ENHANCING OFFICIAL STATISTICS WITH NEW DATA SOURCES – METHODOLOGICAL DEVELOPMENTS FOR INTEGRATING MOBILE NETWORK OPERATOR (MNO) DATA WITH NON-MNO DATA

Gloria Deetjen[1], Maurice Brandt[1],

Marie-Pierre Joubert[2],

Tiziana Tuoto, Maria Grazia Calza, Maria Francesca D'Ambrogio, Immacolata Fera, Tamara Zangla [3]

[1]Statistisches Bundesamt (Destatis), Germany

[2] Institut national de la statistique et des études économiques (Insee), France

[3]Istituto Nazionale di Statistica (Istat), Italy

Abstract

The use of new data sources in official statistics offers not only the opportunity for new analysis but also to improve and to complement regular statistics. Mobile Network Operator (MNO) data are amongst the most favourable types of new data sources for this purpose because of their high spatial resolution as well as their timely availability. The combination of the so-called MNO data with further traditional or new data sources is a great way to enrich existing statistics as various experimental studies have shown. However, the integration of MNO data into regular official statistical production requires profound solutions and innovative methodologies because of potential errors and quality issues which can mostly arise from either the data itself (e.g. data configuration) or from the way the data is used (e.g. direct or auxiliary). Moreover, it is the role and responsibility of statistical offices to maintain high quality outputs in a world of new digital data.

The ESSnet MNO-MINDS project "Mobile Network Operator – Methods for Integrating New Data Sources" aims to address these needs by proposing a reference frame for methods regarding the combination of MNO and non-MNO data for official statistical production. These new methodologies alongside with guidelines are crucial for a successful linkage of MNO data with non-MNO data in the European Statistical System (ESS). The project is organized via four work packages (WP), each focusing on a specific area: WP1 is responsible for coordination, project management and dissemination, WP2 conducts a landscape analysis of non-MNO sources to be combined with MNO data, WP3 focuses on methodologies and open source tools for integrating MNO and non-MNO data sources, and WP4 offers proof of concept of an ad-hoc survey to improve MNO data. This paper provides a general overview on the project and presents first intermediate results of WP2's landscape analysis with a focus on quality.

Keywords: MNO data, data integration, new data sources, ESSnet MNO-MINDS

1. Introduction

In the digital age, with accelerating amounts of fast and big data, statistical offices represent a reliable source for data and statistics. At the same time, statistical offices explore and incorporate new data sources themselves, e.g. to reduce response burden, to improve spatial resolution and timeliness, and to address needs for new statistics which derive from developments of the digital age itself. However, the high-quality standards to which statistical offices are committed require profound methods and quality frameworks before implementing new data sources into statistical production. In experimental statistics, “Mobile Network Operator data” (short: MNO data) have proven to increase for instance spatial resolution and timeliness. However, the full potential of MNO data appears especially when combining MNO data with further data sources. Here, traditional data sources like census or administrative data usually provide opportunities for quality improvement whereas further new data sources enable widening the scope of application opportunities (although the boundary between the contributions of these two data categories is not always clearly identifiable). The integration of MNO data with non-MNO data sources therefore requires an identification of the most suitable non-MNO data sources to be combined with, a reference frame of methods, and quality improvements of MNO data itself. This project aims to address these needs by the efforts in all work packages which are introduced in the following chapters.

1.1 Organisation and Scope of the Project

During the previous ESSnet projects Big Data I and Big Data II, various new data sources were studied and as a result, MNO data was identified as one of the very promising data sources for official statistics. The term “MNO data” mostly refers to signaling data which is produced when mobile phones connect to cell towers, and which contain time stamps and spatial information. National Statistical Institutes (NSIs) usually receive access to aggregated MNO data through e.g. partnerships with MNO’s. But as described above, further research is necessary to get a few steps closer to the implementation of MNO data and new data sources in official statistics.

In MNO-MINDS, the project consortium consists of national statistical institutes (NSIs) from ten countries: Italy (ISTAT) as the coordinator, Austria (STAT) as WP4 leader, Germany (DESTATIS), Spain (INE), France (INSEE) as WP2 leader, the Netherlands (CBS), Norway (SSB) as WP3 leader, Romania (INS), Sweden (SCB), and Portugal (INE-PT). The project started in November 2023 and lasts for a duration of two years.

1.2 Communication and dissemination activities

To incorporate feedback and input from ESS members that are not involved in this project as well as from further relevant stakeholders, various communication and dissemination activities have been and will be conducted: the project page on the CROS portal[1] contains detailed information about the project and the contents of the individual work packages as well as upcoming events. A dedicated page on Istat official website has been designed to support the project during its whole lifecycle. Some consortium partners have also given visibility to the project on their websites.

To maximise reach and impact, the project activities landed on social media, both LinkedIn[2] @MNO-MINDS ESSnet Project and X[3] @TssMethToo_Pj, whose continuous updates contribute to give high visibility among the general public, the professional network community, other specific target audiences, and relevant multipliers.

Communication and dissemination activities help to receive food for thought from further stakeholders (e.g. other non–participating countries in this project, research organisations, academia, MNOs...) that enrich the work especially in the development phase of the project, and eventually contribute to quality improvements. An important upcoming milestone for such valuable input and discussion is to be highlighted here: The SPRINT[4] event taking place on 10th and 11th June 2024 in Vienna which will be an opportunity to identify and prioritize non-MNO sources to be integrated with MNO data; learn about the main principles of our methodology frame; share feedback and suggestions; discuss objectives and design of an ad-hoc survey to integrate and improve MNO-data statistics.

Many other dissemination activities are envisaged during the lifecycle of the project and all work packages will be involved in disseminating intermediate results and achievements for consultancy at European and international conferences. A project final conference will be held in Rome where the participation from all the ESS, from the academia, and from the MNOs will be promoted.

Finally, another relevant source to collect feedback is the exchange within the Task Force on Mobile Network Operator Data for Official Statistics (TF-MNO), which consists of field experts led by Eurostat and supports the sharing of experiences from national pilot projects involving

---

[1] https://cros.ec.europa.eu/mno-minds

[2] https://www.linkedin.com/in/mno-minds-essnet-project-5a09b12a0/

[3] https://twitter.com/TssMethToo_Pj

[4] Please find more information and how to register on CROS or contact tss-meth-too@istat.it

the use of MNO data for experimental statistics and networking of relevant experts across ESS members. Indeed, in a recent position paper (ESS Task Force on MNO data for Official Statistics, 2023) the Task Force depicts and motivates the case for the integration of MNO data with other non-MNO data sources, reinforcing the link between this project and the other initiatives of the European Commission for the use of MNO data for Official Statistics[5]. The exchange ensures that interplay with further activities on MNO data in the ESS are considered and that any important findings can be incorporated in this ESSnet project as early as possible.

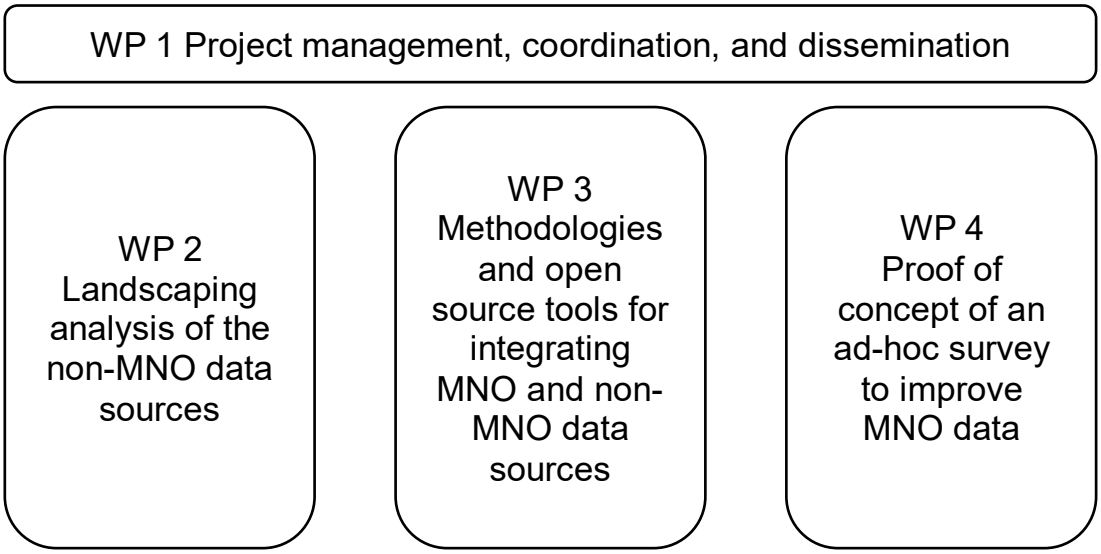
Figure 1: ESSnet MNO-MINDS project logo

1.3 Project structure and overview of WPs

The work is organized in four work packages: ISTAT, as the leader of WP1 is responsible for project management, for the administrative and financial aspects of the project as well as for communication and dissemination activities. As regards the main communication activities of WP1 relevant for quality in official statistics, they have been introduced in the previous subchapter. High quality of the outputs is guaranteed through an effective and smooth coordination. In fact, the project coordinator ensures communication and knowledge flows between work packages as well as communication with Eurostat, monitors project progress with respect to work-plan and ensures high quality of the deliverables by implementing a quality control plan. Moreover, it fosters synergies with other initiatives funded by Eurostat as the Multi-MNO project aiming to develop a complete open end-to-end processing pipeline to produce future official statistics using data from multiple MNOs.

[5] https://cros.ec.europa.eu/MNOdata4OS

Figure 2: Work Packages in the ESSnet MNO-MINDS project



In WP2, a landscape analysis is conducted to identify and to assess non-MNO data sources to be combined with. First, potential data sources are collected based on the experience of the NSIs. In addition, further data sources are added which have been collectively identified as having great potential, but to which no country currently has access. The first round of analysis includes thirteen data sources. For evaluating the data sources, an assessment matrix is developed and filled in for each data source. The first aspect to be assessed is relevance, which is mainly judged by three aspects:

1. Improving population coverage
2. Improving spatial and temporal precision of analysis
3. Broadening the scope of issues covered by official statistics

Further, the matrix includes several questions/aspects regarding data type, data access, metadata, and quality aspects for official statistics. The main results from this first round of analysis are presented in chapter two.

The integration of new data sources with MNO data requires new methodologies and solutions which address potential errors and biases. WP3 suggests a reference frame of such methods and considers different scenarios. For example, whether MNO data is used as a target statistic or as an auxiliary variable. It further considers the data configuration as well as the usage of the data. Eventually, three approaches are followed: Randomisation, quasi-randomisation, and super-population modelling. Methodologies are presented in detail in the conference paper dedicated to work package three.

Whereas work packages two and three focus on the integration of non-MNO data sources, work package four concentrates on improving the quality of MNO data usage in official statistics by developing a survey to reduce bias and to fill potential knowledge gaps. Therefore, three main issues are targeted by the survey: First, device-user mismatch as the user and the contract holder might not be the same person. Then, multiple devices/sim cards as one SIM card does not simply represent one person because some persons may carry regularly more than one phone, e.g. one for private and one for business purposes. Last, there is lack of information on the actual user behaviour, e.g. if phones are always switched on and carried by the person. Before the final survey is suggested, a friendly-user test will be conducted. Intermediate results of work package four are presented in detail in the dedicated conference paper.

The expected outcomes of all work packages together will address the heterogenous application interests and circumstances in the European Statistical System. Further, the results will provide guidance on implementation and solutions regarding the integration of MNO and non-MNO data.

2 Work Package 2

2.1 The process: selecting, describing and scoring the sources

Work Package 2 aims at proposing an assessment matrix, which allows to analyse the relevance of non-MNO data for being combined with MNO data. The WP's final deliverable will explain the reasons that led to the choices of analysis criteria and present these choices. It will also provide a list of non-MNO sources and potential target statistics, along with a systematic identification of their main pros and cons, costs and gains, obtained thanks to the assessment matrix.

All along the two years of the project, the list of the most promising sources and the analysis of their accuracy for being combined with MNO data will be refined thanks to the various feedback gained for instance in the Sprint, or thanks to the comments by the MNO Task Force led by Eurostat. The actual or prospective availability of non-MNO data sources across different countries will also be assessed, potentially through a dedicated survey.

2.2 Main ways in which combining MNO data and non-MNO data can improve the quality of official statistics

Identifying the main use cases of this combination is an important first step to score the accuracy of data to be integrated with MNO data.

The first enhancement brought by this combination is to improve the population's coverage, whether in terms of representativeness or temporal and spatial precision. One important point to consider in the comparative analysis of different sources is the potential difficulty due to some irreconcilable concepts: for instance, different definitions of fundamental units or different time periods.

Combining different data sources also allows to provide more precise analyses. For instance, MNO data alone often do not include socio-demographic information; combining them with fiscal data about people's earnings (even aggregated), will allow to improve the study of socio-spatial segregation. In the same spirit, MNO origin-destination matrix often lacks information about the transportation mode, or it is of very bad quality. This angle of analysis will focus on themes that already exist, where the combination of sources makes it possible to overcome a problem of data quality or data completeness. When examining the interest of combining different sources, limits due to the respect for the confidentiality of personal data and the risk of re-identification will naturally be considered in foreseen use-cases.

Lastly, combining MNO and non-MNO sources allows to cover new topics of interest for official statistics. As explained in the introduction, given the major changes taking place in our environment, this is an essential point for Official Statistics to keep providing up-to-date and relevant information.

2.3 Which criteria to analyze the relevance of non-MNO data?

Some non-MNO data are produced by official statistics, such as survey-based Census. In this case the whole data production process is designed from the outset to meet the needs of NSIs, whether in terms of quality or variables of interest. Others are traditionally used by NSIs although they are originally gathered for administrative purposes. These data need some specific treatments to meet the NSI's quality requirements, yet they are in general structured in a way that allows classical statistical treatments and the data providers have also an interest in offering the most exhaustive view of their population of interest. The last non-MNO data that this report considers are initially produced for objectives far removed from those of the official statistical service. The producer can be either from the private or the public sector and these data are often not structured on a classical way, some lack documentation about the interest variables for NSIs or about the quality issues encountered in the data collection process.

To analyse the potential of these sources, WP2's members have adapted the 'Big Data classification matrix' which was produced by ESSnet BigData II, to the specific question of the combination of MNO and non-MNO sources. Moreover, one important aspect of ESSnet's MNO-MINDS is that it aims at producing official statistics, and not only experimental statistics.

The short list of the most promising data sources will therefore be established considering the availability of sources in all NSIs and their respect for the European statistics code of practice.

2.4 First findings

WP2's members began their focus on thirteen specific sources. Six of them are traditionally used by NSIs: mostly survey and register based data, about socio-demographic themes (census, administrative register, ...), mobility (national travel survey) or tourism (survey about non-resident surveys). Seven data sources are relatively new and often provided by private operators: satellite data, traffic signaling, electronic invoices, social media, credit card transaction data and Google Popular time.

Unsurprisingly, traditional data sources comply almost exactly with official statistics quality standards, whereas new data sources' compliance is both more difficult to assess and potentially not guaranteed. One key element to underline is the issue of access to detailed methodological information on how the data is constructed. Even more when access to individual data is not possible, due for instance to some confidentiality issues, and NSIs therefore only have access to aggregated data. Some private companies are reluctant to share these methodological information, which they consider business sensitive. Others use these data for purposes so far removed from official statistics that they scarcely document any information useful for complying with quality criteria (metadata, missing data, etc.). This raises the importance of building a trusting partnership between the data producer and the NSI. Recent developments in European legislation may help in this area. The regularity of statistical production can also be affected by the commercial uncertainties of a private company.

As regards the relevance for identified use cases: it is hard to envisage the existence of a 100% exhaustive database covering the entire population. Even administrative directories have their limits and most of NSI's usual databases cover only residential population and usual mobilities (such as commuting). The official statistical system lacks information about day-time population and mobility behaviours at a fine geographical and temporal scale. Yet on the contrary new data sources such as mobile phone data, traffic loop sensors or transaction data are not initially collected for statistical purposes and their representativeness of the whole population is often biased due to the data provider's market shares or some technical specificities (such as people switching off their phones). The improvement of population coverage is a typical example of how combining different data can compensate for their individual shortcomings.

3 Conclusion & Outlook

All in all, it can be concluded that there are plenty of aspects to be considered when combining MNO and non-MNO data, whether this refers to the identification of most relevant non-MNO sources to be combined, the various application scenarios which are considered in the methodological developments, or the ad-hoc survey. In this project, the first intermediate results already show a clear direction, and the remaining one and a half years will concretise and test these first developments.

Acknowledgment

This work was co-funded by the European Commission Project “MNO-MINDS” - 101132744 — 2022-IT-TSS-METH-TOO.

References

ESSNet Big Data 2 (2020). Workpackage K Methodology and Quality – Deliverable K7: Typification matrix for big data projects.

ESS Task Force on MNO data for Official Statistics (2023). Reusing Mobile Network Operator data for Official Statistics: the case for a common methodological framework for the European Statistical System 2023 edition, Eurostat Statistical Report, https://ec.europa.eu/eurostat/documents/7870049/17468840/KS-FT-23-001-EN-N.pdf/88ed0175-a8d4-d1c9-bd97-50073bb9d978?version=1.0&t=16944238901833bb9d978?version=1.0&t=1694423890183