eurostat

Reusing Mobile Network Operator data for Official Statistics: the case for a common methodological framework for the European Statistical System

2023 edition

STATISTICAL
REPORTS

Reusing Mobile Network Operator data for Official Statistics: the case for a common methodological framework for the European Statistical System

2023 edition

This document should not be considered as representative of the European Commission’s official position.

Luxembourg: Publications Office of the European Union, 2023

cover © Shutterstock/Kalpana photographics

Print	ISBN 978-92-68-05648-6	ISSN 2529-5233	doi:10.2785/913276	KS-FT-23-001-EN-C
PDF	ISBN 978-92-68-05647-9	ISSN 2529-3222	doi:10.2785/880645	KS-FT-23-001-EN-N

Table of contents

Executive Summary

The European Statistical System (ESS) aims to produce new statistics with improved timeliness and richer information content, for the benefit of the whole society, by leveraging new sources of data, including data generated and held by the private sector. The location data that are routinely collected by Mobile Network Operators (MNO) – or "MNO data" for short – are among the most appealing candidate sources for (re)use in official statistics – but also one of the most challenging ones.

During the last decade several National Statistical Institutes (NSI) in Europe and worldwide have been experimenting with MNO data through case-studies and research projects spanning various statistical domains (e.g., tourism). Such explorative activities have clearly demonstrated the potential of such data to capture and trace changes in human mobility patterns at scale, thus enabling the production of statistics delivering a dynamic view of population presence and mobility.

Moving beyond explorative activities, research projects, one-off case studies and Experimental Statistics towards *regular production of Official Statistics* based on MNO data, several enabling conditions must be put in place, from establishing sustainable models of data access to defining adequate technical and organisational measures to ensure protection of personal data and business sensitive information. Besides all that, methodological aspects must be addressed to ensure that the transformation of "raw" MNO data into statistical figures complies with the principles and requirements of official statistics.

The present paper focuses specifically on the methodological aspects and provides three main contributions. First, starting from a reasoned analysis of the current landscape, it motivates the need for establishing a **common ESS methodological standard defined at the European level** for the transformation of MNO data into official statistics. Second, it establishes the key high-level requirements that such common ESS standard should fulfill: openness, transparency, modularity, evolvability, multi-MNO orientation (i.e., combine information from multiple MNOs) and supporting the fusion of MNO data with non-MNO data. Third, it provides a bird's-eye view of the ongoing activities in the ESS to advance towards the definition of a common ESS methodological standard for MNO data in official statistics.

The present paper is the result of the work of the ESS Task Force on the use of MNO data for Official Statistics. This Task Force was established in 2021 to address specifically the methodological aspects involved in the (re)use of MNO data for official statistics.

1 Introduction

The European Statistical System (ESS) is undertaking several innovation efforts in response to the opportunities and challenges posed by the new digital data era. As part of this transformative process, the ESS and its members are looking with increasing interest towards the integration of new data sources in the statistical production processes, including data generated and held by the private sector, to deliver new statistical products with improved timeliness and richer information content for the benefits of the whole society. In such context, the location data collected by Mobile Network Operators[1] (MNO) – or "MNO data" for short – are among the most promising and appealing data sources with high potential for integration into the statistical production process, despite the challenges and open issues that will be elaborated in this paper.

During the last decade several National Statistical Institutes (NSIs) have been experimenting with MNO data through case-studies and research projects spanning various statistical domains, e.g., in the fields of tourism and demography. Such explorative activities have clearly demonstrated the capability of such data to capture and trace changes in human mobility patterns at scale, thus enabling the production of statistics delivering a *dynamic* view of population presence and mobility. In a few cases the collaboration with MNOs has led some NSIs to publish *Experimental Statistics*[2] based on MNO data. The covid-19 pandemic has highlighted the critical and urgent need for producing timely figures about human presence and mobility, accelerating the interest in statistical products derived from MNO data. During the early stages of the covid-19 crisis, in the absence of official statistics and established quality assurance procedures, critical decisions by the local public authorities on containment measures had in some cases to rely on commercial data of unknown quality provided directly by private companies, produced with proprietary and not completely transparent methods, hence side-lining the NSIs. In other cases, the increased pressure resulting from covid-19 has led to the establishment of new collaborations between MNOs and NSIs, which helped the latter to gain a better understanding of the methodological aspects involved in the processing of such data. As a matter of fact, during the most hectic moments of the covid-19 crisis the timely *availability* of relevant data was given priority, while the issue of their *quality* and *reliability* was often set aside based on the claim that in a contingency situation "*any data is better than no data*". But as we now move forward to consider *regular* production of statistics leveraging the richness of MNO data in non-crisis times – not least to prevent, mitigate or anyway get better prepared for the next crisis – issues of quality, reliability and continuity of these statistics cannot be evaded.

The ESS aims to move beyond explorative research projects, one-off case studies and *Experimental Statistics* towards

[1] Unless differently specified, throughout this paper we use the term Mobile Network Operator (MNO) to refer to the operation of a particular telecom company in a single country. Therefore, one international telecom company operating in multiple countries will be considered as multiple MNOs. This approach is motivated by the need to account for country-level aspects (e.g., national legislation, national policies, national authorities) that do not necessarily allow a full integration of network infrastructures and business processes across different countries, and may possibly lead the same company to follow different management and business strategies in different countries.

[2] The term *Experimental Statistics* refers to statistics that have not reached full maturity in terms of harmonisation, coverage or methodology. For an overview of Experimental Statistics with several examples the interested reader is referred to: https://ec.europa.eu/eurostat/web/experimental-statistics

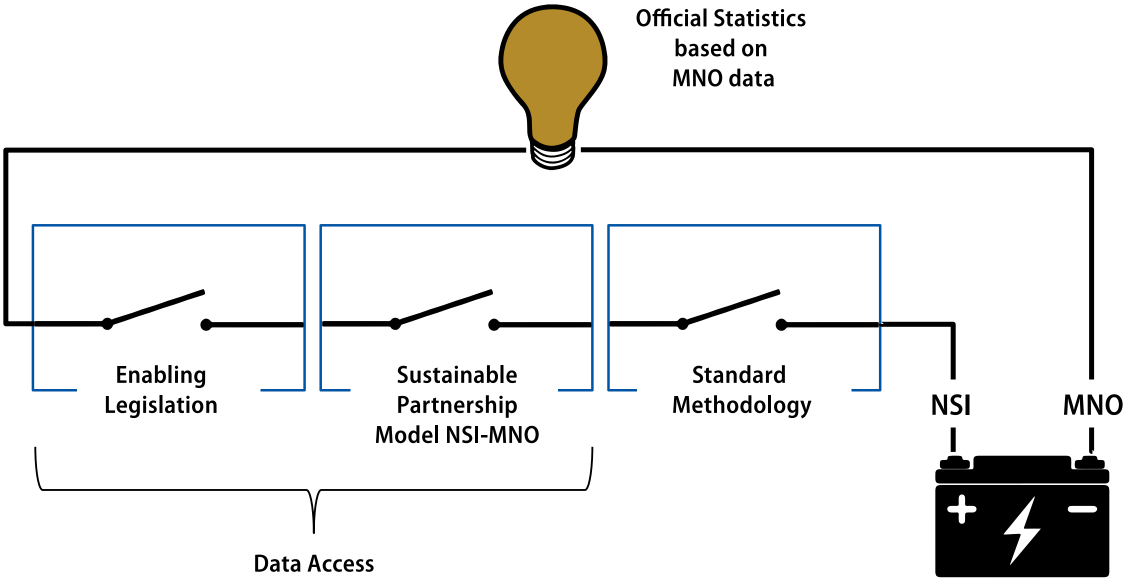
regular production of Official Statistics based on MNO data. This transition is proving quite challenging, due to several distinct but inter-related open issues on the side of data access (legal and business aspects) as well as on the side of data processing methodology. The current situation can be figuratively associated to the analogy sketched in Figure 1: like the bulb in the electrical series circuit will not light on until all switches are closed, regular production of official statistics based on MNO data cannot come to light until all legal, business and methodological issues are solved.

For all these issues, providing a common answer within the ESS is by far more advantageous and effective than seeking independent solutions at the national level. Thus, in full respect of the subsidiarity principle, the ESS is currently working to address these issues at the European level. One line of action focuses on methodological aspects. To this end, the ESS *Task Force on MNO data for Official Statistics* (TF-MNO for short) was setup in 2021 with the mandate to coordinate and steer methodological development related to MNO data within the ESS; to improve coordination and sharing of knowledge and best practices from national experiences; and to advance towards the definition of a common methodological framework for the whole ESS. The present position paper marks a milestone in the ongoing work of the TF-MNO.

Consistently with the TF-MNO mandate, this paper focuses exclusively on the methodological aspects. The main goal is to present the current view of the TF-MNO, advocate the need for a common and open ESS methodology, and outline the key requirements and motivations thereof. For the sake of casting the discussion about methodological aspects in a wider context, the paper touches briefly on the closely related aspects of data access, with a dedicated chapter providing a high-level view of the current orientation of the work by the ESS, along with references to recent documents.

FIGURE 1

Closing the circuit: regular production of Official Statistics based on MNO data can be turned on only when all open issues are "switched on" to a solution. The definition of a suitable methodology is a logically separate issue from the definition of sustainable data access model, which implies the establishment of sustainable partnership models between NSI and MNO in the framework of enabling legislation.


Official Statistics based on MNO data
Enabling Legislation
Sustainable Partnership Model NSI-MNO
Standard Methodology
NSI
MNO
+
−
Data Access

2

What are MNO data and why they are relevant for Official Statistics

Almost every person in Europe nowadays carries one or more personal mobile device(s) with a subscription to some Mobile Network Operator (MNO). To support the various forms of mobile communications (i.e., making or receiving voice calls, exchanging SMS, browsing the Internet, running smartphone apps, etc.) the mobile network must keep track of the radio cell to which the mobile device is connected. To this aim, the mobile communication protocols entail frequent interactions between the mobile devices and the network infrastructure, and every interaction reveals to the network the identifier of the radio cell that was serving the mobile device at that time, among other parameters. By associating the cell identifier to the geographical position of the cell, such events reveal the *approximate* position of the mobile device, hence of its user.

Depending on the configuration of the MNO service and infrastructure, certain types of interactions are recorded and stored for purposes related primarily to mobile service delivery, from billing to network operation and troubleshooting. These data are called "MNO data" because they are collected *on the side of the mobile network* and therefore are held by the telecom operator. MNO data should be distinguished from other types of location data generated *on the side of the mobile device*[3], e.g., GPS data points collected by apps or by the operating system and then delivered to platform operators[4]. We use the term "MNO data" to refer generically to all location data[5] collected on the side of the network, including billing data, i.e., the so-called Call Detail Records (CDRs) and their extensions for data connection (sometimes referred with the terms Data Detail Records, DDRs, or eXtended Detail Records, XDRs) as well as the more informative[6] *signalling data* derived from signalling traffic.

The geo-localisation of individual events (i.e., single CDR or signalling record) requires another stream of MNO data, namely radio network configuration data. Such data encode the geographical position of radio cells along with other radio parameters and are essential to reference geographically the individual records. In the simplest form, radio network configuration data reduce to antenna tower coordinates, but recent research work by Ricciato et al. (2015), Tennekes at al. (2022) and Ricciato et al. (2023) shows that more detailed

[3] The term Mobile Phone Data (MPD) has been used by different authors and across different papers to refer to both types of location data, those generated on the side of the network and of the device. As only network-side data are relevant for this document, to avoid any ambiguity we adopt the more specific term "MNO data".

[4] Some telecom operators invite their mobile subscriber to install their app on their personal devices, and in some cases such apps report GPS location data collected on the side of the mobile device. In this special case the app operator corresponds to the telecom operator.

[5] We focus here on location data. Other types of data held by the telecom operators that are not directly related to the instantaneous location of the mobile devices fall outside the scope of the present contribution. Examples of these other data include customer data (e.g., name, address, demographic details, contact details and bank account of individual mobile subscribers) and caller/callee lists. In principle, some of these data may be used under certain conditions as auxiliary sources, in combination with location data, to improve the quality and enrich the information content of the final statistics. However, doing so poses additional challenges, not least because the mobile subscriber often does not correspond to the actual mobile user. Further exploration of these aspects is left as a point for further study and is not covered in the present paper.

[6] Signalling records are considerably more frequent than CDRs. While CDRs are triggered by events associated to human activities (e.g., making or receiving a phone call or SMS, starting a data connection) signalling events are triggered automatically by the mobile device or by the network and are only loosely coupled with user activity. In one day, a single mobile user may generate several tens of signalling records but only a handful of CDRs.

radio network information (e.g., cell size, tilt and azimuth orientation, operating frequency) can be translated into better spatial accuracy through appropriate probabilistic methods. In the following we will use the term "MNO data" to refer generally to the combination of signalling records or CDR with network configuration data.

In summary, the collection of MNO data embeds information about the (approximate) position and movements of the entire pool of MNO customers, which could serve as basis for extrapolating statistics about presence and mobility patterns of the whole human population in the MNO service area. In principle, even considering the data from a single MNO, the resulting statistics would feature very high degrees of spatial coverage (whole country), temporal continuity (24/7), timeliness (quasi-real time) and population coverage (roughly equal to the MNO market share). Along each of these dimensions, MNO data offer unprecedented possibilities when compared to traditional survey data and travel diaries. On the other hand, transforming MNO data into meaningful and reliable statistics is a difficult and complex task, with several methodological challenges along the path from *raw* MNO data to final statistics.

3 Methodological challenges

In this section we provide a bird's eye account of some fundamental methodological challenges involved in the analysis and interpretation of MNO data. Further high-level methodological requirements are elaborated later in Section 7, while other (non-methodological) issues related to data access, privacy and business aspects are discussed separately in Section 10.

As explained in the previous section, MNO data are collected primarily for purposes related to the delivery of mobile services, and therefore they are designed and optimised for those primary purposes. When seen from the perspective of (re)using such data for producing official statistics (as secondary purpose) some of their characteristics constitute obvious limitations that must be dealt with in the data processing stage. Following Ricciato *et al.* (2015) we distinguish between three main sources of uncertainty affecting MNO data:

- **Temporal uncertainty**: the position of the mobile device is not observed continuously in time but only at discrete observation times that correspond to the occurrence of certain *events* triggering the exchange of signalling messages between the mobile device and the mobile network. The frequency of such events depends on a complex interplay of many factors, including the recent activity of the mobile user, the position of the mobile device, the local configuration of the radio network, etc. Even within the same network, the frequency of events produced by a single mobile user varies greatly, from many events per hour down to only one every few hours. In other words, the position of the mobile device is sampled in time, but the sampling process is uneven, non-uniform and in general not independent from the process we wish to observe. This results in possible biases and incomplete information. In fact, in-between subsequent event times the mobile device position cannot be observed directly but only inferred (or guessed) based on the available data points. Furthermore, when the mobile device is disconnected (e.g., switched off or located outside the mobile service area) the network has no way to learn its position and therefore the device is unobserved, i.e., it disappears completely from the MNO data set.
- **Spatial uncertainty**: the mobile device location is not known precisely as a point in space, but only through the identifier of the serving radio cell that, in turn, maps to the geographical space served by the cell, i.e., the so-called *cell coverage area*. The size of the radio cell spans from a few meters (picocells and femtocells) through hundreds of meters or a few kilometres (micro cells and macro cells in urban areas) up to tens of kilometres (large umbrella cells in rural areas). Even if the radio cell coverage area was known without error, there is no means for the network to know the precise point position of the mobile device within the cell coverage area. On top of that, in practice the actual cell coverage area can be estimated only approximately, and coverage prediction errors add up to the spatial uncertainty budget.
- **Population coverage uncertainty.** The mobile network tracks mobile devices, not people, and people do not map 1:1 to mobile devices. There are persons that do not carry any mobile device (e.g., children) while other people carry two or even more mobile devices, possibly subscribed to different MNOs (e.g., one for work and another for private use). There are also mobile devices that are not associated to any person, often called Internet-of-Things (IoT) or

Machine-to-Machine (M2M) devices[7]. In other words, the *target population* of statistical units (i.e., people) does not correspond to the *observed population* (i.e., mobile devices). Furthermore, both populations change in time, with fluctuations in the observed population due, e.g., to mobile users changing their mobile subscriptions (mobile customer churning) or making use of new devices only for short periods (e.g., foreigners purchasing a local SIM during their stay in the visited country).

The need to deal with these sources of uncertainty introduce additional challenges in the methodological design process and complicate the task of transforming MNO location data into reliable statistics. In particular, the need to project figures about a fluctuating population of mobile devices onto a larger and more stable population motivates the efforts towards (i) the integration of data from multiple MNOs and (ii) the combination of MNO data with other non-MNO data sources – these aspects are discussed later in Section 8 and Section 9, respectively.

Another important difference between MNO data and more traditional statistical data sources (survey or administrative register data) is constituted by the lack of socio-demographical variables. In fact, while basic socio-demographical variables are typically collected by the telecom operators for their customers, there is no guaranteed correspondence between the mobile *user* who carries the mobile device and the mobile *subscriber* who establish the contractual relationship with the MNO (think e.g., to the employer subscribing mobile service contracts for his employees', or to one person subscribing for other family members or relatives, not necessarily in the same household). Therefore, the customer data may be rather inaccurate when taken as proxy for mobile user data. Moreover, the linking between customer data and location data may be regarded as a more critical operation in terms of data protection compliance and protection of business sensitiveness. The limitations of customer data, coupled with the costs and barriers to their secondary (re)use, may well justify resorting to alternative solutions for obtaining socio-demographic variables based on other non-MNO data, as elaborated in Section 9.

[7] Data points from IoT/M2M devices are not relevant for deriving statistics about human mobility and represent a source of confusion. Such data points should be ideally pruned from the data set, but it is not always easy to identify them precisely. Therefore, the presence of some residual data from IoT/M2M device may not be completely excluded.

4

MNO data analytics in the private sector: current state of play

The first research studies leveraging MNO data to study human mobility date back to the late '90s and were typically based on CDR data sets from a single MNO. CDR are by far the simplest form of MNO data, the cheapest to extract from the network infrastructure and the easiest to work with for the researchers. Signalling data are more voluminous, frequent, and informative, but also more complex to extract and to process than CDRs. While the first studies relied exclusively on CDR, research work on signalling data started to appear only later in the mid '00s (see Janecek *et al.* (2015) and references therein). It is somewhat remarkable that the analysis of MNO data took off – and indeed became a fashionable research topic – across a wide variety of research communities, from physicists to geographers, from urban planners to demographers, from traffic engineers to computer scientists, and others. This resulted in a large volume of research papers dealing with MNO data. Such large mass of research work represents a useful knowledge basis but does not offer a consistent and ready-to-use general methodology that is directly applicable to regular statistical production. In fact, research works typically focus on one-off case studies and the analysis methods elaborated there tend to be highly specific *(i)* to the purpose of the study and *(ii)* to the characteristics of the specific data set at hand.

During the last decade MNO data analytics quickly moved out from research labs and became a recognisable business sector that is now growing. On the offer side, we can distinguish two main types of commercial suppliers of analytic products (insights, reports, dashboards, etc.) based on MNO data:

- MNOs themselves: several telecom operators, typically the largest ones, have internal departments or "labs" dedicated to transforming the location data generated by their own network into analytical products and services offered to external customers.
- Specialised companies, external to telecom operator organisations, offering analytic products and services based on the location data collected by MNOs with which they maintain some form of partnership or other kind of business relation. We shall refer to such companies by the term *Mobile Data Analytic Provider* (MDAP for short). MDAP companies are often founded as spin-offs of research teams from public or private laboratories, or as daughter companies of MNOs. Their size is typically small[8]. The business relationships and the partnership models between MNOs and MDAPs vary and may or may not include exclusivity clauses (e.g., preventing one MDAP to partner with mutually competing MNOs, or vice-versa). Unless differently specified, the MDAP term will be used in this paper to refer to both for-profit and non-profit companies[9].

On the demand side of the data analytics market, both private and public organisations are potential customers of MDAPs and MNOs. Examples of public organisations include national bodies (e.g., ministries, national agencies) and local authorities (e.g., tourist offices, city planning departments). Private customers include business consultancies, retailers, construction companies, etc.

[8] Small companies have up to 50 employees and turnover up to 10 million euro, see e.g. the conventional criteria in https://single-market-economy.ec.europa.eu/smes/sme-definition_en

[9] Non-profit organisations offering analytic services based on MNO data often operate in close collaboration with research teams and may form ad-hoc consortia on a per-project basis with MNOs and other partners (e.g., international organisations, local government bodies, NGOs). Their activity tends to be concentrated in developing countries.

5

MNO data and official statistics: current state of play

A bird's eye view of the national activities on MNO data by NSIs over Europe shows a rather heterogeneous and fragmented picture. Not in all European countries the NSIs have been successful to establish some form of collaboration with some local MNOs. Out of those that have managed to do so, there are cases where an established collaboration was later terminated, e.g., due to change of MNO strategic orientation and/or management. The main causes for failing to establish durable NSI-MNO collaboration can be summarised as follows.

- **Legal compliance issues**. Reusing MNO data for Official Statistics touches three legislative domains: data protection legislation (GDPR and national applications thereof), statistical legislation (mainly national) and telecom legislation (ePrivacy Directive[10] and national transpositions thereof). The articulation of national legislations and national applications of European legislation across these three domains creates a very fragmented scenario across the different European countries, with varying degrees of legal support (or lack thereof) for NSI-MNO collaborations. At one extreme, in some countries legal compliance may represent the main obstacle against the reuse of MNO data for (the development of) official statistics. At the opposite extreme, in other countries the legal situation appears to be very favourable, with the national legislation encoding some kind of obligation for MNOs to provide data to the NSI. The situation of most countries is somewhere in between these two extremes, with the issue of legal compliance being neither unsurmountable nor straightforward to sort out.
- **Business aspects**. In most cases NSI-MNO collaboration may only happen on a voluntary basis. Different MNOs have different views about the balance of costs vs. benefits of collaborating with the local NSI even for experimentation and methodological development purposes[11]. Some MNOs demand considerable financial compensation by NSIs and consider them only as potential *customers* to raise financial profit from. Other MNOs are eager to collaborate with NSIs as potential *partners* and value the non-financial benefits resulting from such collaboration, e.g., potential improvement of analytic products and methodologies, or for corporate social responsibility. In several cases MNOs collaborate with NSIs in publicly funded research projects, often along with other partners in larger consortia, e.g., academic institutions and public agencies.
- **Methodology aspects.** Only in rare cases the NSIs have access to the "raw" location data generated by the mobile network infrastructure and therefore can develop – or better co-develop in collaboration with the MNO staff – complete methodologies for transforming the data into statistics. In most cases MNOs tend to apply their own processing methods for producing final or intermediate aggregate data that are then passed to the NSI. Two aspects must be considered in such cases: (i) how much detailed information is *disclosed to* the NSI about the proprietary

[10] Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32002L0058.

[11] This applies to different MNOs operating (and competing) in the same country as well as to MNOs operating in different countries under the same holding or partnership agreement. Recall that in this paper the term MNO identifies the operation of a single company in a single country.

methodology applied by the MNO; (ii) how much the NSI can *influence* (define, vary) the processing methodology. Again, national experiences show different grades of openness and attitudes to collaboration by MNOs.

In addition to the activities at the national level, the ESS has funded two projects at the European level, namely the ESSnet Big Data I (2016-2018) and ESSnet Big Data II (2018-2021), with work-packages dedicated to MNO data where experts from different NSIs worked together to build knowledge and advance methodological development[12]. The work in those projects was focused on a single important use-case, namely the dynamic estimation of present population, for which novel solutions based on probabilistic methods were developed and tested on simulated data[13]. Unfortunately, during those projects it was not possible to involve MNOs directly or gain access to real-world test data[14], nonetheless the project team managed to lay down the high-level view and the initial basis for a common methodological framework.

[12] The project results and deliverables concerning MNO data can be publicly accessed from https://ec.europa.eu/eurostat/cros/content/wp5-mobile-phone-data_en and from https://ec.europa.eu/eurostat/cros/content/WPI_Mobile_networks_data_en respectively, for the project ESSnet Big Data I and ESSnet Big Data II.

[13] An agent-based open-source simulator was developed in the project in order to support the methodological research. The simulator is freely available from https://github.com/MobilePhoneESSnetBigData/simulator.

[14] Like in many other research fields, testing on synthetic data is complementary to, not a replacement for testing based on real-world data. The exploration and assessment of certain methodological aspects require the researcher to maintain full control over the testing conditions (e.g., testing across different parameter settings, comparison against exact ground truth) and therefore can be best addressed with synthetic data. Conversely, other methodological aspects (e.g., validation of model assumptions) require access to, exploration of and testing on real-world data or samples thereof. Therefore, a solid methodological development process must rely on both kinds of data, real and synthetic, for different tasks.

6

The case for methodological standardisation

Official statistics products are defined to *satisfy an information need* by the prospective statistical users (e.g., policy design or assessment) related to some social, economic, or environmental phenomenon of interest. The choice of the statistical products aiming to describe such phenomena must consider several feasibility and cost constraints about the data that could be realistically made available to the statistical production process (by primary collection or secondary reuse). In other words, the statistical product is not an end in itself, but rather a means towards the fulfilment of the target need.

Cost and feasibility constraints influence all aspects of the statistical product design, including *what*, *how* and *how well* target quantities related to the phenomenon of interest is to be measured. In fact, *definitions* (*what* to measure), *methods* (*how* to measure it) and *quality aspects* (*how well* it can and should be measured) are closely interrelated aspects that should be ideally specified together.

Casting this general consideration into the specific context of designing new statistics based on MNO data, leads us to combine **definitions**, **methods**, and **quality** aspects into a single holistic design task. With that in mind, throughout this paper we use the term "*methodology*" in a broad sense to refer to the unified and detailed specification of the three following aspects: (i) the final statistical *product* to be derived from MNO data, including the definition of the quantities (variables) of interest along with their spatial and temporal granularity); (ii) the *methods* by which said product is to be derived from the input MNO data; and (iii) the *quality* requirements related to the whole production process[15]. Therefore, establishing a *methodology* for MNO data amounts to jointly establish (i) **statistical definitions** and indicators that are *coherent with the target information need and at the same time consistent with the nature and the information content of MNO data*; (ii) a **methodological pipeline**[16], i.e., a structured and modular workflow of methods and intermediate data elements by which those statistics and indicators are to be derived from MNO data; and (iii) an overarching **quality framework** providing a set of requirements – on the input data, on the methods, on the production process and therefore the final output – and possibly guidelines for continuous improvement.

We highlight the importance of defining explicitly reasonable **quality requirements on the input data** that are pragmatically balanced between the conflicting goals of (i) not causing excessive additional burden on the data provider, and (ii) enabling the production of statistics with sufficient levels of quality.

[15] In the context of the traditional data sources used in official statistics, the quality aspects are typically addressed in a stand-alone set of guidelines, i.e., the so-called Quality Assurance Framework (QAF), that complements the description of methods and definitions. In the scope of the present paper, we consider the definition of quality aspects (guidelines, requirements, criteria) as a sub-task within the general methodological development process. In other words, the quality framework is seen as one part of the overall methodological framework.

[16] We use the term "methodological pipeline" to refer to the detailed definition of the data workflow as a modular chain of methods and data elements that are specified unambiguously. The notion of methodological pipeline refers to a set of methodological choices and methods and is technology-agnostic, i.e., it does not imply the use of any specific software platform or programming language. One methodological pipeline may correspond to multiple different implementations thereof, based on different data processing technologies and platforms. The term "software pipeline" (IT object) may be used to refer to those specific implementations of the same reference "methodological pipeline" (methodological object).

Given the enormous amount of data at hand and their intrinsic complexity, the transformation of raw MNO data points into final statistics must be automatised to the largest possible extent. Therefore, a large part of the statistical methodology that specifies how such transformation should take place must be encoded into software components. But *automatisation* implies *formalisation*, since being *executable* by machines implies being *understandable* by machines in the first place. Therefore, the whole data workflow, with all methods (algorithms, functions, etc.) and intermediate data elements occurring along the methodological pipeline must be necessarily represented in some formal language (source code, ontologies, etc.). In addition to enabling automated execution, the *formalisation* and *softwarisation* of statistical methodologies have other important implications, for instance in the direction of better harmonisation and **standardisation of statistical processes** (across different MNOs and countries) and **improved transparency**. Related to the latter point, it seems natural to expect that NSIs shall publish open-source reference code as an integral part of methodological documentation.

When the goal is to perform a one-off study, as in most research work and scientific literature, the applied methodology may well tolerate a non-negligible amount of manual configuration, interactive analysis and ad-hoc adjustment depending on the results. In such case the code that embodies the statistical methodology (or part thereof) tends to be poorly reusable for other processing tasks, and even the application of the same code to other data sets might require another round of manual (re)configuration, analysis, and adjustment. Instead, if the goal is to offer analytic services on a regular basis to multiple customers, as is the case for the business model of MNOs and MDAPs, one may expect the processing code to be (i) highly automated, with much less need for manual adjustment and configuration, and (ii) architected and written in highly professional way to facilitate maintenance, reuse, and expansion.

As discussed earlier in Section 4, the mobile data analytics market is competitive and still relatively young. At the current market stage, every MNO or MDAP company offers its own portfolio of analytic products based on own proprietary methodology. Each company regards its methodology as a business asset to protect from competitors. The high-level definition of the analytic product may be adapted to specific customer needs (e.g., spatial and temporal granularity of the final aggregation, classifications of sub-population of interest) and may be documented publicly to some extent. However, the largest part of the core methodological design choices is often considered a proprietary business secret and remains undisclosed.

The adoption of diverse proprietary methodologies may not be critical in commercial applications if the analytic product that is delivered to the customers meets their demands and expectations. On the other hand, methodological heterogeneity and lack of transparency are problematic for official statistics. In other words, **harmonisation** and **transparency** are key qualifiers of *official* statistics vis-à-vis commercial statistics. In facts, if the analytic products (statistics, indicators) defined by different MNO/MDAP are not sufficiently well aligned, they cannot be *compared*, let alone be *combined* with each other. We can compare the views (final or intermediate data) offered by different MNO/MDAP only if they achieve a minimum level of coherency, i.e., if they are based on the same definitions, classifications, methods, spatiotemporal grids and so on. The same applies *a fortiori* if we want to *combine* the views by multiple MNOs/MDAPs, for instance to gain a more complete view of the mobility patterns within one country, or to better measure international mobility flows across different operators (more on these aspects in Section 8). Therefore, *comparability* and *combinability* of statistical products across different MNOs require, at least, very close alignment of definitions and specifications. But if we consider the path from "raw data" to "final statistics" as a staged (or layered) process, we immediately recognise that the definitions at one stage (or layer) depends closely on the definitions at the previous stage (or lower layer), and so on. For example, if we aim at measuring "*how many people visit regularly a certain place*", we need to provide operational definitions for the notions of "*visit*", "*regular*" and "*place*", among other things. And these notions will be probably expressed by referring to the notions of "*frequency*", "*duration*" and "*location*". It is important to remark that **these quantities cannot be observed directly, continuously, and exactly because of the intrinsic limitations of MNO data** discussed earlier in Section 3. They may be only inferred (or guessed) based on the available MNO data points by resorting to heuristic methods and inference models based on assumptions. Consequently, even if two MNO/MDAP companies adopt the same *nominal* definition of "*visit*" and "*regular*", but still rely on different definitions of "*frequency*" and "*duration*", they will be effectively producing two rather different analytic products. Therefore, to enforce comparability and combinability, the alignment among MNO/MDAP should extend also to the more basic notions of "*frequency*" and "*duration*". But the latter cannot be measured directly and continuously, due to the temporal uncertainty affecting the data, and are typically derived from the sequence of the available data points according to some heuristic logic that is typically encoded in a piece of software. Therefore, aligning these definitions across different MNOs implies adopting the very same logic.

In this way it becomes evident that truly aligning analytic products (statistics, indicators) requires adopting the same methodology *as much as possible* along the whole methodological pipeline. In other words, as far as MNO data are concerned, *measuring the very same thing implies measuring in the very same way.*

The above considerations make the case for applying a single common methodology, for a given target statistics, to data originated by different MNOs. On the other hand, the specification of a common standardised methodology should be pursued with a grain of salt. While it is desirable in principle to maximise as much as possible methodological harmonisation across MNOs, at the same time it must be considered that there are practical limits to the level of methodological uniformity that can be achieved in practice. Such limits stem from various MNO-specific aspects that affect the "raw" data generation stage and introduce a certain degree of variability across data generated by different MNOs. In fact, neither the format (syntax) nor the information content (semantic) of the "raw" data extracted from the operational infrastructure can be assumed to be perfectly uniform across MNOs, as the data generation process involves pieces of proprietary equipment and proprietary configurations, of the mobile services as well as of the underlying infrastructure, and is closely interlinked with business-related aspects that differ across MNOs. Such heterogeneity in the technical and business characteristics of the MNO infrastructure has an influence on the information content of the "raw" data points (including their spatial granularity and temporal frequency, among other dimensions). In other words, some **leeway must be designed into the common standard** to accommodate for the unavoidable heterogeneity on the side of data generation. This consideration should not be taken as a counterargument against standardisation, but rather as a high-level requirement to be considered in the process of standard design.

In addition to maximising methodological *harmonisation*, the ESS considers of utmost importance the maximisation of methodological *transparency* as a necessary condition to ensure *interpretability* and *reliability* of the result, in line with the prescription by the European Statistics Code of Practice[17]. In fact, an interesting pattern (or lack thereof) observed in the final statistics may represent a genuine phenomenon occurring "on the ground" that was captured by the data, or it may be due to some artifact introduced unintentionally by the data processing logic. The risk of the latter is non negligible considering the complexity of the processing methodology for this type of data – complexity that results also from the need to deal with multiple sources of uncertainty and limitations of the data, as discussed above. The complete and detailed knowledge of the underlying methodology by which the statistical result is produced is a necessary pre-condition for the correct interpretation of that result. As the official statistical results are made openly accessible, so should be the methodology that was adopted to produce those results. For a methodology that is designed to be executed by machines and is therefore encoded into some formal languages (software code, ontologies), ensuring methodological transparency entails *inter alia* the public release of open-source reference code.

All these aspects – comparability, combinability, transparency and interpretability – may not be critical in business applications, and therefore resorting to proprietary undisclosed methodologies may not problematic as far as commercial analytics are concerned. However, the situation is markedly different in official statistics, where these aspects are of fundamental importance and indeed represent distinguishing elements of *official* statistics versus commercial analytic products.

[17] European Statistics Code of Practice – revised edition 2017, https://ec.europa.eu/eurostat/web/products-catalogues/-/KS-02-18-142

7

An open methodological standard for official statistics based on MNO data

Casting the above considerations in the context of the European Statistical System (ESS) implies that, should the ESS undertake regular production of *official statistics* based on MNO data – and not merely *experimental statistics* – such statistics must be based on a *common* and *open* methodology defined and applied consistently across the European space. This requirement is further reinforced by the importance of adopting a multi-MNO perspective, within and across different countries, as elaborated separately in Section 8.

Moving beyond experimental statistics and one-off feasibility studies towards *regular* production of *official* statistics requires putting in place several enabling conditions around MNO data access for official statistical purposes, from legislation to cost compensation aspects. These aspects are discussed separately in Section 10. The remainder of the present section focuses instead on the development of a common and open ESS methodology, **under the working assumption that a sustainable data access model has been established** based on a fair partnership between MNOs and statistical offices – in other words, returning to the electrical circuit analogy shown in Figure 1, we address here the "methodology switch" under the working assumption that all other "switches" are successfully closed.

The development of a common ESS methodology for MNO data processing should be seen as a **co-development** task involving statisticians and industry specialists, to be conducted through the collaboration of ESS experts and industry experts from MNO/MDAP. Only in this way the methodological development process can truly consider the needs of official statistics alongside the peculiar aspects and limitations of real-world MNO data. Therefore, we envision a common open methodology for the (re)use of MNO data for official statistics that is (co)developed by the ESS in collaboration, or at least in close consultation with industry experts. The ownership of said methodology should stay with the ESS that should also take care of its maintenance and continuous adaptation to follow the physiological evolution of mobile technologies and business models, to the extent that such evolution influences the generation patterns, hence the information content, of future MNO data. Said methodology should be **formally adopted by the ESS**, and ideally endorsed also by the industry, as a standard for all official statistics based on MNO data within the European space[18].

The ESS methodology should be **fully open**, **transparent** and **reproducible**, in other words it shall represent an **open standard**. As such, it must be documented in a way that allows external parties (e.g., MNOs, MDAPs, researchers, NSIs) to reproduce it independently, i.e., develop independent software implementations that are fully compliant with the open standard. Compliance is meant here in the sense of functional equivalence: a software implementation is compliant to the standard methodology if it always delivers the very same statistical result for the same input data. To this aim, it is important to ensure that a **reference software implementation** and **reference test data sets** are developed

[18] It may be expected that the ESS methodological standard will have some influence on the methodological progress also outside Europe. However, it should be noted that regional peculiarities in the characteristics of the mobile telecommunication market, hence on the characteristics of the generated MNO data, may justify different methodological choices in different regions. In other words, at least for certain methodological components, a solution tailored for the European scenario may not be optimal for other extra-European regions, and vice versa.

and released as part of the documentation package for the methodological standard. Furthermore, it may be necessary for the ESS to also develop a suite of **formal rules and functional test procedures** to verify compliance (i.e., functional equivalence) to the standard ESS methodology. With all that in place, it may be expected that different industry players will develop and offer **proprietary software implementations that are compliant (functionally equivalent) to the standard methodology**, but whose use is subject to licensing on commercial terms. Such proprietary software implementations may compete along other business dimensions (e.g., deployment and maintenance cost, scalability, offer of auxiliary non-standard features) but their deployment would not jeopardise methodological transparency and interpretability. Furthermore, the adoption of an open standard methodology does not jeopardise the possibility for industry players to develop a competitive market of **software products and services** around it, on the contrary, it may foster the development of such a market[19].

The data workflow from "raw" MNO data to one final statistical product can be represented as a chain of data processing blocks composed in a single methodological pipeline, as represented graphically in Figure 2. It is essential to design the overall methodology in a highly modular way[20]. The functional specification of each block and of the interfaces between the blocks is part of the methodological specification. In general, part of this methodological pipeline may be executed at MNO premises, up to a certain "split point". The resulting intermediate data are then passed to the NSI for further processing. It is important to highlight that **standardising the methodology amounts to specifying the whole methodological pipeline**, i.e., the function and methods for each data processing block, **regardless of the physical location where each block will be executed**. In other words, the task of designing the methodological pipeline is logically antecedent and independent from the decision of where the execution of each component is placed. Generally speaking, it may be expected that the initial processing blocks upstream and the final processing blocks downstream will be executed at the premises of the MNO and of the NSI, respectively. It is also possible that some intermediate blocks performing combination of confidential information from different sources e.g., fusion between data from multiple MNOs and/or between MNO data and NSI data, will be executed in secure computation environment designed around PET technologies (to avoid excessive graphical burden, Figure 2 does not represent technological measures that may be adopted to protect data confidentiality of selected blocks or sequences thereof).

Once the complete pipeline is specified, the split point in processing execution between the MNO's and NSI's premises may be determined based on other considerations, e.g., availability of IT infrastructure, costs, risks to data confidentiality, etc. The "optimal" choice is to be agreed between the NSI and the MNO. In principle, different MNOs may agree on different split points with their respective NSIs. We expect that in practice the same split point will be agreed with all MNOs in the same country, but strictly speaking this choice is a matter of agreement, not methodological standardisation.

The standard ESS methodology should be sufficiently flexible to work with different MNOs across the European space. It is important to consider that each MNO builds, configures and operates its mobile network infrastructure, and the mobile services running over it, independently and in general differently from the other MNOs. As discussed in the previous section, the detailed characteristics of the "raw" data are influenced by multiple MNO-specific technical and business choices, and such heterogeneity is reflected into the data. Furthermore, while the mobile communication protocols shaping the interaction between mobile devices and networks are highly standardised, the "raw" MNO data that are relevant for statistics are produced by proprietary network equipment in proprietary unstandardised formats. Accordingly, the methodology for processing MNO data should be designed with a sufficient degree of flexibility to accommodate the unavoidable diversity[21] across different MNOs of multiple data dimensions (including quality, format, semantic, granularity in space and time, etc.).

One possible approach to deal with the given data heterogeneity while still pursuing maximal methodological harmonisation is to differentiate the tightness of the methodological specification. Consider again the end-to-end

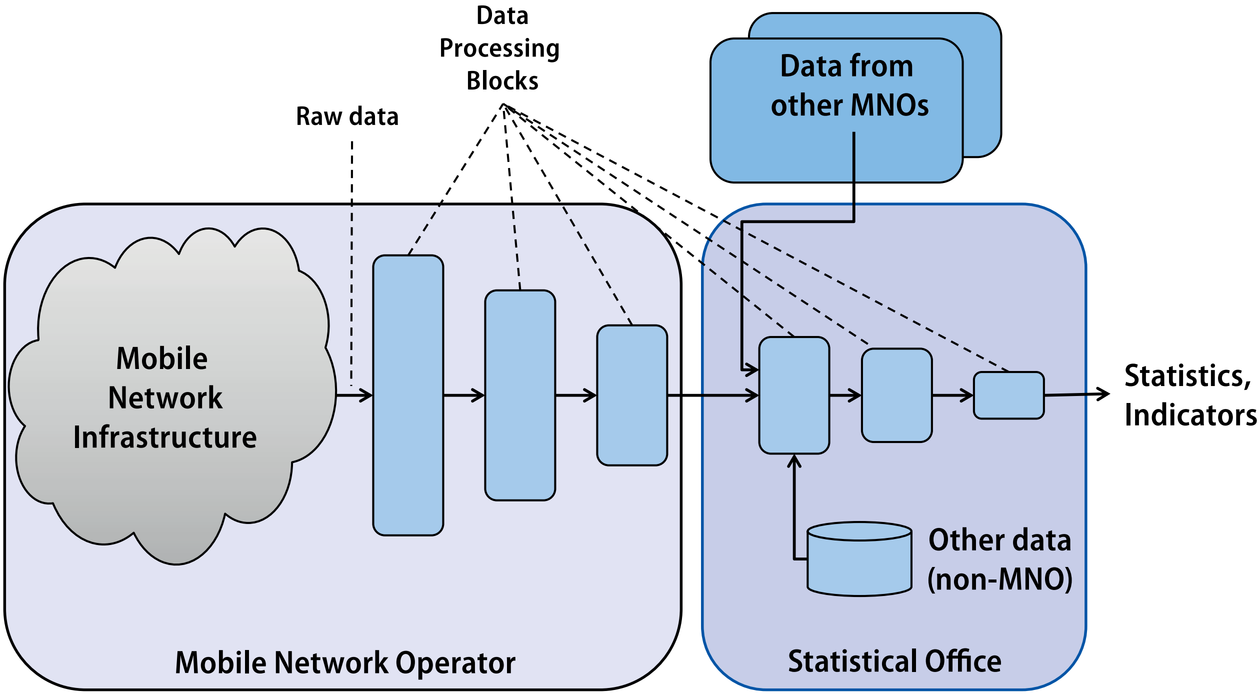
[19] In many ICT fields standardisation is a key driver towards improved competition and market development.

[20] Supporting multiple statistical products requires adding new branches that depart from one or the other module, depending on the overall design, and terminate to the final statistics or indicators (as leaves). Therefore, the linear chain grows into a tree-like structure. The "methodological pipeline" then identifies a single path in the overall "methodological tree" towards some statistical product. However, for the sake of simplicity and to avoid unnecessary burden in the presentation, we shall refer throughout the paper to a single generic (unspecified) statistical product and to the associated "methodological pipeline", with the understanding that all considerations made here are fully applicable also to the other statistical products and to the whole "methodological tree".

[21] Furthermore, as technical and business choices of a single MNO vary in time, a certain degree of temporal diversity should be expected on top of inter-MNO diversity.

FIGURE 2

Modular representation of the methodological pipeline transforming "raw" MNO data into final statistics from the perspective of a single statistics product. The height of each module represents the level of data granularity: lower height corresponds to lower level of granularity, i.e., higher degrees of aggregation.



methodological pipeline from "raw" MNO data to final statistics as a series of processing modules, as graphically sketched in Figure 2. Specifying a standard methodology requires, inter alia, (i) defining the modular architecture itself along with a high-level description of "*what*" each module is supposed to achieve; (ii) define interfaces between the modules in terms of data formats and semantics; and (iii) define a detailed low-level description of "*how*" each module operates, i.e., the exact method or algorithm. In relation to the latter point, the tightness of the definition may vary, from human-readable guidelines (loose description, subject to be interpretated and adapted to the specific situation in each MNO) to machine-executable code represented in some formal language (tight description, perfectly equal for all MNOs), with a whole range of intermediate grades in between these two extremes (e.g., semi-formal pseudo-code).

From Figure 2 it should be clear that the very first modules upstream the methodological pipeline, i.e., the ones interfacing with the unstandardised system component generating the "raw" data in the first place (e.g., monitoring probes) can be standardised only in the sense of providing a loose description, leaving room to the operators to adapt the actual implementation to the peculiar configuration of their mobile network infrastructure. This is because one cannot standardise tightly a module designed to interwork with a preceding module that is not standardised. But as we move downstream the methodological pipeline, one can increasingly tighten up the level of description and reach a point where all modules can be standardised in the sense of providing a reference open-source implementation. This *graduated transition* through increasingly tighter levels of standard description is key to reconcile the goal of producing maximally standardised statistics in output with the unstandardised nature of the raw data in input.

Another important requirement for the design of the ESS methodology is its *evolvability*, i.e., the degree by which the methodology is designed upfront with the perspective of being continuously improved and adapted to a changing environment. Changes occur across various dimensions. First, as mobile technologies, services and usage patterns change, so do the MNO data that result from the use of those technologies and services. Second, the needs by statistical users change over time, with emerging demands tomorrow to measure more phenomena and/or with better accuracy than what is considered sufficient today. Third, the capabilities and costs of hardware and software resources change, possibly enabling tomorrow the application of highly sophisticated methods that may be considered unfeasible today. Fourth, our own understanding of the data advances based on lessons learned and experience from past measurements, and we want such advancement to be reflected in methodological improvements. For all such reasons, the ESS methodology for MNO data should be seen as an **evolving standard** that, like many other ICT standards, progresses through successive releases.

Methodological development and standardisation are not one-off tasks, but rather continuous processes, and this is even more true when coping with data produced by one of the fastest evolving business sectors, namely mobile communications. It is of outmost importance that the ESS methodological is **designed upfront with the perspective of anticipating possible future directions and therefore facilitate evolution**.

8

The importance of taking a multi-MNO perspective

The envisioned ESS methodological standard should be designed upfront to produce statistics based on the *combination of information*[22] *sourced from multiple MNOs* (multi-MNO statistics) within and across different European countries.

From the perspective of statistics referred to a single country, relying on data from all the major MNOs active in that country (typically between two and four) brings the following benefits:

- Better representativeness of the total population, lower exposure of the final statistics to population coverage bias and improved stability. As MNOs tend to specialise and target different customer segments, certain population subgroups may be under- or over-represented in the customer base of one individual MNO. There are also geographical differences, as one MNO may have better or worse coverage than its competitors in some geographical area, and temporal fluctuations, as one MNO may suddenly attract or lose mobile subscribers from/to its competitors due to the launch of new tariffs or marketing campaign. By integrating information from all main MNOs, the differences across MNOs are mutually compensated, the total bias is reduced, the temporal stability and geographical coverage are intrinsically improved.
- Increased robustness of the total statistics to anomalies, glitches and interruptions of data availability caused by MNO-specific technical glitches, e.g., due to failures of data acquisition system or other problems with the mobile network infrastructure.
- Equal treatment of competing MNOs. Engaging in the process of statistical production with *all* the major MNOs competing in the same country prevents introducing asymmetries and differences of treatment (level playing field). Furthermore, if the statistics that are finally disseminated by the NSI are built from three or more MNOs of comparable size, the task of preventing inference of business-sensitive information referred to any individual MNOs from the combined aggregate figures is greatly simplified. This aspect, alongside the standard practices of business confidentiality protection that are already in place for official statistics, contributes to avoid any risk of interfering with competition dynamics between MNOs.

The capability of integrating information from different MNOs, within each country and across different countries, is key to improve the quality of cross-border statistics about international mobility. This is due to the characteristics of roaming data, i.e., MNO data produced by subscribers of country A visiting a foreign country B. The following considerations apply:

- It may be practically very difficult, if at all possible, to extrapolate the number of inbound visitors from country A to country B based solely on the roaming data collected by a single MNO operating in B. This is because the *actual* roaming shares, i.e., the actual fraction of visitors to country B that are "seen" by each MNO fluctuate and do not always correspond to the *nominal* shares negotiated between operators.

[22] We use the general term "information" with the purpose of leaving unspecified the actual modalities by which data integration could be implemented across MNOs and the level of granularity of the data to be combined, that could range from the linking individual records to fusion of aggregate data. Different statistical use-cases may require different combination modalities, and their specification is part of the methodological development process.

- Consider a mobile user from country A that is visiting country B. The home operator[23] from country A and the visited operator from country B hold complementary views of the visited locations in their respective countries. To simplify, we can say that the home operator knows that the outbound roamer is visiting country B but does not know exactly which geographical region thereof. Conversely, the visited operator knows that the inbound roamer originates from country A but not from which region thereof. For certain important statistics it is important to leverage spatial information about both countries (e.g., to measure touristic flows between two cities in different countries, or between border regions in neighbouring countries). In such cases, the information held by the two operators operating in both countries must be combined.

For all these reasons it is desirable to develop a methodological framework that can support the integration of information from multiple MNOs, within and across different countries.

[23] It should be noted that a mobile user subscribed to a telecom operator in country A may not necessarily live or be officially registered in that country, therefore the "home operator" of a mobile subscription does not necessarily correspond to the "home country" of the mobile subscriber as defined for administrative or statistical purposes. For instance, a trans-national worker that commutes regularly between two countries may have a mobile subscription in the home country, in the work country or in both countries. In all such cases, taking a multi-MNO perspective is a necessary (but not sufficient) condition to produce correct statistics in the presence of transnational mobility.

9

Fusion of MNO data with other data sources

Experience with past experimental and research projects has shown the necessity to combine MNO data with other kinds of non-MNO data to stabilise, correct or calibrate the final statistics. In fact, the *observed* population of mobile devices does not correspond exactly to the *target* population of humans. Moreover, the gap between the two populations is not static but rather dynamic since, as explained earlier in Section 3, the observed population fluctuates in time due to various behavioural aspects that are not linked to changes in the underlying human population, e.g., mobile users changing their mobile subscription to a different operator, adding a second subscription, or simply switching off their phones for a certain period.

The spurious fluctuations in the observed population should be mitigated in the final statistics, while the total figures obtained from MNO data should be related to the total target population. This is not at all a simple task, and various approaches are possible that are currently the object of ongoing research within the ESS.

One simplistic approach is to weight the total figures obtained from MNO data with market share estimates for individual MNOs. In this case the market shares serve as "external" data element. This approach has several limitations. First, the total market share does not capture the possible over- or under-representation of certain population groups in the customer base of each MNO. Second, long-term average shares may fail to deal with short-term fluctuations (e.g., a massive flow of subscriber to one MNO due to the launch of a particularly appealing mobile offer).

A considerably more sophisticated alternative is to anchor MNO data-based figures to reliable hot-spot or section-crossing data. To illustrate, consider the example where some other data source makes available the exact count of how many people have moved between two points of interest (e.g., the number of railway passengers from station A to station B based on train ticket sales), or have crossed a gate (e.g., the count of people entering a shopping mall) or have visited a place (e.g., a concert, again based on ticket sales) during a particular time interval, and assume that the same quantity is also measured from the set of MNO data. The ratio between the two counts can be used to calibrate the figures obtained from MNO data also for other flows, under certain assumptions. The viability of this approach depends in the first place on the actual availability of sufficiently accurate hot-spot data, and on the successful establishment of additional agreements with the respective data providers for making such data available for statistical production, with the necessary levels of continuity and quality.

An alternative approach would seek to use data that are already held by NSI, e.g., spatial census data or population registers, to "anchor" the figures obtained from MNO data to a stable reference. Adjustment weights may be geographically localised, i.e., differentiated based on the territory of residence. This approach relies on the accuracy of spatial census data as the "external" non-MNO data element. Furthermore, it does not solve the problem of coping with the sub-population of mobile users that are not represented in the census data, e.g., visitors or non-registered residents.

Another possible strategy is to rely on sample survey data *designed and collected for the purpose of calibrating statistics based on the secondary (re)use of MNO data*. This approach is probably the most natural for statistical authorities in many respects. On the methodological side, it gives to statistical authorities full control over the design of survey data—a task that is at the core of their mandate and for which they have well-established internal competencies—resulting in high levels of data quality even for moderate numbers of respondents (small sample size). In other words, a small random sample of "designed" survey data is leveraged to boost the quality of a much larger volume of non-random non-designed data (such as MNO data). On the side of data access, since survey data would be produced natively for statistical purposes and by statistical authorities, they would fall within the scope of existing statistical legislation and would not require the establishment of additional relationships with external data providers. From a more strategic point of view, the perspective of integrating NSI data and MNO data would lead to a more balanced relationship between the two organisations – with NSI acting not merely as consumers of MNO data but as co-providers of other data – which could facilitate the establishment of mutually beneficial partnerships.

In conclusion, fusing together large-scale MNO data, rich of spatiotemporal information and covering almost the entire population (assuming multi-MNO data combination), with small-scale survey data designed ad-hoc for this task (possibly including basic socio-demographic variables) appears to be a particularly promising strategy to deliver high-quality and information-rich statistics at acceptable costs. While further work is needed to define the detailed methods of data fusion, and to assess quantitatively the costs and benefits of this approach, the ESS methodological framework must be prepared to embrace this perspective.

10 Data access

The availability of a common and open methodology fulfilling the requirements outlined above is a necessary but not sufficient condition towards the regular production of official statistics based on MNO data. Recalling the analogy in Figure 1, other switches must be closed alongside the methodological one.

First and foremost, the processing of MNO data for official statistics must comply with the applicable data protection legislation. There is no doubt that processing MNO data qualifies as processing of Personal Identifiable Information (PII) and therefore must comply with GDPR in the first place. The goal of statistical authorities is to produce aggregate statistics referred to large population groups, therefore the whole transformation process, from personal "raw" data to non-personal aggregate statistics must be implemented in a way that ensures the privacy risks (actual or perceived) are minimised and anyway kept down to an acceptably low level. This is not only a legal requirement, but also an ethical imperative! Considering the **European scope of the envisioned methodology**, that is developed purposedly for application in all European countries and to produce European statistics in addition to national statistics, it appears natural to verify compliance to data protection rules directly at the European level, by consultation with the European Data Protection Supervisor (EDPS) and European Data Protection Board (EDPB). Such consultation may possibly lead to the recommendation of specific supplementary technical and organisational measures to safeguard the processing of MNO data for official statistics purposes. Some of these measures may involve the adoption of Privacy Enhancing Technologies (PET) to increase the level of protection during the processing stage[24]. Input privacy solutions seem to be particularly well suited to play a role in the context of MNO data processing for official statistics, particularly to secure the integration of data from different sources, i.e., between data from different MNOs and between MNO and non-MNO data[25]. If PET are used to protect the integration of (intermediate) data from different MNOs, they would at the same time concur to protect the confidentiality of the business sensitive information embedded in the data. This would further reassure the mobile network operators and the mobile subscribers alike. In other words, the adoption of PET would result in a stronger protection of data confidentiality both in terms of personal privacy and protection of business secrecy, and therefore increase acceptability by data holders and data subjects. Moving from general considerations to the specification of deployable PET solutions (*how* to protect) is not possible without having first established in detail the overall processing workflow (*what* to protect), and in this sense the definition of a standardised and fully open methodological framework represents an enabling condition for the definition, implementation and deployment of adequate protection measures based on PET.

Another important pre-condition for the regular production of statistics based on privately held data, such as MNO data, is the establishment of a *sustainable* model for data access,

[24] The interested reader is invited to refer to the recent Final Report of the UNECE Project on Input Privacy Preservation, https://tinyurl.com/4rmpj354, and to the UN Guide on Privacy-Enhancing Technologies for Official Statistics, https://unstats.un.org/bigdata/task-teams/privacy/guide/index.cshtml.

[25] See e.g. https://cyber.ee/resources/case-studies/mobile-phone-data-meets-sharemind-hi:-tourism-statistics-innovation-in-indonesia/ and https://cros-legacy.ec.europa.eu/content/eurostat-cybernetica-project_en for two examples of relevant experimental projects in this field.

tailored to the peculiar requirements of official statistics (among other public purposes) and respecting the legitimate interests of the data holders. The ESS recognises that relying solely on voluntary participation by the data providers has proven to be insufficient even for experimental purposes, let alone regular production, and has advocated the need to establish a clear set of rights and duties, obligations and limitations for data holders and statistical authorities (European Statistical System (2021)).

As far as the European legislation is concerned, the European Commission has recently started the process of revising Regulation 223/2009 on European Statistics[26] to "*tap the potential of new data sources*". The new revised regulation will expectedly introduce legal enablers for the reuse of privately held data, including MNO data, for the development and production of European statistics. In parallel, the proposal for a new ePrivacy Regulation[27] that should replace the current ePrivacy Directive is currently progressing through trilogue negotiations and will hopefully include similar enablers on the side of telecom legislation.

A new legislative framework that is supportive of MNO data reuse for official statistics is a central *conditio sine qua non* regular statistical production can take place. According to the final recommendations of the *Expert Group on facilitating the use of new data sources for official statistics*[28] (B2G4S for short), a **revised legislative framework** would be instrumental to foster the definition and implementation of fair partnership models between statistical authorities and private data holders, with a balanced distribution of costs and responsibilities. Indeed, the processing of MNO data with guaranteed minimum levels of quality and continuity involves non-negligible **costs** along multiple dimensions (IT, organisational, business processes, human resources) for both, NSI and MNO sides. How these costs shall be quantified, distributed and compensated remains an open issue. Elaborating on these aspects falls outside the scope of the present paper (and of the mandate of the TF-MNO). However, it should be self-evident that the detailed specification of a standard methodological framework should at least facilitate the quantification of the involved execution costs.

Another open issue, intimately related to all the aspects presented above, concerns the **operational modalities** of data access. Considering the data processing workflow as a chain of processing stages (modules), the question amounts to determine how the *execution* of the processing pipeline is divided between the MNO and the NSI premises. Based on the knowledge gained through past experiences, transferring large amounts of granular data (i.e., individual records and detailed cell configuration data) to the NSI does not appear to be the most convenient option, if at all a feasible one. Besides the obvious duplication of IT resources, such an approach may unnecessarily increase the risks (actual or perceived) to data confidentiality and business secrecy. To protect more effectively (i) the personal data of mobile users and (ii) the business sensitive information of the MNO, it would be preferable to execute at least some part of the processing pipeline locally, close to where the data are generated, i.e., at the premises of the MNOs, and then export to the NSI intermediate data at some level of aggregation. In general, higher levels of aggregation correspond to lower levels of sensitiveness and risk. Accordingly, exporting aggregate (pre-processed) data outside the MNO premises is less risky, more resource-efficient and more socially acceptable than moving very granular "raw" data. This approach complies with the GDPR principle of data minimisation and is also fully in line with the paradigm of "*pushing computation out*" (towards the data source) instead of "*pulling data in*" (into the NSI) that is central to the Trusted Smart Statistics concept (Ricciato *et al.* (2019)). Based on these considerations, it may be expected that in most cases a considerable part of the processing pipeline will be executed at the MNO premises.

[26] See this page for up-to-date information regarding the legislative process: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13332-European-Statistical-System-making-it-fit-for-the-future_en

[27] https://www.europarl.europa.eu/legislative-train/theme-connected-digital-single-market/file-jd-e-privacy-reform

[28] This was a group of experts with diverse professional backgrounds that was established by the European Commission in 2021 and tasked with providing a set of specific recommendations to enable the reuse of privately held data for official statistics. The final report "*Empowering society by reusing privately-held data for official statistics — A European approach*" was published on 28 June 2022 and is available from https://ec.europa.eu/eurostat/en/web/products-statistical-reports/-/ks-ft-22-004

11 The way forward

In the present paper we have outlined the TF-MNO vision of how the (future) production of official statistics based on MNO data should be organised in Europe, with particular focus on the methodological aspects. The ESS is already taking concrete steps towards implementing this view, capitalising on the knowledge and experience acquired through past projects.

In line with the view elaborated by the TF-MNO and presented in this paper, Eurostat has already launched two complementary projects in the field of methodological development. The first project, started in January 2023 following an open call for tenders[29], gathers a **mixed team of industry experts and statistical experts** from multiple organisations. The project team is given the task to **co-develop** under the supervision of Eurostat the initial version of a fully open methodological pipeline for transforming MNO data into official statistics. The project consortium includes *mutually competing* business companies working alongside NSIs – two MDAPs and five MNOs from four different countries contribute to the project work. The project is committed to release an open-source reference implementation of the methodological pipeline and to demonstrate its feasibility across multiple MNOs networks – hence the short name "Multi-MNO" for this project[30].

The second project is a **research grant** directed at NSIs and is expected to start in the final quarter of 2023. The project scope is focused on the integration of MNO and non-MNO data and its future results are expected to provide guidance as to (i) which sources of non-MNO data are most appealing to complement MNO data for regular statistical production and (ii) which methods to use for the integration of these data sources with MNO data. These two projects lie at the core of methodological development on MNO data in the ESS[31].

[29] https://etendering.ted.europa.eu/cft/cft-display.html?cftId=10149

[30] https://cros-legacy.ec.europa.eu/content/multi-mno-project_en

[31] For more information on future activities and projects funded by Eurostat in the field of MNO refer to https://cros-legacy.ec.europa.eu/content/mobile-network-operator-data-official-statistics-mnodata4os_en

List of References

Janecek, A., Valerio, D, Hummel, K. A., and Hlavacs, H. (2015). The Cellular Network as a Sensor: From Mobile Phone Data to Real-Time Road Traffic Monitoring. IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 5, pp. 2551-2572. doi: 10.1109/TITS.2015.2413215. https://ieeexplore.ieee.org/document/7079458

Ricciato, F., P. Widhalm, M. Craglia, and F. Pantisano (2015). Estimating Population Density Distribution from Network-based Mobile Phone Data. JRC Technical Report https://doi.org/10.2788/162414.

European Statistical System Committee (ESSC). European Statistics Code of Practice – revised edition (2017). https://ec.europa.eu/eurostat/web/quality/european-quality-standards/european-statistics-code-of-practice

European Statistical System Committee (ESSC). Quality Assurance Framework – version 2.0 (2019). https://ec.europa.eu/eurostat/web/quality/european-quality-standards/quality-assurance-framework

Ricciato, F., Wirthmann, A., Giannakouris, K., Reis, F. and Skaliotis, M. (2019). Trusted smart statistics: Motivations and principles. Statistical Journal of the IAOS 35, 589–603. DOI 10.3233/SJI-190584. https://cros-legacy.ec.europa.eu/system/files/sji190584.pdf

European Statistical System (2021). Position paper on the future Data Act proposal. https://ec.europa.eu/eurostat/documents/13019146/13405116/main+ESS+position+paper+on+future+Data+Act+proposal.pdf

Tennekes, M., Gootzen, Y.A.P.M. (2022) Bayesian location estimation of mobile devices using a signal strength model. Journal of Spatial Information Science, 25, 29-66. doi:10.5311/JOSIS.2022.25.166. https://josis.org/index.php/josis/article/view/166/169

Ricciato, F. and Coluccia, A. (2023). On the estimation of spatial density from mobile network operator data. IEEE Transactions on Mobile Computing, 22(6), 3541-3557. 10.1109/TMC.2021.3134561. https://ieeexplore.ieee.org/document/9647984

List of Acronyms

CDR	Call Detail Records
DDR	Data Detail Records
GPS	Global Positioning System
ICT	Information and Communication Technologies
IOT	Internet of Things
IT	Information Technologies
ESS	European Statistical System
MDAP	Mobile Data Analytic Provider
MNO	Mobile Network Operator
MPD	Mobile Phone Data
M2M	Machine-to-Machine communication
NGO	Non-Governmental Organisation
NSI	National Statistical Institute
PII	Personal Identifiable Information
PET	Privacy Enhancing Technologies
TF-MNO	ESS Task Force on MNO data for Official Statistics
XDR	eXtended Detail Records

Contributors

This paper is the result of the work of the ESS Task Force on MNO data for Official Statistics (TF-MNO for short) and represents the view of all its members. The official members of the TF-MNO as of 20th June 2023 are listed[32] below along with the additional ESS experts that have contributed to shape the content of this paper.

Fabio Ricciato (Chair) – Eurostat

Laust Hvas Mortensen - Danmarks Statistik, Denmark

Natalie Rosenski, Sandra Hadam - Statistisches Bundesamt, Germany

Patrick Gregg - Central Statistics Office, Ireland

David Salgado, Sandra Barragán - Instituto Nacional de Estadística, Spain

Marie-Pierre Joubert - **De Bellefon** - Institut National de la Statistique et des Étude Économiques, France

Mateković Franjo - Croatian Bureau of Statistics, Croatia

Tiziana Tuoto, Roberta Radini - Istituto Nazionale di Statistica, Italy

Vilma Nekrašaitė-Liegė - Statistics Lithuania, Lithuania

Marc Pauly - Institut National de la Statistique et des Etudes Economiques, Luxembourg

May Offermans, Martijn Tennekes - Centraal Bureau voor de Statistiek, Netherlands

Johannes Gussenbauer, Alexander Kowarik - Statistik Austria, Austria

Marek Cierpiał-Wolan - Statistics Poland, Poland

Pedro Sousa - Instituto Nacional de Estatística, Portugal

Bogdan Oancea, Marian Necula - National Institute of Statistics, Romania

Igor Kuzma - Statistical Office of the Republic of Slovenia, Slovenia

Pasi Piela - Statistics Finland, Finland

Pieter Vlag - Statistics Sweden, Sweden

Johan Fosen - Statistics Norway, Norway

Thanks also to Peter Struijs and Monika Wozowczyk from Eurostat for helping to improve earlier versions of this manuscript.

[32] The official name of each institution as well as the order of listing follows the official list available from the Eurostat website https://ec.europa.eu/eurostat/web/european-statistical-system.

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

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

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696,
- via the following form: european-union.europa.eu/contact-eu/write-us_en.

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

EU open data

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

Print: KS-FT-23-001-EN-C
PDF: KS-FT-23-001-EN-N

Reusing Mobile Network Operator data for Official Statistics: the case for a common methodological framework for the European Statistical System

This position paper was prepared by the ***ESS Task Force on the use of Mobile Network Operator (MNO) data for Official Statistics***. The Task Force was established in 2021 to address the methodological aspects involved in the (re)use of MNO data for official statistics. The paper provides three main contributions. First, starting from a reasoned analysis of the current landscape, it motivates the need for establishing a common ESS methodological standard defined at the European level for the transformation of MNO data into official statistics. Second, it establishes the key high-level requirements that such common ESS standard should fulfill: openness, transparency, modularity, evolvability, multi-MNO orientation (i.e., combine information from multiple MNOs) and supporting the fusion of MNO data with non-MNO data. Third, it provides a bird's-eye view of the ongoing activities in the ESS to advance towards the definition of a common ESS methodological standard for MNO data in official statistics.

For more information

https://ec.europa.eu/eurostat/

Publications Office
of the European Union

Print: ISBN 978-92-68-05648-6
PDF: ISBN 978-92-68-05647-9