Towards the usage of Mobile Network Operators' data for European official statistics production:

improving quality through standardisation

Erika Cerasti[1], Tiziana Tuoto[1], Roberta Radini[1], Giorgia Simeoni[1], Gabriele Ascari[1], Cristina Faricelli[1], Paolo Mattera[1], Luca Valentino[1], Florabela Carausu[2] Matthias Offermans[3], Edwin de Jonge[3], Ricardo Herranz[4], Miguel Picornell[4], Cristina Escribano[4], Margus Tiru[5], Villem Tonnison[5],

[1]Istat - National Institute of Statistics, Italy

[2]Gopa Worldwide Consultants, Germany

[3]CBS - National Institute of Statistics, Netherlands

[4]Nommon Solutions and Technologies – Spain

[5]Positium – Estonia

Abstract

In recent years, the widespread adoption of personal mobile devices has opened new opportunities to collect rich spatio-temporal data about human activity. In this context, governments are embracing data-driven approaches as a powerful tool to inform policy-making. As an example, during the COVID-19 pandemic different European governments used data obtained from Mobile Network Operators (MNOs) to monitor the mobility and presence of people across the territory. These data proved to be useful for a variety of purposes, such as epidemic modelling and mobility planning, which fostered the interest of National Statistics Institutes (NSIs) in using MNO data for the production of official statistics. To leverage the proved richness of MNO data while acknowledging that the processing of these by private companies, through an undisclosed methodology, may not reach the quality and reliability standards requested by official statistics, the European Statistical System (ESS) is investing in creating the conditions to integrate MNO data in the production of statistics featuring increased timeliness and information content. In this paper, we present the work done in the context of the Multi-MNO project funded by Eurostat with the aim of developing a standardised pipeline for MNO data processing. We define the high-level principles followed by the proposed pipeline — which has been designed to be modular and general enough to fit the needs of different countries, multiple MNOs, and several statistical purposes — and describe the methodological steps of the proposed pipeline, with particular focus on possible quality assessment measures to be implemented throughout the different processing phases.

Keywords: MNO, Standardised process, mobile data processing, European Statistical System, official statistics.

1. Introduction

The utilization of data collected by mobile network operators (MNOs) holds significant promise for producing official statistics across various policy domains within the European Statistical System (ESS). These data can provide valuable insights about population presence and mobility patterns, crucial for areas such as spatial planning, transport, health, environment,

economy, and tourism. Recognizing this potential, in recent years several National Statistical Institutes (NSIs) started to explore the usage of MNO data for innovative experimental statistics. However, the processing of MNO data for production of official statistical indicators need to satisfy requirements and principles of official statistical production such as quality, transparency, privacy and scientific rigour. The development of standardised methods serving as reference for NSIs and MNOs is essential to this scope.

Standardization offers numerous benefits: it ensures consistency and comparability of output, needed for data integration and data exchange at national and international levels; it promotes transparency, privacy and accountability, helps policy formulation and evaluation and foster new business models. On top of that, standardization is crucial for the improvement of the accuracy and reliability of statistical products based on MNO data. Uniform methodologies, guidelines and quality standards for data collection, processing, and dissemination will reduce errors, biases, and inconsistencies, enhancing the credibility and trustworthiness of official statistics based on MNO data.

In response to these needs, the Multi-MNO project aims to develop a reference methodological framework combined with a processing pipeline, as a proposal for the ESS standard in the future production of official statistics based on MNO data. The project includes the implementation of the pipeline in an open-source software, its application to a set of use cases and the testing on real data from multiple MNOs across EU countries.

Use cases in the project covers different domains, like tourism, population, mobility and environmental risk.

The Multi-MNO project is funded by Eurostat and it is carried out by a consortium led by GOPA (Germany), in collaboration with industry partners NOMMON (Spain) and POSITIUM (Estonia), and the National Statistical Institutes ISTAT (Italy) and CBS (Netherlands). The consortium includes five MNOs from different countries: Orange Spain, Vodafone Spain, A1 Slovenia, Post Luxembourg and Vodafone Italy [Ref 2]. The multi MNO project is placed within a set of initiatives founded by Eurostat on the exploration of MNO data potential in official statistics [Ref 1, Ref3].

2. Design principles and reference scenario

The methodological framework underlying the processing pipeline development follows high-level design principles decided in agreement with Eurostat.

1.1. Design Principles

The following principles ensure a robust, flexible, and privacy-compliant data processing pipeline capable of adapting to various analytical needs and maintaining high standards of data integrity and utility.

Multiscale longitudinal analysis: data processing is organized into three timescales: short-term (one day), mid-term (one month or quarter), and long-term (one year). Granular event data (nano-data) is processed daily to produce summaries, which are then aggregated monthly or quarterly for mid-term analysis. Long-term data labels (e.g., place of residence) are derived from sequences of mid-term summaries. This approach supports GDPR compliance by minimizing storage duration for raw data and retaining aggregated data, which has a lower risk of re-identification, for longer periods.

Input data accessibility from any point of the pipeline: any data element can be read by any functional module.

Bottom-up one-way processing: data is processed from lower to higher timescales, with each timescale's data derived solely from the preceding timescale. This prevents bias from rewriting real-time data based on long-term summaries. Higher timescale data can be used to create new intermediate data, adhering to the principle of input data accessibility.

Separate integration along different dimensions: the workflow involves operations along time, space, and units (devices). Integration functions are separated into distinct modules to maintain modularity and clarity. When integration across multiple dimensions is necessary, it should be decomposed into sequential modules, each handling one dimension.

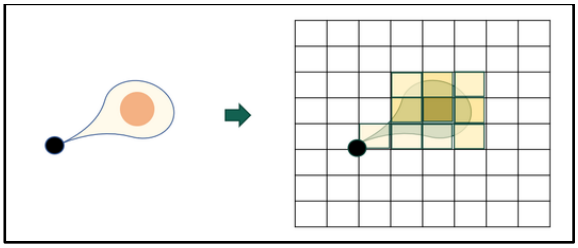
Balancing flexibility and parsimony: the pipeline supports multiple parallel methods within sub-modules, allowing for flexibility without a one-size-fits-all approach. This flexibility is balanced by limiting the number of variants to avoid complexity, grouping similar use cases to reuse methods effectively.

Soft classification and rigorous uncertainty assessment: decisions based on observed data should include an 'unknown' option when confidence is insufficient, avoiding misleading hard decisions. This approach ensures accurate representation of uncertainty in both functional blocks and data objects.

1.2. Standardized spatial grid

The data spatial representation is a crucial element for a standardization process. Hence, among the fundamental principles, the adoption of a common reference grid is included. In the project the INSPIRE geographical grid system is used throughout the pipeline, enabling data from different Mobile Network Operators (MNOs) to be combined and compared consistently over time. Intermediate data will adhere to this grid, while final statistics can be projected onto other grids (e.g., administrative units) as needed. This transformation process is detailed in the project documentation [Ref 2].

Figure 1: From cell coverage area to cell footprint on the spatial reference grid



1.3. Reference scenario

The pipeline high-level design has been conceived according to a hypothetical *reference scenario*, while the actual implementation refers to a *demonstration scenario* shaped by current privacy laws and data access agreements with participating MNOs.

The *reference scenario* assumes that conditions for accessing MNO data will favor the re-use of data by statistical authorities under a sustainable partnership model, ensuring strong data confidentiality to protect individual privacy and business-sensitive information. Microdata will remain within the protected MNO environment and be processed locally. Aggregate data will be accessed by NSIs without additional protection measures, under the assumption of a legal basis allowing the NSI to combine data from all major MNOs and apply protection measures before publication

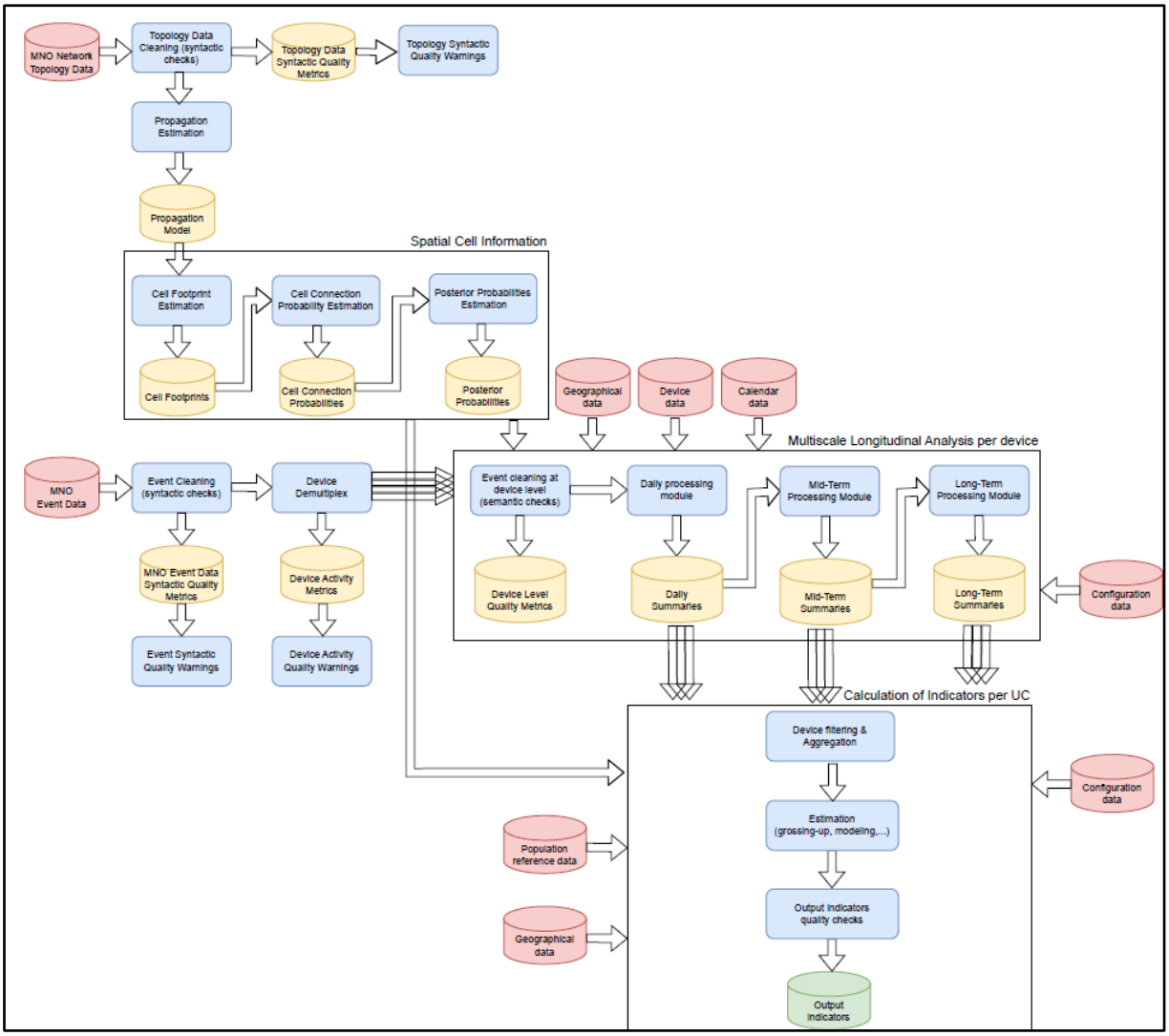
On the other hand, in the *demonstrator scenario*, aggregate data may need to undergo some statistical disclosure control (SDC) measures, such as k-anonymity, before NSI access. The necessity and specifics of these measures will be discussed with participating MNOs. Both scenarios involve combining data at the aggregate level, with micro-level integration only considered for inbound roaming statistics. The pipeline is designed to be modular, adaptable to both scenarios, and tested within the consortium, which includes Orange Spain and Vodafone Spain, ensuring flexibility. Overall, while the *reference scenario* allows more direct access to aggregate data, the *demonstrator scenario* requires

additional SDC measures, but both ensure data combination occurs at the aggregate level, maintaining privacy and business confidentiality.

3. Data processing flow – Pipeline

All data objects consist of various sub-objects with different formats and structures, accessible at any point in the pipeline for processing. MNO data enter at the pipeline's start and undergo processing through different stages. The framework is modular and flexibel, capable of incorporating further metods and more advanced techniques.

Figure 2: High- level representation of the pipeline



Arrows represent data flows, including both input data and data flows generated by the functional blocks of the pipeline. *Cylinders* represent (sets of) data objects, used to store information, with the color corresponding to a type of data objects (red for input data, yellow

for data generated by pipeline functions, and green for the final output). *Rectangles* are (sets of) functions representing actions performed on data objects..

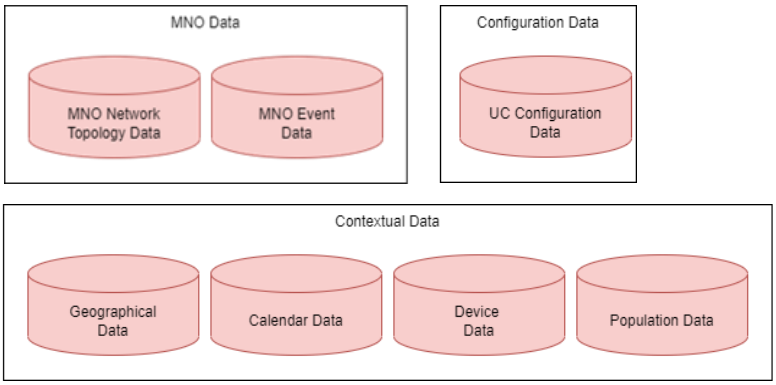
3.1. Input Data and preprocessing

The pipeline operates with three main categories of data: *MNO Data, Contextual Data, and Configuration Data*. These data are continuously ingested by the pipeline and enter at the beginning, flowing through various stages.

MNO Data includes *MNO Network Topology Data* and *MNO Event Data*. *MNO Network Topology Data* includes cell_id and coverage areas or data for estimating coverage, like antenna locations and characteristics. *MNO Event Data* consists of transactions/events between mobile devices and the network infrastructure collected from CDR/IPDR data and signalling data.

Contextual Data are composed of different types of data: *Geografical Data* cover grids, administrative boundaries, topography, street networks, etc., which can be enriched with MNO network topology data (e.g., cells dedicated to specific locations); *Calendar Data* provides information such as working days, weekends, bank holidays and special events (they can be enriched with MNO network topology data, like temporary cell placement or cell re-purposing, expecially for special events); *Device Data* includes data about the users of the mobile device, like foreign visitors, special client users, etc; *Population Data* like census records, tourism statistics, aggregated to a geospatial level. In the demonstration scenario they are used only at an aggregated level and exclusively in the final step for validation, benchmarking, grossing up and extrapolation.

Figure 3: Data objects representation of input data



Configuration Data define use case specific information, like specific requirements, selected indicators, time resolution and zoning system.

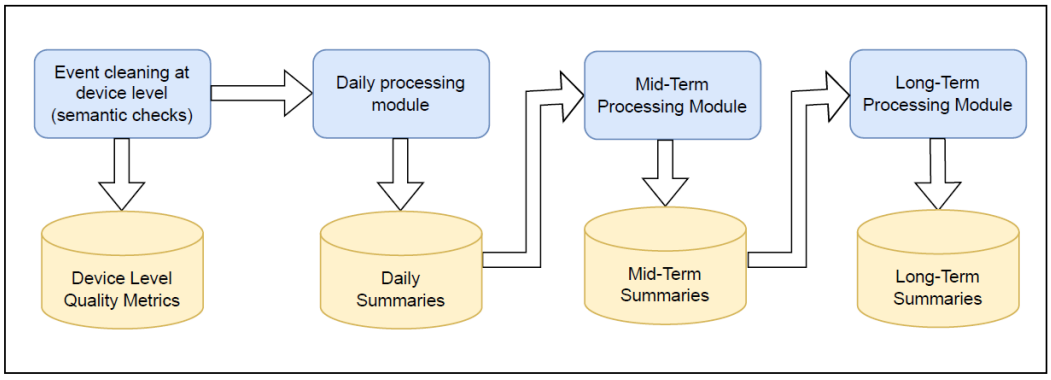
The first steps of the preprocessing provide the spatial cell coverage information for each cell_id at the given time and the cleaning of the input flow of events data received daily. The clean event data are then grouped per device and per day, generating a separate sub-flow of temporally ordered events for each device that is later processed by the longitudinal analysis module.

3.2. Multi-scale longitudinal analysis

This module implement the parallel analysis of individual device events longitudinally and at a multi-scale level. The data flow moves from smaller to larger scales, ensuring a one-way bottom-up processing approach. Smaller-scale analysis provides input for larger-scale analysis, with summaries at one timescale available for any module but not influenced by larger-scale summaries.This modular structure ensures extensibility for new case studies.

Daily Processing module: daily batches of event data are first processed by the *Daily Processing module* that allows functions to produce all the syntheses of the individual data that are required, as intermediate results, for the computation of the statistical indicators in the various use cases. The daily syntheses, denoted as *Daily Summaries,* provide information on events occurring whithin each single day for each device. These information are stored in separate data sub-objects serving as input for the next modules.The design of the *Daily Processing module* allows for the inclusion of additional functions by inserting other sub-modules and will support the availability of more than one single method.

Figure 4: Multi-scale longitudinal analysis at the device level (in the pipeline processing)



Longitudinal Processing modules: daily summaries are further analysed in the processing flow, first at a mid-term temporal scale and then at a long-term temporal scale.The *Mid-*

Term Processing module aggregate daily data and synthesizes daily summaries into mid-term summaries (e.g., monthly or quarterly), containing measures relevant for specific use cases.

Similarly, the *Long-Term Processing module* captures the long-term behavior of devices by analyzing daily and mid-term summaries over several months. Both the mid-term and long-term modules are designed modularly, with separate functions and data objects, ensuring the pipeline's flexibility and evolvability to accommodate new use cases and requirements as they emerge.

3.3. Output indicators and quality metrics

Output indicators are obtained after the last steps of the data processing that provide for data integration and for the application of statistical disclosure control methods. The development of sophisticated data integration methods is not the focus of this project, as this is pursued in a parallel research project funded by Eurostat (ESSnet MNO-MINDS [Ref 3]). However the modularity and flexibility of the methodological framework allow to incorporate future methodological improvements and new requirements, without the need of a full redesign.

Moreover, the framework incorporate mechanisms aimed at ensuring that the quality of the resulting statistical products is in line with official statistics principles and practices. Quality controls are implemented at different stages of the data processing flow, encompassing input data, intermediate results and final outputs.

4. Conclusions

By establishing standardized approaches to data collection, processing, and dissemination, the multi MNO project aims to ensure the consistency, reliability, and credibility of statistical products derived from MNO data. Ultimately, the project seeks to empower policymakers, researchers, and the public with accurate, up-to-date, and comparable statistical insights.

According to the principle of flexibility, evolvability and generality the proposed methodological framework includes a processing pipeline that can be applied to a wide range of use cases, with the production of population presence and mobility indicators in various statistical domains. Moreover, by incorporating state-of-the-art approaches, leveraging previous findings and adhering to principles of modularity, consistency, quality assurance and privacy protection,

the framework provide a solid foundation for the production of reliable and valuable statistical products.

References

1. Eurostat (2023). Reusing mobile network operator data for official statistics: the case for a common methodological framework for the European Statistical System – 2023 edition - *Products Statistical reports - Eurostat (europa.eu).* https://ec.europa.eu/eurostat/en/web/products-statistical-reports/w/ks-ft-23-001

2. Multi-MNO Project: “Development, implementation and demonstration of a reference processing pipeline for the future production of official statistics based on multiple Mobile Network Operator Data” https://cros.ec.europa.eu/landing-page/tss-multi-mno

3. ESSnet MNO-MINDS “Mobile Network Operator Methods for Integrating New Data Sources” https://cros.ec.europa.eu/dashboard/mno-minds