Rome, Italy, 6-7 December 2023

ISTAT SECOND WORKSHOP ON METHODOLOGIES FOR OFFICIAL STATISTICS

Quality aspects using Mobile Network Operators data for Official Statistics

G. Ascari , E. Cerasti , C. Faricelli , P. Mattera , S. Piombo , R. Radini , G. Simeoni and T. Tuoto, Istat, Italy

Presenter: Giorgia Simeoni

Background and objectives

- Over the last years, increasing interest in the use of Mobile Network Operator (MNO) data for the production of Official Statistics in several domains
- Several research projects, case studies, experiments carried out over the world
- Many potentialities: improve timeliness, reduce costs, improve relevance
- … but could all the quality requirements for Official Statistics be fulfilled? Not only a matter of accuracy!
- We do not have yet the answer but we started analysing the issue in a systematic way, in order to build, in the long term, a **comprehensive quality framework** for official statistics based on MNO data
- How?

  First identifying the main characteristics of MNO data that differentiate them from more traditional statistical sources and processes and what risks they imply for quality

MNO Data

- As in Ricciato et al. (2023), we use the term “MNO data” to refer generically to all location data collected on the side of the network
- ***Event data***: generated by the mobile devices directly due to their activities
  - calling, receiving a call, sending and receiving text messages (call detail records or CDR), connecting to the internet (data detail records or DDR)
  - connecting to the telco network (signalling data) not voluntary activities of the device user, but information exchange concerning the establishment and control of the communication and the management of the network

  each event data corresponds to a particular device which contain a determined Subscriber Identity Module (SIM)

  - Big data in terms of volume, remain personal even after psudonymisation
- ***Network data***: technical data referring to the kind of technologies, the state of the antennas and network. They also contain the position of the antennas. This kind of data is at the base of all algorithms used to retrieve the position of the device.

No business data, or Mobile Phone data (e.g. GPS data from apps) are considered

Peculiarities in using MNO data in Official Statistics

- **Privately held data**: data are generated out of National Statistical Institute (NSI) – similiar to administrative data - but also out of National Statistical System.
- Due to their “Big data” and confidential nature, the **raw data processing usually is made in MNO-premise. NSI usually receives aggregated data.**

Not only data generation is out of the control of NSIs and with different purposes, but also part of data processing (pros and cons)

- Big impact on outputs of sequence of processing steps, algorithms that model them and version of the software that implements them

Main references

How did we approach the quality framework?

- Looking at the elements in the existing and **consolidated quality frameworks** for survey and administrative data based statistics and their applicability in this context (e.g.: ES Code of Practice, ESS Quality Assurance Framework, Total Survey Error approach and its estensions, BLUE-ETS project,etc.)
- Taking stock of the relevant experiences and proposals related to quality management of MNO data in statistics already formulated in other international and European projects (E.g. ESSnet Big data II)
- Actively participating in current EU funded projects on this topic:
    - “Development, Implementation and Demonstration of a Reference Processing Pipeline for the Future Production of Official Statistics Based on Multiple Mobile Network Operator Data” (TSS Multi-MNO)
    - “Trusted Smart Statistics: methodological developments based on new data sources ” (TSS-METH-TOOLS).

Overview of the proposed quality framework

Structured quality framework including:

- ○ Quality requirements at **Institutional level:** analysis of the ES Code of Practice principles of the Institutional Environment area and their application in the case of MNO data.

- ○ Quality layer that follows the production process:

  - **Input (raw MNO data)** quality: start from the approach developed for the quality of administrative data, and evaluate its applicability and possible extension or adaptation to the case of MNO data;
  - **Lower throughput** quality: quality of the processing made by MNO on their premises. More peculiar and complex part of the framework to be developed.
  - **Upper throughput** quality: the processing and analysis steps carried out by NSIs on pre-aggregated data. Challenges related to the integration with other data (also) and other methods to improve quality. Use of data for multiple MNO
  - **Output** quality: here the reference are the traditional quality criteria, but the way to evaluate and to report them to the users should be adapted.

Quality at Institutional Level

ES Code of Practice principles in the Institutional Environment area

1. Professional independence

1bis. Coordination and cooperation

2. Mandate for data collection and access to data

3. Adequacy of resources

4. Commitment to quality

5. Statistical confidentiality and data protection

6. Impartiality and objectivity

- Promote the **cooperation** with MNOs
- Facilitate data access to NSI through **legislation** or defining harmonised templates for **agreements**, including requirements for documentation and **transparency** and also clauses to assure **confidentiality**
- Develop ad-hoc **methods** for quality monitoring and assessment.

Input quality

Approach used in the framework for evaluating quality administrative data used for statistical purposes (Daas et al. 2009, BLUE-ETS project, SN MIAD)

Hyperdimension	Dimension	Quality Indicator	Method

Example

• Metadata	• Clarity	• Variable definition	• Clarity score of the definition

Input quality

Approach used in the framework for evaluating quality administrative data used for statistical purposes (Daas et al. 2009, BLUE-ETS project, SN MIAD)

Hyperdimension

Dimension

- Source
- Metadata
- Data

Input quality

Approach used in the framework for evaluating quality administrative data used for statistical purposes (Daas et al. 2009, BLUE-ETS project, SN MIAD)

Hyperdimension

- Source
- Metadata
- Data

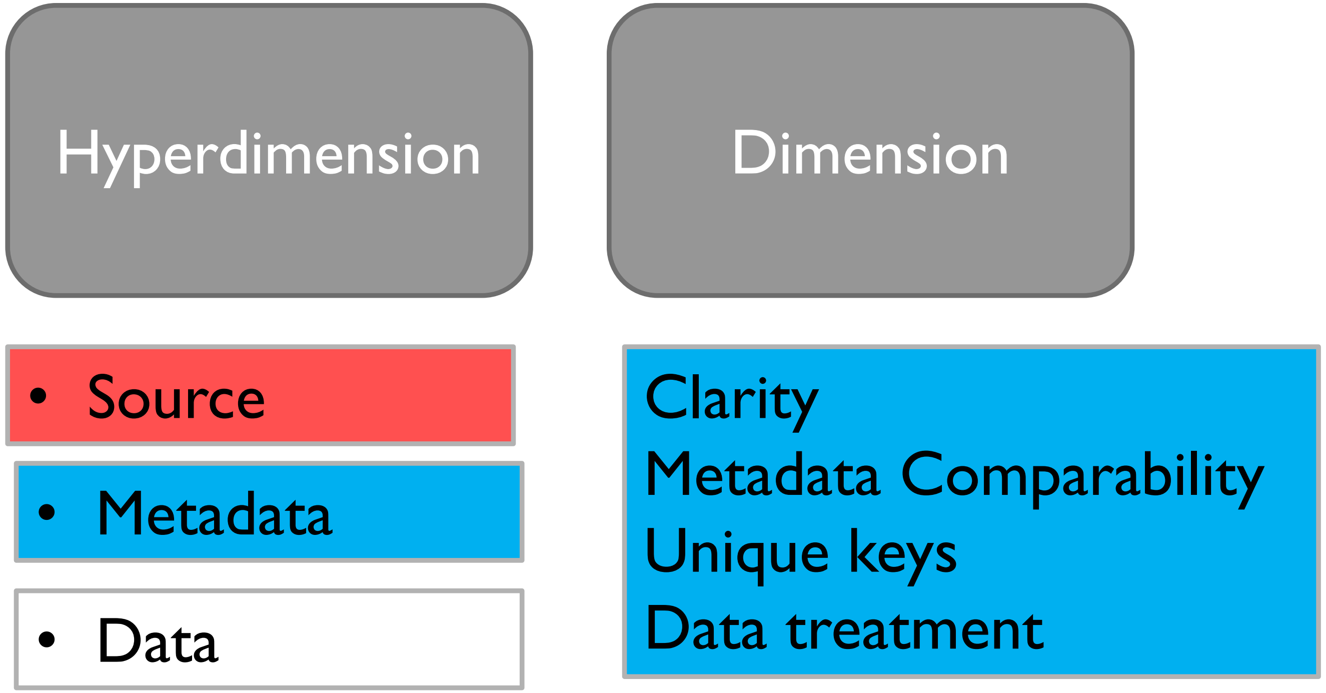
Dimension

Supplier
Relevance
Privacy and Security
Delivery
Procedure

While supplier and relevance dimensions can be easily translated from the application to administrative data to MNO data, the remaining 3 dimensions are more complex.
Privacy and security of raw data is managed by MNO at their premises and they are not usually delivered to NSI.
The application of this dimensions to the pre-processed intermediate aggregated output transmitted to NSI is more straightforward.
Information on procedures should not only refer to data generation but to data processing

Input quality

Approach used in the framework for evaluating quality administrative data used for statistical purposes (Daas et al. 2009, BLUE-ETS project, SN MIAD)



Hyperdimension

- Source
- Metadata
- Data

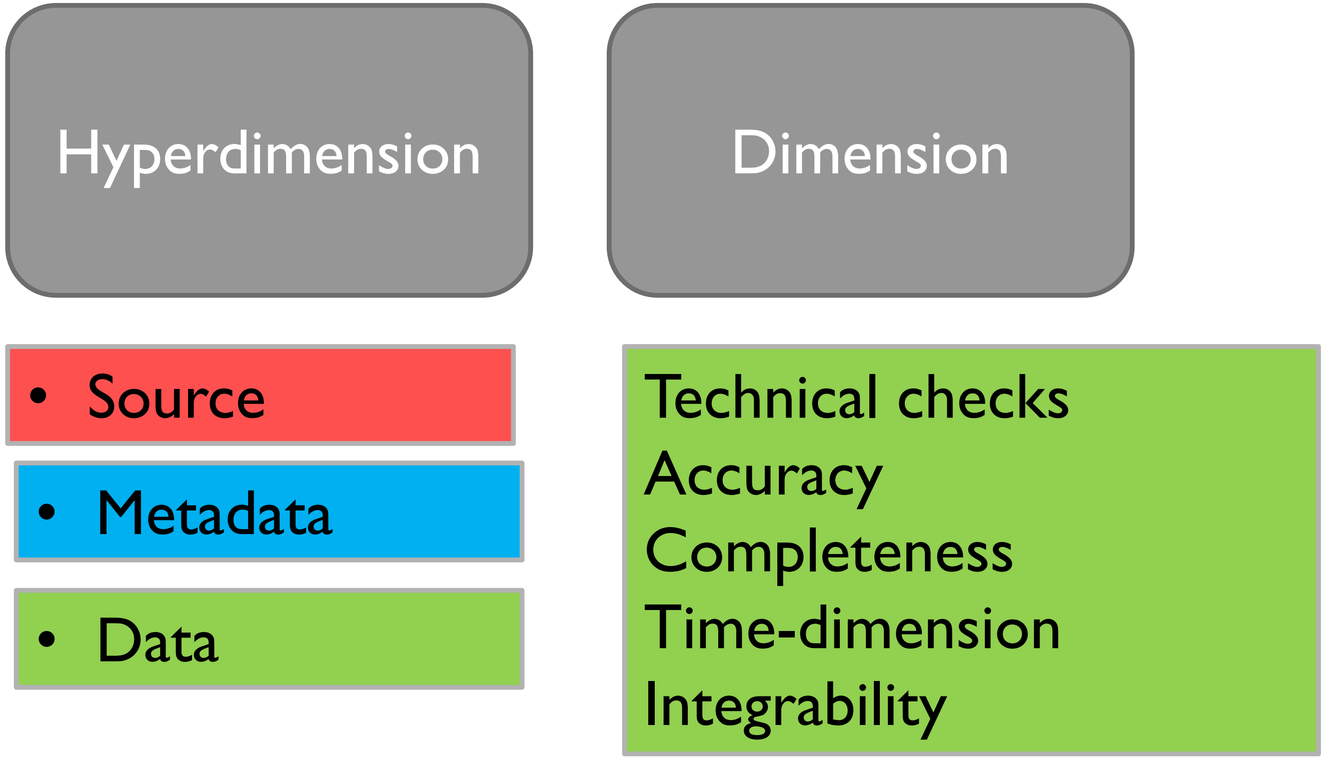
Dimension

Clarity
Metadata Comparability
Unique keys
Data treatment

Clarity of metadata describing e.g. the events and the variables in the raw data are particularly relevant, while their comparability with metadata for Official statistics should be «built» applying the right transformation to raw data. The presence and stability of unique keys in raw data is a re-requisite for their use in this context.
The information on Data treatment by the MNO is to be transformed from metadata provision to NSI-MNO dialogue, with NSI providing indication on how to proceed and MNO answering with the output of agreed checks on data and methods

Input quality

Approach used in the framework for evaluating quality administrative data used for statistical purposes (Daas et al. 2009, BLUE-ETS project, SN MIAD)



Hyperdimension

Dimension

- Source
- Metadata
- Data

Technical checks
Accuracy
Completeness
Time-dimension
Integrability

All the dimensions of data quality are extremely relevant, with technical checks applied by the MNO instead of the NSI, that should provide indication and receive feedback on that. Accuracy and completeness lead to reflections more connected to the desired statistical output. Undercoverage of MNO data for some part of the population is well known. Inaccurate network location data can provoke substantial errors in positioning devices and erroneous conclusions. This make emerge another relevant dimension for this data that is the space-dimension.

Lower Throughput quality

- Distinction between lower and upper introduced ESSnet big data II
- Lower throughput should transform raw data in statistical data
- It is carried out by MNO, possibly under NSI indication on the methods and to the quality check
- How to manage quality issue:
  - Set up a collaborative system with a continuous dialogue MNO-NSI
  - Obtain a detailed documentation of the process, the methods, the algorithms and the software used
  - Promote standardisation of the process, supported by Enterprise Architecture, with integrated quality evaluation of each step and method
  - Analyse the process from a statistical point of view to identify main sources of errors in a Total Survey Error-like approach.

Next steps

- The definition of the quality framework for MNO data is a long journey that we have only started
- The analysis of institutional aspects and raw data quality provide already an insight on some specific quality issues
- Next steps
  - Develop lower throughput quality: identify error sources arising, possible mitigating actions and quality measures.
  - Start developing upper throughput quality analyzing quality characteristics of intermediate aggregated data provided by MNO to NSI
  - After methods and procedures for upper throughput have been defined, quality assurance system should be developed also for this part, that is under the NSI control
  - Output quality analysis will then mainly focus on accuracy, clarity, comparability and coherence dimensions.

Thank you for your attention!

Giorgia Simeoni | simeoni@istat.it