# Towards a methodological framework for the integration of mobile phone data in the production of official statistics

Dr. David SALGADO

*Head of Unit of Dept. Methodology and Development of Statistical Production, Statistics Spain (INE), Spain*

Dr. Bogdan OANCEA

*Director of Dept. Innovative Tools, INS, Romania*

**Abstract.** Mobile phone data are widely recognized as a potential data source for producing official statistics in several fields such as population estimations or tourism statistics. In this paper we proposed some first elements for a methodological framework for the integration of mobile phone data in the production of official statistics, starting from an adaptation of the two-phase life-cycle model for statistical microdata used in the case of administrative data sources. Thus, our proposal considered all phases of the statistical production, beginning with the processing of raw telecommunication microdata into appropriate microdata sets, followed by an aggregation procedure according to a well-established methodology into data sets for each territorial cell and time instants, and a final inference stage from these aggregated data sets to the target population. We emphasized the similarities between the treatment of administrative data and mobile phone data sets. For the final inference stage, we designed a hierarchical model borrowed from ecological sampling techniques, which is based on a Bayesian approach. This model combines the aggregated mobile phone data sets with other data sources such as a population register to produce estimates for the target population. We implemented this model in a software package and conducted experiments which allowed us to identify those areas that need further developments. Investigating the quality of the estimates we concluded that several principles from the European Statistics Code of Practice are affected since (i) an important part of the production process is carried out, at least at this moment and in the near future, by MNOs, (ii) the inferential paradigm is changed, and (iii) an unprecedented degree of spatial and time disaggregation will be reached.

## 1. Introduction

Mobile phone data have been an outstanding promising data source to produce official statistics for the last decade with successful proofs of concept both in the public [1], the academic [2] (and multiple references therein), and the private sector (see e.g. [3]). However, integrating this data source into the standard production system of national and international statistical offices has been proved to be remarkably difficult.

Within the European Statistical System (ESS), the VIP project ESSnet on Big Data [4] was launched from February 2016 to May 2018 to investigate the many aspects of this integration. The project seeked to implement the production of concrete actual statistics from mobile phone data sets. It followed a hands-on bottom-up approach focusing on this ensuing natural sequence of milestones: (i) getting access to real data, (ii) developing the necessary statistical methodology, (iii) implementing these methods in concrete software tools and related IT infrastructure, and (iv) producing outputs upon which quality issues were to be investigated.

Regarding mobile phone data, this research should have driven us close enough to implement the production of these statistical outputs paving the way for a wide range of statistical domains. Despite the big efforts in the right direction, the access issue remains as an outstanding obstacle to achieve a full integration of mobile phone data in the standard production. Nonetheless, important elements for the construction of a production framework have been identified and put into place. We provide a high-level summary of these elements aiming at a standardized production framework within the ESS and underline strategic issues identified so far in the research.

## 2. Big Data for Official Statistics and administrative data

To begin with, the classical definition of Big Data in terms of volume, velocity, and variety [6] was found at least misleading for Official Statistics. We identified another three features of Big Data for Official Statistics as more relevant [7]: (i) they are data about third people not about data providers/holders themselves, (ii) they are at the core of the economic activity of data providers/holders, and (iii) they lack statistical metadata (since they are generated for other reasons). Survey data fail to meet these criteria whereas novel data sources satisfy them. These characteristics have been found important not only in negotiating the access but also in the development of statistical methods and IT infrastructure to process data not to mention diverse quality aspects.

Curiously enough, the definition of administrative data in the UNECE Terminology on Statistical Metadata [8] defines them as “data collected by sources external to statistical offices”. Thus, we see a first similarity here between administrative data and mobile phone data in a wide sense. This similarity will again arise later on.

## 3. Access to mobile phone data

In September 2016, the project organized an international workshop in Luxembourg gathering ESS National Statistical Institutes (NSIs), European Mobile Network Operators (MNOs), and some other international agencies and institutions [5]. This meeting ran parallel to the bipartite negotiations

between NSIs in the project and their national MNOs to reach an agreement on accessing mobile phone data. This rich experience brought a lot of empirical knowledge to the ESS.

A mobile telecommunication network generates a huge amount of different types of data. The network can indeed be understood as a complex sequence of nested information systems rooted in many local antennae connecting to mobile devices which ultimately ends up in a centralized network management system (billing system included) (see figure 1). This entails far-reaching consequences in accessing mobile phone data.

### *3.1. What data?*

When approaching an MNO to request access to mobile phone data, more often than not we must face the question "what data"? This arises from the complexity of the network and the huge amount of data produced by the subscribers' communication activity. As a rule of thumb, with figure 1 in mind, we can state that the closer the source of data generation to the antennae, the richer the data for statistical purposes. In this sense, data coming only from the billing system will offer fewer possibilities than data coming from the interaction between each mobile device and the antennae (which e.g. does not totally depend on the communication pattern of subscribers). This is the basis for the distinction between so-called Call Detail Records (CDRs) and signaling data and it shows a reverse side: the closer the source to the antennae, the higher the technical intricacy in retrieving the data. In addition, this entails a potential higher level of disruption in the routinely operation of the network and a need for investment in the data retrieval system. Many MNOs have already monetized their data as statistical products, thus this infrastructure is already present in many cases.

The ESS should strive for accessing and using those data with better potentialities for statistical purposes, i.e. for accessing and using signaling data.

### *3.2. Preprocessing*

Raw telecommunication data retrieved from a network lack statistical metadata because they are generated for a very different reason. For the statistical exploitation they need to be preprocessed. This means that we need to produce data at the mobile device level **for statistical purposes**, which are a combination of a pseudonymised identification variable, space and time attributes, and complementary variables per mobile device. Details will depend on the concrete raw telco data (CDRs, signaling data…). By and large, this preprocessing must be currently undertaken in the MNOs' information systems. Complementarily, these data can be further aggregated (e.g. number of mobile devices per territorial cell), although this further step can be carried out by NSIs in those cases where they have access to the statistical microdata. The aggregated data will finally be

analyzed to infer statistical aggregates for the target population. This final inference step will be conducted by NSIs. We arrive at a very important finding for this data source: MNOs will be partially integrated in the official statistical production process. We envisage the extension of this fact to many new data sources (e.g. Trusted Smart Statistics) thus portraying far-reaching consequences for the production of official statistics.

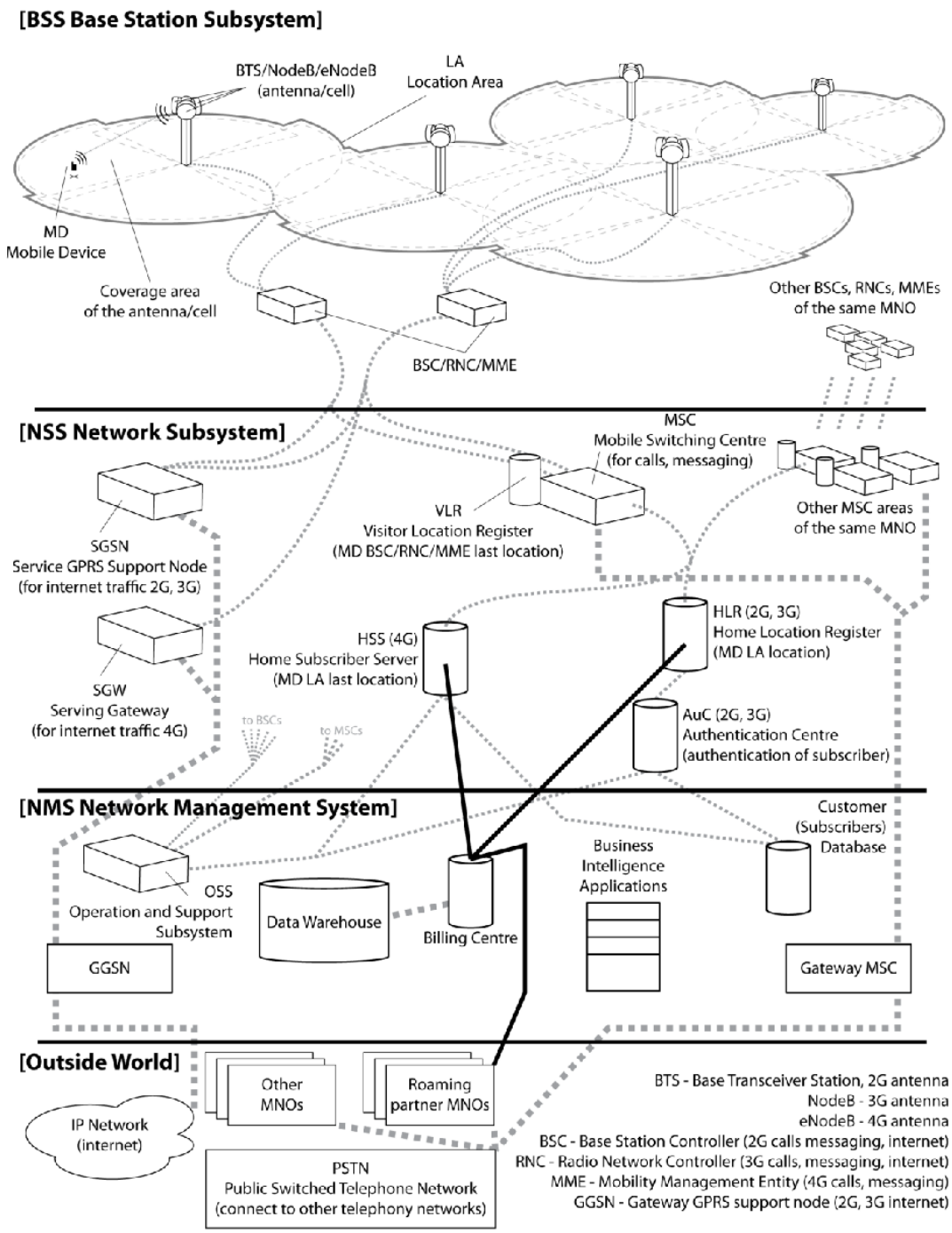


Figure 1. Schematic representation of a mobile telecommunication network (taken from [9]).

### *3.3. Obstacles*

From our experience, we claim that granting access to mobile phone data for official statistical purposes is a business decision, not a technical decision. This decision has been partially positive only in a few countries with some MNOs and only for research under restrictive conditions. The decision for standard production conditions has been clearly negative so far across the ESS.

MNOs not interested on monetizing their data in the form of statistical products show an evident reluctance to collaborate with Official Statistics. For MNOs with this business line, however, an entangled set of risk perceptions lies behind their negative answer. As main issues identified both in the bilateral contacts and in the workshop with MNOs in Luxembourg we highlight the following:

- Legal issues are adduced in terms of lack of clarity of both European and/or national official statistical regulations and in terms of telecommunication regulations. More clarity is demanded. In our positive experiences, the concurrency of Data Protection Agencies has been proved to be essential in this respect providing legal coverage and support. The ESS Task Force on Big Data reports no definitive legal barrier upon analyzing the situation across Europe.

- Intellectual property rights and industrial secrecy requirements are posed to deny access. Since preprocessing is needed by MNOs, they perceive a high risk in disclosing their know-how in their extremely competitive market. Indeed, this potentially collides with the transparency of the official statistical production process. In our view, standard common production processes can be put into place for all MNOs so that no real disclosure threatens their position in the market.

- Costs associated to data retrieval and data preprocessing must be carefully analyzed. Market prices not only of these activities but of data themselves are offered to NSIs. The situation is indeed complex, since MNOs are actually part of the production process. Thus, the associated (marginal) costs, still to be objectively and empirically determined, should be shared by NSIs, but never violating the golden rule in Official Statistics of not paying for the data (this is a public good) [10].

- The public perception of the privacy and confidentiality of subscribers' mobile phone data is posed also as an important issue. Clearly, a communication strategy and visible transparency for citizens are needed. A joint communication strategy must be put into place.

- NSIs seem to be perceived as dangerous competitors to MNOs in their new business line of monetizing their data potentially ruining their activity. More collaboration is needed in this

line around concrete use cases to show empirically that public and private interests are not only different but also complementary.

All in all, the construction of mutual trust through the joint work upon concrete use cases stands up as the most promising solution. Technical solutions do exist for many issues so that the decision on this collaboration is indeed a strategic business decision in the private sector with a unified institutional action in the public sector both at the national and European levels. A high-level institutional action from the public sector is further needed.

## 4. Statistical methodology

New data sources entail also the need for new statistical methods, which is the case of mobile phone data too. For an immediate understanding of these needs, it is highly convenient to have a broad view of the whole statistical production process using these data. This is represented in figure 2.


Figure 2. Schematic representation of the production process with mobile phone data.

Oval shapes represent the sequence of data along the process from the retrieved raw telco data to the final statistical product. In rectangular shapes we denote the production activities already mentioned in section 3. The aggregation of statistical microdata can be undertaken by MNOs or NSIs (preferred option) depending on the operational agreement to access data. A similar structure with many common points is suggested by Eurostat [11]. Notice how the execution role of the

production process steps is now secondary leaving the primary concern on the design of these tasks, which, in our view, must be the responsibility of NSIs.

The understanding of the generation of data and identification of error sources is a key element to achieve high-quality products. This situation is already present with administrative registers, which are formed in terms of *events* (a person is added to an employment register because he/she is demanding a job and later on is unsubscribed because he/she has found the job). Events are then transformed into persons (*statistical units*) and finally used to make an inference about the target population.

We claim that methodological tools such as the two-phase life-cycle model for administrative data can be adapted to the production process with mobile phone data. Indeed, the first steps to identify error sources in the process have been already taken in the project. However, a closer collaboration with MNOs is needed to analyze the initial phases of the process (preprocessing). Thus, we see again a remarkable degree of similarity with administrative data.

The need for new statistical methodology derives from the impossibility of using traditional sampling designs. Three main steps have been identified as crucial in the process [12]:

- The geolocation of network events (calls, SMS, Internet connections…). The network does not provide an exact location of subscribers. This is crucial to assign the spatial attributes to each mobile device. We want to underline that a key ingredient is the combination of telco data with auxiliary information (official data as land use, population figures…). We propose to use Bayesian techniques as natural toolbox to carry out this data integration. Novel computational and statistical skills are needed in the office to accomplish this task.

- The construction of a data model. Mobile phone data are basically a combination of a pseudonymised identification variable and space and time attributes[1] possibly complemented with some other information depending on the agreement with the MNO. This is the basis to construct a database where important statistical concepts are derived such as country of residence, anchor points (home/work/leisure/…), usual environment, trips, etc. which is accomplished using special (possibly machine learning) algorithms. Again, new skills are needed in the office.

- The inference stage. Once data have been aggregated in territorial cells (number of individuals, number of inbound tourists, number of commuters…), these data have to be

[1] Essentially where and when network events occur.

connected with the target population under study (for population statistics, for tourism statistics, for mobility statistics…). The traditional design-based inference techniques are useless. We have proposed to adapt statistical modelling methods already used in the species abundance problem in Ecology. We also propose the Bayesian approach as a natural toolbox to combine mobile phone data with other data sources (such as official population figures, survey data, administrative data …). The proposed hierarchical model rests on two assumptions, namely (i) at a given time period $t_0$ population figures according to both mobile phone data and administrative/survey data can be assimilated and (ii) the mobility patterns are uncorrelated with the specific MNO which individuals are subscribed to. In mathematical terms, essentially, for each territorial cell i the number of individuals is given by:

$$N_i(t_n) = \left[ N_i(t_0) + \sum_{\substack{j=1 \\ j \neq i}}^{I} p_{ji}(t_0, t_n) N_j(t_0) - \sum_{\substack{j=1 \\ j \neq i}}^{I} p_{ij}(t_0, t_n) N_i(t_0) \right]$$

$$N_i(t_0) \cong Poisson\left(\lambda_i(t_0)\right)$$

where the transition probabilities $p_{ij}(t_0,t_n)$ between cells i and j in the time interval $(t_0, t_n)$ and the parameters $\lambda_i(t_0)$ are hierarchically modelled by choosing prior distributions integrating as much auxiliary information as possible in their construction. Notice that we obtain a posterior distribution for the number of individuals in each cell i at every time period $t_n$.

Curiously enough, the statistical models so far proposed points in the same direction of similar approaches with administrative data [13]. Again, a new similarity with these data arises.

## 5. IT infrastructure

The empirical knowledge about the IT infrastructure needed to produce official statistics with mobile phone data is currently severely limited by the restricted access to data in the ESSnet project. However, diverse important issues have been clarified and partially addressed. Most of them are actually conditioned by how data access and data preprocessing are to be conducted [14].

Firstly, from our experience, large amounts of mobile phone data have not been transmitted to NSIs to demand a specific deployment of distributed platforms in the office. In those cases where microdata at the level of mobile devices have been transmitted, these have been processed in more

or less traditional computing environments. For the cases of transmission of aggregated data, the computational requirements coming from data volume are indeed looser.

However, as pointed out in preceding sections, a realistic process will need both data retrieval and data preprocessing in the information systems of the MNOs themselves. This in-situ process step need an agreement regarding both the hardware/software infrastructure and the staff to manage it. Now the technical requirements in terms of data volume, data generation velocity, and data processing are indeed tighter and distributed computing platforms are to be put into place.

In the ESSnet project we have partial experience with such an in-situ platform. This was not a specific infrastructure for this project nor for Official Statistics purposes. An agreement for optimal collaboration conditions must be pursued.

The transition from a local computing environment in a PC with user-friendly tools like Excel to distributed computing platforms in a cluster with client/server architectures and with statistical computing languages (such as Python, R, Scala…) can be deduced from our original software solutions for the geolocation of network events and statistical models under the Bayesian paradigm. The proof of concept of the underlying statistical methodology has been provided with R packages. Their deployment in production at large scale will again need novel skills for the production staff.

## 6. Quality issues

Quality has been a distinction of official statistics with traditional survey and administrative data for the last decades. Furthermore, it should be still the ultimate goal in the production of official statistics with new data sources, in general, and with mobile phone data, in particular.

In the ESS realm, the guiding quality framework is established with the European Statistics Code of Practice (CoP) [15]. Thus, an assessment of the impact of using mobile phone data together with the associated changes in the production must begin with an analysis of the CoP in the light of the novelties brought by these changes. A complete revision of the CoP is beyond the scope of the project, but a first immediate bottom-up analysis of these impacts entailed by the former access issues and the novel statistical methodology has been undertaken [16].

Not all the fifteen principles of the CoP are equally affected but three main details have been identified in the root of all these impacts:

- Part of the statistical production process will have to be executed by the MNOs.
- There will be a partial change in the inferential paradigm.

- There will be an unprecedented scale of spatial and time breakdown in the disseminated statistics.

The access issues described in preceding sections together with the data retrieval and data preprocessing by MNOs necessary in the first stage of the statistical process clearly introduce a novel ingredient in the production of official statistics impinging on quality principles as:

- The professional independence (decisions by MNOs for diverse reasons cannot have an unsupervised influence on the production of official statistics);
- Mandate for data collection (legal support to guarantee the sustainability of the data);
- Commitment to quality (quality policy to be agreed with and respected by MNOs);
- Statistical confidentiality (especially when a combination with official data is undertaken e.g. through a multiparty secure computation channel);
- Impartiality and objectivity (e.g. openly informing about the methodology);
- Non-excessive burden on respondents (burden now placed on MNOs themselves).

The use of statistical models for the production of official estimates forces statistical offices to take decisions about the modelling exercise (prior distributions, etc.) impinging directly e.g. on the professional independence and calling for new indicators about the quality of these decisions (e.g. about the goodness of fit of both data and estimates, which as a matter of fact can be naturally derived from the use of these methods). Most of the CoP is clearly oriented towards the traditional inferential paradigm based on sampling designs. This needs a careful revision to update the CoP.

Relevance, timeliness, and punctuality will be immediately affected in the positive sense by the unprecedented degree of breakdown both in space and time in the statistical outputs. But this also means that statistical confidentiality and accessibility and clarity poses tighter challenges.

## 7. Future research

Invaluable empirical knowledge has been gained in the ESS with this project in many aspects, especially, regarding access to data, needs for novel methodology and IT tools and infrastructure, and a bottom-up revision of the quality framework. The ESS should pursue this direction to build a common production framework with mobile phone data and the research will presumably continue in the forthcoming second ESSnet on Big Data.

Since accessing data and reaching agreements with MNOs have been proved to be very difficult, we are proposing a more modular approach so that the different elements of the production framework (geolocation methods, statistical models, software tools, visualization tools, quality indicators…) can be investigated using simulated mobile phone data in parallel to the painstaking activities of getting access to data. Upon success on these agreements, simulated data can be substituted by real data and the framework will be ready for production.

A decisive highest-level institutional action both at the European and the national levels is strongly advised for the ESS to have access to mobile phone data so that they are integrated into the standard production of official statistics.

## 8. References


[1] Eurostat, NIT, University of Tartu, Statistics Estonia, Positium, IFSTTAT, and Statistics Finland (2014). Feasibility study on the use of mobile positioning data for tourism statistics. Available at http://ec.europa.eu/eurostat/web/tourism/methodology/projects-and-studies.

[2] Netmob (2017). Conference on the scientific analysis of mobile phone datasets. Available at http://netmob.org.

[3] Luca (2018). LUCA – data driven decisions. Available at https://luca-d3.com/technology-smart-steps/index.html.

[4] ESSnet on Big Data (2018a). ESS.VIP Project Big Data: from exploration to exploitation. Available at https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Main_Page.

[5] ESSnet on Big Data (2016). Workshop on Public-Private Partnerships for Mobile Phone Data for use in Official Statistics, Luxembourg, 22-23 September 2016

[6] Laney, D. (2001). 3D Data management: controlling data volume, velocity and variety. Application Delivery Strategies by META Group Inc. (2001, February 6), p. 949. Available at https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[7] ESSnet on Big Data (2018b). Deliverable 5.2. Guidelines for the access to mobile phone data within the ESS. Available at https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/a/ac/WP5_Deliverable_5.2.pdf.

[8] UNECE (2000). Terminology on Statistical Metadata. Conference of European Statisticians. Statistical standards and studies -- no. 53. Available at http://ec.europa.eu/eurostat/ramon/coded_files/UNECE_TERMINOLOGY_STAT_METADATA_2000_EN.pdf.

[9] Positium (2016). Technical documentation for required raw data from mobile network operator for official statistics. ESSnet WP5 internal technical report.

[10] United Nations Global Working Group on Big Data (2016). Recommendations for access to data from private organizations for Official Statistics. https://unstats.un.org/unsd/big-data/conferences/2016/gwg/Item%202%20(i)%20a%20-%20Recommendations%20for%20access%20to%20data%20from%20private%20organizations%20for%20official%20statistics%20Draft%2014%20July%202016.pdf

[11] F. Ricciato (2018). Towards a reference methodological framework for MNO data processing for official statistics. ESSnet BD WP5 Internal meeting 20-21 March, 2018.

[12] ESSnet on Big Data (2018c). Deliverable 5.3. Proposed Elements for a Methodological Framework for the Production of Official Statistics with Mobile Phone Data. Available at https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/4/4d/WP5_Deliverable_5.3_Final.pdf.

[13] Bryant, J.R. and P. Graham (2015). A Bayesian approach to population estimation with administrative data. Journal of Official Statistics 31 (3), 475-487.

[14] ESSnet on Big Data (2018d). Deliverable 5.4. Some IT elements for the use of mobile phone data in the production of official statistics. Available at https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/c/ce/WP5_Deliverable_5.4_Final.pdf.

[15] ESS (2011). European Statistics Code of Practice. Available at http://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-32-11-955.

[16] ESSnet on Big Data (2018e). Deliverable 5.5. Some Quality Aspects and Future Prospects for the Production of Official Statistics with Mobile Phone Data. Available at https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/7/71/WP5_Deliverable_5.5_Final.pdf.