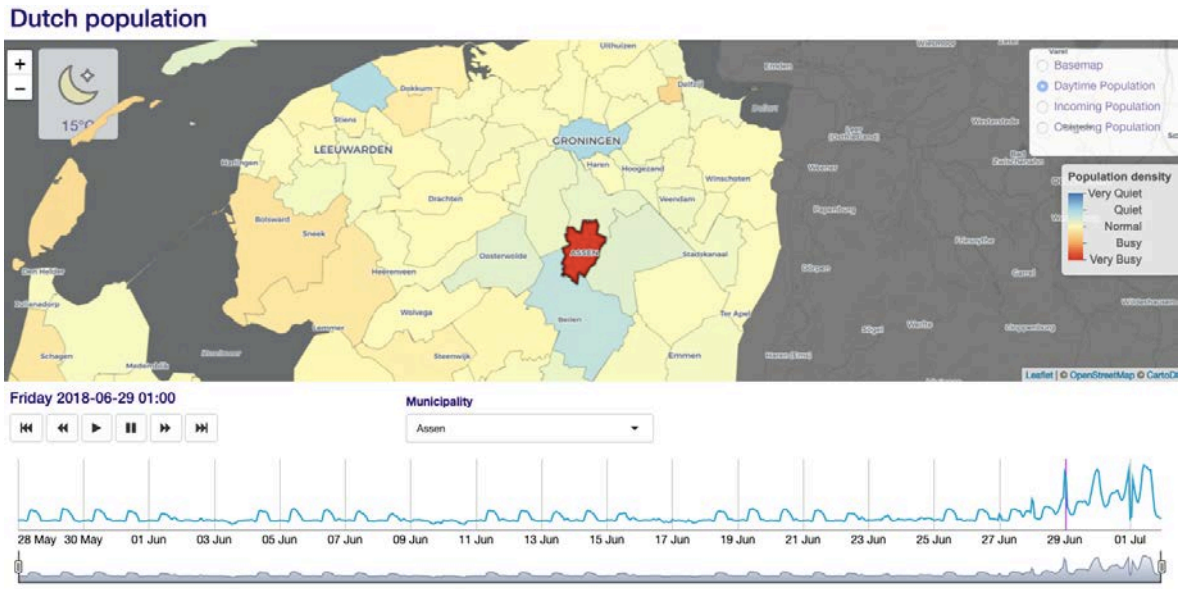


Figure 26: Dutch population in municipality Assen

The information based on anonymised aggregated mobile phone data allows to produce population flows as well. This means that it is possible to estimate the incoming and outgoing population. For example, we can explore where people are coming from who visited Assen on Friday the 29th of July at 01:00. Note that this is just a selection of the flow cube where we fixed time and place of presence. The visual representation of this may be seen on Figure 27.

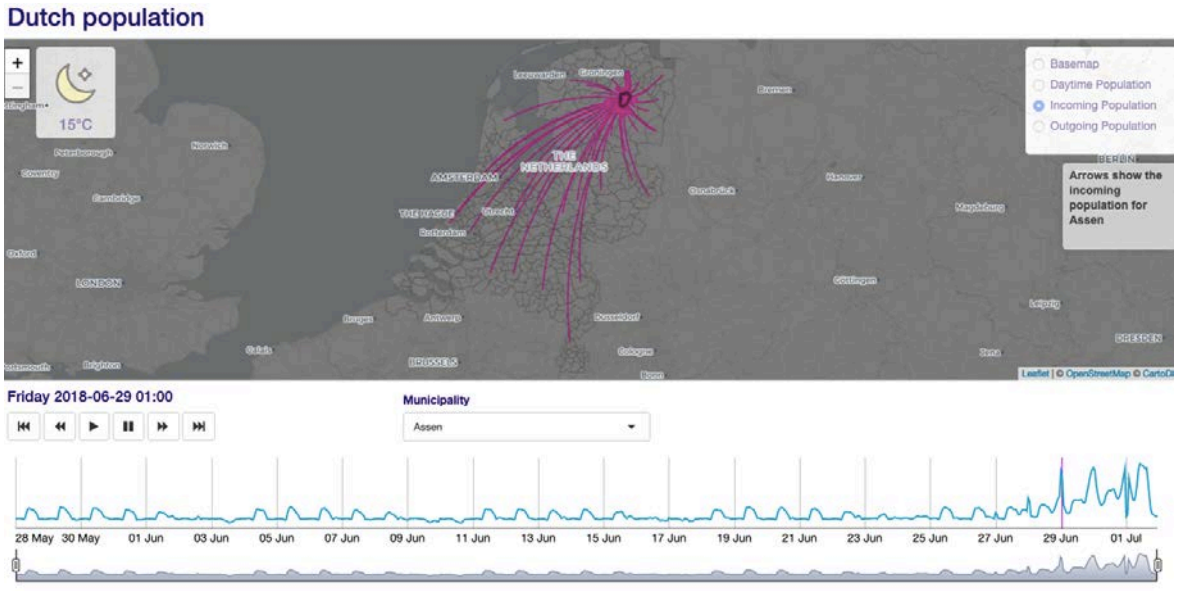


Figure 27: Incoming population of Assen on Friday 29 of June at 01:00

Respectively selecting in a column place of residence municipality of Assen, column place of presence will identify estimated number of visitors from Assen in other municipalities. The visual representation of this may be seen on Figure 28.

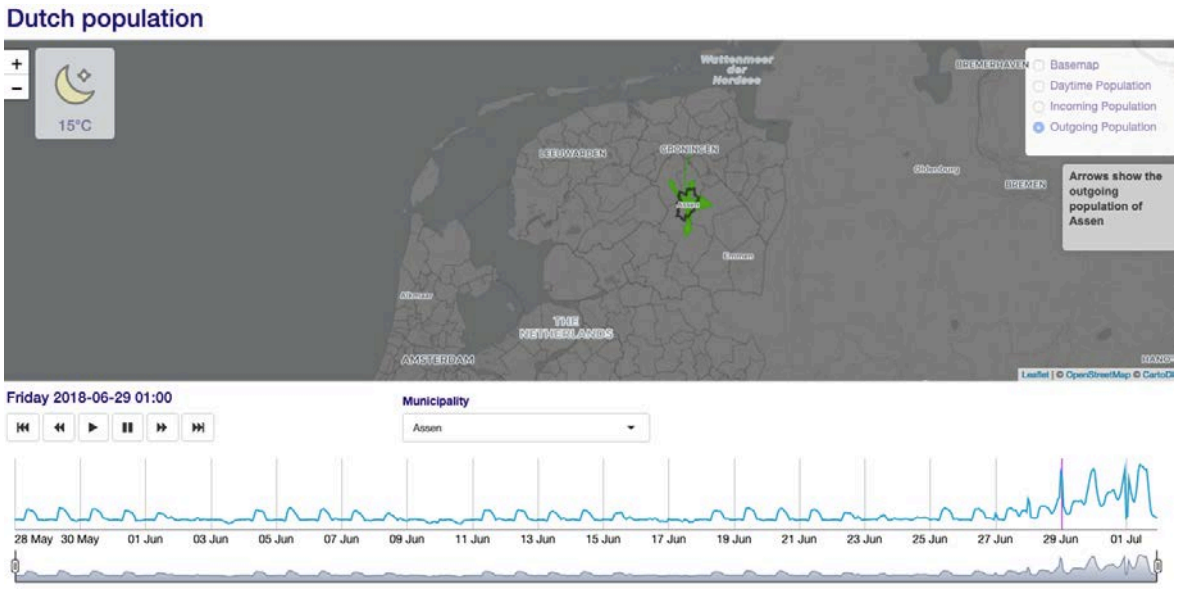


Figure 28: Outgoing population of Assen on Friday 29 of June at 01:00

Conclusions, challenges and prospects of using anonymised aggregated mobile phone data for producing information on population and its flows

From the results mentioned above it is clear that, at the moment, information based on anonymised aggregated mobile phone data maybe is not perfect, but it has huge potential for statistical purposes. Even data from one MNO shows logical patterns of population behaviour and accrued events (festivals, racings etc.). The data (which is never traceable to a single customer) allows to estimate day- and night-time population and its flows. As mentioned, information can be produced hourly on grid and/or various administrative levels. At the moment for this project to ensure the most representative results district level data was used for analysis and municipal level data was used for the output. Exploring anonymised aggregated mobile phone data is an innovative process with many challenges:

- Matching the night-time population to the PRD residential population often does not work well at the district level, because P(x|a) is still too coarse for fine spatial level.
- The mass lost from the cube by the threshold minimum of 15 devices.
- Time, human resource and computation limitations.
- Organizational and work cultural differences between NSI and MNO.

During the process of getting grip on anonymised aggregated mobile phone data the following possible way of its further development were identified:

- Improving methodology of estimating the geolocation of devices.
- Involving radio technical experts to get better understanding of the technology.
- Improving the calibrating method by taking into account that not every person owns a mobile phone and some people use multiple phones.
- Improving the calibration method by taking into account the Dutch persons who are abroad. We can use auxiliary data for this, for instance from the Continuous Holiday Survey.
- Getting access to a data from all Dutch MNOs. This will solve the problem on getting the complete picture on all Dutch mobile devices.
- Conducting research on how, next to 4G, the use of 2G and 3G generated records can improve aggregates.
- Research on Timing Advance, which allows GSM base transceiver station to control the signal delays in their communication with the mobile device.

For this project, data on Dutch devices were used but since signalling data also includes data from roaming devices, statistics on foreign visitors can be produced as well. This can be done by extracting information on originating country from the variable imsi of

signalling data. In this case in a devices' flow cube the place of residence will be defined as foreign country. For calibration in this example, when no populations registers of foreign countries are available, the market share from the foreign MNO can be used. MNOs have bilateral agreements between countries on which mobile network is preferred abroad. For instance, a German device from Deutsche Telecom Germany will by default use the mobile network from T-Mobile Netherlands. So in case of Germany we would calibrate German devices on market share of Deutsche Telecom Germany to estimate German population visiting the Netherlands. However, we do not take into account that not all foreign people use mobile phones. Further research is needed to solve this issue.

3.5. Experiences of using mobile phone data in Belgium

Mobile phone data for statistics: splendours and miseries

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%[39]. All three have been contacted by Statbel with proposals to jointly exploring their data and to combine them with statistical data for their commercial and Statbel's statistical use cases.

Talks with Telenet/Base in 2015 and 2018, respectively, both before and after the takeover of Base by Telenet, did not lead to any concrete plans or projects in spite of the interest expressed by the network operator on both occasions.

A first contact with Orange in 2016 ended with a similar non-result. In 2018, a more concrete proposal by MIT to combine mobile network signalling data and Statbel fiscal income data to study social segregation has not produced concrete results yet.

With Proximus and Eurostat a quite successful collaboration was started in December 2015 to explore and analyse the Proximus network signalling data for their information value and potential use. This resulted in 10 publications (see De Meersman e.a., 2016, for an overview of the project) and a joint press conference in September 2016 presenting the promising results to the general public. Unfortunately, this collaboration was ended and access to the mobile phone data was blocked by the Proximus management in March 2017. Contacts with Proximus were maintained but no further datasets were made available for study. However, in the context of the above-mentioned MIT project to study social segregation in Brussels, Proximus provided mobile network signalling data and Statbel created a fiscal income dataset which could be linked in a novel way and at a very detailed though not individual level to the mobile phone data, thus avoiding privacy or confidentiality issues. This was done by aggregating the income data using the geographical format of the mobile network antenna cells (Voronoi mobile network cell shapefiles provided by Proximus), ensuring a one-to-one mapping at a detailed level of the mobile phone and fiscal income data. This method can of course also be applied to many other personal statistical variables, most of which are geocoded to a person's domicile, without any privacy problem as only aggregated data leave Statbel. It offers a paradigm for linking mobile phone and other data, both for the commercial use of mobile network operators and the production of official statistics. Furthermore, the method presents one of the typical characteristics of so-called 'smart statistics' in which data are not taken out of an owner's data warehouse, but a computation is sent in and aggregated results are exported, solving two major concerns, of guaranteeing privacy and maintaining data protection, confidentiality and ownership.

This non-threatening approach may convince mobile network operators they have nothing to lose and a lot to win by exchanging and combining tailored datasets for commercial use by the operator and for statistical by the statistical office. At this moment, after more than three years of trying, this is still not the case, not in Belgium and not in the entire European Statistical system: mobile phone data are not freely available at this moment for producing official statistics.

---

[39] https://www.internetproviders.be/overzicht-mobiel-internet-op-belgische-markt/

A new approach not exclusively relying on creating win-win situations and convincing the network operator may be in order. The problem is arguably not with the innovation or business development units within the mobile network operator, they easily see the advantages far outweighing the costs and risks. Unfortunately, they do not make the decision to open up data access, and the higher management which seems to listen mainly to the legal department trying to minimise risk and the sales department not interested in co-creation of new valuable products but only in selling the data.

Only high-level intervention supported legal obligations is likely to change this situation, in a way similar to scannerdata from supermarkets and chain stores having become available for calculating consumer prices. A paper recently presented in October 2018 at the DGINS Conferene in Bucharest (Debusschere, Waeyaert, Van Loon, 2018) argues for a four-way approach to obtaining access to 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.

Statistical use cases

Numerous pilot studies have investigated the potential of mobile phone data for compiling or validating official statistics. Results, amongst others from the joint Statbel-Eurostat-Proximus project, look very promising. The logical next step is then the elaboration of statistical use cases, by identifying a concrete 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.

Key to a statistical use case is a concrete 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, it should be sustainable, i.e. repeatable with the agreed frequency without additional effort, and thus able to support long-term 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, etc. 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.

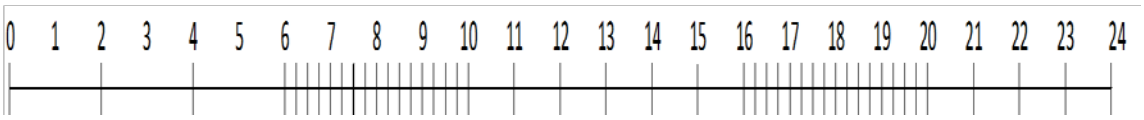
Improving the statistics on living place and workplace

Mobile phone data were already used for validating the Belgian Census results on population as to residence, based on the Belgian Population Register (De Meersman e.a., 2015). 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), 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.

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 the 24-hour timeline below, 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 a 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
- 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, every 15 minutes, instead of the proposed 45 would double the size of the dataset, but may simplify the query parameters.

Matrix living place x workplace (origin & destination of commuting)

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 cell 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 aggregates these observations to a dataset.

This mobile phone dataset could serve, in the context of the Census, to validate the Matrix Living Place X Workplace 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 Matrix Living Place X Workplace is being used, amongst others, by the social partners (employers and unions) represented in the Central Economic Council and by the Central Planning Office for their economic models.

The required 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.

In concrete, the request is for the results of a calculation on the basis of the data for the latest month of October, considered to be the most 'normal', 'typical' month as to length and absence of holidays or holiday periods. 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 cells. 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 cells.

Scenario 1

- Step 1: determine, for each mobile device, the Voronoi cell 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 cell in which the device is located at 04:00 is identified. The most probable living place then is the Voronoi cell 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 cells measured at 04:00 every day in October, for a given device).

- Step 2: determine, for each mobile device, the Voronoi cell which can be considered the most probable workplace of the person using the device.
  For every device the Voronoi cell 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 cell 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 cells 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						

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 cells. Let us note the number of times (the *frequency*) that the *i*-th device is found in the *j*-th Voronoi cell (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 cell 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:

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

We calculate the probability *p* that the *i*-th mobile device is located in Voronoi cell 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 cell if p_i is larger than or equal to a certain minimum value α. If $p_i < \alpha$, the 'most probable workplace' is considered as non-defined.
α could be set as the lowest decile of p. Or differently expressed

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

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 workplace 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						

Conclusions on using mobile phone data in Belgium

In theory, the use cases presented above should be able to give rise to experimental statistics and even to improved or wholly new statistical products, at a considerably greater level detail and much more timely if not practically in real time. However, this can only be tested with real mobile phone data.

In spite of the very promising situation at the beginning, Statbel has not been able to convince Proximus or any other mobile network operator to provide the data for testing the use cases. It is hoped that recent new discussions focusing on combining mobile phone data and statistical for commercial use cases will provide access to data for testing.

However, the advantage of mobile phone data is that networks, network signals and the way these are stored as data records, are quite similar in all countries. So it is hoped that the first one in the European Statistical System gaining access to the data needed will test these use cases.

3.6. Experiences of using mobile phone data in Austria

Statistics Austria established an intense cooperation with one of the MNOs to improve the estimation in the area of tourism statistics and investigate possible usage in our statistical production process. At the moment, there is no access to the microdata, but only to aggregated numbers.

The MNO works together with an analytical company to create a product for analysis in the area of tourism statistics and reached out to Statistics Austria's subject matter specialists. A written agreement between Statistics Austria and the MNO was established which defines the mutual handling on the exchanged data. In a progressing work the goals for this cooperation of Statistics Austria and the involved MNO should be clearly stated. This is a of course a good situation when a MNO wants to work in an area which is of interest also to the statistical office.

Statistics Austria got aggregated data on **inbound** tourism, a nightly stay is approximated with a stationary position in the night hours (between 02 am and 06 am) and this is currently only done for foreign sim-cards. The spatial resolution of the aggregates:

- Regional distribution: municipalities,
- Origin: all countries by two letter code
- Timely distribution: daily basis

For **outbound** tourism test data were sent which include the number of all resident simcards that at least had one connection with a foreign MNO.

The aggregation level was chosen to have appropriate comparison values in the traditional statistical products. For inbound tourism, Statistics Austria only received test data sets for 2 months with the structure of month X, municipality and the number of night stays in this cell. For outbound tourism, data structure looks as following: month X, country with the number of sim-cards active in the specific country in a given month. At the moment aggregated data and aggregation methods are being studied.

Due to given situation, mapping day- and night-time (residential) population as well as population flows are not implemented at the moment. The most feasible at the moment is to map night-time population which could be done with a similar methodology to the inbound tourisms, but the target group would have to be changed to Austrian sim-cards.

As Statistics Austria is in a preliminary phase, methodology is in a process of development. Due to the fact that it is definitely limiting to have data from only one MNO, we are always in a process of trying to establish cooperation with the other two MNOs in Austria to get the whole picture of mobile data. Using the data from only one MNO could potentially lead to biased results derived.

4. CONCLUSIONS AND RECOMMENDATIONS

The project confirmed that using micro data from the ***Labour Force Survey*** (LFS) to produce information on geographical areas other than NUTS 2 regions, like Functional Urban Areas (FUA's), is not without issues. Sample sizes, sampling designs and weighting procedures of the LFS are tailored to deliver outputs for NUTS 2 areas only. The consortium looked at the possibilities of using on the LFS to produce basic indicators (employment rate, unemployment rate and educational attainment) for the countries involved (BE, AT, DE, FR, AT and NL). Concrete reliability issues and limitations of using the LFS were identified. In addition, we organised a workshop for LFS experts to discuss the possibilities for using LFS data for FUA's and other geographies using the analysis as input.

The general conclusion was that one should be very careful in producing outputs in a creative way for varying geographies. Sample sizes should be larger enough. And even when they are large enough quality issues remain. The experts recommended that NSI's are involved in producing such outputs using the best methods available and checking if the results are of acceptable quality. Since only a limited set of indicators is required it would be feasible to define a set of tables that can be produced and verified by the NSI's. It would not involve a lot of work.

In this project we learned several lessons attempting to use anonymised aggregated **mobile phone data** to produce statistics on low regional level in general and city data in particular. It makes sense to present those lessons following the three process steps involved in producing statistics: acquiring, processing and disseminating data. We present them below.

Let us start with the first phase of acquiring mobile phone data. This is quite problematic. Almost all institutes participating in this project faces problems in getting hold of mobile phone data. Austria was still in the stage of discussing the possibilities to get data, Belgium had data in the past but the MNO they worked with terminated the cooperation, France could only use old data that they got hold on in the past. Germany was able to test the potential of mobile phone data based on a detailed table they received from a major MNO based on a cooperation with this MNO. The methodology how the table was produced was determined by the MNO or mediating company. It is a black box. Only the Netherlands has a live cooperation with one of Dutch MNOs. They work together on developing methodology. In return they are able to design and received tables with anonymised aggregated data. Having learned from that experience, Statistics Netherlands is cautious in dealing with MNO's. At all times, continuation of the fragile cooperation should be guaranteed. MNO's are extremely sensitive to privacy issues. The public and their customers should be certain that privacy is guaranteed under all circumstances. Even sharing anonymised micro data with statistical institutes that have a record of dealing with micro data in a reliable way is considered a bridge too far at this stage. For this reason, Statistics Netherlands arranged that the anonymised micro data will stay at the premises of the MNO. They can run programs on the MNOs' database resulting in detailed tables that contain counts of events with a minimum threshold of a certain number of cases. In addition, for the time being they decided not to produce tables that involves following telephones longitudinally. With these measures one can easily explain that there is not risk of sharing sensitive data. This approach is deliberately designed to be on the save side.

One can conclude several matters. First of all, statistical institutes will never be able to acquire mobile phone micro data because of legal, privacy and ethical reasons. They will have to settle with receiving anonymised aggregated information from MNOs as input to produce statistics. Not possessing the original micro data is an important change in the statistical production process. Furthermore, purchasing such an information is not a valid option for producing official statistics. An alternative is that this information is provided within the context of a cooperation agreement. The important question here is of course: what is in it for the MNO? It is still early to answer this question. The future will tell. At this stage one can mention a few points and elements that seem to have potential. The community of statistical institutes are able to develop and maintain a methodology as open standard how to produce statistics from anonymised aggregated mobile phone

data. This means that the MNOs do not have to do this work but can benefit from it. This methodology could be applied by the MNOs to produce customised statistical information. In addition, they could build new modules on this general software and methodology to produce more sophisticated products. Finally, the open standard methodology is internationally and universally usable which is beneficiary to all parties concerned. This approach has to be explored and needs to be tested in practice but it seems to have potential for a bright future.

The second part of the process is the methodology how to produce statistical output starting with the raw signalling or CDR data. The section on mobile phone data deals with that part of the process. It does not make sense to repeat all the details here. The bottom line is that quite some methodology has been developed but is still not yet at a mature level. A lot needs to be done. However it is do-able. A basic method is ready. It can already be used to produce outputs. The method must be improved obviously in the next phase. An important step to take is to test it in other countries using data from other MNOs. This will all helps to improve the methodology and define new and better outputs. It is important that this work is carried out by several institutes in a coordinated way. The resources has to be pooled, innovation has to be done jointly and the methods must be tested in several countries. This is required in order come to an international standard in an efficient way.

The final element of the process concerns the output mobile phone data should produce. The specific features and possibilities of anonymised aggregated mobile phone data forces us to define products that go beyond the standard and traditional products. Producing a set of predefined tables is not sufficient. First of all, we have the regional component of the data. Anonymised aggregated mobile phone data can be used to produce output on all regional levels. This ranges to national and even international level to very detailed local level like grid cells. These levels involve too much data to fit in a simple set of tables. Some of this information should be made available as open data that can accessed and processed by software and application. Methods, formats, standards and platforms have to be defined, developed, maintained and improved. In addition, we need interactive visualisation tools that enables a range of users to produce tailored outputs or carry out all sorts of analysis. These tools have to developed making use of state of the art technologies.

Both making the information as open data available and the accompanying visualisation tools require involvement of users to find out which functionalities these tools and systems will need to have. It requires an longstanding and never ending innovation process that has to be organised. This task should not be underestimated. The statistical community will have to organise the involvement of a broad range of stakeholders. Varying from national and international organisations and authorities to local players like cities or regional governments.

We argue that using mobile phone data is a game changer for the way statistics are produced. We will have to adapt the way we work in all stages of the statistical production process. In the past, NSI's could work quite independently to produce statistics. They collected data themselves and defined to a large extent the products they supply. Circumstances of data providers of users could largely be ignored. With many big data sources in general and mobile phone data in particular this is not the case anymore. The methods, products the products they create as the organisation of the work is depending on other actors forcing NSI's to cooperate with the actors involved. Acquiring data requires collaboration with MNO's in dealing privacy sensitive data and find creative solutions for a partnership that is beneficiary to all parties without disturbing their main processes. It could even lead to a situation that NSI's have to accept, at least for the time, that a part of the process will be a black box where they have no saying on. Developing methodologies will require international cooperation between statistical institutes, collaboration with scientists and international organisations since setting open standards is a common goal concerning all. Finally, new kind of products will have to be development in close cooperation with all kinds of users.

REFERENCES

References Germany

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. Journal of the American Statistical Association, 74(366):269-277

Brown, G., Chambers, R., Heady, P., and Heasman, D. (2001). Evaluation of small area estimation methods: An application to unemployment estimates from the UK LFS. Symposium 2001 - Achieving Data Quality in a Statistical Agency: A Methodological Perspective, Statistics Canada.

Schmid, T., Bruckschen, F., Salvati, N., and Zbiranski, T. (2017). Constructing sociodemographic indicators for national statistical institutes using mobile phone data: Estimating literacy rates in Senegal. Journal of the Royal Statistical Society Series A, 180(4):1163-1190.

Gross, M. (2018): Kernelheaping: Kernel Density Estimation for Heaped and Rounded Data. R package version 2.0.0. https://cran.r-project.org/web/packages/Kernelheaping/Kernelheaping.pdf, accessed 12 February 2018.

Gross, M., U. Rendtel, T. Schmid, S. Schmon, and N. Tzavidis (2017). Estimating the density of ethnic minorities and aged people in Berlin: Multivariate kernel density estimation applied to sensitive geo-referenced administrative data protected via measurement error. Journal of the Royal Statistical Society: Series A (Statistics in Society), 180(1):161-183.

Hadam, S. (2018). Use of mobile phone data for official statistics. METHODS – APPROACHES – DEVELOPMENTS
Information of the German Federal Statistical Office, 2018(2):6-9. https://www.destatis.de/EN/Methods/MethodologicalPapers/Download/mad2_2018.pdf;jsessionid=4F30D71F69DC8F918E54DB1F152C7:A3B.InternetLive1?__blob=publicationFile , accessed 15 January 2019.

References Belgium

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)

References The Netherlands

Tennekes, M., Gootzen, Y., Scholtus, S. (2019). Geographic Location of Mobile Phone Events. (In preparation.)

ANNEX 1 REPORTS ON USING LFS FOR INFORMATION ON FUA'S

ANNEX 1.1.

Report on the use of LFS for information on Functional Urban areas

Country: AUSTRIA

Date: 31-08-2018

Author(s):

Kowarik Alexander

Gussenbauer Johannes

Weinauer Marlene

1. Introduction

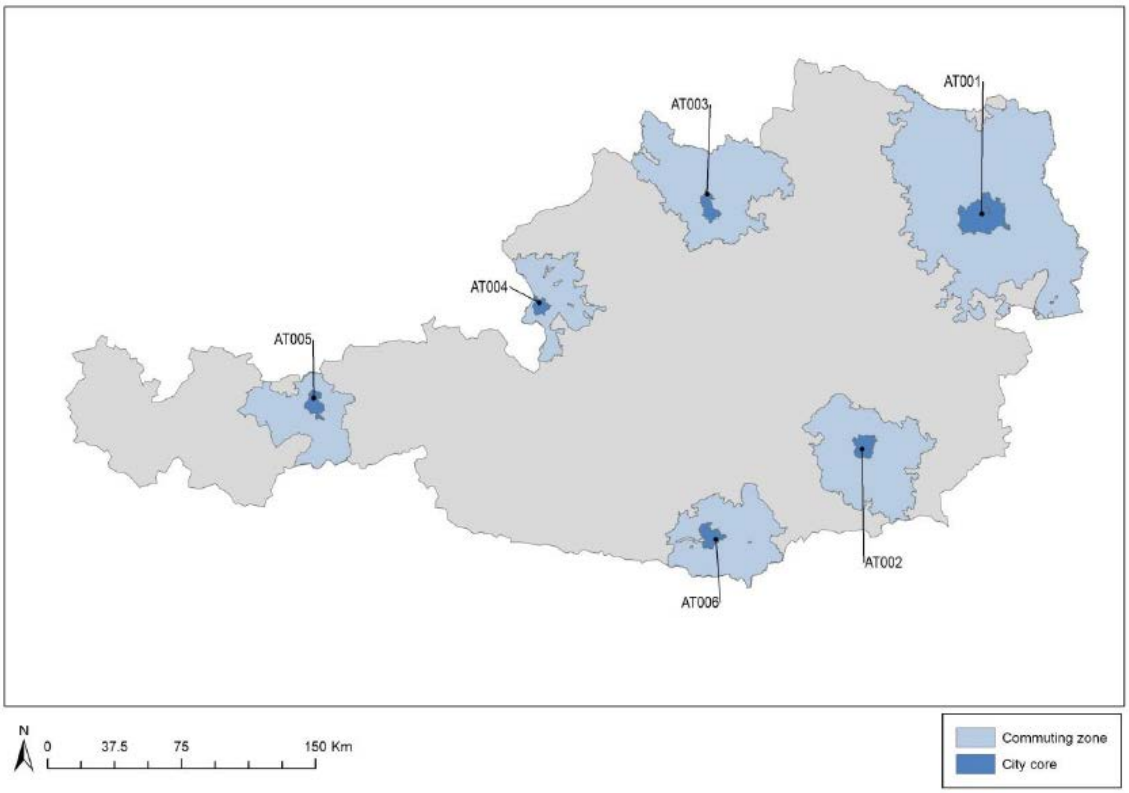
The Labour Force Survey (LFS) is the main source in the EU for harmonised labour market statistics. This survey is based on a sample and therefore only allows a limited degree of regional breakdown. Information on NUTS 2 is in general available. UN and OECD developed Functional Urban Areas (FUA's)[40] which play an important role in current policies dedicated to urban regions. One would expect that the populations of Metropolitan urban areas, with population between 500,000 and more, are of the same order as many NUTS 2 regions. The sample size would therefore be in principle large enough to provide LFS statistics. The main issue is that the sample of the LFS was not designed to generate statistics for these specific regions. This fact has implication on the quality. Small urban areas are excludes from the analysis, since the corresponding sample sizes are too small. For medium-sized urban areas we will investigate the possibilities.

This report presents the results of the analysis carried out for the country: AUSTRIA

2. Functional urban areas in Austria

In Austria the functional areas are constituted by the capitals of six federal states and their commuting area (Vienna; Linz of Upper Austria; Graz of Styria, Salzburg of Salzburg, Innsbruck of Tyrol and Klagenfurt of Carinthia).

[40] The boundaries of cities and Functional Urban Areas are available on the Eurostat-GISCO website. The version referred to as "Urban Audit 2011-2014" is still valid for France, Germany and Belgium. Revisions have taken place in the Netherlands and Austria. The revised FUA boundaries are included in the FUA 2015-2018 dataset. { HYPERLINK "http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/urban-audit" }



Country	ID on the map	Name FUA	Class type	Total population (2000)	Total population (2014)
Austria	AT001	Vienna	Large metropolitan areas	2,430,340	2,793,631
Austria	AT003	Linz	Metropolitan areas	582,430	616,829
Austria	AT002	Graz	Metropolitan areas	553,115	633,168
Austria	AT004	Salzburg	Medium-sized urban areas	334,913	362,455
Austria	AT005	Innsbruck	Medium-sized urban areas	273,373	304,173
Austria	AT006	Klagenfurt	Medium-sized urban areas	243,336	251,383
Total functional urban areas				4,417,507	4,961,639
Share of national population in functional urban areas				55.2%	58.3%
Number of functional urban areas					6

3. Results of LFS

In the following Austrian figures are presented for the indicators:

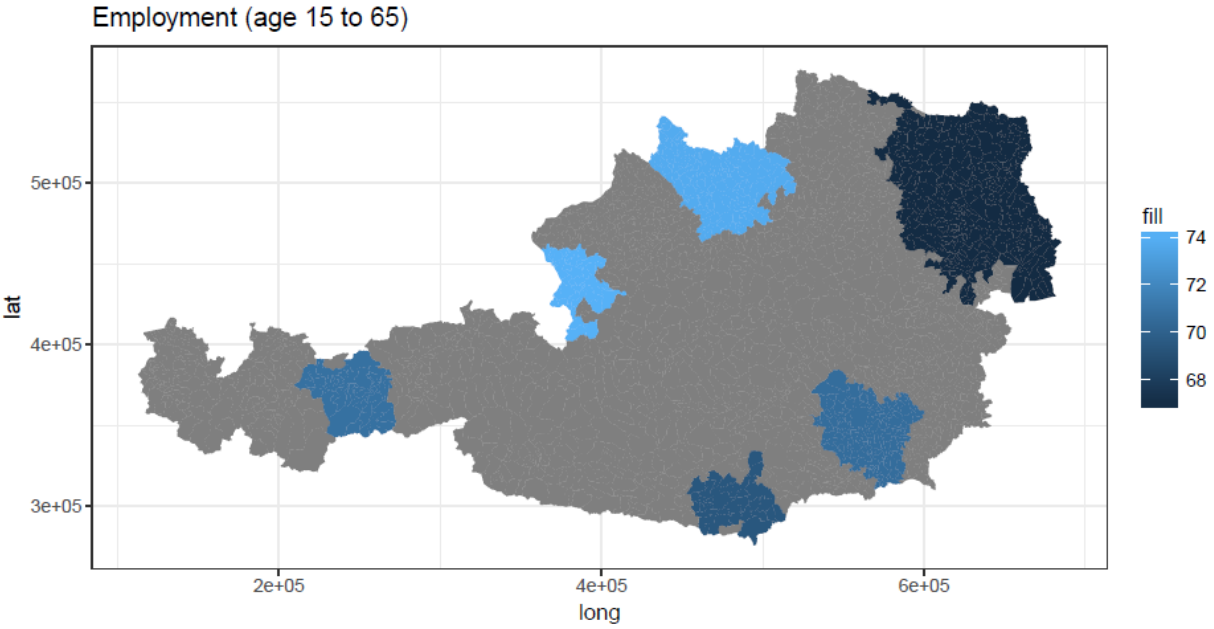
- Employment,
- Unemployment and
- Educational attainment

Each indicator is presented by a map indicating the shares of the population aged between 15 and 65 of each FUA and tables indicating totals and shares for breakdowns by age and sex for each FUA respectively. Closing a comparison with the shares of the associated federal state is given.

Employment rate

The employment shares of the Austrian FUAs proclaim a relatively homogeneous picture ranging from 66.8% to 74.1%. The FUA around Vienna records the lowest share.

High standard errors are noticed, especially for persons aged 15 to 24, but this is not special to the estimations for the FUAs.



Innsbruck	Total population	Employed persons			
	in 1,000	in 1,000	se in 1,000	in percent	se in percent
All Persons	**227,2381**	**160,9705**	**4,4437**	**70,8378**	**0,7619**
Age					
15 to 24 years	42,4875	19,8678	1,1088	46,7615	1,9117
25 to 44 years	90,0823	74,4944	2,8359	82,6959	1,1670
45 to 65 years	94,6683	66,6083	2,2778	70,3597	1,2900
Sex, Age					
Male	111,6038	83,9104	2,6175	75,1860	0,9964
15 to 24 years	21,4148	10,0552	0,8226	46,9542	2,9271
25 to 44 years	45,6665	39,1424	1,7015	85,7136	1,5102
45 to 65 years	44,5225	34,7129	1,3408	77,9670	1,5714
Female	115,6342	77,0601	2,3090	66,6412	1,0034
15 to 24 years	21,0727	9,8127	0,7300	46,5657	2,7836
25 to 44 years	44,4158	35,3520	1,5374	79,5933	1,6506
45 to 65 years	50,1457	31,8954	1,2725	63,6055	1,6761

Graz	Total population	Employed persons			
	in 1,000	in 1,000	se in 1,000	in percent	se in percent
All Persons	**450,2186**	**317,9187**	**6,8127**	**70,6143**	**0,6330**
Age					
15 to 24 years	76,5901	37,7401	2,0085	49,2754	2,0158
25 to 44 years	178,4224	149,7222	4,1611	83,9144	0,9985
45 to 65 years	195,2061	130,4565	3,9038	66,8301	1,1811
Sex, Age					
Male	227,2879	167,0288	3,9454	73,4878	0,8454
15 to 24 years	36,8198	17,2537	1,3945	46,8598	2,9938
25 to 44 years	93,1785	80,6499	2,5757	86,5541	1,3398
45 to 65 years	97,2896	69,1253	2,3158	71,0510	1,5201
Female	222,9307	150,8899	3,6055	67,6847	0,8530
15 to 24 years	39,7703	20,4864	1,2724	51,5118	2,5621
25 to 44 years	85,2439	69,0723	2,2366	81,0290	1,3876
45 to 65 years	97,9165	61,3312	2,2175	62,6363	1,5583

Klagenfurt	Total population	Employed persons			
	in 1,000	in 1,000	se in 1,000	in percent	se in percent
All Persons	**144,7360**	**100,3985**	**3,3146**	**69,3666**	**0,8837**
Age					
15 to 24 years	19,0445	8,9432	0,8187	46,9593	2,6265
25 to 44 years	56,6493	47,5735	1,8961	83,9790	1,2381
45 to 65 years	69,0423	43,8818	1,9283	63,5579	1,5029
Sex, Age					
Male	69,9415	50,1781	1,9162	71,7429	1,2085
15 to 24 years	9,3603	4,1396	0,5030	44,2253	3,5374
25 to 44 years	27,0452	23,7141	1,0929	87,6833	1,6060
45 to 65 years	33,5359	22,3243	1,1391	66,5683	2,0213

Female	74,7946	50,2205	1,7531	67,1445	1,1857
15 to 24 years	9,6842	4,8035	0,5647	49,6018	3,9114
25 to 44 years	29,6040	23,8594	1,1087	80,5950	1,7866
45 to 65 years	35,5064	21,5576	1,0777	60,7147	1,9102

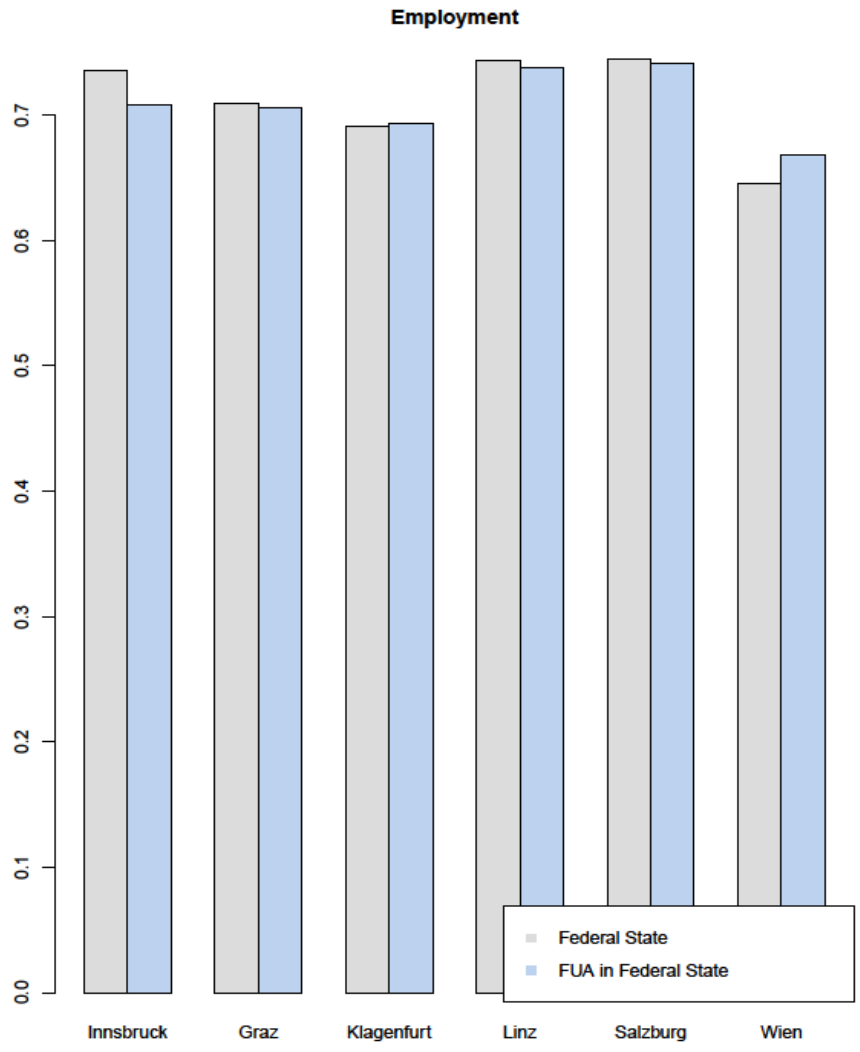
Linz	Total population	Employed persons			
	in 1,000	in 1,000	se in 1,000	in percent	se in percent
All Persons	**409,5013**	**302,3790**	**7,9070**	**73,8408**	**0,7755**
Age					
15 to 24 years	62,6690	32,4779	2,0705	51,8245	2,1967
25 to 44 years	167,1342	145,3051	4,8356	86,9392	1,1056
45 to 65 years	179,6982	124,5960	4,2839	69,3363	1,2502
Sex, Age					
Male	203,6495	158,6105	4,6024	77,8840	1,0348
15 to 24 years	30,1954	16,2733	1,3306	53,8934	3,1082
25 to 44 years	83,5861	75,2746	2,8259	90,0564	1,4678
45 to 65 years	89,8681	67,0626	2,5442	74,6234	1,5510
Female	205,8518	143,7685	4,2031	69,8408	1,0193
15 to 24 years	32,4736	16,2046	1,3100	49,9008	2,9095
25 to 44 years	83,5481	70,0306	2,8081	83,8206	1,5489
45 to 65 years	89,8301	57,5334	2,3785	64,0469	1,6425

Salzburg	Total population	Employed persons			
	in 1,000	in 1,000	se in 1,000	in percent	se in percent
All Persons	**236,5245**	**175,2340**	**3,8001**	**74,0871**	**0,5646**
Age					
15 to 24 years	40,9494	21,3041	1,1510	52,0255	1,9221
25 to 44 years	97,1170	82,8078	2,2943	85,2660	0,9948
45 to 65 years	98,4580	71,1221	2,0656	72,2360	1,1399
Sex, Age					
Male	117,6979	91,2169	2,1698	77,5008	0,7904
15 to 24 years	20,8657	10,6509	0,7530	51,0450	2,6486
25 to 44 years	48,7182	43,0420	1,3325	88,3489	1,2703
45 to 65 years	48,1140	37,5240	1,2612	77,9897	1,5772
Female	118,8265	84,0171	2,0380	70,7057	0,8394
15 to 24 years	20,0838	10,6533	0,7721	53,0441	2,7431
25 to 44 years	48,3988	39,7657	1,2894	82,1627	1,3644
45 to 65 years	50,3440	33,5982	1,1279	66,7372	1,5007

Wien	Total population	Employed persons			
	in 1,000	in 1,000	se in 1,000	in percent	se in percent
All Persons	**1.904,8876**	**1.272,6497**	**10,2986**	**66,8097**	**0,2794**
Age					
15 to 24 years	315,8811	125,9848	3,5148	39,8836	0,9983
25 to 44 years	790,0377	621,4242	6,7884	78,6575	0,5800
45 to 65 years	798,9689	525,2407	6,7946	65,7398	0,6108
Sex, Age					
Male	941,4558	657,5594	5,8933	69,8450	0,3848

15 to 24 years	159,4058	59,9170	2,3274	37,5877	1,3613
25 to 44 years	391,3891	318,8047	4,1916	81,4547	0,7628
45 to 65 years	390,6608	278,8378	3,9840	71,3759	0,7716
Female	963,4319	615,0903	5,8150	63,8437	0,3696
15 to 24 years	156,4753	66,0678	2,5313	42,2225	1,4326
25 to 44 years	398,6485	302,6195	4,2545	75,9114	0,7926
45 to 65 years	408,3080	246,4029	4,1379	60,3473	0,8099

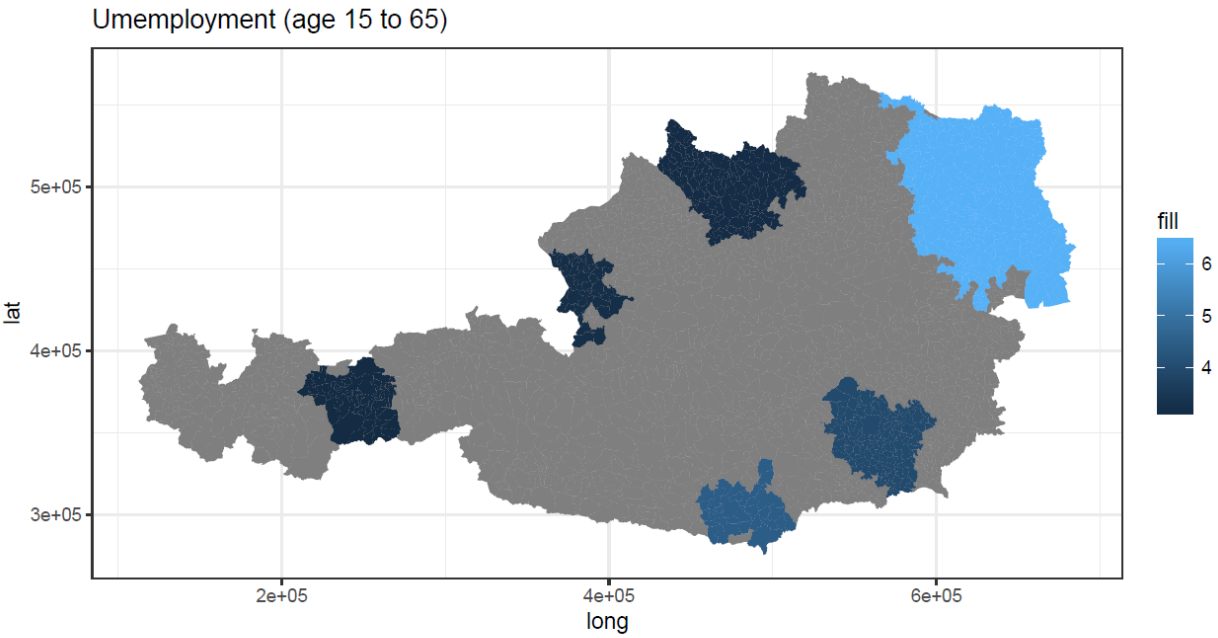
The following barplot compares employment shares of FUAS with those of the corresponding federal states. The fairly equal shares could be explained by the small area of Austria and therefore the opportunity to have a medium or large FUA within reach for commuting.



Unemployment rate

In accordance to the employment rate, the unemployment rate is the highest in the FUA around Vienna (6.5%). The other FUAS are again characterized by homogenous figures (3.1% to 4.5%).

Again high standard errors are noticed, especially for persons aged 15 to 24.



Innsbruck	Total population	Unemployed persons			
	in 1,000	in 1,000	se in 1,000	in percent	se in percent
All Persons	**227,2381**	**7,0436**	**0,6989**	**3,0997**	**0,2958**
Age					
15 to 24 years	42,4875	1,7809	0,2824	4,1916	0,6505
25 to 44 years	90,0823	3,9629	0,5640	4,3992	0,6071
45 to 65 years	94,6683	1,2998	0,3265	1,3730	0,3423
Sex, Age					
Male	111,6038	3,3669	0,5043	3,0169	0,4376
15 to 24 years	21,4148	0,8516	0,2176	3,9766	0,9957
25 to 44 years	45,6665	1,9636	0,4296	4,2999	0,9194
45 to 65 years	44,5225	0,5517	0,1985	1,2392	0,4433
Female	115,6342	3,6767	0,4445	3,1796	0,3778
15 to 24 years	21,0727	0,9293	0,1910	4,4101	0,9062
25 to 44 years	44,4158	1,9993	0,3569	4,5014	0,7766
45 to 65 years	50,1457	0,7480	0,2421	1,4917	0,4821

Graz	Total population	Unemployed persons			
	in 1,000	in 1,000	se in 1,000	in percent	se in percent
All Persons	**450,2186**	**17,7547**	**1,3450**	**3,9436**	**0,2916**
Age					
15 to 24 years	76,5901	4,0177	0,6167	5,2458	0,7698
25 to 44 years	178,4224	7,1415	0,8804	4,0026	0,4853
45 to 65 years	195,2061	6,5955	0,9494	3,3787	0,4790
Sex, Age					
Male	227,2879	10,7865	0,9855	4,7457	0,4272
15 to 24 years	36,8198	2,7059	0,5475	7,3491	1,4234

25 to 44 years	93,1785	4,3381	0,6910	4,6556	0,7394
45 to 65 years	97,2896	3,7425	0,6389	3,8468	0,6436
Female	222,9307	6,9683	0,8068	3,1257	0,3543
15 to 24 years	39,7703	1,3118	0,2949	3,2985	0,7360
25 to 44 years	85,2439	2,8035	0,5899	3,2888	0,6790
45 to 65 years	97,9165	2,8530	0,5983	2,9137	0,6060

Klagenfurt	Total population	Unemployed persons			
	in 1,000	in 1,000	se in 1,000	in percent	se in percent
All Persons	**144,7360**	**6,4712**	**0,6852**	**4,4710**	**0,4500**
Age					
15 to 24 years	19,0445	0,9721	0,1742	5,1043	0,9119
25 to 44 years	56,6493	3,0742	0,4423	5,4267	0,7602
45 to 65 years	69,0423	2,4249	0,4517	3,5122	0,6349
Sex, Age					
Male	69,9415	3,5318	0,4960	5,0496	0,6781
15 to 24 years	9,3603	0,5275	0,1126	5,6352	1,2209
25 to 44 years	27,0452	1,4234	0,3155	5,2629	1,1394
45 to 65 years	33,5359	1,5810	0,3895	4,7143	1,1293
Female	74,7946	2,9394	0,3797	3,9300	0,4951
15 to 24 years	9,6842	0,4446	0,1403	4,5911	1,4448
25 to 44 years	29,6040	1,6509	0,3047	5,5765	1,0077
45 to 65 years	35,5064	0,8440	0,2077	2,3769	0,5742

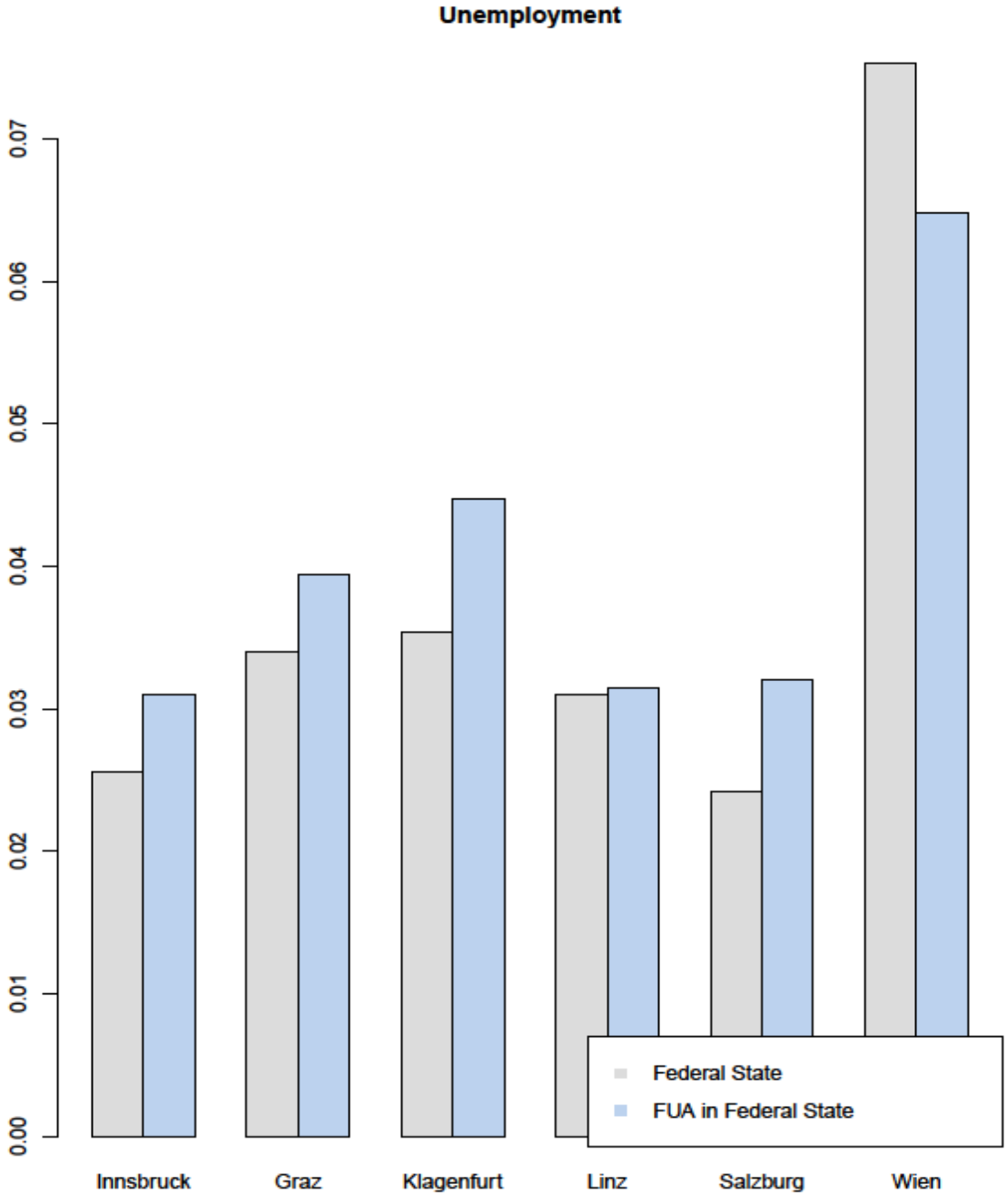
Linz	Total population	Unemployed persons			
	in 1,000	in 1,000	se in 1,000	in percent	se in percent
All Persons	**409,5013**	**12,8742**	**1,3767**	**3,1439**	**0,3195**
Age					
15 to 24 years	62,6690	2,6705	0,5287	4,2613	0,8233
25 to 44 years	167,1342	6,9184	1,1709	4,1394	0,6750
45 to 65 years	179,6982	3,2854	0,6890	1,8283	0,3785
Sex, Age					
Male	203,6495	7,8691	1,1138	3,8640	0,5268
15 to 24 years	30,1954	1,4465	0,3715	4,7906	1,1754
25 to 44 years	83,5861	4,5932	1,0298	5,4952	1,1775
45 to 65 years	89,8681	1,8293	0,4924	2,0356	0,5428
Female	205,8518	5,0052	0,7333	2,4314	0,3439
15 to 24 years	32,4736	1,2240	0,3633	3,7691	1,1020
25 to 44 years	83,5481	2,3252	0,4624	2,7830	0,5428
45 to 65 years	89,8301	1,4561	0,4599	1,6209	0,5100

Salzburg	Total population	Unemployed persons			
	in 1,000	in 1,000	se in 1,000	in percent	se in percent
All Persons	**236,5245**	**7,5818**	**0,6225**	**3,2055**	**0,2538**
Age					
15 to 24 years	40,9494	2,2669	0,3544	5,5358	0,8476
25 to 44 years	97,1170	3,1568	0,3884	3,2505	0,3958

45 to 65 years	98,4580	2,1582	0,3794	2,1920	0,3745
Sex, Age					
Male	117,6979	4,4090	0,4344	3,7461	0,3584
15 to 24 years	20,8657	1,3033	0,2733	6,2460	1,3106
25 to 44 years	48,7182	1,6015	0,2771	3,2874	0,5633
45 to 65 years	48,1140	1,5042	0,3006	3,1264	0,6081
Female	118,8265	3,1728	0,4298	2,6701	0,3549
15 to 24 years	20,0838	0,9636	0,2247	4,7980	1,1060
25 to 44 years	48,3988	1,5552	0,2668	3,2133	0,5457
45 to 65 years	50,3440	0,6540	0,2362	1,2990	0,4614

Wien	Total population	Unemployed persons			
	in 1,000	in 1,000	se in 1,000	in percent	se in percent
All Persons	**1.904,8876**	**123,4875**	**3,5509**	**6,4827**	**0,1855**
Age					
15 to 24 years	315,8811	24,6613	1,7217	7,8071	0,5354
25 to 44 years	790,0377	58,8109	2,7798	7,4441	0,3498
45 to 65 years	798,9689	40,0153	2,3151	5,0084	0,2875
Sex, Age					
Male	941,4558	72,4710	2,7301	7,6978	0,2847
15 to 24 years	159,4058	15,1062	1,3754	9,4766	0,8421
25 to 44 years	391,3891	33,6213	2,1977	8,5902	0,5570
45 to 65 years	390,6608	23,7435	1,7817	6,0778	0,4546
Female	963,4319	51,0166	2,0466	5,2953	0,2140
15 to 24 years	156,4753	9,5550	0,9628	6,1064	0,6077
25 to 44 years	398,6485	25,1896	1,6132	6,3188	0,4049
45 to 65 years	408,3080	16,2719	1,3642	3,9852	0,3325

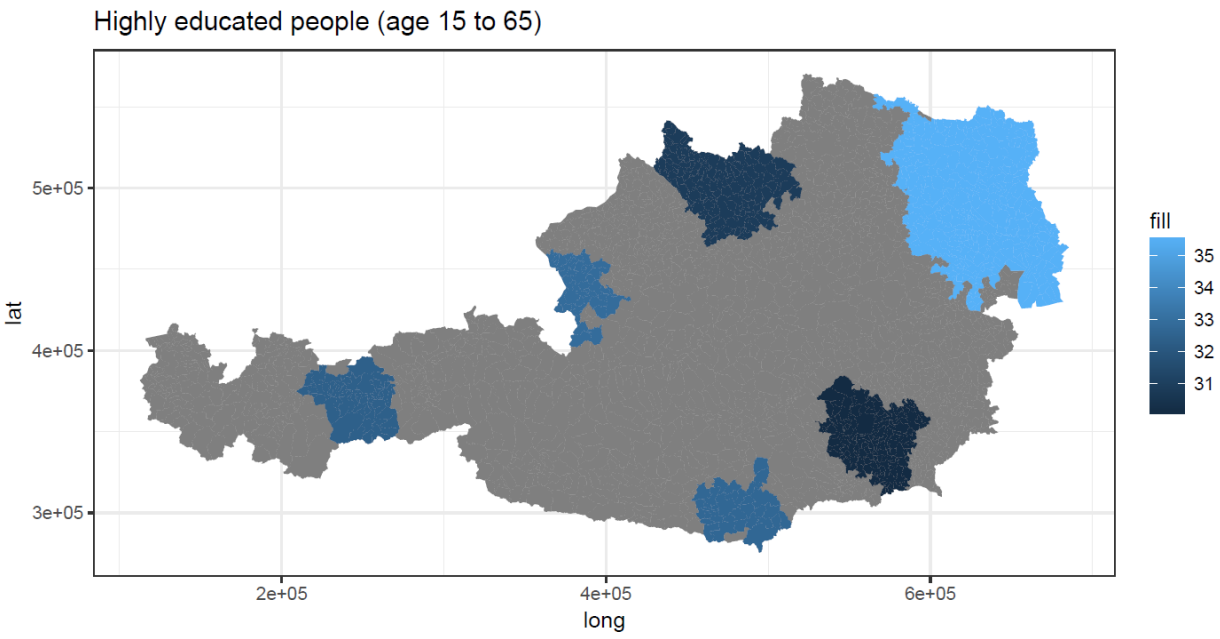
The following barplot compares unemployment shares of FUAs with those of the corresponding federal states. The shares of unemployment are always higher in the FUA than in the federal states, except for the FUA around Vienna. Probably unemployment is generally higher in urban areas and FUAS are more urban than their state, except in Vienna, where the state is the city and the difference between the FUA 'Vienna' and the state is less urban than the city.



Rate of highly educated people

The highest share of highly educated people by far is observed in the FUA around Vienna (35.5%). Lowest shares are noticed in the FUAs around Graz and Linz.

Again high standard errors are observed.



Innsbruck	Total population	Highly educated people			
	in 1,000	in 1,000	se in 1,000	in percent	se in percent
All Persons	**227,2381**	**73,7357**	**3,1033**	**32,4487**	**1,0793**
Age					
15 to 24 years	42,4875	6,4455	0,7483	15,1703	1,7149
25 to 44 years	90,0823	37,2731	2,2112	41,3767	1,8336
45 to 65 years	94,6683	30,0172	1,6548	31,7077	1,5034
Sex, Age					
Male	111,6038	36,3107	1,7620	32,5353	1,3033
15 to 24 years	21,4148	3,0797	0,4897	14,3811	2,2353
25 to 44 years	45,6665	17,7542	1,3347	38,8779	2,4029
45 to 65 years	44,5225	15,4768	1,0394	34,7617	2,0820
Female	115,6342	37,4250	1,9871	32,3650	1,4604
15 to 24 years	21,0727	3,3658	0,5153	15,9723	2,3719
25 to 44 years	44,4158	19,5189	1,3846	43,9458	2,4141
45 to 65 years	50,1457	14,5404	1,0549	28,9962	1,8668

Graz	Total population	Highly educated people			
	in 1,000	in 1,000	se in 1,000	in percent	se in percent

All Persons	**450,2186**	**135,6094**	**4,9583**	**30,1208**	**0,9512**
Age					
15 to 24 years	76,5901	15,8967	1,4946	20,7555	1,7633
25 to 44 years	178,4224	73,5642	3,6046	41,2303	1,7069
45 to 65 years	195,2061	46,1485	2,6570	23,6409	1,2438
Sex, Age					
Male	227,2879	68,8877	3,0673	30,3086	1,1762
15 to 24 years	36,8198	7,2812	0,9630	19,7753	2,3292
25 to 44 years	93,1785	36,9127	2,2262	39,6150	2,1234
45 to 65 years	97,2896	24,6938	1,5927	25,3817	1,5243
Female	222,9307	66,7217	2,8850	29,9293	1,1451
15 to 24 years	39,7703	8,6155	1,0185	21,6630	2,4157
25 to 44 years	85,2439	36,6515	2,1702	42,9960	2,1900
45 to 65 years	97,9165	21,4547	1,6116	21,9112	1,5322

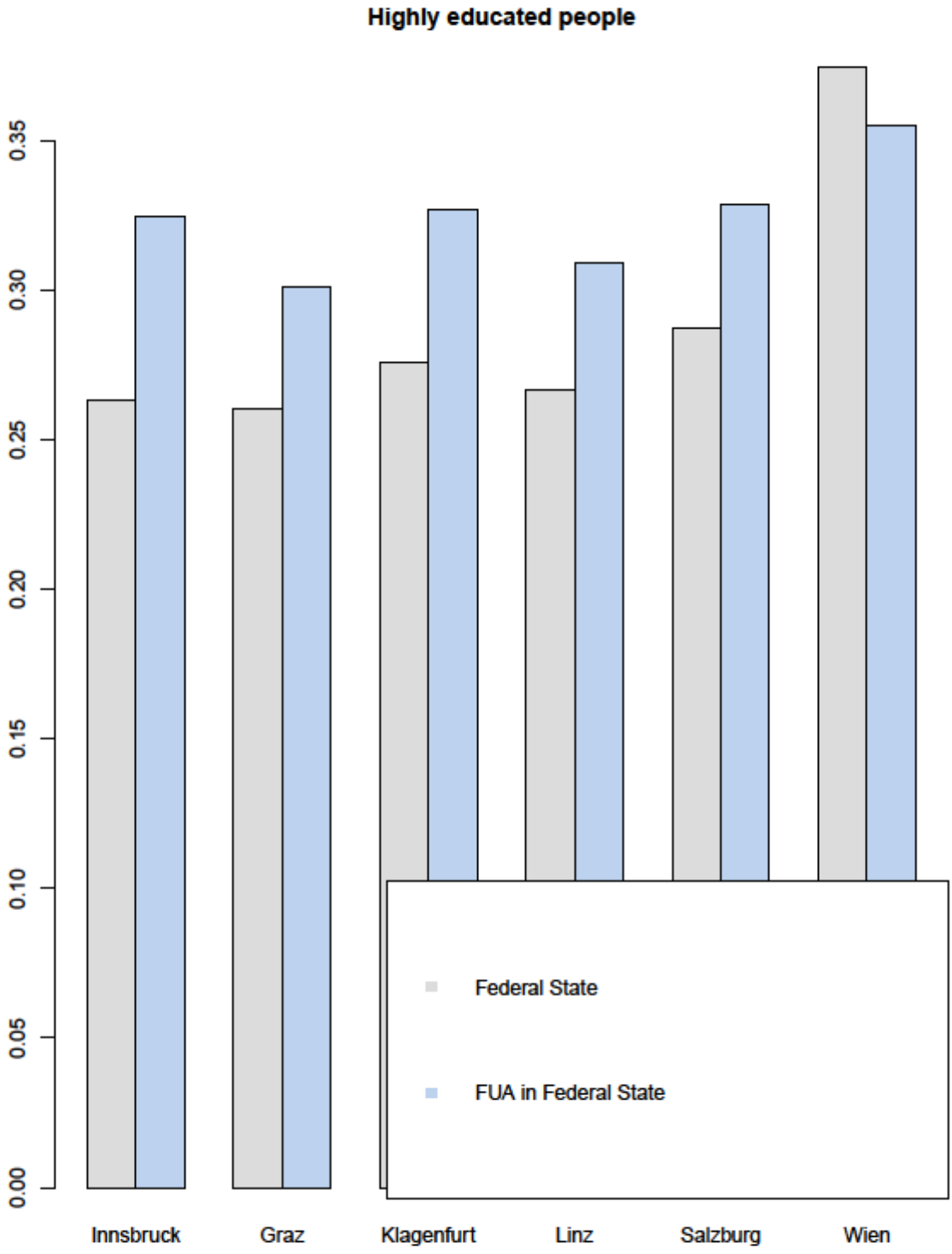
Klagenfurt	Total population	Highly educated people			
	in 1,000	in 1,000	se in 1,000	in percent	se in percent
All Persons	**144,7360**	**47,3435**	**2,1976**	**32,7103**	**1,1102**
Age					
15 to 24 years	19,0445	4,1084	0,6313	21,5727	2,8094
25 to 44 years	56,6493	24,1178	1,4568	42,5739	2,0180
45 to 65 years	69,0423	19,1173	1,3050	27,6893	1,5723
Sex, Age					
Male	69,9415	22,1625	1,3554	31,6872	1,5358
15 to 24 years	9,3603	1,7455	0,4093	18,6476	3,7522
25 to 44 years	27,0452	10,4680	0,9222	38,7055	2,8681
45 to 65 years	33,5359	9,9490	0,8369	29,6668	2,1236
Female	74,7946	25,1810	1,3281	33,6669	1,4106
15 to 24 years	9,6842	2,3629	0,4282	24,3999	3,7706
25 to 44 years	29,6040	13,6498	0,8953	46,1080	2,4490
45 to 65 years	35,5064	9,1683	0,7482	25,8215	1,8570

Linz	Total population	Highly educated people			
	in 1,000	in 1,000	se in 1,000	in percent	se in percent
All Persons	**409,5013**	**126,5704**	**5,2377**	**30,9084**	**1,0851**
Age					
15 to 24 years	62,6690	11,7528	1,5153	18,7538	2,1244
25 to 44 years	167,1342	66,1881	3,6271	39,6018	1,7751
45 to 65 years	179,6982	48,6296	2,9270	27,0618	1,4209
Sex, Age					
Male	203,6495	65,6328	3,1734	32,2283	1,3851
15 to 24 years	30,1954	4,7226	1,0077	15,6403	3,0629
25 to 44 years	83,5861	30,7351	2,1767	36,7706	2,3163
45 to 65 years	89,8681	30,1751	2,0005	33,5771	1,9412
Female	205,8518	60,9376	3,0473	29,6026	1,3277
15 to 24 years	32,4736	7,0302	0,9783	21,6488	2,7049
25 to 44 years	83,5481	35,4530	2,2383	42,4342	2,1917
45 to 65 years	89,8301	18,4545	1,5718	20,5437	1,6381

Salzburg	Total population	Highly educated people			
	in 1,000	in 1,000	se in 1,000	in percent	se in percent
All Persons	**236,5245**	**77,7368**	**2,7965**	**32,8663**	**0,9448**
Age					
15 to 24 years	40,9494	8,3881	0,8612	20,4840	1,9174
25 to 44 years	97,1170	39,3124	1,8401	40,4795	1,5325
45 to 65 years	98,4580	30,0363	1,5025	30,5067	1,3264
Sex, Age					
Male	117,6979	38,7577	1,6426	32,9298	1,1973
15 to 24 years	20,8657	3,3616	0,4770	16,1105	2,2499
25 to 44 years	48,7182	18,0901	1,0423	37,1321	1,8719
45 to 65 years	48,1140	17,3060	1,0213	35,9687	1,8551
Female	118,8265	38,9792	1,6972	32,8034	1,2092
15 to 24 years	20,0838	5,0265	0,6869	25,0277	3,0450
25 to 44 years	48,3988	21,2223	1,1713	43,8489	2,0357
45 to 65 years	50,3440	12,7303	0,8771	25,2867	1,6201

Wien	Total population	Highly educated people			
	in 1,000	in 1,000	se in 1,000	in percent	se in percent
All Persons	**1.904,8876**	**676,0880**	**10,8384**	**35,4923**	**0,5250**
Age					
15 to 24 years	315,8811	50,9998	3,0101	16,1453	0,9195
25 to 44 years	790,0377	359,7847	7,4555	45,5402	0,8988
45 to 65 years	798,9689	265,3035	6,3922	33,2057	0,7671
Sex, Age					
Male	941,4558	321,6448	6,7458	34,1646	0,6795
15 to 24 years	159,4058	18,6892	1,8637	11,7243	1,1615
25 to 44 years	391,3891	163,6340	4,5426	41,8085	1,1228
45 to 65 years	390,6608	139,3216	4,0622	35,6631	0,9869
Female	963,4319	354,4432	6,4202	36,7897	0,6422
15 to 24 years	156,4753	32,3106	2,1136	20,6490	1,2807
25 to 44 years	398,6485	196,1507	4,5303	49,2039	1,1029
45 to 65 years	408,3080	125,9819	3,9105	30,8546	0,9452

The following barplot compares the rates of people of highly educated people of FUAs with those of the corresponding federal states. The rates of highly educated people is always higher in the FUAS than in the corresponding federal state – except for Vienna (for which the whole federal state is included in the FUA).

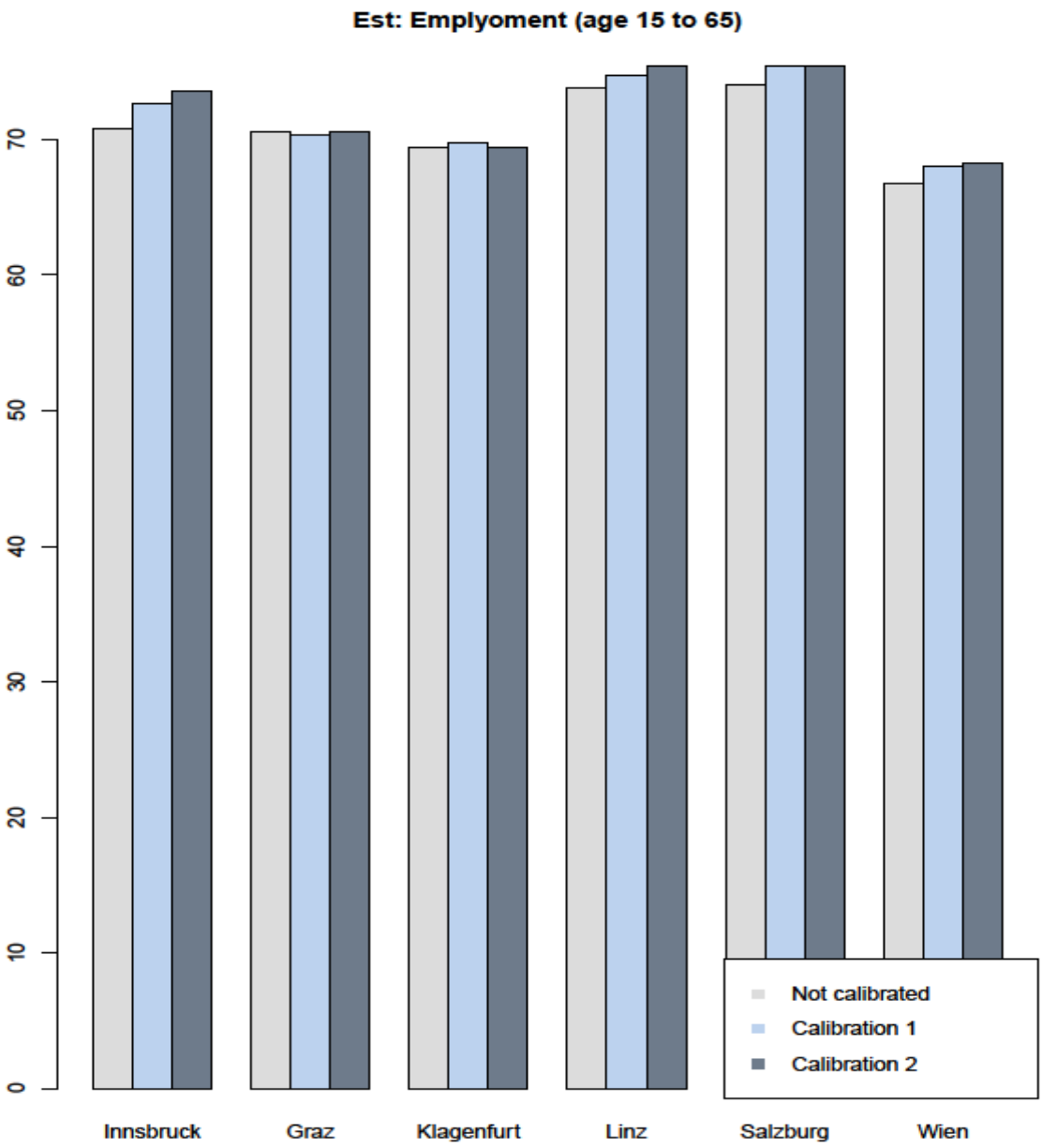

Highly educated people
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Innsbruck
Graz
Klagenfurt
Linz
Salzburg
Wien
Federal State
FUA in Federal State

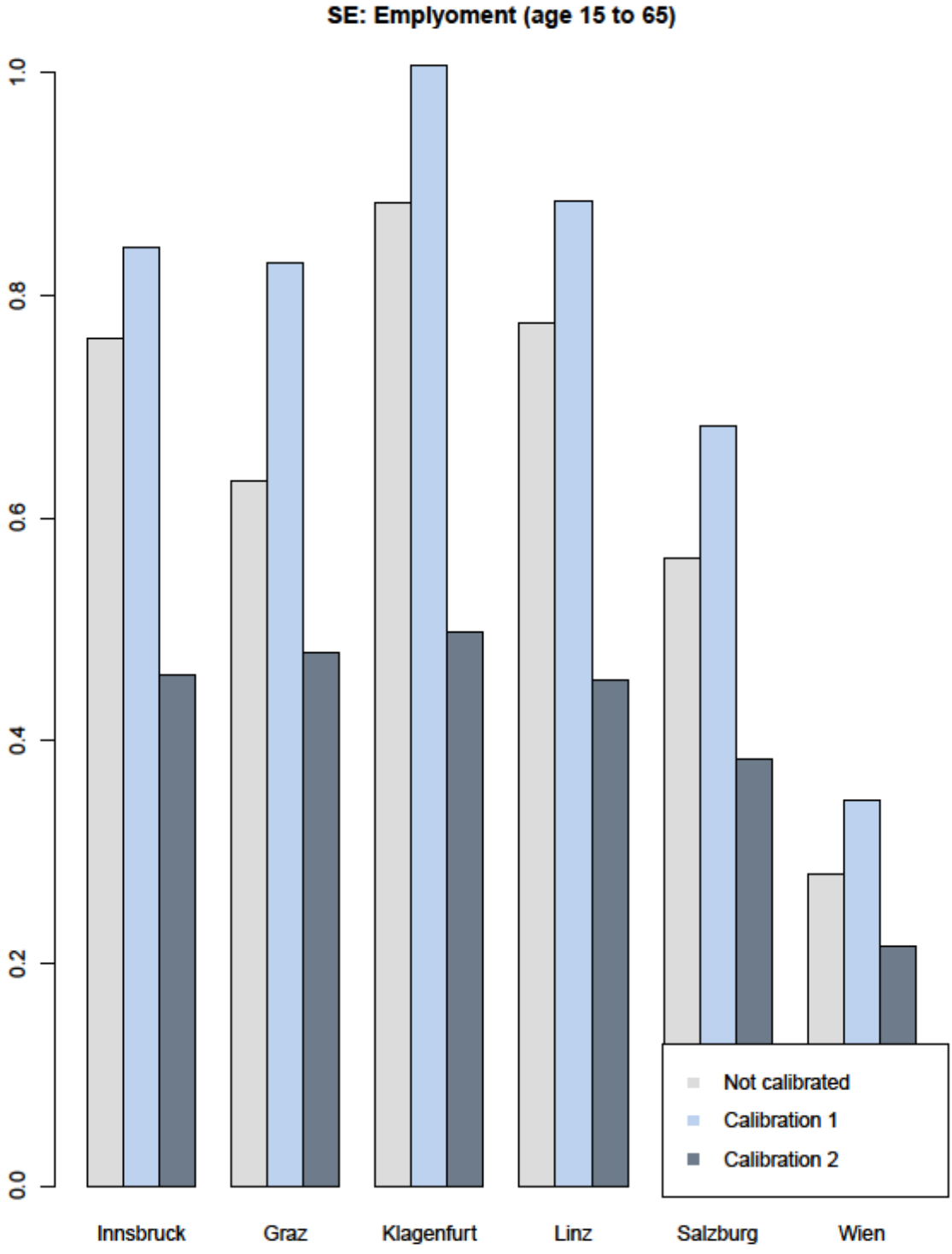
Issues and concerns in producing indicators for FUA's

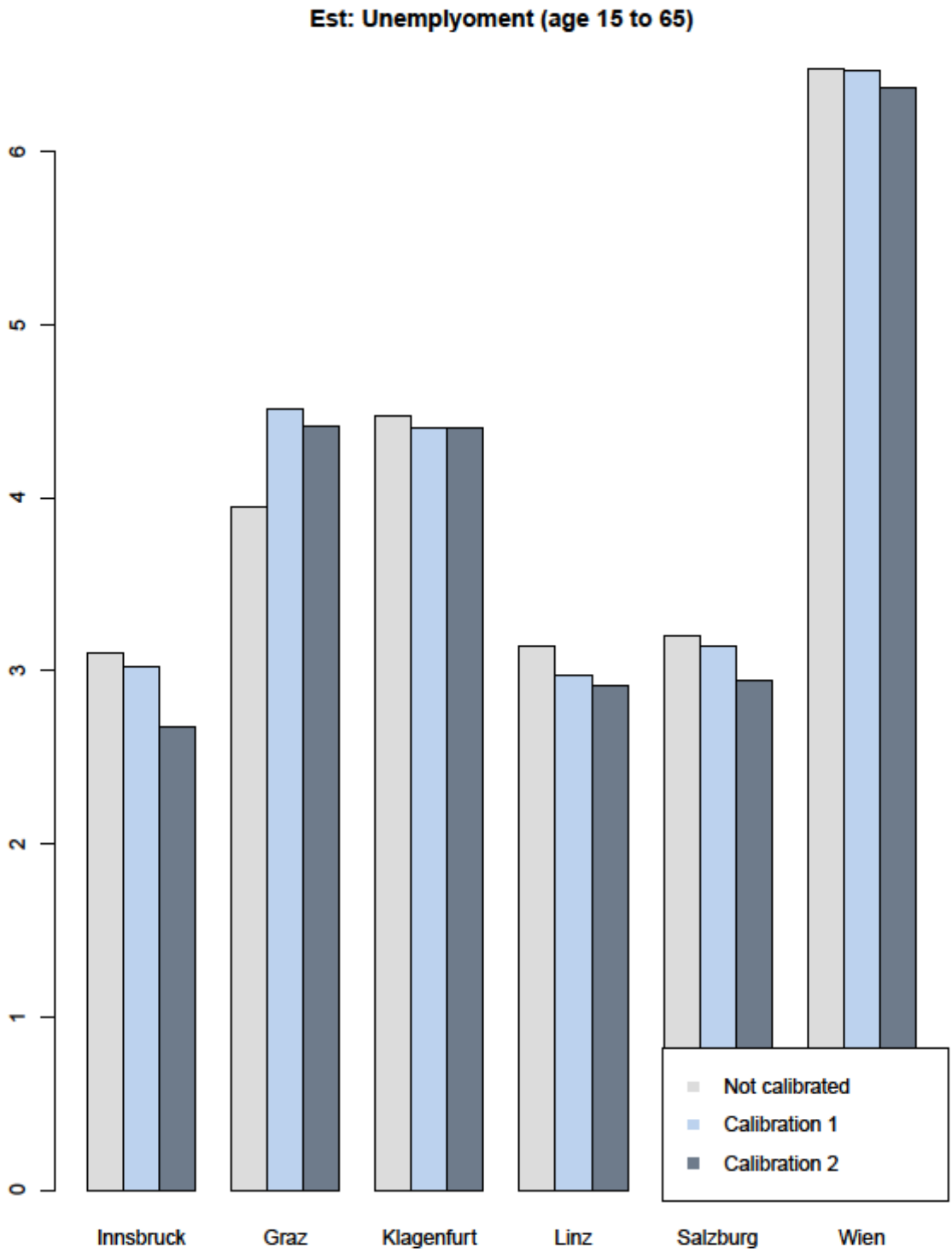
The estimates presented in the previous section do, in some cases, yield high standard errors, which raise the question if the use of LFS is enough to produce reliable estimates for FUAs. Nevertheless there is room for improvement for instance using variables from other administrative sources which highly correlate with the education and employment rate. Taking the corresponding population margins from these sources and specifically calibrating the sampling weights can further improve the estimates for this specific problem. Statistics Austria tested multiple approaches for calibrating the sampling weights for which the results are presented for 2 of those approaches in the following. The approaches correspond to

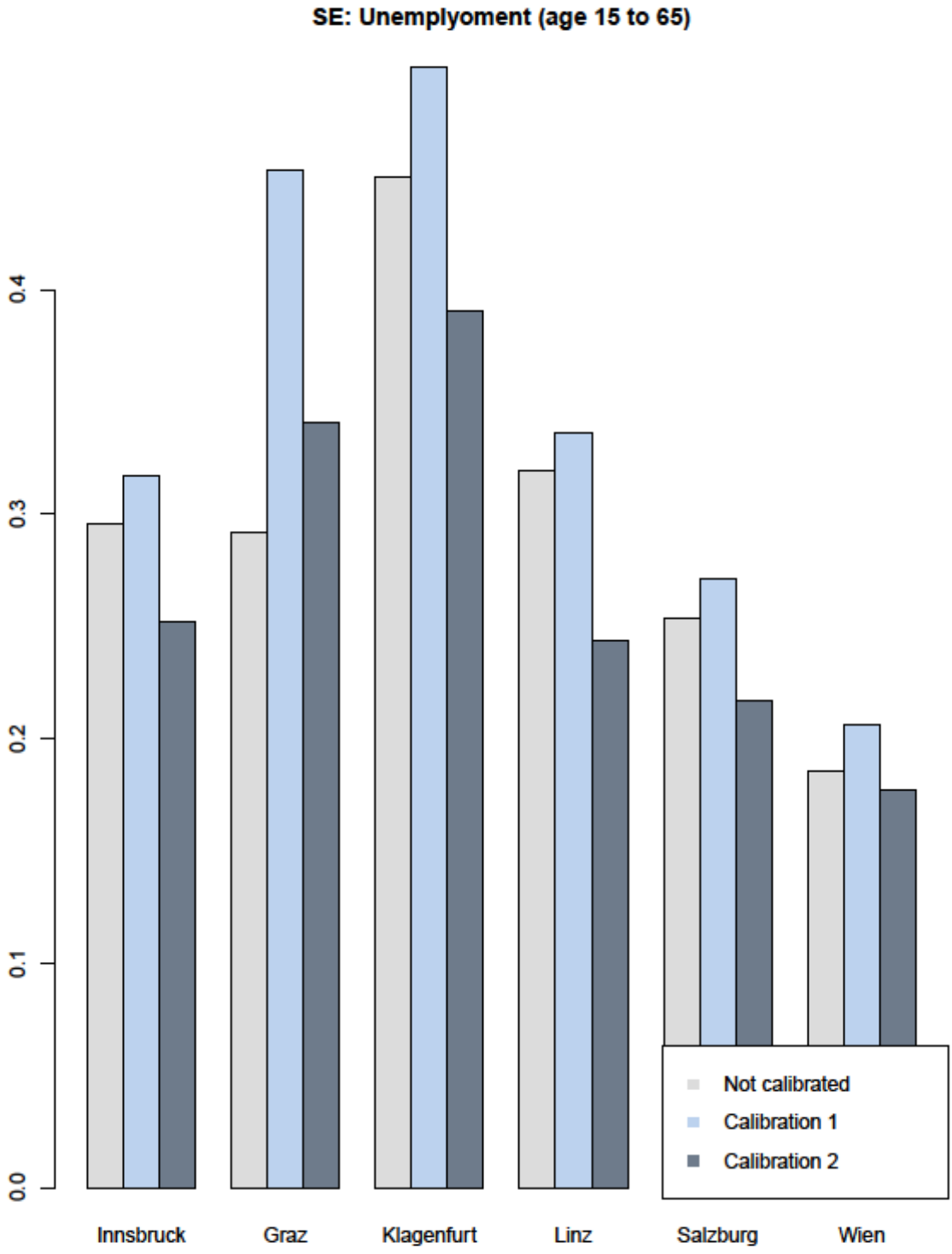
1. Calibrating on sex and age
2. Calibrating on sex, age and employment status (taken from administrative sources)

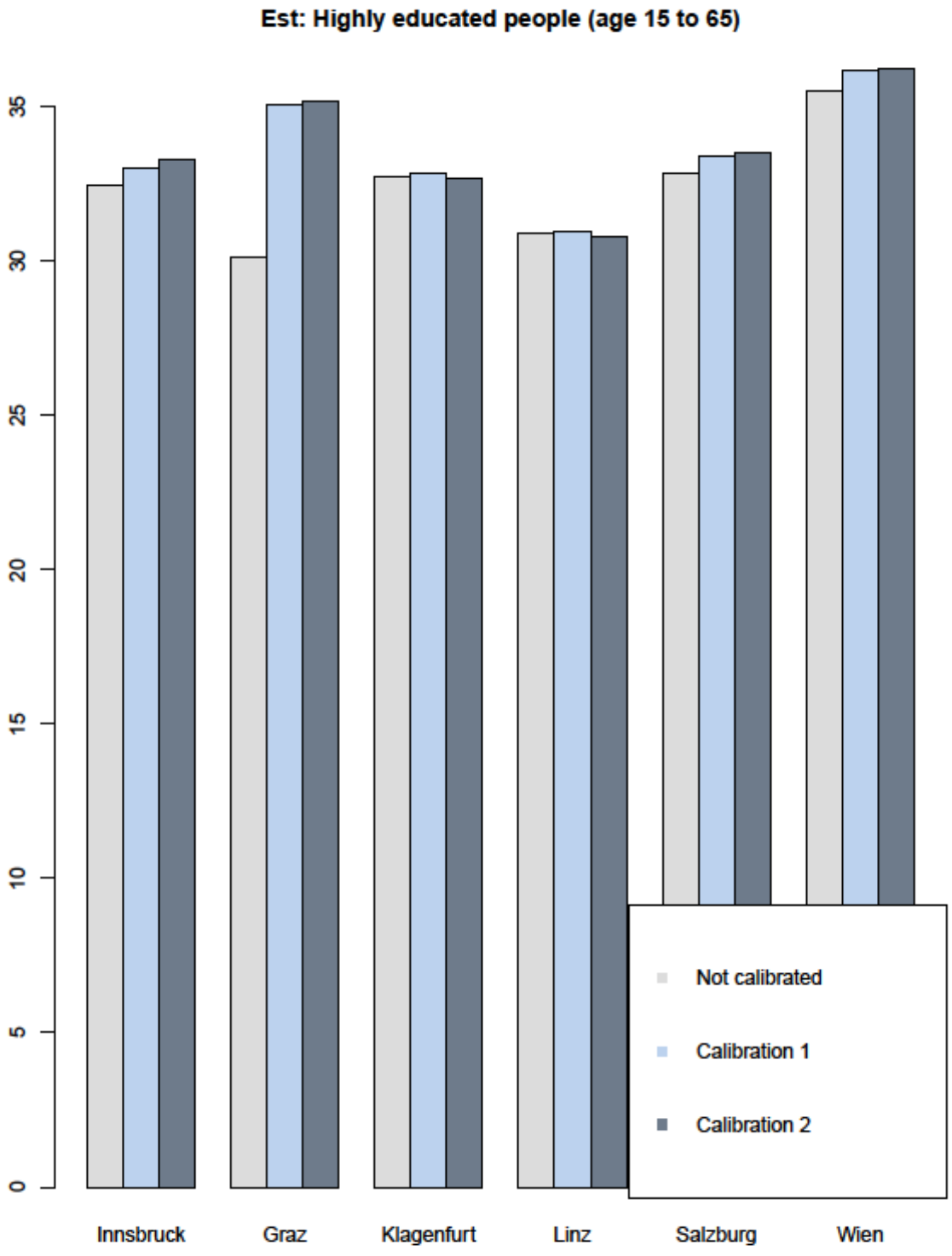
Regarding the results we observe slight changes in the point estimates (see barplots below), which is to be expected due to the recalibration of weights. Moreover we see that, especially for the second approach, the standard errors decrease significantly in most cases (see barplots below). However there are also cases where calibration did not improve the precision of the estimates. Refining this recalibration step could possibly mitigate these issues, but this would require more tests and analysis. Important to note is the issue of recoding administrative data to correspond with the variables of interest in the LFS. For Statistics Austria this did not work in all cases and is still an open issue.

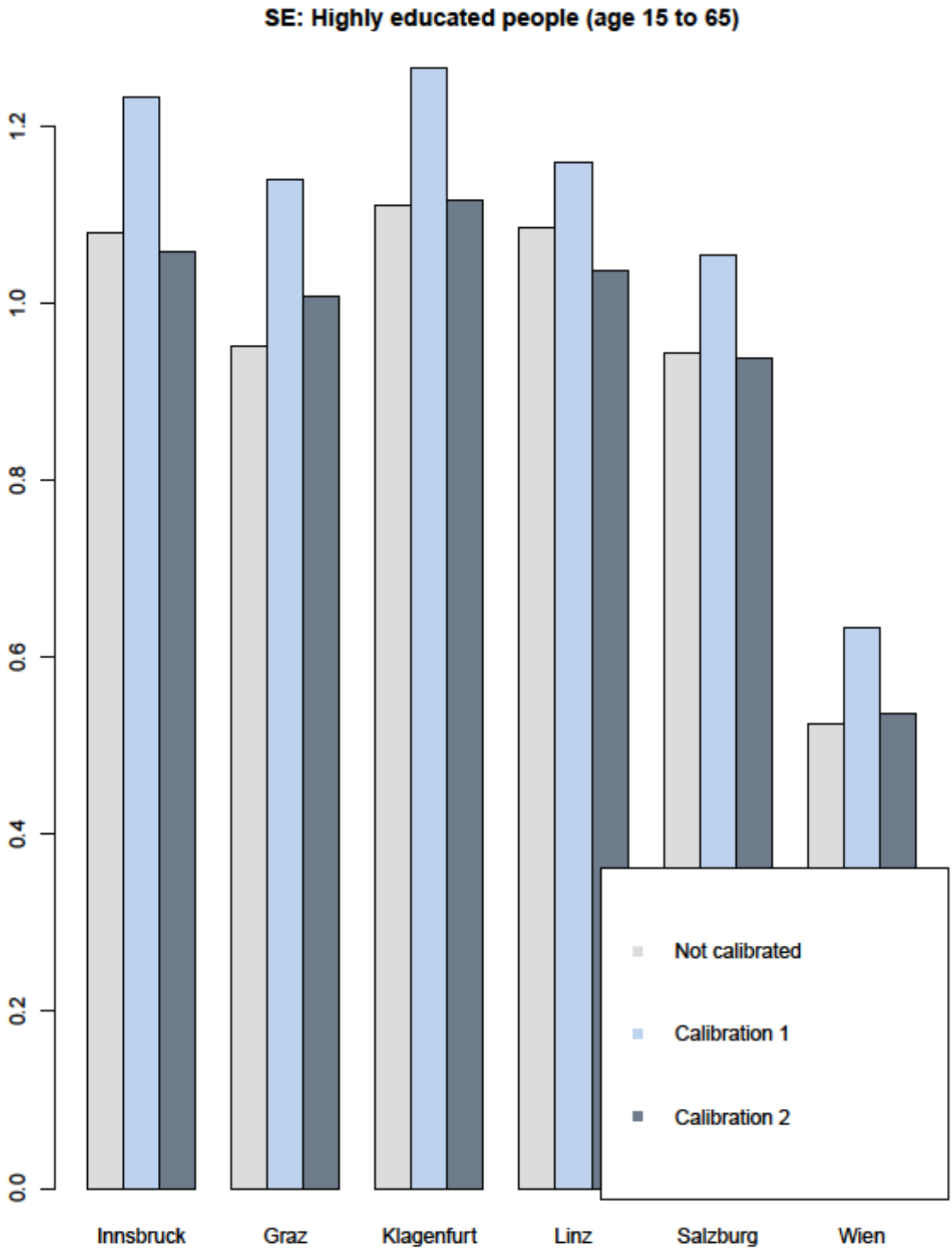

Est: Emplyoment (age 15 to 65)
0
10
20
30
40
50
60
70
Innsbruck
Graz
Klagenfurt
Linz
Salzburg
Wien
Not calibrated
Calibration 1
Calibration 2


SE: Emplyoment (age 15 to 65)
1.0
0.8
0.6
0.4
0.2
0.0
Not calibrated
Calibration 1
Calibration 2
Innsbruck
Graz
Klagenfurt
Linz
Salzburg
Wien


Est: Unemplyoment (age 15 to 65)
0
1
2
3
4
5
6
Innsbruck
Graz
Klagenfurt
Linz
Salzburg
Wien
Not calibrated
Calibration 1
Calibration 2


SE: Unemplyoment (age 15 to 65)
0.0
0.1
0.2
0.3
0.4
Innsbruck
Graz
Klagenfurt
Linz
Salzburg
Wien
Not calibrated
Calibration 1
Calibration 2


Est: Highly educated people (age 15 to 65)
0
5
10
15
20
25
30
35
Innsbruck
Graz
Klagenfurt
Linz
Salzburg
Wien
Not calibrated
Calibration 1
Calibration 2


SE: Highly educated people (age 15 to 65)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Innsbruck
Graz
Klagenfurt
Linz
Salzburg
Wien
Not calibrated
Calibration 1
Calibration 2

Conclusions

Looking at the results for Austria the use of LFS for indicators in FUA's is on one hand very straight forward on the other hand could lack required precision since the resulting standard errors can be quite high. This issue can however be addressed with the use of additional information from administrative data. In our case we used the population margins for variables which highly correlate with education and employment from administrative sources to recalibrate the sampling weights. This did improve the precision of the estimates in almost all cases. However, the improvement was rather limited. In addition we encountered issues when trying to recode the variables from administrative data sources to correspond better with the variables in LFS, which we were not able to overcome in all cases.

ANNEX 1.2

Report on the use of LFS for information on Functional Urban areas

Country: **Belgium**

Date: 10-09-2018

Author(s): Ellen Quintelier and Anja Termote

1. Introduction

The Labour Force Survey (LFS) is the main source in the EU for harmonised labour market statistics. This survey is based on a sample and therefore allows for a limited degree of regional breakdown. Information on NUTS 2 is generally available. UN and OECD developed Functional Urban Areas (FUA's)[41] which play an important role in current policies dedicated to urban regions. One would expect that the populations of Metropolitan urban areas, with population between 500,000 and more, are of the same order as many NUTS 2 regions. The sample size would therefore be in theory large enough to provide LFS statistics. The main issue is that the sample of the LFS was not designed to generate statistics for these regions. This impacts on the quality. Small urban areas are excluded from the analysis, since the corresponding sample sizes are too small. For medium-sized urban areas we will investigate what is possible.

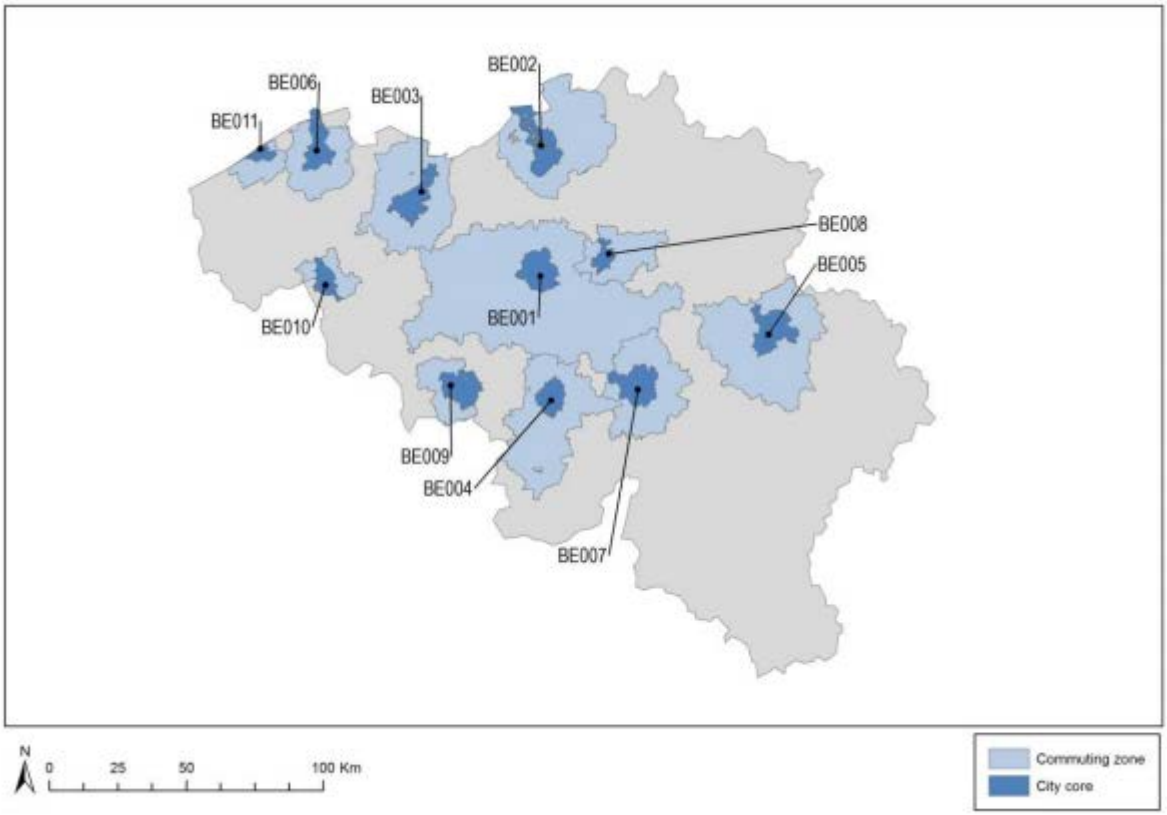
This report presents the results of the analysis carried out for country: **Belgium.**

2. Functional urban areas in Belgium

Belgium has 11 Functional Urban Areas. Brussels is the capital of Belgium and the only large metropolitan area in Belgium, with a total population above 1.5 million (table 1). Antwerpen, Liège and Gent are metropolitan areas with population between 500,000 and 1.5 million. Charleroi, Brugge, Namur and Leuven are medium-sized urban areas with population between 200,000 and 500,000. Mons, Kortrijk and Oostende are small urban areas with population between 50,000 and 200,000. Antwerpen, Brugge, Gent, Kortrijk, Leuven and Oostende are situated in the Dutch-speaking part of Belgium. Charleroi, Liège, Mons and Namur are French-speaking. Brussels has two official languages: both Dutch and French. Figure 1 presents the FUA's on a map.

[41] The boundaries of cities and Functional Urban Areas are available on the Eurostat-GISCO website. The version referred to as "Urban Audit 2011-2014" is still valid for France, Germany and Belgium. Revisions have taken place in the Netherlands and Austria. The revised FUA boundaries are included in the FUA 2015-2018 dataset. . http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/urban-audit

Figure 2. Functional Urban Areas in OECD countries: Belgium



Source: https://www.oecd.org/cfe/regional-policy/functional-urban-areas-all-belgium.pdf

Table 1 : Total population by functional urban area

ID on the map	Name of functional urban area	Total population (2014)
BE002	Antwerpen	1,081,904
BE006	Brugge	224,480
BE001	Brussels	2,588,102
BE004	Charleroi	487,240
BE003	Gent	592,100
BE010	Kortrijk	170,223
BE008	Leuven	220,979
BE005	Liège	739,798
BE009	Mons	194,876
BE007	Namur	219,218
BE011	Oostende	128,083

Source: https://www.oecd.org/cfe/regional-policy/functional-urban-areas-all-belgium.pdf

3. Results of LFS

In 2017 almost 98,000 respondents aged 15-64 participated to the Belgian LFS (= sample for the whole country, sum of all quarterly samples). 59,600 of them are unique respondents. Table 2a presents the number of respondents that were surveyed in each FUA in 2017. The number of respondents aged 15-64 varies from 986 in Oostende to 27,271 in Brussels.

If we compare the size of the 11 FUA's with the size of the 11 provinces (NUTS II) in Belgium (table 2b), we see that the smallest province in Belgium (Luxemburg) has a population aged 15-64 of 183,000 which is more than the population aged 15-64 in 6 FUA's (Oostende, Mons, Kortrijk, Namur, Brugge and Leuven). But the sample size (15-64) of the smallest province (Luxemburg) (5,100) is much higher than the sample size of all FUA's, except Brussels (27,300), Antwerp (5,800) and Liège (5,600).

Table 2a. Number of respondents and weighted population within each FUA – population aged 15-64 (2017)

Name of functional urban area	Number of respondents	% of respondents	weighted population in FUA	% of weighted population in FUA
Antwerpen	5,807	10.1	688,265	15.6
Brugge	2,421	4.2	154,623	3.5
Brussels	27,271	47.2	1,732,280	39.4
Charleroi	4,573	7.9	325,594	7.4
Gent	3,927	6.8	408,715	9.3
Kortrijk	1,712	3.0	112,394	2.6
Leuven	1,937	3.4	171,465	3.9
Liège	5,568	9.6	469,189	10.7
Mons	1,340	2.3	111,567	2.5
Namur	2,196	3.8	139,998	3.2
Oostende	986	1.7	85,144	1.9
Total/Mean (all FUA's)	57,738	100	4,399,234	100

Table 2b. Number of respondents and weighted population within each province (NUTS II) – population aged 15-64 (2017)

Name of province (NUTS II)	Number of respon-dents	% of respondents	weighted population in province	% of weighted population in province
Antwerpen	10,016	10.2	1,171,201	16.1
Brussel	13,835	14.1	795,839	11
West-Vlaanderen	11,134	11.4	735,154	10.1
Oost-Vlaanderen	9,942	10.2	957,983	13.2
Henegouwen	10,299	10.5	858,771	11.8
Luik	10,341	10.6	707,225	9.7
Limburg	7,575	7.7	561,826	7.7
Luxemburg	5,079	5.2	183,012	2.5
Namen	5,318	5.4	316,849	4.4
Vlaams-Brabant	9,224	9.4	721,405	9.9
Waals-Brabant	5,183	5.3	256,512	3.5
Total/Mean (all FUA's)	97,946	100	7,265,778	100

3.1. Employment rate (15-64)

On average, 62% of the population aged 15-64 is employed. This ranges from less than 55% in Charleroi, Liège and Mons to more than 70% in Brugge, Kortrijk and Leuven (**Table** 3 and figure 2). It has to be noted that the trends in these FUA over time seem not identical: while the employment rate has increased in some FUA's, the employment rate has stabilized or even decreased in others (Table 4). But the evolutions are not significant so we have to interpret them very carefully (see confidence intervals in part 4). Brussels' employment rate (62%) is more or less equal to the average. This probably has to do with the large area that is considered as FUA Brussels, and which combines areas with high and low employment rates (see also point 4. Issues and concerns in producing indicators for FUA's). Brussels Capital Region (which is much smaller than the FUA of Brussels) had a mean employment rate of 56,2%[42] in 2017, which would add them to the FUA's with rather low levels of employment.

[42] Employment rate, https://statbel.fgov.be/en/themes/work-training/labour-market/employment-and-unemployment%23figures%20

Table 3. Employment rate (15-64) within each FUA (2017)

Name of functional urban area	Number of employed in sample	Employment rate (based on unweighted numbers) (%)	Number of employed (weighted numbers)	Employment rate (%)
Antwerpen	3,714	64.0	442,902	64.4
Brugge	1,751	72.3	111,719	72.3
Brussels	16,625	61.0	1,076,236	62.1
Charleroi	2,479	54.2	173,102	53.2
Gent	2,741	69.8	278,112	68.1
Kortrijk	1,187	69.3	79,054	70.3
Leuven	1,437	74.2	124,320	72.5
Liège	3,017	54.2	246,904	52.6
Mons	642	47.9	51,040	45.8
Namur	1,355	61.7	83,838	59.9
Oostende	631	64.0	53,918	63.3
Total/Mean (All FUA's)	35,579	61.6	2,721,144	61.9

Table 4. Employment rate (15-64) within each FUA per year (2015-2017)

Name of functional urban area	2015 (%)	2016 (%)	2017 (b) (%)
Antwerpen	63.3	62.3	64.4
Brugge	66.7	70.5	72.3
Brussels	60.5	60.5	62.1
Charleroi	49.1	51.5	53.2
Gent	67.5	69.5	68.1
Kortrijk	69.3	69.2	70.3
Leuven	68.7	72.6	72.5
Liège	54.1	53.7	52.6
Mons	50.4	48.8	45.8
Namur	59.7	61.0	59.9
Oostende	61.9	59.1	63.3
Total/Mean (All FUA's)	60.3	60.9	61.9

Figure 3. Employment rate in the FUA's (2017)

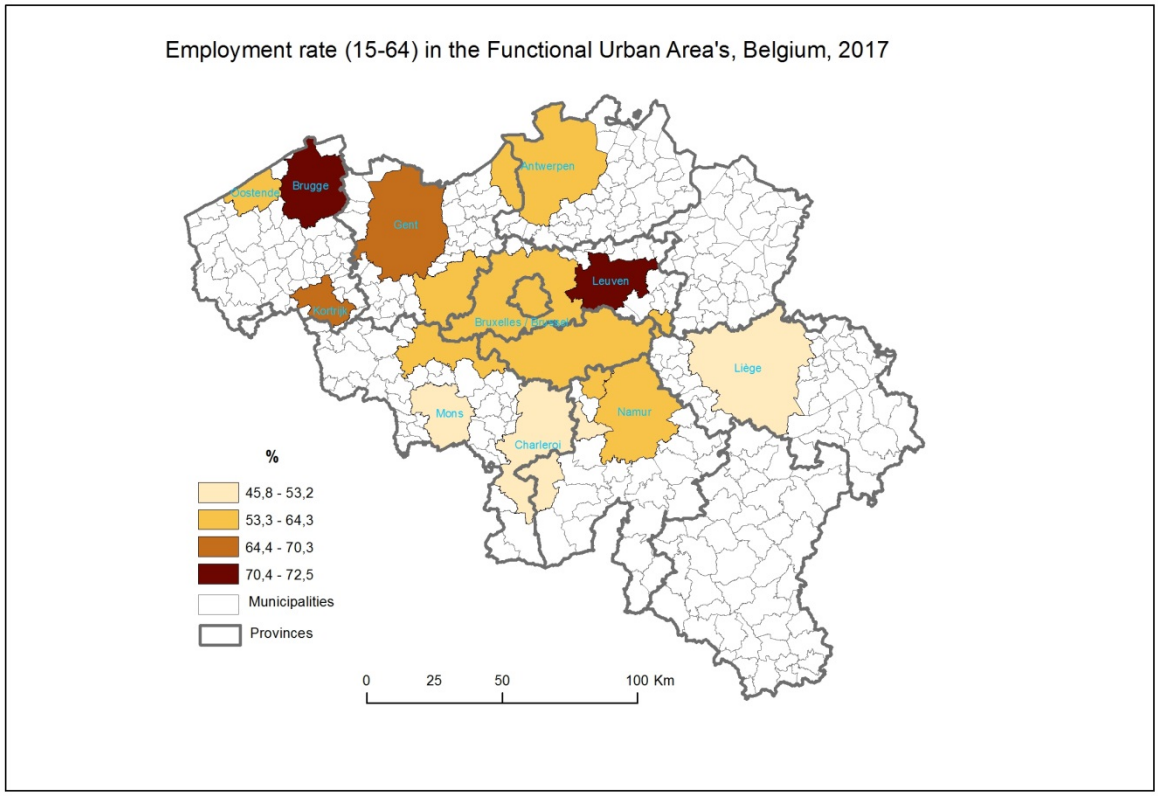


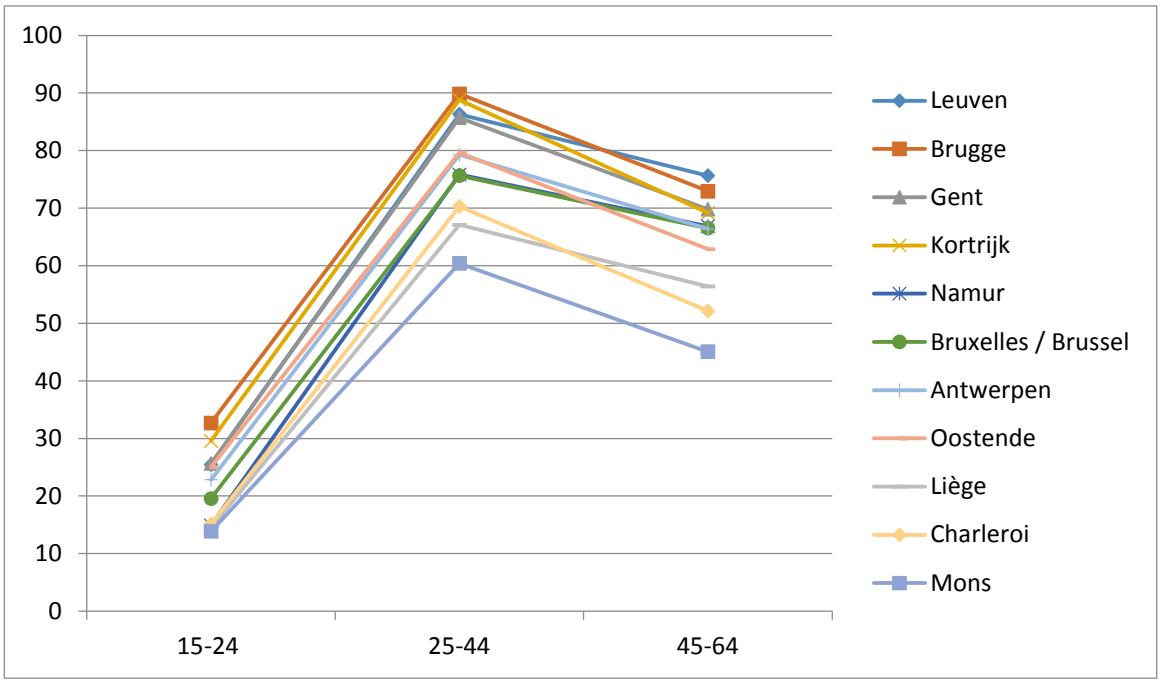
Table 5 and figure 3 present the employment rate among 3 age groups: 15-24, 25-44 and 45-64 years olds. We observe very similar differences between FUA's with respect to their employment rate, within the three different age groups. Given the strong impact of age on employment rate, this suggest that the differences between FUA's cannot be explained to the different composition of FUA's in terms of age.

However, the variation is smaller among the youngest age groups (about 20 percentage points) than the older ones (about 30 percentage points). The mean employment rate of the 45-64 is about 12 percentage points lower than those of the 25-44 year olds. The Europe 2020 strategy targets 75% of the people between 20 and 64 to be employed. Most FUA's reach this threshold among the 25-44 year olds, yet only one FUA also among the 45-64 year olds.

Table 5. Employment rate within each FUA by age group (2017)

	15-24		25-44		45-64	
Name of functional urban area	Number of employed	Employment rate (%)	Number of employed	Employment rate (%)	Number of employed	Employment rate (%)
Antwerpen	26,717	22.8	227,729	79.3	188,456	66.4
Brugge	8,899	32.7	52,583	89.8	50,237	73.0
Brussels	60,030	19.6	564,069	75.6	452,137	66.5
Charleroi	8,658	15.1	94,736	70.3	69,707	52.2
Gent	19,210	25.7	139,540	85.7	119,362	69.8
Kortrijk	5,548	29.6	39,564	88.8	33,942	69.2
Leuven	7,049	25.5	68,808	86.3	48,462	75.7
Liège	13,005	14.5	124,304	67.1	109,594	56.4
Mons	2,853	13.9	28,277	60.4	19,910	45.1
Namur	4,214	14.9	41,647	75.8	37,977	66.8
Oostende	3,118	25.0	24,218	79.6	26,582	62.9
Total/Mean (All FUA's)	159,301	20.4	1,405,475	76.8	1,156,366	64.7

Figure 3. Employment rate within each FUA by age group (2017)

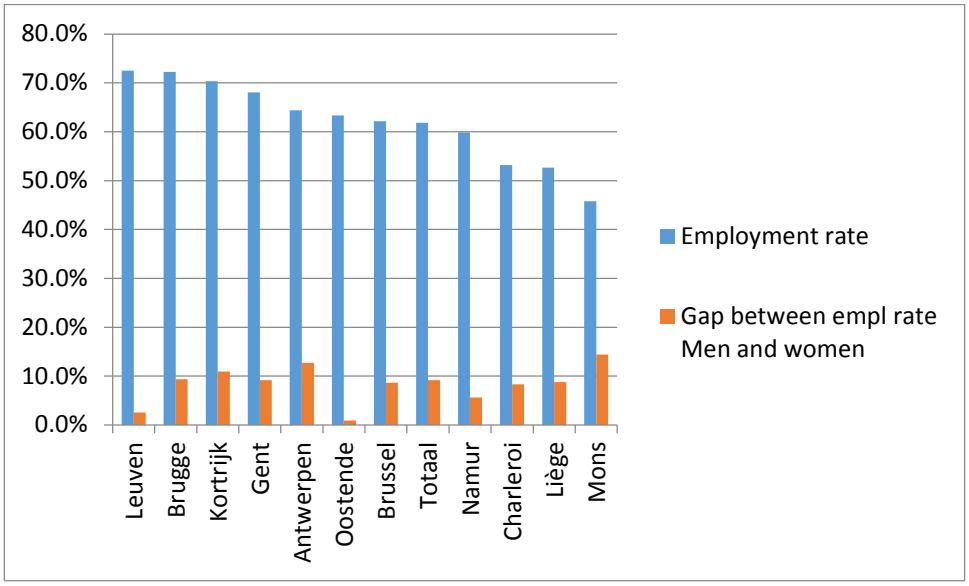


Looking at differences according to gender (table 6), we see that, on average, the employment rate for men is around 9 percentage points higher than the figure for women. In every FUA, the male employment rate is higher than the overall percentage of employed. The gap is largest in Mons (14 percentage points) and smallest in Oostende (1 percentage point). The gap is above 10 percentage points in Mons, Antwerp and Kortrijk (figure 4) and less than 3 percentage points in Oostende and Leuven. There seems no clear relation between the level of the employment rate and the size of the gap between the male and female employment rate.

Table 6. Employment rate within each FUA by gender (2017)

	Men		Women	
Name of functional urban area	Number of employed	Employment rate (%)	Number of employed	Employment rate (%)
Antwerpen	247,282	70.6	195,619	57.9
Brugge	59,169	77.0	52,551	67.6
Brussels	574,731	66.5	501,506	57.8
Charleroi	95,760	57.2	77,342	48.9
Gent	146,382	72.7	131,730	63.5
Kortrijk	43,162	75.7	35,892	64.8
Leuven	65,983	73.7	58,336	71.2
Liège	131,525	57.1	115,378	48.3
Mons	28,115	53.4	22,925	39.0
Namur	43,142	62.8	40,695	57.1
Oostende	27,572	63.8	26,346	62.9
Total/Mean (all FUA's)	1,462,823	66,4	1,258,319	57,3

Figure 4: Employment rate (15-64) and gap between the employment rate of men and women by FUA (2017)



3.2 Unemployment rate

Charleroi, Liège and Mons do not only have the lowest employment rates, they also have the highest unemployment rate (more than 11%) (Table 7 and Figure 5). Kortrijk and Brugge on the other hand, have an unemployment rate of around 3%, which is seen as a sort of lower limit so there is almost full employment in these functional urban areas. The unemployment rate in Leuven and Oostende is around 5%. Brussels and Antwerp are somewhat in the middle with an unemployment rate of 9,5% in Brussels and 7,1% in Antwerp.

Table 8 present the unemployment rates for 2015-2017. About the trends in unemployment rate between 2015-2017 we cannot say a lot since the precision of the estimates is limited (see also part 4).

Table 7. Unemployment rate (15-64) within each FUA (2017)

Name of functional urban area	Number of unemployed in sample	Unemployment rate (based on unweighted numbers) (%)	Number of unemployed (weighted)	Weighted unemployment rate (%)
Antwerpen	230	5.8	33,676	7.1
Brugge	47	2.6	3,402	3.0
Brussels	1,821	9.9	113,350	9.5
Charleroi	278	10.1	21,864	11.2
Gent	88	3.1	10,198	3.5
Kortrijk	30	2.5	2,152	2.7
Leuven	57	3.8	6,535	5.0
Liège	349	10.4	36,563	12.9
Mons	88	12.1	8,242	13.9
Namur	100	6.9	7,760	8.5
Oostende	28	4.3	2,887	5.1
Total/Mean (all FUA's)	3,116	8.1	246,628	8.3

Figure 5. Unemployment rate in the FUA's (2017)

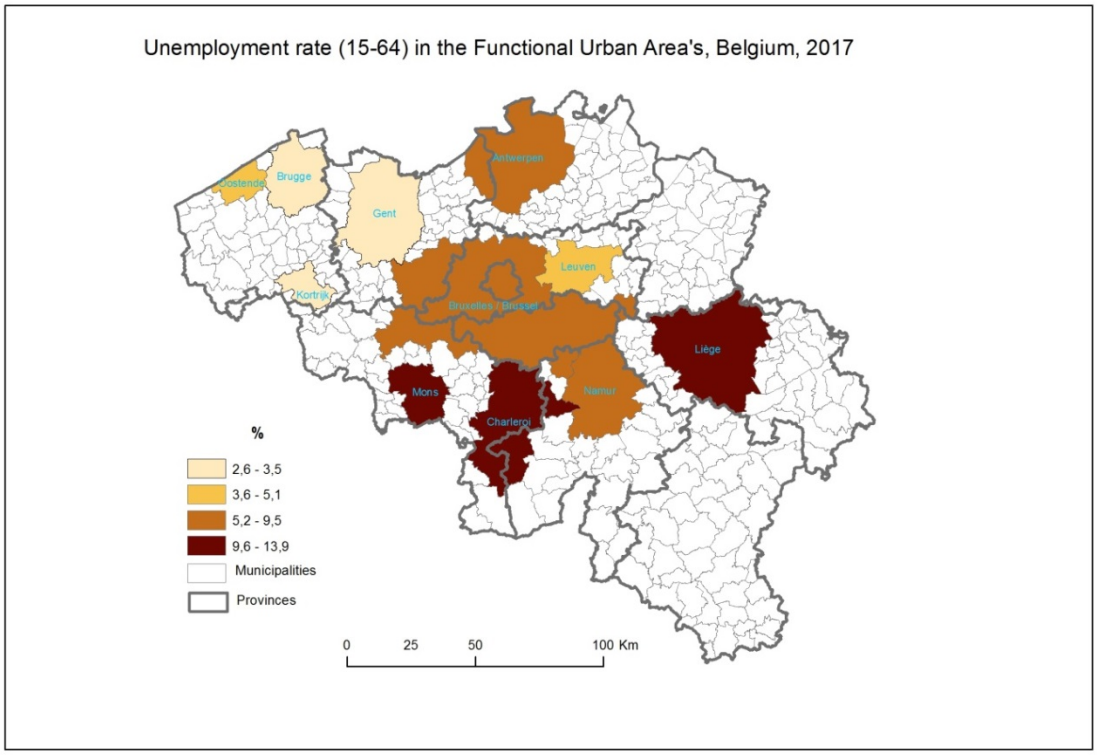


Table 8. Unemployment rate (15-64) within each FUA per year (2015-2017)

Name of functional urban area	2015 (%)	2016 (%)	2017 (b) (%)
Antwerpen	7.2	7.0	7.1
Brugge	3.8	3.7	3.0

Brussels	11.1	11.4	9.5
Charleroi	15.0	13.2	11.2
Gent	4.8	4.1	3.5
Kortrijk	4.9	3.3	2.7
Leuven	5.0	3.5	5.0
Liège	14.4	12.6	12.9
Mons	15.9	14.4	13.9
Namur	10.1	9.1	8.5
Oostende	6.1	6.9	5.1
Total/Mean (all FUA's)	9.9	9.3	8.3

The unemployment rate is in general highest among young people, ranging from less than 6% in Kortrijk to 31% in Charleroi. Among 25 to 44 year olds these percentages and consequently the gap are/is much lower, ranging from 2 to 15% and 2 to 10% among the persons aged 45 to 64. Brussels and Antwerp are quite different in this respect: they both have a high unemployment rate among young people, but this declines to 6% among the 25-44 year olds in Antwerp and to 10% within this age group in Brussels.

Table 9. Unemployment rate within each FUA by age (2017)

	15-24		25-44		45-64	
Name of functional urban area	Weighted number of unemployed	Weighted unemployment rate (%)	Weighted number of unemployed	Weighted unemployment rate (%)	Weighted number of unemployed	Weighted unemployment rate (%)
Antwerpen	5,958	18.2	14,041	5.8	13,677	6.8
Brugge	1,127	11.2	1,130	2.1	1,145	2.2
Brussels	14,994	20.0	63,048	10.1	35,308	7.2
Charleroi	3,878	30.9	12,481	11.6	5,505	7.3
Gent	2,958	13.3	4,681	3.3	2,559	2.1
Kortrijk	328	5.6	1,038	2.6	786	2.3
Leuven	1,437	16.9	2,656	3.7	2,442	4.8
Liège	6,009	31.6	19,213	13.4	11,341	9.4
Mons	1,168	29.1	4,881	14.7	2,193	9.9
Namur	1,712	28.9	4,446	9.6	1,603	4.1
Oostende	236	7.0	1,324	5.2	1,327	4.8
Total/Mean (all FUA's)	39,805	20.0	128.939	8.4	77.886	6.3

Whereas the employment rate of men is higher than that of women in every FUA, the situation for unemployment is different for different FUA's. For instance women are more frequently unemployed than men in Leuven, Antwerp and Kortrijk, while men are more unemployed in Charleroi, Liège and Namur. So while the gender gap is overall limited, the figures/observations are different for different FUA's.

Table 10. Unemployment rate within each FUA by gender (2017)

	Men		Women	
Name of functional urban area	Weighted number of unemployed	Weighted unemployment rate	Weighted number of unemployed	Weighted unemployment rate

		(%)		(%)
Antwerpen	15,654	6.0	18,022	8.4
Brugge	1,831	3.0	1,571	2.9
Brussels	63,733	10.0	49,617	9.0
Charleroi	13,312	12.2	8,552	10.0
Gent	6,011	3.9	4,187	3.1
Kortrijk	974	2.2	1,178	3.2
Leuven	2,479	3.6	4,056	6.5
Liège	21,714	14.2	14,849	11.4
Mons	4,661	14.2	3,581	13.5
Namur	4,714	9.9	3,047	7.0
Oostende	1,760	6.0	1,127	4.1
Total/Mean (All FUA's)	136,843	8.6	109,787	8.0

3.3. Educational attainment of population 25-64 years of age

Table 11 presents the number of respondents aged 25-64 by educational attainment and FUA. Table 12 and Figure 6 present the weighted figures.

Table 11. Number of respondents aged 25-64 within each FUA by educational attainment (2017)

Name of functional urban area	Low	Medium	High	Total
Antwerpen	1,139	1,992	1,753	4,884
Brugge	401	811	813	2,025
Brussels	5,164	6,962	10,520	22,646
Charleroi	1,193	1,598	1,059	3,850
Gent	709	1,07	1,533	3,312
Kortrijk	297	489	0,655	1,441
Leuven	141	475	1,034	1,650
Liège	1,436	1,564	1,566	4,566
Mons	368	400	360	1,128
Namur	361	669	776	1,806
Oostende	212	379	271	862
Total/Mean (all FUA's)	11,421	16,409	20,340	48,170

All large Belgian cities have one or more institutes for higher education. Antwerp, Brussels, Gent, Leuven, Liège, Mons and Namur have one or more universities. So in general, the Belgian citizens have many opportunities to obtain a higher education degree close to home. However, large discrepancies arise: in Leuven 67% of persons aged 25-64 has a higher education degree, followed by 49% in Gent and 47% in Brussels. On the lower side of the spectrum we have Charleroi with 28% and Mons with 30% higher educated people.

Table 12. Educational attainment (25-64) within each FUA (2017)

	Low		Middle		High	
Name of functional	Weighted number	Weighted %	Weighted number	Weighted %	Weighted number	Weighted %

urban area						
Antwerpen	124,376	21.8	236,115	41.3	210,668	36.9
Brugge	23,939	18.8	51,064	40.1	52,401	41.1
Brussels	308,709	21.7	445,613	31.3	671,060	47.1
Charleroi	83,806	31.2	110,774	41.3	73,837	27.5
Gent	67,904	20.3	103,530	31.0	162,382	48.6
Kortrijk	19,221	20.5	31,829	34.0	42,590	45.5
Leuven	10,062	7.0	37,604	26.2	96,111	66.8
Liège	123,467	32.5	123,752	32.6	132,389	34.9
Mons	31,291	34.4	32,422	35.6	27,276	30.0
Namur	22,488	20.1	39,323	35.2	49,935	44.7
Oostende	16,229	22.3	31,383	43.2	25,063	34.5
Total/mean (all FUA's)	831,492	23.0	1,243,409	34.4	1,543,712	42.7

Figure 6. % higher educated (25-64) in the FUA's (2017)

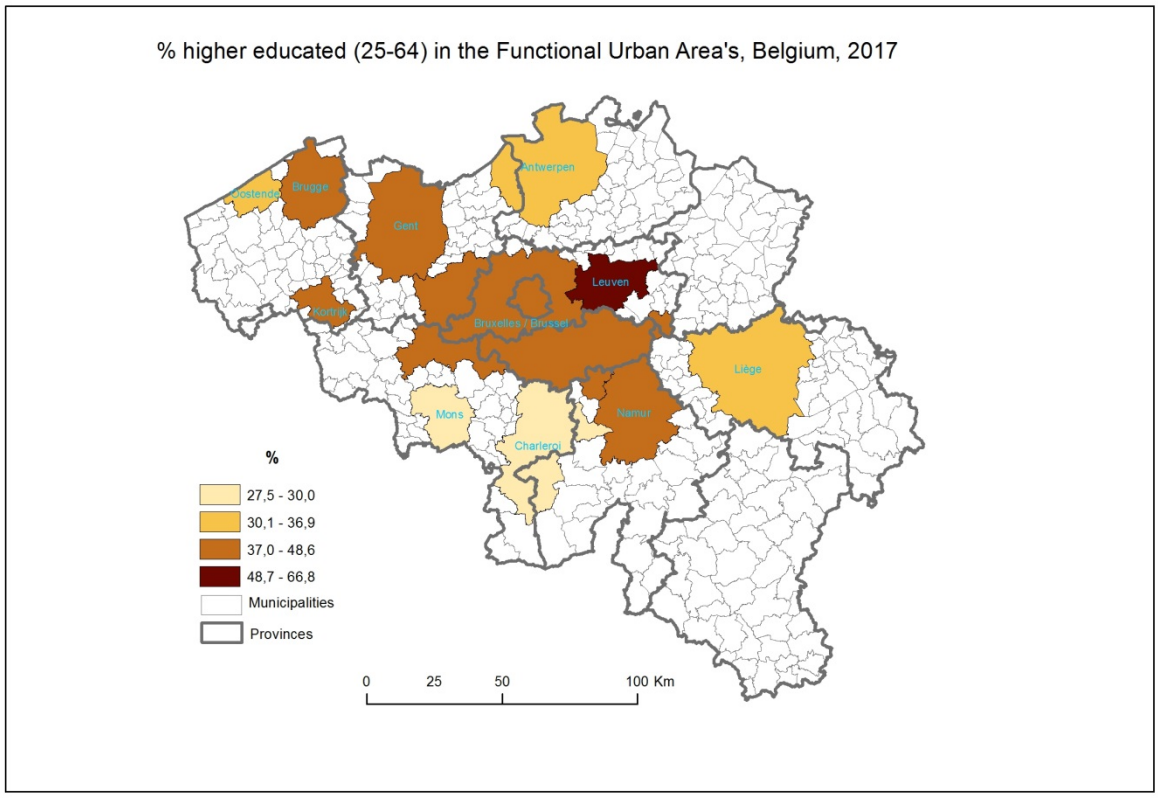


Table 13 presents the percentage of higher educated persons aged 25-64 for the period 2015-2017. In most of the FUA's this percentage raised between 2015 and 2017 except in Antwerp. But these trends have to be interpreted very carefully (see further). Confidence intervals for 2016 are presented in part 4.

Table 13. % higher educated (25-64) within each FUA per year (2015-2017)

Name of functional urban area	2015 (%)	2016 (%)	2017 (b) (%)
Antwerpen	38,5	34,7	36,9
Brugge	36,6	40,2	41,1
Brussels	44,2	44,5	47,1
Charleroi	24,5	25,0	27,5
Gent	43,1	47,3	48,6
Kortrijk	38,1	41,8	45,5
Leuven	52,3	59,9	66,8
Liège	34,2	32,2	34,9
Mons	29,7	32,5	30,0
Namur	41,8	41,4	44,7
Oostende	27,3	28,5	34,5
Total/mean (All FUA's)	39.6	39.9	42.7

Table 14 shows the educational attainment of persons aged 25-64 by age groups 25-44 and 45-64. Among the oldest age group, people in Leuven, Gent, Brussels and Namur have much more higher educated diploma's (i.e. more than 40%) than in Oostende, Mons and Charleroi (i.e. less than 30%). The Europe 2020 strategy targets that at least 40% of people aged 30–34 having completed higher education. Most FUA's almost reach this threshold among the 25-44 year olds.

Table 14. Educational attainment within each FUA by age group (2017)

	25-44			45-64		
Name of functional urban area	Low (%)	Middle (%)	High (%)	Low (%)	Middle (%)	High (%)
Antwerpen	17.2	43.1	39.7	26.5	39.5	34.0
Brugge	10.9	39.6	49.6	25.5	40.5	33.9
Brussels	17.5	30.5	52.0	26.2	32.1	41.7
Charleroi	25.9	40.4	33.7	36.6	42.1	21.3
Gent	13.6	30.7	55.7	26.8	31.3	41.9
Kortrijk	10.1	34.8	55.1	30.0	33.2	36.8
Leuven	4.7	24.0	71.2	9.8	28.8	61.4
Liège	29.3	34.1	36.6	35.6	31.2	33.2
Mons	29.2	34.7	36.2	39.9	36.7	23.4
Namur	14.3	37.2	48.5	25.8	33.3	41.0
Oostende	16.7	40.8	42.6	26.4	44.9	28.7
Total/mean (All FUA's)	18.2	34.2	47.6	27.9	34.5	37.6

In every FUA, women are higher educated than men. This gap is largest in Brugge: 12 percentage points more high educated women than men, and smallest in Kortrijk (1 percentage point).

Table 15. Educational attainment (25-64) within each FUA by gender (2017)

	Men	Women

Name of functional urban area	Low	Middle	High	Low	Middle	High
Antwerpen	23.0	42.0	35.0	20.5	40.6	38.9
Brugge	19.3	45.6	35.1	18.3	34.5	47.2
Brussels	22.1	33.6	44.3	21.2	29.0	49.8
Charleroi	30.9	45.2	23.9	31.5	37.2	31.3
Gent	21.8	33.6	44.7	19.0	28.6	52.5
Kortrijk	22.4	32.5	45.1	18.6	35.5	45.9
Leuven	7.0	28.1	64.9	7.0	24.1	69.0
Liège	34.2	34.7	31.2	31.0	30.6	38.4
Mons	35.6	38.3	26.2	33.3	33.3	33.4
Namur	21.0	39.2	39.8	19.4	31.4	49.3
Oostende	23.4	46.6	30.1	21.4	39.9	38.7
Total/mean (All FUA's)	23.7	36.6	39.6	22.2	32.1	45.7

4. Issues and concerns in producing indicators for FUA's

First, we give more information about the sampling method, stratification, extrapolation and publication thresholds concerning the Belgian LFS. After that we look at the consequences at the level of FUA's.

General issues about the Belgian LFS: sampling method, stratification, extrapolation, publication thresholds

The first stage sampling frame, i.e. the PSU sampling frame, consists of geographic areas, which are either 'statistical sections' or regroupings of statistical sections within 'statistical letters' or sub-municipalities. Each PSU must contain a minimum number of eligible private households: if a PSU is sampled in the first stage, it must be possible to select enough households (26 in Brussels Capital Region, and 23 in other sampling strata) to form a 'group of households' (which will be assigned as a whole to an interviewer). There are about 6,354 PSUs in the sampling frame, containing 676 households on average; 'small' sections only represent 0.15 % of the total number of households. Systematic probability proportional to size sampling (PPS-SYS; where a proxy for the number of eligible private households is used as size measurement) is applied in each first stage sampling stratum. The number of PSU drawn in each sampling stratum is fixed in advance, such that 6,695 HHs are selected in total per quarter in the second stage. Larger PSUs can be selected more than once, while smaller PSUs are probably not selected.

The number of households selected in stage 2 (the final sampling stage) in a selected PSU equals the number of times the PSU is selected in the first stage times the size of the groups of households to be formed in the PSU (i.e. 26 or 23). Basically, simple random sampling (SRS) is applied to select households (SSUs or FSUs) in each selected PSU. If a PSU is selected only once in the first sampling stage, SRS is applied to a frame of eligible households, where eligibility means that a household contains at least one person in age class 15-76 years at the reference Sunday (i.e. the last day of the reference week, which is determined after the first sampling stage, as mentioned before). If a PSU is selected more than once, the 'groups' of households are selected (by SRS) one after another, where the frame of eligible households is adapted to the specific group (i.e. to its reference week/Sunday), taking into account that households should be selected for only one group.

In the first stage of sampling, the PSUs are explicitly stratified by NUTS II regions (i.e. provinces; the Brussels Capital Region and the German community are separate strata). The PSU sampling frame, within each stratum, is further sorted on (1) the quintile of the number of private households, (2) the quintile of the unemployment rate and (3) the quintile of the average household income. This implies implicit stratification on the latter three PSU characteristics within each explicit stratum.

The data are extrapolated to the population at the individual level, including adjustments for non-response (in the first and consecutive waves). The population and the sample are the subject of a posteriori stratification by NUTS II (province), sex and age (5-year age groups).

In our publications and when disseminating data we advise users to interpret results based on small sample sizes (extrapolated figures < 5,000) very carefully. The largest geographical detail that is published by Statbel (Statistics Belgium) is the level of NUTS II but the number of variables and their detail is very limited (ILO-status by sex, 3 age groups and educational attainment (low, medium, high).

It has to be noted that a break in time series occurred in 2017 as the Belgian LFS has undergone a major reform in 2017 with the introduction of a 2(2)2 panel design, the introduction of mixed mode data collection, the introduction of the wave approach and the change of the calibration method.

Consequences of the Belgian LFS methodology at the level of FUA's

As explained above, the stratification of the Belgian LFS is done at NUTS II level (provinces), not at NUTS III level or lower. So the Belgian LFS was not designed to have representative figures at the level of FUA's. The number of PSU's per FUA varies between less than 30 in Oostende (17), Kortrijk (20) and Mons (26) to more than 330 in Brussels (figures for 2017).

Looking at the estimates based on LFS we see that the estimates are quite reliable for **employment and educational attainment**: FUA's are sufficiently large (in terms of population), and it is even possible to divide into smaller categories (as age, sex, etc.) for most FUA's. On the other hand, the **unemployment** becomes quickly too small to divide into smaller categories. The global unemployment rates of Oostende and Kortrijk for instance are based on only 28 and 30 respondents.

The confidence intervals permit to distinguish between higher and lower 'scoring' FUA's (see figures 7, 8 and 9). But we see that for most of the FUA's the confidence intervals are too large to distinguish real evolutions from year to year. Tables 16, 17 and 18 present estimates of the employment rate, the unemployment rate and the percentage of higher educated people for 2015, 2016 and 2017 together with the confidence intervals of 2016. We note large intervals for medium-sized and small urban areas so the precision of the estimates is limited.

Important to notice is that it is known that Brussels Capital has high unemployment rates and low employment rates which puts them at the resp. higher/lower end. The surrounding municipalities however show quite different figures for both employment and unemployment. As a consequence, the numbers for the FUA Brussels represent averages for

Brussels Capital Region on the one hand and the surrounding municipalities on the other. By focusing on the FUA instead of Brussels Capital Region, we get a quite different picture and very valuable information is lost.

Figure 7. Confidence intervals employment rate (2016)

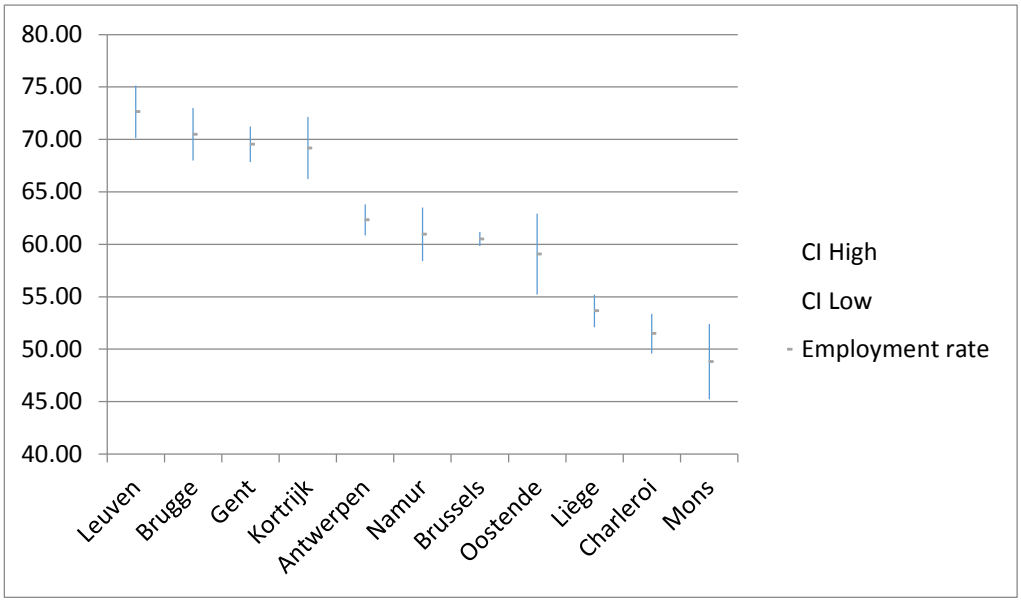


Table 16: Employment rates (15-64) by FUA with confidence interval for 2016

	2015	2016			2017
		Estimate 2016	CI lower limit	CI higher limit	
Antwerpen	63,3%	62,3%	60,9%	63,8%	64,4%
Brugge	66,7%	70,5%	68,0%	73,0%	72,3%
Brussels	60,5%	60,5%	59,8%	61,2%	62,1%
Charleroi	49,1%	51,5%	49,6%	53,4%	53,2%
Gent	67,5%	69,5%	67,8%	71,2%	68,1%
Kortrijk	69,3%	69,2%	66,2%	72,1%	70,3%
Leuven	68,7%	72,6%	70,1%	75,1%	72,5%
Liège	54,1%	53,7%	52,1%	55,2%	52,6%
Mons	50,4%	48,8%	45,2%	52,4%	45,8%
Namur	59,7%	61,0%	58,4%	63,5%	59,9%
Oostende	61,9%	59,1%	55,2%	62,9%	63,3%

Figure 8. Confidence intervals unemployment rate (2016)

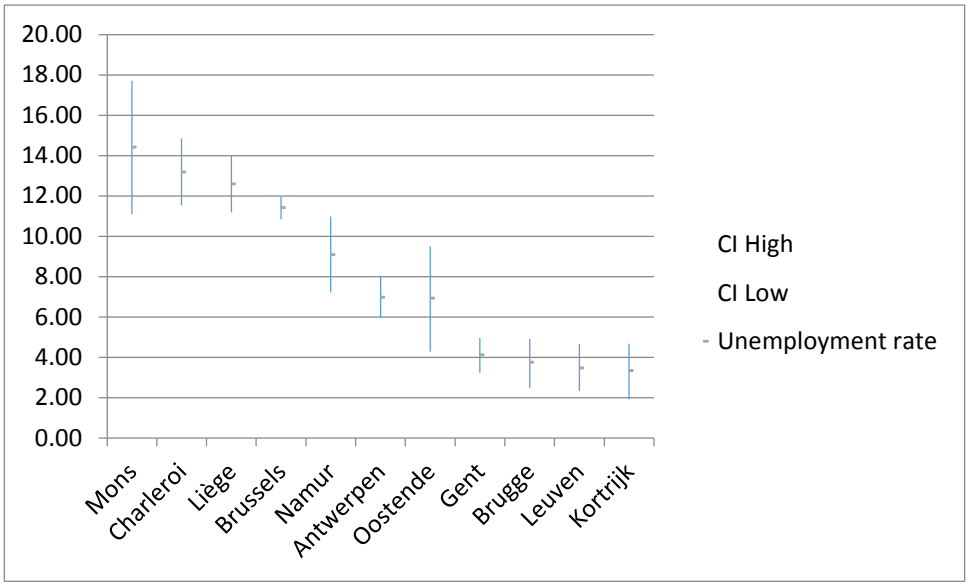


Table 17: Unemployment rates (15-64) by FUA with confidence interval for 2016

	2015	2016			2017
		Estimate 2016	CI lower limit	CI higher limit	
Antwerpen	7,2%	7,0%	6,0%	8,0%	7,1%
Brugge	3,8%	3,7%	2,5%	4,9%	3,0%
Brussels	11,1%	11,4%	10,8%	12,0%	9,5%
Charleroi	15,0%	13,2%	11,5%	14,9%	11,2%
Gent	4,8%	4,1%	3,2%	5,0%	3,5%
Kortrijk	4,9%	3,3%	1,9%	4,7%	2,7%
Leuven	5,0%	3,5%	2,3%	4,7%	5,0%
Liège	14,4%	12,6%	11,2%	14,0%	12,9%
Mons	15,9%	14,4%	11,1%	17,7%	13,9%
Namur	10,1%	9,1%	7,2%	11,0%	8,5%
Oostende	6,1%	6,9%	4,3%	9,5%	5,1%

Figure 9. Confidence intervals % higher educated (2016)

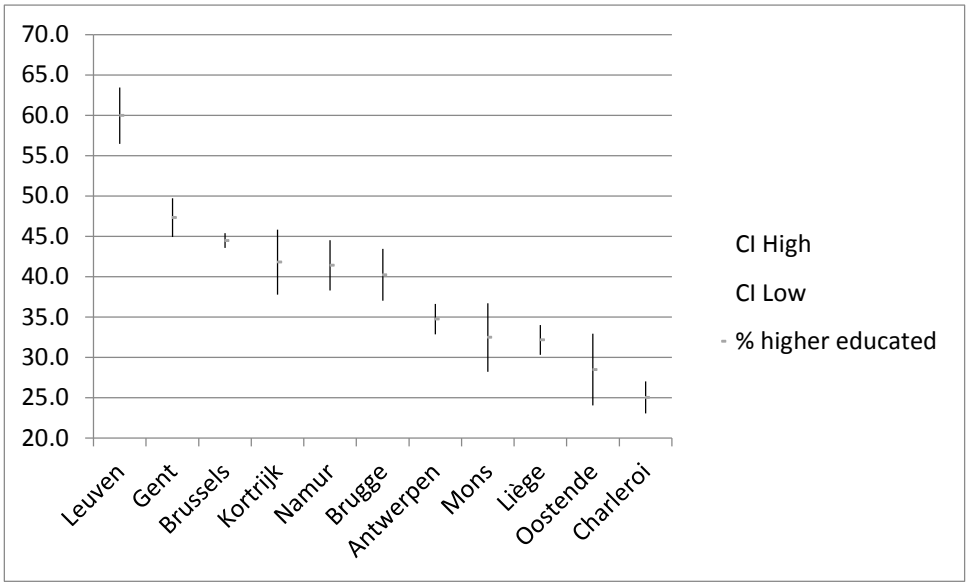


Table 18: % higher educated (25-64) by FUA with confidence interval for 2016

	2015	2016			2017
		Estimate 2016	CI lower limit	CI higher limit	
Antwerpen	38,5%	34,7%	32,8%	36,6%	36,9%
Brugge	36,6%	40,2%	37,0%	43,5%	41,1%
Brussels	44,2%	44,5%	43,5%	45,4%	47,1%
Charleroi	24,5%	25,0%	23,0%	27,0%	27,5%
Gent	43,1%	47,3%	44,9%	49,7%	48,6%
Kortrijk	38,1%	41,8%	37,8%	45,8%	45,5%
Leuven	52,3%	59,9%	56,5%	63,4%	66,8%
Liège	34,2%	32,2%	30,3%	34,0%	34,9%
Mons	29,7%	32,5%	28,2%	36,7%	30,0%
Namur	41,8%	41,4%	38,3%	44,5%	44,7%
Oostende	27,3%	28,5%	24,0%	32,9%	34,5%

Conclusions

A first analysis of LFS-indicators for the Belgian FUA's shows that the FUA's largely differ. There are FUA's with high employment, low unemployment rates and high percentages of high educated persons. Other FUA's perform less well for one or more of these indicators.

The FUA's Brussels and Antwerp, the biggest Belgian FUA's, score average. Brussels and Antwerp have average levels of employment, unemployment and higher education diploma's.

Despite the fact that the Belgian LFS is not designed to generate statistics on the level of FUA's, first analyses show that LFS can be used to a certain extent for large and medium-sized FUA's (with the exception of unemployment). But it is difficult to interpret evolutions from year to year because of the sometimes large confidence intervals. The presented results should be interpreted very carefully and further analysis is needed.

ANNEX 1.3

Report on the use of LFS for information on Functional Urban areas

Country: Netherlands

Date: 31-8-2018

Author(s): Martijn Souren

Introduction

The Labour Force Survey (LFS) is the main source in the EU for harmonised labour market statistics. This survey is based on a sample and therefore allow a limited degree of regional breakdown. Information on NUTS 2 in generally available. UN and OECD developed Functional Urban Areas (FUA's)[43] which play an important role in current policies dedicated to urban regions. One would expect that the populations of Metropolitan urban areas, with population between 500,000 and more, are of the same order as many NUTS 2 regions. The sample size would therefore be in principle large enough to provide LFS statistics. The main issue is that the sample of the LFS was not designed to generate statistics for these regions. This impacts on the quality. Small urban areas are excluded from the analysis, since the corresponding sample sizes are too small. For medium-sized urban areas we will investigate what is possible.

This report presents the results of the analysis carried out for country: Netherlands.

2. Functional urban areas in the Netherlands

For the Netherlands in total 35 FUA's can be identified. These cover 82.3 percent of the total population of 15-64 years old. More than half of these areas are so-called small urban areas which will not be discussed. The remaining 16 metropolitan or medium-sized areas cover 65.7 percent of the population in the Netherlands. The metropolitan FUA's can be found mostly in the Western part of the country: Amsterdam, Rotterdam, The Hague and Utrecht. One area is located in the Southeast: Eindhoven. The medium-sized areas can be found all across the country.

[43] The boundaries of cities and Functional Urban Areas are available on the Eurostat-GISCO website. The version referred to as "Urban Audit 2011-2014" is still valid for France, Germany and Belgium. Revisions have taken place in the Netherlands and Austria. The revised FUA boundaries are included in the FUA 2015-2018 dataset. http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/urban-audit

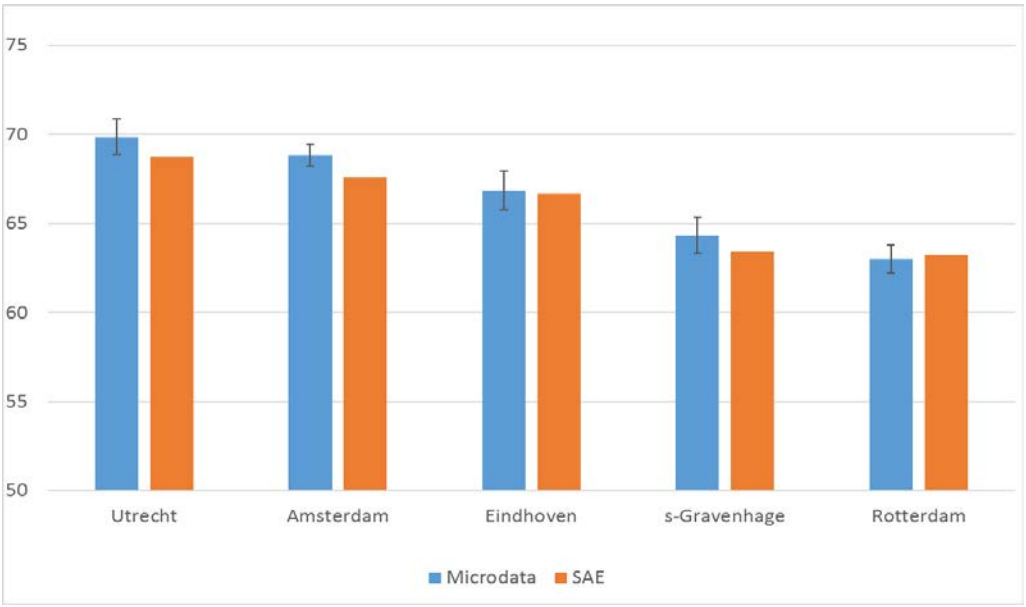
3. Results of LFS

The results presented in this report cover the population of 15 to 74-years old. The reason is that for this population also estimates can be produced with Small Area Estimators (SAE). SAE is a model-based estimator which uses the GREG-estimates in combination with estimates based on auxiliary information from registers. The larger the region the stronger the estimate relies on the GREG-estimator. Nevertheless, also SAE estimates for Metropolitan regions can be different from GREG-estimates. This is due to the fact that in SAE auxiliary information is used on a regional level, also for the GREG-estimates. Whereas for the regular GREG estimator this is only done on the country level. This way the SAE does not only improve the precision of the estimates but also the bias can be reduced. In the following paragraphs the results from the GREG estimates using the microdata are compared with SAE.

3.1. Employment rate

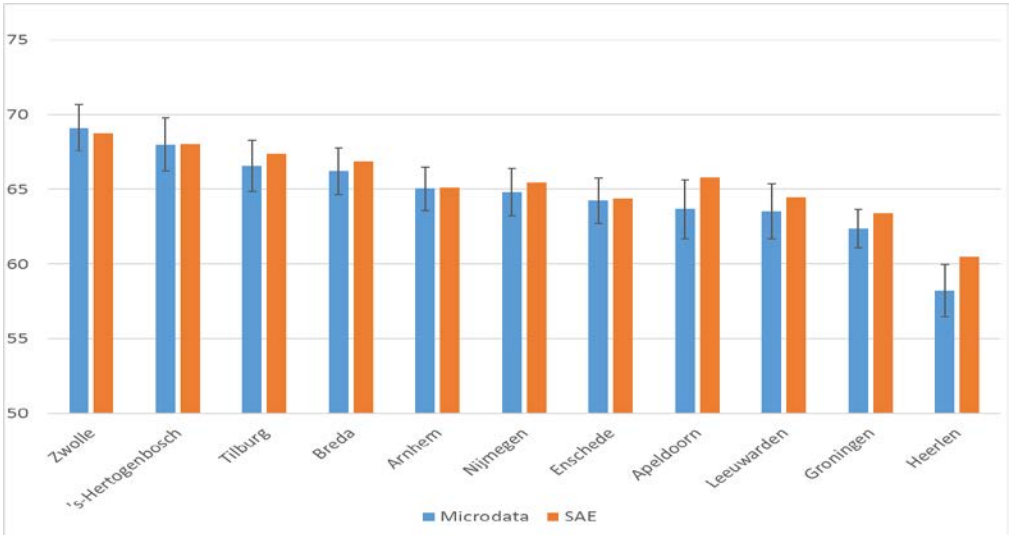
The highest employment rate of all FUA's in 2016 was in Utrecht where nearly 70 percent of the population is at work. The lowest employment rate of the metropolitan areas was in Rotterdam with 63 percent. The use of SAE does not change the order in which the regions can be placed according to the employment rate in 2016. Nevertheless, the difference between the two methods estimates can be larger than 1 percentage point. Because the regions are large, this can be a sign that the bias is larger in one metropolitan area compared to the other. This could be further investigated by calculating time series in order to see if the difference between the estimates is relatively stable over a longer period of time.

Figure 1. Employment rate by Dutch metropolitan FUA 15-74, 2016



For the medium-sized FUA's the difference between the areas in employment rate are comparable to the metropolitan area's. The employment rate ranges from 62 to 69, except for the region of Heerlen where the employment rate of 58 is relatively low. When the microdata estimates are compared with the SAE it can be seen that the difference between them is mostly within the reliability limits of the microdata estimate. This means that either the reliability limits are so large that a bias falls also within these limits or that the bias in these smaller areas is actually smaller than for the large cities. Only the regions of Apeldoorn and Heerlen show SAE for the employment rate which cross the reliability limits of the GREG estimate. For the region of Heerlen the difference is again the largest where the SAE is more than 2 percentage points higher than the GREG estimate. So part of the explanation why the employment rate is so low in Heerlen has probably to do with an estimation bias. However, also when the SAE are used the employment rate remains the lowest in Heerlen. Also for the other medium-sized regions the order does not change dramatically when the SAE are used instead of estimating directly from the microdata.

Figure 2. Employment rate by Dutch medium sized FUA 15-74, 2016



If the focus lies on a partical age group one can see that the reliability limits are larger. It becomes even harder to say that the employment rate in one area is satistically significant from another area. Also the SAE lie with the limits of the GREG estimates so it is hard to say if the difference between the two estimates are due to precisison or bias issues. Nevertheless the estimates still give some insights for particular regions. In Heerlen for example, the employment rate is also in this age group the lowest but the difference with other regions is smaller compared to the diffence for the total population. Also for the youngsters this is the case. In the oldest age group the difference is again relatively large. This shows that oldest age group creates the difference: Their particpation rate is low and in fact they are also overrepresented in Heerlen, which causes the total employment rate to be relatively low. Similar effects can be analysed for other region. In Zwolle for example, the employment rate is the highest and difference with other regions become even larger for 25 to 45-years old.

Figure 3. Employment rate by Dutch medium sized FUA 25-44, 2016

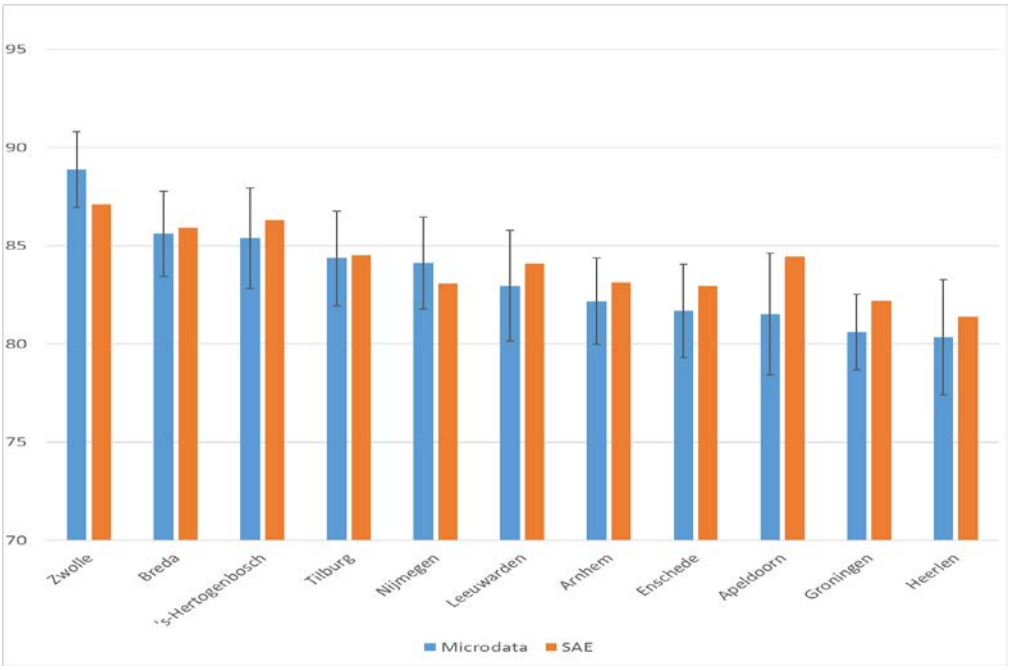


Figure 4. Employment rate by Dutch medium sized FUA 15-24, 2016

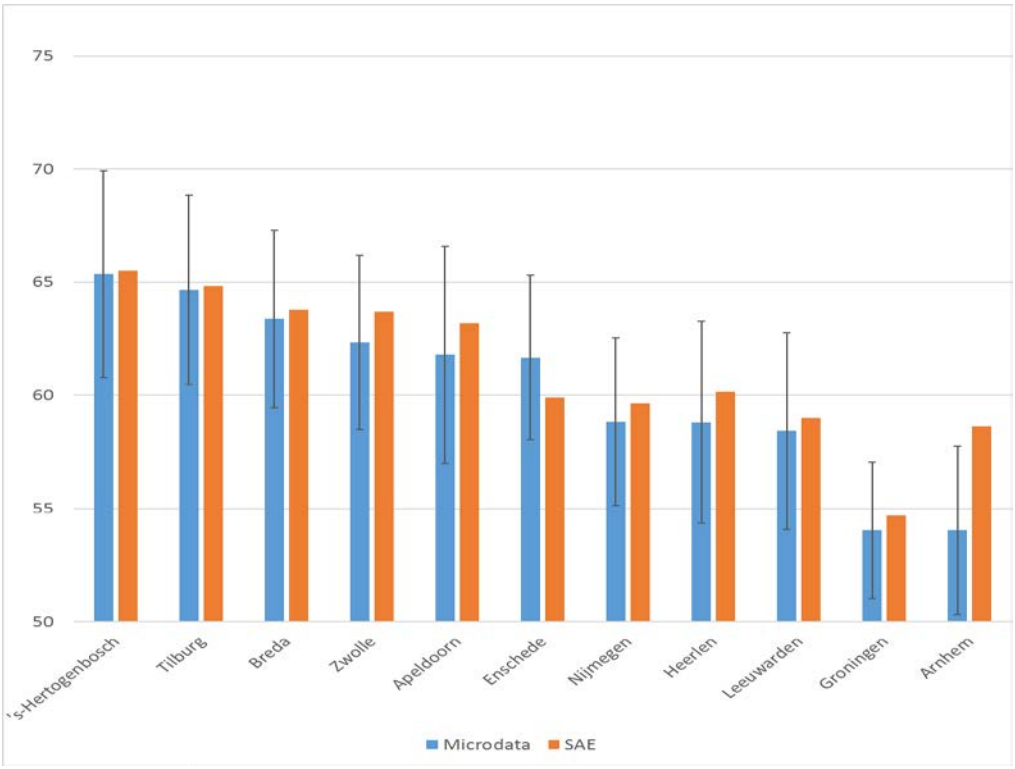
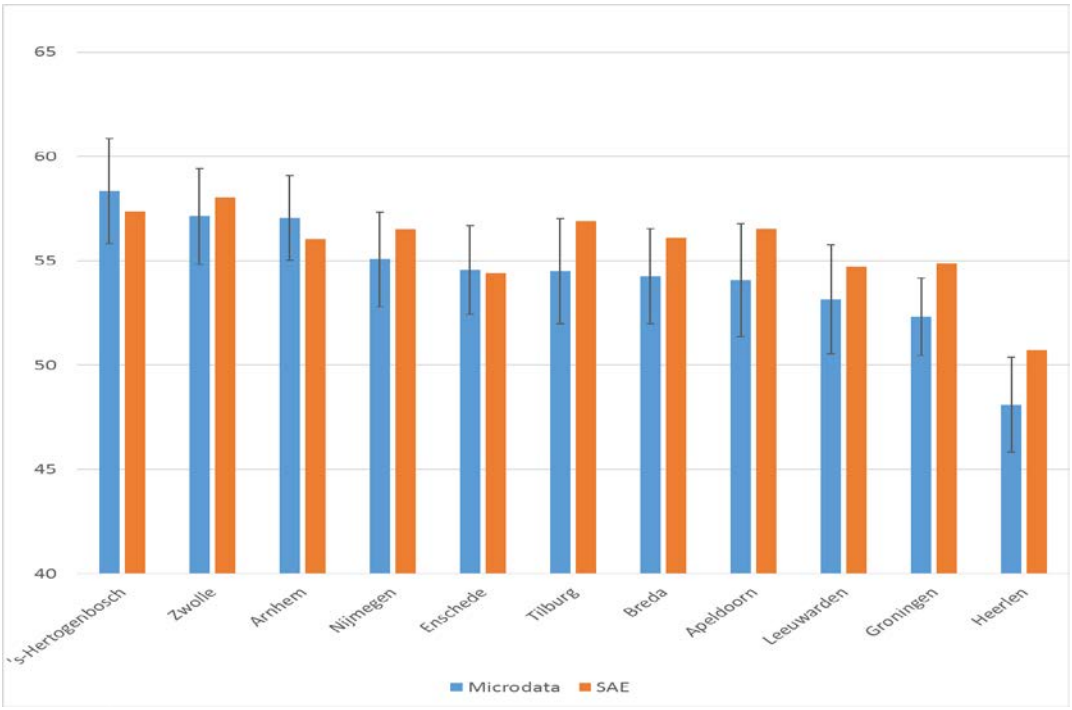


Figure 5. Employment rate by Dutch medium sized FUA 45-74, 2016



3.2 Unemployment rate

For the unemployment rate the same pictures can be displayed. The results are somewhat similar compared to the employment rate, but there are also differences. First of all, the SAE are less different from the GREG estimates. This has to do with the fact that the auxiliary information for employment on a regional level correlates better than for unemployment. It is more difficult to reduce the possible selection bias in unemployment statistics. Furthermore, it is interesting to see when an area is on top or bottom of the employment rate statistics but it is not (in reverse of course) for unemployment statistics. For example Heerlen again, is in the middle part as far as unemployment is concerned while it is in the very bottom as far as employment is concerned.

Figure 6. Unemployment rate by Metropolitan FUA 15-74, 2016

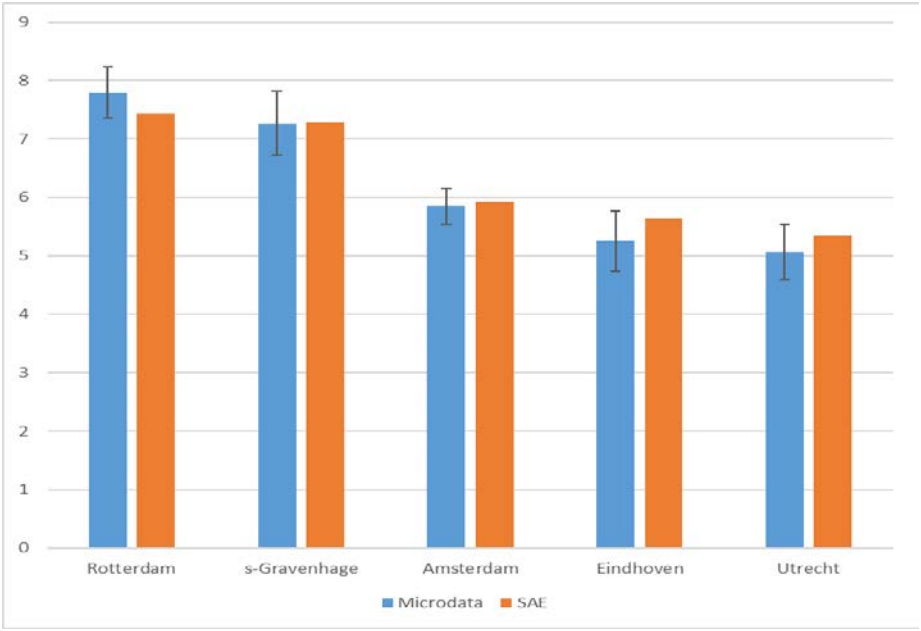


Figure 7. Unemployment rate by Medium sized FUA 15-74, 2016

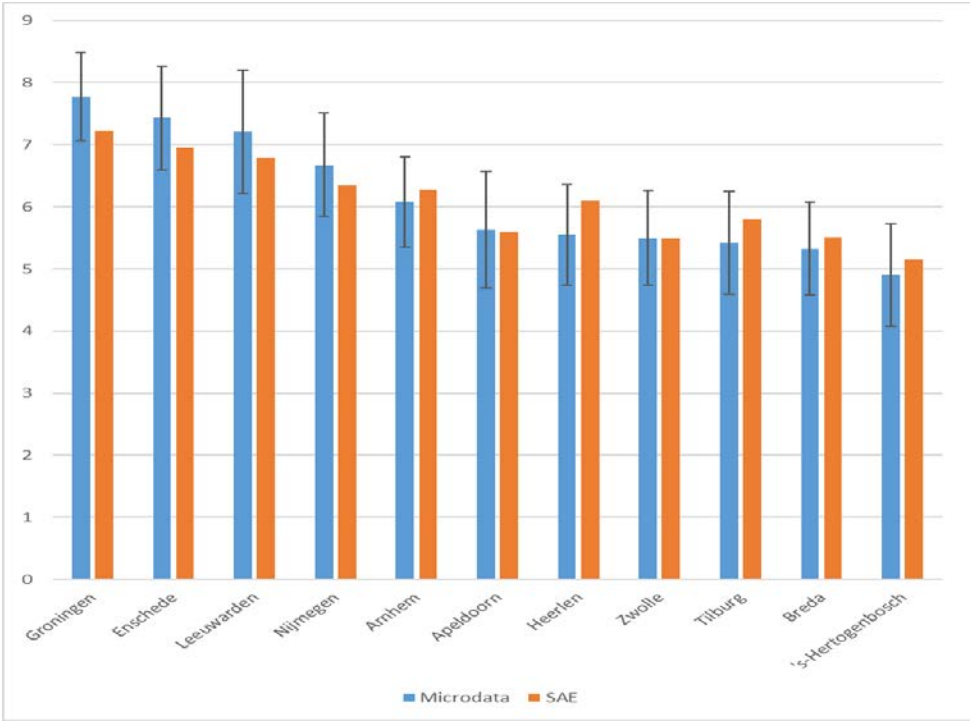
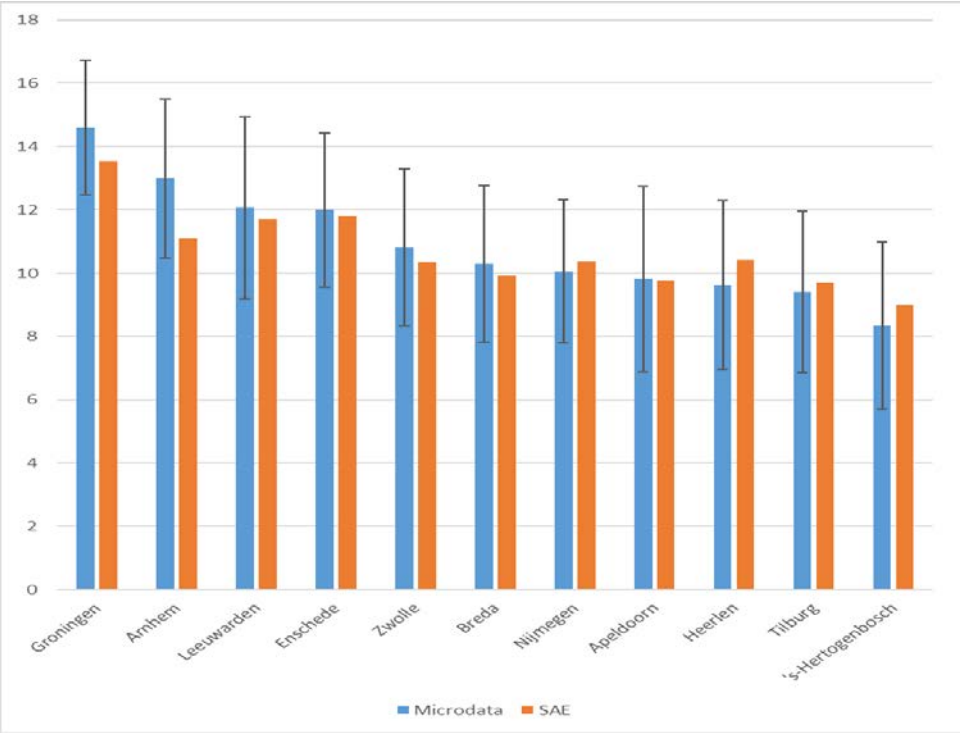


Figure 8. Unemployment rate by Medium sized FUA 15-24, 2016



3.3. Educational attainment

For education statistics the results are the most promising. Differences between areas are quite large and the precision is rather high. Again, SAE can reduce the bias and moreover for education it seems helpful on the level of medium sized areas as well.

Figure 9. Rate of highly educated people by Metropolitan FUA 15-74, 2016

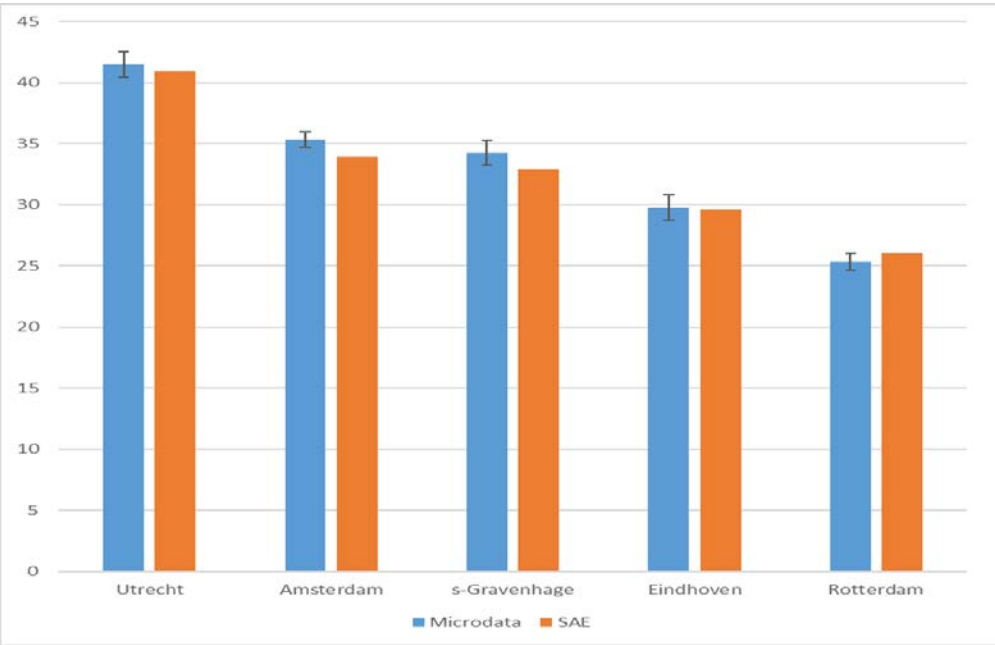
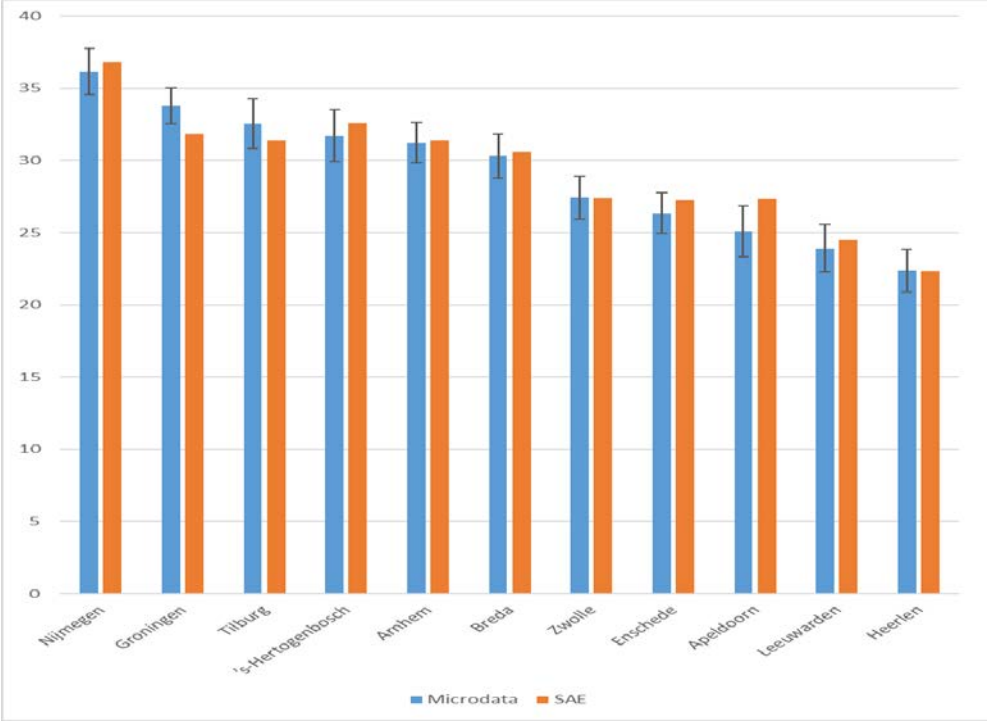


Figure 10. Rate of highly educated people by Medium sized FUA 15-74, 2016



4. Issues and concerns in producing indicators for FUA's

The biggest issue for regional statistics in The Netherlands is the precision of the estimates. This was the main reason why SAE were introduced. When analysing the results it showed that some areas also have more selection biases than others. The bias can be tackled by introducing auxiliary information on a regional level. At the moment this is incorporated in the SAE but it could also be tackled by improving the weighting on regional level. Producing the GREG estimates is a fairly straightforward. For SAE this much more complicated and time-consuming.

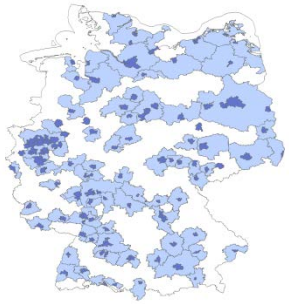
5. Conclusions

The results show that the labour market can be very different between areas within a rather small country as the Netherlands. It is therefore very useful to have precise and unbiased statistics. GREG estimates can be used but should be handled with care. SAE can help to improve precision and reduce the bias, especially for employment and education statistics.

ANNEX 1.4

Report on the use of LFS for information on Functional Urban Areas

Country: Germany



Date: 29-11-2018

Author(s): Sandra Hadam, Dr. Martina Rengers

List of figures

List of tables

1. Introduction

The Labour Force Survey (LFS) is the main source in the EU for harmonised labour market statistics. This survey is based on a sample and therefore allow a limited degree of regional breakdown. Information on NUTS 2 in generally available. UN and OECD developed Functional Urban Areas (FUAs)[44] which play an important role in current policies dedicated to urban regions. One would expect that the populations of Metropolitan urban areas, with population between 500,000 and more, are of the same order as many NUTS 2 regions. The sample size would therefore be in principle large enough to provide LFS statistics. The main issue is that the sample of the LFS was not designed to generate statistics for these regions.

[44] The boundaries of cities and Functional Urban Areas are available on the Eurostat-GISCO website. The version referred to as "Urban Audit 2011-2014" is still valid for France, Germany and Belgium. Revisions have taken place in the Netherlands and Austria. The revised FUA boundaries are included in the FUA 2015-2018 dataset. http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/urban-audit

This impacts on the quality. Since Germany has many small urban areas they are not excluded from the analysis. For medium-sized urban areas we will investigate what is possible.

This report presents the results of the analysis carried out for country: Germany.

2. Functional Urban Areas in Germany

The geographic information system of the commission GISCO is explained on the following Eurostat website:

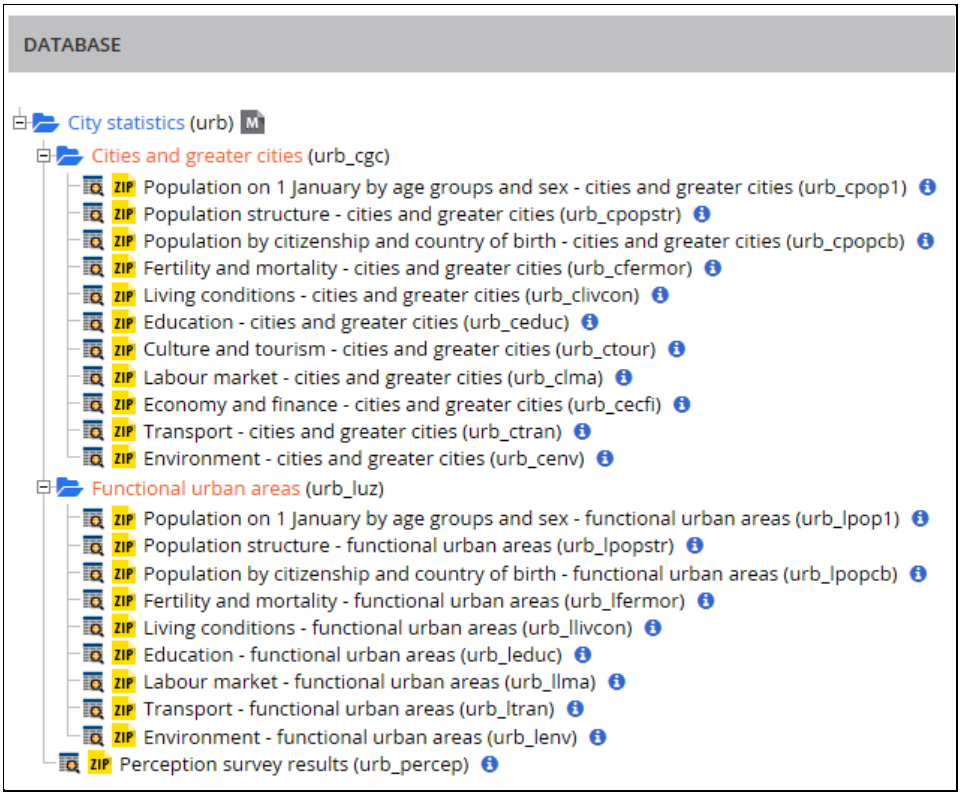
- ***https://ec.europa.eu/eurostat/web/gisco/overview***.

You can find there methodological information and statistics. As subdomains there are links to "Cities (Urban Audit)" and "Regions". The latter one provides statistics for NUTS 2 level and in some instances for NUTS 3 level. In total, Germany has 38 NUTS 2 regions (partly based on Regierungsbezirke or on Bundesländer) and 402 NUTS 3 regions (Landkreise, Stadtkreise and kreisfreie Städte).

The subdomain „Cities (Urban Audit)" contains an Excel-File with the list of Cities, greater Cities and Functional Urban Areas (FUAs) and a "Methodological manual on city statistics" (Eurostat 2017). It is explained, that until 2011 there was no harmonized definition of a city for European and OCED countries. From 2011 onwards a new EC-OECD definition "identified more than 900 cities with an urban centre of at least 50,000 inhabitants in the EU, Switzerland, Iceland and Norway. Each city is part of its own commuting zone or a polycentric commuting zone covering multiple cities. These commuting zones are significant, especially for larger cities. The cities and commuting zones put together are called Functional Urban Areas." (Eurostat 2017, p. 11). "The Functional Urban Area consists of the city and its commuting zone." (Eurostat 2017, p. 12).

More details can be found in Eurostat (2017), but the above mentioned definition can lead to confusions: Is "city" and "commuting zone" meant as "**either-or" principle** or as **"all-or-nothing" principle**? This uncertainty will be reinforced by looking on the structure of the database Eurostat has chosen for the city statistics. As **Figure 12** shows, there are two separated categories (1) cities and greater cities and (2) functional urban areas.

Figure 4: Structure of EUROSTAT database for cities (Urban Audit)



Source: https://ec.europa.eu/eurostat/web/cities/data/database (19-11-2018).

In the category "cities and greater cities" you can find **125** units for Germany in 2016 and **94** units in the other category "functional urban areas". [45] A closer look shows that these so called FUAs contain only commuting zones and not cities. That means, that more information is needed and as Figure 12 shows, the database also includes metadata for "city statistics". In the German metadata the contact organization **KOSIS-Gemeinschaft Urban Audit** with the URL *www.urbanaudit.de* is mentioned. In the meantime this URL is obsolete and is replaced by

- ***http://www.staedtestatistik.de/urban-audit.html*.**

This website contains some methodological documents. Urban Audit (2017) explains Eurostat's definition of Urban Audit cities and FUAs. FUAs comprise a combination of cities and suburban areas [in German: Umland] that share commuting patterns. Cities are defined by the degree of urbanisation, whereas suburban areas are defined by the commuting pattern: "The degree of urbanisation plays only an indirect role in determining a Functional Urban Area (FUA). The focus is on the economic interdependency of a city with its suburban areas in the form of commuters. A functional urban area consists of the city area and the corresponding commuting zone." (Urban Audit 2017 p. 11). The German KOSIS Association Urban Audit point out that statistical data especially for the difference between the FUAs and the Urban Audit cities are of high interest. For this reason published data have the objective to calculate the values for suburban areas. But the difference between the FUAs

[45] For the year 2017 the above mentioned Excel-File lists a number of 127 cities for Germany. Compared to the year 2016 this is an increase of two cities. The names of these two cities are "Düren, Stadt" (code DE548C1) and "Bocholt, Stadt" (code DE549C1).

and the Urban Audit cities cannot be calculated for Urban Audit cities whose area is per definition identical to that of their FUA. This is the case for 11 cities.

Figure 13 shows the relevant areas for Germany for the reference year 2016: There are in total 208 units which are relevant for determining FUAs. The FUAs are composed of **125** city cores called category 'C', and **83** commuting areas of these cores, known as category 'F'. Especially through the agglomeration of the cities in North-Rhine Westphalia, e.g. the Ruhr area, there exist more city cores than commuting zones. Furthermore, **11** commuting areas are indicated as city cores at the same time – as mentioned above (see Figure 13, red highlighted areas). Therefore it should be decided for a clear assignment of these areas in order to avoid double counting.

The FUAs consists of city cores and their commuting zones.[46] City cores are urban centres with at least 50,000 inhabitants. The commuting zones contain the surrounding travel-to-work areas of the city cores where at least 15 percent of their employed residents are working in this city. Urban areas are in general defined by the population size. Small urban areas contain a resident population between 50,000 and 200,000, medium-sized urban areas between 200,000 and 500,000, metropolitan areas between 500,000 and 1,5 million and large metropolitan areas with a resident population above 1,5 million inhabitants. The FUAs in Germany contain all four urban areas with the smallest city core named "Speyer" with 50,551 inhabitants. Germany has quite a lot small urban areas in the FUAs compared to other European countries. Therefore, we will not exclude small urban areas from the analysis and investigate the estimation of LFS indicators for all types of urban areas.

[46] The commuting areas consist of several NUTS 3 level areas. Most of the city cores consist of one municipality. By aggregating the NUTS 3 areas on FUAs as a result some FUAs have almost the same size as NUTS 2 level regions.

Figure 5: 94 commuting areas and 125 cities and greater cities in Germany

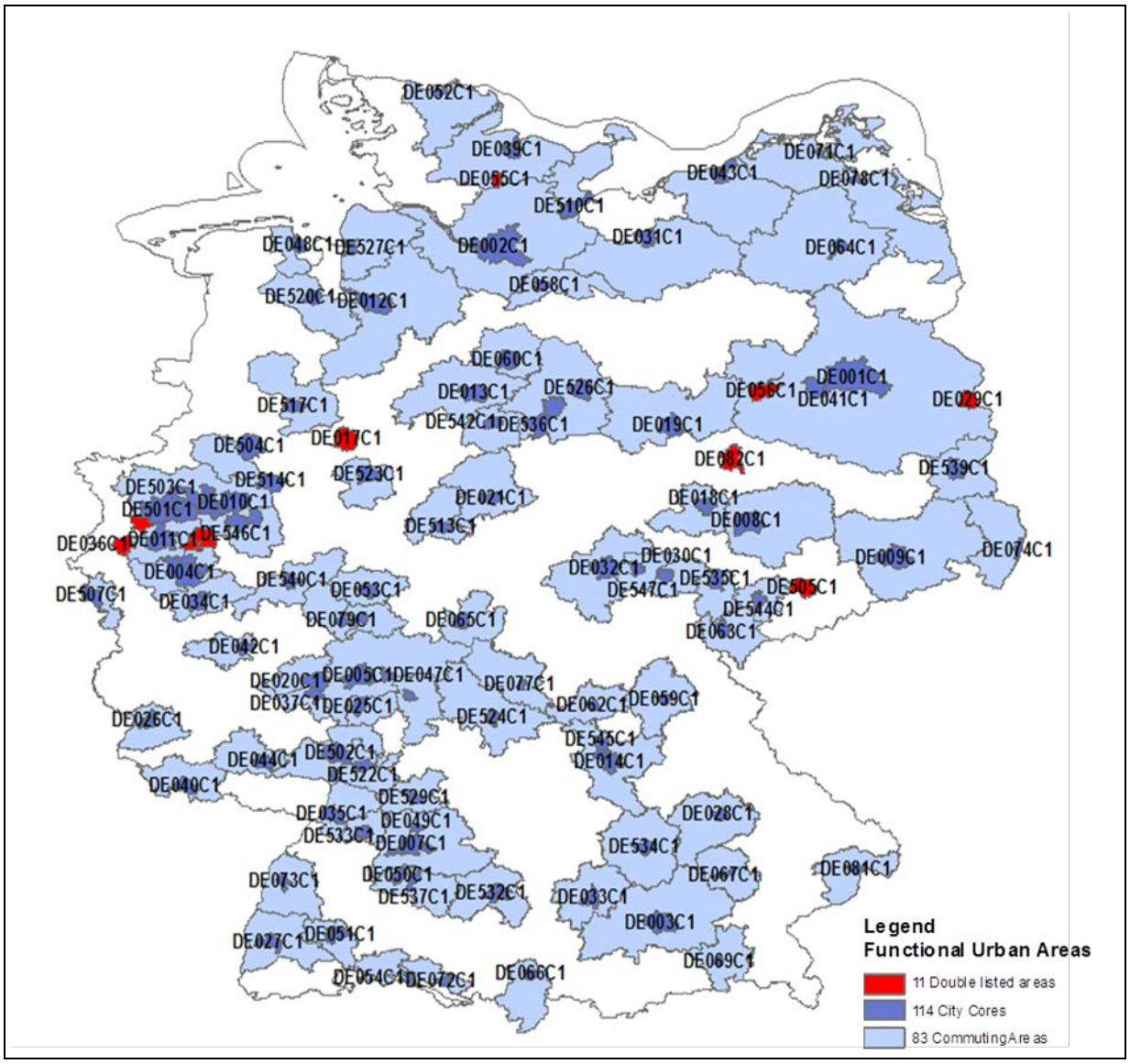


Table 3 shows the names of all the 208 relevant areas for Germany. The list is ordered by the FUA coding and different colours separate the three types of areas/units: (1) dark blue for 114 city cores, (2) light blue for 83 commuting areas and (3) light red for the 11 double listed areas, meaning Urban Audit cities whose area is per definition identical to that of their FUA. Furthermore, the LFS sample sizes of each area are reported in **Table 3**. This is the basis for the following analysis in chapter 3, where we present some results of a direct estimation of labour market and educational indicators for the 208 city cores, commuting zones and FUAs.

Table 2: Functional Urban Areas in Germany: code, names and LFS sample size, 2016

No.				FUA	Name	LFS sample size
1	1			DE001C1	Berlin	30,983
2		1		DE001L1	Berlin	13,239
3	2			DE002C1	Hamburg	15,040
4		2		DE002L1	Hamburg	13,022
5	3			DE003C1	München	12,731
6		3		DE003L1	München	12,395
7	4			DE004C1	Köln	9,163
8		4		DE004L1	Köln	5,487
9	5			DE005C1	Frankfurt am Main	5,715
10		5		DE005L1	Frankfurt am Main	14,422
11	6			DE006C1	Essen	5,040
12	7			DE007C1	Stuttgart	4,887
13		6		DE007L1	Stuttgart	16,896
14	8			DE008C1	Leipzig	4,658
15		7		DE008L2	Leipzig	4,165
16	9			DE009C1	Dresden	4,890
17		8		DE009L2	Dresden	7,438
18	10			DE010C1	Dortmund	5,163
19	11			DE011C1	Düsseldorf	5,185
20		9		DE011L1	Düsseldorf	7,082
21	12			DE012C1	Bremen	5,239
22		10		DE012L1	Bremen	6,184
23	13			DE013C1	Hannover	4,382
24		11		DE013L1	Hannover	6,780
25	14			DE014C1	Nürnberg	4,522
26		12		DE014L1	Nürnberg	5,512
27	15			DE015C1	Bochum	3,317
28			1	DE017C1	Bielefeld	2,897
29	16			DE018C1	Halle an der Saale	2,067
30		13		DE018L1	Halle an der Saale	1,783
31	17			DE019C1	Magdeburg	2,115
32		14		DE019L2	Magdeburg	2,466
33	18			DE020C1	Wiesbaden	2,358
34		15		DE020L1	Wiesbaden	1,489
35	19			DE021C1	Göttingen	998
36		16		DE021L1	Göttingen	2,401
37	20			DE022C1	Mülheim a.d.Ruhr	1,486
38	21			DE023C1	Moers	890
39	22			DE025C1	Darmstadt	1,261
40		17		DE025L1	Darmstadt	2,478
41	23			DE026C1	Trier	939
42		18		DE026L1	Trier	1,364
43	24			DE027C1	Freiburg im Breisgau	1,806
44		19		DE027L1	Freiburg im Breisgau	3,782
45	25			DE028C1	Regensburg	1,418
46		20		DE028L1	Regensburg	3,010
47			2	DE029C1	Frankfurt (Oder)	537
48	26			DE030C1	Weimar	552
49		21		DE030L1	Weimar	753
50	27			DE031C1	Schwerin	903
51		22		DE031L1	Schwerin	1,842
52	28			DE032C1	Erfurt	1,963
53		23		DE032L1	Erfurt	2,857
54	29			DE033C1	Augsburg	2,660
55		24		DE033L1	Augsburg	3,553
56	30			DE034C1	Bonn	2,815
57		25		DE034L1	Bonn	4,462
58	31			DE035C1	Karlsruhe	2,366
59		26		DE035L1	Karlsruhe	3,675
60			3	DE036C1	Mönchengladbach	2,284
61	32			DE037C1	Mainz	1,866
62		27		DE037L1	Mainz	2,067
63		28		DE038L1	Ruhrgebiet	13,444
64	33			DE039C1	Kiel	2,026
65		29		DE039L1	Kiel	3,601
66	34			DE040C1	Saarbrücken	1,540
67		30		DE040L1	Saarbrücken	5,774
68	35			DE041C1	Potsdam	1,494
69	36			DE042C1	Koblenz	1,004
70		31		DE042L1	Koblenz	2,030
71	37			DE043C1	Rostock	1,778
72		32		DE043L2	Rostock	1,830
73	38			DE044C1	Kaiserslautern	841

No.				FUA	Name	LFS sample size
74		33		DE044L1	Kaiserslautern	1,576
75	39			DE045C1	Iserlohn	801
76		34		DE045L1	Iserlohn	2,697
77	40			DE046C1	Esslingen am Neckar	818
78	41			DE047C1	Hanau	633
79	42			DE048C1	Wilhelmshaven	690
80		35		DE048L1	Wilhelmshaven	911
81	43			DE049C1	Ludwigsburg	791
82	44			DE050C1	Tübingen	689
83		36		DE050L1	Tübingen	1,231
84	45			DE051C1	Villingen-Schwenningen	750
85		37		DE051L1	Villingen-Schwenningen	1,069
86	46			DE052C1	Flensburg	794
87		38		DE052L1	Flensburg	1,760
88	47			DE053C1	Marburg	628
89		39		DE053L1	Marburg	1,346
90	48			DE054C1	Konstanz	769
91		40		DE054L1	Konstanz	1,562
92			4	DE055C1	Neumünster	616
93			5	DE056C1	Brandenburg an der Havel	576
94	49			DE057C1	Gießen	674
95		41		DE057L1	Gießen	1,518
96	50			DE058C1	Lüneburg	734
97		42		DE058L1	Lüneburg	892
98	51			DE059C1	Bayreuth	661
99		43		DE059L1	Bayreuth	1,661
100	52			DE060C1	Celle	670
101		44		DE060L1	Celle	1,049
102	53			DE061C1	Aschaffenburg	578
103		45		DE061L1	Aschaffenburg	2,725
104	54			DE062C1	Bamberg	681
105		46		DE062L1	Bamberg	1,314
106	55			DE063C1	Plauen	600
107		47		DE063L1	Plauen	1,455
108	56			DE064C1	Neubrandenburg	547
109		48		DE064L1	Neubrandenburg	2,054
110	57			DE065C1	Fulda	527
111		49		DE065L1	Fulda	1,265
112	58			DE066C1	Kempten (Allgäu)	643
113		50		DE066L1	Kempten (Allgäu)	1,466
114	59			DE067C1	Landshut	569
115		51		DE067L1	Landshut	1,456
116	60			DE068C1	Sindelfingen	482
117	61			DE069C1	Rosenheim	586
118		52		DE069L1	Rosenheim	2,294
119	62			DE070C1	Frankenthal (Pfalz)	454
120	63			DE071C1	Stralsund	457
121		53		DE071L1	Stralsund	1,435
122	64			DE072C1	Friedrichshafen	472
123		54		DE072L1	Friedrichshafen	1,112
124	65			DE073C1	Offenburg	499
125		55		DE073L1	Offenburg	3,013
126	66			DE074C1	Görlitz	547
127		56		DE074L1	Görlitz	1,833
128	67			DE075C1	Sankt Augustin	523
129	68			DE076C1	Neu-Ulm	495
130	69			DE077C1	Schweinfurt	493
131		57		DE077L1	Schweinfurt	2,061
132	70			DE078C1	Greifswald	439
133		58		DE078L1	Greifswald	1,474
134	71			DE079C1	Wetzlar	517
135		59		DE079L1	Wetzlar	1,664
136	72			DE080C1	Speyer	415
137	73			DE081C1	Passau	493
138		60		DE081L1	Passau	1,830
139			6	DE082C1	Dessau-Roßlau	846
140		61		DE083L1	Braunschweig-Salzgitter-Wolfsburg	4,586
141		62		DE084L1	Mannheim-Ludwigshafen	5,447
142	74			DE501C1	Duisburg	4,195
143	75			DE502C1	Mannheim	2,476
144	76			DE503C1	Gelsenkirchen	2,184
145	77			DE504C1	Münster	2,928
146		63		DE504L1	Münster	1,928

No.				FUA	Name	LFS sample size
147			7	DE505C1	Chemnitz	2,339
148	78			DE506C1	Braunschweig	2,142
149	79			DE507C1	Aachen	2,014
150		64		DE507L1	Aachen	2,664
151			8	DE508C1	Krefeld	2,141
152	80			DE509C1	Oberhausen	1,787
153	81			DE510C1	Lübeck	1,737
154		65		DE510L1	Lübeck	1,694
155	82			DE511C1	Hagen	1,647
156	83			DE513C1	Kassel	1,603
157		66		DE513L1	Kassel	2,035
158	84			DE514C1	Hamm	1,686
159	85			DE515C1	Herne	1,283
160			9	DE516C1	Solingen	1,434
161	86			DE517C1	Osnabrück	1,471
162		67		DE517L1	Osnabrück	3,261
163	87			DE518C1	Ludwigshafen am Rhein	1,614
164	88			DE519C1	Leverkusen	1,464
165	89			DE520C1	Oldenburg (Oldenburg)	1,383
166		68		DE520L1	Oldenburg (Oldenburg)	2,122
167	90			DE521C1	Neuss	1,283
168	91			DE522C1	Heidelberg	1,123
169		69		DE522L1	Heidelberg	4,831
170	92			DE523C1	Paderborn	1,255
171		70		DE523L1	Paderborn	1,431
172	93			DE524C1	Würzburg	1,025
173		71		DE524L2	Würzburg	3,334
174	94			DE525C1	Recklinghausen	1,013
175	95			DE526C1	Wolfsburg	1,039
176	96			DE527C1	Bremerhaven	1,141
177		72		DE527L1	Bremerhaven	1,895
178	97			DE528C1	Bottrop	975
179	98			DE529C1	Heilbronn	1,049
180		73		DE529L1	Heilbronn	2,878
181			10	DE530C1	Remscheid	958
182	99			DE531C1	Offenbach am Main	875
183	100			DE532C1	Ulm	1,041
184		74		DE532L1	Ulm	2,598
185	101			DE533C1	Pforzheim	1,026
186		75		DE533L1	Pforzheim	1,681
187	102			DE534C1	Ingolstadt	1,226
188		76		DE534L1	Ingolstadt	3,371
189	103			DE535C1	Gera	853
190		77		DE535L1	Gera	915
191	104			DE536C1	Salzgitter	939
192	105			DE537C1	Reutlingen	865
193		78		DE537L1	Reutlingen	1,396
194	106			DE538C1	Fürth	1,125
195	107			DE539C1	Cottbus	918
196		79		DE539L1	Cottbus	990
197	108			DE540C1	Siegen	934
198		80		DE540L2	Siegen	2,778
199	109			DE541C1	Bergisch Gladbach	984
200	110			DE542C1	Hildesheim	934
201		81		DE542L1	Hildesheim	1,645
202	111			DE543C1	Witten	937
203	112			DE544C1	Zwickau	932
204		82		DE544L1	Zwickau	2,167
205	113			DE545C1	Erlangen	1,001
206			11	DE546C1	Wuppertal	3,055
207	114			DE547C1	Jena	1,006
208		83		DE547L2	Jena	727

Legend

- 11 double listed areas with two codes that only differ in the two last digits: DExxxC1 vs. DExxxL0.
- 114 city cores
- 83 commuting areas

3. Results of LFS

We regarded the data set of the LFS from 2016 with an overall sample size of 725,829 observations. By considering the FUAs, that does not cover the whole German territory, the sample size decreases to 533,356 observations. Furthermore, this report covers the population of 15 to 64 years old.
The LFS indicators were normally published at NUTS 2 level (see chapter 2). Since each observation of the LFS also contains information of NUTS 3 level regions as well as of communities, we can use this information to estimate indicators on smaller areas.
Based on it 73.5 % (=533,356/725,829) of the LFS sample covers the FUAs, whereas 26.6% are located outside these areas.

At the beginning a direct estimation is applied. This includes a direct estimation of weighted means of the variable of interest, which is solely based on the sampling and survey design of the LFS. Therefore, a Horvitz-Thompson estimator is used. Considering the one-stage clustered sample or area sample, one gets an unbiased estimator for the LFS design for any breakdown.
For simplicity and clarity and to ensure a comparison between all indicators we focus on the areas with the highest and lowest sample sizes. The following results and figures represent five areas of the FUAs with the highest sample size and five areas with the lowest sample size. The areas in the figures are ordered by their sample size in descending order from left to right.

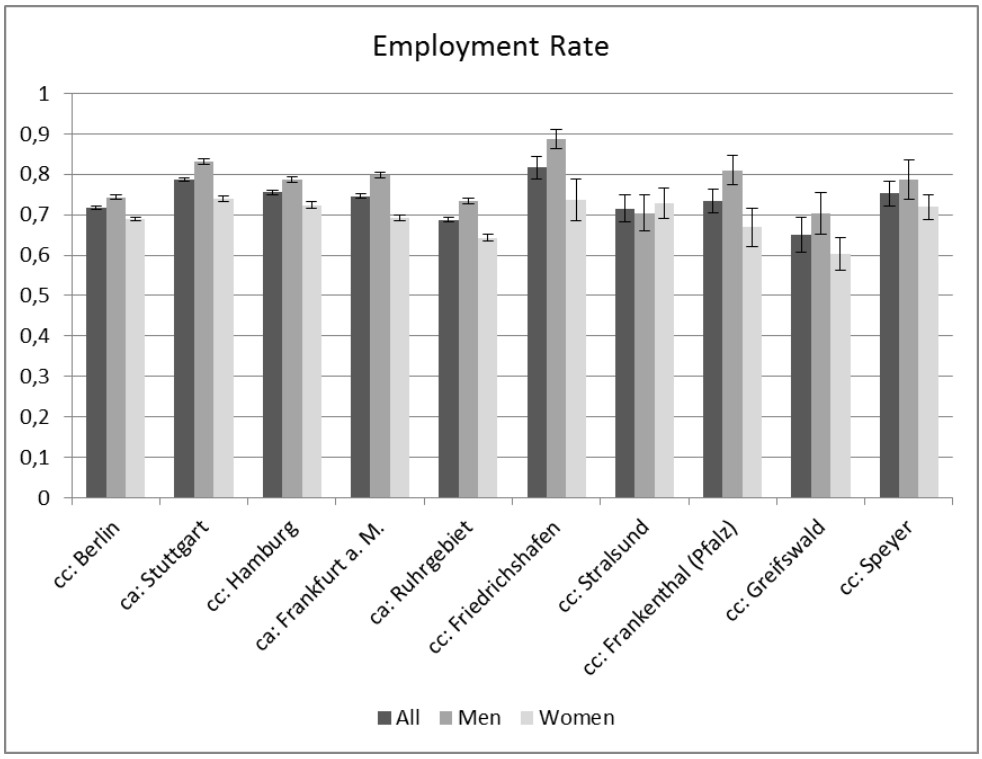
3.1. Employment rate

On average, 73.6% of the population in the FUAs is employed. As Figure 14 shows, the employment rate among women is in most cases lower than the employment rate among men. On average, the employment rate of men is 77.1% and 70.0% of women. We obtain higher standard errors in areas with fewer observations (right side) compared to areas with more observations (left side) as well.
The standards errors are within an interval of [0.004, 0.042] for the employment rate of the population and within an interval of [0.005, 0.064] by gender.[47] For the evaluation the relative standard errors are determined. They are also known as Coefficient of Variation (CV). We obtain CVs of minimum 0.59% to a maximum of 6.45% without breakdowns and CVs of minimum 0.69% maximum 8.6% by gender. They increase as soon as the sample size and the number of observations decrease. Furthermore, we obtain higher standard errors if we are taking breakdowns into account.
Compared to NUTS 2 level regions one can see a big increase of the highest CV for employment rate: in 2016 the maximum CV for all 38 NUTS 2 regions reached a value of 1.09% (region “Trier”). The minimum CV for NUTS 2 regions is 0.33% (region “Oberbayern”) and is as well better than that for the FUAs.[48]

[47] The standard errors are illustrated as black vertical lines in the presented bar charts.

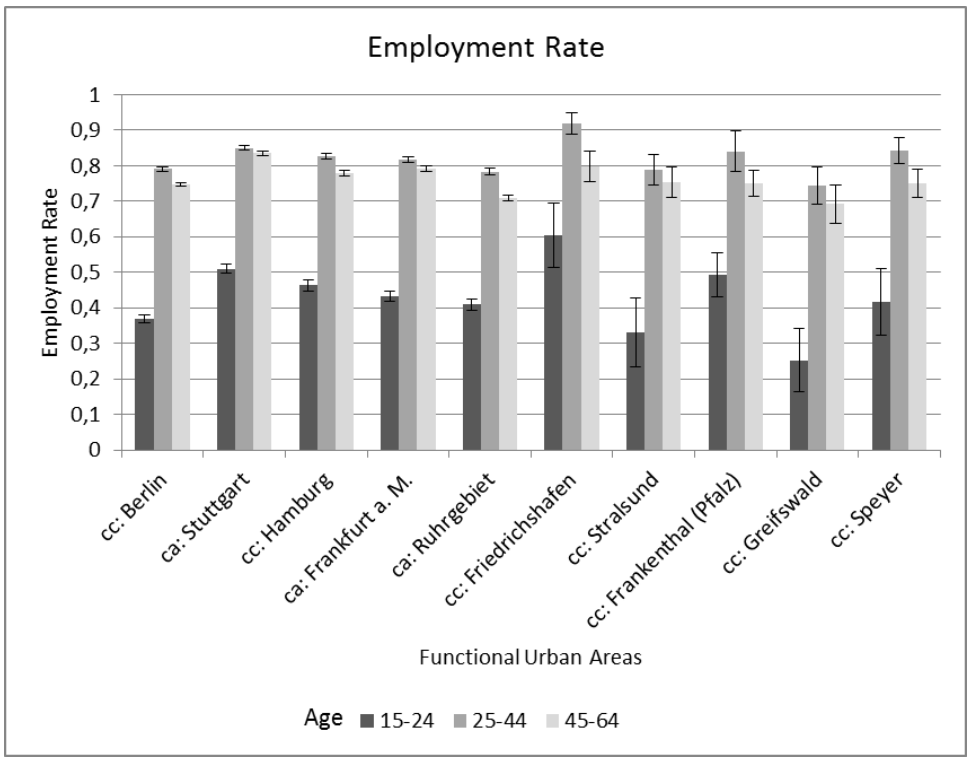
[48] See Eurostat (2018).

Figure 6: Employment rate by sex in selected FUAs: direct LFS estimation



Others: ca: commuting area cc: city core

Figure 7: Employment rate by age groups in selected FUAs: direct LFS estimation



Others: ca: commuting area cc: city core

Regarding the employment rate by age groups in Figure 15 there are similarities between the rate of the 25-44 years old and those of the 45-64 years old. The average employment rate of the 15-24 years old is 37.2 percentage points lower than those of the 25-44 years old. Furthermore, the deviation is higher for the youngest age group and the CVs are between

2.9% and 19.02%. Compared to this the CVs for the other two age groups are between 0.7% and 7.8%.

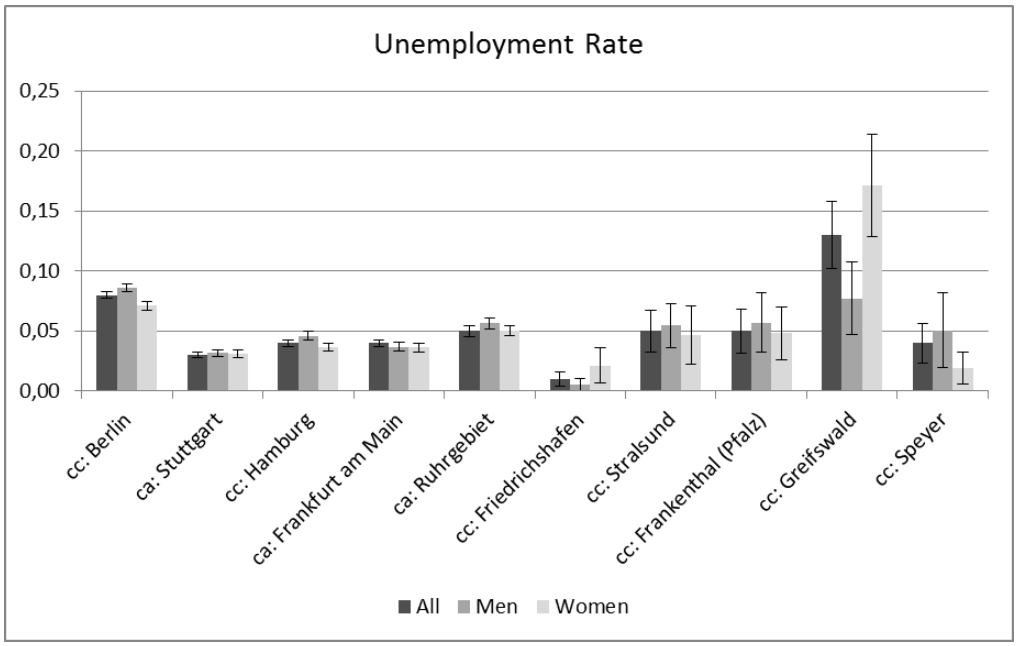
3.2. Unemployment rate

On average, 4.6% of the population in the FUAs is unemployed. The unemployment rate of women in **Figure 16** is in most cases lower than the unemployment rate of men. We obtain an average unemployment rate of 5.1% for men and 3.9% for women. As well we obtain higher standard errors in areas in areas with fewer observations compared to areas with more observations.

Expressed as relative standard errors we obtain CVs within a range of minimum 9.8% and maximum 22.57% for the unemployment rate of the population, within a range of minimum 12.3% to a maximum 25.01% for women and even until 56.25% for men. Compared to NUTS 2 regions this is again a big increase. We have there a maximum CV of 13.92% (region "Trier") and a minimum of 3.32% (region "Berlin").[49]

Compared to the employment rates for FUAs in chapter 3.1 we obtain a strong increase in the deviation and uncertainty and also in varying intensity depending on gender. The percentage points of the CVs increase as soon as the sample size and the number of observations decrease. Furthermore, we obtain higher standard errors if we are taking more breakdowns into account.

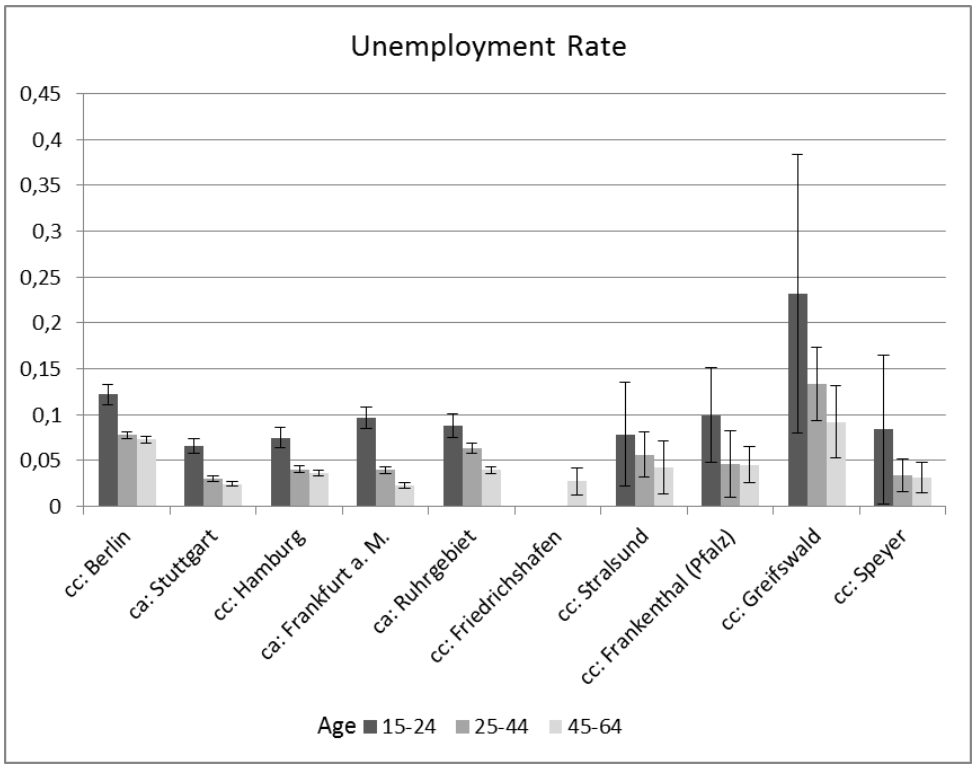
Figure 8: Unemployment rate by sex in selected FUAs: direct LFS estimation



Others: ca: commuting area cc: city core

[49] See Eurostat (2018).

Figure 9: Unemployment rate by age groups in selected FUAs: direct LFS estimation



Others: ca: commuting area cc: city core

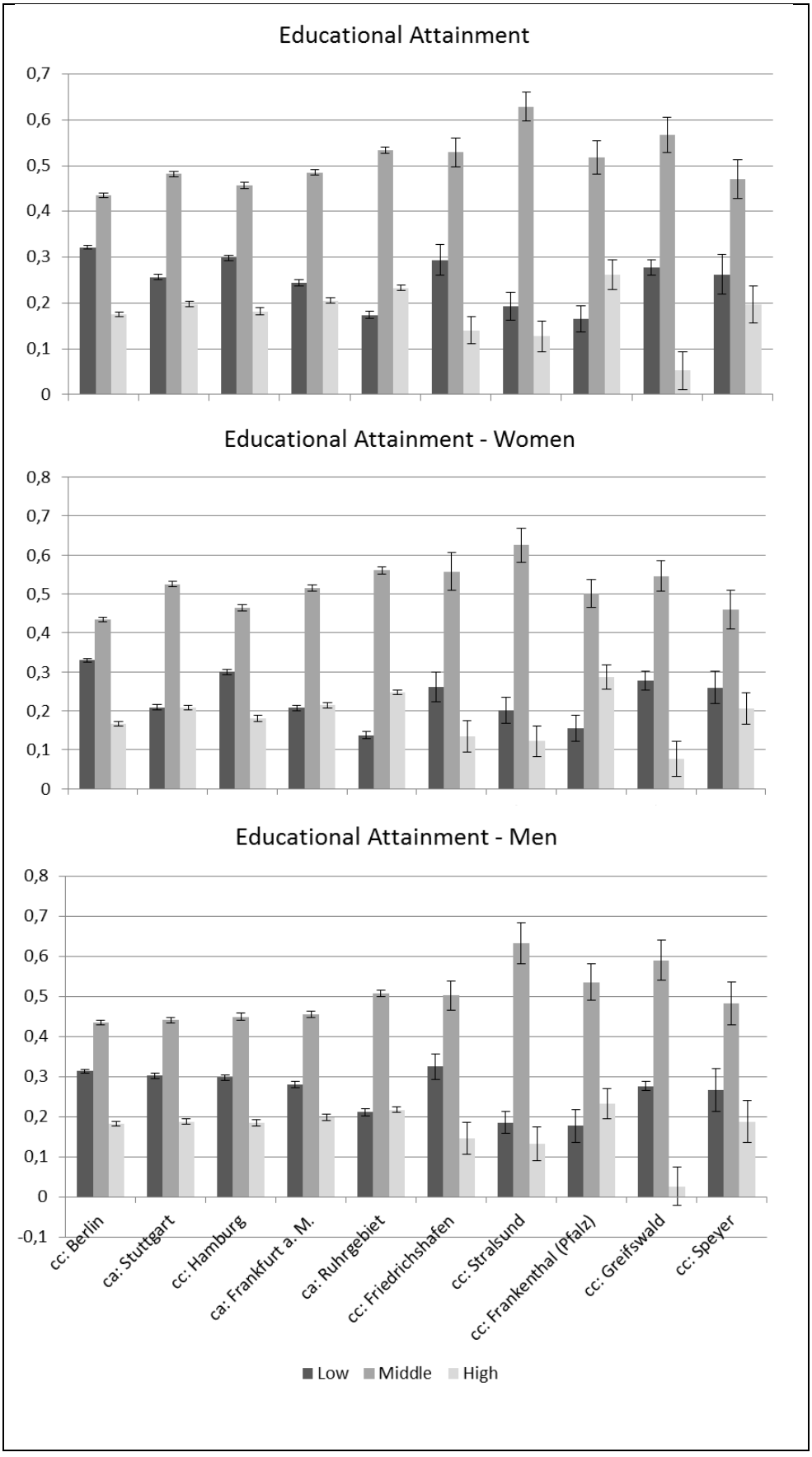
Figure 17 represents the unemployment rate by age groups. Because of the decreasing number of observations due to the breakdown, some out-of-sample domains arise. Out-of-sample domains are areas without observations. We receive six out-of-sample domains for the youngest age group and CVs between a minimum of 12.78% and a maximum of even 77%, one out-of-sample domain for the 25-44 years old and associated CVs between 13.9% and 29.57% and for the 45-64 years we obtain three out-of-sample domains with CVs between a minimum of 15.07% and a maximum of 42.8%. This indicates greater uncertainty in estimating unemployment rate for the youngest and oldest age group. “Greifswald” is a small urban area with very few observations. That’s why we receive very high CVs for this area compared to large metropolitan areas like “Berlin”.

3.3. Educational attainment of population 25-64 years of age

To obtain the educational attainment of the population the ISCED-classification was used, where related qualifications were summarized to the categories low, middle and high. **Figure 18** reports the educational attainment for the whole population and by gender.

In all cases a greater proportion of population with middle educational attainment is obvious compared to the other levels. The average rate of middle educational attainment is 52.2% of the population in FUAs, 53.6 % by women and 50.9 % by men.

Figure 10: Educational attainment



Others: ca: commuting area cc: city core

The CVs are similar for each educational attainment compared by gender. We obtain CVs of minimum 4% up to a maximum of 10%, where we obtain lower CVs for the middle educational attainment and the highest in case of low educational attainment. The CVs of the educational attainment by men is between a minimum of around 2% and a maximum of around 46%, whereas the CV is in average 9.7% for higher educational attainment, 5.2% for middle and 12.4% for lower educational attainment.
Compared to the variation in case of males the results for females are in average somewhat similar. The CVs of educational attainment for women are around a minimum of 2% up to a maximum of 32%, whereas the CV is in average 10.5% for higher educational attainment, 4.8% for middle and 12% for lower educational attainment.

4. Issues and concerns in producing indicators for FUAs

On the **Eurostat database for cities** (see **Figure 12**) there can also be found some metadata for German city statistics:

- ***https://ec.europa.eu/eurostat/cache/metadata/EN/urb_esms_de.htm.***

There is an overview of the availability of data and a description of the data in separate annexes under:

- https://ec.europa.eu/eurostat/cache/metadata/Annexes/urb_esms_de_an2.xlsx
- https://ec.europa.eu/eurostat/cache/metadata/Annexes/urb_esms_de_an3.xlsx
- https://ec.europa.eu/eurostat/cache/metadata/Annexes/urb_esms_de_an4.xlsx

A lot of indicators are published. But hereinafter we focus only on figures regarding the issues (1) employment, (2) unemployment and (3) educational attainment.

For persons employed, as data collection and data source respectively register data of employed persons with social insurance and of unemployed from the Federal Employment Agency FEA and 1% Microcensus (note: the LFS survey is included in the Microcensus) data and data on statistics on civil servants from the German Federal Statistical Office Destatis are mentioned. The type of statistics is described as “modelling based on administrative register and sample survey”. For persons unemployed it’s exactly the same.

The number of persons with ISCED level low, middle or high as the highest level of education, which is needed for the calculation of the educational attainment, is modelling based on sample survey, or more exactly on 1% Microcensus.

Due to these circumstances there is the possibility to do an additional quality check for two of the directly estimated LFS indicators on FUAs of chapter 3. That means that a direct LFS estimation for employment rate and for unemployment rate can be compared with the published results based on a complex system which combines the above mentioned administrative register data with LFS sample survey results.[50]

In the following the results based on a direct estimation and the already published Urban Audit indicators will be compared and discussed (see chapter 2, especially **Figure 12**). **Figure 19** and **Figure 20** compares the unemployment rate by gender of selected FUAs with the published estimators of the Urban Audit. At first, the unemployment rate of males and

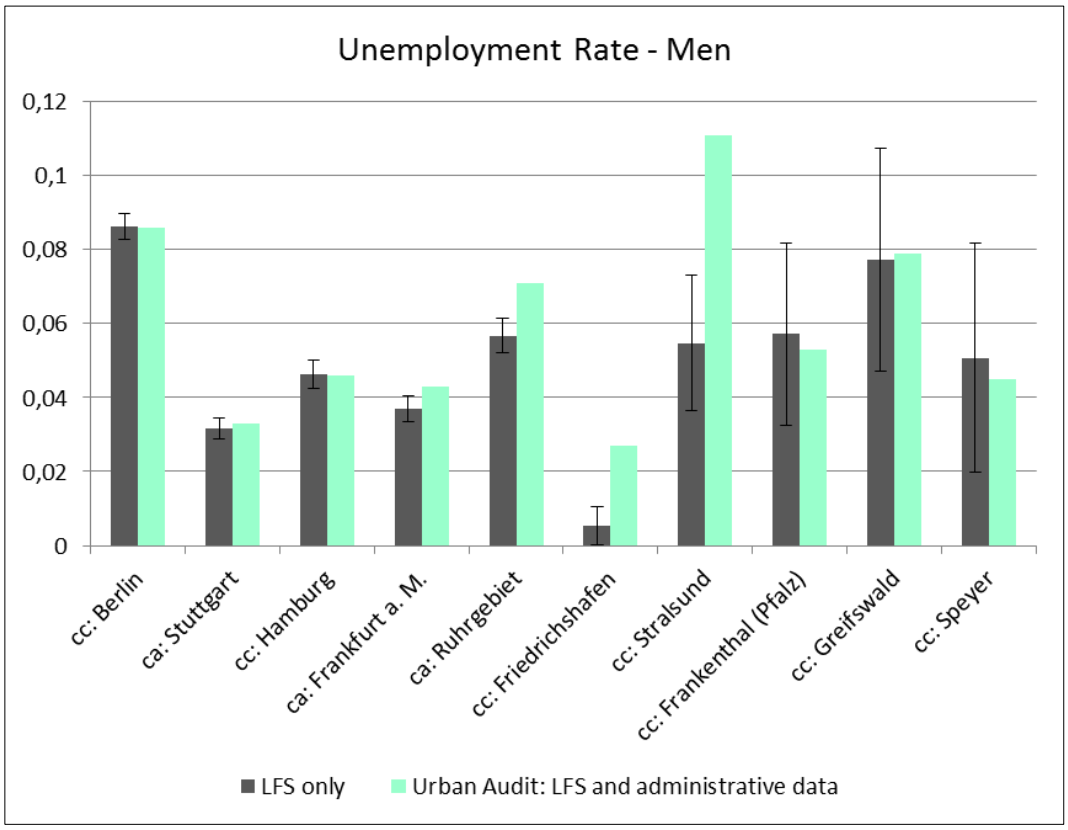
[50] A detailed description of this model based estimation can be found in Urban Audit 2016 (only available in German).

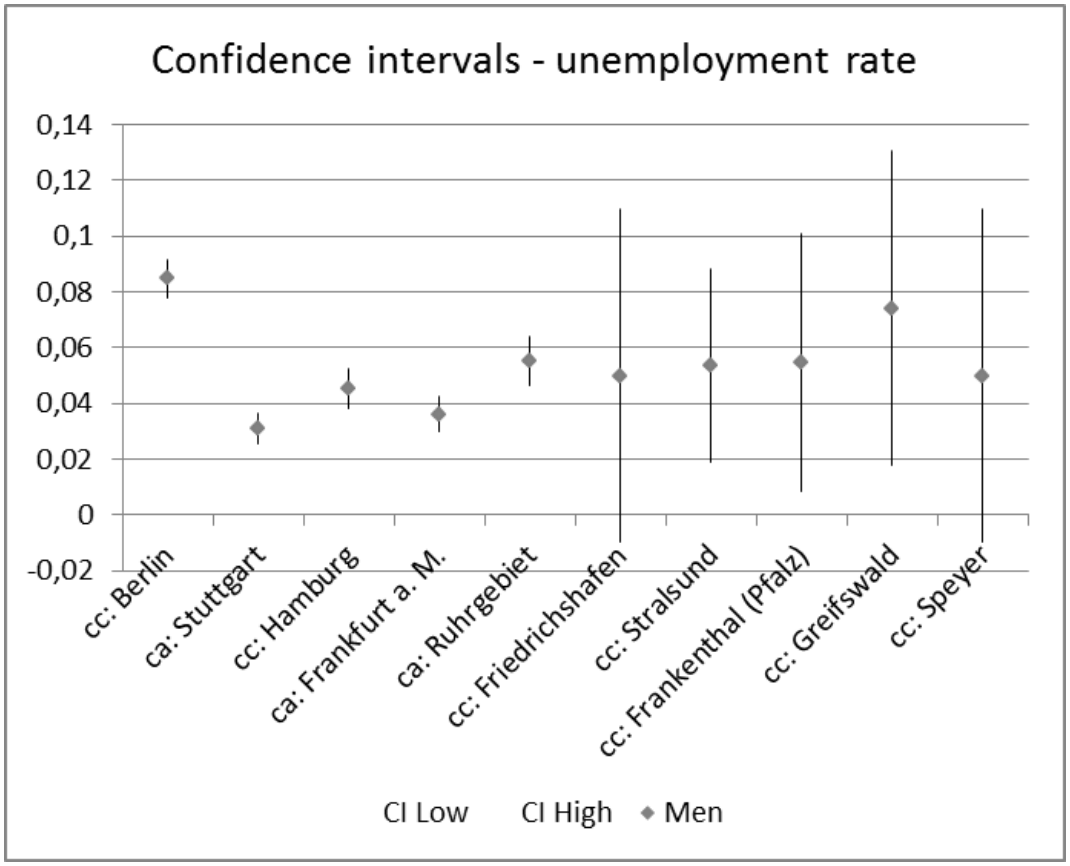
females are very similar for the areas with higher sample sizes like for Berlin, Stuttgart and Hamburg due to lower uncertainty in the data, which can be seen in the corresponding confidence intervals. The opposite is the case for smaller urban areas where the sample size becomes smaller. Since the direct estimator uses only LFS information, it does not produce reliable results for all units of FUAs, what is obvious for areas like “Stralsund” or “Friedrichshafen”. The probability to over- and underestimate the rate increases in these cases.
The 95% confidence interval for the unemployment rate for both genders is between -0.97% and 25.55%. As a result, the confidence interval increases too, compared to NUTS 2 regions. The 95% confidence interval of NUTS 2 regions for the unemployment rate is, for example, between 4.06% and 4.21%.[51]

Furthermore, Figure 21 compares the unemployment rate for selected medium sized urban areas with the published Urban Audit indicators. The areas are again ordered by their sample size in descending order from left to right. As described before, medium sized urban areas are defined urbanised areas with a resident population between 200,000 and 500,000 inhabitants.
The results indicate that the direct estimators are in all cases different from the Urban Audit estimators except the unemployment rate for women in “Magdeburg”. The difference between the estimated rates is highest for the urban area “Karlsruhe” where we overestimate the unemployment rate by 2.6 percentage points. For “Darmstadt” we underestimate the unemployment rate by 2.1 percentage points. With regard to the unemployment rate by men the differences are somewhat smaller. We overestimate the rate for the urban area “Darmstadt” by 1.4 percentage points and underestimate it by 1.7 percentage points for “Aschaffenburg”.
Furthermore, we obtain high CVs of minimum 17% up to a maximum of 47% for the unemployment rate for women and CVs of minimum 16% up to a maximum of 55% for the unemployment rate for men.

[51] See Eurostat (2018).

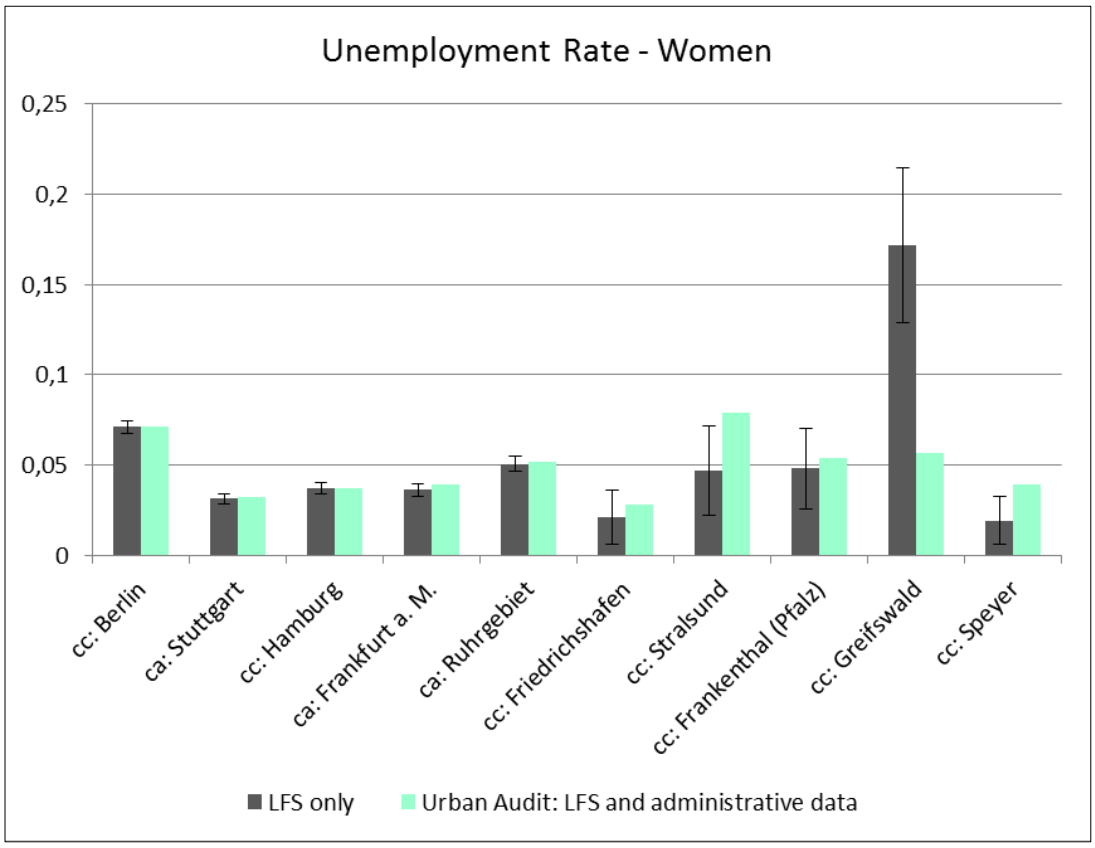
Figure 11: Unemployment rate and confidence intervals by men in selected FUAs: comparison of different estimations

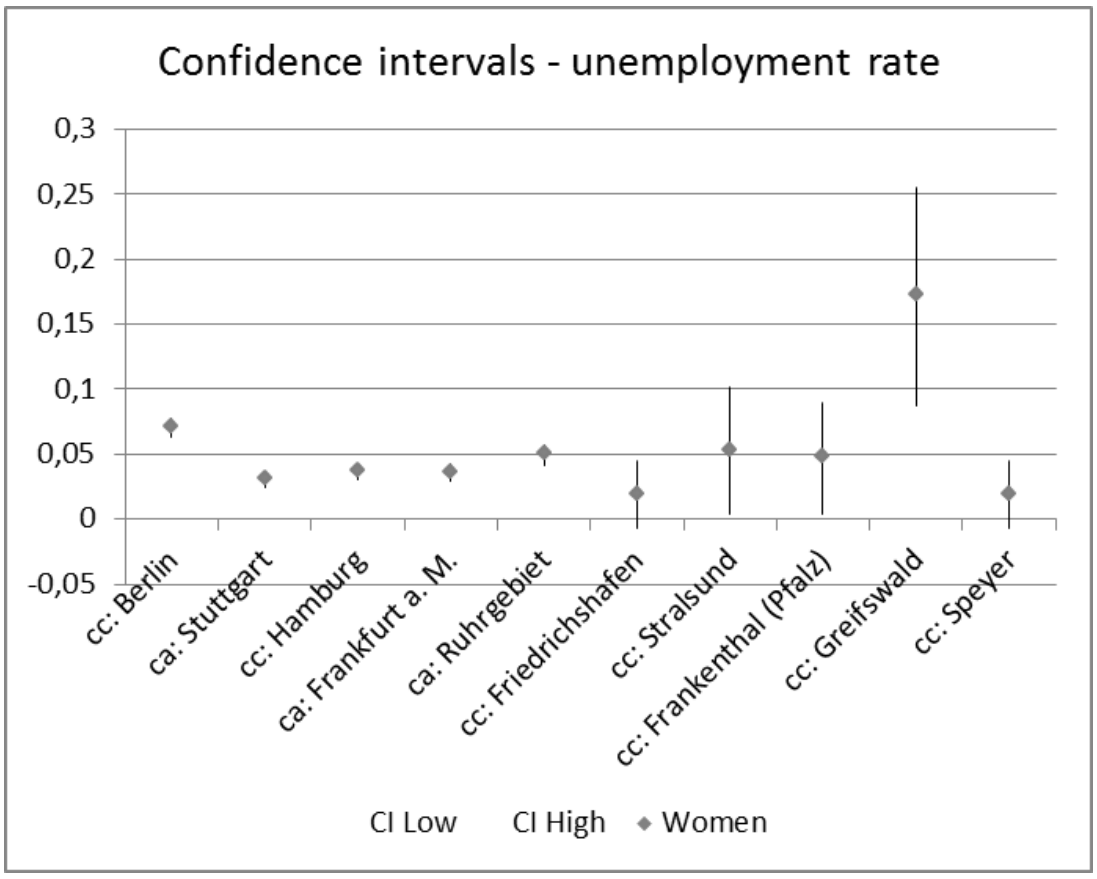




Others: ca: commuting area cc: city core

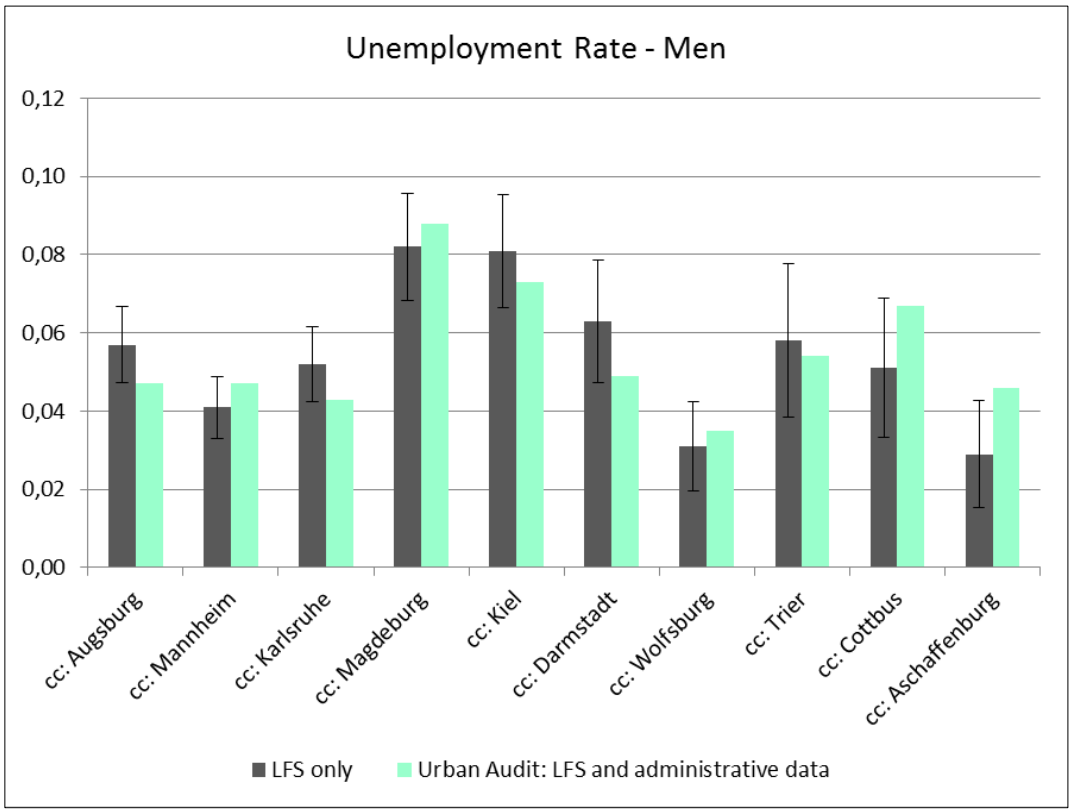
Figure 12: Unemployment rate and confidence intervals by women in selected FUAs: comparison of different estimations

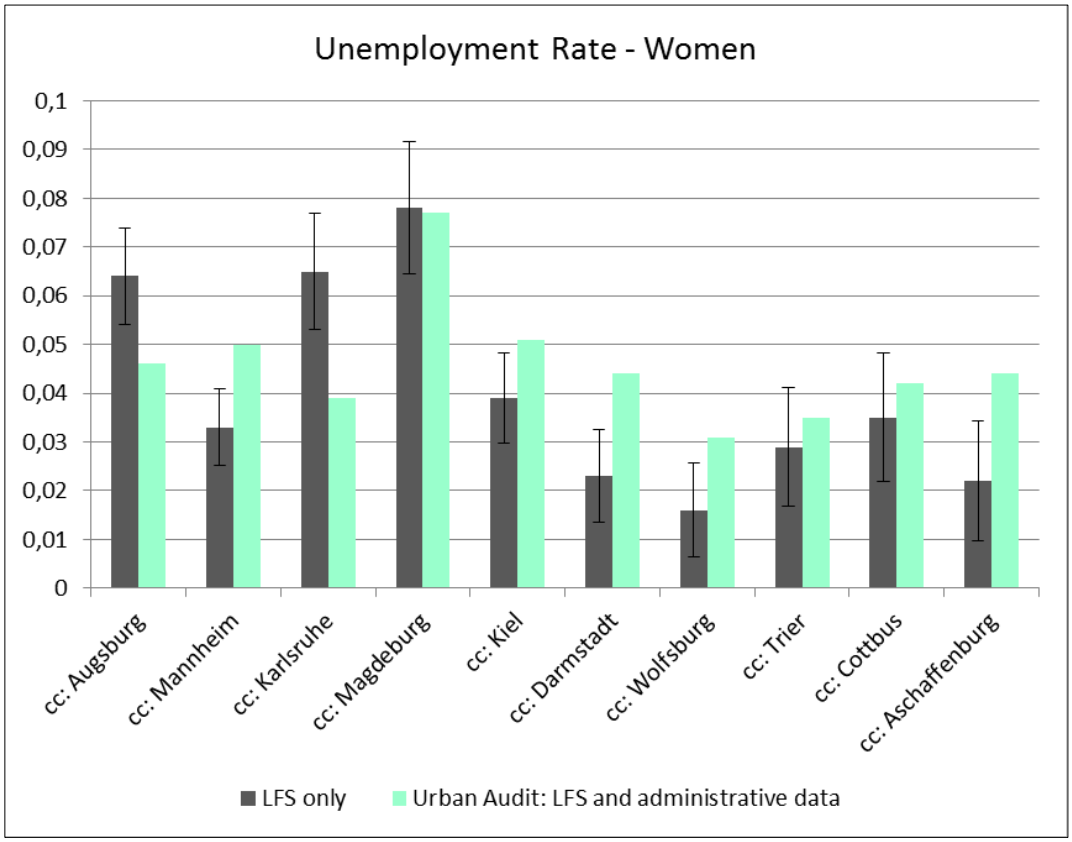




Others: ca: commuting area cc: city core

Figure 13: Unemployment rate by men in selected medium sized urban areas: comparison of different estimations





Others:

cc: city core

5. Conclusions

In summary, there is no clear trend in the indicators which are due to the areas. Uncertainty and deviation in the data and indicators depend on the number of observations. In general we obtain lower standard errors and CVs in areas with higher sample sizes and vice versa. But this is not necessarily the case for all indicators and areas. It depends above all on the number of observations of the interested variable, which is partly determined by the sample size. The areas with the smallest sample sizes are in general small urban areas. In comparison, metropolitan areas or large metropolitan areas are accompanied by higher sample sizes, lower CVs and unbiased estimates. Due to the high number of observations for estimating the employment rate, the direct estimation works quite well. The relative standard errors are comparatively low in comparison to those of the unemployment rate for FUAs. Conversely, the results of direct estimation for the unemployment rate go together with larger standard errors because of lower numbers of observations and biased estimators. Regarding the indicators with breakdowns increases the likelihood that out-of-sample domains arise because of missing observations in the sample. That problem does not happen in general if we are estimating the indicators without breakdowns. Especially in the distinction of age groups the problem of zero sample sizes occurs.

Overall, the methodological approach of the Urban Audit provides reliable indicators for the FUAs. Next to this approach alternative estimation methods can be used like small area estimation in combination with register data or administrative data. A direct estimation by using solely LFS information is not recommended for medium and smaller urban areas. But it is useful to get an overview of indicators for metropolitan and large metropolitan areas.

References

Eurostat (2018). Employment and unemployment (Labour force survey) (employ). National Reference Metadata in ESS Standard for Quality Reports Structure (ESQRS). Germany. Available at: https://ec.europa.eu/eurostat/cache/metadata/EN/employ_esqrs_de.htm#conf1538490502267

Eurostat (2017). Methodological manual on city statistics – 2017 edition, Luxembourg 2017. Available at: http://www.staedtestatistik.de/fileadmin/urban-audit/2017/2017manual_KS-GQ-17-006-EN-N.pdf also available at: http://ec.europa.eu/eurostat/documents/3859598/8012444/KS-GQ-17-006-EN-N.pdf/a3f1004f-cfae-4cc4-87da-81d588d67ae2

Urban Audit (2017). The German Urban Audit: Quality of Life and Suburban Areas, Joint project with the Federal Statistical Office and the Statistical Offices of the Federal States (Länder), promoted by Eurostat, Mannheim 2017. Available at: http://www.staedtestatistik.de/fileadmin/urban-audit/2018/UA_Broschuere_2017_eng_web.pdf

Urban Audit (2017). Das Deutsche Urban Audit – Lebensqualität in Stadt und Umland, Gemeinschaftsprojekt mit den Statistischen Ämtern des Bundes, Gefördert von Eurostat, Mannheim 2017. Available at: http://www.staedtestatistik.de/fileadmin/urban-audit/2017/UA_Broschuere_2017_Web.pdf

Urban Audit (2016). Regionalisierung des Mikrozensus für den europäischen Städtevergleich, Gemeinschaftsprojekt mit den Statistischen Ämtern des Bundes und der Länder, Gefördert von Eurostat, dem Statistischen Amt der Europäischen Union, Mannheim 2016. ONLY in GERMAN available at: http://www.staedtestatistik.de/1161.html?&K=630&F=1%20 und bei EUROSTAT

ANNEX 1.5 LFS DATA AT FUA LEVEL : THE FRENCH CASE

The French LFS : a survey for national analysis

The French LFS is a survey designed to provide results at national level. The sampling rate is about 1/400 in terms of dwellings. The rotational scheme of the sample is 6 consecutive waves. The national quarterly unemployment rate is estimated from the LFS with a 95 % confidence interval of ± 0,3 percent.

The sampling frame and the weighting procedure take into account theNUTS2 dimension. As a consequence, some indicators can be produced at NUTS2 level, but caution is needed regarding their reliability. At NUTS2 level, the confidence interval of the unemployment rate is about 2 percent for regions like Champagne-Ardenne, Picardie or Haute-Normandie ; it is more than 3 percent in Corse.

The French LFS does not allow analysis at a lower geographical level, including NUTS3 or FUA. Indeed, the sample is not meant to be representative at this level.

In France, analysis at local level are conducted using alternative sources (census, administrative data, or combined approach). For instance, the national unemployment rate is declined at infra-national level using administrative data on employment and registered unemployment (down to « employment zone » https://www.insee.fr/en/metadonnees/definition/c1361).

Data : https://www.insee.fr/fr/statistiques/1893230 (in French only)
Methodology : (https://www.insee.fr/en/metadonnees/source/indicateur/p1660/description

LFS data at FUA level : some quantitative results

Some results at FUA level from the French LFS :

- **the sample size is very small for some FUAs** : over a total of 64 FUAs, the quarterly sample size[52] is less than 1000 units for 53 FUAs, less than 500 units for 27 FUAs and less than 200 units for 7 FUAs.
- **the population at FUA level estimated from the LFS differs from the population estimated from the census[53]**, as this dimension is not taken into account in the weighting procedure : the absolute deviation exceeds 10 % for 37 FUAs, 20 % for 28 FUAs, 50 % for 5 FUAs.
- **the estimations of population at FUA level from the LFS show a high level of variability** : the absolute evolution of the total population estimated from the LFS between 2015 and 2017 exceeds 10 % for 26 FUAs, 20 % for 12 FUAs and 50 % for 3 FUAs.
- **the labour market indicators estimated from the LFS at FUA level show a high level of volatility** : while the national unemployment rate varies from 1 percent on the period 2014-2017, the range between the min and the max values on the period 2014-2017 exceeds 5 percent for 17 FUAs and 10 percent for 5 FUAs.

These evolutions cannot be explained by any change in the labour market situation. For example, if we consider the FUA of Nantes , the unemployment rate rises sharply between 2014 and 2015, then falls sharply between 2015 and 2017, while we observe a trend decline at national level. The smaller the FUA is, the greater the variability. For example, for the FUA of Pau, the unemployment rate rises from 5% to 10% between 2015 and 2016.

The same variability is observed when considering the employment rate of people aged 15 to 64.

---

[52] The quarterly sample size is more informative than the annual sample size as the annual sample counts several times a given sample unit according to the sample design.

[53] We compare the total population living in private households estimated from the LFS with the total population (including collective households) estimated from the census.

ANNEX 2 GERMANY MOBILE PHONE ANALYSIS

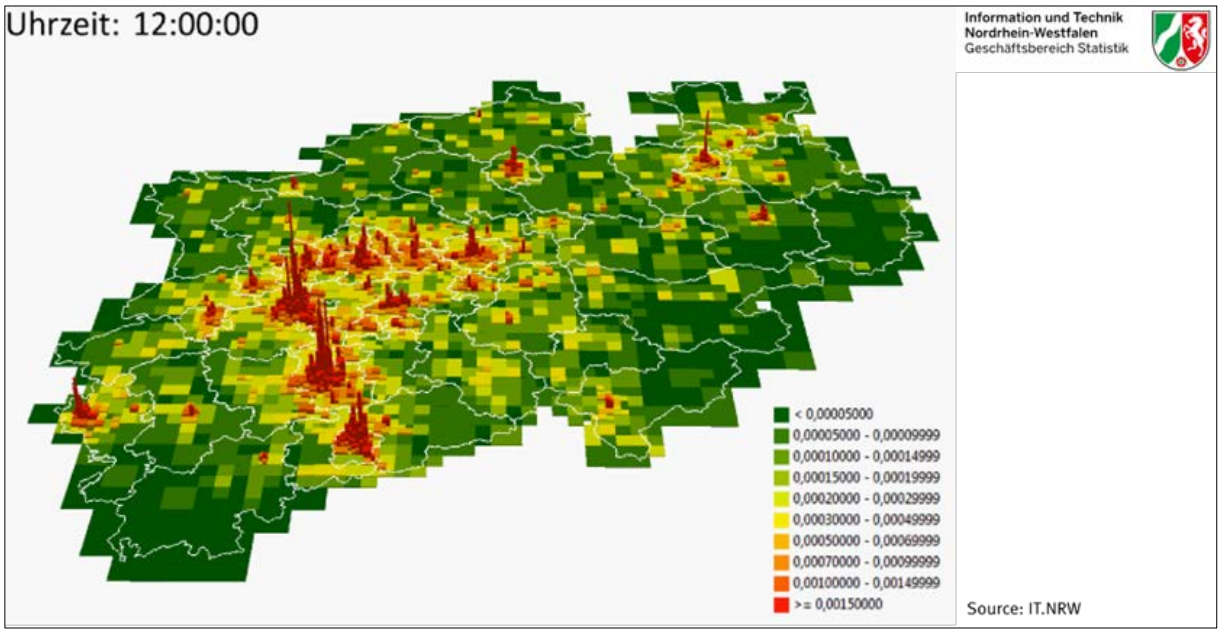


Figure 29: Change of mobile activities in the course of the day.

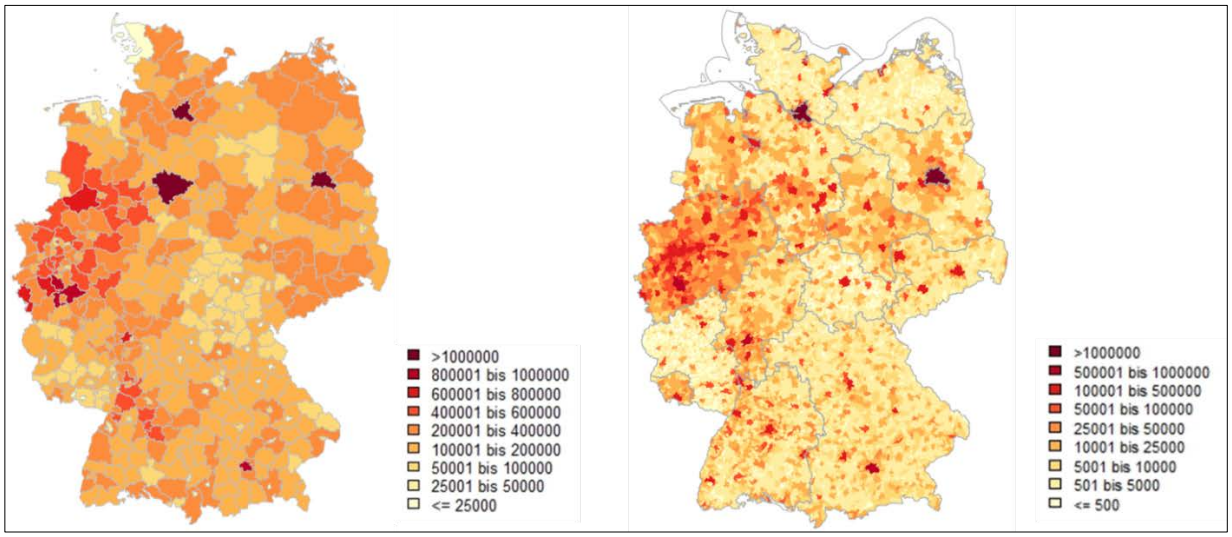


Figure 30: Extrapolated mobile phone activities at district and municipal level.

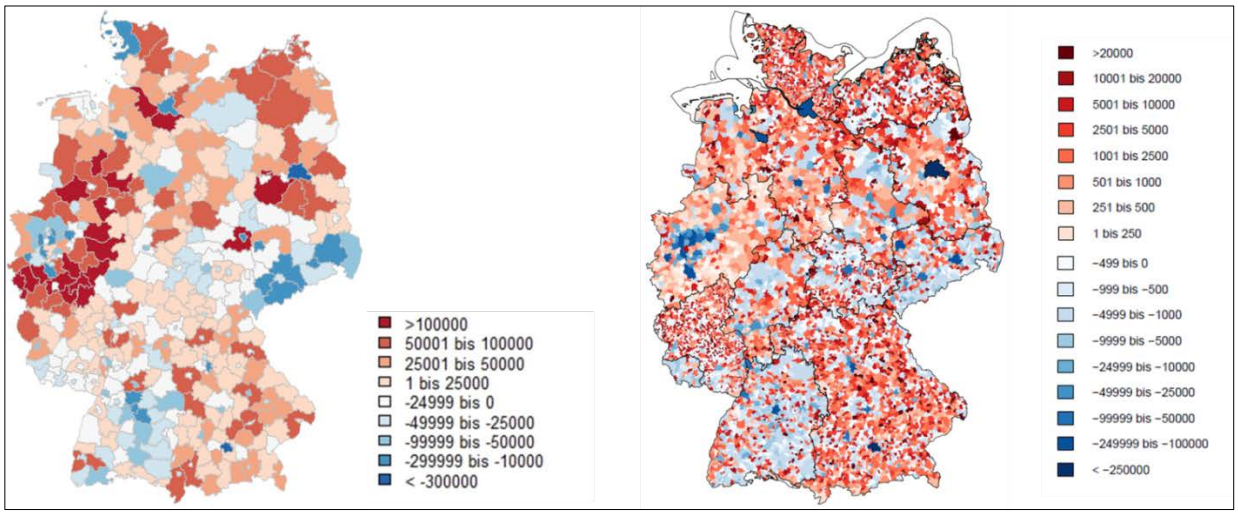


Figure 31: Difference between mobile phone activities and population at district and municipal level. Blue areas: less mobile activity counts than population figures from the 2011 census. Red areas: more mobile activity counts than population figures from the 2011 census.

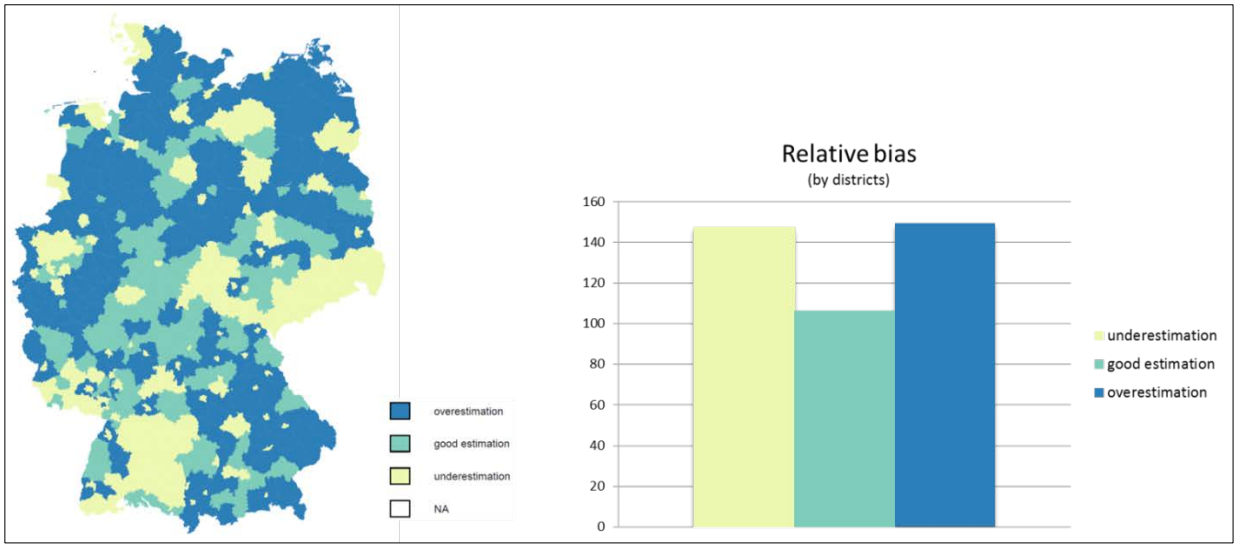


Figure 32: Relative bias between mobile counts and population figures from the 2011 census by districts.

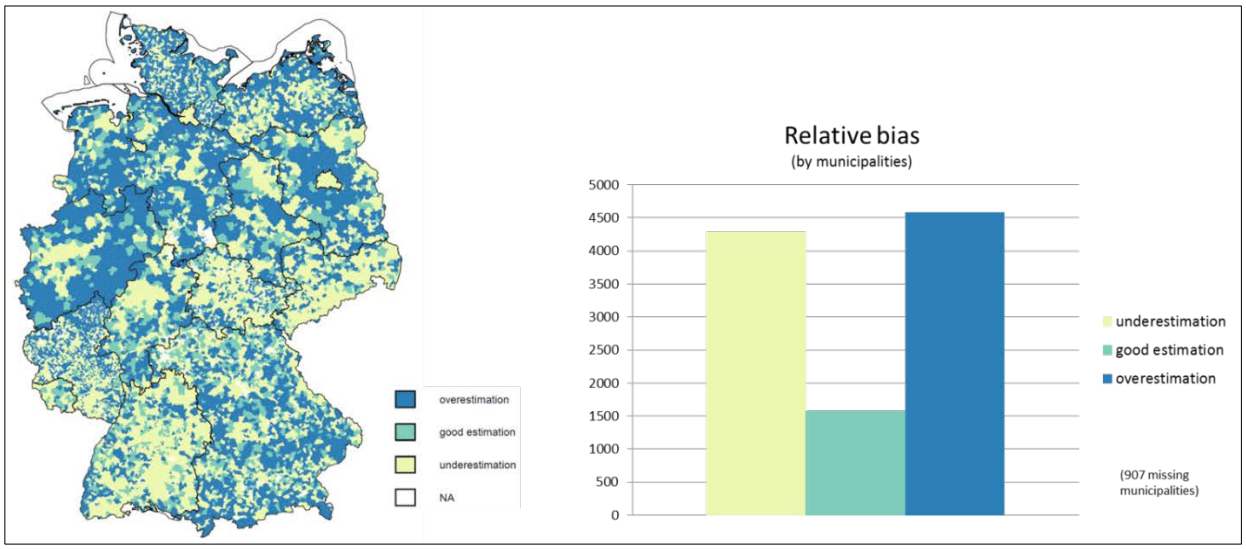


Figure 33: Relative bias between mobile counts and population figures from the 2011 census by municipalities.

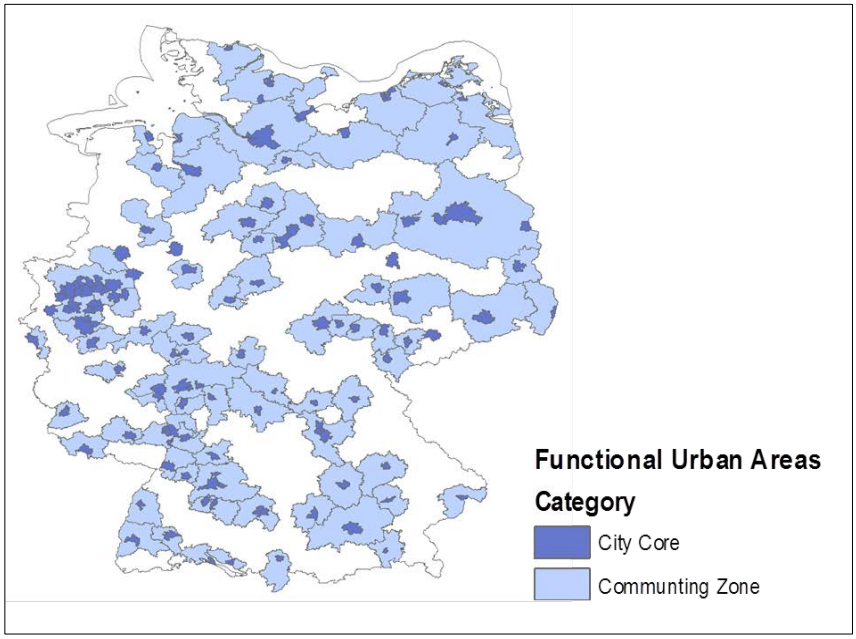


Figure 34: Functional Urban Areas in Germany (FUAs).

ANNEX 3 OVERVIEW OF EXPERIENCES OF USING MOBILE PHONE DATA BY MEMBERS OF THE PROJECT

Country: France
Spatial resolution of the aggregates
Aggregates are computed at 500m*500m grid cells. It appears to be a fair balance between the accuracy available in the data that may be at a finer level in city centres but lower in the suburbs. The mobile phone dataset used has very low spatial information, only cell towers coordinates. So in itself the geographic precision available is heterogeneous and quite vague. Yet with ancillary data from land use register it is possible to disaggregate mobile phone data at 500m*500m grid cells level with a gain in the accuracy of home detection. Besides at this level there is very few confidentiality issues with geocoded tax data in urban areas.
Mapping night-time (residential) population
We computed per month residential densities. Estimates are assessed against a tax register for validation. Correlation in the XX urban area is XX %. We detect quite efficiently the largest densities in the very centre of cities but there are still some discrepancies in the intra urban variation of population densities. The reason for that probably comes from the age of the mobile phone dataset available (2007), its nature (Call Details Records) and the lack of additional information on the spatial attributes of the data (antennae type or configuration…).
Mapping day-time population and flow data
Because the average number of detection per mobile phone per day is around 4, we did not produce any day-time population densities nor flow data and concentrated on validating the residential estimates.
Data access circumstances
We had access to the data at the MNO premises, working on the MNO infrastructure. Only aggregates could be exported to the NSI. The infrastructure is a big data platform built on an HDFS and programs were written in pySpark.

Country: Germany

Spatial resolution of the aggregates:

In principle, the Deutsche Telekom can supply any spatial resolution of the aggregates as long as the minimum number of 30 mobile activities of German Telekom customers per geometry is reached. In the present case, the data is available on grid cells of different sizes for North-Rhine Westphalia (NRW), which can be converted to the municipal and NUTS 3 level (district level) with the aid of a cartographic kernel density estimation. The grid cells conform with INSPIRE and correspond to the grid cells of the 2011 Census Atlas. Reason(s) for the use of this resolution, limitations and opportunities (confidentiality, pre-aggregation,…): In compliance with data protection rules, the mobile phone activities were anonymised aggregated and already distributed to the desired geographical level, in the present case on grid cells by the Deutsche Telekom. Since the activities are aggregated and have already been distributed to specific geometries, it is not advisable to include other data sources to distribute the aggregated activities. Only values based on a minimum of 30 activities per grid cell or geometry were transmitted to Destatis. The number of mobile phone activities depends on the location and number of cell towers in the various grid cells. Depending on the cell towers' location (rural or urban), their frequencies differ and, as a result, they are sometimes distributed unevenly across the regions. Consequently, an existing geometry may contain 5 to 20 cell towers. For that reason, some grid cells are combined to ensure a minimum of 30 activities per grid cell, which vary from 500 x 500 to 8000 x 8000 meters.

Mapping night-time (residential) population

The population figures from 2011 census were used to determine the correlation between mobile phone activities and census values by type of day and time of day for NRW. Overall, the values reveal a high correlation of 0.8 between mobile phone activities and census values throughout Saturday and Sunday. In order to represent the resident population, the period of a statistical Sunday evening from 8 p.m. to 11 p.m. is selected to count only activities at the place of residence. For a unified and visualisable comparison, it is helpful to transfer the aggregated data to a higher geometric level with the help of the core density estimation in order to make further content analyses of the distributions of the mobile activities. To enable a direct comparison with the population figures of the census and mobile activities, the latter was extrapolated with a correction factor, which is calculated from the sum of the population divided by the sum of the mobile phone data in the period under consideration. This allows considering the current distribution of the entire German population on the basis of mobile phone data. The results show that, to some extent, the mobile phone data available could provide a good picture of the population. The differences observed between the population figures based on mobile phone data and those based on census values may partly be explained by the time difference between the mobile phone data from summer/autumn 2017 and the census data from 2011, but they may also result from the extrapolation method used by mobile network operator. This allows considering the current distribution of the entire German population on the basis of mobile phone data.

Mapping day-time population

The population figures from 2011 census were used to determine the correlation between mobile phone activities and census values by type of day and time of day for NRW. On weekdays, the correlation declines to less than 0.7 between 5 a.m. and 4 p.m., which indicates significant differences in the resident population according to the 2011 census and to the location of mobile phone activities within the given period, which refers to the daytime population. The afternoon or noon in NRW is well suited to represent the day-time population with mobile phone data. See for further information ‘Results, opportunities, limitations’ from ‘Mapping night-time (residential) population’.

Mapping flow data (especially between night-time and day-time)
Using the current available data, it is however not yet possible to describe the commuting patterns as such, i.e. the movement in space. Nevertheless, the results allow deducing commuter regions.
Data access circumstances
Access to data in Germany only through cooperation agreements. The German mobile communications market currently consists of the three providers Telekom, Vodafone and Telefónica with a respective market share of one-third each. In order to research the use of mobile phone data for official statistics, the Federal Statistical Office of Germany (Destatis) entered into a cooperation with T-Systems International GmbH and Motionlogic GmbH (both wholly-owned subsidiaries of Deutsche Telekom AG)

Country: The Netherlands
Spatial resolution of the aggregates
For objectives of this project, due to methodological challenges, aggregates were computed at municipal level. However aggregates can be defined in different spatial resolution as administrative regions such as provinces, municipalities, its districts (wijken) and neighborhoods (buurten), as well as grid cells. The size of a grid cells and the level of administrative regions' detail depends on the antennas density and number of devices connecting to antenna, accuracy of the geolocation method and the mass lost from the cube by the threshold of 15 devices .
Mapping day- and night-time residential population and its flows
As generated flow cube of persons includes place of residence and presence, we were able to map day- and night-time residential population as well as its flows at municipal level (see Section 4.4.4). Limitations: more detailed special resolution requires improvement of geolocation method; data in complicity (SN works with data from one of three MNOs); mass lost as a result of preventing disclose etc. Opportunities: improving methodology by conducting research on time advance parameter and use of 2G and 3G, next to 4G, generated records, establishing collaboration with other MNOs.
Data access circumstances
Statistics Netherlands (SN) works with MNO T-Mobile directly, what allows to study but not export the anonymised micro data made available, understand its structure and get acquainted with the technical infrastructure surrounding it. We would like to stress the fact that none of these data are shared with SN. Only the anonymised aggregated outcome data were delivered to SN to produce statistics. A collaboration was started on the basis of a contract for half a year with an option for extension under agreement of both parties, and this has been successfully continued.

Country: Belgium
Spatial resolution of the aggregates
Proximus mobile network cells (Voronoi shapefiles which can also serve to aggregate statistical variables), on average 2.78 km² (surface of Belgium 30.528 km² divided by ~11,000 network cells), greatly varying in size from very small in city centre to very large in forest areas. Reason(s) for the use of this resolution, limitations and opportunities (confidentiality, pre-aggregation,…): smallest possible geographical level to which device can be assigned with the present technology, amply sufficient for most statistical purposes
Mapping night-time (residential) population
most frequent network cell presence (modus of distribution of cells) of device at 04:00 during the 'typical' month of October. Results, opportunities, limitations: fairly simple but probably sufficient as algorithm, to be tested and validated in practice
Mapping day-time population
Number of devices per network cell at 45 (or 96) points in time during every day of the last 12 months (365 days), extrapolated for the local Proximus market share (approximately 180 million records) Results, opportunities, limitations: this should provide a very granular view, both in time and space, of the evolution of present population and allow linking to other spatiotemporal datasets (seasonal, weekday-weekend, weather conditions, land use, etc.).
Mapping flow data (especially between night-time and day-time)
After having established living place (see above) and workplace (=most frequent network cell at a selection of business hours in October) a 11,000x11,000 matrix linking both at aggregate level can be compiled. These should ideally be adjusted for local market share at the living place network cell. Results, opportunities, limitations: this fairly limited dataset will provide a geographically very detailed view of commuting movements across the whole Belgian territory.
Data access circumstances
No data access at present, terminated by higher management, but talks still ongoing on commercial use case which may interest sales department of network operator, and could be the basis for obtaining the data mentioned above

Country:Austria
Spatial resolution of the aggregates
Inbound: municipalities, tourism regions; Outbound: countries. The aggregation level was chosen to have appropriate comparison values in the traditional statistical products.
Mapping night-time (residential) population
Not implemented at the moment, but this could be done with a similar methodology to the inbound tourisms, but the target group would have to be changed to Austrian sim-cards.
Mapping day-time population
Not implemented
Mapping flow data (especially between night-time and day-time)
Not implemented
Data access circumstances
Currently Statistics Austria is in an intense cooperation with one of the MNOs to improve the estimation and investigate possible usage in our statistical production process. There is no access to microdata, but only to aggregated numbers.

Getting in touch with the EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

– by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

– at the following standard number: +32 22999696 or

– by email via: https://europa.eu/european-union/contact_en

Finding information about the EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1952 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

Open data from the EU

The EU Open Data Portal (http://data.europa.eu/euodp/en) provides access to datasets from the EU. Data can be downloaded and reused for free, for both commercial and non-commercial purposes.

KN-02-19-484-EN-N

Publications Office

doi: 10.2776/120066