# A PROCESSING PIPELINE FOR EUROPEAN OFFICIAL STATISTICS: TOWARDS STANDARDISATION OF MOBILE NETWORK OPERATOR DATA PROCESSING

**Erika Cerasti** et al.

**Erika Cerasti, ISTAT - Istituto Nazionale di Statistica**
*erika.cerasti@istat.it*

Consortium leader

GOPA WORLDWIDE CONSULTANTS

positium NOMMON

Consortium partners

Subcontractors

Istat

Methodology and quality in official statistics

Advisory Board for the Use of MNO's Data for Official Statistics

REPUBLIC OF SLOVENIA STATISTICAL OFFICE

Statistics Poland

Individual experts

Testing MNOs

POST LUXEMBOURG

A1

Vodafone Italy

orange

Orange España

Vodafone España

GOPAcom.

Communication and dissemination

## Specific objectives:

- Development of an **open end-to-end methodological framework** for the processing of MNO data for official statistics;
- Development of a **reference open-source software pipeline** implementing the proposed methodological framework;
- Practical **demonstration of the processing** pipeline across five MNOs in four EU countries to produce a set of experimental statistics.

Use-cases domains: inbound tourism, outbound tourism, population, spatial statistics (e.g., functional urban areas), local commuting, and international commuting

*Period: February 2023 - June 2025*

***Develop a complete, open end-to-end PROCESSING PIPELINE as a PROPOSAL for the production of future official statistics based on MNO data, and to demonstrate it across data from multiple MNOs.***

Specific ***Design principles*** defined in coordination with Eurostat – EU statistical office

*Tailored to the European Statistical System (ESS) context and needs, the project develops a proposal for a methodological framework that will be reviewed by the ESS TF MNO and may be adopted as ESS standard.*

## Context

- **ESS Task Force on the use of MNO Data for official statistics:** established in 2021 to address the methodological aspects involved in the reuse of MNO for official statistics
  Chaired by Eurostat – 18 EU NSIs

- **Research grant**: **TSS Methodological development based on new data sources**
  **Object**: methodologies for Integration of MNO and non-MNO data sources
  Funded by Eurostat – 10 EU NSIs

Reference Position paper:

***"Reusing mobile network operator data for official statistics: the case for a common methodological framework for the European Statistical System – 2023 edition"***

https://ec.europa.eu/eurostat/en/web/products-statistical-reports/w/ks-ft-23-001

## Key considerations to ensure the framework's *flexibility, adaptability, and applicability* to diverse scenarios:

- **Methodological soundness**
- **Integration of previous findings** (*from past work by the ESS and NSIs at EU at national level)*
- **Stakeholder consultation**
- **Evolvability**.
- **Methodological challenges and recommendations**.
- **Modularity**
- **Use case specific descriptions of the methods using the pipeline**
- **Consistency**
- **Quality assurance**
- **Explainability**
- **Adherence to standards**
- **Openness and reproducibility**
- **Multi-MNO orientation**
- **Privacy protection**

# REFERENCE and DEMONSTRATOR SCENARIO

Set of assumptions about the context in which the proposed pipeline will be used by NSIs.

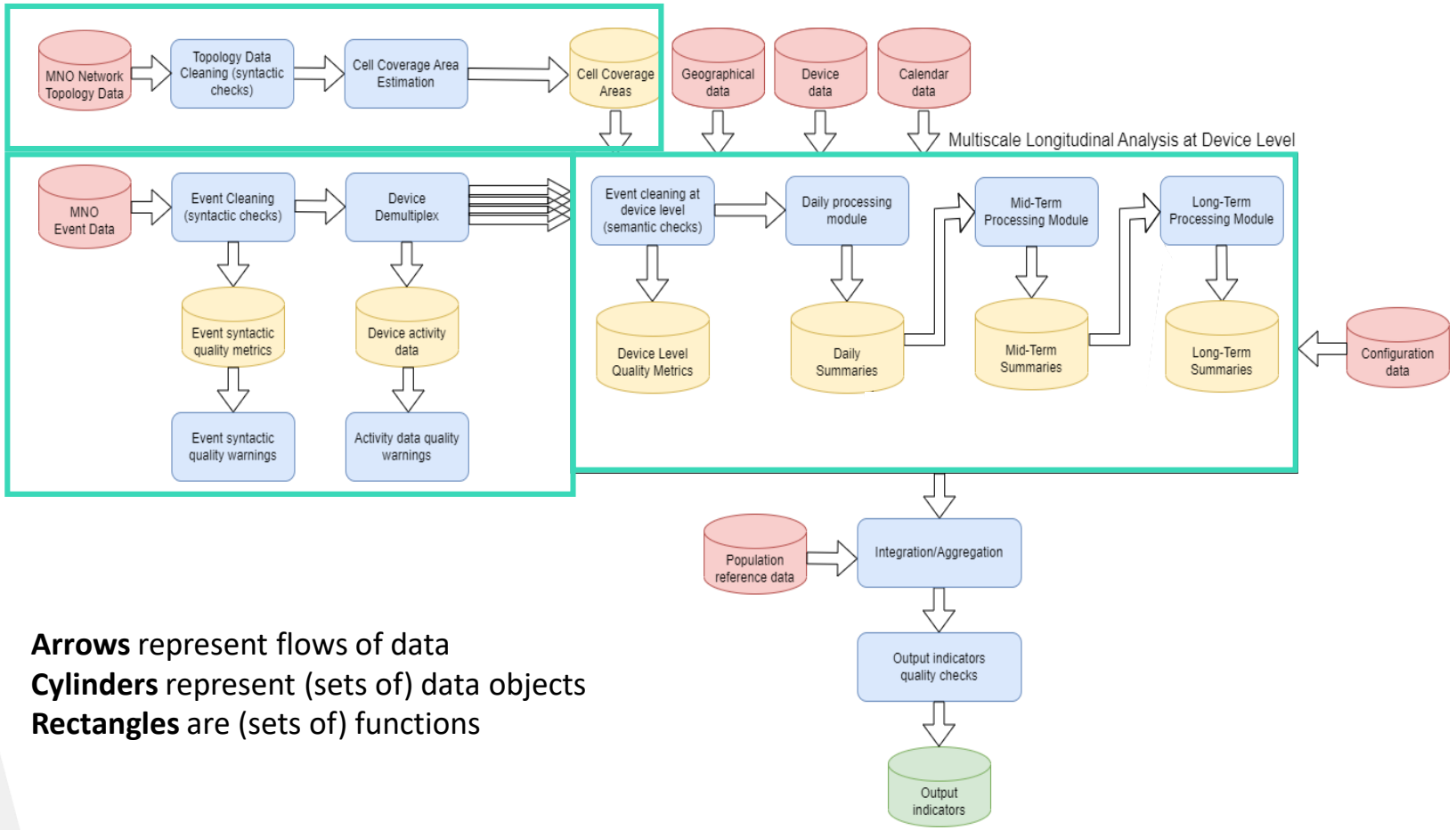
- In the **demonstrator scenario**:
  - Data processing environment - Disaggregated data will always remain within the MNO infrastructure.
  - SDC methods are implemented at the MNO level. The reasons for this approach encompass legal, technical, and privacy considerations.

- In the **reference scenario**:
  - SDC methods are implemented at the NSIs level.
  - Advancements in integration and privacy preserving techniques.

This project is not devoted to the investigation of methods for integrating MNO data with other non-MNO data sources or SDC methods

(other Eurostat projects will be devoted to that)

# PIPELINE PROCESS FLOW DIAGRAM

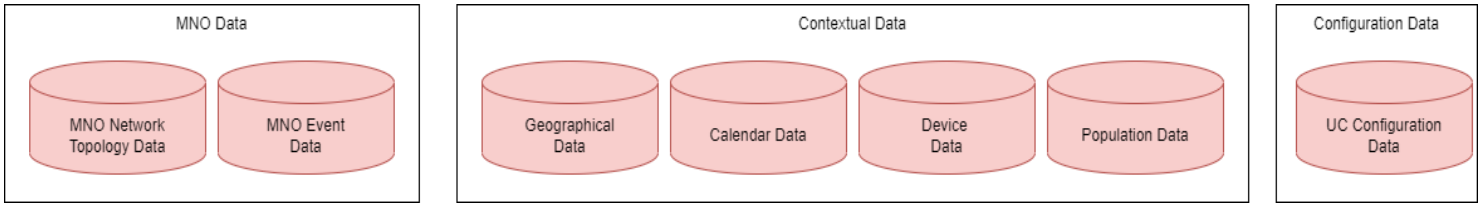



**Arrows** represent flows of data
**Cylinders** represent (sets of) data objects
**Rectangles** are (sets of) functions

# INPUT DATA OBJECTS

The process take as input three data categories containing different types of data objects:

- **MNO Data:** MNO network topology data, MNO event data
- **Contextual Data:** geographical data, Calendar Data, Device Data, Population Data
- **Configuration Data**: data used to specify the use case (i.e. selected indicators, time resolution, zoning system, use case specific requirements, etc)



# PROCESSING NETWORK TOPOLOGY DATA

## Cell footprint estimation module

The cell footprint determines how much a cell "covers" each tile of a spatial grid.

The aim of this process if to generate ***a standardised spatial representation*** of the cell footprint for each cell based on the available MNO Network Topology Data. When latitude and longitude information about the device location are not provided, we use the estimation of the geographic area, covered by the cell the device is connected to, as the probable location of the device.

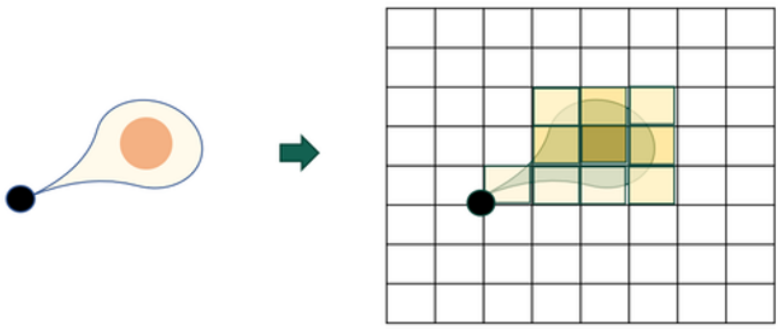
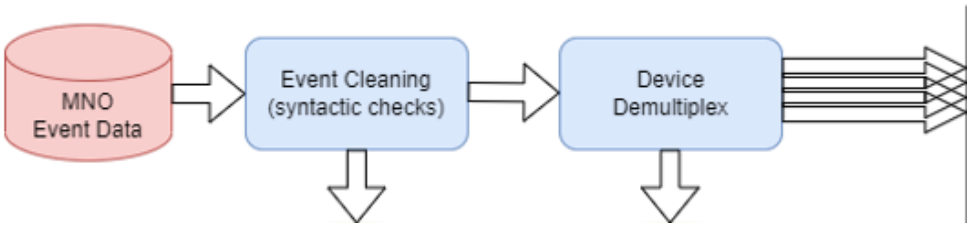


*Figure 12: From coverage area to cell footprint*

# PROCESSING EVENT DATA

The pipeline receives daily flows of Event Data. Each flow is processed according to the following chain:

- **Event Cleaning function**
  - Filter out malformed or missing data
  - A syntactic check is applied
- **Device Demultiplex module**
  - Cleaned Data are **grouped per device and per day**
  - sub-flow of temporally ordered events for each device




MULTI-SCALE
LONGITUDINAL
ANALYSIS PER DEVICE

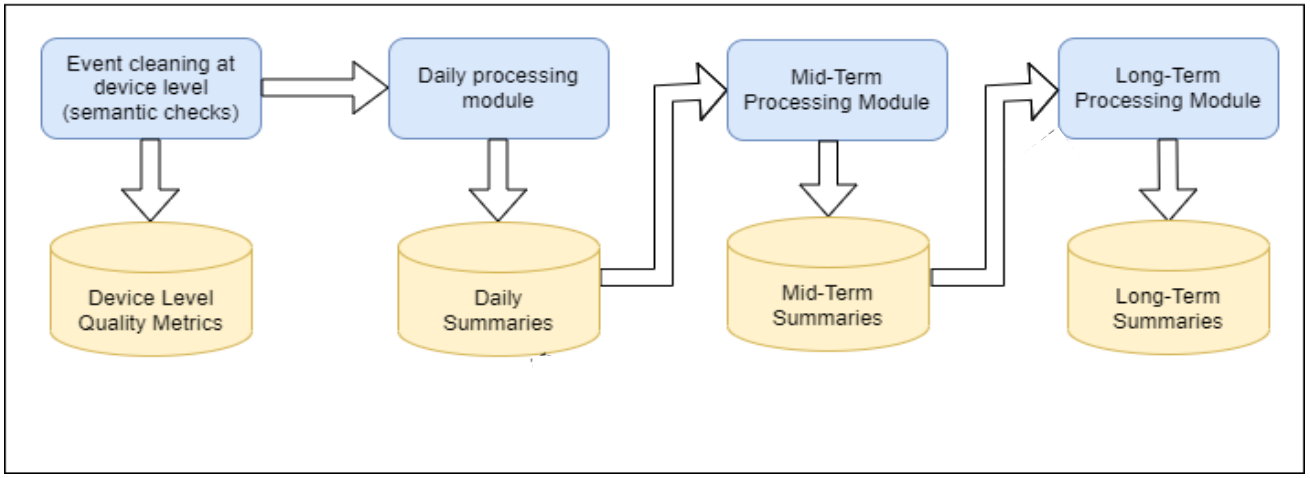
Quality measures

# MULTI-SCALE LONGITUDINAL ANALYSIS PER DEVICE

Events of each device are processed independently in a multi-scale dimension:

- **Event cleaning at device level**
- **Daily processing**
- **Longitudinal functions: Mid-Term processing and Long-Term Processing modules**

In the multi-scale analysis the data flow is unidirectional, from the smaller to larger scales (no feedback). Information at the individual device level is never exposed to outside the secure environment, only aggregate data are passed to NSI.




summary object may need to be integrated with several types of contextual information

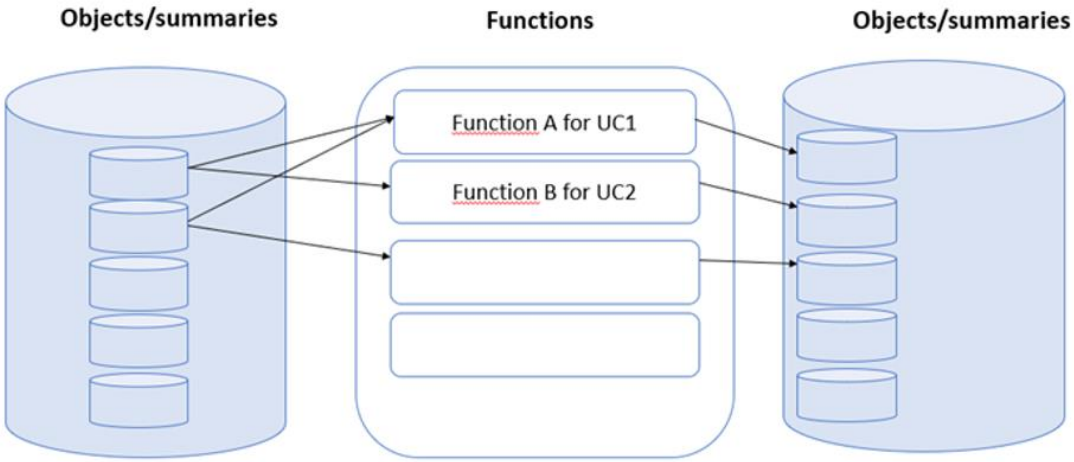
# MODULARITY OF FUNCTIONS AND DATA OBJECTS

**From Daily processing → Mid/Long-Term processing**

Daily summaries serve as input for different functions of the mid/long-term processing module.

The modularity of the pipeline enables different use cases providing different statistics of interest and facilitates the evolvability of the pipeline if new needs arise.

daily summaries

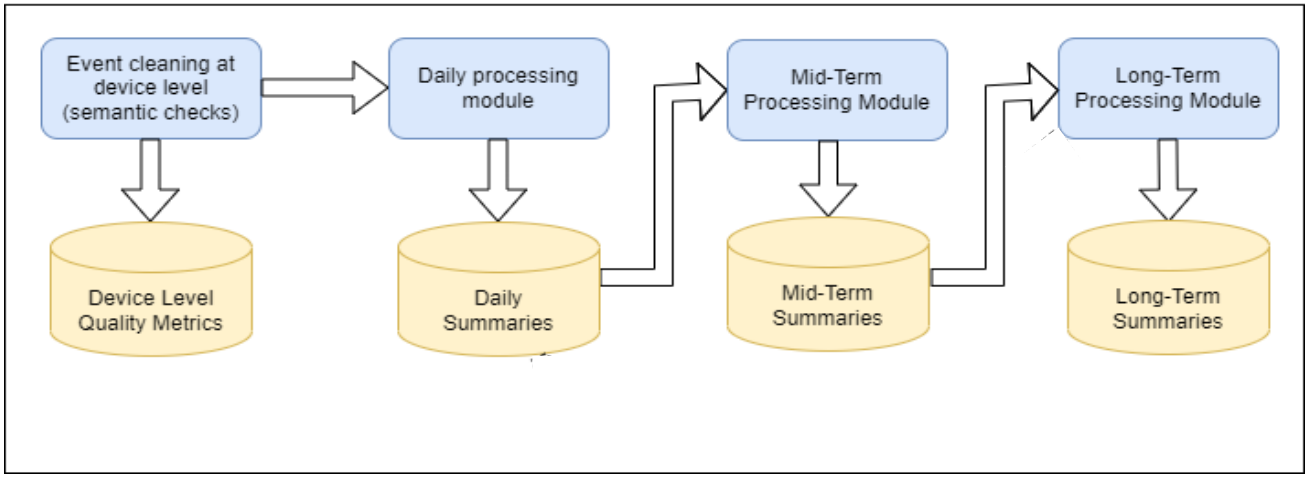



mid-term summaries

# MULTI-SCALE LONGITUDINAL ANALYSIS PER DEVICE

The periodicity of data ingestion and of software runs (implementation level) can be different from the periodicity of logical processing (semantic level)

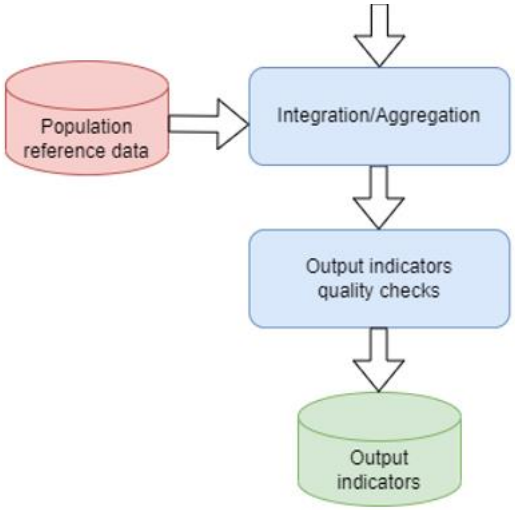
- **Mid-Term processing output example:**
  - the most common overnight place of a user, over a certain month or season of the year. Some use cases will require the production of outputs related to the mid-term interval, as the tourism statistics use case.
- **Long-Term processing output example:**
  - anchor points (e.g. home-location, usual environment) and other labelled/tagged information that is required by the use cases



# AGGREGATION AND INTEGRATION MODULES

Though aggregation and integration methods are not the focus of the project, the pipeline includes steps corresponding to these processes in order to provide an end-to end workflow.

Some advanced methods will be proposed in the separate research project:
***TSS Methodological development based on new data sources***



# For further details or if you wish to be informed on the progress of the work please:

**Follow our website:** https://cros-legacy.ec.europa.eu/content/multi-mno-project_en
*(temporary landing page – new project website will be launched in January 2024)*

***and/or***

**Contact: Florabela Carausu /** Florabela.Carausu@gopa.de
*(Project Manager, Service Contract Eurostat ref. 2021.0400 TSS Multi-MNO (Trusted Smart Statistics, multi-Mobile Network Operators)*

GOPA
WORLDWIDE CONSULTANTS

# Thank you for your attention

NetMob