Handbook on the use of Mobile Phone data for Official Statistics

UN Global Working Group on Big Data for Official Statistics

DRAFT

November 2017

Table of Contents

1. Introduction

[Context of the GWG]

[Big Data documents as reference / introductory document for capacity building]

[General overview of work already done on the use of Mobile Phone data for official statistics: Eurostat, Positium, Orange, Telenor, with forward reference to Chapter 2]

[Overview of the Chapters in this document]

Mobile phone data has surfaced in recent years as one of the big data sources with a lot of promise for its use in official statistics. There is an expectation that mobile phone data could fill data gaps especially for developing countries given their high penetration rates. In its 2014 report[1] the International Telecommunication Union (ITU) shows that the average mobile subscription rate is 96.4 per 100 inhabitants world-wide, with some lower averages in Asia (89.2) and Africa (69.3). Nevertheless, these numbers show how pervasive mobile phone use is. ITU elaborates that rural areas are still lacking behind urban areas, and this should be considered in studies using mobile phone data, but it is clear that the coverage of these data is global. Almost every person in the world lives within reach of a mobile-cellular signal.

The Task Team on Mobile Phone Data has been reviewing the development in the use of mobile phone data for official statistics and drafted this guide to give elaborations for countries which are in the process of designing project for using mobile phone data for their official statistics or are in the implementation process. The guide is to be published only in electronic format and in English and is expected to be reviewed and amended as projects are being advanced.

The guide composes of four chapters (Applications, Data sources, Access to mobile phone data and partnership models, Methods) but also contains country case studies in the Annex.

[1] https://www.itu.int/en/ITU-D/Statistics/Documents/publications/mis2014/MIS2014_without_Annex_4.pdf

2. Applications

Businesses have been using big data for a longer period of time meanwhile national governments including the national statistical offices as well as international organizations have started checking how this data source could be used for producing official statistics and as such to support the decision making more efficiently. Many applications have been developed in the past decade but our target to review in this chapter is some of those applications which particularly based on the use of mobile phone data and are relevant in producing official statistics, including tourism statistics, events statistics, population statistics, migration, commuting statistics, traffic flow/monitoring, transportation statistics, urban planning, employment statistics or on statistics of border/seasonal workers.

2.1. Tourism and event statistics

In the following sub-chapter several studies focusing on the use of mobile phone data for tourism and event statistics are reviewed.

2.1.1. Use of mobile positioning data in tourism statistics, a study by Eurostat

The aim of this study[2] commissioned by Eurostat was to assess the feasibility of using mobile positioning data for generating statistics on domestic, outbound and inbound tourism flows, and to address the strengths and weaknesses related to access, trust, cost and the technological and methodological challenges inherent in the use of this data source. They noted that inbound, outbound roaming and domestic data stored by mobile network operators (MNO) clearly corresponds to the respective inbound, outbound and domestic domains of tourism, however, not without some methodological reservations. National statistical institutes (NSI) perceive this data source mostly as complementary and in some cases, it is also seen as a potential replacement for existing data sources and methodologies. The main findings of the study are: 1. Access to mobile positioning data is currently very limited mainly because of the regulatory limitations. There are major differences between countries in this regard. 2. The study concludes that there is a need for a central framework for NSIs and other stakeholders, in order to obtain the data legally and according to an approved methodology, in order to be able to produce comparable and reliable tourism statistics. 3. Longitudinal data is a must for reliable tourism statistics in order to assess the whereabouts of the subscribers over a longer period of time (eg. Usual environment, differentiation of the trips by length, identification of overnight stays, etc). 4. Based on the outcome of this study, it can be concluded that at present mobile positioning data can be used as a supplement rather than as a replacement source of data for tourism indicators. 5. However, the use of mobile data as a source for tourism indicators introduces several aspects of improvement compared to the existing statistical processes, such as: timeliness (in some cases up to near real time), access to statistical information previously not available (new indicators), calibration

[2] "Feasibility study on the use of mobile positioning data for tourism statistics. Consolidated report Eurostat contract No 3051.2012.001-202.452" http://ec.europa.eu/eurostat/documents/747990/6225717/MP-Consolidated-report.pdf

opportunities for existing data, better resolution, and accuracy in time and space. 6. Mobile positioning data can complement the currently used methods through mixed-mode data collection enabling the sample size of the conventional survey to be decreased.

The methodology explored in the study was the statistical use of location information from technical databases and registries that concern the historical location of mobile devices within the network of mobile network operators (MNOs). The aim of the study was to assess the feasibility of using such data for estimating domestic, outbound and inbound tourism flows and to address the strengths and weaknesses that are related to access, trust, cost and the methodological challenges of using mobile positioning data in tourism statistics. The specific reports of the study include:

Report 1. Stock-taking makes an inventory of all of the research that has been conducted to date, along with applications and experience, in EU member states, EFTA, candidate countries and around the world, and provides an up-to-date description on the state of the art use of mobile positioning data in research and applications in tourism statistics and related domains. The results of the usage case analysis highlight the fact that there is an increasing number of mobile data-based studies, research papers, projects, applications and businesses being created. MNOs are overcoming privacy and business confidentiality-related concerns as they see the appearance of new revenue possibilities, and also the value that can be gained from the internal use of such data. The shortcomings of the data include a lack of qualitative information on the user, such as the purpose of the visits and the means of transport; as well as privacy issues and challenges in processing large amounts of data. It was discovered that a major problem for statistical bodies was not methodological, but rather issues concerning access to data, privacy concerns, and the relatively high 'entry cost' of using new ICT-based data sources.

Report 2. Feasibility of access mainly focuses on access to the data in order to produce official tourism statistics for the NSIs, but other usages are also considered. The report concentrates on regulatory, business and technological barriers along with practical access to the data. MNOs seem to realize the potential of using this data source but see many obstacles in terms of privacy protection and legislation, user adaptation (the switch to new data sources from traditional practices), the representative nature of the data, and various methods for translating mobile data so that it represents the 'real world.'

The focus of *report 3b. feasibility of use: coherence* is on carrying out a quantitative and qualitative comparison of mobile phone based tourism statistics with reference statistics that include official tourism statistics and other available indicators related to tourism activities. There are several examples presented to illustrate the different aspects of coherence. One of them is regarding analysing the total volume of inbound and outbound ferry passengers travelling between Estonia and Finland/Sweden. In this case, mobile positioning data underestimates real passenger volumes as reflected by reference statistics due to the other nationalities and transit passengers on board. The study also found that when compared to demand statistics, mobile positioning data also provides a consistent estimate of the total number of trips of Finns to Estonia when compared to Finnish demand statistics. The

consistency of mobile positioning data against outbound trips in Estonian demand statistics is only moderate; many Estonians commute to Finland, making Finland part of their usual environment. The over-coverage in mobile positioning data is highest in the first and fourth quarters, both of which are outside the summer holiday season, which indicates that part of the outbound trips in mobile positioning data are made by these commuters. One of the weaknesses of mobile positioning data is the potential misclassification of same-day and overnight trips, as at least two mobile phone activities are necessary at long enough time intervals (on different calendar dates) for a trip to be classified as an overnight trip. Mobile positioning data clearly overestimates the volume of Finnish same-day inbound trips to Estonia and underestimates the volume of overnight trips. When comparing Estonian domestic mobile positioning data against accommodation statistics, the trends shown in overnight trips coincide with arrivals in accommodation statistics. For demand statistics, when making a comparison against Estonian tourism demand statistics, the difference in domestic overnight trips is relatively small.

Regarding overnight trips, mobile positioning data provides a consistent estimate over time and the correlation to existing tourism statistics, both in terms of demand and supply, is very good for the most part. While mobile positioning data cannot provide the split into paid and non-paid accommodation (a problem only if the split as such is relevant for users), it is nevertheless a more comprehensive data source for overnight trips. As it can be assumed that traditional sources tend to underestimate real tourism flows, it can be argued that mobile positioning data provides a more realistic picture of complete tourism flows. Although mobile positioning data provide fairly good estimates for the number of trips, the nights spent and the destination, they do not produce any additional data about trips such as the purpose of the trip, the type of accommodation or the expenditure. Mobile positioning data can be used to potentially strengthen current tourism demand surveys through mixed-mode data collection. In such a scenario the number and duration of trips are based on mobile position data, while tourism expenditure and ratios (purpose of the trip, type of accommodation, means of transport, etc) still rely on demand survey. The sample size of the demand survey could be decreased considerably since the survey does not need to support breakdown by destination, thereby reducing the cost and burden of data collection. Also more countries and even sub-regions could be included in the statistics since the sample size is not an issue in mobile positioning data. When compared to the workload induced by traditional data sources and processes, mobile positioning data can be obtained and processed rather more efficiently. When compared to surveys, mobile positioning data shows a number of advantages in terms of accuracy (smaller sampling error, no memory gaps), regardless of the selection bias. It has been shown in pilot studies that mobile positioning data can be used in terms of official statistics, specifically in the travel item of the Balance of Payments in transport and commuting statistics. A higher impact will probably arise from the transmission of data from MNOs to the NSI—but this only after the necessary automation processes have been carefully planned and thoroughly tested. The heterogeneity of rules and regulations concerning access to mobile positioning data does not allow for useful application in all countries. Some existing geographical situations may render the use of such data particularly

hard (e.g. a high share of migrants producing roaming data without being part of any tourism activities, as might be the case in Luxembourg).

2.1.2. Use of mobile positioning data in tourism statistics, an Estonian case study

In their research project[3], Ahas et al.used passive mobile positioning data (meaning, location data that is stored automatically in the databases of mobile operators when a person uses the mobile phone) recorded at EMT, the largest network operator in Estonia. The data was recorded for all call activities made by foreign phones in the EMT network (roaming) with the precision of Cell ID. Call activity means any active use of the phone, including making or receiving calls, SMSs, using the internet or location-based services. Normally the phone switches to the closest or strongest radio coverage antenna, which allows to roughly estimate the location of the phone owner. In the data used, all phone identifiers remained anonymous, and only the country of registration of the phone was recorded. The researchers found that the national distribution of call activities in Estonia is similar to the official accommodation statistics. They also found that visitors from farther destinations stay longer in Estonia, and that for these visitors the discrepancy between call activities and accommodation statistics is smaller. In addition, when looking specifically at tourists in the city of Tartu, the researchers found that the differences between the 2 data sets (mobile positioning and accommodation statistics) are minimal, and that mobile positioning data tends to underestimate the number of overnight visitors from farther origin countries, while overestimating the number of overnight visitors from neighbouring countries. In addition, the researchers also analysed the other locations in Estonia that were visited by the same tourists who had visited Tartu, which could be useful in tourism marketing and developing interregional tourism cooperation in the country. The researchers also examined event statistics and their effect on tourism statistics. They discovered that the effect of agricultural fairs and international events on tourism was very strong.

The researchers concluded that mobile positioning data is a promising source of information for the investigation and management of tourism, with the advantages of this data source being low cost and high quantity, as well as better temporal and spatial precision compared to standard tourism statistics such as accommodation statistics. Moreover, the digital records of a tourist's movements make it possible to analyse space-time movement, which allows the linking of routes used by tourists with the places they visited. In addition, mobile phone data is excellent for studying tourism in less visited areas where it is difficult to use other methods such as accommodation statistics or questionnaires. However, they also note that while the data allows one to analyse tourist flows to some extent, it does not answer important questions, such as why, how and who (eg, this data does not offer demographic information such as age or gender). In addition, the researchers note that the cost of roaming calls and

[3] "Mobile positioning data in tourism studies and monitoring: case study in Tartu, Estonia". Ahas et al. https://pdfs.semanticscholar.org/492b/53aa1c540ac9db823faefef3b1fef5f229b9.pdf

different income rates in home countries may influence the use of phones in the destination country, and that this needs further study.

2.1.3. Sport and cultural events and destination loyalty, an Estonian case study

Kuusik A, Ahas R. and Tiru M. investigate in this paper[4] the ability of the sport and cultural events to generate destination loyalty and repeat visitations. Passive mobile positioning (PMP)—data that is automatically stored in mobile operators' memory files as locations of telephones or call activities in network—method was used to analyse the behaviour of tourists during a couple of years. The data of one of the three Estonian mobile network owners, EMT was used in the study, which has a penetration rate of 40-45%. Ten different events were studied. For each event, a number of visitors of the event were filtered based on the proximate antennae and time of the event. The filtered visits dataset combines all visits of those tourists who visited selected events and shows how many visitors had previously visited Estonia, how many of them visited Estonia for the first time and how many of them came back later. The dataset combines the total of 4.5 years of visitor data, so the visitors might have changed their phone number during this period, which would cause them to be misrepresented in the data. However, because all the events have the same risk of bias, the comparison between them is possible in relative terms. The locations of antennae and the dates of the increase were compared against the list of events. Ten specific events were matched this way and people who made call activities during the period in that location were flagged as event visitors of these events. Visitors from neighbouring countries are most common visitors of events in Estonia, with majority of tourists being from Finland, Latvia, Sweden, Russia, Lithuania and Norway. The nationality of events visitors is usually similar to an average Estonian visitors' nationality, but there are also some niche events attracting also non-traditional nationalities to Estonia (eg Karate Europe Championship). Visitor numbers to those events often reflect the organizers and participants of the event rather than ticket-buying visitors, though from the perspective of tourism industry they are no less important. For all events investigated in the current paper, the rate of the first-time visitors was above 35%. The most productive events in generating repeat visitors were Alexela Rally and the Kalev Horse Show. On the other hand, very specific sports events such as football match or Karate EM practically did not generate repeat visitors. In addition, most of the events that generated repeat visitors, generated country-based destination loyalty, with the exceptions being periodic specific sports and cultural events which generated both country-based and event based destination loyalty. Overall, the results revealed that all events are good generators of new visitors. This has practical implications for tourism authorities, as it is cheaper to keep existing customers than trying to get new ones. Thus, instead of generating expensive advertising campaigns to get new tourists, it should be more reasonable to use events and pay more attention to the effort to get visitors of the events back to Estonia.

[4] "The ability of tourism events to generate destination loyalty towards the country: an Estonian case study." Andres Kuusik, Rein Ahas, and Margus Tiru. https://ojs.utlib.ee/index.php/TPEP/article/download/878/855

2.1.4. Destination choice based on weather and climate, an Estonian case study

Studies show that weather and climate is one of the most significant factors for destination choice. The authors, Järv O., Aasa A., Ahas R. and Saluveer E., analysed tourists' spatial behaviour using passive mobile phone positioning. This paper[5] tried to find answers to the questions: how weather affects tourist spatial movement within a local destination? Do tourists' behaviour patterns in inland areas resemble those in coastal areas? Does weather affect tourists' behaviour in bigger cities? For this study, the inbound tourist was defined as the foreign visitor who visited Estonia (irrespectively of the purpose of the visit) and during that time used his/her mobile phone in the biggest mobile network provider in Estonia. The strongest correlations between temperature and number of foreign visitors were found for daily mean air temperature with a correlation of 0.48. In the bigger cities, the relationship between temperature and the number of tourists was weak (rho=0.16). Excluding the data of bigger cities, the correlation coefficient in the rest of Estonia increased up to 0.58. Network cells with the strongest positive correlations were located in the coastal areas along the northern Estonian coast, along the north coast of Lake Peipsi and in western Estonia. The results of the paper showed that for Estonia, inbound tourist numbers in the summer period are positively correlated to air temperature. The locations with the highest correlation are holiday destinations along the coast where beach tourism dominates. Tourists who visit inland areas with fewer tourist attractions do not appear to be affected by weather.

2.2. Population statistics

Human populations are dynamic, moving daily, seasonally, and annually, resulting in rapidly changing densities. During the past few decades, the mobile phones are being widely used, and now have an extremely high penetration rate across the globe. With the increasing penetration of mobile phones, mobile positioning data as a new data source can be used as supplement for the current official population statistics. As well as in operational and governmental decisions, this new data should improve the quality and efficiency of making urban planning and management.

2.2.1. Improving population statistics with mobile data

As we all know, traditional methods of population statistics are Population Census, population administrative register, and household surveys. The datasets derived from current population statistics are accurate, detailed, static, and spatial. Although mobile data does not provide the accuracy of a census, a comparison of mobile data with a census and the population register shows a high correlation[6,7] and much more dynamic in the aspect of rapid changes. So, in terms of timeliness, quick and short-time indicators could be deprived from using mobile phone data as an additional source of population information.

[5] "Weather dependence of tourist's spatial behaviour and destination choices: case study with passive mobile positioning data in Estonia." O. Järv, A. Aasa, R. Ahas and E. Saluveer.

[6] "Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics Report 4. Opportunities and Benefits"(2014): http://unstats.un.org/unsd/trade/events/2014/Beijing/documents/phonedata/MP_Task%204%20report.pdf

[7] "A Study on Urban Mobility and Dynamic Population Estimation by Using Aggregate Mobile Phone Sources" (Chapter 6): http://www.csis.u-tokyo.ac.jp/dp/115.pdf

In countries where detailed human population census data are available at high resolution, the main value of using mobile data is not only so much in the gain in spatial resolution, but also more in the ability to estimate population numbers and densities at high spatial resolution for any time period. Using mobile data allows us to follow how population distribution changes through time in relation to the week, the season, or any particular event affecting population over large spatial extents.

Research shows that in low-income countries where population distribution data may be scarce, outdated, and unreliable, the mobile positioning data is particularly important.[8]In Malawi for example, censuses have been performed once per decade for the past three decades and data are readily available at the level of enumeration areas (i.e., administrative units of 9.38 km2 on average).In contrast, in the Democratic Republic of the Congo (DRC), the most recent census was undertaken in 1984 and data are available only at the level of territories (i.e., administrative units of 12,466 km2 on average). However, in the DRC, the mobile phone penetration rate, although biased toward certain demographic groups, is relatively high [69% on average by the end of 2014 in Africa (31)], and the mobile phone approach would produce considerable improvements in current knowledge of how population is distributed in the country.

Based on findings from the examined use cases we can conclude that mobile data is being used increasingly in a number of different fields. Most of the active usage is within academia, with some already established applications on state level. There are a few direct use cases and examples about using mobile positioning data for generating population statistics.

2.2.2. Population statistical indicators generated from mobile data

Based on home, work-time and other types of anchors, regular and irregular movements, and tourism travels, monitoring some indicators for population statistics can be very fast and reliable. Mobile data could be used as of the intermediate sources for population statistics between population censuses, for assessing internal migration, for monitoring temporary population statistics, and other purposes. Based on the population statistics, the historical analysis, real-time monitoring and prediction of the people present in specific areas can be applied in risk assessment analysis, and for generating different emergency situation scenarios. In the paper "Overview of the sources and challengers of mobile positioning data for statistics", the authors proposed population statistical indicators which can be generated from mobile positioning data[9]:

- The number of residences geographically distributed according to available accuracy;
- The number of workplace, school, secondary home, and other regular locations;
- Internal migration based on the change of the residences within the country;
- Change of workplace over time;
- Cross-border migration based on the regular travels between different countries;
- Population grid statistics (1 km2);
- Temporary population statistics;
- Assessing temporary population (hourly, daily, weekly, monthly, etc.) (Figure 1);

[8] "Dynamic population mapping using mobile phone data" (2014): http://www.pnas.org/content/111/45/15888.full.pdf
[9] "Overview of the sources and challengers of mobile positioning data for statistics": http://unstats.un.org/unsd/trade/events/2014/Beijing/Margus%20Tiru%20-20Mobile%20Positioning%20Data%20Paper.pdf

- Real-time assessment for specific location during the large-scale event, gathering of people or actual emergency situations (e.g. what is the consistence of the crowd in specific location, how many people are affected by an earth-quake of hurricane) (Lu et. al. 2012);
- Risk assessment for law enforcement (planning the number of patrol units in the area based on the consistency of the temporary population).

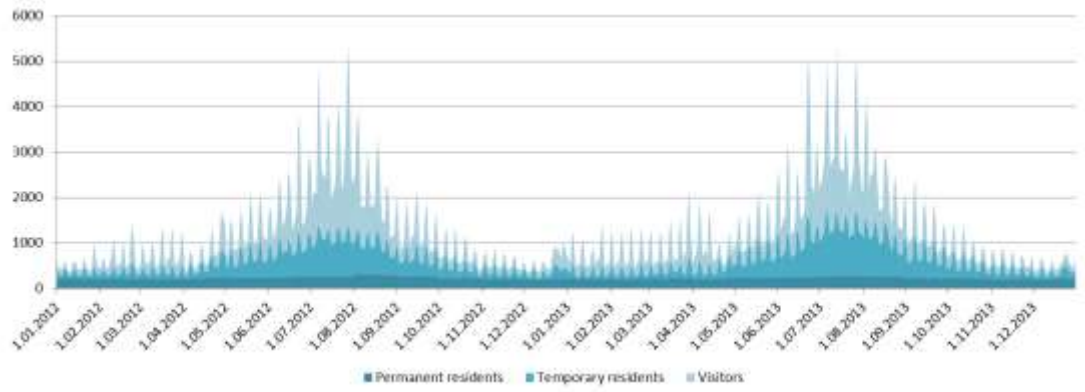


Figure 2.1 Daily temporary population statistics for a small rural municipality with many summer-hoses in Estonia. The chart represents the number of people present within the municipality during a 2 year period. The seasonality and weekly cycle can be clearly distinguished for second home residents (temporary residents) and visitors.

2.2.3. Population density and population mapping

Population density is one of the essential concepts in population statistics. During the past decade, official institutions, national and international organizations, research institutions have been exploring methodology on how to estimate population density based on mobile network datasets, and produced some methods with practice. Here, we introduce some work and studies on measuring population by using mobile phone data.

In the paper "*Estimating population density distribution from network-based mobile phone data*" (2015)[10], the research group propose a systematic methodological framework for population density estimation based on mobile network data. The study addresses the problem of leveraging network-based data (CDR and/or VLR) for the task of estimating population density distribution at pan-European level. The primary goal of the study was to develop a methodological framework for the collection and processing of network-based data that can be plausibly applied across multiple MNOs. The main challenge of this task is to design a methodology that achieves general applicability in a highly heterogeneous scenario, where several technical details of network configuration and data organization remain highly MNO-specific. To this aim, the authors pursue the design of an "resilient" methodological framework, whereas the core set of functions does not rely on any non-standard MNO-specific configuration — hence, it can be implemented by any MNO — and, at the same time, it is flexible enough to optionally leverage additional MNO-specific network and/or

[10] "Estimating population density distribution from network-based mobile phone data" (2015): https://ec.europa.eu/eurostat/cros/content/estimating-population-density-distribution-network-based-mobile-phone-data_en

data characteristics so as to improve the fidelity of the final results to the "ground truth". Owing to such flexibility, the proposed methodology lends itself to be extended and further refined, by taking advantage of the future evolutions of mobile network infrastructures (e.g., availability of additional data sources).

In the article "*Dynamic population mapping using mobile phone data*"(2014)[11], by using the datasets of more than 1 billion mobile phone call records from Portugal and France, the research team developed population mapping methods to estimate population density. The results of study explicitly show the spatial and temporal of estimated population densities at national scales, and also demonstrate seasonal changes in population distribution. This study demonstrates how the analysis of mobile phone data that are collected readily every day by phone network providers can complement traditional census outputs. Not only can population maps as accurate as census data and existing downscaling methods be constructed solely from mobile phone data, but these data offer additional benefits in terms of measuring population dynamics. Further, a combination of both the mobile phone and remote sensing methods facilitates the improvement of both spatial and temporal resolutions and demonstrates how high-resolution population datasets can be produced for anytime period.

2.2.4. Measuring urban population and inter-city mobility – a study by ISTAT, Italy

In the paper "*Use of mobile phone data to estimate mobility flows. Measuring urban population and inter-city mobility using big data in an integrated approach*" ( )[12], the team investigate the massive and constantly updated information carried by mobile phone call data (CDRs) for estimating population statistics related to residence and mobility. The study focused on the 39 municipalities in the province Pisa, Tuscany. The dataset used in this work consists of 7.8 million CDRs collected from Jan 9th to Feb 8th, 2012. The dataset contains calls corresponding to about 232,200 users with a national mobile phone contract (no roaming users are included). In this work the researchers propose and experiment an analysis process built on top of the so-called *Sociometer*, a data mining tool for classifying users by means of their calls habits. The results obtained show flow estimates with the method in the study are more accurate with larger towns.

2.2.5. Daytime population estimations – a study by Statistics Netherlands

The concept of the daytime population refers to the number of people who are present in an area during normal business hours, including workers. This is in contrast to the "resident" population, which refers to people who reside in a given area and are typically present during the evening and nighttime hours. Further details can be found in the studies below:

- "Visualization and Big Data in Official Statistics": http://www.inegi.org.mx/eventos/2014/big-data/doc/P-MartijnTennekes.pdf
- "Big data, the future of statistics": http://www.riksbank.se/Documents/Forskning/Konferenser_seminarier/2015/Big%20data%20the%20future%20of%20statistics%20Experience%20from%20Statistics%20Netherlands.pdf

[11]This work forms part of the WorldPop Project (www.worldpop.org.uk) and Flowerminder Foundation, (www.flowerminder.org)

[12] The study has been jointly developed by Istat, CNR, University of Pisa in the range of interest of the "*Commssione di studio avente il compito di orientare le scelte dellIstat sul tema dei Big Data*": http://www.cisstat.com/BigData/CIS-BigData_06_Eng%20%20IT%20Mobile%20phone%20data.pdf.

2.2.6. Dynamic population monitoring platform by Beijing Municipal Bureau of Statistics

In recent years, Beijing Municipal Bureau of Statistics of China establishes a dynamic population monitoring platform based on mobile phone data and develops methods of estimating population. The platform provides a new data source and acts important supplement for traditional population statistics. By using mobile data, Beijing Municipal Bureau of Statistics of China obtains the monthly population data and implements the data coherence between municipal and districts. Meanwhile, the platform measures dynamic population flow, meets the needs of government management. And furthermore, the platform is economical and timely.

2.3. Migration statistics

With the rapid urbanization, the numbers of people migrating have been greater than before[13]. In some developing and undeveloped counties/regions, some natural and economic factors or livelihood patterns may also cause the large-scale migrations.

Till now, the common data of migration is derived from Population Censuses, register, household surveys or other administrative collection. With the popularity of mobile phones around the world, mobile positioning data is becoming a novel data source to describe human migration, particularly in developing and un-developed counties/regions which lack of accurate and timely population data. In view of the characters of real-time, dynamic, positioning, the mobile phone data is a strong supplement for traditional population data in terms of urban management, emergency events, epidemic prevention and control, disasters disposal, etc. Here, we introduce several cases which using mobile data to observe, measure and analyze human migration in developing countries.

2.3.1. Climate-induced migration: a case study in Bangladesh

In the paper[14] "*Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in Bangladesh*", the researchers establish a methodological framework in studies of human migration and climate change by using mobile phone data. Climate change is likely to drive migration from environmentally stressed areas. However quantifying short and long-term movements across large areas is challenging due to difficulties in the collection of highly spatially and temporally resolved human mobility data. In this study, the team uses two datasets of individual mobility trajectories from six million de-identified mobile phone users in Bangladesh over three months and two years respectively. Using CDR data collected during Cyclone Mahasen, which struck Bangladesh in May 2013, the study shows that analyses based on mobile network data can describe important short-term features (hours–weeks) of human mobility during and after extreme weather events, which are extremely hard to

[13] The migration as used in this draft refers to internal migration.

[14] "Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in Bangladesh" (2016): http://www.flowminder.org/publications/unveiling-hidden-migration-and-mobility-patterns-in-climate-stressed-regions-a-longitudinal-study-of-six-million-anonymous-mobile-phone-users-in-bangladesh

quantify using standard survey based research. The study also demonstrates how to analyze the relationship between fundamental parameters of migration patterns on a national scale based on the mobile phone data. The study concurrently quantifies incidence, direction, duration and seasonality of migration episodes in Bangladesh.

2.3.2. Measuring migration in developing countries: evidence from Rwanda

In the paper[15] "*Inferring Patterns of Internal Migration from Mobile Phone Call Records: Evidence from Rwanda*", the authors describe mobile phones can provide a new source of data on internal migration. Using Rwanda as a case study, the researchers demonstrate the methods of using mobile data in practice. The study develops and formalizes the concept of inferred mobility, and compute this and other metrics on a large dataset containing the phone records of 1.5 million Rwandans over four years. The analysis reveals more subtle patterns that were not detected in the government survey. Namely, the authors observe high levels of temporary and circular migration, and note significant heterogeneity in mobility within the Rwandan population. The study provides a new quantitative perspective on certain patterns of internal migration in Rwanda that are unobservable using standard survey techniques, and explores methods on new forms of information and communication technology can be used to better understand the behavior of migration in developing countries.

2.4. Commuting statistics

With rapid urbanization, nowadays the scale of commuting is growing, and commuters are travelling long times and distances to get to work, especially in metropolis. The common statistical methods of measuring commuting are Census and surveys. The traditional dataset could provide amount information on travel to work flow and the characteristics of workers based on where they live (origin) and where they work (destination). However, obtaining detailed commuting data within cities has traditionally been proved challenging. This is typically achieved using infrequent and expensive large-scale Census and surveys. In face of such difficulty of the lack of large-scale surveys or government statistics (particularly in developing countries), researchers have proposed using location information from cell phone data to estimate users' home and work locations, and commuting flows. This approach can complement traditional collection techniques, which are often outdated by the time they're available to policy makers and the general public.

2.4.1. A pilot study of Estonia

In the article[16] "*Application of mobile phone location data in mapping of commuting patterns and functional regionalization: a pilot study of Estonia*", the researchers present initial steps into the research of commuting patterns and functional regions using mobile phone location data. The main aim is to introduce and discuss the potential of mobile phone location data as an alternative data sources to censuses for mapping commuting flows and subsequent

[15] "Inferring Patterns of Internal Migration from Mobile Phone Call Records: Evidence from Rwanda" (2011): http://www.un.org/en/development/desa/population/publications/pdf/technical/TP2013-1.pdf

[16] "Application of mobile phone location data in mapping of commuting patterns and functional regionalization: a pilot study of Estonia": http://www.tandfonline.com/doi/pdf/10.1080/17445647.2012.762331?needAccess=true

functional regionalization. A set of analytical maps covering various aspects of regular daily movements of population and functional regionalization is provided. Estonia is serving as a pilot laboratory for analyses based on commuting flows derived from mobile phone location data. The maps give to reader a synthetic overview of contemporary settlement system in Estonia and introduce the potential of mobile phone location data for research in this field.

Through analysing, the authors draw the conclusions that mobile phone location data represent a significant alternative source in cases where traditional data are missing or are not appropriate or the method of collection is not cost effective. Mobile phone location data can provide commuting information for majority of population in almost freely decided point of time and frequency of repetition, in greater spatial detail and much cheaper, than traditional means of data gathering. Moreover, the time gap between the moment of survey and the moment of availability of data could be shorten to almost real-time mode.

2.4.2. Urban Commuting and Economic Activity

In another paper[17] "*Estimation of urban commuting patterns using cellphone network data*", the researchers explore to analyze the relationship of human mobility and economic activity in the urban by using mobile phone data. The article indicates that understanding this relationship is particularly relevant in large cities in developing countries, which are dynamic, connected and dense, yet poorly covered by conventional data sources. To make progress on this issue, the study develops a data set of commuting flows extracted from cell phone transaction data from Sri Lanka, and analyzes this data using a simple urban economics model. The model maps the commuting flows to the geographic distribution of key economic indicators such as economic output and residential income. The study shows that commuting flows, as well as the geographic distribution of economic output and income estimated using the model, capture the main features of Colombo's urban structure. The study also shows that the constructed income measure has strong predictive power for nighttime lights, an established proxy of residential income.

2.5. Traffic flow statistics

2.5.1. Mobile phones for traffic flow measurement – an Estonia case study

Järv O, Ahas R, Saluveer E, Derudder B, Witlox F published a study on analyzing mobile call detail records during rush hour.[18] A number of various intelligent transportation systems (ITS) had been developed in order to understand the causes of traffic congestion and arrive at possible solutions. While it had been proven to be effective in relieving worsening traffic conditions, it was also believed that methods could further be improved by leveraging on

[17] LIRNEasia, "Estimation of urban commuting patterns using cellphone network data"：
https://pdfs.semanticscholar.org/4af0/36dcbd2f0a26f01c0e1b90fede1c94821782.pdf

[18] Järv O, Ahas R, Saluveer E, Derudder B, Witlox F (2012) Mobile Phones in a Traffic Flow: A Geographical Perspective to Evening Rush Hour Traffic Analysis Using Call Detail Records, PLOS Research Atricle, November 14, 2012. , https://doi.org/10.1371/journal.pone.0049171

other available applications of information and communications technology (ICT). One such application that was proposed and applied was the use of mobile phone positioning data as an encouraging movement data collecting method. The use of this data had enabled the provision of larger amounts of data over longer study periods, wider study coverage with unlimited sample size potential and could also be processed into real-time movement information. Nonetheless, it was also recognized that using this type of data had some challenges in the areas of privacy, data access and sampling related to phone ownership and usage.

The use of mobile phone data had also proven to be more cost-efficient as there was already a strong presence of mobile phones in daily life and it was able to provide more in-depth information on individuals' mobility patterns. Such information had proven to be complimentary to traditional approaches in transport statistics. The authors of this study pointed out that in one previous study[19] on traffic flows, a comparison of results based on the extracted mobile phone data and on-road survey was in sync and therefore encouraged them to apply mobile phone data in transportation studies.

The paper provided new insights regarding the composition of traffic flows and explained by what means and to what degree suburbanites' travelling had affected rush hour traffic. It presented an alternative methodological approach using data extracted from mobile phones, particularly call detail records, in their assessment of traffic flow composition during evening rush hours in Tallinn, Estonia. They discovered that daily commuting and suburbanites had impacted transportation demand through increased evening rush hour traffic, albeit daily commuting comprised only 31% of the movement during that period. The paper also pointed out that the evening rush hour traffic on Fridays were different from other days, and assumed that it was related to domestic tourism and leisure time activities, which suggested that the general movement of individuals attributed to changes in social behavior played a greater role in evening rush hour traffic conditions as compared to the impact of suburbanization.

The authors believed that the method would enable the possibility to study geographical distribution and temporal variability of trips in a traffic flow. When such data and methodology were applied, it would be easy to detect road users in a given area and certain timeframe, and link their origin and destination with their trip indication. Information from traditional methods would be complicated and expensive. However, at the time, the authors recognized that there were two limitations to this approach. The first was the inability to know the actual purpose of the movement or to reveal the modal split for road users. The second was the issues related to privacy in the use of such cellphone data. In essence, this study presented an approach that was cost-effective in providing supplementary information for traffic studies, particularly in terms of the composition of road users in spatial and temporal contexts. Further, this method could be adopted in real-time traffic monitoring tools for ITS in the future.

[19] Saluveer E, Järv O (2008) Analyzing Traffic Flow and Geographical Distribution of Origin-Destination on Highway Using Passive Mobile Positioning Data. Presentation in the International Workshop SPM2008 in Tartu, Estonia.

The authors' next steps for this research would be the real-time applicability along with the comparison of the traffic composition between different road sections as promising avenues for further research of implementing mobile phone data.

2.5.2. Mobile Phone Data for Real-Time Road Traffic Monitoring

A. Janecek, D. Valerio, K. A. Hummel, F. Ricciato and H. Hlavacs proposed in their study[20] a fresh approach to monitoring real-time traffic based on the signaling data between mobile devices and a mobile cellular network. The researchers had estimated travel times in road sections by mapping the sequence of encrypted signaling messages (to preserve anonymity) for each mobile device to the physical movement along the road. The advantages of this approach were that it was based on data available 24/7 and that it did not require costly road sensor installation.

Even if idle devices contributed a large volume of spatially coarse-grained mobility data, the researchers believed that active devices provided finer-grained spatial accuracy for a limited subset of devices. They recognized that the combined use of data from idle and active devices improved detection performance in terms of coverage, accuracy and timeliness. The difference of this study as compared to other studies on call data records (CDR) data was that this approach was not limited to observe the small segment where mobile devices were actively engaged in voice calls or data connections. On this note, this approach would achieve a wider coverage and more accurate estimation; yet, it would require advanced methods to handle a more heterogeneous set of signaling data.

The researchers had validated their method against different conventional data sources - namely road sensor data, toll data, taxi floating car data, and radio broadcast messages. They had considered one full month of data and had focused on a sample highway of 36km that spanned urban, semi-urban, and non-urban areas. With optimal parameter tuning, their method had identified all road congestion episodes without any false positive. On average, their approach had given results 3 minutes faster than the traditional monitoring approach. In addition, their approach had also provided a 25% improvement over the smallest average road segment length that was observable by other traditional and existing road monitoring systems. Finally, estimates for travel time that had been delivered through this method could be manually inspected to predict signals for a possible classification of congestion episodes.

2.5.3. Mobile phone data to measure traffic variability caused by holidays

Gang Liu, Chenhao Wand and Tony Z. Qiu study[21] used anonymous cell phone data to evaluate the change in traffic patterns caused by holiday traffic, and discussed how traffic

[20] A. Janecek, D. Valerio, K. A. Hummel, F. Ricciato and H. Hlavacs, "The Cellular Network as a Sensor: From Mobile Phone Data to Real-Time Road Traffic Monitoring," in IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 5, pp. 2551-2572, Oct. 2015. doi: 10.1109/TITS.2015.2413215
https://www.researchgate.net/publication/275273390_The_Cellular_Network_as_a_Sensor_From_Mobile_Phone_Data_to_Real-Time_Road_Traffic_Monitoring

[21] Gang Liu, Chenhao Wand and Tony Z. Qiu: Understanding Intercity Freeway Traffic Variability at Holidays Using Anonymous Cell Phone Data, ResearchGate, Oct 2016,
https://www.researchgate.net/publication/304620561_Understanding_Intercity_Freeway_Traffic_Variability_at_Holidays_Using_Anonymous_Cell_Phone_Data

patterns vary each day during holiday periods (before the holiday, during the holiday, and after the holiday). It was anticipated that the findings of this study would help transportation engineers and program managers implement appropriate congestion-related countermeasures for mitigating heavy congestion on a subject roadway during the busiest holiday periods. Drivers could also choose to avoid the congestion and change their holiday travel schedules based on the information about holiday traffic. The study and analysis was based on cell phone data collected from the stretch in China's freeway from Hangzhou to Quzhou over 12 months. The traffic pattern during the Labor Day holiday in 2014 was investigated using the cell phone sample and speed data. Results showed that holiday affected daily and hourly traffic volume and speed, and that the effects were different for southbound and northbound traffic in the study area. The researchers believed that good understanding of temporal and spatial traffic patterns due to holiday effects could assist in developing appropriate countermeasures for congestion mitigation.

2.5.4. Mobile phone data in transportation and urban planning – a case study in Sri Lanka

The authors, Lokanathan S., Kreindler G.E., de Silva N.H.N, Miyauchi Y. Dhananjaya D. and Samarajiva acknowledged that an increasing problem in countries experiencing growth was road congestion in their article[22]. Data was needed to identify the bottlenecks and to prioritize improvements. While data-centric approaches to transportation management based on sensor data was already a reality in many developed countries, many developing countries still relied on traditional data sources such as questionnaires. Those survey-based methods were very costly and intrusive. Other conventional methods such as traffic recorders were less intrusive but did not yield actual route information. The authors believed that mobile network big data (MNBD) had enormous potential for traffic planning as data streams were continuously flowing and therefore the effects on changes in traffic routes (one-way schemes, new roads) had the potential to be tracked easier. Even though there was the potential of additional costs related to data storage, BTS or base transceiver station hand-off data could serve as speed and traffic disruption sensors. In this study, the authors wanted to understand whether mobile network big data could assist in transportation planning in the city of Colombo, Sri Lanka through the creation of origin-destination matrices that described the flow of traffic between different areas and determined where the daily commuting population of Colombo came from. The dataset included four months of encrypted (to protect anonymity) passive positioning CDR data of voice calls from 5 to 10 million SIMs from the Sri Lankan mobile operator.[23] Each CDR contained the following information: (1) call direction - a code to denote if the record was an incoming or outgoing call; (2) subscriber identifier: pseudonymised identifier for the subscriber in question; (3) identifier of the other

---

[22] Lokanathan, S., Kreindler, G. E., de Silva, N. H. N., Miyauchi, Y., Dhananjaya, D., & Samarajiva, R. (2016). The potential of mobile network big data as a tool in Colombo's transportation and urban planning. Information Technologies & International Development [Special Issue], 12(2), 63–73. https://www.researchgate.net/publication/269465996_Using_Mobile_Network_Big_Data_for_Informing_Transportation_and_Urban_Planning_in_Colombo

[23] Agreements with the Sri Lankan operator required that they do not disclose the actual mobile operator and actual number of SIMs that were analyzed.

party: pseudonymised identifier for the other party to the call; (4) cell ID: identification of the antenna that the subscriber was connected to at the time of the call; (5) date and time that the call was initiated; and (6) duration of the call. A unique home and work location was assigned to each SIM in the dataset at the level of the divisional secretarial division (DSD), which was the third sub-national administrative level (after provinces and districts) in Sri Lanka. Based on the home and work assignments, it had been possible to find regular inter-DSD mobility patterns, which formed the basis of the study.

The researchers had devised a methodology to find the "home" and "work" location for each sim by finding the time bracket within which an average individual would have spent at either location. They had also assumed that there would be two peaks of daily human movement – the first was in the morning when people were commuting to work and the second was in the evening when people were commuting from work to their home. On that note, they had assumed that the time that fell outside the "home" and "work" brackets were the individuals' commuting time between locations and the time between the two peaks would have been the "work" hours. With that assumption, the researchers had followed a three-step process in creating a mobility graph from the dataset by (1) calculating the average position (latitude, longitude) for each individual SIM for each hour of the day (24 points in all), (2) obtaining the distance travelled by each individual SIM during each hour bracket through the calculation of the Euclidian (i.e. straight-line) distance between two consecutive hour-wise average locations, and (3) obtaining the average hour-wise distance measure for all the SIMs that had showed movement during a specific hour. The results from these datasets were plotted on a graph.

Their next step was to find home and work DSDs. They had used the logic behind the "tower days" concept (as was proposed by Isaacman, Becker, and Cáceres (2011))[24] to generate a "DSD count" concept. Accordingly, the researchers followed a three-step process as follows: (1) A DSD was assigned to each individual SIM based on the DSD that was most used by each individual SIM, for each home or work time slot; (2) For each SIM all potential DSDs were obtained for both the home and the work time slots over a four-month period, and were listed along with its associated frequency; (3) An overall home location and work location were assigned for each SIM based on the DSD that had occurred most in the frequency table generated in the previous step.

The researchers found that most people were living and working within the same DSD. In Colombo city, which was comprised of the Colombo and Thimbirigasyaya DSDs, the researchers had also examined the extent and relative rank of each home DSDs contribution to Colombo's working. Their analysis had found that nearly 47% of Colombo city's working population came from outside the city, as just over 53% of the SIMs had both a home DSD count and a work DSD count in Colombo city.

[24] Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., & Varshavsky, A. (2011). Identifying Important Places in People's Lives from Cellular Network Data. In K. Lyons, J. Hightower, & E. M. Huang (Eds.), Pervasive Computing: Lecture Notes in Computer Science (Vol. 6696, pp. 133–151). Berlin, Heidelberg: Springer Berlin Heidelberg.

The findings of their research suggested that in developing countries having limited sensor networks such as Sri Lanka, active and passive positioning data from mobile network operators could provide relevant insights that could improve the efficiency and reliability of transportation system. Data extracted from mobile phones could detect human behavioral patterns relating to mobility, and therefore showed promise as a source of timely and cost-effective data for transportation planning. The researchers also believed that insights from mobile phone data was not exhaustive, and thus would not remove the need for surveys. However, any incremental increase in the type and frequency of insights that was possible, especially in resource-constrained developing economies could potentially facilitate a more effective planning process.

2.5.5. Mobile phone data for traffic and urban spatial pattern analysis – a Dutch case study

Analyzing and forecasting road traffic are two significant areas where mobile network data can be used by Steenbruggen J, Borzacchiello M.T., Nijkamp P and others in their article[25]. Growing traffic volumes have led to traffic congestion, mobility issues, especially during peak travel hours, in both local road and highway networks.

Traditional traffic data collection methods are effective and precise, but also have practical and financial limitations. The increasing need for a cost-efficient and reliable information system led to the growing interest in the use of data derived from cellular networks to support traffic parameter estimations without the need to install complex and expensive measurement systems.

This paper provided a broad overview of the main studies and projects which addressed the use of data derived from mobile phone networks to obtain location and traffic estimation of individuals, as a starting point for further research on incident and traffic management. They also presented the findings of the Current City project, which was a test system in Amsterdam for the extraction of mobile phone data and for the analysis of spatial network activity patterns.

This paper is the first step of a study whose aim is to further investigate data deriving from the project Current City Amsterdam; the next phases of the research will include the development of a validation methodology of this data using loop detector data as ground truth, and the study of new applications in the field of traffic management. Especially for contingency management (e.g. traffic accidents, network disturbances caused by terror attacks or nature catastrophes) the use of cellular phone data may be of strategic importance in the future.

[25] Steenbruggen, J., Borzacchiello, M.T., Nijkamp, P. et al."Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities, GeoJournal (2013), Issue 2. April 2013. https://link.springer.com/article/10.1007/s10708-011-9413-y

2.6. Employment statistics on border and seasonal workers

2.6.1. Tracking employment shocks using mobile phone data

Toole J.L., Lin Y-R, Muehlegger E, Shoag D, Gonza'lez M.C. and Lazer D. study[26] demonstrated how mobile phone data could be utilized to quickly and accurately detect, track and predict economic changes at multiple levels. It emphasized the potential of mobile phone data as a reliable source of information in improving forecasts in critical economic indicators relevant to policymakers in the private and public sector.

Call detail records may facilitate the identification of macroeconomic statistics quicker and in more detail than traditional methods of tracking the economy. Administrative databases and surveys could be slow to collect, costly to administer and could fail to capture significant segments of the economy. Surveys quickly face sample size limitations and would require strong assumptions about the consistency of responses over time. Data from mobile phones had already been proven extremely helpful in understanding the dynamics of society. With a fundamental understanding of regular behavior, it would be easier to detect deviations caused by collective events. In this regard, linking this data to economic behavior might provide a methodology to infer changes to measure employment shocks at extremely spatial and temporal resolutions and improve critical economic indicators.

In this study, researchers show how mobile phone data can provide a quick insight into employment levels, precisely because people's communication patterns change when they are not working. The researchers harnessed the power of algorithms to analyze and to study data from two undisclosed European countries.

In the first country, they present how it was possible to observe mass layoffs and identify the affected demographic through mobile phone records. They used call data records spanning a 15-month period and designed a structural break model to identify mobile phone users who had been laid off. Then they tracked the mobility and social interactions of the affected workers, looking at several quantities related to their social behavior, including total calls, number of incoming calls, number of outgoing calls, and calls made to individuals physically located at the plant. The findings revealed that job loss had a 'systematic dampening effect' on their mobility and social behavior. For example, the researchers found that the total number of calls made by laid-off individuals dropped 51 percent following their layoff when compared with non-laid-off residents while their number of outgoing calls decreased 54 percent. What's more, the month-to-month churn of a laid-off person's social network -- that is, the fraction of contacts called in the previous month that were not called in the current month -- increased approximately 3.6 percentage points relative to control groups. In terms of

[26] Toole JL, Lin Y-R, Muehlegger E, Shoag D, Gonza´lez MC, Lazer D. 2015 Tracking employment shocks using mobile phone data. J. R. Soc. Interface 12, The Royal Society Publishing: May 8, 2017. http://rsif.royalsocietypublishing.org/content/royinterface/12/107/20150185.full.pdf

mobility, they found that the number of unique mobile phone towers visited by people who had lost their jobs decreased 17 percent relative to a random sample.

In the second country, one that had experienced macroeconomic disruptions during the period in which the data was available, the researchers analyzed the call detail records of thousands of subscriber and looked for behavioral changes that may have been caused by layoffs -- fewer outgoing calls, for example, or an increase in churn -- to determine whether those changes could predict general unemployment statistics. They did find that changes in mobility and social behavior predicted unemployment rates before the release of official reports and more accurately than traditional forecasts. Specifically, the researchers noted that their novel methods allowed them to predict present unemployment rates two to eight weeks prior to the release of traditional estimates and forecast future employment rates up to four months ahead of official reports.

The findings were believed to have great practical importance as it potentially facilitates the identification of macroeconomic statistics with much finer spatial granularity and faster than traditional methods of tracking the economy. Their algorithms also predicted future states and correct for current uncertainties. However, the researchers noted and cautioned against viewing their methods as a substitute for survey-based approaches to detecting future unemployment rates. They believed that mobile phone data was a powerful and complementary tool – fast and inexpensive – but the norms against phone use were constantly changing and would force the researchers to adjust and calibrate their methods.

2.7. Other applications or areas

"A brilliant example of how the application of big data analysis to mobile technologies can be used to accelerate emergency aid, and provide intelligence to help prepare for future disasters." – **Global Mobile Industry Recognition (Feb 2016)**

A number of research papers and cases show that mobile phone data is playing very important role on areas of health, socio-economics, disaster response, urban management, etc. is listed below without further details:

- Mobile phone network data for ***development***", Global Pulse, 2013: http://www.unglobalpulse.org/sites/default/files/Mobile%20Data%20for%20Development%20Primer_Oct2013.pdf
- "State of mobile data for ***social good***", Global Pulse, GSMA, 2015: http://www.gsma.com/mobilefordevelopment/wp-content/uploads/2017/02/State-of-Mobile-Data-for-Social-Good-Preview-Feb-2017.pdf
- "On the use of human mobility proxies for ***modeling epidemics***", 2014: http://perso.uclouvain.be/adeline.decuyper/docs/journal.pcbi.1003716.pdf
- "Migration statistics relevant for ***malaria transmission*** in Senegal derived from mobile phone data and used in an agent-based migration model" (2016): http://www.geospatialhealth.net/index.php/gh/article/view/408/357
- "Mobile phone data highlights the role of mass gatherings in the ***spreading of cholera outbreaks***" (2015): http://www.pnas.org/content/113/23/6421.full.pdf

- "Predicting ***poverty and wealth*** from mobile phone metadata": http://www.uvm.edu/~cdanfort/csc-reading-group/blumenstock-science-2015.pdf
- "Analysing ***seasonal mobility*** patterns using mobile phone data", Global Pulse project series, no.15, 2015: http://www.unglobalpulse.org/sites/default/files/UNGP_ProjectSeries_Mobility_Senegal_2015_0.pdf
- "Mobile data access for ***public benefit/SDGs***: Cases and caveats", Flowminder: http://www.oecd.org/std/Flowminder-OECD-2015-Dec.pdf
- "Rapid and Near Real-Time Assessments of Population Displacement Using Mobile Phone Data Following Disasters: The 2015 ***Nepal Earthquake***", 2016: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4779046/

3. Data sources

3.1. Data from MNO's systems

Wireless technologies (such as 2G, 3G and 4G) that are used in mobile telephones and telecom networks worldwide follow different standards and specifications composed and maintained by organizations such as ETSI, 3GPP, Qualcomm and IEEE. In Europe (and also Japan and China), the most prevalent mobile communication technologies today are GSM (2G) and UMTS (3G), which follow the 3rd Generation Partnership Project's (3GPP[27]) technical specifications. For this reason, most of the descriptions of mobile network operators' (MNOs) network architecture and data access from the MNOs' systems in this report will be based on 3GPP specifications.

Data extraction from the MNO's systems depends on the specific technical solutions the MNO is using. In general, every MNO has a billing centre and, most likely, a data warehouse where data is periodically stored for billing and further analyses for the purposes of network planning and management. Besides central storage systems where data is readily available, it is possible to extract data in the process of probing on the BSS and NSS levels.

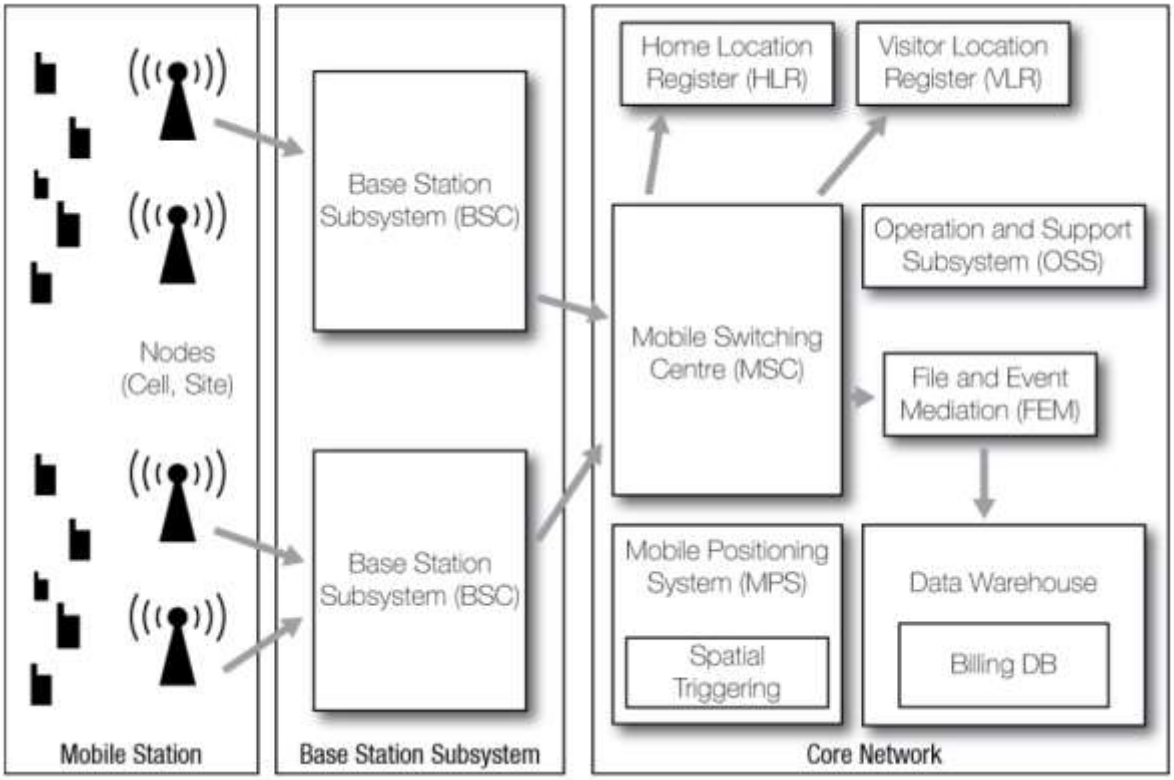


Figure 3.1. Sources of mobile positioning data in the operators' system. Source: Tiru & Ahas

[27] http://www.3gpp.org/about-3gpp/about-3gpp

3.1.1. Central storage systems

Data extraction from central storage systems is the easiest way to obtain data for national-level statistics. There are several places where such data can be found, collected for different purposes and containing different kinds of information:

- The billing domain (BD) stores CDR and IPDR for charging purposes. Data will be stored here after a successful charging record generation procedure.
- Customer databases contain information about users, which can add extra value to CDR or IPDR data – e.g., socio-demographics and place of residence.
- Data warehouses host collections of data from different sources but might not always be easily accessible due to the huge amount of data stored there.

Databases in the billing domain are generally considered most easily accessible. Data stored in the BD is gathered by a mediation system from different network entities responsible for providing various types of services.

3.1.2. Probing and signaling data

There are two types of probing – active and passive – for monitoring and acquiring data from telecommunications networks. Active probing is used by the network operator to generate traffic to monitor the overall network performance, whereas passive probes monitor data flows between different network entities. Because of this, passive probing is one possibility to gather information from MNOs' networks.

Table 3.1. Comparison of different data sources.

Data source	**Accessibility**	**Major features**	**Adequacy for statistics**
Cell activity	Easy, standard outlet from the operator's system	Phone use intensity at the cell level	Low
Call detail records (CDR)	Privacy problem, software development needed	Call activities	High
Probes	Privacy problem, major software development needed	Call activities and handover logs; personal features from the operator	High
Mobile Positioning System	Easy for small samples, usually requires opt-in	Accurate positioning, custom frequencies; questionnaire with the respondent	High, can be used for benchmarking

		possible	

Source: Tiru & Ahas, 2012

Capabilities of passive probing are usually achieved by deploying licensed software together with required hardware to the network. There are several possible locations for such probes, most commonly between the BTS/Node-B/eNode-B and BSC/RNC/MME or BSC/RNC/MME and MSC. These probes can also query databases, such as the VLR, that store relevant user data.

Pros:

- Gives access to data types that usually won't be stored in billing centres and data warehouse systems
- Increasing the number of records per subscriber helps to minimize diurnal and daily differences in record counts that are the result of subscribers' calling patterns

Cons:

- Expects monitoring or data acquisition systems to be "online" in order to constantly gather and store the acquired data that is temporary in nature
- Additional network load for MNOs
- Not all information is mandatory for an MNO to store during probing (e.g. LAI is mandatory, but the more precise CGI is not)
- Increases the number of events per subscriber considerably, meaning that more resources need to be dedicated for the storage and processing of such data
- Involving system-generated data in statistical analyses might result in high levels of background noise that needs to be addressed in the data processing phase

The cost of installing the probing systems in the MNOs is usually high and it is not implemented simply for generating statistical indicators for the NSIs. But if the MNOs have implemented some sort of a probing system, it is a good source for data for statistical indicators as it involves much more data compared to "traditionally" collected CDR from a billing system of data warehouse. Probing data is limited to inbound roaming and domestic data, excluding outbound roaming data.

Signalling data is generally referred to obtaining transmission signals from the radio network directly and storing it to database. Similarly to probing data, signalling data is also very voluminous, and can be limited to inbound roaming and domestic data, excluding outbound roaming data.

3.1.3. Active positioning data

Active positioning data is generated by positioning the mobile subscriber through device- or network-centric methods, as well as via satellite (i.e. GPS). These methods are being used either for providing location-based services or in response to national regulations requiring the capture of highly accurate location data (e.g. for rescue and police purposes). Active

positioning capabilities are common in developing economy operators' systems and are often used merely on a case-by-case basis. Common methods of active positioning include cellular triangulation or assisted GPS.

Researchers have used active mobile positioning with small samples for benchmarking and in studies that require constant accurate positioning.

	Data source	How easy?	What?	How much data?	How adequate for statistics
PASSIVE	**Cell activity**	Easy, standard	Phone use intensity at the cell level	↓	↘
	Call detail records (CDR + DDR)	Needs privacy protection and processing methods	Call activities	→	↗
	Probes / Signalling	Needs collection, privacy protection and processing methods	Call activities and handover logs; personal features from the operator	↗	↑
ACTIVE	**Mobile Positioning System**	Easy for small samples, usually requires opt-in	Accurate positioning, custom frequencies; questionnaire with the respondent possible	↑	↗

3.2. Mobile phone event data – Passive positioning data

A mobile phone event is seen as action that is initiated by or targeted to a subscriber or mobile device (MD). Most of these events generate passive positioning data. When these events occur, they are registered by different network entities. For example, when a mobile device initiates a location area update, the MSC/VLR store new LAC for the mobile device. Similarly, CDR and IPDR are generated every time a subscriber uses services such as calling (in/out), messaging (in/out) or accessing the Internet. Event data can also originate from probing or signalling sources which include CDRs and IPDRs, and also other technical events like location area update or other operational activities that generate event along with cell tower reference. For every such event, an MNO gathers data that is relevant for the event, meaning that the list of attributes is often event-specific. Moreover, only an MNO can record some of the attributes, making their availability MNO-specific as well.

Minimal attributes needed for population statistics are subscriber identifier, time attribute and location. Any additional attributes available might help to increase the value and usability of mobile positioning data and are therefore desirable, if it is possible to collect them.

3.2.1. Forms of the mobile data

There are three forms of mobile phone data which might be stored and provided separately:

1. Domestic data – any location events occurring within the network of the specific MNO. These are CDR's or other events where home subscriber (a customer) of the

specific MNO is involved. A local subscriber calling another local subscriber in the same MNO network will generate at least two domestic events (one call initiation, one call receiving).

2. Outbound roaming data – any location event by a local MNO subscriber conducted in another network (usually a foreign MNO roaming service). These data usually represent the local subscribers using mobile phones while travelling in foreign countries.
3. Inbound roaming data – any location event by a foreign MNO subscriber. These data usually represent foreign subscribers using a local roaming service. This data can also include domestic subscribers from another MNO using a roaming service because there is no reception by their own MNO.

3.2.2. Subscriber-related identities

The most relevant identities to use for statistics are subscriber related. These identities help to distinguish individual subscribers from one another and are commonly long-lived compared to equipment related identities (people tend to change their mobile phones fairly often compared to their number or SIM card). The most important characteristics to consider in choosing identifiers for population statistics purposes are:

a. the identifier should be unique in the time span of the analysis;
b. the identifier's life span should ideally be as long as the analysis period chosen;
c. every person should ideally have only one identifier (in reality people may carry several devices with them – one person may be referred to as several different subscribers).

There are a number of different subscriber related identities generated by and forwarded to different network entities. Some of the identities are used system-wide and some have only local importance for a limited time span. Each of those identities has different limitations regarding their usage for statistical purposes. Usually IMSI (International Mobile Subscriber Identity), MSISDN (Mobile Subscriber Integrated Services Digital Network number) or MNO-specific customer/subscriber ID serve as permanent keys to identify a user from the databases. As long as a subscriber has not changed his SIM card, the IMSI will remain the same and as long as subscriber has not changed his telephone number, the MSISDN will remain the same.

3.2.3. Equipment related identities

Every mobile device is uniquely identifiable with either its IMEI (International Mobile Station Equipment Identity) or IMEISV (International Mobile Station Equipment Identity and Software Version number). These identities are not permanently related to a subscriber, since devices retain their IMEI and IMEISV even after transmission to another user. By 3GPP standards the IMEI and IMEISV are considered temporary information that might be stored in the HLR, SGSN or VLR.

3.2.4. Time attributes

While generating CDR/IPDR for billing, event time is another mandatory field that needs to be included pursuant to the 3GPP standard. Every temporal attribute should at least include the date, hours, minutes and seconds. The 3GPP standard defines three different types of call handling timestamps.

- Seizure time is the time when resources are seized to provide service to the subscriber. This field is mandatory only for calls that were unsuccessful.
- Answer time is the time when a call is answered – the connection was successful. Answer time is a mandatory field for successful calls.
- Release time is the time when seized resources are released again. Release time is an optional field.

Time attributes are generated by entities that are responsible for providing specific services. They are typically not stored in entities such as the HLR or VLR but are rather accessed during the CDR generation process by using specific functions triggered when a chargeable event occurs.

3.2.5. Location-related attributes

In order to add a spatial aspect to statistical analyses that use mobile positioning data, it is crucial to be able to geographically reference events that occur in such data. By 3GPP standards location related identities are typically mandatory (or to be included when certain conditions are met) during the process of charging information generation (ETSI TS 132 250). Roaming is one of the cases where adding a location related identity is not mandatory by standard. For this reason, outbound data might not include location related identities other than country code.

For domestic and inbound roaming data, the geographical reference is the location and/or the coverage area of the network cell (initially the ID of the cell). For outbound roaming data, the initial geographical reference is the country of the roaming partner MNO. In some cases, the outbound data also includes the network Cell ID of the outbound roaming event.

Box: Important concepts: Network cell, cell sites, location area

For cellular systems, the smallest structural element is a network cell (Figure 3.2). Every cell can be described by a number of attributes such as azimuth, sector angle, shape and size of the coverage area, type of antenna and location. A mobile telecommunications network can contain thousands of individual cells. Since every cell can serve only a limited number of users, they are unevenly distributed over the MNO's service area – more densely in urban areas, while rural areas are sparsely populated by cells.

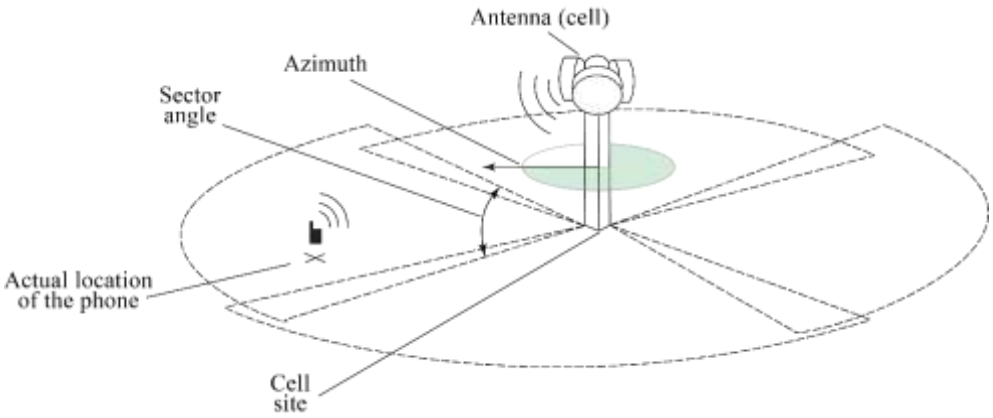


Figure 3.2 Cell, cell site and location area, and their relations to one another

A cell tower or mast is a collection of transmitters at one specific location that can be referred to as a cell site (Figure 3.2). This means that one location can host several different cells when we group them by point coordinates.

Adjacent cells, typically 30 or 40, are grouped together into one location area (LA[28]). This is done to balance signalling load so a device needs to send information about a user's location (location area updates) only when movement into another LA occurs (Sauter, 2010) and periodically with predefined time intervals.

- Location and routing area identity

Location area is a definition that refers to a circuit switched (CS) network. In this network, a group of cells form a location area (LA) or tracking area (TA) for LTE networks. To be able to distinguish between several LAs/TAs, location area identity (LAI)/tracking area identity (TAI) will be assigned. The LAI/TAI consists of three elements where the first two parts of the code are always the same for the MNO.

- The Mobile Country Code (MCC) is a three-digit identification of the country where the LA is located. The MCC is defined by the ITU-T Recommendation E.212.
- The Mobile Network Code (MNC) is a two- or three-digit identification code for the MNO pursuant to the ITU-T Recommendation E.212.
- The Location Area Code (LAC)/Tracking Area Code (TAC) is an identifier of the location area within an MNO's network. This part of the code can be represented using hexadecimal values with a length of two octets.

[28] In LTE networks, term tracking area (TA) is used instead of location area (LA).

For example, the LAI/TAI for Telia in Estonia could be 248010001, where:

- 248 is MCC for Estonia;
- 01 is MNC for Telia;
- 0001 is LAC/TAC.

Considering the data format, the LAI/TAI might be added as one field (MCC+MNC+LAC/TAC) or as three fields (MCC; MNC; LAC/TAC). If the first two fields are missing but the data source is one specific MNO, the MCC and MNC can be derived from the ITU-T Recommendation E.212. Since the LAC code is unique only inside one specific MNO location area pool, datasets that have multiple MNOs' data (such as outbound), should contain the MCC and MNC (or equivalent attributes that would enable to distinguish operators and countries).

Routing area (RA) is the counterpart of LA in packet-switched (PA) networks. The RA is usually a smaller area compared to the LA because using multimedia services requires more frequent paging messages. Reducing the area that needs to be paged helps to lower the number of paging messages that are sent out. Routing areas are identified by Routing Area Identity (RAI), which is a composition of a previously described LAI/TAI plus Routing Area Code (RAC), which is a one octet long code.

The LAI, TAI and RAI are temporary subscriber data and stored in the VLR and SGSN respectively.

- Cell identity

Inside every LA, TA and RA there are a number of cells that can be identified by Cell Identity (CI, Cell ID). This identity is unique only inside the location area. For these reasons, the term of cell global identification (CGI) has been introduced. CGI is a composition of LAI/TAI/RAI and CI, where CI is two octets long and can be coded using hexadecimal representation similarly to LA. CGI is the unique identifier of a single cell. As can be seen, identifying a single cell inside a network requires quite lengthy identification codes. For this and several other reasons, MNOs might create their own cell identity codes.

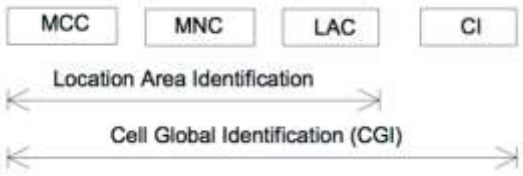


Figure 3.3 Comparison of location area identification (LAI) and cell global identification (CGI) Source: ETSI TS 123 003

The last known Cell ID is temporarily stored along with the subscriber identifier, for example in the VLR and SGSN. These registries are used by the mediation system for creating CDR files that would be transferred to a billing system and data warehouse.

Unique Cell IDs (CGIs or MNO-created) are commonly also stored in a database on the OSS level with several other attributes that describe the cell. Since the MNOs' network can change frequently (e.g., addition of new antennae, directional changes for individual cells), it is important to establish an understanding between an MNO and NSI on how often such data should be sent. Ideally, updated cell information files should be sent together with event data.

- Coverage area of the antennae

Usually, the minimum geographical information provided by the MNOs is the network antennae location (coordinate pair of the antenna point). If the geographical breakdown and accuracy of the resulting statistical indicators is low-scale (country-level maps, NUTS 3, LAU 1), the geographical representation of the network antennae as coordinate points is sufficient. However, if the planned geographical breakdown is to be more precise (LAU 2, LAU 3, municipality, village, city districts), the coverage information of the antennae is important, as antennae might cover areas over several municipalities (i.e., the antennae are located in one municipality but the majority of the coverage area covers another municipality). Therefore, it is important that either the MNOs also provide the geographical coverage area of each antenna (part of data preparation), or the theoretical cell coverage area be calculated during the pre-processing of the data (see Figure 3.4).

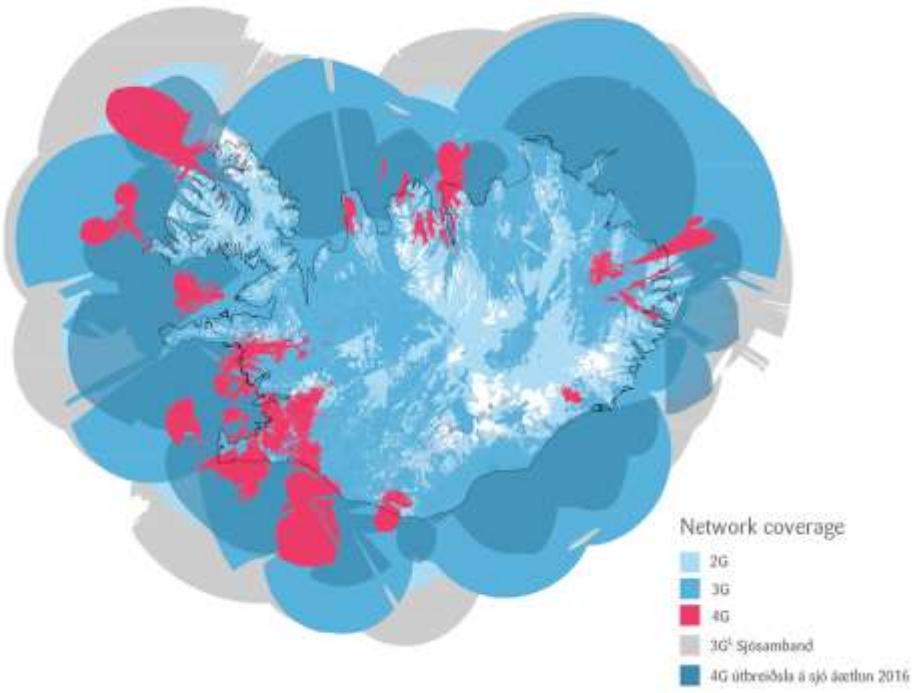


Figure 3.4 Illustration of one MNO's network antennae coverage area. Source: SIMINN

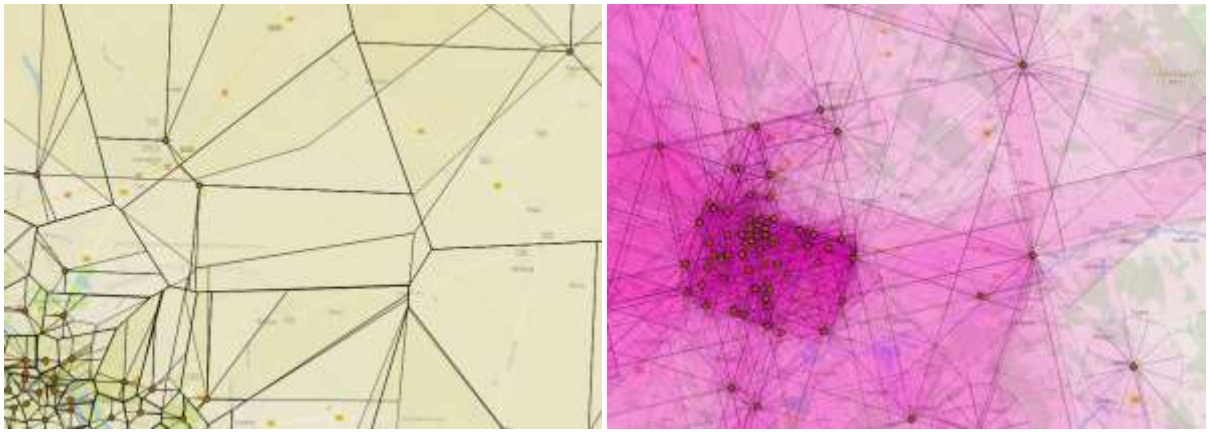


Figure 3.5 Example of two different methods for calculating the theoretical coverage area of a network antenna: a) Voronoi/Thiessen tessellation; b) theoretical sectors of the coverage area

3.2.6. Events data additional attributes

The list of possible attributes in these specifications is exhaustive and, alongside with useful attributes, it contains many elements that are of little use for enriching the data for population statistical analyses. The most common attributes to consider for inclusion are:

- record type – helps to differentiate between data types (CDR and IPDR), service types (e.g., SMS, MMS, outgoing call, etc.) and distinguish between events that are not subscriber generated (e.g., location updates);
- call duration or data volume for analysing mobile phone usage patterns;
- subscriber or equipment identity for receiving/sending parties.

3.2.7. Network data additional attributes

Additional parameters for a network are generally accessed from the OSS where the MNO stores data that is relevant for network maintenance purposes. For improving the location accuracy of the events, attributes for the cells are of great value.

3.2.8. Subscribers' additional attributes

The MNOs store information about their customers in subscriber databases (for domestic and outbound customers). The information that is collected and stored is highly MNO-specific, but typically includes (Eurostat 2014):

- socio-demographic characteristics of the subscriber (owner of the contract) that might be age, gender, preferred language, etc.
- details on the contract and service:
    - private or business client
    - invoice address
    - average cost of the service
    - contract type (pre-paid, post-paid SIM, machine-to-machine SIM).

This information should be used with great caution, as it is highly sensitive and not always accurate (phone user may differ from the contract holder). The value of these attributes lies in

the possibility to add extra dimensions to statistical analysis (e.g., gender), as well as in the option of performing data cleaning processes (e.g., removal of M2M data).

The socio-demographic attributes of the subscribers are usually collected in the CRM system of the MNO and related to the customer's profiling system. The socio-demographic attributes are mostly collected for post-paid private (non-commercial) customers, as there is often no information on the profiles of pre-paid customers or the information is limited (unless they are required to register upon the purchase of the SIM card, and even then it is not always available in the CRM). For commercial contracts it is often not possible to know which specific person is using the phone (and therefore the attributes of the person are unknown). With post-paid private customers, data is also often biased in case of family plans – the socio-demographics of the person who signed the contract (father, mother) are extended to the whole group (meaning the age and gender of the children using the phone are by extension those of the father's).

3.3. General data extraction process

Individual steps of the data extraction process depend on who is responsible for data processing and where it will be done. If the data will be processed by the MNO or on the MNO's premises using a Sandbox-like platform, several extraction steps, like encrypting the data, can be omitted.

3.3.1. Data preparation

The data preparation step typically involves creating a data extraction script to extract necessary data from the storage unit. It should be done by the MNO, most likely with database communication languages such as SQL. The extraction script preferably extracts data automatically at fixed intervals previously agreed upon between the MNO and NSI. This way it is possible to minimize delays that can be caused by the human factor.

The data preparation script should take into account details such as the type of data to be sent (e.g., CDR, IPDR or both), time period and attributes that are needed, also the file format and need for anonymization (for reasons of clarity, this will be discussed in a separate chapter). The data preparation script also involves some basic data processing steps that the MNO might undertake to provide high-quality data. These steps would involve the removal of non-representative data or subscribers that are black-listed for security reasons. The removal of non-representative data would include the exclusion of subscribers that do not represent real humans.

As was stated before, the data preparation step would require the MNO to assign a person to write this script, and, depending on the execution, either run it at certain times manually or make sure that automatic processes are working as expected. For example, if there are delays or data is missing, this person should be able to tell why these problems occurred.

3.3.2. Data anonymization

The purpose of the data anonymization process is privacy protection. During this process, a subscriber's personal identity code can be modified or data can be aggregated to give anonymity to the subjects. Although there are several ways to do that, no good method

currently exists to maintain all the aspects of the data needed for longer study periods while making it anonymous to the required extent. While receiving data where a subscriber is identifiable by the IMSI would be the ideal case, different legislation and MNOs' policies limit the methods that can be actually used. lists possible options for anonymization. Alongside with options presented in , sampling (decreasing the possibility for a person to be included in data set) and obfuscation (masking/hiding original data) could be used as alternatives or employed for further increasing the level of privacy protection.

When considering these options, the scope of the project where the data will be used should be discussed in detail to understand what anonymization method would be feasible for the NSI as well. Anonymization (if present) should be part of the data preparation script. If pseudonymous unique codes will be used, the MNO needs to create and maintain hash functions to guarantee the same key and value pairs for the period agreed upon.

Table 3.2 Limitations that various levels of anonymization bring to the usefulness of the dataset

Anonymization level of subscriber identifier	Level of protection of personally identifiable information (anonymity)	Limitations									
		ID combining with other MNOs	Domestic activity space recognition	Usual environment	Trip identification	Same-day / overnight visits	Border bias	Long-term visitors	Transit visits	Repeating visits	Visit routes
Permanent unique code (IMSI)	Not protected at all.										
Permanent pseudonymous unique code	Very weakly protected. Pseudonymous but not anonymous. Indirect identification is possible.										
Temporary pseudonymous unique code 2 years	Very weakly protected. Indirect identification is possible.										
Temporary pseudonymous unique code 6 months	Very weakly protected. Indirect identification is possible.										
Temporary pseudonymous unique code 1 month	Very weakly protected. Indirect identification is possible.										
Temporary pseudonymous unique code 1 week	Weakly protected. Indirect identification is most possible.										
Temporary pseudonymous unique code 24 hours	Very well protected. Indirect identification is possible but in very few cases.										
Temporary pseudonymous unique code 90 minutes	Excellent protection of identity (can be considered fully anonymous). Indirect identification almost impossible.										
Aggregated raw data	Subscribers' identities totally protected if threshold is also used (e.g., not showing aggregates under 10).										

3.3.3. Data encryption

Data encryption is a small but important step that is needed when data processing takes place outside the MNOs' premises. Its' purpose is to ensure that unauthorized parties are not able to read the data in case of a security breach during data transmission.

During data encryption, a key is generated that will define how data will be encrypted into a cipher text. In case of symmetric encryption, the same key will be used to decrypt the data. For asymmetric encryption, a key pair is generated where the public key will be used to encrypt the data and the private key to decrypt it. It is generally advised to use asymmetric encryption if keys will be exchanged over the Internet.

There are several cryptographic libraries available to use for encrypting and decrypting data. Examples include OpenSSL and GNU Privacy Guard, both of which can be used for commercial and non-commercial purposes.

3.3.4. Data transmission

During transmission, data will be made available for the NSI either for in-house processing or processing outside the MNO's premises (centralized). In the first case, data is generally transported into a database dedicated for this purposes or made available in the form of a data file on a server that the receiving end is able to access. If centralized data processing is used, the easiest ways to transport data is:

- the MNO will dedicate server space to extracted data and make it accessible to the receiving end. The receiving end will connect to the server and transfer the data files to their dedicated server space
- the receiving end will dedicate server space to extracted data and make it accessible to the MNO. The MNO will connect to the server and transfer the data files.

As can be seen from above, in either case, both parties would need to dedicate some server space to transfer the data.

3.3.5. Data archiving

In case of an accidental data loss due to technical problems or any other unforeseeable cause, it is important that historical data could be reacquired and used for recalculation. For the MNOs this simply means the re-extraction of historical (archived) data. However, if data is transmitted to an external processing facility after the extraction, all extracted raw data should be kept in a dedicated storage unit that is not connected to the same unit where the data processing takes place. These archives should follow the legislation concerning data retention and kept only for the period of time allowed by regulations.

3.3.6. The logical order of steps in the process of data extraction

1. Where can data be processed? At MNO or outside MNO
2. What is the origin of the data? Probing or database
3. What data sources are available? Inbound, outbound, domestic, CRM, geographical reference database
4. What is the data type coverage? CDR, IPDR, other

5. What is the cellular system generation coverage? 2G, 3G, 4G
6. What is the record type coverage? Call, SMS, MMS, data traffic, location area updates, other
7. What is the subscriber coverage? All or sample
8. What is the user coverage by payment type? Pre-paid or post-paid
9. What is the data coverage by charging system type? Offline or online charging
10. What attributes are available in the data? Subscriber identifier, location, time, other attributes
11. What are the formats of identifiers for the above?
12. How will references to countries be handled (inbound and outbound data)? MCC from subscriber identifier or separate field
13. What is the method for privacy protection? Data aggregation, unique pseudonymous identifier, sampling, obfuscation, no anonymization
14. How will data be organized in files?
15. What are the specifications for data file elements? Attribute names, types, file names, file type
16. What are the specifications for data extraction script? Removal of non-representative data (M2M / IoT, duplicates), removal of black-listed subscribers
17. How will encryption be handled? Encryption software or key exchange
18. How often can data be extracted from the storage systems? Every day, month, other
19. How should the data be transferred? Push or pull
20. Where should raw data be archived? At MNO or NSI premises or not at all

3.4. Coping with under/over coverage

In case of mobile positioning data, there is one discrepancy that is immediately clear from comparing the definition of the target populations and the population frame. In the frame are all of those subscribers who use a mobile phone, whilst the target population includes all of those individuals who reside in the country. This leads to a large number of various coverage problems, a list of which is given in Table 3.3.3.

Some coverage issues that are listed in the table can be avoided during frame formation or data compilation (before estimation is carried out) by identifying those observations that do not form part of the population and by excluding them from the frame.

Table 3.3 List of possible coverage issues with data from MNOs.

Issue	**Under- or over-coverage**	**Possible solution**
1. Data from selected MNOs only i.e. not all data from MNOs is available.	Under-coverage	Apply external information regarding the penetration rates and customers' profile.
2. People who do not use mobile phones	Under-coverage	At estimation stage use information from additional sources to make and check assumptions about the

Issue	Under- or over-coverage	Possible solution
		behaviour of the under-covered group and then apply appropriate models to adjust estimates. Both model based and model-assisted estimators can be used.
3. Use of more than one mobile device: • all devices use the same network causing duplication in one of the MNO's datasets; • devices use different networks causing duplication in the datasets supplied by different MNOs.	Over-coverage	At estimation stage use information from additional sources if available to make and check assumptions about the number of people having several mobile phones and then apply appropriate models to adjust estimates. Adjustment depends on the frequency of the phenomenon.
4. Different penetration rates for MNOs among foreign subscribers of specific countries.	Under-coverage or over-coverage	Assuming that information is available that allows to create separate MNO-based models that have country-specific parameters then weight adjusted estimates can be computed.
5. Different regional and socio-demographic penetration rates for MNOs.	Under or over-coverage	Surveys covering regional and socio-demographic penetration rates of MNOs, if available, can be used in estimations.
6. Machine-to-machine communication devices that were not removed by MNOs during the preparation process.	Over-coverage	Possible removal based upon event patterns. Based upon the Estonian inbound roaming data, the percentage has been from 0.04% to 2% depending on the MNO and the algorithm used. Based on Estonian outbound roaming data the number of such devices is very small, less than 0.01%.
7. Inbound: national roaming subscribers.	Over-coverage	Can be fully excluded as the country of roaming partner is the same as the country of reference.
8. Inbound: Visitors not actually entering/exiting the country of reference (not crossing the border) but who are using the MNO	Over-coverage	Can be excluded partially based upon border bias recognition algorithms. Based upon the Estonian data the

Issue	Under- or over-coverage	Possible solution
roaming service of the country of reference or foreign country.		percentage of excluded trips among inbound roaming has been 10.4% of all trips. A total of 58% of such trips are classed as trips with only one event.
9. Tourism: same-day visits are over-represented when compared to overnight visits. This is caused with CDR data because it is likelier that visitors do not use their mobile phones on every single day of the trip.	Over-coverage of shorter visits on the account of longer visits. Under-coverage of the duration of the trips.	Model the likelihood of single trip being a same-day trip and use modelled values in the estimation. In case reliable reference data concerning the difference between the same-day and overnight visitors exists, such data should be used to correct the mobile data.
10. Inbound: visitors who use local SIMs for some reason and are not represented in the inbound/outbound roaming registry (see also Point 18).	Under-coverage	Calibration or similar approach where known totals from other sources are used to correct for the under-covered part of the population. Use data from CRM about origin of SIM card owners if the country has mandatory registration requirements.
11. Technological limitations on the use of mobile phones from/in specific countries due to technological barriers, limitations on roaming service (no roaming agreement between MNOs), high roaming costs, or network coverage is bad.	Under-coverage	Has to be a part of different MNO-based country-specific estimations compensating the technological limitations. In Estonia, subscribers from US and Japan are in such a situation and therefore require higher correction coefficients.
12. International roaming (travel) SIMs that are sold globally and where the country of the MNO does not match the actual country of origin of the subscriber.	Over-coverage of the country of the MNO, under-coverage of the subscribers from the countries that are using this service.	Exclusion of such subscribers should be possible for outbound roaming as MNOs usually know who their subscribers using their international travel roaming service are. Possible compensation on the account of the actual residents of the country of reference might be required. The number of such roaming cards is very different in countries and it is very difficult to

Issue	Under- or over-coverage	Possible solution
		measure.
13. Differences in phone usage patterns depending upon the location peculiarities; therefore generating different volumes of events.	Under-coverage in rural areas	At estimation stage use information from additional sources to make and check assumptions about the travel behaviour of the under-covered group and then apply appropriate models to adjust estimates.
14. Limitations on the use of mobile phones in specific areas within a single country in which network coverage is poor.	Under-coverage	If the area is large and is relevant from additional sources to make and check assumptions about the travel behaviour of the under-covered group and then apply appropriate models to adjust small area estimation methods e.g. a synthetic estimator could be applied.
15. The smaller the geographical level, the less representative the mobile data will be when it concerns travel to or through those places.	Under-coverage on lower administrative levels	Higher correction coefficients for lower administrative levels when compared to higher administrative levels.
16. Cross-roaming, i.e. a single device using several roaming services during the trip causing duplication of the data for a single subscriber in two or more MNOs' datasets.	Over-coverage	Unless subscriber records can be linked across different MNOs, this has to be dealt with during the estimation process.
17. Inbound: residents of the country using foreign phones.	Over-coverage	Can be excluded partially based upon the duration of the presence within the country of reference. Based upon the Estonian data approx. 0.2-0.4% of the total number of trips can be identified as being trips made by residents and not tourism.
18. Tourism: Visitors passing through the country (transit visits).	Over-coverage	Can be excluded partially based on identifying short trips within transit corridors. Depending upon the month, the number of transit trips

Issue	Under- or over-coverage	Possible solution
		in the Estonian data vary from 2% to 10% of total inbound trips and 9.2% of the stays in foreign countries can be considered as transit pass-through.
19. Tourism: non-resident subscribers whose usual environment and residence are outside the country of reference but who are using local pre-paid SIMs for various reasons (see also Points 10 and 20).	Over-coverage	Exclude subscribers with a short lifetime or if ID linkage to outbound data is possible, compare the duration of stays within the country of reference and abroad for exclusion from the domestic and outbound data and inclusion to inbound data. Based upon Estonian data 2.9% of all domestic subscribers spend more time abroad, representing 24% of all outbound trips, and therefore should be considered as foreign residents and excluded from domestic tourism dataset.
20. Resident subscribers who change their pre-paid cards very often (short longevity of the *subscriber_id*) and whose residence and/or usual environment cannot be identified (see also Point 19).	Under- or over-coverage	Exclude subscribers with a short lifetime.
21. Tourism: subscribers whose place of residence and/or usual environment within the country is calculated incorrectly.	Under- or over-coverage	The usual residency and usual environment are 'compensated' by similar incorrect calculations for other subscribers. The quantity of incorrect calculations depends on the geographical level.
22. Outbound: residents of foreign countries who are using SIM cards from the country of reference;	Over-coverage	If ID linkage to outbound data is possible, compare the duration of stays within the country of reference and abroad for exclusion from the domestic and outbound data and inclusion to inbound data. Also estimation. Based on Estonian data roughly 3% of combined

Issue	Under- or over-coverage	Possible solution
		domestic-outbound subscribers spend more time abroad than in the country of reference.

There are many contributors to the coverage bias, but due to the co-effect some bias components cancel each other out (over-coverage versus under-coverage), some contribute very little, and some may contribute a lot. Many problems, however, are inherent in the mobile positioning data and therefore cannot be avoided. Furthermore, their total effect, i.e. the total size of the coverage bias of an estimate of interest, needs to be evaluated or bias-corrected estimates need to be computed.

Some information about mobile phone usage while travelling is available for Europe. The Special Eurobarometer 414 (2014) shows that 28% of the travellers in the EU switch off their mobile phones when visiting another EU country (the corresponding specific figures are 33% for DE and 41% for FR but only 13% for EE and 15% for FI).

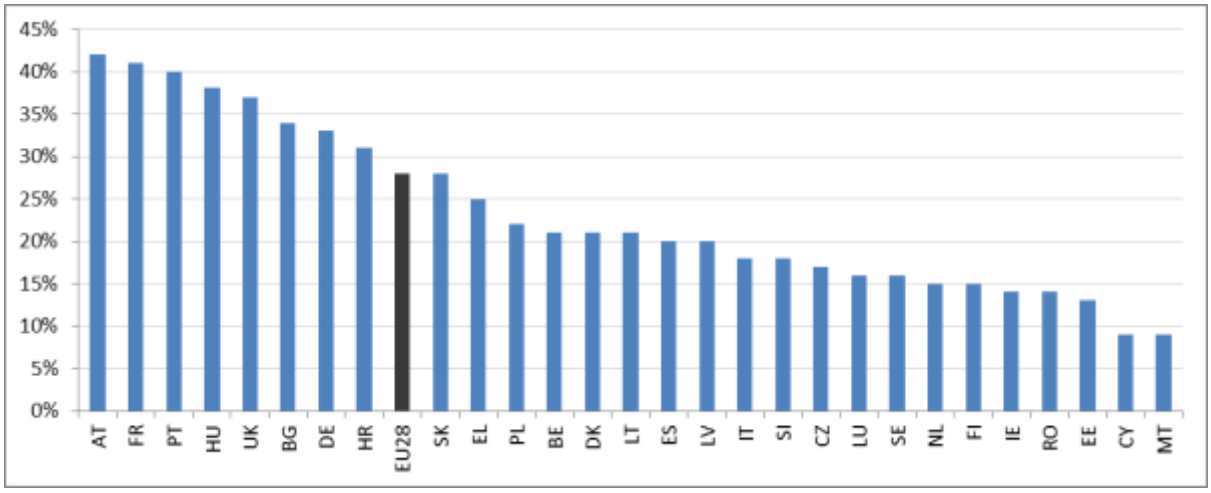


Figure 3.6 Percentage of the population (aged 15+, who have travelled and own a personal mobile phone) who generally switch off their mobile phone and never use it while visiting another EU country (Source: Special Eurobarometer 414).

The best but also the costlier way to evaluate the bias would be to carry out sample surveys for quality assessment purposes for each coverage issue, providing the basis for individual estimations. This survey would give the estimates for the proportion of people not having or using mobile phones, proportion of people using two or more devices etc. This information would allow to compute the total size of the bias under certain assumptions. Linking survey data with mobile positioning data at a micro level (which would reveal quality problems not only at an aggregate level but also at a record level) would allow more precise bias estimations to be carried out. However, this solution is ideal from a theoretical point of view, but it is not a very realistic solution in practice.

A less costly way to reduce the bias is to use the information from the other sources (e.g. results from the surveys covering the relation between the number of same-day trips and overnight trips to the country for inbound, the usage of mobile phone in different countries for outbound). Sample survey literature has shown that a coverage bias can be dealt with by using a method such as, for example, a calibration estimation if suitable auxiliary information exists. In addition, different model-based estimates using aggregate data from other sources can be constructed. Both model-assisted and model-based estimates are well covered in literature in which tools for carrying out model validation and methods of computing precision measures are also supplied.

If no coverage problems existed and the data from MNOs represented the perfect frame for the population of interest, estimation would not be necessary. However, this is never the case and therefore using data from all MNOs from the country of reference decreases the overall coverage bias, but the estimation process is still required.

The evaluation of the size of all the listed coverage problems needs to be carried out for each country separately, as many problems listed here depend on the environment (e.g. the price of the calls) and culture (e.g. children having mobile phones, having many mobile phones).

3.5. References

Tiru, M., Ahas, R. (2012) "Passive Anonymous Mobile Positioning Data for Tourism Statistics". 11th Global Forum on Tourism Statistics http://congress.is/11thtourismstatisticsforum/papers/Session3.pdf

ITU (2014) “The role of big data for ICT monitoring and for development” Measuring the Information Society Report 2014 https://www.itu.int/en/ITU-D/Statistics/.../mis2014/MIS2014_without_Annex_4.pdf

Eurostat (2014a) Report 2 − Feasibility of access. Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics

Eurostat (2014b) Report 3a − Feasibility of use, methodological issues. Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics.

Special Eurobarometer 414 (2014) e-Communications Household Survey and Telecom Single Market Survey Roaming Results. TNS Opinion & Social at the request of European Commission: https://ec.europa.eu/digital-agenda/en/news/e-communications-household-survey-and-telecom-single-market-survey-roaming-results-special

4. Access to mobile phone data and partnership models

4.1. Introduction

Mobile phone data is considered to be one of the most promising types of Big Data that can be used to generate official statistics (Hilbert 2016; Blumenstock, Cadamuro & On 2015; Letouzé, Vinck & Kammourieh 2015). Mobile phones are widely adopted – global subscription rates are now estimated to be almost 100 per cent (ITU 2016). As such, mobile phone data represents a rich source of information on a variety of human activities in a wide spectrum of countries, both developed and developing. The steady growth in subscription rates means that as mobile phone use becomes more universal, mobile phone users are also becoming more representative of national populations. At present, mobile phone data is already generating a more pervasive data footprint in developing and least-developed countries, compared to other well-known sources of Big Data such as social media, e-commerce transactions, and other Internet-based activities. Its widespread availability make it a desirable source of globally comparable statistics. However, for a number of reasons, obtaining access to mobile phone data for use in official statistics is seen to be a major challenge (UNSD 2016).

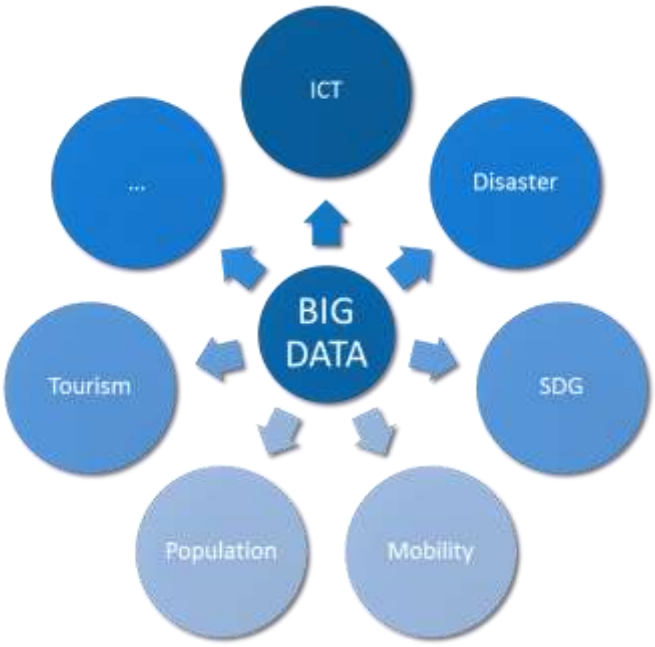


Figure 4.1 Multiple uses of big data for official statistics
Source: Tiru (2016)

First, there are capacity gaps: various stakeholders such as national statistics offices, telecommunication regulators, ICT ministries, data privacy authorities and telecommunication operators may possess differing mandates and varying levels of technical capacity and resources. The mismatch in mandates, resources and expertise lead to capacity gaps that potentially hinder access to mobile phone data. Government agencies who are

responsible for collecting and publishing official statistics from mobile phone data may not have the computing infrastructure and technical staff to transfer, store, process and analyze Big Data, especially in developing countries. Processing large amounts of data is expectedly within the purview of national statistical offices but they may not possess domain specific knowledge of mobile phone data nor have ongoing contact with telecommunication operators. The ICT ministries and telecommunication regulators are generally better positioned to negotiate access to mobile phone data but they may not have the technical expertise to facilitate the transfer and storage of the data. It is the telecommunication operators who already collect, store, process, and analyze mobile phone data as part of their daily business operations who are most likely to have the Big Data infrastructure and knowhow. These capacity gaps exist in varying forms and degrees in both developed and developing countries, but may present more significant challenges in the latter.

Second, there is a privacy gap arising from shortcomings in the legal framework governing the processing, use, and transmission of data. In many countries, laws and regulations concerning cybersecurity, data privacy or data protection are either missing or inadequately enforced. The absence of adequate legal safeguards makes the use of mobile phone data for official statistics more risky; affected parties, whether they be individual customers or businesses, will not be protected by law should there be intended or unintended consequences. Rightly or wrongly, telecommunication operators could cite increased risk as a reason for refusing access to mobile phone data for purposes outside of their own business operations. Yet, the promulgation of cybersecurity, data privacy or data protection laws does not necessarily guarantee fast and easy access to mobile phone data for official statistics. Rather, the primary role of these laws is to lower the risks associated with using data and to provide assurance to stakeholders.

Building partnerships are crucial towards closing the capacity and privacy gaps that hinder access to mobile phone data for use in official statistics. Mobile phone data is collected, stored and processed by telecommunication operators whose main interest is in business uses of the data. Government agencies interested in using mobile phone data to generate official statistics on various aspects of national life, whether it be tourism, education, migration, poverty, and etc., will require cooperation from telecommunication operators and other network providers. Such cooperation can be facilitated, first and foremost, by laying the legal foundation for operators and other network providers to give access to data in a manner that is secure and respects the privacy rights of individuals. Ideally, this means that rules to protect privacy must first be legislated and enforced, if there were none to start with. Cooperation can be further facilitated by working hand in hand with telecommunication regulators and data privacy authorities. In many countries, telecommunication regulators have the mandate to regularly request data from telecommunication operators in the course of their regulatory work and are therefore well-positioned to obtain access to mobile phone data. On the other hand, data privacy authorities, where they exist, are mainly responsible for enforcing national data privacy or data protection laws. They are also crucial partners in providing confidence to stakeholders over the use of mobile phone data for official statistics. Government agencies and national statistics offices will greatly benefit from partnering

closely with telecommunication regulators and data protection authorities when seeking to use mobile phone data.

In this chapter, we distill best practices in building partnerships for accessing mobile phone data to derive official statistics. We detail the main ways in which these partnerships can be configured, the challenges to be overcome, and the roles of four key stakeholders: national statistical offices, telecommunication regulators, telecommunication operators/network providers, and data privacy authorities. We also give examples of these partnerships using the experience of the International Telecommunication Union (ITU), the United Nations specialized agency for information and communication technologies. The ITU is actively exploring ways of using big data from the ICT industry to improve and complement existing statistics and methodologies to measure the information society through various country pilots. Much of the discussion in this chapter is drawn from the insights generated by ITU's experience with the country pilots.

4.2. Enabling environment for access to mobile phone data for official statistics

Partnerships, whatever their form, are more likely to flourish when there are enabling institutions and supportive legal frameworks. To ensure productive partnerships among key stakeholders, it is also necessary to tailor the partnership model such that each stakeholder plays a role that complements each other in terms of capacity and mandates. In the case of telecommunication operators, it is crucial to understand the market incentives they are facing and whenever possible, to harness these market incentives in order to motivate their cooperation. An often overlooked but equally critical component of a productive partnership is support from the data privacy authority, where it exists. Partnerships for accessing mobile phone data for official statistics are more likely to achieve its goals without unforeseen delays (and costs) when there is buy-in from and involvement of the data privacy authority early in the partnership. If data privacy authorities are involved at the start of the undertaking, they can help stimulate an atmosphere of trust among different stakeholders and inspire confidence in the safe, secure, and socially beneficial use of mobile phone data. In the proceeding sections, we elaborate on these ideas by outlining different types of partnership models and discussing roles, capacities, mandates and incentives of the four key stakeholders. We also investigate the challenges of using mobile phone data in the normal course of producing official statistics, especially those arising from the privacy and security of mobile phone records as well as their continuing availability and suitability for statistical purposes.

4.2.1. Partnership Models for Using Mobile Phone Data for Official Statistics

Partnerships for accessing mobile phone data for official statistics can come in many shapes and forms. There is no one-size-fits-all partnership model; each has its own strengths and weaknesses. Partnerships can differ in terms of management, funding, objectives and allocation of data processing tasks. When objectives are long term and broadly defined, i.e. the set of official statistics covers a wide range of sectors and will be periodically derived

from mobile phone data for many years in the future, national statistics offices will likely be found taking a management role. In partnerships where objectives are narrowly defined and focused on generating official statistics on specific sectors, the government agency responsible for the sector may also choose to take a management role and fund most of the activities. For example, when the goal is to generate official statistics for telecommunication and ICT, telecommunication regulators and ICT ministries are often the ones who take the initiative to build the partnership, fund most activities, and manage the outcomes.

Funding, of course, is a critical ingredient in effective partnerships. There are many different approaches to funding activities geared towards the use of mobile phone data for official statistics, each with its own set of challenges and advantages. Again, there is no one correct approach for all circumstances in different countries. Whatever the chosen approach, it is good practice to inventory and anticipate at the beginning of the partnership, all the direct and indirect costs that may be faced by each stakeholder through the entire course of their participation. Consideration must be taken, not only of new capital and administrative outlays, but also of the opportunity costs of existing personnel time and equipment that will also be used in partnership activities. In many partnership models, it is inevitable that the responsibility for funding activities will be shared by all stakeholders, whether in cash or in kind. Although the partnership may have access to dedicated funds from central government or other sources, to be used for capital outlays or for the hiring of specialized expertise, each stakeholder will inescapably have to divert personnel time and resources to work on partnership activities. Recognition of these types of opportunity costs from the outset will greatly aid stakeholders to plan and effectively manage their deliverables.

However, no matter who takes the lead or funds most of the activities, and what the statistical objectives are, ultimately the key distinguishing feature of a partnership model lies in the allocation of responsibility for data processing. In particular, who are the stakeholders responsible for processing raw data, e.g. Call Detail Record (CDR) and Internet Protocol Detail Record (IPDR), into statistical indicators? Where is the data processing done? Based on the allocation of data processing tasks, we can identify two main types of partnership models: a) telecommunication operators mainly responsible for data processing, or b) government stakeholders/non-operators mainly responsible for data processing. The choice in the allocation of data processing tasks will have implications on the distribution of the costs as well as on the needed measures to ensure security, privacy and data protection. To better understand the differences between these two partnership models, we first need to define the different tiers of data corresponding to various stages of data processing, from extraction of raw data to the computation of publishable indicators.

1) **Tier I** data consists of initial, raw, and not aggregated data containing business confidential and personally identifiable information. The structure of Tier I raw data will depend on the specific database from which it was extracted and will vary between telecommunication operators. Tier I data from different operators is

not yet ready to be merged. It will need to be further cleaned, formatted and aggregated into Tier II data.

2) **Tier II** data is initially aggregated data with no personally identifiable information and some business confidential information. Tier II data is already prepared in such a way that data from different operators is ready to be merged.

3) **Tier III** data consists of aggregated indicators that can be publicly shared and no longer contains any personally identifiable or business confidential information.

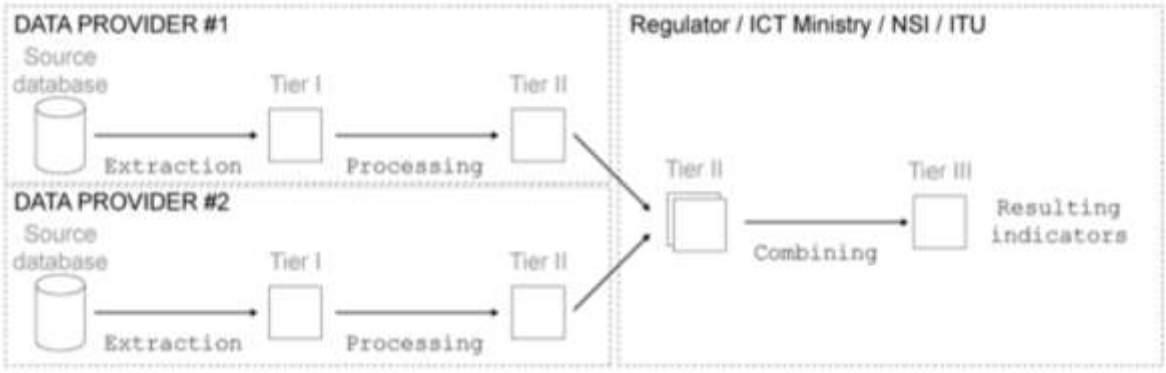


Figure 4.2 Type A – Telecommunication operators mainly responsible for data processing
Source: ITU 2016

In a **Type A partnership model**, as illustrated in Figure 4.2, the bulk of data processing is done by telecommunication operators within their premises. Operators are essentially responsible for extracting raw data and processing them until the Tier II stage. The Tier II data from different operators is then merged and processed further into Tier III data outside the premises of the operator, by a responsible government agency or a consortium of government agencies and non-government organizations. In some cases where the partnership involves international organizations or foreign entities (and contingent on data sovereignty requirements), processing of Tier II to Tier III data may even occur outside national borders.

The main advantage of the Type A partnership model is that it gives data providers (in this case, telecommunication operators) greater confidence that business confidential and personally identifiable information will not be accidentally disclosed to irrelevant third parties. Given that the bulk of data processing is done by their own personnel within their premises, the operators will have more control over the risks to data privacy and security. For this reason, they may prefer the Type A partnership model. Consequently, it may be easier to secure the participation of telecommunication operators using this model. The downside for the operators is that they will have to share more of the costs and devote more personnel time and equipment to data processing tasks. However, since operators are, more often than not, already undertaking Big Data processing as part of their regular business operations, their

marginal costs will probably be lower compared to government stakeholders who may need to make new investments in computing infrastructure and technical expertise.

In terms of statistical outcomes, i.e. validity and accuracy of computed indicators, the Type A partnership model will pose additional challenges for data verification. Before indicators computed from Big Data can be adopted as official statistics, it will be necessary to check scripts and algorithms used to extract raw data and transform them into Tier II data. This means that the government stakeholder that receives the Tier II data will need to have the required technical expertise to perform such checks and evaluate the completeness and accuracy of initially aggregated data. For the Type A partnership model, it will be particularly crucial to put in place mechanisms to ensure that the processing of the data, from extraction to Tier II, are transparent, documented, and replicable.

Data scientists who are not working for the network operator can only access Tier II aggregated data under the Type A model, so there is insufficient granularity in the data to perform detailed location-based analysis. For example, CDRs reveal patterns in population movements that is useful for predicting outbreaks of deadly epidemics in specific locations. However, this potentially life-saving information is lost when data is aggregated.

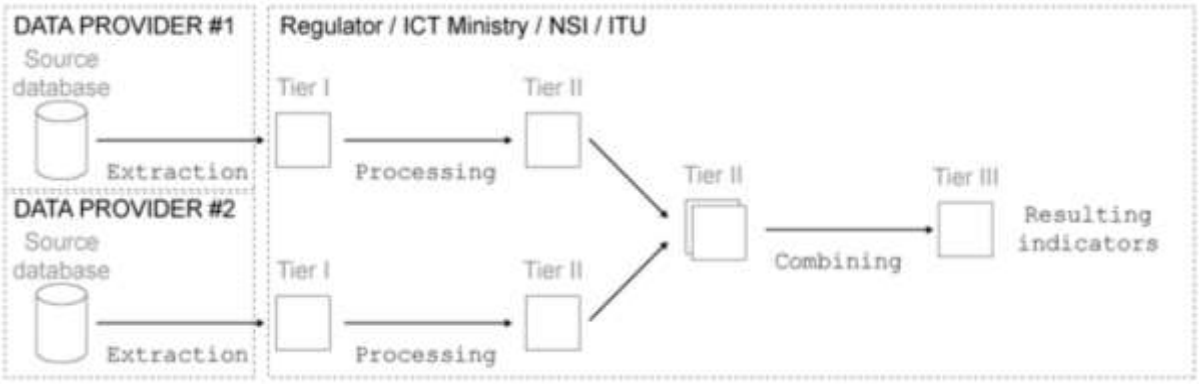


Figure 4.3 Type B – Government stakeholders mainly responsible for data processing

In a **Type B partnership model**, as illustrated in Figure 4.3, the bulk of data processing is done by government bodies, outside the premises of telecommunication operators. Operators are responsible for extracting raw data and transferring them as Tier I data to responsible government stakeholders. The government stakeholder (or consortium of government and non-government organizations, as the case may be) will then be responsible for cleaning and aggregating Tier I data into Tier II data, merging Tier II data from multiple operators, and finally, computing aggregated indicators contained in Tier III data. In some cases where the partnership involves international organizations or foreign entities (and contingent on data sovereignty requirements), processing of Tier I to Tier III data may even occur outside national borders. However, given the potentially sensitive and private information contained in Tier I data and to some extent, Tier II data, processing outside of national borders will require extra care with redaction, anonymization and security.

From the perspective of ensuring the veracity and continuity of deriving official statistics from mobile phone data, the Type B partnership model hold multiple advantages. For one, government stakeholders publishing official statistics from mobile phone data will be better positioned to verify and evaluate the correctness of the computation of aggregated indicators. Since they themselves are processing the data, they will be able to not only examine the content of the algorithms and scripts to determine whether they are in accordance with prescribed methodologies, but they can also re-trace directly how the scripts were run on the data to produce the indicators. Given that most data processing is done by government stakeholders, there is also potentially less disruption to the production flow of official statistics when operators enter and exit the market. Since operators are limited to the task of extracting raw data, there will be no need to orient new entrants on prescribed methodologies beyond specifying the required data fields.

The Type B partnership model will require government stakeholders to already possess the infrastructure and technical expertise to undertake Big Data processing, or if not, spend significantly more on setting up the computing infrastructure and on acquiring the expertise. The additional expenditure required on the part of government stakeholders are not necessarily drawbacks. Done strategically and with a view to the future, these expenditures can be understood as long-term public investments that can yield benefits beyond its immediate goals. For example, Type B partnerships initially aimed towards producing official statistics for specific sectors, can extend the same approach to other sectors, using the newly acquired infrastructure and expertise. Extending to other sectors can more easily be done if data processing tasks, infrastructure and expertise is concentrated in national statistical offices, since they are the government body normally mandated to collect data and produce official statistics on a wide variety of economic and social activities.

Whatever the legal jurisdiction and the privacy paradigm being applied, the Type B partnership model will pose an additional challenge to compliance with data privacy requirements, given that it requires the transfer from one entity to another, of mobile phone records containing personally identifiable information. In most jurisdictions, the act of transferring such data is a sensitive and tightly regulated processing activity. This is especially true for countries with European-style data protection rules, in which the scope of regulated data and regulated processing activities are more broadly defined.

Under European-style rules, businesses and other organizations (including government agencies) are generally forbidden from processing personal data, where personal data is broadly defined as any information that relates to an identifiable human individual, whether or not the data itself allows identification of the data subject (European Union Agency for Fundamental Rights 2013). Similarly, data processing is also broadly defined to include any activity related to personal data, from collection to destruction and everything in between. Processing is permitted only when data subjects give their consent or when statutory exemptions apply (e.g. cases of public interest). Under such data privacy regimes, stakeholders in the Type B partnership model may need to seek exemptions. If exemptions are not granted, stakeholders will need to take additional steps to redact Tier I data so as to no

longer constitute personal data. Depending on country-specific rules, redaction may involve anonymization or aggregation, and not just pseudonymization. At the minimum, safeguards will need to be put in place to ensure that Tier I data is securely transmitted and stored.

Table 4.1. Comparing partnership models to access mobile phone data for official statistics

Characteristics	Type A	Type B
Responsibility for most of the data processing	Data providers/network operators	Government/non-government organizations
Granularity of data available for computing statistical indicators	Low	High
Perception of risks to privacy and business confidentiality arising from use of data for official statistics	Low	High
Range of potential statistical indicators that can be derived	Low	High
Ease of verifiability and statistical quality control	Low	High
Impact of the entry and exit of new operators in the market on the production of official statistics	High	Low

4.2.2. Understanding Stakeholders: Roles, Capacities, and Mandates

Effective and sustainable partnerships will require buy-in and commitment from four groups of key stakeholders: national statistics offices and/or sector ministries, telecommunication regulators, telecommunication operators and other network providers, and data privacy authorities (where they exist). Stakeholders have complementary roles depending on their mandates or organizational objectives and their technical capacities. In the case of operators, their market and regulatory environment will have a bearing on their incentives to share data. It is advisable for government stakeholders to firmly grasp what these incentives are and whenever feasible, shape these incentives through regulatory *quid pro quo*, in order to secure the cooperation of all potential data providers. In the discussion that follows, we will explore the potential role of each group of stakeholders based on their mandates or objectives, and their technical capacities.

National Statistical Offices (NSOs). In the era of Big Data, official statistics is at a crossroad and so are the NSOs who are their main purveyors (Skaliotis and Wirthmann 2014). A number of NSOs worldwide are now awake to the potential of Big Data to increase the timeliness and cost efficiency of producing official statistics, as well as provide new statistical products and services. According to the latest inventory of the United Nations Global Working Group on Big Data for Official Statistics, there are now more than 180 projects in both developed and developing countries, that explore or pilot the use of Big Data

for public statistical purposes. Many NSOs are already well-positioned to exploit Big Data sources since they usually possess the legal mandate to compel the provision of information from data providers and integrate sensitive data (Tam and Clarke 2014). They also have the experience and expertise in collecting and processing large amounts of data. Last but not least, NSOs are uniquely qualified to assess the quality and representativeness of Big Data sources, since they often already possess the required information to determine statistical benchmarks (Ibid., p.8).

However, not all NSOs are created equal. In some countries, NSOs were founded with a broad and clear mandate to collect data and produce official statistics impartially and independently. Such NSOs are usually vested with the right to access data from both public and private sectors, within the bounds of privacy and confidentiality. In contrast, there are countries where the NSO have the right to access data from public authorities only. Unfortunately, more often than not, the right of NSOs to access data are not accompanied by specific obligations for data holders and adequate penalties for non-compliance. Hence, an NSO may find that its legislative mandate is unsupported in practice. The legislative framework for NSOs may first need to be strengthened to enable them to take the lead in exploiting mobile phone data to produce official statistics.

The processing of Tier I mobile phone data will present technical challenges for NSOs due to its size. One call detail record has an average size of 200 bytes. If a country has 6 million mobile subscriptions of which about 20 per cent makes a call each day, the total file size for one month of records will be about 200 GB at minimum. Therefore, Tier I data will require the use of higher performance database systems and computing equipment to allow processing within an acceptable time frame. This will in turn also require new skillsets in data engineering. In comparison, the processing of Tier II data can be done using traditional computers and database systems. It is important to note that technical challenges are surmountable. The existing technical capacity of NSOs can be augmented by investments in equipment, software, and personnel. For least-developed countries, the limiting factor may be the availability of data engineering skillsets in the local job market at public sector wage rates.[29] Under such circumstances, the Type B partnership model may be more feasible, since the heavy lifting will be done by operators who are more likely to already have the technical capacity.

The particular role of NSOs in partnerships for accessing mobile phone data will depend on their mandate and capacity. Ideally, the NSO should take the initiative to forge the partnership and manage the statistical outcomes. If mobile phone data is to be used sustainably in the actual production of official statistics, the input and statistical oversight of NSOs will be crucial. At the minimum, the NSO should be informed about the activities and outcomes of the partnership, since mobile phone data is a potentially relevant source of statistics for a wide range of sectors that fall under the purview of the NSO.

[29] The scarcity of such skillsets can form a rationale for international cooperation and capacity-building.

Telecommunication operators and other network providers. For usage tracking and billing purposes, telecommunication operators generate CDRs containing the mobile phone transactions of their subscribers. These CDRs contain personally identifiable information and can also be used to derive sensitive business information (e.g. financial profile, subscriber profile). If mobile phone data is to be used for official statistics, it will be crucial to obtain CDRs from all operators. If not, the data will only represent the characteristics of the subscribers of participating operators and inadvertently reveal sensitive business information to non-participating competitors. This partial data cannot be used to produce official statistics.

The willingness of operators to provide access to CDRs will be influenced by their market and regulatory environment. Key factors could include: regulatory requirements to share data, government ownership, barriers to entry, total market size, number of firms, and relative market shares. If operators are already required by law to submit records to a regulatory agency as a condition of their license or franchise then it will be easier to access mobile phone data from both existing operators and new entrants (if any). Otherwise, it may be necessary to see whether it would be feasible to put in place a records submission requirement. This will be particularly crucial in telecommunication markets where there is minimal government equity since there will be less channels to influence record sharing. In markets where there are high rents from telecommunication services but minimal government equity, i.e. significant barriers to entry resulting in fewer operators relative to total market size, the resistance to records sharing could be strong. Therefore, such resistance will need to be anticipated and resolved at the beginning of the partnership, otherwise it will present a significant roadblock to accessing mobile phone data for official statistics. There need to be explicit mechanisms to assure operators that the records will be used for no other purposes outside of the production of official statistics. There also need to be explicit protocols and agreements for ensuring security, privacy, and confidentiality in the use of mobile phone records, for both operators and government stakeholders.

Providing access to mobile phone data to produce official statistics could be of benefit to operators. It can help foster regulatory goodwill. It can help demonstrate compliance and corporate social responsibility. It can be used as an opportunity to strengthen their internal capacity to analyze Big Data for business decision-making. It can improve the timeliness and quality of publicly available datasets and statistical indicators that can be used as inputs to guide business strategy. With the right approach, providing access to mobile phone data can be a winning scenario for data providers.

Telecommunication regulators. Partnerships to access mobile phone data will be more effective when telecommunication regulators are there to liaise with operators and facilitate access to mobile phone data. In some cases, they may already have the mandate to request mobile phone records as a requisite for monitoring licensing conditions, and may also have already invested in equipment and expertise to store and process these records. If so, they will be important conduits for NSOs to gain access to mobile phone data to produce

official statistics beyond the telecommunication and ICT sectors. Even if they do not already collect mobile phone records, they are the government agency that frequently interact with operators in the course of their regulatory work. As such, they are likely to be well positioned to negotiate and mediate access to mobile phone data.

The appropriate role of regulators in the partnership will depend, among other things, on the amount of resources at their disposal, their relative level of technical expertise, and their legal mandate to compel submission of mobile phone records. In countries where NSO resources are overstretched and technical capacity is limited but where the regulator is already actively collecting and handling mobile phone records, the initiative to set up the partnership may fall on the regulator. However, once the partnership is established and the pilots completed, leadership and management responsibilities can be turned over to the NSO whose core business, after all, is official statistics. The key thesis is that partnerships are more sustainable and productive when regulators are committed participants, regardless of whether or not they are playing the lead role.

Data privacy authorities. Countries are increasingly legislating data privacy rules and setting up national data privacy authorities. As of 2017, about 90 countries have promulgated at least some type of data privacy rules enforced by data privacy authorities or other designated government bodies (DLA Piper 2017). Although, the specific content of privacy laws vary across countries, these can be classed under one of two distinct legal paradigms: continental European-style data protection laws or Anglo-Saxon style data privacy laws (Determann 2015). The former is often more comprehensive and restrictive compared to the latter. Under continental European-style laws, the default approach is to forbid and minimize the processing of data relating to identifiable humans except when there is a justified exemption. Data processing for statistical, scientific and other public interest purposes are generally allowed, provided safeguards are in place, including whenever feasible, pseudonymization or anonymization. In contrast, Anglo-Saxon style laws focus on protecting individuals from interception of confidential communication. This implies that unless there is reasonable expectation of privacy, there are no explicit restrictions on data processing.

Partnerships to access mobile phone data will have to take into account the specific rules in force and work together with the data privacy authority, when there is one. In providing guidance and oversight for lawful data processing, the data privacy authority must strike a balance between the public interest value in statistical uses of mobile phone data and the perceived privacy risks.

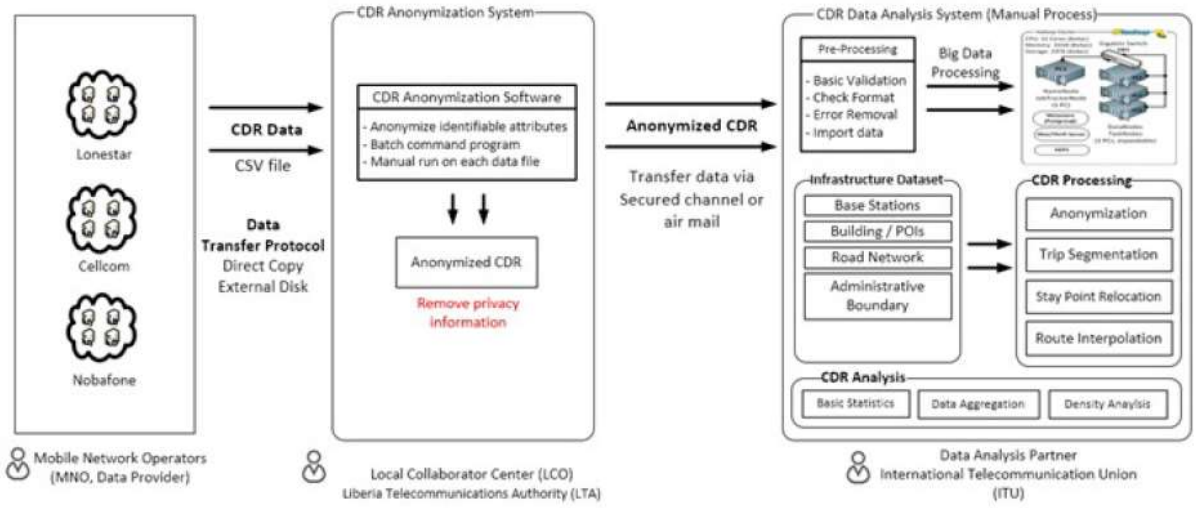
4.3 Case studies

4.3.1. ITU Ebola project

In the last outbreak of Ebola in West Africa over 11 000 people died. National-health authorities in West Africa have struggled to contain Ebola, especially how to estimate the affected population in the outbreak prone area and how to understand outbreak patterns spreading from one area to another. Currently, population movement simulations are based on statistical data that cannot provide reliable and dynamic population flows from an area where an outbreak has occurred, nor can they predict where outbreaks were likely happen next. This type of simulation can be better served by using call data-records (CDR) generated from mobile phone data. The CDR data contains time and location information for voice, messaging, and data communication of each handset – collected for billing purposes – and it can also be used to estimate time-based and up-to-date dynamics of population movements, leading to better support and preparation and more lives saved.

In 2015, ITU launched a big data project to showcase the potential of big data to facilitate the timely exchange of information to combat the Ebola epidemic - which had gripped West Africa in 2014 - and future health crises. The project used Call Detail Record (CRD) data, which includes information on the use of the mobile phone, including the location, from mobile network operators in Liberia, Guinea and Sierra Leone. The project demonstrated how analyzed CDR data can provide information on human mobility, including cross-border movement, and the spatiotemporal distribution of people, while safeguarding individual privacy. In the case of the outbreak of a disease this information is critical for governments as well as for humanitarian aid agencies, for effective intervention, and to tackle the disease. It can further be used to build models of population flow patterns over time, and at specific events, and to combine these data with other information.

For each of the three countries, two months of CDR data – from June to July 2015 – were collected to demonstrate how dynamic population movements could be estimated. In Guinea, data was prepared by the mobile network operators: Cellcom, Intercel, MTN and Orange. In Liberia, data was prepared by the mobile network operators: Lonestar, Cellcom and Novafone. In Sierra Leone, data was prepared by the mobile network operators in Sierra Leone, Africell, Airtel and Smart. This meant that in all three countries, data included the majority of the mobile subscriber population. Subscriber privacy issues were addressed by the creation an anonymous set of information. The unique identification number was replaced using a cryptographic hash algorithm that generated new random numbers and there is no way to rebuild the original identities. Not all operators prepared the data in the specified format, and in these cases, data pre-processing was necessary.



Source: ITU

The process of CDR analysis for the ITU Ebola project required steps that could eventually be automated, including the use of anonymization software, which has been developed and shared to data providers to remove personally identifiable information from their data sets, so that the people remain anonymous. Data transfer was also carried out manually online using secure channels or offline using external disks. Details of each step and module is described below:

• **Mobile network operator (MNO):** Data is collected from operators. In this case, the mobile operator collects position data of the mobile device stored on its server. Data is then exported and transferred to local collaborator centres (LCOs) for further processing, generally provided in a compressed CSV file format. For the purposes of this project, data transferring was carried out via direct copy.

• **Local collaborator centre (LCO):** This country specific unit stores and sanitizes data (eliminates the risk of personal data disclosure). Usually, this role is carried out by the regulator or mobile operator licence provider, which is also in charge of transferring anonymized data to a designated data analysis partner.

• **Data analysis partner (DAP):** The role of the DAT is to keep and maintain all sanitized CDR data received from operators through the LCO, and is in charge of processing and analysing CDR data:

– CDR anonymization: This function handles the anonymization process on CDR data. The LCO retrieves raw CDR data in csv format and manually processes it to remove all privacy related information. The CDR data can then be transferred to the data analysis partner. In this project, a macro-programme was run (a command line application) once certain parameters such as path of input, path of output, seed data and CDR format parser had been set.

– CDR data analysis: This function collects all CDR data from the LCO, manually checks the data and imports it to a big data platform before it carries out deeper analysis. This analysis

by DAP staff incorporates many modules including: pre-processing, big data processing, and CDR processing.

4.3.2. ITU project on big data for measuring the information society

ITU is currently undertaking a pilot project to use big data from the telecom industry to improve and complement existing statistics and methodologies to measure the information society. The results of this project are expected to help countries and ITU to produce official ICT statistics and to develop new methodologies that combine new and existing data sources. The following six countries are participating in the pilot: Colombia, Georgia, Kenya, Philippines, Sweden and the United Arab Emirates. The pilot projects are still ongoing; however, data has already been successfully accessed through either a Type A or a Type B partnership model as shown in Figure 4.6.

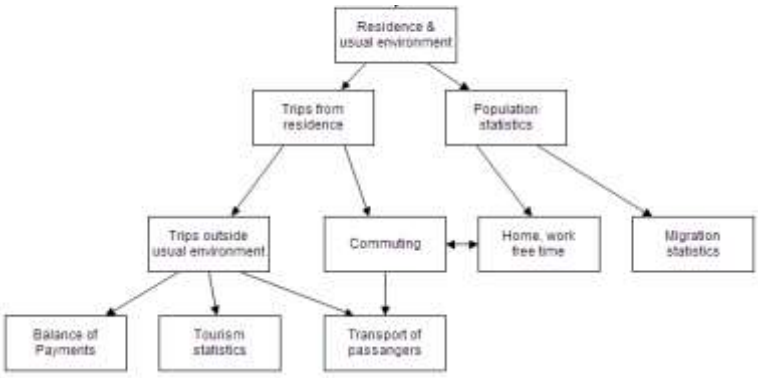
Figure 4.6. Accessing mobile phone data in six pilot countries: ITU's big data project for measuring the information society

Country	Partnership Model	Responsibility for Data Processing
Colombia	B	National Statistics Office
Georgia	B	Telecom Regulator
Kenya	A	Network Operators
Philippines	A	Network Operators
Sweden	B	ITU
United Arab Emirates	A	Network Operators

5. Methods

5.1. Concepts and definitions

Mobile network data, in particular mobile positioning data, may be used in several statistical domains, which would benefit from a common conceptual framework. Statistical domains previously identified as possibly benefiting from the use of this data source include Tourism statistics, Balance of Payments travel item, Tourism Satellite Account, Passenger Transport, Population, Migration and Commuting Statistics.



In economy and finance statistics, mostly the Balance of Payments is relevant for assessing synergies with tourism statistics. Statistically, 'tourism' is a subset of 'travel' and consequently, (tourism) 'visitors' is a subset of 'travellers'. Moreover, the issues for tourism activities, namely the problem to accurately delineate usual environment, does not apply to the travel item in the Balance of Payments. Methodologically, during the processing of data, several objectives can be achieved depending on the system's setup. For example, although domestic tourism concentrates on travel outside the usual environment, the same process can be extended to identify any trips that are taken within the usual environment.

For calibrating transport demand and organising transport supply, it is very important to have accurate estimates of origin-destination matrices. However, it is quite difficult and very costly to obtain these matrices through conventional survey methods. Therefore, MNOs can provide less costly and much more accurate matrices from mobile positioning data. However, the data will not show the mode of transport or the purpose of the trip.

In population statistics, the spatial distribution of population (living, working) and mobility aspects such as commuting will be relevant. As opposed to census-based statistics, mobile positioning data will always lack accuracy and will not offer the required level of detail. Nevertheless, the data is timely and can provide overall indications concerning the commuting, migration and internal migration information.

The implementation of a system of statistics production based on mobile positioning data is rather expensive; however, if the system is implemented for several domains (tourism activities including BoP, transportation and population), the additional costs for adding processing components is relatively lower.

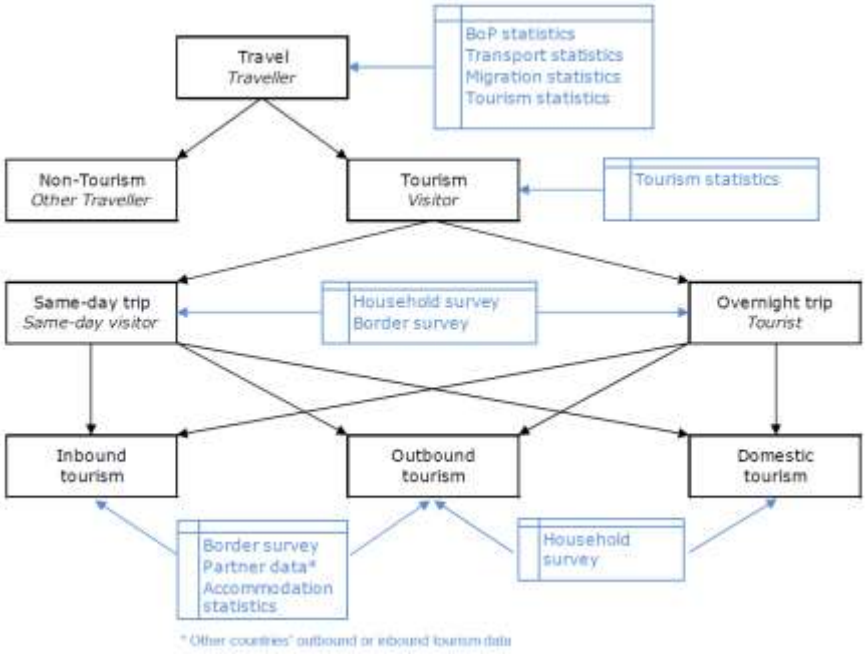


Figure 5.1. Scheme illustrating concepts in tourism statistics and the sources in which these concepts are used

Travel - refers to the activity of travellers (which is similar to the official definition).

Traveller - someone who moves between different geographic locations, for any purpose and for any duration (which is similar to the official definition).

Tourism - the activity of visitors who are taking a trip to a main destination which is outside the usual environment, which lasts less than a year, and which is for any main purpose, including business, leisure or other personal purpose, other than being employed by a resident entity at the location that has been visited. The main characteristics are similar to the official definition, although persons who are employed by a resident entity in the location that has been visited cannot be excluded from mobile positioning data.

Visitor - a traveller taking a trip to a main and/or secondary destinations outside their usual environment, for less than a year (differs from the official definition - see Section 3.1).

Tourist (overnight visitor) - a visitor whose visit includes an overnight stay (which differs from the official definition - see Section 3.1).

Overnight visitor/visit/trip - as for a tourist, the term is used to specifically distinguish visits that focus on the duration of the stay at a specific place in a country (new concept).

Same-day visitor/visit/trip (excursionist) - a visitor whose visit does not include an overnight stay (similar to the official definition).

Inbound tourism (tourism form) - comprises the activities of a non-resident visitor within the country of reference on an inbound trip (similar to the official definition).

Domestic tourism (tourism form) - comprises the activities of a resident visitor within the country of reference either as part of a domestic trip or part of an outbound trip (similar to the official definition).

Outbound tourism (tourism form) - comprises the activities of a resident visitor outside the country of reference, either as part of an outbound trip or as part of a domestic trip (similar to the official definition).

Internal tourism (tourism category) - comprises domestic tourism and inbound tourism, that is, the activities of resident and non-resident visitors within the country of reference as part of domestic or international trips (similar to the official definition).

National tourism (tourism category) - comprises domestic tourism and outbound tourism, that is, the activities of resident visitors within and outside the country of reference either as a part of domestic or outbound trips (similar to the official definition).

International tourism (tourism category) - comprises inbound tourism and outbound tourism, that is, the activities of resident visitors outside the country of reference either as a part of domestic or outbound trips and the activities of non-resident visitors within the country of reference on inbound trips (similar to the official definition).

Country of reference - a country that is linked to the forms (inbound, domestic, outbound) and categories (internal, national, international) of tourism. Outbound tourism from the country of reference to a foreign country is inbound tourism for a foreign country (new concept).

Foreign country - a country outside the country of reference in respect of the forms (inbound, outbound) and categories (internal, national, international) of tourism. Outbound tourism from a foreign country to the country of reference is inbound tourism to the country of reference (new concept).

Trip - refers to the journey of an individual from the time at which that individual departs from their place of residence until they return; it therefore refers to a round trip. A trip is made up of visits to different places. Trips consist of one or more visits during the same round trip (similar to the official definition).

(Tourism) visit - refers to a stay in a place visited during a tourism trip. The stay does not need to be overnight to qualify as a tourism visit. Nevertheless, the notion of stay supposes that there is a stop. Entering a geographical area without stopping there does not qualify as a visit to that area. It is recommended that countries define the minimum duration of stops to be considered as being tourism visits. The concept of a visit depends on the level of the geography in which it is used. It can mean either the whole tourism-related trip or only a part of it, depending on the perspective (origin-based or destination-based) (similar to the official definition).

Overnight stay - the criterion to distinguish tourists (overnight visitor, overnight visits) from same-day visitors. A visitor is considered to have had an overnight stay/visit in a place if the

visitor is believed to have stayed there during a change of calendar dates (a place in which a night is spent regardless of the actual rest/resting place). If during a change of dates, a visitor is in the middle of moving between Points A and B within a country of reference, a night might be assigned (depending on the national criteria of the specific country) to Point A, Point B, or it might not be assigned at all. However, from the perspective of country (the place is the country of reference), a visitor spent a night within a country (which differs from the official definition - see Section 3.1).

Trip section - a trip consists of the stay and movement sections. All sections (stay and movement) are aggregated into stay sections if viewed from a higher geographical level. Stay section in place A; movement section between A and B; and a stay section in place B in country X combine a single stay section when viewed from the country level (new concept).

Duration of the trip/visit/stay - mobile positioning provides a means to measure the duration of the visit in total hours, days present, nights spent. Duration of travelling to and from the destination can be identified and excluded. Total hours and nights spent per trip can be summarised for all aggregation levels; however, days present cannot be summarised (which differs from the official definition - see Section 3.1).

Country of (usual) residence - the country in which a person spends the majority of the year.

Place of (usual) residence - the geographical location of the person's place of residence. In cases in which this is a foreign country, the place of residence is the country of residence as it is not possible to determine a more accurate/specific location of the residence within a foreign country when using mobile positioning data. In the case of the country of reference, the place of residence is a specific location within the country of reference with accuracy depending upon the method of identifying the location's actual point (e.g. the smallest administrative level, the smallest identifiable grid unit, or geographical point).

Usual environment - each form of tourism has a specific definition and method of defining the place of residence and usual environment. By default, the place of residence and usual environment for subscribers of inbound data is the foreign country of the subscriber unless identified differently. For domestic and outbound subscribers, usual environment can be defined with precision of country of reference, county, municipality or some other geographical areas or administrative units. The level of detail used depends on the data available and producers' needs (which differs from the official definition - see Section 3.1).

Main destination - the main destination of a tourism trip is defined as the place visited that is central to the decision to take the trip. However, if no such place can be identified by the visitor, the main destination is defined as the place at which they spent most of their time during the trip. Again, if no such place can be identified by the visitor, then the main destination is defined as the place that is the farthest from the place of residence. In mobile positioning data, a distinction between the main destination for a trip (which is similar to the official definition) and a secondary destination (a new concept) has to be made. The main destination for the trip can be identified using the official criteria; however, during overnight trips, each day might have a different main destination (which is usually the one at which the

night is spent). On a same-day trip, the main destination for a trip is the place in which most of the time was spent. A visitor can visit one main destination and several secondary destinations during one day.

Secondary destination/visit - as opposed to the main destination, a secondary visit is a place to which a visitor makes a visit (stays) in addition to the main destination for a period longer than the minimum duration of stops to be considered as being tourism visits (new concept).

Transit pass-through - as opposed to main destination and secondary destination, a transit pass-through is the place that visitors pass through or stop during a period of time that is less than the minimum duration of stop to be considered as being tourism visits. A transit pass through does not count as a tourism visit. At a country level, transit pass-through or transit trips/visits are considered as being trips for which the purpose is passing through that country on one's way to or from the country that is their main destination (similar to the official definition).

5.2. Data processing methodology

The methodology consists of the following phases: the additional preparation of event data, frame formation, data compilation and estimation. The initial data that is extracted and prepared by MNOs is based upon network events that specify a specific subscriber's presence in time and space. Additional preparation may include geographical referencing, the elimination of non-human-operated mobile devices, checking the time and area coverage of the data, dealing with missing values, etc. After the data has been prepared by MNOs, the following processing steps are set out:

- Frame formation:
    - The application of trip identification algorithms - identifying each subscriber's individual trip to the country in question with the start and end time for each trip;
    - Identifying the population of interest (distinguishing tourism activities from non-tourism activities):
        - Defining roaming subscribers not actually crossing the border and entering the country (inbound, outbound);
        - Defining residents (inbound, outbound);
        - Defining the place of residence and the usual environment (domestic);
        - Identifying country-wide transit trips (inbound);
        - Identifying destination and transit countries (outbound);
- Data compilation:
    - Spatial granulation (visits at the smallest administrative level for inbound);
    - Defining variables (number of visits, duration of trips, classification, etc.);
- Estimation (from an MNO-specific sample to the whole population of interest) contains:
    - Time and space aggregation of the data (day, week, month, quarter/grid-based (one km2), LAU-2, LAU-1, country);
    - Combining data from various MNOs and computing final statistical indicators.

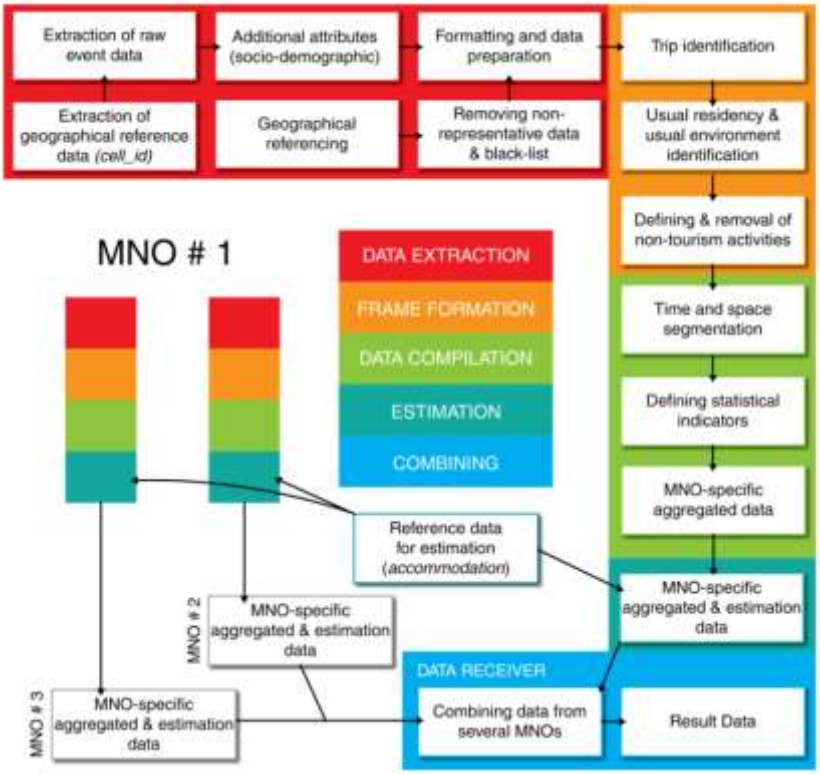


Figure 5.2 Data processing steps

The specifics of methodology might depend upon the characteristics and origin of the data (time and space frequency, the geographical accuracy of the events, and the available attributes of the events or the subscribers). The processes described in Report 3a (trip calculation, identification of usual environment, non-residents, transit trips, duration of stays, etc.) assume that longitudinal calculations for single subscribers are possible.

Vital procedures that are carried out during the frame formation process include the identification of the county of residence and the usual environment of the subscribers in order to be able to define tourism trips (i.e. trips outside the usual environment). These calculations require the historical time series (longevity) and interconnectivity between the different data forms (same subscriber ID in domestic and outbound data) for the subscribers in order to be able to define frequently visited places and countries.

This description of methodology in Report 3a could be taken as a step-by-step guideline how to produce tourism statistics as it is rather detailed but at the same time general enough for broad use. There is an assumption that several variables relevant for the forming of new variables exist in the dataset and that the activities of anonymous subscribers can be followed over longer period of time to establish their residency and/or usual environment. The assumption regarding availability of longitudinal data is crucial for the production of tourism statistics. Algorithms identifying the usual environment and country of usual residence rely upon the availability of past data. If the available data describes only a short period of time

then the issue of processing errors arises and, as simulations show, the data quality can be too low to produce reliable results. Such limited data can be used as comparison indicators in some unofficial domains (e.g. the number of unique foreign subscribers on the site of attraction or concert) and for relative comparison.

5.3. Quality assessment of statistics based on mobile network data

5.3.1. Populations observed in mobile network data

To give context on how selectivity could be measured in mobile network data, we start with the presentation of the relation between populations. Figure 1 presents connections between the CDR population denoted by $\mathbf{\Omega}_{CDR}$ with associated frames denoted by $A_{CDR,1}, A_{CDR,2},$, the BTS population denoted by $\mathbf{\Omega}_{BTS}$ with associated frames denoted by $B_{CDR,1}, B_{CDR,2}$, the Mobile Numbers (MN) population, the Mobile Users (MU) Population and the Target population. Dashed lines refer to different frames observed in the above populations. For instance, $A_{CDR,1}$ refers to the frame of CDR in, say Orange, and $A_{CDR,2}$ refers to, say T-Mobile. Naturally, these two frames will overlap because people can call each other and phone numbers will be present in both frames. In terms of MN population, the overlap is presented due to a possible move of mobile phone numbers between providers. For instance, at the beginning of a period Orange could be observed, while at the end it could be T-Mobile. Solid grey lines refer to connections between objects observed in populations and shapes refer to either objects or statistical units. Black shapes represent objects/units observed that are in frames, grey shapes refer to objects/units that are in frames but were not observed and finally white shapes denote objects/units that are not covered by frames. For instance, in a given period, not all BTS might be present in the data that were made available to statisticians. Links between the mobile phone users' population and the target population present cases of over-coverage and under-coverage. White squares represent the MU population that is not observed in available frames (e.g. uses small provider).

The relation between populations gives context on how these data are created and geo-located. Each row in the call detail record has an assigned BTS. This relation gives spatial context and enables to assign a given mobile phone to a certain location. An important issue from a statistical point of view is the fact that these relations are complex and that multiple frames are observed. However, from a practical point of view, it is unlikely to obtain data from more than one provider which limits the analysis to the problem of a single and imperfect frame.

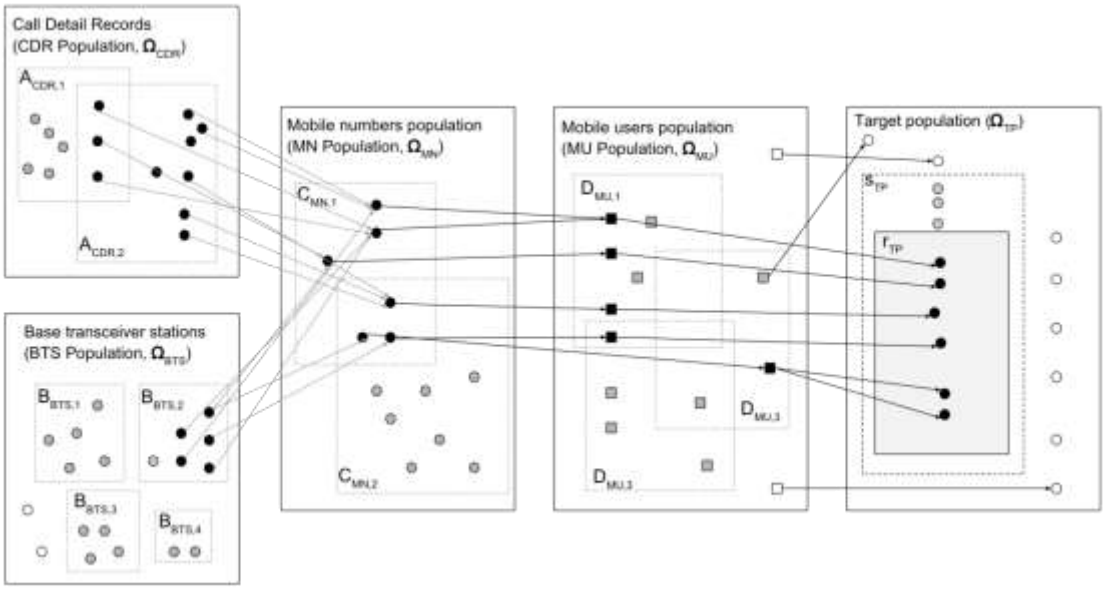


Figure 5.3 Relation between populations of different statistical units in mobile network data

Figure 5.4 presents the relation between the target population and the final sample that is obtained. Here we distinguish the population with phones (any kind) and the population with mobile phones. The observed population Ω_MP is the population that is observed within mobile frames and s_MP is the sample that we have access to (e.g. limited to a certain period). We do not specify that Ω_MP=s_MP because the sample observed is limited to members of the target population only, while Ω_MP might have an overcoverage (e.g. objects that are not referring to any population member). Finally, we end with r_MP because for certain units we might observe missing data in the target variable.

It might happen that, for all units from the observed sample, we do not have information on a target variable (e.g. resident location) which must be imputed. In this case, we might consider the problem of latent variables which are not observed directly but indirectly (e.g. by analysis of location by BTS stations).

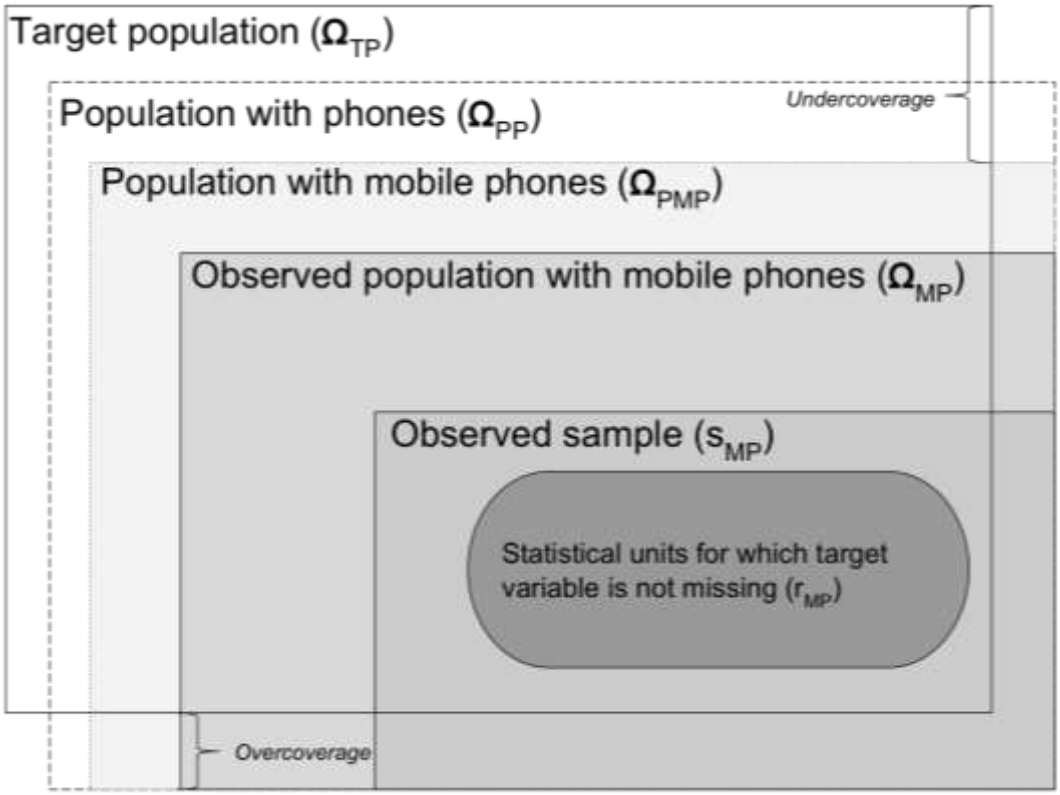


Figure 5.4 – From target to observed population in mobile network data

Finally, we would like to justify why self-selection (selectivity) is presented in terms of sub-setting the target population. Regardless of the target population definition (e.g. trips, persons), some units will always be left out of big data limit. For instance, the coverage of mobile infrastructures limits the exact identification of trips; however, this should be classified as a coverage error. Observed trips are a subset of the population of all trips (e.g. limit to one provider, only to those with mobile phones) and trip can be identified ($R_i = 1$) which is also related to infrastructure coverage.

5.3.2. Assessing coverage and selectivity

The majority of mobile phone populations refer to objects rather than directly to the target population. Mobile operators are in possession of systems that provide background on the structure of their clients (cf. CRM). However, in practice, the analysis is limited to the population of CDR, which is further related to SIM cards (numbers) and infrastructure (BTS).

Figure 5.5 presents hypothetical self-selection mechanism observed in mobile network data. Here we distinguish three phases that refer to using the mobile phones/devices which refer to coverage error (access to phone). The second phase refers to selecting a given mobile provider and tariff. And finally, we refer to the sample of pseudo-respondents for which we have a response (observed directly or indirectly). We should point that if there will be no data within a given time frame there will also be data missing.

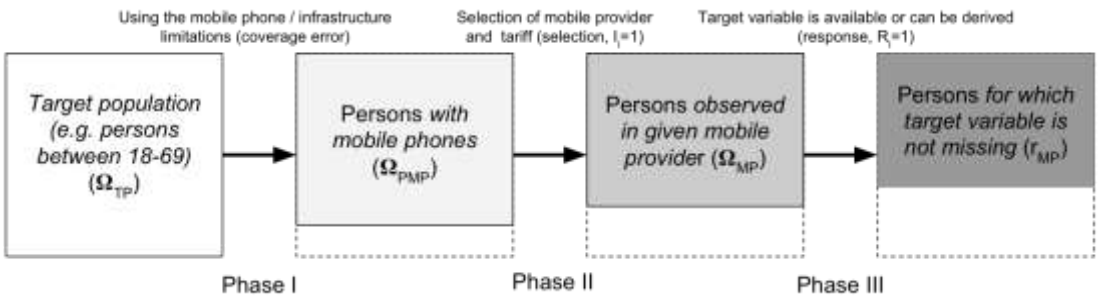


Figure 5.5 – Self-selection mechanism in mobile network data

Hence, from a practical point of view we would like to answer the following questions:

- what is the overall coverage of the target population in mobile operator frame (population coverage)?
- what is the overall coverage of the infrastructure coverage? (coverage by BTS and cells)

The first question refers to the coverage of the target population (regardless its definition). One should take into account that the populations observed in mobile network operators refer to businesses and natural persons which might overlap (e.g. one person with both a private and a business mobile phone). The second question concerns the selectivity of the stations and at which level of aggregation these data can be used. Certainly, the BTS is located according to the density of the population and in the literature it is mainly presented using Voronoi diagrams (Ricciato et al., 2015).

5.3.3. Selectivity of infrastructure - BTS and cells

To assess the selectivity of the infrastructure it is important to understand how it is created. The infrastructure of the telecommunication companies' network can, on the most basic level be divided into:

- BTS – Base Transceiver Stations, which handles the communication within a geographic area and can have multiple directional antennas linked to one BTS (each BTS has an assigned exact location),
- cell-geographic area covered by the same antenna, where each BTS is responsible for one or more cells.

BTS have different technologies and might change overtime. For instance, in Poland we observe changes in the technologies from standard GSM to LTE. The modernisation of the infrastructure introduces additional coverage problems. However, due to competition on the market, MNO are extending the number of BTS or upgrade the existing ones to the new technologies.

Telecommunication companies can only operate on limited radio frequencies and if all the frequencies are already taken by other users' calls or other services in the target area, then the next incoming calls will be handled by a more distant BTS. This situation is called a hand-off. BTSs use a low power signal which, in conjunction with the use of multiple frequencies, allows the re-usage of radio frequencies without signal interferences between the stations. As only certain frequencies are available for the operators, they need to adjust the placement of

the BTS, so they don't interfere with themselves. The areas covered by the stations can be divided by the usage of directional antennas working by a certain angle.

Another important issue relates to the area covered by the BTS which is defined as a cell. The shape of the cells are not consistent with the administrative and statistical areas. In most cases, these cells are created using voronoi diagrams which represents the area covered by one and only one BTS. An example diagram with voronoi diagrams with cells referring to BTS is presented below. These cells present hypothetical coverage of a given area by 50 points. The coverage of BTS is correlated with the population density which means that cities are highly covered by antennas Compared to rural regions.

Because of the infrastructure of BTS it is not possible to provide the exact location of a mobile device. To resolve this problem, we could consider using Wi-Fi or cell signals as Facebook does. These signals might provide the location with more precision (e.g. 100 meters).

Another issue is the existence of BTS frames in general. For example, according to the Polish regulations, only a list of approvals published by the Office of Electronic Communications is available. The data presented in such lists do not refer to actually working BTS. The final number of active BTS are in the possession of the MNO.

Not all BTS are working at all time. For instance, in Poland there are several stations that are enabled only on special occasions (e.g. football matches, concerts). Therefore, the main idea is to verify what fraction of BTS are working continuously. This might enable to verify the possibility of longitudinal observations.

To measure the coverage and selectivity of BTS and cells one should:

- provide a list of all BTS in a given country (if available),
- provide a list of all active BTS of a given MNO (if available),
- measure the coverage of cells (voronoi diagrams) with regard to existing statistical and administrative regions,
- create a list of BTS working constantly or from time to time.

5.3.4. Self-selection process on mobile phone market — Can it be ignored?

As we stated previously in the second part, we treat mobile network data in the context of opt-in panels. The motivation is that each person decides whether to have a mobile phone and furthermore decides which mobile provider to select (including tariff). Of course we might have situations where the decision to select a provider is not independent (e.g. within marriage). Finally, the self-selection related to the target variable is the usage of the mobile phone which directly influences the possibility to measure mobility.

In general, we summarise potential outcomes of the self-selection on the mobile market by:

⇒ The possession of mobile device (one or more),

⇒ The selection of mobile provider (one or more),

⇒ The selection of the tariff (contract, prepaid, other) — does prepaid and contract users differ?

⇒ The selection of device type (smartphone, standard cell phone),

⇒ The usage of a mobile phone data (one or several SIM cards).

The main issue in the self-selection study is to verify whether the mechanism is ignorable or not, with respect to auxiliary variables. To answer this question one should imagine situations in which the target variable is not recorded by a mobile device. Let's assume that the target variable is mobility. The question is: when will such actions not be recorded? We provide below possible examples assuming a full coverage of mobile phones and no attrition:

- people forgot to take their mobile phones — this might be considered as a random event, however one should take into account the scale of such situations
- people do not take their mobile phones on purpose — this cannot be treated as a random event. Another question that could be asked is the reason for such situations to occur?
- people who are not addicted to mobile phones (e.g. older people),
- people who do not want to be located (e.g. criminals).

However, having in mind how important mobile phones are for young people as well as for the population in labour force age, these situations might be very unlikely.

Now, considering the fact we have access to the data from one mobile provider. Once more we consider the context of an opt-in panel. We start with a general question — when is a missing data included in the target variable (e.g. mobility)? We could explain this issue in the context of a panel attrition. In particular we could consider the following cases:

- churn – a mobile user switches to another company/provider on purpose, this could be considered as a *voluntary attrition*,
- stop using a mobile phone — this could be considered as 'churn' (migration to another provider but without providing the information), not addicted to mobile phones (*passive attrition*) or the death/relocation of a customer (*mortality attrition*, a given unit is out of the target population).

To sum up, we could consider the term *churn*, widely used in marketing, as a *voluntary attrition*. However, we argue that it might not be connected to the target variable (e.g. mobility, economic activity) but to financial or private factors which do not influence the target variable.

Moreover, the mobile phone market selectivity has two main causes: (1) limitations of infrastructure and (2) individuals. The first cause brings problems identifying trips (if trips are regarded as the target population) where the infrastructure is sparse or in the case of overcrowded areas due to technical issues. The second cause refers to decisions of individuals; whether to use a given mobile provider and mobile phones. Hence, we should carefully study these two sources to distinguish MAR or MNAR mechanism.

5.3.5. Limitations of inference

Now we will summarise the main limitations of inference based on mobile data users. First of all, the availability of background information on mobile device users might be related to the regulations of a given country. For instance, Poland recently approved the 'anti-terrorists' law that required the interdiction of selling pre-paid cards without checking and recording the

identification of the buyer.

Secondly, mobile phones are often registered to one person while other persons can be using it. For instance, to have a contract, one should have an ID and should be over 18 years old (in Poland). This problem also refers to pre-paid SIM cards that might be bought by one person but used by another.

These problems are referring to the concept of 'unit-error theory' that was introduced by Zhang (2011) in the context of register data. To sum up, we observe the following problems regarding inference, based on mobile data:

- ⇒ problems regarding the definition and derivation of populations observed in mobile network data,
- ⇒ limited access to background information on mobile devices users (place of residence, gender, age, LFS, marital status) — need for imputation (or profiling, as it is described in the literature),
- ⇒ identification of over-coverage in frame(s),
- ⇒ identification of statistical units,
- ⇒ identification of exact locations,
- ⇒ including uncertainty of imputation into estimates.

Statistics Quality Assurance Framework

Completeness	No complete coverage of any sector relevant for tourism statistics
Timeliness	Full integration and automatisation → Much quicker than traditional sources
Validity	No specific advantages/disadvantages
Accuracy	Advantages over traditional sources (Smaller sampling error, no memory gaps) – need to re-define 'usual environment'
Consistency	High grade of consistency in variation compared to traditional sources.
Resolution	Finer granulation of space and time → new possibilities (again, need to re-define 'usual environment')

Another, I suggest to add/mention also about the important to have QA/QAF in every stage/phase of MPD data (like in Estonia)

Annex 1 - Case Study: France

Introduction

In July 2015, Bank of France and the department responsible for tourism statistics at the French ministry of Economy[30] launched an experiment in order to study the possible use of mobile phone positioning statistics for the compilation of tourism statistics. The aim of this experiment was to assess whether mobile phone positioning statistics might progressively replace the traffic data currently used for the estimation of the non-resident visitors flows (tourists and one-day travellers) in France. This experiment has allowed French compilers to deal with issues related to big data: access to the data, regulatory constraints, methodology and quality of the estimates.

Context

France is one of the leading countries for tourism in the world, with a huge diversity in the origin of visitors, and rapidly evolving patterns. Receipts in the current account of the balance of payments exceed 40 USD bn per year, and are a significant component in the balance of the current account. Against this background one of the challenges for French compilers is to maintain over time the quality of tourism statistics (number of visitors and travel receipts). Currently, these statistics are based on exiting traffic data by transport mode combined with a border survey. The border survey includes a manual counting of travellers in airports and at a sample of road exit points aimed at estimating the proportion of non-residents travellers in the total exiting traffic. The by-country breakdown is then estimated through the questionnaires collected by the border survey. The compilation method is challenged by exogenous factors such as the importance of transiting traffic in hub airports and the hurdles to measure road traffic in an open border area like “Schengen”. In this context, mobile phone positioning data promises a potential for the purpose of estimating the number of non-resident visitors by country of origin. As at now, mobile phone positioning data is already being used for tourism statistics mainly by Estonia and by some French local administrations.

Needs of the French tourism NSO and Bank of France

The statistics needed by Bank of France and the French ministry of Economy are the number of visitors’ arrivals and the number of tourists spending a night in France, per month, by country of origin and if possible by transport mode. The border survey should in any case be necessary to provide estimates of visitors’ expenses.

Legal framework and cooperation between NSO and MNO

In France, MNOs are not allowed to save and to sell individual data derived from mobile phone positioning. MNOs are however allowed to collect and process individual data

[30] Bank of France is responsible for the compilation of the Balance of Payments (including the Travel item) while the Directorate for Enterprises (DGE) of the French Ministry for Economy is in charge of estimating tourism statistics.

whenever they respect legal legislation about privacy protection. This is why MNOs have started to offer statistical services to private companies and administrations. Bank of France and the French Ministry for Economy therefore launched a public procurement in June 2015 in order to purchase the needed estimates from one MNO. The MNO was asked to deliver the needed information on a monthly basis with a one-month delay during one year. Two MNOs made an offer and one candidate was chosen, taking into account, among other criteria, the capacity to adjust its own data for the market share. After the first year, we decided to pursue the experiment for one more year because the first year had not been enough to conclude the experiment.

In terms of intellectual property, the algorithms developed by the MNO and related to the transformation of the mobile phone signal into statistical data are kept secret and remain its own property while the statistics produced from this raw statistical data are the NSOs' property.

Design of the project

The design of the project consists in defining the behavioural criteria that match tourism: an overnight stay is defined as a presence between midnight and 6am; a tourist arrival is defined as an overnight stay, without being there the day before; finally a tourist departure is defined as an overnight stay followed by an absence the following night. As a strong proxy, the country of residence of a visitor is assumed to be the country of his SIM card.

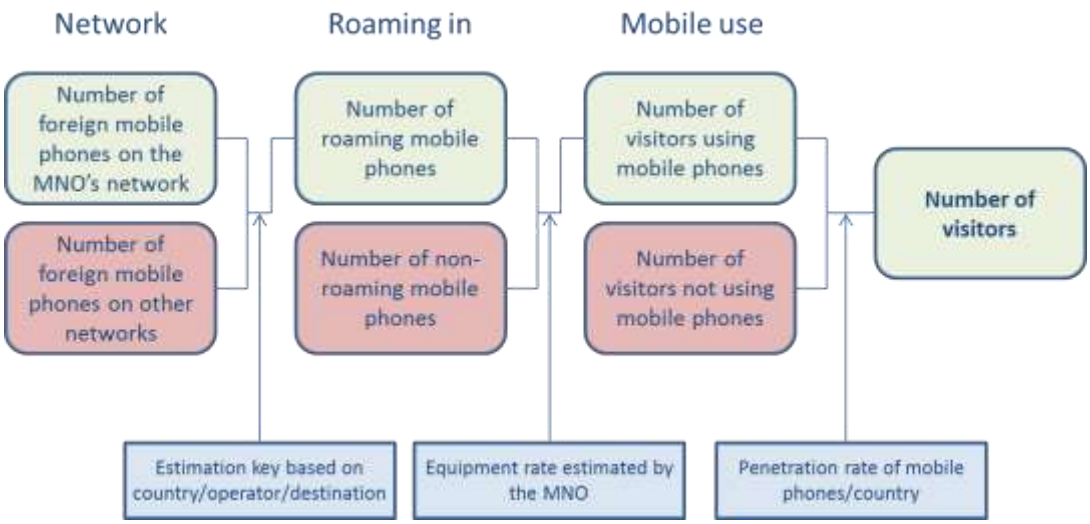
Data Sources and Methodology

The MNO builds an algorithm that takes as an input the individual data about connexions between mobile phones and antennas (SIM card, date and time, antenna, mobile phone activity[31]). The individual data is not saved: the algorithm works in real time and counts the number of people who meet the different criteria (arrival, overnight stay...). The main drawback of this methodology is that changes in the criteria cannot be backward deducted.

The algorithm includes different grossing-up that are necessary to estimate a number of visitors based on the observation of mobile phones. Some of these grossing-up, related to the MNO's market share on roaming activity, cannot be isolated and are part of the anonymization process. The other grossing-up factors are linked to mobile phones usage and ownership. They can be modified. This setup allowance appeared fundamental in practice, since MNOs have no precise knowledge of mobile phones habits among the population of foreign visitors. The equipment rate observed in the country of origin of the visitor is mainly used as a proxy for usage rate abroad; one should easily imagine how strong could be the gap between equipment rate of the whole population and the specific one of tourists travelling abroad for some countries.

Grossing-up: from a number of mobiles to the number of visitors

[31] This individual data includes both the passive signalling and mobile phone activity (calls, messages, etc.)



Quality

Issues and quality improvements

During the experiment, we first observed huge discrepancies between mobile phone positioning estimates and our survey estimates. Consequently, we worked regularly with the MNO in order to identify potential issues and improve quality. The main issues that affected the quality of mobile positioning estimates and the actions taken are summarized below:

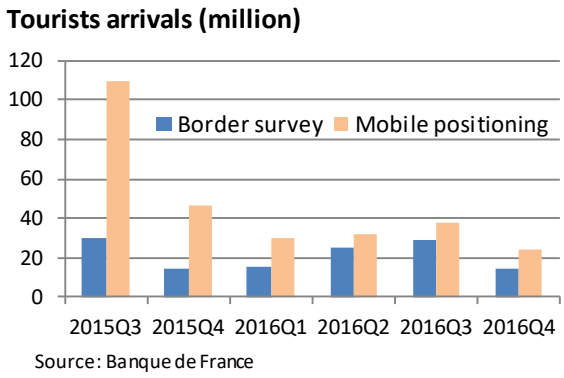
1. French residents may use foreign SIM cards (border workers for example) and can be identified as foreign tourists (the main examples are Luxemburg and Switzerland).
2. Random network switching: despite privileged roaming agreements between foreign and local MNOs, foreign mobile phones can change network, which introduces a bias in estimates due to the market share grossing up.
3. Mobile usage: the use of mobile phones among tourists from distant countries is not well estimated. Consequently, the grossing up led to unreliable estimates for countries like the United States, Canada and China. In fact, the high roaming costs charged to visitors from distant countries may favour the use of local SIM cards when travelling and no precise data is available to take this into account.
4. Wrongly interrupted stays: it has been observed that a large part of arrivals were due to visitors who already were in the country some days before. This phenomenon can be linked to effective stays (truck drivers for instance) but can also be caused by visitors who did not leave the country (change of network, mobile switched-off during a period…).

Different methodological changes have been adopted in order to take into account these phenomena:

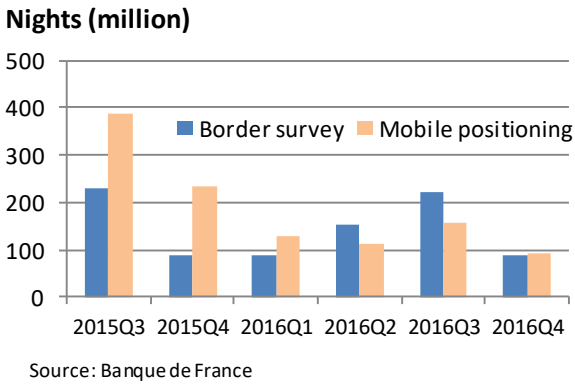
1. French residents with foreign SIM cards: a new criterion was introduced, assuming that persons who have spent 30 nights in France out of the past 60 nights are considered as residents. Because of the learning process, results were only made available late in February 2016.
2. Random network switching: the idea was to introduce different criteria in order to ensure that counted mobile phones are part of the real network users of the MNO. In November 2015, a criterion linked the cumulated connexion to the network was

added. Finally, in February 2016, a criterion linked to network usage was added (calls, messages...).
Current working axis is the selection of mobile phones whose foreign MNO has a privileged roaming agreement with the French MNO partner of the experience.

3. Mobile phone usage: in order to improve grossing up factors linked to mobile phone usage, some questions about this matter were added in our border survey. The first results will only be available in late 2017.



Source: Banque de France



Source: Banque de France

4. Interrupted stays: different conditions on the absence that has to be observed (from one to six days) before an arrival were experimented; no optimal criterion has been found yet.

Comparison between mobile positioning and border survey estimates

Mobile phone positioning estimates are not supposed to exactly match our border survey estimates, especially because mobile phone positioning was expected to improve the quality of our estimates. However, the comparison with our border survey estimates[32] is a way of measuring the reliability of mobile positioning estimates. As explained, we first observed huge discrepancies that were then reduced thanks to the methodological improvements described above. These methodological improvements, although suggested by the discrepancies with the survey, are not dependent on data from the survey. The number of nights appeared to be more reliable than the number of tourists' arrivals because the latter is more sensitive to interrupted stays. On the available period, we observe a convergence between mobile positioning estimates and our border survey estimates for aggregated indicators. However, there are still huge differences when it comes to country level figures. Basically, distant countries are currently underestimated by mobile positioning due to the bad knowledge of roaming habits for these countries. We expect our future survey results to help us to solve this issue.

[32] The estimates presented in this document are those of Bank of France for the purpose of estimating the travel item of the Balance of Payments. They can differ from the official tourism statistics published by the Ministry of Economy.

Tourists arrivals and nights (million)

		2015Q3	2015Q4	2016Q1	2016Q2	2016Q3	2016Q4
Arrivals	Border survey	29,5	14,5	15,3	24,6	28,5	14,3
	Mobile positioning	109,6	46,0	29,6	31,4	37,2	24,5
	spread	*272%*	*217%*	*93%*	*28%*	*30%*	*71%*
Nights	Border survey	231,2	88,1	87,0	154,7	223,6	86,6
	Mobile positioning	387,0	234,1	128,6	114,4	155,2	93,4
	spread	*67%*	*166%*	*48%*	*-26%*	*-31%*	*8%*

Source : Banque de France

Lessons learned

Among the advantages of mobile positioning, one can say that this data source is largely of some interest in particular, since it allows measurement of short term variations according that daily data is available. It is also at disposal with a short delay and is relatively cost-effective compared to the quarterly traffic data currently used.

As far as these properties (measure of short term variations, timeliness, cost-effectiveness) are concerned, it remains to be assessed to what extent mobile phone data brings something more than the available credit card data already used to capture short term trends (although monthly and not daily). However, these two data sources are not really substitutes as credit card data concerns directly expenses while mobile phone data aims at estimating a number of people and seems therefore better suited for tourism statistics.

Our view so far is that mobile phone data would bring a specific and undisputable advantage if reliable estimates of "number of nights by country of origin" could be derived. Crossing this data with information derived from the survey and credit card data could improve the process of counting visitors and estimating, country-by-country, more precise "expense per visitor" ratios.

At the same time, the experiment gave us the opportunity to observe that getting reliable estimates is time consuming. Therefore, after 18 months, Bank of France and its partner's still believe that mobile phone positioning estimates are not ready to replace their current compilation system. Should the extension of the experimental period not be successful, we should have to decide whether it is worth using this data as a complementary benchmark.

Important: The data and conclusions presented in this case study are communicated for the purpose of the Mobile Phone Data Task Team work. They should not be communicated outside this Task Team and may change before the final publication due to changes in figures and necessary approvals from the different institutions concerned by this case study.

Annex 2 - Case study: Indonesia

Introduction

BPS-Statistics Indonesia is the institution that produce and publish tourism data in Indonesia. Prior to the use of mobile positioning data (MPD), the main source of inbound tourist data is administrative data from the Directorate General of Immigration, which covered all main airports and entrance. Other data source is Cross-Border (Shuttle) Survey conducted by BPS in order to obtain tourists/visitors data at a certain border entrance/gate where the administrative data is not available or border survey cannot be conducted due to geographical condition.

In October 2016, BPS-Statistics Indonesia started to implement Mobile Positioning Data (MPD) for several border areas. The aims are to increase the coverage and to be able to accurately capture inbound tourist data. So, the MPD data is used as a complement of other sources, that are administrative data and survey.

Purpose

Tourism in Indonesia has been growing in a positive direction from year to year, it is showed by the increasing number of foreign tourists visiting Indonesia from time to time. As mentioned before, prior to October 2016, the inbound tourist statistics are based on administrative data and surveys. The survey is Cross Border (Shuttle) Survey and Passenger Exit Survey (PES). The Cross-Border Survey is conducted in the border entrances where there are no immigration check point, the entrance/gate is only guarded by army personnel's or head of village. However, Indonesia is quite vast area and many of the border is difficult to reach, so the Cross-Border Survey required a high cost especially for transportation.

Therefore, BPS-Statistics Indonesia started to use MPD in the data released in October 2016 in order to be able to accurately capture and increase the coverage of international visitor arrival. Data is obtained from one Mobile Network Operator (MNO) which has the highest market share at cross border area (around 92 percent at the border and 70 percent nationally).

The MPD is used to correct cross-border posts in 19 regencies in which Immigration check point is not available and the Cross-Border Survey is difficult to be conducted due to geographical difficulties or need a very high cost to be surveyed. By using the MPD technology, we can record and monitor foreign tourists' arrivals at the borders effectively and accurately.

For BPS-Statistics Indonesia, the use of Mobile Positioning Data (MPD) in tourism statistics is in line with BPS-Statistics Indonesia Strategic Planning 2015-2019 (Optimizing the Use of ICT) and the United Nation recommendation on data revolution that is stated in the UN Expert Advisory Group report called "A World that Counts". Data revolution is defined as an explosion in the volume of data, the speed with which data are produced, the number of producers of data, the dissemination of data, and the range of things on which there is data, coming from new technologies such as mobile phones and the "internet of things", and from

other sources, such as qualitative data, citizen-generated data and perceptions data. Data revolution for sustainable development means the integration of these new data with traditional data to produce high-quality information that is more detailed, timely, and relevant for many purposes and users, especially to foster and monitor sustainable development

Methodology

As mentioned above, the MPD data is obtained from (only) one of the biggest MNOs in Indonesia. Currently, BPS-Statistics get the MPD from Ministry of Tourism since there is MoU between BPS Statistics Indonesia and Ministry of Tourism. The arrangement is the MNO process the MPD to become statistics and give it to Ministry of Tourism (MoT), the MoT give to BPS Statistics Indonesia. The rule and filter were set up by BPS-Statistics Indonesia, Ministry of Tourism prepare the budget for MPD data processing. So, the raw individual data is remained with MNOs, the MNO gives the statistics to BPS-Statistics Indonesia every month with n-1 lag.

The filter used is that every mobile phone staying at least seven consecutive days or accumulative less than 20 days in the region (district) is consider as a tourist. If two or more number (IMSI) are always stay in the same area and close to eachother then it calculated as one tourist instead of two or three tourists. For inbound tourists/visitors, the MPD used is the signal data not the Call Detail Record (CDR) data. However, for outbound the MPD used is the CDR, since there is no signal data.

The statistics provide by MNO is a table which contain district name (proxy of place of entrance), country (from SIM card/MCC) and number of tourists/visitors. BPS-Statistics Indonesia is then compare the statistics with data from Immigration and Border Survey (if available), and the difference is added to the data or the MPD data is used whenever there is no data from immigration or no survey at the border entrance.

Timeline

Project time-line for 2016-2017

Activity	Time
1. Meeting to use MPD for Tourist data at the Border	July 2016
2. Meeting on deciding the MNOs to collaborate	August 2016
3. Discussing Methodology and presenting data calibration	September 2016
4. Deciding Methodology and Border Area that will use MPD	October 2016

5. Implementing MPD to October data	November 2016
6. Discussing and drafting MoU between BPS-MoT and MoCI (Based on Statistical Act No. 16/1997)	December 2016
7. MoU between BPS-MoT and MoCI (Based on Statistical Act No. 16/1997) is signed	January 2017
8. Presenting MPD Methodology and Implementation to Statistical Society Forum	February 2017
9. Initiating and Planning the Mobile Phone Usage Survey at the Border	March 2017
10. Workshop on the Use of MPD (Participants: Provincial and Regency Statistical Office, Ministry of Planning, MoT)	April 2017
11. Preparing the Mobile Phone Usage Survey (Enumerator Training etc)	May 2017
12. Conducting the Mobile Phone Usage Survey (Field work)	June 2017

Future plans (What is next ?)

Activity	**Time**
1. Discussing and drafting the Technical Partnership Agreement for Mobile Positioning Data Access	2017
2. Signing the Technical Partnership Agreement for Mobile Positioning Data Access	2017
3. Technical Workshop/Consultation on MPD	August 2017
4. Developing MPD Methodological Handbook for Tourism Statistics	2017

5. Memorandum of Understanding with one MNO	2017
6. Technical Workshop/Training on Data Processing	October 2017
7. Preparing Server, Network etc for data exchange	2017
8. Developing QAF for MPD	2017

Advantages and Disadvantages of MPD

Mobile positioning data is considered as one of the most promising ICT (Information and Communication Technologies) data sources for measuring the mobility of people, including mobility of tourists. Almost every person on the planet now has a mobile phone that he/she uses to communicate, access internet, conduct everyday banking and entertainment. The digital footprint left by the users is very sensitive, but also highly valuable, as it provides new possibilities to measure and monitor the spatio-temporal activities of the population. For statistical purposes, mobile data positioning provides new possibilities in terms of quality of the data as well as new opportunities. Statistics based on Big Data i.e. Mobile Positioning Data can be compiled automatically, in some cases almost in real time and requires less manual labour. Obviously, the job of analyzing and interpretation of the statistical indicators is left for statisticians and researchers, but the new concept of fast and expansive data collection should improve the quality of decision-making process and results in public and private sectors.

In order to optimally benefit from the MPD technology, we also need to understand that this technology also has its own strengths and weaknesses. For instance, on one side, MPD has fairly good consistency over time for the number of trips and nights spent compared to data based on ‘traditional’ methodologies. It also improves the timeliness of statistics (up to near-real time) and the possibility to use mobile data as unconfirmed quick indicators. It is also more accurate, since the survey is only conducted one week to estimate a month data. There will be an estimation problem if there is peak or low inbound visitors. On the other hand, we might also face difficulty in assessing the quality of statistics based on the MPD because mobile phone usage during travel is largely unknown and if the methodology is not firm. There might also relatively lack of information on the purpose of the trip, expenditure, type of accommodation and means of transport used.

As a new data source, there are many challenges that have to be taken into account by BPS-Statistics Indonesia in order to be able to produce valuable and high-quality statistics. The main challenge is the access to the data where legal, administrative, and commercial barriers have to be overcome.

In terms of legal, BPS-Statistics Indonesia has Statistical Law (Statistical Act no. 16/1997) that mentioned about cooperation of BPS and line ministries or institution. Therefore, in order to secure data continuity and data access of MPD, BPS Statistics Indonesia make Memorandum of Understanding (MoU) with Ministry of Tourism and Ministry of Communication and Information. Then, the MoU will be followed with Technical Partnership Agreements for data access.

Technical Issues

Although MPD has been implemented, there are still some issues regarding the data quality. Some main technical issues are:

1. Indonesia residents may use foreign SIM cards when they across the border and can be identified as foreign tourists (the main examples are Malaysia). Vice versa, Malaysia residents may use national SIM card. Since there are sim card seller at the border gate, who sell both SIM card (Indonesia and Malaysia)

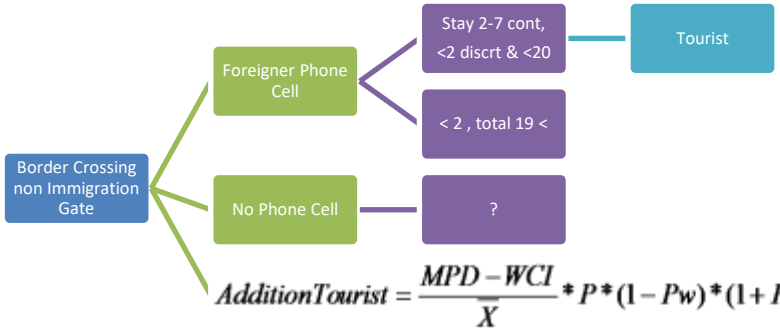
2. Blank spot: there are possibility that a visitors who already were in the country some days before which entering through immigration check point, then disappear (the mobile is switched-off). Then the signal is catch by the antennae at the border are (possible double counting).

3. MCC/MNC: false country due to investment (one example is a lot of Vietnamese number in Timor Leste, which is possible are not Vietnamese but Timor Leste.

Improvements

Improvement conducted:

1. Mobile phone usage survey: in order to correct mobile phone and SIM card usage, some questions about this matter were added in the cross border (shuttle) survey. The cross-border survey will give the result for the following formula



$$AdditionTourist = \frac{MPD - WCI}{\overline{X}} * P * (1 - Pw) * (1 + Pnonpc)$$

Where :
MPD=Number of foreigner phone cell signal
WCI = Number of tourist entering Imigration Post
= Averege of phone cell actively used by tourist
P = Rasio of foreigner traveler
P_w = Rasio of foreigner traveler for work or student
Pnonpc= Ratio of foreigner traveler not bringing phone cell

- Developing Methodological Handbook : Methodological Handbook for MPD use in Tourism Statistics should be developed and then followed by Methodological Handbook of MPD use for other Official Statistics.
- Developing Quality Assurance Report:The Quality Assurance Report of the use MPD for Tourism Statistics should be developed.
- Memorandum of Understanding with MNO: Memorandum of Understanding between BPS and one of MNOs which has the biggest market share.
- Technical Partenrships Agreement between BPS-Statistics Indonesia, Ministry of Tourism and Ministry of Communication and Information to obtain MPD from other MNOs.