

# Handbook on the Use of Mobile Phone Data for Official Statistics

*UN Global Working Group on Big Data for Official Statistics*

September 2019

# Table of Contents

|   |    |
|---|----|
| Introduction.....   | 4  |
| List of Definitions and Abbreviations .....   | 5  |
| 1 Applications .....  | 9  |
| 1.1 Mobile positioning data used in official statistics production .....                | 9  |
| Travel statistics in Estonia .....  | 9  |
| Inbound cross-border visitor arrival statistics in Indonesia .....                      | 9  |
| 1.2 Mobile positioning data used in research and pilots applicable for statistics ..... | 10 |
| Domain: Tourism and event statistics .....  | 10 |
| Domain: Population statistics.....  | 12 |
| Domain: Migration statistics.....   | 14 |
| Domain: Commuting statistics.....   | 15 |
| Domain: Traffic flow statistics.....  | 16 |
| Domain: Employment statistics on border and seasonal workers.....                       | 18 |
| Other applications.....   | 19 |
| 2 Data sources .....  | 21 |
| 2.1 Data from systems of mobile network operators .....                                 | 21 |
| Central storage systems .....   | 22 |
| Probing and signalling data.....  | 22 |
| Active positioning data .....   | 23 |
| 2.2 Mobile phone event data – passive positioning data.....                             | 24 |
| Forms of mobile data.....   | 25 |
| Subscriber-related identities.....  | 25 |
| Equipment-related identities .....  | 26 |
| Time attributes.....  | 26 |
| Location-related attributes .....   | 27 |
| Location and routing area identity.....   | 28 |
| Cell identity.....  | 29 |
| Coverage area of the antennae .....   | 30 |
| Additional attributes for events data.....  | 31 |
| Additional attributes for network data.....   | 32 |
| Additional attributes for subscribers .....   | 32 |
| 2.3 General data extraction process.....  | 33 |
| Data preparation .....  | 33 |
| Data anonymisation .....  | 33 |

|  |    |
|--|----|
| Data encryption .....  | 36 |
| Data transmission .....  | 36 |
| Data archiving .....   | 36 |
| The logical order of steps in the process of data extraction.....                      | 37 |
| 3 Access to mobile phone data and partnership models .....                             | 38 |
| 3.1 Introduction.....  | 38 |
| 3.2 Enabling environment for access to mobile phone data for official statistics ..... | 40 |
| Partnership models for using mobile phone data for official statistics.....            | 41 |
| ITU project on big data for measuring the information society .....                    | 47 |
| Understanding Stakeholders: Roles, Capacities, and Mandates .....                      | 49 |
| 4 Methods .....  | 54 |
| 4.1 Concepts and definitions.....  | 54 |
| Concepts related to trips .....  | 55 |
| Concepts relevant for several domains of statistics.....                               | 56 |
| Some concepts related to tourism domain of statistics.....                             | 57 |
| 4.2 Data processing methodology .....  | 58 |
| 4.3 Quality assurance framework for producing statistics with mobile network data .... | 61 |
| 4.4 Coping with under-/over-coverage.....  | 62 |
| Annex 1 - Case Study: France.....  | 71 |
| Annex 2 - Case study: Indonesia.....   | 78 |

## Introduction

Mobile phone data has surfaced in recent years as one of the big data sources with great 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 2018 report<sup>1</sup>, the International Telecommunication Union (ITU) shows cellular telephone subscriptions keep growing, even when there are already more subscriptions than people on Earth. Developing countries and least developed countries will continue to reduce the gap with developed countries. This goes to show how pervasive mobile phone use is. ITU elaborates that rural areas are still lagging behind urban areas, and this should be considered in studies using mobile phone data, but it is clear that the coverage of this 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 a 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 comprises four chapters (Applications, Data sources, Access to mobile phone data and partnership models, Methods), followed up by two country case studies.

This handbook has been composed by and edited by members of the Global Working Group on Big Data and specifically the Task Team on Mobile Phone Data.

---

<sup>1</sup> Measuring the Information Society Report 2018:  
<https://www.itu.int/en/ITU-D/Statistics/Pages/publications/misr2018.aspx>

# List of Definitions and Abbreviations

## Definitions used in the handbook

1. Mobile phone data or mobile positioning data (MPD) in the context of this handbook is any type of mobile phone event data that includes a subscriber identifier, time attribute and location.
2. Mobile network operator (MNO) also known as a wireless service provider, wireless carrier, cellular company, or mobile network carrier, is a provider of wireless communications services that owns or controls all the elements necessary to sell and deliver services to an end user. A key defining characteristic of a mobile network operator is that an MNO must own or control access to a radio spectrum license from a regulatory or government entity. A second key defining characteristic of an MNO is that it must own or control the elements of the network infrastructure necessary to provide services to subscribers over the licensed spectrum.
3. Domestic data – Any location events occurring within the network of the specific MNO.
4. Inbound roaming data – Any location event by a foreign MNO subscriber in the local MNO's network
5. Outbound roaming data – Any location event by a local MNO subscriber conducted in another network (usually a foreign MNO roaming service).
6. Call detail record (CDR) – Metadata record of a telecommunications transaction (call, text message, etc.) that is usually recorded for billing purposes.
7. Internet Protocol Detail Record (IPDR) or Data Detail Record (DDR) – metadata records of data exchanges in the mobile network
8. Signalling data – data captured through passive probing (see below)
9. Passive probing – monitors data flows between different network entities (as opposed to active probing, which generates traffic to monitor the network). This is the preferred possibility to gather information from MNO's network.
10. Record type – Every event in the network can be differentiated between data types (CDR and IPDR), service types (e.g., SMS, MMS, outgoing call, etc.) and events that are not subscriber generated (e.g., location updates)
11. Hashing or masking or pseudonymization – De-identification procedure by which personally identifiable information fields within a data record are replaced by one or more artificial identifiers, or pseudonyms.

12. Network cell – the smallest structural element for cellular systems. Can be described with azimuth, sector angle, shape and size of the coverage area, type of antenna and location. Every cell can serve only a limited number of users.
13. Base Transceiver Stations (BTS) – the node that handles the communication within a geographic area and can have multiple directional antennas linked. In LTE networks referred to as eNodeB.
14. Cell tower or mast – collection of transmitters at one specific location that can be referred to as a cell site
15. Cell site – see cell tower
16. Longitudinal data – Data about the subscribers over a longer period of time
17. Usual environment – The geographical boundaries within which an individual displaces himself/herself within his/her regular routine of life

### **General abbreviations**

1. National statistical institutes (NSI) = National statistics office (NSO)
2. International Telecommunication Union (ITU)
3. Telecommunication regulatory authority (TRA)
4. National Statistics Office (NSO)
5. Data Protection Authority (DPA)
6. Information and Communication Technologies (ICT)
7. Sustainable development goals (SDGs)
8. Quality Assurance Framework (QAF)
9. Machine-to-Machine communication (M2M)
10. Internet of Things (IoT)

### **Technical abbreviations related to mobile network operation**

1. Call detail record (CDR)
2. Internet Protocol Detail Record (IPDR) or Data Detail Record (DDR) – metadata records of data exchanges in the mobile network
3. Visitor location register (VLR)
4. Global System for Mobile Communications (GSM) refers to 2G networks
5. Universal Mobile Telecommunication (UMTS) refers to 3G networks
6. Long-Term Evolution (LTE) refers to 4G networks

7. 3<sup>rd</sup> Generation Partnership Project's (3GPP)
8. Base station subsystem (BSS)
9. Network switching subsystem (NSS)
10. Mobile switching centre (MSC)
11. Base station subsystem (BSC)
12. Home location register (HLR)
13. Business/Operations support system (B/OSS)
14. Mobility Management Entity (MME)
15. International Mobile Subscriber Identity (IMSI) – SIM card identifier. As long as subscriber has not changed SIM card, it will remain the same.
16. Mobile Subscriber Integrated Services Digital Network number (MSISDN) – Subscriber identifier. As long as subscriber has not changed his telephone number, it will remain the same.
17. International Mobile Station Equipment Identity (IMEI) – Device identifier. As long as the device is used, it will remain the same.
18. International Mobile Station Equipment Identity and Software Version number (IMEISV)
19. HLR
20. SGSN
21. Cell Global Identity (CGI)
22. Location area (LA)
23. Tracking area (TA)
24. Location area identity (LAI) – temporary subscriber data
25. Tracking area identity (TAI) – temporary subscriber data
26. Location Area Code (LAC)/Tracking Area Code (TAC) – identifier of the location area within an MNO's network
27. Circuit switched (CS) network – a network, where a group of cells form a location area or tracking area for LT networks
28. Mobile country code (MCC)
29. Mobile network code (MNC)
30. Routing area (RA)
31. Packet-switched (PA) network
32. Routing Area Identity (RAI) – composition of LAI/TAI plus RAC. Temporary subscriber data

33. Routing Area Code (RAC)
34. Cell Identity (CI, Cell ID) – unique only inside the location area
35. Cell global identification (CGI) – composition of LAI/TAI/RAI and CI. This is the unique identifier of a single cell



# 1 Applications

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

## 1.1 Mobile positioning data used in official statistics production

Currently, there are two countries where the statistics providers use mobile positioning data (MPD) as part of the regular production of official statistics – Estonia and Indonesia.

### Travel statistics in Estonia

Since 2008, Eesti Pank, the central bank of Estonia, has been cooperating with private companies to source big data for official statistics. The central bank uses mobile phone data from Estonian mobile network operators to quantify inbound and outbound travel, and credit card payment data to calibrate expenditure figures. The combination of these two datasets and several other sets of reference data allows the timely and efficient production of statistics on trade in services, used in the compilation of the balance of payments and the national accounts.

Travel statistics from mobile phone data have been separately disseminated as official statistics on Eesti Pank's website on a quarterly basis since 2012 with a time series running back to 2008. The cooperation has stood the test of time as part of the statistical business process at the central bank.

### Inbound cross-border visitor arrival statistics in Indonesia

To increase the data coverage on the number of foreign visitor arrivals, especially in border areas where the number has not been recorded with official immigration checkpoints, the BPS Statistics Indonesia and the Ministry of Tourism have been trying to improve the methodology of calculating the number of foreign visitors using mobile positioning data since October 2016.

MPD is used in cross-border posts in districts where immigration checks are not available and organising cross-border postal surveys is difficult due to geographical conditions.

Published by BPS Statistics Indonesia, the time series of inbound visitor arrival statistics that include mobile positioning data reaches back to October 2016. The cooperation model has changed over the years, as reflected in Annex 2.

## **1.2 Mobile positioning data used in research and pilots applicable for statistics**

There is now an extensive body of knowledge covering mobile positioning data, which is increasing by hundreds of academic articles per year. Most of the research and pilots are one-off projects, but increasingly, there are research institutions that have regular access to MPD, allowing them to perform work that builds on methods developed over time. To provide a glimpse of the work that has been done, this chapter discusses only some of the research and pilots with mobile positioning data which have been conducted so far.

*Domain: Tourism and event statistics*

| Description  | Source   |
|--|--|
| <p>The aim of this study 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.</p> <p>Country: Estonia</p>   | <p>Feasibility study on the use of mobile positioning data for tourism statistics. Consolidated report Eurostat contract No 3051.2012.001-202.452 (2014)</p> <p>Authors: Rein Ahas Jimmy Armoogum, Siim Esko, Maiki Ilves, Epp Karus, Jean-Loup Madre, Ossi Nurmi, Françoise Potier, Dirk Schmücker, Ulf Sonntag, Margus Tiru</p> <p>Link: <a href="http://ec.europa.eu/eurostat/documents/747990/6225717/MP-Consolidated-report.pdf">http://ec.europa.eu/eurostat/documents/747990/6225717/MP-Consolidated-report.pdf</a></p> |
| <p>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. The researchers also examined event statistics and their effect on tourism statistics. They discovered that agricultural fairs and international events have a very strong impact on tourism.</p> <p>Country: Estonia</p>  | <p>Mobile positioning data in tourism studies and monitoring: case study in Tartu, Estonia (2007)</p> <p>Authors: Rein Ahas, Anto Aasa, Siiri Silm, Margus Tiru</p> <p>Link: <a href="https://pdfs.semanticscholar.org/492b/53aa1c540ac9db823faefef3b1fef5f229b9.pdf">https://pdfs.semanticscholar.org/492b/53aa1c540ac9db823faefef3b1fef5f229b9.pdf</a></p>   |
| <p>Researchers investigated the ability of sports and cultural events to generate destination loyalty and repeat visitations. 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 making an effort to get the visitors of previous events to come back to Estonia</p> <p>Country: Estonia</p> | <p>The ability of tourism events to generate destination loyalty towards the country: an Estonian case study (2010)</p> <p>Authors: Andres Kuusik, Rein Ahas, and Margus Tiru</p> <p>Link: <a href="https://ojs.utlib.ee/index.php/TPEP/article/download/878/855">https://ojs.utlib.ee/index.php/TPEP/article/download/878/855</a></p>   |
| <p>Using passive mobile phone positioning, the researchers analysed tourists' spatial behaviour regarding weather and</p>  | <p>Weather dependence of tourists' spatial behaviour and destination choices: case study</p>   |

|  |  |
|--|--|
| <p>climate. 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.</p> <p>Country: Estonia</p> | <p>with passive mobile positioning data in Estonia (2007)</p> <p>Authors: Olle Järv, Anto Aasa, Rein Ahas and Erki Saluveer</p> <p>Link: <a href="https://www.semanticscholar.org/paper/WEATHER-DEPENDENCE-OF-TOURIST'S-SPATIAL-BEHAVIOUR-J%C3%A4rvAasa/f2d5966c48a4a1db8f3a3f3828604737444371aa">https://www.semanticscholar.org/paper/WEATHER-DEPENDENCE-OF-TOURIST'S-SPATIAL-BEHAVIOUR-J%C3%A4rvAasa/f2d5966c48a4a1db8f3a3f3828604737444371aa</a></p> |
|--|--|

### Domain: Population statistics

| Description   | Source   |
|---|--|
| <p>For this research paper, personal trip survey data was mainly used to generate substantial daily trajectories for the simulation system and for the validation of estimated population results.</p> <p>Country: Japan</p>  | <p>A Study on Urban Mobility and Dynamic Population Estimation by Using Aggregate Mobile Phone Sources (Chapter 6)</p> <p>Link: <a href="http://www.csis.u-tokyo.ac.jp/dp/115.pdf">http://www.csis.u-tokyo.ac.jp/dp/115.pdf</a></p>  |
| <p>The author proposed population statistics indicators which can be generated from mobile positioning data, such as the number of residences, workplaces, schools, secondary homes and other regular locations, changes of workplaces, cross-border migration, population grid statistics, temporary population statistics. In addition, real-time assessment for specific locations and risk assessment for law enforcement were presented.</p> <p>Country: Estonia</p> | <p>Overview of the sources and challengers of mobile positioning data for statistics</p> <p>Author: Margus Tiru</p> <p>Link: <a href="https://unstats.un.org/unsd/trade/events/2014/Beijing/Margus%20Tiru%20-%20Mobile%20Positioning%20Data%20Paper.pdf">https://unstats.un.org/unsd/trade/events/2014/Beijing/Margus%20Tiru%20-%20Mobile%20Positioning%20Data%20Paper.pdf</a></p> |
| <p>The research group proposed a systematic methodological framework for population density estimation based on mobile network data. The study addressed the problem of leveraging network-based data (CDR and/or VLR) for the task of estimating population density distribution at the pan-European level. The primary goal of the study was to develop a methodological framework for the collection and</p>   | <p>Estimating population density distribution from network-based mobile phone data (2015)</p> <p>Authors: Fabio Ricciato, Peter Widhalm, Massimo Craglia, Francesco Pantisano</p>  |

|  |   |
|--|---|
| <p>processing of network-based data that can be plausibly applied across multiple MNOs.</p> <p>Pan-European level</p>  | <p>Link:<br/> <a href="https://ec.europa.eu/eurostat/cros/system/files/Final-%20jrc-AIT-MNO-study-compressed_1.pdf">https://ec.europa.eu/eurostat/cros/system/files/Final-%20jrc-AIT-MNO-study-compressed_1.pdf</a></p>   |
| <p>The research team developed population mapping methods to estimate population density. The results of the study explicitly showed estimated population densities at national scales, and also demonstrated seasonal changes in population distribution. This study demonstrated how the analysis of mobile phone data, which is collected every day by mobile network providers, can complement traditional census outputs. Not only can population maps be 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 mobile phone data and remote sensing methods facilitates the improvement of both spatial and temporal resolutions and demonstrates how high-resolution population datasets can be produced for any time period.</p> <p>Countries: Portugal, France</p> | <p>Dynamic population mapping using mobile phone data (2014)</p> <p>Authors: Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R. Stevens, Andrea E. Gaughan, Vincent D. Blondel, Andrew J. Tatem</p> <p>Link:<br/> <a href="https://www.pnas.org/content/pnas/111/45/15888.full.pdf">https://www.pnas.org/content/pnas/111/45/15888.full.pdf</a></p>  |
| <p>The team investigated the massive and constantly updated information carried by mobile phone call data (CDRs) for estimating population statistics related to residence and mobility. In this study, the researchers proposed and experimented an analysis process built on top of the so-called Sociometer, a data mining tool for classifying users by means of their call habits. The results obtained show that flow estimates employing the method presented in the study are more accurate with larger towns.</p> <p>Country: Italy</p>   | <p>Use of mobile phone data to estimate mobility flows. Measuring urban population and inter-city mobility using big data in an integrated approach (2014)</p> <p>The study has been jointly developed by Istat, CNR, University of Pisa in the range of interest of the “Commissione di studio avente il compito di orientare le scelte dell’Istat sul tema dei Big Data”</p> <p>Authors: Barbara Furletti, Lorenzo Gabrielli, Fosca Giannotti, Letizia Milli, Mirco Nanni, Dino Pedreschi, Roberta Vivio, Giuseppe Garofalo</p> |

|   |  |
|---|--|
|   | Link: <a href="http://www.cisstat.com/BigData/CIS-BigData_06_Eng%20%20IT%20Mobile%20phone%20data.pdf">http://www.cisstat.com/BigData/CIS-BigData_06_Eng%20%20IT%20Mobile%20phone%20data.pdf</a>  |
| 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 night-time hours<br><br>Country: the Netherlands  | Daytime population visualisation by CBS Netherlands<br><br>Link: <a href="https://www.cbs.nl/en-gb/our-services/innovation/project/how-many-people-here-">https://www.cbs.nl/en-gb/our-services/innovation/project/how-many-people-here-</a> |
| In recent years, Beijing Municipal Bureau of Statistics of China establishes a dynamic population monitoring platform based on mobile phone data and develops methods for estimating population numbers. The platform provides a new data source and acts as an important supplement to traditional population statistics. By using mobile data, Beijing Municipal Bureau of Statistics of China obtains monthly population data and adjusts for data coherence between municipalities and districts. The platform measures dynamic population flows, meets the needs of government management and, furthermore, the platform is economical and timely.<br><br>Country: China | N/A  |

### Domain: Migration statistics

| Description  | Source  |
|--|---|
| The researchers established a methodological framework in studies of human migration and climate change by using mobile phone data. The study showed 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 quantify using standard survey-based research. The study also demonstrated how to analyse the relationship between fundamental parameters of migration patterns on a national scale based on mobile phone data. The study concurrently quantified the incidence, direction, duration and seasonality | Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in Bangladesh (2016)<br><br>Authors: Xin Lu, David J Wrathall, Pål Roe Sundsøy, Md. Nadiruzzaman, Erik Wetter, Asif Iqbal, Taimur Qureshi, Andrew Tatem, Geoffrey Canright, Kenth Engø-Monsen, and Linus Bengtsson |

|  |  |
|--|--|
| <p>of migration episodes in Bangladesh.</p> <p>Country: Bangladesh</p>   | <p>Link: <a href="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">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</a></p> |
| <p>The authors described how mobile phones can provide a new source of data on internal migration. The analysis revealed more subtle patterns that were not detected in a government survey on migration. The study provided a new quantitative perspective on certain patterns of internal migration in Rwanda that are unobservable using standard survey techniques and explored the ways how new forms of information and communication technology can be used to better understand the essence of migration in developing countries.</p> <p>Country: Rwanda</p> | <p>Inferring Patterns of Internal Migration from Mobile Phone Call Records: Evidence from Rwanda (2011)</p> <p>Author: Joshua Blumenstock</p> <p>Link: <a href="http://jblumenstock.com/files/papers/jblumenstock_itd2012_wp.pdf">http://jblumenstock.com/files/papers/jblumenstock_itd2012_wp.pdf</a></p>   |

### Domain: Commuting statistics

| Description   | Source  |
|---|---|
| <p>The researchers presented initial steps into the research of commuting patterns and functional regions using mobile phone location data. The main aim was to introduce and discuss the potential of mobile phone location data as an alternative data source to censuses for mapping commuting flows and subsequent functional regionalisation. A set of analytical maps covering various aspects of regular daily movements of the population and functional regionalisation was provided. Mobile phone location data can provide commuting information for the majority of population at an almost freely selected point of time and frequency of repetition, in a greater spatial detail and much more cheaply than traditional means of data gathering. Moreover, the time gap between the moment of survey and the moment of availability of data could be shortened to an almost real-time mode.</p> | <p>Application of mobile phone location data in mapping of commuting patterns and functional regionalization: a pilot study of Estonia (2013)</p> <p>Authors: Jakub Novak, Rein Ahas, Anto Aasa &amp; Siiri Silm</p> <p>Link: <a href="http://www.tandfonline.com/doi/pdf/10.1080/17445647.2012.762331?needAccess=true">http://www.tandfonline.com/doi/pdf/10.1080/17445647.2012.762331?needAccess=true</a></p> |

|   |   |
|---|---|
| Country: Estonia  |   |
| The researchers analysed the relationship between human mobility and economic activity in the urban environment by using mobile phone data. The article indicated 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 developed a dataset of commuting flows extracted from cell phone transaction data from Sri Lanka, and analysed 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 showed 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 showed that the constructed income measure has strong predictive power for night-time lights, an established proxy of residential income. | Estimation of urban commuting patterns using cellphone network data (2012)<br><br>Authors: Vanessa Frías-Martínez, Cristina Soguero-Ruiz, Enrique Frías-Martínez<br><br>Link: <a href="https://pdfs.semanticscholar.org/4af0/36dcbd2f0a26f01c0e1b90fede1c94821782.pdf">https://pdfs.semanticscholar.org/4af0/36dcbd2f0a26f01c0e1b90fede1c94821782.pdf</a> |
| Country: Sri Lanka  |   |

#### Domain: Traffic flow statistics

| Description   | Source  |
|---|---|
| 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. 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 was different from that of 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 behaviour played a greater role in evening | Mobile Phones in a Traffic Flow: A Geographical Perspective to Evening Rush Hour Traffic Analysis Using Call Detail Records (2012)<br><br>Authors: Olle Järv, Rein Ahas, Erki Saluveer, Ben Derudder, Frank Witlox<br><br>Link: <a href="https://doi.org/10.1371/journal.pone.0049171">https://doi.org/10.1371/journal.pone.0049171</a> |



|   |   |
|---|---|
| <p>rush hour traffic conditions as compared to the impact of suburbanisation.</p> <p>Country: Estonia</p>   |   |
| <p>The researchers estimated travel times in road sections by mapping the sequence of encrypted signalling messages (to preserve anonymity) for each mobile device to the physical movement along the road. On average, their approach gave results 3 minutes faster than the traditional monitoring approach. In addition, their approach 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 which were delivered using this method could be manually inspected to predict signals for a possible classification of congestion episodes.</p>   | <p>The Cellular Network as a Sensor: From Mobile Phone Data to Real-Time Road Traffic Monitoring (2015)</p> <p>Authors: Andreas Janecek, Danilo Valerio, Karin Anna Hummel, Fabio Ricciato, Helmut Hlavacs</p> <p>Link: <a href="https://www.researchgate.net/publication/275273390_The_Cellular_Network_as_a_Sensor_From_Mobile_Phone_Data_to_Real-Time_Road_Traffic_Monitoring">https://www.researchgate.net/publication/275273390_The_Cellular_Network_as_a_Sensor_From_Mobile_Phone_Data_to_Real-Time_Road_Traffic_Monitoring</a></p> |
| <p>The researchers used anonymous cell phone data to evaluate the change in traffic patterns caused by holiday traffic and discussed how traffic 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. Results showed that holidays affected daily and hourly traffic volume and speed, and that the effects were different for southbound and northbound traffic in the study area.</p> <p>Country: China</p> | <p>Understanding Intercity Freeway Traffic Variability at Holidays Using Anonymous Cell Phone Data (2016)</p> <p>Authors: Gang Liu, Chenhao Wand, Tony Z. Qiu</p> <p>Link: <a href="https://www.researchgate.net/publication/304620561_Understanding_Intercity_Freeway_Traffic_Variability_at_Holidays_Using_Anonymous_Cell_Phone_Data">https://www.researchgate.net/publication/304620561_Understanding_Intercity_Freeway_Traffic_Variability_at_Holidays_Using_Anonymous_Cell_Phone_Data</a></p>  |
| <p>The authors acknowledged that road congestion was an increasing problem in countries experiencing growth. Data was needed to identify the bottlenecks and to prioritise improvements. In this study, the authors wanted to understand whether mobile network big data could assist in transportation planning in the</p>   | <p>The potential of mobile network big data as a tool in Colombo's transportation and urban planning (2016)</p> <p>Authors: Sriganesh Lokanathan, Gabriel Kreindler, Nisansa de Silva, Yuhei Miyauchi, Dedunu Dhananjaya, Rohan Samarajiva</p>  |

|  |   |
|--|---|
| <p>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 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 which could improve the efficiency and reliability of the transportation system.</p> <p>Country: Sri Lanka</p> | <p>Link: <a href="https://www.researchgate.net/publication/269465996_Using_Mobile_Network_Big_Data_for_Informing_Transportation_and_Urban_Planning_in_Colombo">https://www.researchgate.net/publication/269465996_Using_Mobile_Network_Big_Data_for_Informing_Transportation_and_Urban_Planning_in_Colombo</a></p>  |
| <p>This paper provided a broad overview of the main studies and projects which address the use of data derived from mobile phone networks to obtain location and traffic estimation for individuals, as a starting point for further research on incident and traffic management. The authors 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.</p> <p>Country: the Netherlands</p>                               | <p>Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities (2013)</p> <p>Authors: John Steenbruggen, Maria Teresa Borzacchiello, Peter Nijkamp, Henk Scholten</p> <p>Link: <a href="https://link.springer.com/article/10.1007/s10708-011-9413-y">https://link.springer.com/article/10.1007/s10708-011-9413-y</a></p> |

### Domain: Employment statistics on border and seasonal workers

| Description  | Source   |
|--|--|
| <p>The study demonstrated how mobile phone data could be utilised to quickly and accurately detect, track and predict economic changes at multiple levels. In this study, researchers showed how mobile phone data can provide a quick insight into employment levels, precisely because people's communication patterns change when they are not working. Specifically, the researchers noted that their novel methods allowed them to predict present unemployment rates</p> | <p>Tracking employment shocks using mobile phone data (2015)</p> <p>Authors: Jameson L. Toole, Yu-Ru Lin, Erich Muehlegger, Daniel Shoag, Marta C. Gonzalez, David Lazer</p> |

|   |  |
|---|--|
| two to eight weeks prior to the release of traditional estimates and forecast future employment rates up to four months ahead of official reports. Their algorithms also predicted future states. | Link:<br><a href="http://rsif.royalsocietypublishing.org/content/royinterface/12/107/20150185.full.pdf">http://rsif.royalsocietypublishing.org/content/royinterface/12/107/20150185.full.pdf</a> |
|---|--|

### Other applications

A number of research papers and cases show that mobile phone data plays a very important role in various domains such as health, socio-economics, disaster response, urban management, etc. A list of such research papers is given below:

| Topic                          | Source   |
|--------------------------------|--|
| Development                    | Mobile phone network data for development, Global Pulse, 2013<br>Link: <a href="http://www.unglobalpulse.org/sites/default/files/Mobile%20Data%20for%20Development%20Primer_Oct2013.pdf">http://www.unglobalpulse.org/sites/default/files/Mobile%20Data%20for%20Development%20Primer_Oct2013.pdf</a>   |
| Social good                    | State of mobile data for social good, Global Pulse, GSMA, 2015<br>Link: <a href="http://www.gsma.com/mobilefordevelopment/wp-content/uploads/2017/02/State-of-Mobile-Data-for-Social-Good-Preview-Feb-2017.pdf">http://www.gsma.com/mobilefordevelopment/wp-content/uploads/2017/02/State-of-Mobile-Data-for-Social-Good-Preview-Feb-2017.pdf</a>  |
| Modelling epidemics            | On the use of human mobility proxies for modeling epidemics, 2014<br>Link: <a href="http://perso.uclouvain.be/adeline.decuypere/docs/journal.pcibi.1003716.pdf">http://perso.uclouvain.be/adeline.decuypere/docs/journal.pcibi.1003716.pdf</a>   |
| Malaria transmission           | Migration statistics relevant for malaria transmission in Senegal derived from mobile phone data and used in an agent-based migration model, 2016<br>Link: <a href="http://www.geospatialhealth.net/index.php/gh/article/view/408/357">http://www.geospatialhealth.net/index.php/gh/article/view/408/357</a>   |
| Ebola outbreak                 | 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 (CDR) data, which includes information on the use of mobile phones, including the location, from mobile network operators in Liberia, Guinea and Sierra Leone. The project demonstrated how analysing CDR data can provide information on human mobility, including cross-border movement, and the spatiotemporal distribution of people, while safeguarding individual privacy. The website for the ITU Ebola Project is available at: <a href="http://www.itu.int/net4/ITU-D/CDS/projects/display.asp?ProjectNo=7RAF15087">http://www.itu.int/net4/ITU-D/CDS/projects/display.asp?ProjectNo=7RAF15087</a> |
| Spreading of cholera outbreaks | Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks<br>Link: <a href="http://www.pnas.org/content/113/23/6421.full.pdf">http://www.pnas.org/content/113/23/6421.full.pdf</a>  |

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

## 2 Data sources

### 2.1 Data from systems of mobile network operators

Data extraction from the system of a mobile network operator (MNO) depends on the specific technical solutions the MNO is using<sup>2</sup>. 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 (Figure 2.1). Besides central storage systems where data is readily available, it is possible to extract data in the process of probing on the base station subsystem (BSS) and network switching subsystem (NSS) levels.

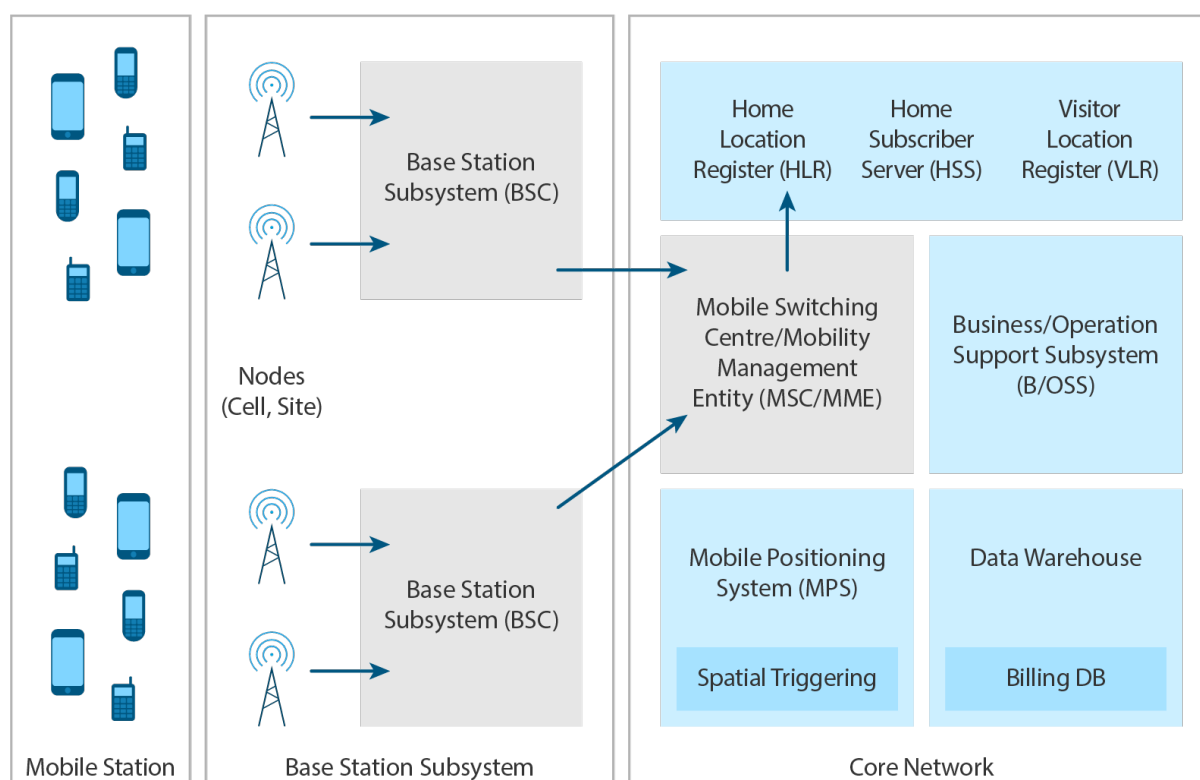


Figure 2.1 Sources of mobile positioning data in an operator's system<sup>3</sup>

<sup>2</sup> 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 organisations such as ETSI, 3GPP, Qualcomm and IEEE. In Europe (and also Japan and China), the most prevalent mobile communication technologies today are GSM (2G), UMTS (3G) and LTE (4G), which follow the technical specifications of the 3rd Generation Partnership Project (3GPP). For this reason, most of the descriptions of MNOs' network architecture and data access from the MNOs' systems in this report will be based on 3GPP specifications.

<sup>3</sup> Tiru, M., Ahas, R. (2012) "Passive Anonymous Mobile Positioning Data for Tourism Statistics". 11<sup>th</sup> Global Forum on Tourism Statistics <http://congress.is/11thtourismstatisticsforum/papers/Session3.pdf>

## 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 stores call detail records (CDRs) and internet protocol detail records (IPDRs) 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 billing domain is gathered by a mediation system from different network entities responsible for providing various types of services.

## Probing and signalling 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, and is usually the preferred option.

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, but they are most commonly as local as possible in the system. These probes can also query databases, such as the Visitor Location Register (VLR), that store relevant user data.

In terms of increasing the granularity of observations in data, there are many benefits to using data from probes:

- Gives access to data types that will usually not be stored in billing centres and data warehouse systems;

- Increasing the number of records per subscriber helps to minimise diurnal and daily differences in record counts which are the result of subscribers' calling patterns.

However, using data from probes (signalling data) has its drawbacks:

- 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. location area identity (LAI) is mandatory, but the more precise cell global identity (CGI) is not);
- Increases the number of events per subscriber considerably, meaning that more resources need to be dedicated to 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' systems is usually high and it is not implemented simply for generating statistical indicators for the NSIs. But if the MNOs have already implemented some sort of a probing system and the required data storage, it is a good source of data for statistical indicators as it involves much more data compared to “traditionally” collected CDR from a billing system or a data warehouse. Probing data is limited to inbound roaming and domestic data, normally signalling data is not exchanged between MNOs, which is why it is excluded from outbound roaming data.

Signalling data generally refers to directly obtaining transmission signals from the radio network and storing it in a 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.









### 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 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 the systems of operators in developing economies and are often

used merely on a case-by-case basis. Common methods of active positioning include cellular triangulation or assisted GPS (Table 2.1).

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

Table 2.1 Summary of main types of data sources for mobile positioning data. Source: Authors

|         | 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             | Cell activities  |    |    |
|         | Probes/signalling             | Needs collection, privacy protection and processing methods | Cell 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 |  |  |

## 2.2 Mobile phone event data – passive positioning data

A mobile phone event is seen as an action that is initiated by or targeted to a subscriber or mobile device. 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, a database like the VLR stores a new location area code (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 updates or other operational activities that generate an event along with a 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 include a 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.

### Forms of 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 a specific MNO. These are CDRs or other events where a 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 receipt).
2. **Outbound roaming data:** Any location event by a subscriber of a local MNO conducted in another network (usually a foreign MNO roaming service). This data usually represents local subscribers using mobile phones while travelling in foreign countries.
3. **Inbound roaming data:** Any location event by a subscriber of a foreign MNO. This data usually represents 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.

### 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:

- The identifier should be unique in the time span of the analysis;
- The identifier's life span should ideally be as long as the analysis period chosen;

- 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 an IMSI (International Mobile Subscriber Identity), an MSISDN (Mobile Subscriber Integrated Services Digital Network Number) or an MNO-specific customer/subscriber ID serve as permanent keys to identify a user from the databases. As long as a subscriber has not changed their SIM card, the IMSI will remain the same and as long as a subscriber has not changed their telephone number, the MSISDN will remain the same.

#### Equipment-related identities

Every mobile device is uniquely identifiable by 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.

#### Time attributes

When CDRs/IPDRs are generated 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 a service to a 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.

### 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 (ETSI TS 132 250), location-related identities are typically mandatory (or to be included when certain conditions are met) during the process of generating charging information. 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 a 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, outbound data also includes the network cell ID of the outbound roaming event.

For cellular systems, the smallest structural element is a network cell. Every cell can be described by a number of attributes such as its 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.

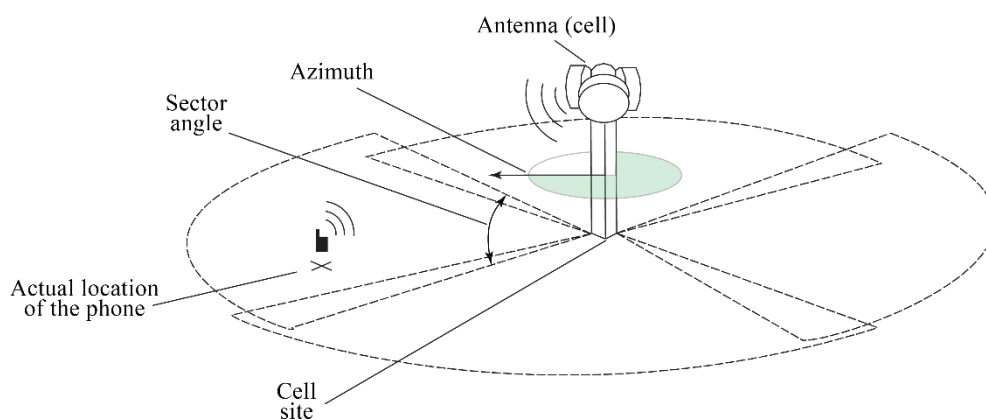


Figure 2.2 Cell, cell site and location area, and their relations to one another

A cell tower or mast is a collection of transmitters at one specific location that can be referred to as a cell site (Figure 2.2). 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<sup>4</sup>). This is done to balance the signalling load, so a device needs to send information about a user's location (location area updates) only when movement into another LA occurs<sup>5</sup> and periodically at predefined time intervals.

### Location and routing area identity

'Location area' is a term 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) is 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.

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

- 248 is the MCC for Estonia;
- 01 is the MNC for Telia;
- 0001 is the 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

---

<sup>4</sup> In LTE networks, the term 'tracking area' (TA) is used instead of 'location area' (LA).

<sup>5</sup> Sauter, M. (2010) From GSM to LTE: An Introduction to Mobile Networks and Mobile Broadband. John Wiley & Sons, 414p

the ITU-T Recommendation E.212. Since the LAC is unique only inside one specific MNO's 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 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 the Routing Area Identity (RAI), which is a composition of a previously described LAI/TAI plus the 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 is 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 '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 a hexadecimal representation similarly to LA (Figure 2.3). 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.

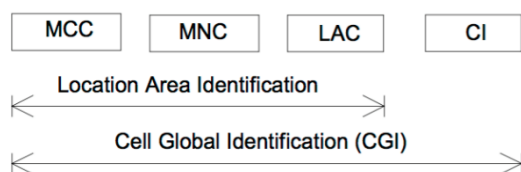


Figure 2.3 Comparison of location area identity (LAI) and cell global identity (CGI) Source: ETSI TS 123 003<sup>6</sup>

<sup>6</sup> The ETSI Technical Specification document defines the principal purpose and use of International Mobile Equipment Identities (IMEI) within the digital cellular telecommunications system and the 3GPP system: [https://www.etsi.org/deliver/etsi\\_ts/123000\\_123099/123003/14.03.00\\_60/ts\\_123003v140300p.pdf](https://www.etsi.org/deliver/etsi_ts/123000_123099/123003/14.03.00_60/ts_123003v140300p.pdf)

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 an MNO's network can change frequently (e.g., with the addition of new antennae, directional changes for individual cells), it is important to establish an understanding between an MNO and the NSI on how often such data should be transmitted. Ideally, cell information files follow event data.

### Coverage area of the antennae

Usually, the minimum geographical information provided by the MNOs is the location of network antennae (coordinate pair for the antenna point). If the geographical breakdown and accuracy of the resulting statistical indicators is of low resolution (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 (as part of data preparation), or the theoretical cell coverage area be calculated during the pre-processing of the data (Figure 2.4, Figure 2.5).

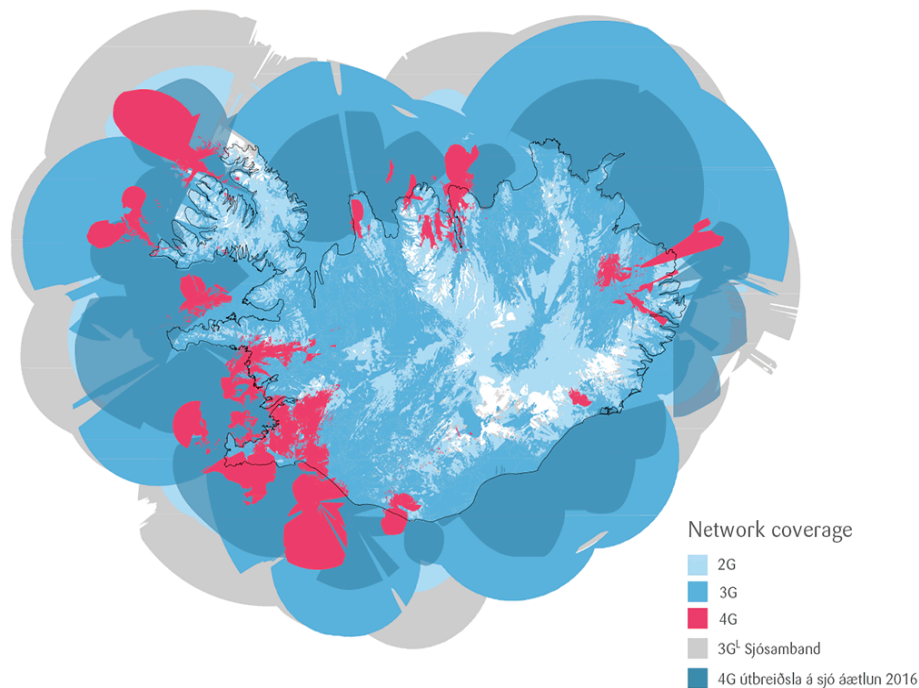


Figure 2.4 Illustration of the coverage area of one MNO's network antennae. Source: SIMINN<sup>7</sup>

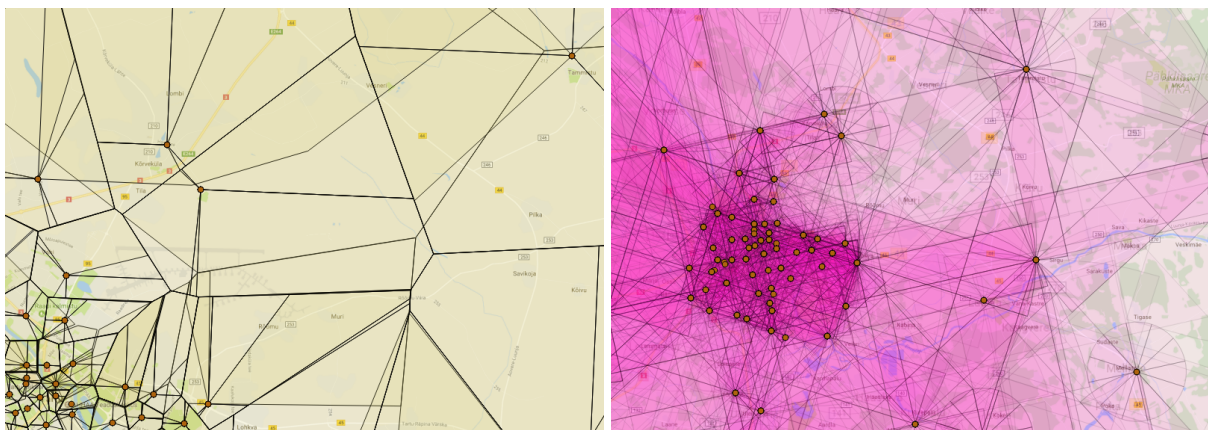


Figure 2.5 Example of two different methods for calculating the theoretical coverage area of a network antenna: a) Voronoi/Thiessen tessellation; b) theoretical sectors of the coverage area

#### Additional attributes for events data

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:

<sup>7</sup> Siminn in Iceland is one MNO that publishes its network's coverage map online: <https://www.siminn.is/en/prepaid/network>

- Record type – helps to differentiate between data types (CDR and IPDR), service types (e.g., SMS, MMS, outgoing call, etc.) and identify 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.

#### Additional attributes for network data

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.

#### Additional attributes for subscribers

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:

- 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 (the phone user may differ from the contract holder). The value of these attributes lies in the possibility of adding extra dimensions to statistical analyses (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 customer relationship management system of the MNO and are related to the customer profiling system. 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 the information is not always available about the customer). For enterprise



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 the 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 that of the father/mother).

## 2.3 General data extraction process

The individual steps of the data extraction process depend on who is responsible for data processing and where it is done. If the data is 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.

### Data preparation

The data preparation step typically involves creating a data extraction script to extract the necessary data from the storage unit. It should be done by the MNO, most likely employing a database communication language such as SQL. The extraction script preferably extracts data automatically at fixed intervals previously agreed upon between the MNO and the NSI. This way, it is possible to minimise 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, but also the file format and the need for anonymisation (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 blacklisted 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.

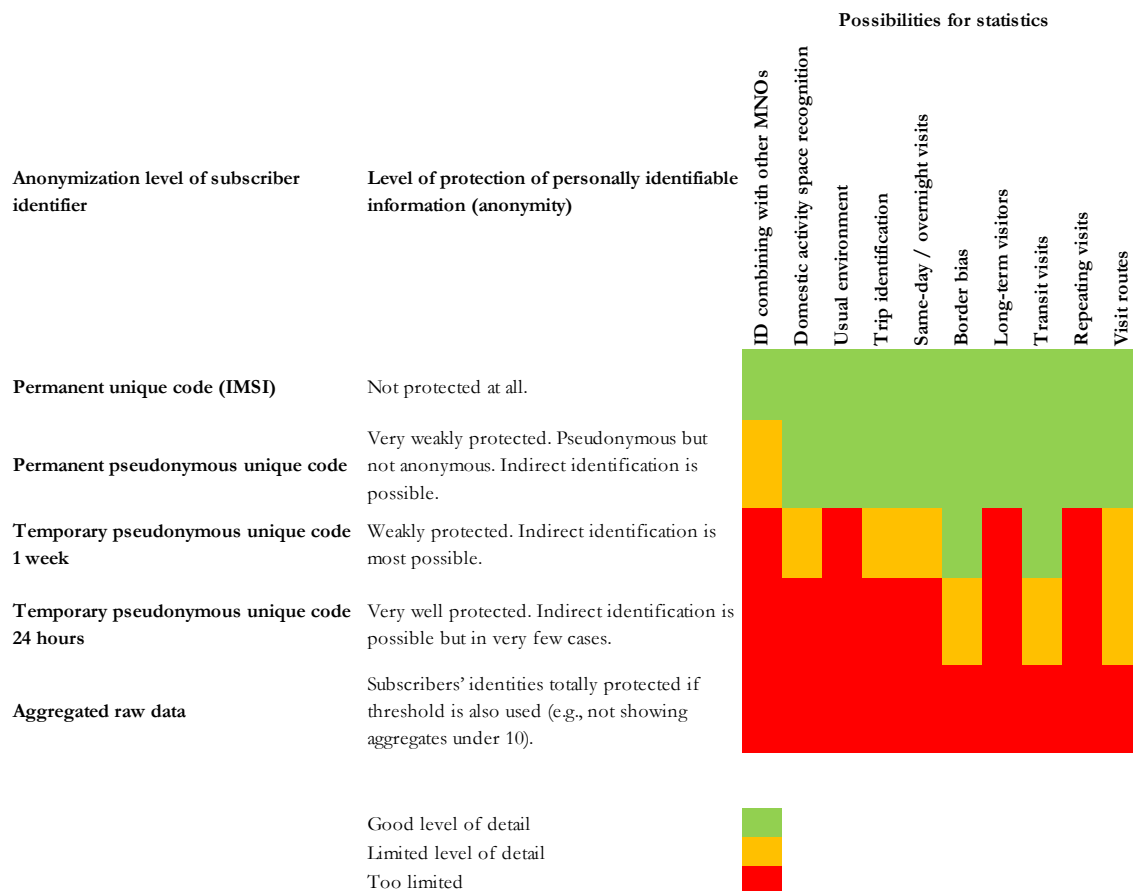
### Data anonymisation

The purpose of the data anonymisation 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 actually be used. Sometimes MNOs choose to hash the ID and thus mask it for direct identification (pseudonymise). If possible, this hashing should allow following the ID over a long period of time, but Table 2.2 lists the limitations of shorter refresh rates.

Alongside the options presented, sampling (decreasing the possibility for a person to be included in a dataset) 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 anonymisation method would be feasible for the NSI as well. Anonymisation (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 2.2 Limitations that various levels of anonymisation impose on the usefulness of the dataset for tourism statistics<sup>8</sup>



<sup>8</sup> Eurostat (2014) Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics

## Data encryption

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

During data encryption, a key is generated that will define how the data will be encrypted into a cipher text. In the 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.

## Data transmission

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

- 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.

## Data archiving

In the 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.

#### 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 anonymisation
14. How will the data be organised in files?
15. What are the specifications for data file elements? Attribute names, types, file names, file type
16. What are the specifications for the data extraction script? Removal of non-representative data (M2M / IoT, duplicates), removal of blacklisted 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 Access to mobile phone data and partnership models

### 3.1 Introduction

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 (Figure 3.1). Its widespread availability makes 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.

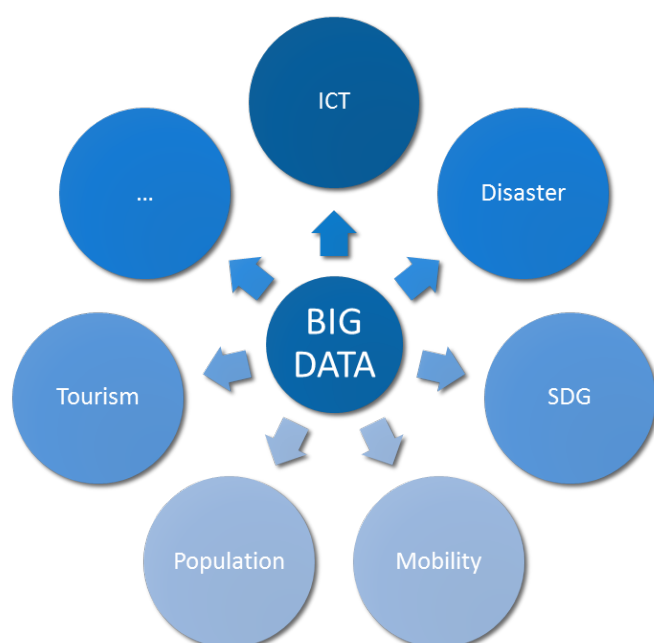


Figure 3.1 Multiple uses of big data for official statistics

First, there are capacity gaps: various stakeholders such as national statistical 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 analyse 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 analyse 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 riskier; 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 is 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 lies 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, 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 statistical 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 distil the best practices in building partnerships for accessing mobile phone data to produce 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 specialised 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.

### **3.2 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 their goals without unforeseen delays (and costs) when there is a 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 the 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.

### 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 statistical 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, at the beginning of the partnership, it is good practice to take inventory of and anticipate 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 specialised 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 the extraction of raw data to the computation of publishable indicators.

1. **Tier I** data consists of initial, raw, and non-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 MNOs. Tier I data from different operators is not yet ready to be merged. It will need to be cleaned further, 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 it no longer contains any personally identifiable or business confidential information.

In a **Type A partnership model**, as illustrated in Figure 3.2, the bulk of data processing is done by MNOs within their premises. Operators are essentially responsible for extracting raw data and processing it 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 organisations. In some cases where the partnership involves international organisations or foreign entities (and is contingent on data sovereignty requirements), processing of Tier II to Tier III data may even occur outside national borders.

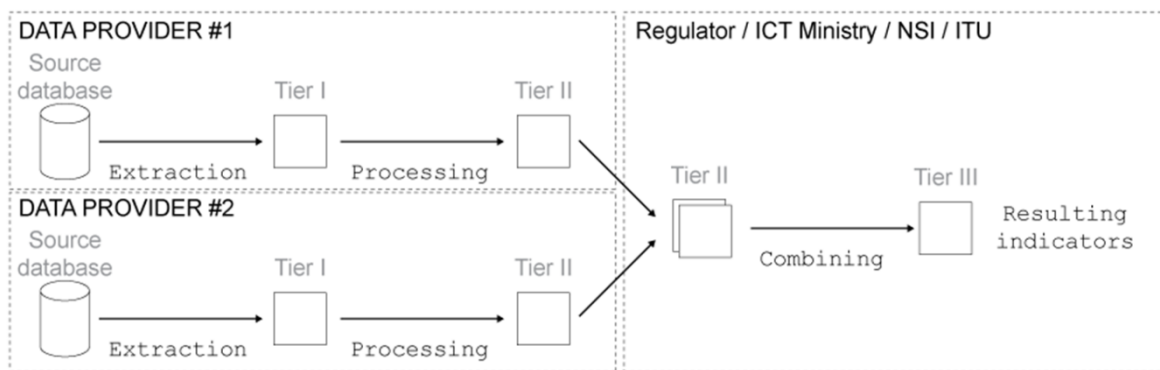


Figure 3.2 Type A partnership model – Mobile network operators mainly responsible for data processing<sup>9</sup>

The main advantage of the Type A partnership model is that it gives data providers (in this case, MNOs) 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 MNOs 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. the 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 the 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.

<sup>9</sup> This and the following figure are derived from ITU work on the project Big Data for Measuring the Information Society: <https://www.itu.int/en/ITU-D/Statistics/Pages/bigdata/default.aspx>

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 a detailed location-based analysis. For example, CDRs reveal patterns in population movements, which is useful for predicting outbreaks of deadly epidemics in specific locations. However, this potentially life-saving information is lost when data is aggregated.

In a **Type B partnership model**, as illustrated in Figure 3.3, the bulk of data processing is done by government bodies, outside the premises of MNOs. 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 organisations, 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 organisations or foreign entities (and is 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, in Tier II data, processing outside of national borders will require extra care with redaction, anonymisation and security.

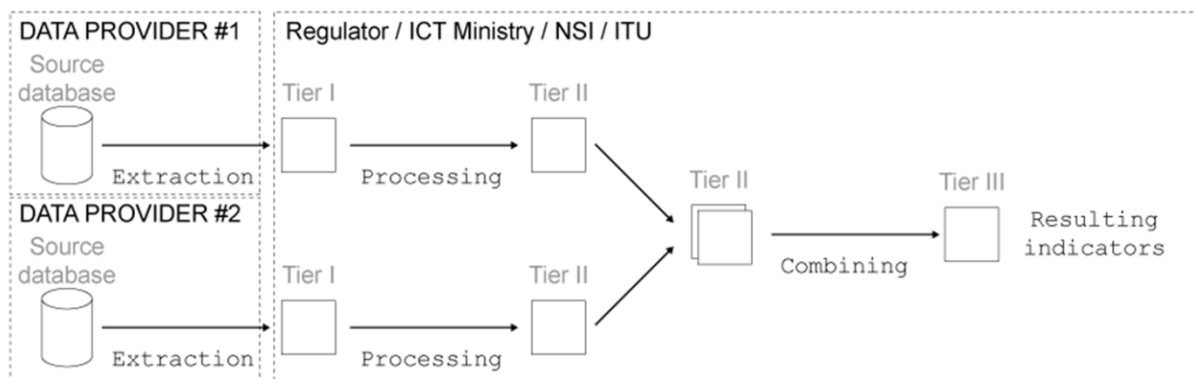


Figure 3.3 Type B partnership model – Government stakeholders mainly responsible for data processing

From the perspective of ensuring the veracity and continuity of deriving official statistics from mobile phone data, the Type B partnership model holds 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 process the data, they will not only be able to 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 orientate new entrants towards 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 is not necessarily a drawback. Borne strategically and with a view to the future, these costs can be understood as long-term public investments that can yield benefits beyond their 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 be done more easily 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 of mobile phone records containing personally identifiable information from one entity to another. 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 organisations (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<sup>10</sup>. 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

---

<sup>10</sup> European Union Agency for Fundamental Rights 2013

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 anonymisation or aggregation, and not just pseudonymisation. At the minimum, safeguards will need to be put in place to ensure that Tier I data is securely transmitted and stored.

In most cases, the preferred mode of operation is in fact a hybrid mode – Type AB, where the data remains at the MNO’s premises but is fully accessible by government-assigned data scientists. This approach is commonly used in the telecommunications sector, where the MNOs provide access to data to third-party vendors, either physically on location or through a remote secure connection. This mode of access has been used for many successful projects with mobile positioning data, especially in cases where data is used in official statistics (see Case study of Indonesia in Annex). The comparison of the types can be found in Table 3.1.

Table 3.1 A comparison of partnership models for accessing mobile phone data for official statistics

| Characteristics  | Type A                        | Type B                            | Type AB                           |
|--|-------------------------------|-----------------------------------|-----------------------------------|
| Data location  | MNO                           | Government                        | MNO                               |
| Entity responsible for data processing   | MNO                           | Government                        | Government                        |
| Granularity of data  | Low granularity               | High granularity                  | High granularity                  |
| Perception of risks to privacy and business confidentiality arising from use of data for official statistics | Low risk                      | High risk                         | Medium risk                       |
| Range of potential statistical indicators that can be derived  | Pre-defined statistics        | Wide range of possible statistics | Wide range of possible statistics |
| Ease of verifiability and level of statistical quality control   | Low control                   | High control                      | High control                      |
| Impact of the entry and exit of new operators in the market on the cost of production of official statistics | High cost of adding operators | Low cost of adding new operators  | Medium cost of adding operators   |

### ITU project on big data for measuring the information society

ITU has undertaken 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<sup>11</sup>. The following six countries participated in the pilot: Colombia, Georgia, Kenya, the Philippines, Sweden and the United Arab Emirates. Data has already been successfully accessed through either a Type A or a Type B partnership model as shown in Table 3.2 (Type AB was not used).

Table 3.2 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 Statistical Office        |
| Georgia              | B                 | Telecom Regulator                  |
| Kenya                | A                 | Network Operators                  |
| Philippines          | A                 | Network Operators                  |
| Sweden               | B                 | ITU                                |
| United Arab Emirates | A                 | Network Operators                  |

There are several lessons learned in terms of data access and preparing the field for a project on mobile positioning data.

1. First, there is a need to prepare all administrative and legal procedures to access the data before the project starts. This means there is an agreement between the stakeholders on the processing model for calculation (by data providers or by TRA/NSO). Regardless of where the processing happens, the method for data transfer

---

<sup>11</sup> The latest information on the project can be accessed at the ITU Big Data for Measuring the Information Society project website: <https://www.itu.int/en/ITU-D/Statistics/Pages/bigdata/default.aspx>

should also be agreed on in advance – see Section 2.3 for a logical order of topics to discuss before data transfer can happen.

2. Second, the method should be standard, clear and unambiguous. This includes a detailed description of the data source (input data, see Section 2.3). There should be a clear concept for the calculation methodology, which can be based on experience from other countries. The data processing teams also require example algorithms for calculation.
3. Third, the infrastructure and human resources (data scientists, data and system engineers, and project managers) for data processing should be in place for the project to be a success.

In addition, the project also showed that coordination with all stakeholders (on access to data, validation of results, analyses) should be continuously maintained and this will be covered next.



## Understanding Stakeholders: Roles, Capacities, and Mandates

Effective and sustainable partnerships will require buy-in and commitment from four groups of key stakeholders: national statistical offices and/or sector ministries, telecommunication regulators, MNOs and other network providers, and data privacy authorities (where they exist). Stakeholders have complementary roles depending on their mandates or organisational 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. The discussion that follows discusses 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 crossroads and so are the NSOs who are their main purveyors<sup>12</sup>. 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 to integrate sensitive data<sup>13</sup>. 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.

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 NSOs have the right to access data from public authorities only. Unfortunately, more often than not,

---

<sup>12</sup> Florescu, Karlberg, Reis, Castillo, Skaliotis and Wirthmann (2014) Will 'big data' transform official statistics?

<sup>13</sup> Tam, S.-M. and Clarke, F. (2014) Official Statistics and Some Initiatives by the Australian Bureau of Statistics. Paper presented at International Conference on Big Data for Official Statistics, Beijing, China, 28–30 Oct 2014 <http://unstats.un.org/unsd/trade/events/2014/Beijing/documents/other/Australia%20Bureau%20of%20Statistics>

the right of NSOs to access data is 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 the size of the dataset. One call detail record has an average size of 200 bytes. If a country has 6 million mobile subscribers of which about 20 per cent make a call each day, the total file size for one month's worth 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<sup>14</sup>. 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.

**Mobile network operators (MNOs):** For usage tracking and billing purposes, MNOs 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 CDRs are not obtained

---

<sup>14</sup> The scarcity of such skillsets can form a rationale for international cooperation and capacity-building.

from all operators, the data will only represent the characteristics of the subscribers of participating operators and inadvertently reveal sensitive business information to non-participating competitors. In the case of ICT indicators, partial data is not suitable for producing 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 enterprises
- Relative market shares

If operators are already required by law to submit records to a regulatory agency as a condition of their licence 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 fewer 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, 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 beyond 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 the MNOs' internal capacity to analyse 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 regulatory authority (TRA):** 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 they may also have already invested in equipment and expertise to store and process these records. If so, the regulators 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 agencies 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 the 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 have been 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 protection authority (DPA):** 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<sup>15</sup>. Although, the specific content of privacy laws varies 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<sup>16</sup>. The former is often more comprehensive and restrictive compared to the latter. Under continental European-

---

<sup>15</sup> DLA Piper has a regularly updated map of privacy rules around the world: <https://www.dlapiperdataprotection.com/>

<sup>16</sup> Determann (2015) Determann's Field Guide to Data Privacy Law: International Corporate Compliance

style laws, the default approach is to forbid and minimise 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 is generally allowed, provided safeguards are in place, including, whenever feasible, pseudonymisation or anonymisation. In contrast, Anglo-Saxon style laws focus on protecting individuals from the interception of confidential communication. This implies that unless there is a 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 Methods

### 4.1 Concepts and definitions

Mobile phone 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 benefitting from the use of this data source include tourism statistics, the balance-of-payments “travel item”, the Tourism Satellite Account, passenger transport, population, migration and commuting statistics (Figure 4.1).

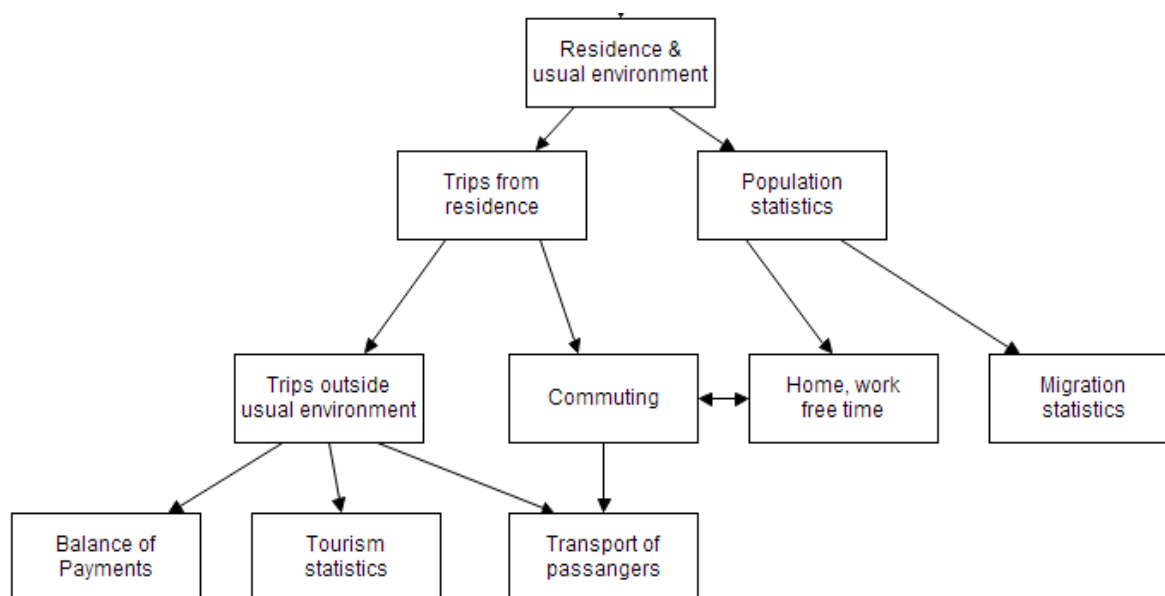


Figure 4.1 Connection of various domains of statistics<sup>17</sup>

In economy and finance statistics, it is mostly the balance of payments (BoP) that 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.

<sup>17</sup> Eurostat (2014) Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics. Report 3a – Feasibility of use, methodological issues.

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 the 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 are relatively lower.

The following chapters will introduce some useful concepts in the methodology of producing various human mobility statistics and how they apply when mobile positioning data is used.

### Concepts related to trips

**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 (similarly to the official definition).

**Visit/stay/dwelling:** 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. In tourism, 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) (similarly to the official definition).

**Overnight stay:** A concept that is used as a 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 the country concerned (the place is the country of reference), a visitor spent a night within the country (which differs from the official definition).

**Trip section:** A trip consists of 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. The 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).

#### Concepts relevant for several domains of statistics

**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 where 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 a 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 origin of the subscriber unless identified differently. For domestic and outbound subscribers, the usual environment



can be defined with the 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 on the producers' needs (which differs from the official definition).

#### Some concepts related to tourism domain of statistics

**Tourism:** The activity of visitors who are taking a trip to a main destination which is outside their 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.

**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 the course of one day.

**Secondary destination/visit:** As opposed to the main destination, a secondary destination 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 for a stop to be considered as a tourism visit (new concept).

**Transit pass-through:** As opposed to the main destination and the secondary destination, a transit pass-through is the place that visitors pass through or stop at during a period of time that is shorter than the minimum duration for a stop to be considered as a tourism visit. 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 trips the purpose of which is passing through that country on one's way to or from the country that is their main destination (similar to the official definition).

## 4.2 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 (Figure 4.2):

- Data extraction and raw data quality assurance
- 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)
  - 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 (e.g. one km<sup>2</sup>), LAU-2, LAU-1, country)
  - Combining data from various MNOs and computing final statistical indicators

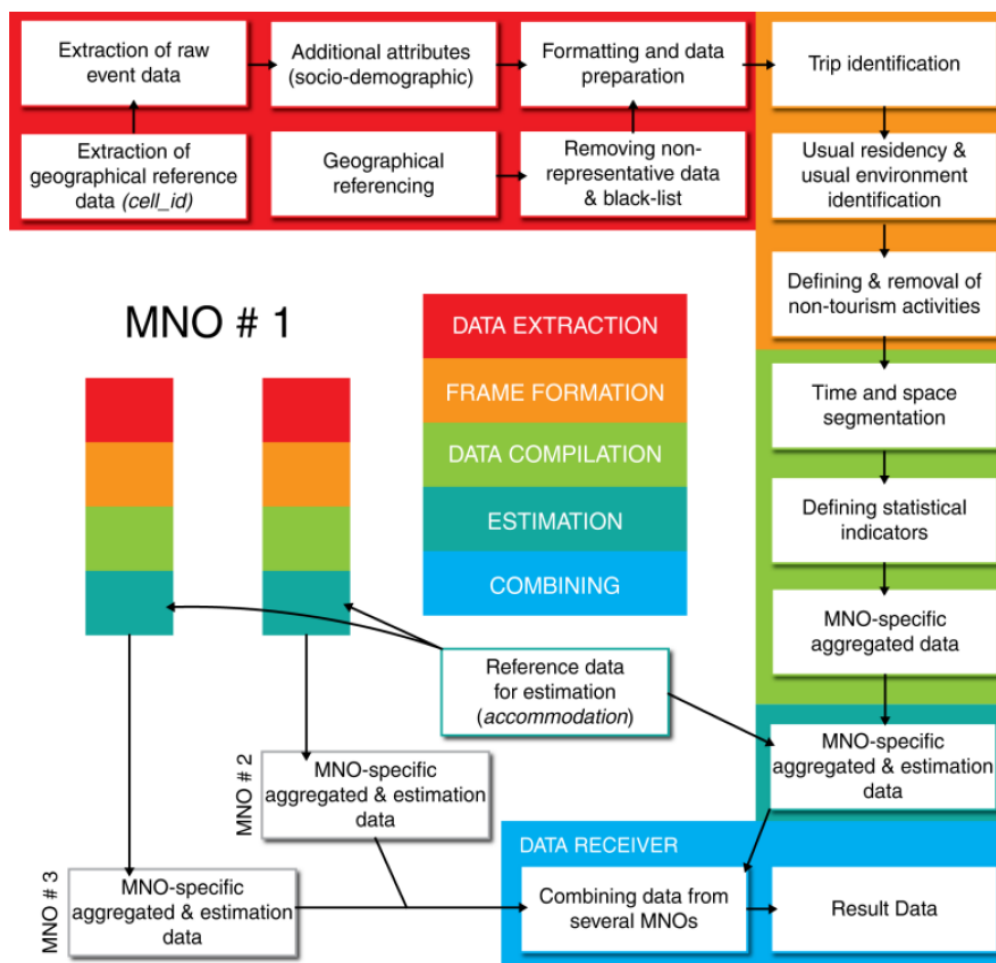


Figure 4.2 Data processing steps to create tourism statistics from mobile phone data<sup>18</sup>

The specifics of the 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 of the Eurostat Feasibility Study (trip calculation, identification of the 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 country 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 a historical time series (longevity) and interconnectivity between the different datasets

<sup>18</sup> Eurostat (2014) Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics. Report 3a – Feasibility of use, methodological issues.

(same subscriber ID in domestic and outbound data) for the subscribers in order to be able to define frequently visited places and countries.

One of the most important methodological considerations is the detection of the usual place of residence. Applying official definitions to mobile data is complicated as there are several criteria that need to be decided to build the home detection algorithm. According to research in France<sup>19</sup>, the choice of criteria in home detection algorithms significantly influences the results – up to 40%. In defining the algorithms for a specific country, the criteria depend on the data source and local conditions. In Estonia, the University of Tartu and Positium have developed home anchor point models<sup>20</sup> that reach up to 99% accuracy for the county level, with over 90% accuracy at higher levels of administrative units.

The Feasibility Study of the Use of Mobile Positioning Data for Tourism Statistics commissioned by Eurostat provides a description of the methodology in Report 3a that could be taken as a step-by-step guide to producing tourism statistics as it is rather detailed but at the same time general enough for broad use. There is an assumption that the dataset includes all the data necessary to construct tourism variables, e.g. the activities of anonymous subscribers can be followed over a longer period of time to establish their residency and/or usual environment. Availability of longitudinal data is crucial for the production of tourism statistics as algorithms identifying the usual environment and country of usual residence rely upon the availability of past data. If the data available 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 at a site of attraction or at a concert) and for relative comparisons.

---

<sup>19</sup> Vanhoof, M et al (2018) Assessing the quality of home detection from mobile phone data for official statistics. It is a validation study where the results of different home detection algorithms were compared to census counts at the cell tower level. The study is available at: <https://arxiv.org/abs/1809.07567>

<sup>20</sup> Ahas et al (2010) Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. The article validates the algorithm's accuracy at the aggregate level for the level of administrative units. There is also another validation study conducted with Statistics Estonia, where a comparison was made for volunteers, between their mobile data anchors calculated by Positium, the Population Registry data and the volunteers' true home address.

### 4.3 Quality assurance framework for producing statistics with mobile network data

Quality has an important role in building confidence in the products produced by an institution. Quality assurance forms an integral part of statistics production at any statistical office.

In mobile positioning data (MPD) projects, ensuring quality has become an issue of grave importance. A quality assurance framework (QAF) advocates for the positive assurance of quality at each stage of processing: 1) assurance of the quality of input data, 2) assurance of the quality of statistical processes, and 3) assurance of the quality of the product (output). Additionally, the institutional environment should be set up to be conducive to statistical quality. Through these three stages of quality assurance, the quality of statistical data generated will be well maintained because it is always checked at every stage of the process of statistical production.

A good QAF reflects the quality concepts related to the production of official statistics (e.g. ESS 2012<sup>21</sup>, UN 2015<sup>22</sup>, UNECE 2014<sup>23</sup>). The QAF should address:

1. Institutional environment – The institutional-level agreements between the stakeholders, details of communication, privacy aspects, regulations, roles, and other general questions.
2. Input quality – General guidance on the quality of this phase ensures the readiness of the raw input data to be processed. At the first access or due to important changes in raw input data general metadata and quality measures are collected (including measures of precision and selectivity). Afterwards this phase focuses on metadata accompanying regular data transfers.
3. Processing quality – Details of the statistical processing predominantly deal with the quality of the methodology and processing algorithms – verification of algorithms, validity of modelled data, metadata, logging and reporting throughout the processing.
4. Output quality – Quality aspects of the outcomes of data processing, validity of the transferred data files, adherence of the output to quality principles, usually measured

---

<sup>21</sup> EU, (2012), European Statistical System “Quality Assurance Framework of the European Statistical System” [http://epp.eurostat.ec.europa.eu/cache/ITY\\_PUBLIC/QAF\\_2012/EN/QAF\\_2012- EN.PDF](http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/QAF_2012/EN/QAF_2012- EN.PDF)

<sup>22</sup> UNECE (2014) A Suggested Framework for the Quality of Big Data

<sup>23</sup> United Nations (2015) Fundamental Principles of Official Statistics: Implementation guidelines

by levels of coherence (comparability over time and geography and with reference data).

The last phase – measuring output quality – is the most familiar to statistical offices. However, if data quality is assured in the previous phases in an expertly and consistent way, output quality measures should not throw up any surprises. It is the first phases of processing mobile positioning data where quality is to be assured, and output is reflective of that.

There are no standards for setting up a quality assurance framework, but work is in progress at the UNECE for creating a QAF for big data<sup>24</sup> which is applicable to mobile positioning data with some modifications. In Indonesia, BPS Statistics Indonesia has a comprehensive QAF for the methodology of deriving inbound cross-border tourist numbers from MPD (see Annex for the respective case study).

#### **4.4 Coping with under-/over-coverage**

In the case of mobile positioning data, there is one discrepancy that is immediately clear from comparing the definition of the target population and that of the population frame. In the frame, there are all those subscribers who use a mobile phone, whilst the target population includes all of the individuals who reside in the country.

Figure 4.3 presents the relationship between the target population and the final sample that is obtained. Here we distinguish between the population with phones (any kind) and the population with mobile phones. 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 base transceiver stations).

---

<sup>24</sup> UNECE (2014) A Suggested Framework for the Quality of Big Data

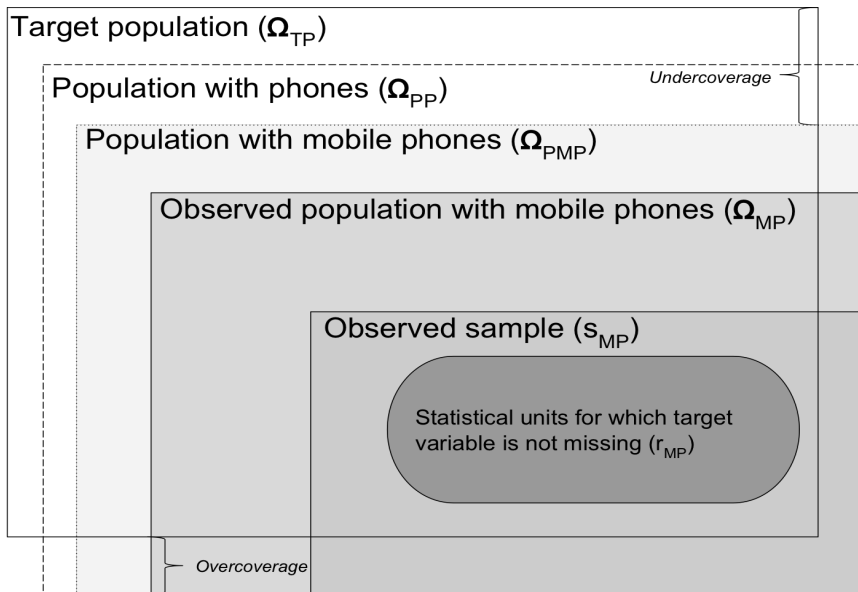


Figure 4.3 Moving from target to observed population in mobile network data<sup>25</sup>

Why is self-selection (selectivity) presented in terms of subsetting the target population? Regardless of the definition of the target population (e.g. trips, persons), some units will always be excluded. For instance, the mobile network infrastructure limits the exact identification of trips. However, this should be treated 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 a trip can be identified, which is also related to infrastructure coverage.

The two main reasons for coverage and selectivity problems stem from the following:

- Coverage of the target population (regardless its definition). One should take into account that the populations observed among the subscribers of the MNOs refer to businesses and natural persons, and they 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 this 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<sup>26</sup>.

<sup>25</sup> Maciej Beręsewicz, Risto Lehtonen, Fernando Reis, Loredana Di Consiglio, Martin Karlberg (2018) An overview of methods for treating selectivity in big data sources:

<https://ec.europa.eu/eurostat/documents/3888793/9053568/KS-TC-18-004-EN-N.pdf>

<sup>26</sup> Ricciato et al., 2015

One of the most important activities at the start of any mobile positioning data project is input data quality assurance, to have a very good description of the dataset and its coverage aspects, from the points of view of the subscriber profile, event data and cell coverage.

To sum up, we observe the following problems regarding inference, based on mobile data:

- Problems regarding the definition and derivation of target populations observed in mobile network data
- Limited access to background information on the users of mobile devices (place of residence, gender, age, 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

Some coverage issues 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 target population and by excluding them from the frame (Table 4.1).

Table 4.1 List of possible coverage issues with data from MNOs<sup>27</sup>

| Issue  | Under- or over-coverage | Possible solution  |
|--|-------------------------|--|
| 1. Data from selected MNOs only, i.e. not all data from MNOs is available. | Under-coverage          | Estimate other MNO shares by applying external information regarding the penetration rates and customer 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 behaviour of the under-covered group and then apply appropriate models to adjust the estimates. Both model-based and model-assisted estimators can be used. |
| 3. Use of more than one mobile device:                                     | Over-coverage           | At estimation stage, use information from additional sources – if available – to make  |

<sup>27</sup> Based on: Eurostat (2014) Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics. Report 3a – Feasibility of use, methodological issues.



|   |  |  |
|---|--|--|
| <ul style="list-style-type: none"> <li>• All devices use the same network causing duplication in one of the MNO's datasets</li> <li>• Devices use different networks causing duplication in the datasets supplied by different MNOs.</li> </ul> |  | <p>and check assumptions about the number of people having several mobile phones and then apply appropriate models to adjust the estimates. Adjustment depends on the frequency of the phenomenon.</p>   |
| <p>4. Different penetration rates for MNOs among foreign subscribers of specific countries.</p>   | <p>Under-coverage or over-coverage</p>   | <p>Assuming that information is available that allows the creation of separate MNO-based models that have country-specific parameters, weight-adjusted estimates can be computed.</p>  |
| <p>5. Different regional and socio-demographic penetration rates for MNOs.</p>  | <p>Under or over-coverage</p>  | <p>Surveys covering the regional and socio-demographic penetration rates of MNOs, if available, can be used in estimations.</p>  |
| <p>6. Machine-to-machine (M2M) communication devices that were not removed by MNOs during the data preparation process.</p>   | <p>Over-coverage</p>   | <p>Possible removal based upon event patterns.<br/>Based upon the Estonian inbound roaming data, the percentage of such devices in raw data has been from 0.04% to 2% depending on the MNO and the algorithm used. Based on Estonian outbound roaming data the share of such devices is very small, less than 0.01%.</p> |
| <p>7. Inbound: national roaming subscribers.</p>  | <p>Over-coverage</p>   | <p>Can be fully excluded as the country of roaming partner is the same as the country of reference.</p>  |
| <p>8. Inbound: Visitors not actually entering/exiting the country of reference (not crossing the border) but who are using the MNO roaming service of the country of reference or foreign country.</p>  | <p>Over-coverage</p>   | <p>Can be excluded partially based upon border bias recognition algorithms.<br/>Based upon the Estonian data, the percentage of excluded trips among inbound roaming has been 10.4% of all trips. A total of 58% of such trips is classed as trips with only one event.</p>  |
| <p>9. Tourism: same-day visits are over-represented when compared to overnight visits. This occurs with CDR data when visitors do not use their mobile phones on every single day of the trip.</p>  | <p>Over-coverage of shorter visits on the account of longer visits.<br/>Under-coverage of the duration of the trips.</p> | <p>Model the likelihood of a single trip being a same-day trip and use the modelled values in the estimation. In the case reliable reference data concerning the difference between the same-day and overnight visitors exists, such data should be used to correct the mobile phone data.</p>                           |

|   |  |  |
|---|--|--|
| 10. Inbound: visitors who use local SIMs for some reason and are not represented in the inbound/outbound roaming registry (see also Point 18 in this table).  | Under-coverage   | Calibration or a similar approach where known totals from other sources are used to correct the under-covered part of the population. Use data about the 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 poor network coverage. | Under-coverage   | Estimations have to take into account country-specific differences, compensating the technological limitations. In Estonia, subscribers from the 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 differs greatly by country and is very difficult to 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 the 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, use relevant information from additional sources to make and check assumptions about the travel behaviour of the under-covered group and 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 phone 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,   | Over-coverage  | Cross-roaming can be estimated by using privacy-preserving techniques and  |

|   |                         |  |
|---|-------------------------|--|
| causing duplication of the data for a single subscriber in the datasets of two or more MNOs.  |                         | comparing lists of roaming IMSIs between operators – it has been done in Indonesia to come up with market share ratios.  |
| 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).  | Over-coverage           | Can be excluded partially based on identifying short trips within transit corridors. Depending upon the month, the share of transit trips in the Estonian data varies from 2% to 10% of the total number of 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 in inbound data. Based on Estonian data, 2.9% of all domestic subscribers spend the majority of their time abroad, representing 24% of all outbound trips, and should therefore be considered as foreign residents and excluded from the domestic tourism dataset. |
| 20. Resident subscribers who change their pre-paid cards very often (short longevity of the <i>subscriber_id</i> ) 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 for’ by similar incorrect calculations for other subscribers. The quantity of incorrect calculations depends on the level of spatial resolution.  |

|  |                      |   |
|--|----------------------|---|
| <p>22. Outbound: residents of foreign countries who are using SIM cards from the country of reference;</p> | <p>Over-coverage</p> | <p>Based on Estonian data, roughly 3% of combined domestic-outbound subscribers spend more time abroad than in the country of reference. 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 in inbound data. The share of such persons can also be estimated.</p> |
|--|----------------------|---|

There are many contributors to 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 to a great extent. Many problems, however, are inherent in mobile positioning data and therefore cannot be avoided. Furthermore, their total effect, i.e. the total extent 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. European Commission regularly conducts a Eurobarometer surveys, and some are to study roaming behaviour. Flash Eurobarometer 468<sup>28</sup> showed that four fifths now use roaming in other EU countries after roaming became free in the EU (Figure 4.4). Previously, a Special Eurobarometer 414<sup>29</sup> showed that 28% of the travellers in the EU switch off their mobile phones when visiting another EU country (Figure 4.5 shows the data by country).

<sup>28</sup> Flash Eurobarometer 468 (2018) survey report on the end of roaming charges within the EU: <https://ec.europa.eu/digital-single-market/en/news/eurobarometer-survey-report-end-roaming>

<sup>29</sup> 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>

Q2a The last time you visited another European Union country, how often did you use the following services on your mobile phone?

(% - EU, 2018)

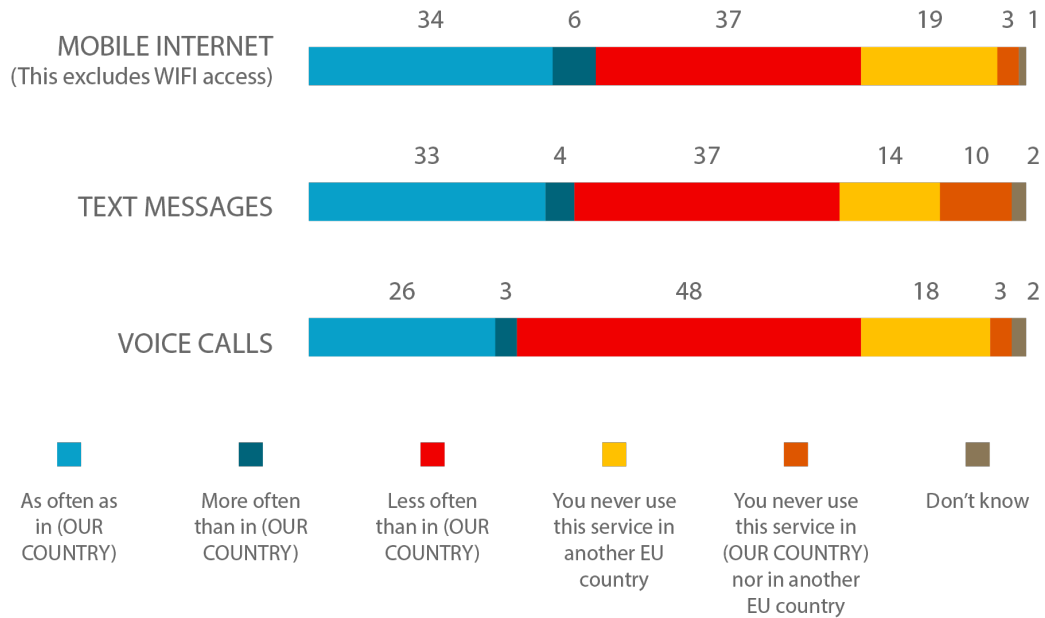


Figure 4.4. Percentage of people surveyed in the EU in 2018 on how often they use various roaming services while travelling to another EU country (a fifth do not).

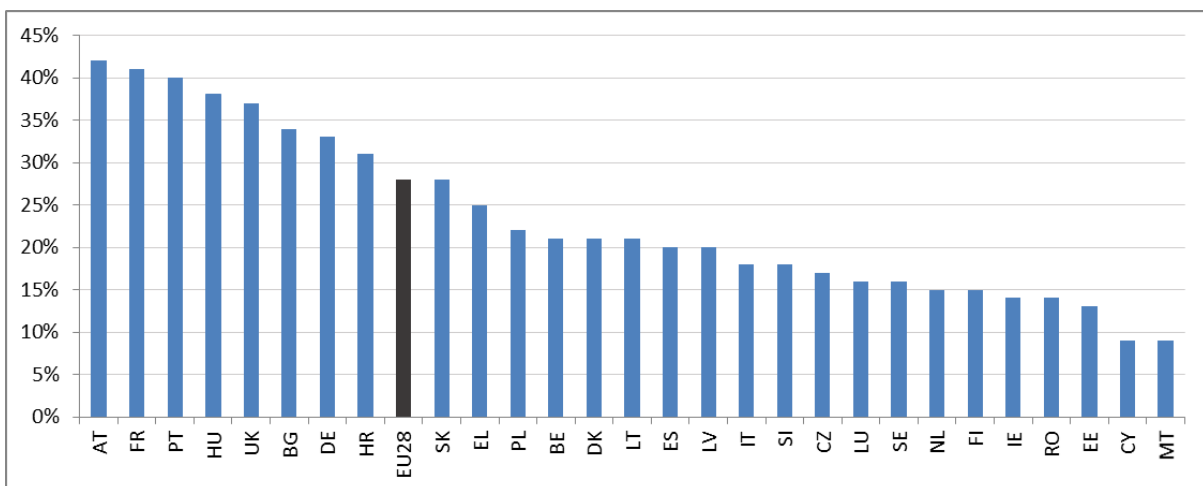


Figure 4.5 Percentage of the population (aged 15+, who have travelled and own a personal mobile phone) who generally switch off their mobile phone and never use it while visiting another EU country<sup>30</sup>

The best but also the costlier way to evaluate bias would be to carry out sample surveys for quality assessment purposes for each coverage issue, providing the basis for individual

<sup>30</sup> Ibid.

estimations. This survey would give the estimates for the proportion of people not having or using mobile phones, the proportion of people using two or more devices etc. This information would allow to compute the total extent of 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 bias is to use information from other sources (e.g. results from surveys covering the relationship between the number of same-day trips and overnight trips to the country for inbound tourism, the usage of mobile phone in different countries for outbound tourism). Literature on sample surveys has shown that coverage bias can be dealt with by using methods such as 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 provided.

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 making estimations is still required.

The evaluation of the extent of all the coverage problems listed 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, subscribers having many mobile phones).

## **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<sup>31</sup> 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 visitor 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**

---

<sup>31</sup> 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.

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 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<sup>32</sup>). The individual data is not saved: the algorithm works in real time and counts the

---

<sup>32</sup> This individual data includes both the passive signalling and mobile phone activity (calls, messages, etc.)



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 (Figure 0.1). Some of these grossing-up, related to the MNO's market share on roaming activity, cannot be isolated and are part of the anonymisation 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.

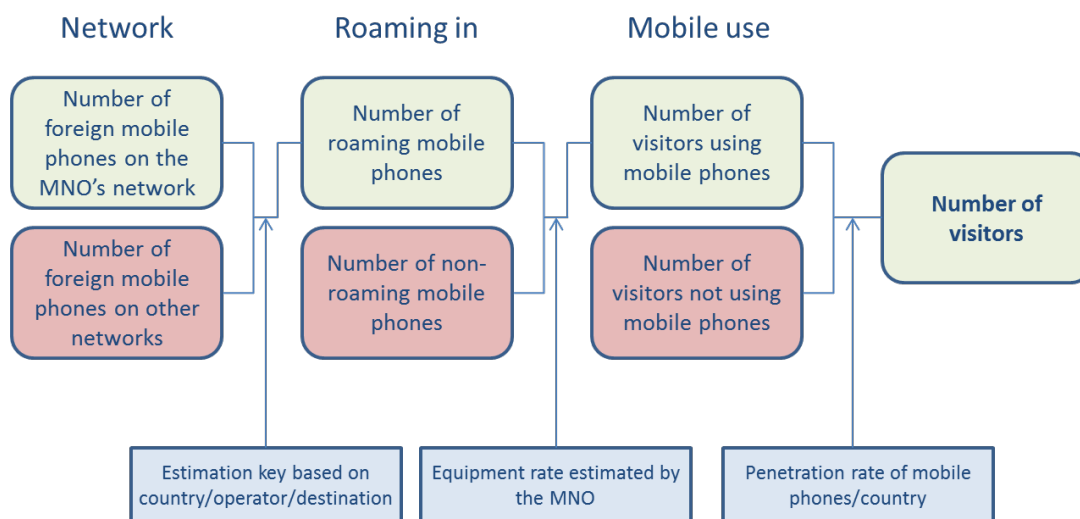


Figure 0.1 Grossing-up: from a number of mobiles to the number of visitors

## 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 summarised below:

- **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)
- **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
- **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.
- **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:

- **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.
- **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
- **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

- **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 (Figure 0.2).

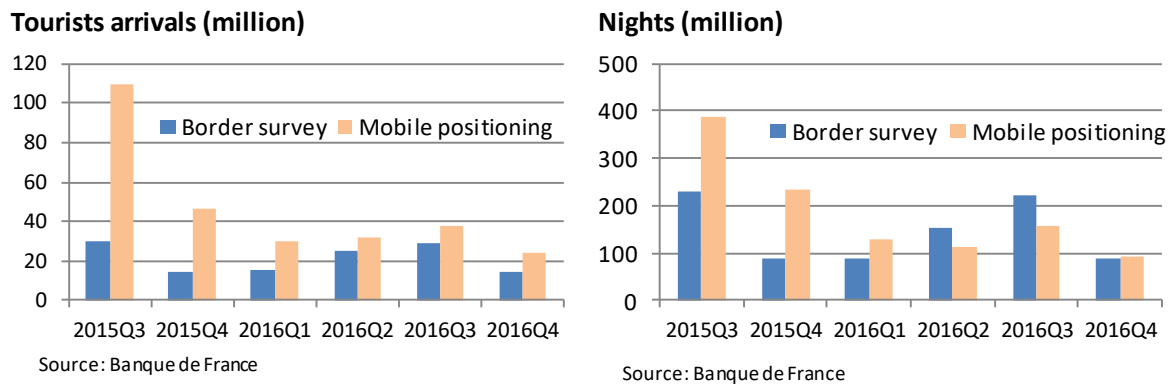


Figure 0.2

### 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 (Table 0.1). However, the comparison with our border survey estimates<sup>33</sup> 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.

<sup>33</sup> 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.

Table 0.1

**Tourists arrivals and nights (million)**

|                 |                    | 2015Q3 | 2015Q4 | 2016Q1 | 2016Q2 | 2016Q3 | 2016Q4 |
|-----------------|--------------------|--------|--------|--------|--------|--------|--------|
| <b>Arrivals</b> | 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   |
|                 | <i>spread</i>      | 272%   | 217%   | 93%    | 28%    | 30%    | 71%    |
| <b>Nights</b>   | 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   |
|                 | <i>spread</i>      | 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 produces and publishes tourism data in Indonesia. Prior to the use of mobile positioning data (MPD), the main source of inbound tourist data was administrative data from the Directorate General of Immigration, which covered all main airports and entrances. Other data source was Cross-Border (Shuttle) Survey conducted by BPS in order to obtain tourists/visitors data at a certain border entrances/gates, 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 were based on administrative data and surveys. The surveys were 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 points, the entrance/gate is only guarded by army personnel's or head of village. However, Indonesia is quite vast area and many of the borders are 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 are costly 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 such: the MNO processes the MPD to become statistics and give it to Ministry of Tourism (MoT), the MoT gives it to BPS Statistics Indonesia. The rules and filters were set up by BPS-Statistics Indonesia, Ministry of Tourism prepared the budget for MPD data processing. So, the raw individual data remains 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 staying in the same area and close to each other, then they are 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 provided by MNO is in a form of a table which contains district name (proxy of place of entrance), country (from SIM card/MCC) and number of tourists/visitors. BPS-Statistics Indonesia then compares 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

Table 0.1 Project time-line for 2016-2017

| Activity  | Time           |
|---|----------------|
| Meeting to use MPD for Tourist data at the Border   | July 2016      |
| Meeting on deciding the MNOs to collaborate   | August 2016    |
| Discussing Methodology and presenting data calibration  | September 2016 |
| Deciding Methodology and Border Area that will use MPD  | October 2016   |
| Implementing MPD to October data  | November 2016  |
| Discussing and drafting MoU between BPS-MoT and MoCI (Based on Statistical Act No. 16/1997)                     | December 2016  |
| MoU between BPS-MoT and MoCI (Based on Statistical Act No. 16/1997) is signed                                   | January 2017   |
| Presenting MPD Methodology and Implementation to Statistical Society Forum                                      | February 2017  |
| Initiating and Planning the Mobile Phone Usage Survey at the Border   | March 2017     |
| Workshop on the Use of MPD (Participants: Provincial and Regency Statistical Office, Ministry of Planning, MoT) | April 2017     |
| Preparing the Mobile Phone Usage Survey (Enumerator Training etc)   | May 2017       |
| Conducting the Mobile Phone Usage Survey (Field work)   | June 2017      |

Table 0.2 Future plans (What is next ?)

| Activity   | Time        |
|--|-------------|
| Discussing and drafting the Technical Partnership Agreement for Mobile Positioning Data Access | 2017        |
| Signing the Technical Partnership Agreement for Mobile Positioning Data Access                 | 2017        |
| Technical Workshop/Consultation on MPD   | August 2017 |



|   |              |
|---|--------------|
| Developing MPD Methodological Handbook for Tourism Statistics | 2017         |
| Memorandum of Understanding with one MNO                      | 2017         |
| Technical Workshop/Training on Data Processing                | October 2017 |
| Preparing Server, Network etc for data exchange               | 2017         |
| 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 the 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 analysing 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 be a relative 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 concerns, 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 makes 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**

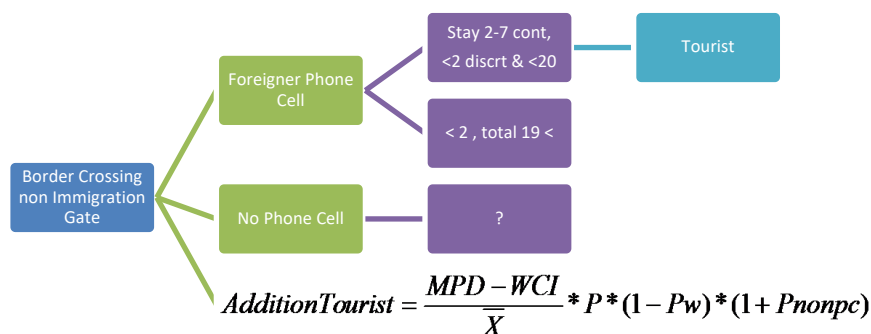
Although MPD has been implemented, there are still some issues regarding the data quality. Some main technical issues are:

- **Indonesia residents** may use foreign SIM cards when they cross the border and can be identified as foreign tourists (the main examples are Malaysia). Vice versa, Malaysia residents may use national SIM card. There are sim card sellers at the border gate, who sell both SIM cards (Indonesia and Malaysia).
- **Blank spot**: there is a possibility that a visitor, who already was in the country some days before entering through immigration check point, disappears (the mobile is switched off). Then the signal is caught by the antennae at the border (possible double counting).
- **MCC/MNC**: false country due to investment (one example is a lot of Vietnamese number in Timor Leste, which is possible that is not Vietnamese but Timor Leste).

### **Improvements**

Improvements conducted:

- **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:



Where :

MPD=Number of foreigner phone cell signal

WCI = Number of tourist entering Immigration Post

= Averege of phone cell actively used by tourist

P = Rasio of foreigner traveler

P<sub>w</sub> = Rasio of foreigner traveler for work or student

P<sub>nonpc</sub>= 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 Partnrships Agreement** between BPS-Statistics Indonesia, Ministry of Tourism and Ministry of Communication and Information to obtain MPD from other MNOs