ESSnet Big Data II

Grant Agreement Number: 847375-2018-NL-BIGDATA

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata
https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en

Workpackage I

Mobile Network Data

Deliverable I.6 (Quality)

A Proposal for a Statistical Production Process with Mobile Network Data

Final version, 18 March, 2021

Prepared by:

Roberta Radini, Tiziana Tuoto (Istat, Italy)

- Raffaella M. Aracri (Istat, Italy)
- David Salgado (INE, Spain)

Workpackage Leader:

David Salgado (INE, Spain)
david.salgado.fernandez@ine.es
telephone : +34 91 5813151
mobile phone : N/A

Contents

1. General Introduction and Motivation

This document contains a modelling exercise of the proposed end-to-end production process from raw telco data to final estimates of the target population according to the BREAL model developed by WPF in this project (see below for details). For a more comprehensive view of quality issues regarding mobile network data, the reader is strongly encouraged to read the deliverables from WPK on Methodology and Quality, where an important feedback from WPI regarding the description of metadata according to an adaptation of the SIMS v2 to new digital data sources has been provided. In this document, to avoid unnecessary repetition of material, we shall concentrate on the milestone of using the innovative BREAL to describe the production process with mobile network data.

The National Statistics Institutes (NSIs) have been investing for several years in standardization projects, with the aims of describing their specific activities through a standard framework, comparable models, and harmonised terminology. A standard framework that describes the statistical processes is crucial for process documentation, and for process quality assessment and improvement.

A well-known model for process standardization is the GSBPM (Generic Statistical Business Process Model [1]), which is dedicated to the generalization of the statistical production process, based on the business process model. The GSBPM was firstly developed taking in mind the survey-based production process. It was later adapted to define production processes based on the use of administrative data and statistical records [2]. The ESSnet project EARF (Enterprise Architecture Reference Framework), funded by Eurostat, defined not only the Business Process layer, but also all the components that describe the principles, artifacts, processes, relationships that allow NSIs to achieve their business objectives, according to the principles of the Enterprise Architecture [3].

The ESSnet project Big Data II devoted a Work Package, (WPF 'Process and Architecture') to the definition of a European reference architecture for Big Data, serving the purpose to (i) guide Big Data investments by NSIs and (ii) help the development of standardized solutions and services.

In parallel, the project also dedicated the present Work Package (WPI 'Mobile Network Data') to the incorporation of mobile network data in the production of official statistics. The proposed methodology in this package reveals the need for new business functions, since survey-based production techniques cannot provide adequate solutions for the challenges in this data source. Business functions are the activities carried out by an enterprise; they can be divided into core functions and support functions [14]. A full description of the terms and concepts related to what is covered by Mobile Network Data can be found in Deliverable I.5 [5]. There is a need to approach data acquisition in a new fashion, the highly technological nature of this digital data demands a non-negligible amount of preprocessing for statistical purposes, new statistical methods for Official Statistics need to be put in place to produce new insights, and visualization tools are needed beyond classical tables and graphs, just to mention a few. Thus, in order to configure a production process with this data source with the expected quality, the process must be designed and implemented according to the new architecture proposed in WPF. This document provides a first proposal in this line.

The outcome of WPF is what we called **BREAL** (*Big Data REference Architecture and Layers*) [4]. BREAL serves the purpose of guiding Big Data investments by NSIs and helping the development of standardized solutions and services to be shared within the ESS and beyond. BREAL has started an analysis of all architectures and frameworks on producing official statistics to define reusable processes and methods for Big Data. Therefore for a detailed analysis we refer to [4] on the applicability of the standards: Generic Statistical Business Process Model (GSBPM), Generic Activity Model for Statistical Organizations (GAMSO), Generic Statistical Information Model (GSIM), Common Statistical Production Architecture (CSPA), Common Statistical Data Architecture (CSDA), ESS Statistical Production Reference Architecture (SPRA), Eurostat Big Data Task Force (BDTF), NIST Big Data Reference Architecture (NBDRA), Cross-industry standard process for data mining (CRISP-DM).

BREAL is a set of artifacts organized according to the different layers that typically compose enterprise architecture, namely:

- **The Business Layer**, dealing with '**what'** NSIs do with respect to Big Data management. The artifacts of this layer are: (i) a set of principles, (ii) a set of business functions, (iii) a description of the Big Data based production process called Big Data Life Cycle and (iv) a set of Actors and Stakeholders.

- **The Application Layer**, dealing with '**how'** NSIs could / should realize the business functions and the Big Data Life Cycle in terms of application components and services.

- **The Information Layer**, dealing with '**how'** NSIs could / should realize the business functions and the Big Data Life Cycle in terms of data models.

Using the BREAL as the Reference Architecture of the Big Data Statistical Process, in this document we model the production process based on the Mobile Network Data (MND). For this purpose, we use the Big Data Life Cycle defined by the BREAL [4] as well as the description of the business functions that compose the Big Data Life Cycle in terms of application components and services.

The representation of the metadata and the data structure, which are modeled in the BREAL Information Architecture, was described in Deliverable I.5 "First proposed standards and metadata for the production of official statistics with mobile network data" [5] and in chapter 5 the mapping with the layers provided by BREAL is reported.

Finally, chapter 6 introduces the mapping with the BREAL Support Application; they concern communication to citizens, agreements with providers and sharing the knowledge on MND and the data processing algorithms.

These topics within the MND still need further study and analysis but this can be an opportunity to formalize and model them as foreseen in BREAL.

2. The Big Data Life Cycle and Application Architecture

In the deliverable F1 [4] of WPF 'Process and Architecture', the Big Data Life Cycle for official statistics production is detailed: it specifies the list of business functions that composes the single

process of the Life Cycle. The list of the business functions is very rich because it includes all possible activities, we will use only some of these in our model.

For this reason, we consider it appropriate to start with the general description of the Life Cycle for official statistics production on MND and, in a second step, we describe the business functions that are used in the single use cases that we will model.

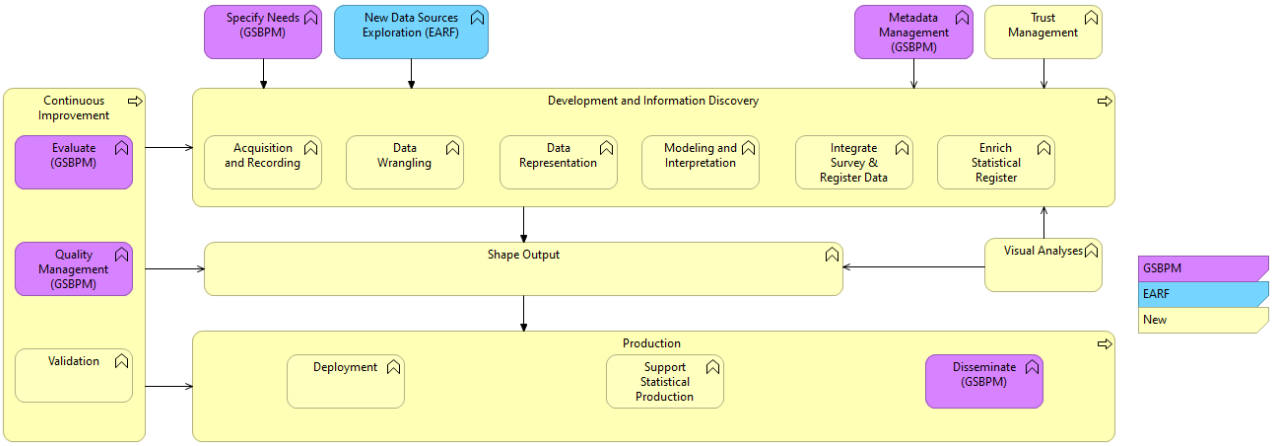
Big Data Life Cycle for official statistics production is organized into three major business process areas:

• Development and Information Discovery – where the exploration of the Big Data source, its integration with other data and the discovery of information take place;

• Production – where actually statistical products are created through the use of Big Data sources;

• Continuous Improvement – where we monitor and assess the Big Data source usage, with a focus on the population coverage issues and the validity of the models used.

Figure 1: BREAL Functions in the Big Data Life Cycle



For the use of MND, we will focus on the Development and Information Discovery phase, since the level of maturity of usage of this data in official statistics seems still not ready for the production phase.

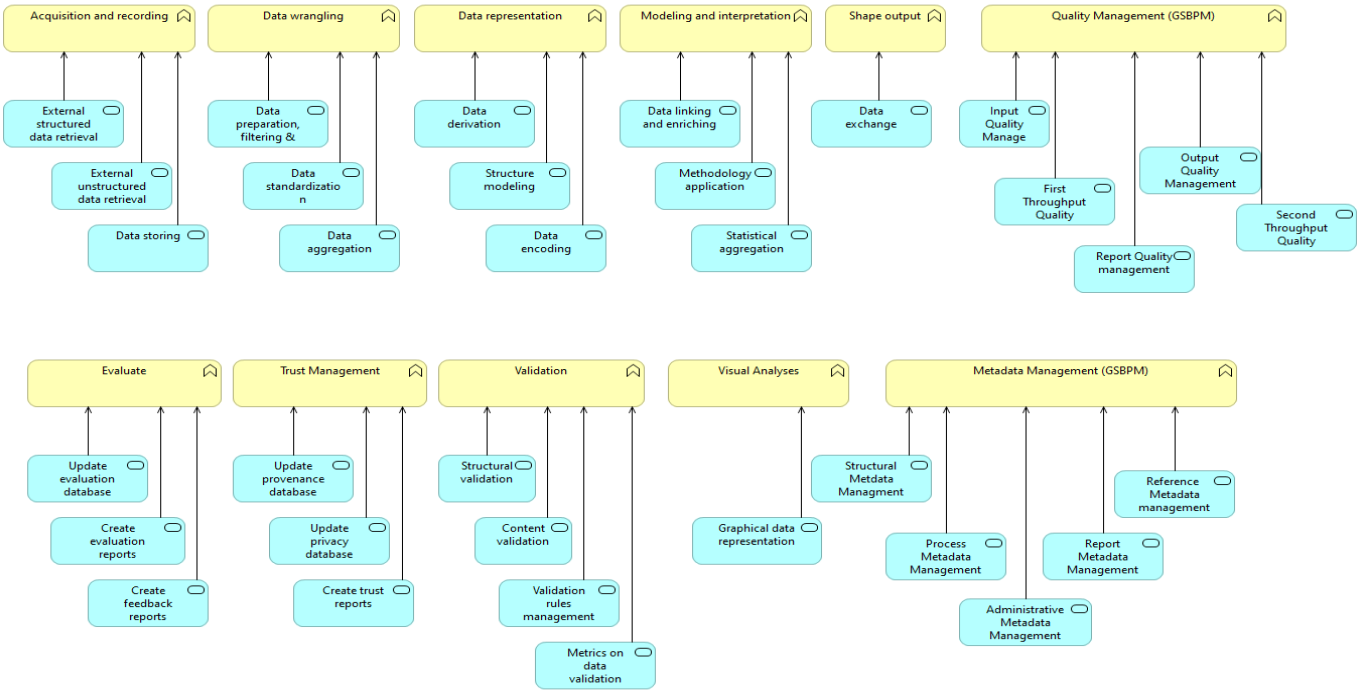
In the first phase of the Big Data Life Cycle, the two most relevant business functions are **Specifying Needs** (GSBPM) and **New Data Sources Exploration** (EARF). Obviously, they group different activities requiring Big Data examination skills, knowledge and resources. In addition, the business functions **Metadata** (GSBPM) and **Trust Management** support and serve the **Development and Information Discovery** phase. The **Trust Management** is crucial to the Big Data Life Cycle, as the overall consistency of measurements is an essential point for the adoption of Big Data in official statistical production. It is important to note that the accuracy and consistency of a Big Data source may be high at a given moment, but its reliability over time may prove to be problematic. This aspect is particularly relevant for the MND, in the absence of a specific regulation that allows NSIs to regularly access to them.

For the business processes "**Development and Information Discovery"** BREAL has specified the main business functions, which are:

(i) Acquisition and Recording - Ensuring access to the data,
(ii) Data Wrangling - Couples three business processes: extraction, cleaning and annotation,
(iii) Data Representation - Necessary when data is unstructured or semi-structured,
(iv) Modeling and Interpretation - Using algorithms and models specific to Big Data.

For some of them, it is relevant to specify some application services, i.e. application components that implement specific functions. The detail of each application service is explained in figure 2.

Figure 2: BREAL Business Functions and application services for the Development and Information Discovery area



The following sections describe the business process and the application architecture for the usage of MND. The representation obviously depends on the application, for this reason we identify two use cases of MND in Official Statistics to map their implementation process on the BREAL model.

These two processes, in addition to providing a high level of processing complexity with many components involved, illustrates two generic applications of MND in different statistical domains as concrete insights.

In particular, they define how to move from SIM information to estimates of number of individuals and also to the classification of the displacement behavior of individuals. Two main inherent outputs arise: *commuting*, i.e. the Origin-Destination Matrix, and *residence or permanence* in a place, i.e. the present population.

3. Application architecture for MND: the case of the Present Population estimation

The MND can be exploited for several purposes to provide insights in many fields of interest for Official Statistics. The WPI has recognized how a common denominator for many possible

applications with the MND is the identification of the so-called present population, i.e. the target statistical population present at a given time in a given space. For this reason, we use the derivation of the present population as the first use case to describe the application architecture to the MND.

To introduce the application architecture for MND we exploit and strictly connect with the deliverable I.3 on Methodology "A proposed production framework with mobile network data" [6], where the business function and the application services/components are deeply described from the logical and statistical perspectives. The deliverable I.3 on Methodology adopts a modular structure for the description of activities belonging to business functions and the application components that fully adhere to the BREAL philosophy. We also illustrate this advantage for some business functions and the application services of this use case.

Shortly, the deliverable I.3 on Methodology identifies the following steps:

1. Geolocation of mobile devices.

2. Device duplicity classification.

3. Statistical filtering.

4. Aggregation.

5. Inference.

Clearly, in the uses of MND for estimating the present population we still start with an "Acquisition and Recording" business function. However, it could be the case that NSIs cannot physically access and record the data, while the Mobile Network Operators (MNOs) might be able to run shared and agreed algorithms and methodologies on their premises and provide the NSIs only with aggregated results.

The "Data Wrangling" business function, declined in "Data preparation, filtering and deduplication" application services still need to be applied. They consist of automatic procedures to check that only the agreed records are included in the analysis. If MND are not physically moved to the NSIs, these services need to be commonly designed and executed to the MNO premises.

The specific modules identified in the deliverable I.3 on Methodology can be organized according to the BREAL as follows:

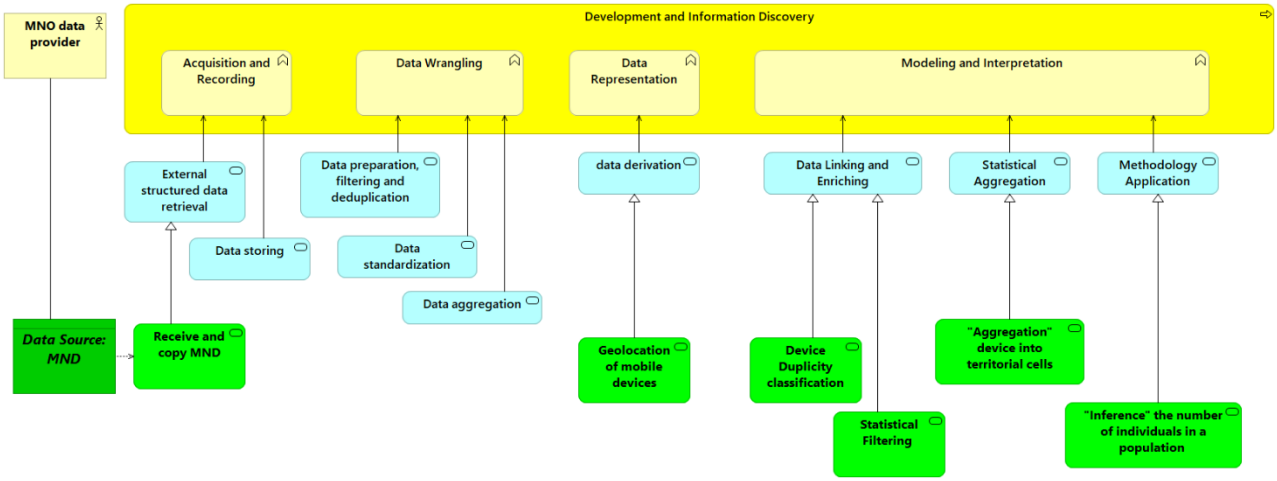
1. The *geolocation of mobile devices* can be included in the Business Function "*Data Representation*", and specifically in the application service "*Data Derivation*".
   Actually, the set of procedures belonging to this task aims at "deriving" (estimating, indeed) the geolocation of mobile devices using as input the network data: the location of the antennas and other network characteristics, possibly enriched with other data in use at the NSIs, i.e. the land-use or GIS data, and data generated from the interaction between antennas and devices. See the Glossary of Terms provided in the deliverable I.5 on Standards and Metadata [6] for the specification of network data and details on network characteristics useful for the usage of MND in official statistics.
   It is worthwhile noting that this specific service of "Data Derivation" is not a *simple* derivation service, like deriving an age variable from a birth date. On the contrary, it might benefit from the application of complex statistical models, as introduced and discussed in deliverable I.3

on Methodology [5]. When the available data or other factors prevent from the application of the models proposed in deliverable I.3 on Methodology, other solutions, still sophisticated, can be adopted, as, for instance a location derived by the shape of the BSA or derived by a Voronoi tessellation based on the position of the antennas.
The level of specialization required by this service is however quite high. This service can be provided on the MNO premises, according to shared algorithm agreed with the NSIs.

2. The *device duplicity classification* can be included in the Business Function "*Modeling and Interpretation Representation*", and specifically in the application service "*Data Linking and Enriching*".
This service aims at identifying multiple devices carried by the same individual. Even in this case, we have multiple choices: we can apply a classification algorithm that compares each device with all other devices, as suggested in deliverable I.3 on Methodology [5]. As an alternative, we can use some aggregated info like the penetration rate. In both cases, the application service "Data Linking" expresses the need of linking the MND with both the target statistical population and the MND itself.
3. The *statistical filtering* can be included in the Business Function "*Modeling and Interpretation*", and specifically in the application service "*Data Linking and Enriching*".
This service aims at identifying and selecting the devices corresponding to the target statistical population. This service requires the design and implementation of algorithms to filter target devices in the MND. In the current use case on the estimation of the present population, this service provides thus population counts at different levels of territorial and time disaggregation.
4. The *aggregation* module can be included in the Business Function "*Modeling and Interpretation*", and obviously in the application service "*Statistical Aggregation*".
This service aims at aggregating the data at the device level into territorial cells. This module uses as input data the results of the previous modules on device location, duplicity, and filtering to estimate the number of individuals detected in each territorial cell of interest.
5. The *Inference* module can be included in the Business Function "*Modeling and Interpretation*", and specifically in the application service "*Methodology Application*".
This service aims at inferring the number of individuals in a target population using the number of individuals detected in the network and auxiliary information.

The previous Big data life cycle for the estimation of present population is depicted in figure 3. In figure 3, we map the business functions and application services of the BREAL on the specific use case of MND for the estimation of present population (in green on the bottom).

Figure 3: BREAL Business Functions and application services for the present population estimation with MND



As already mentioned, the implementation of the "Acquisition and Recording" of data from the MNOs and the "**Data Wrangling**" functions deserve a dedicated discussion as they depend on the supply agreements defined with MNO.

The micro data acquisition of MNDs entails risks for the violation of the citizen's privacy, therefore there are problems of a different nature to be able to deal with them directly in the internal NSIs systems.

So many NSIs are moving towards defining agreement protocols in the processing to be undertaken on the data to obtain pre-processed data from the MNO with a controlled violation risk. In this case, the acquisition function uses "Structured Data Acquisition" application services in aggregate form, while the "Data Preparation, Filtering and Deduplication" function is carried out in the MNO systems according to algorithms prepared in accordance with NSI.

4. Application architecture for MND: the case of the Origin/Destination Matrix

This case study actually reports a real life application that was experimented by the Italian National Institute for Statistics (Istat) in collaboration with an Italian MNO. At the time this report was written, the joint work on the case study was not yet concluded, so the description of the use case cannot be complete; nevertheless, we decided to include it in this deliverable, since it allows highlighting limits and potentialities of the BREAL description of usages of MND in official statistics through a real case. The collaboration was defined with exploratory purposes, however the results of this application will be evaluated for future uses within the production process and the possibility of new experimental statistics for the Institute as well.

4.1 The NSI and MNO agreement

The agreement with the MNO consists of the definition of the various phases of the design and development process without acquiring the data in the NSI premise. The joint working method between MNO and NSI (Istat) was formalized through the execution of a Sprint, i.e. a full immersion project involving a MNO's team, composed by network experts and data analysts, and a NSI's team, composed by thematic experts and data scientists.

According to the BREAL model, the NSI and MNO agreement is an instance of the "**Trust Management**" function as it regulates access to data, and manages the methods of IT security and respect for citizen privacy.

The NSI and MNO agreement actually describes not only the "**Trust Management**" function, as stated before, but also has an impact on the relationships between NSI and MNO and the related tasks of: raw data analysis, preparation of specifications, implementation of processing algorithms and analysis of the quality of the results (BREAL models this with the "**Provision agreement management**" sub-function of Support).

4.2 The Origin/Destination Matrix

The object of the case study is the Origin/Destination (O/D) matrix. In an O/D matrix the rows represent the origins and the columns are the destinations. This kind of matrix can represent different phenomena connected to the human mobility, for instance, we can consider an O/D matrix of trips, collecting the daily traffic flows between municipalities or other territorial units; on the other hand, we can be interested in an O/D matrix of commuters, representing the habitual mobility linked to the usual movements to reach the usual workplaces and schools/universities.

Some uses of the O/D matrix of trips can be: measuring local mobility, highlighting critical situations in terms of traffic peaks in particular days or time slots, facilitating the management of the local mobility service; in general this matrix allows urban planners and policy makers to understand travel patterns in urban mobility [10][11][12]. In fact, it is one of the most requested products by local administrations, such as large cities, to implement Urban Plans for Sustainable Mobility.

On the other hand, the commuting O/D matrix aims at the analysis of habitual mobility, it is therefore used for planning and managing long-term mobility infrastructures at the national level, and for managing local labor systems as well.

Traditionally in Italy the commuting O/D matrix is produced by the Population Census, and identifies all movements within the residence municipality or outside it for study or work reasons [13]. Recently, Istat has reviewed the census strategy, moving forward a register based population counts, integrated with a couple of two large yearly sample surveys. This means the current yearly sample survey may have some limitation in describing all the movements for work and study reasons within and between municipalities, therefore Istat is investigating the potentialities of MND in providing the kind of information traditionally supplied by decennial population census.

In this application, the movement has a particular definition, i.e. it is characterized by being a habitual movement for work or study reasons, the origin is the usual residence accommodation, the destination the place of work or study. Each day trip starts and ends to the same point, the usual residence, called origin.

For this application the MNO proposes to use signaling data (also called Probe Data) [5]. The intention of the collaboration is to create a proxy or a support product for the realization of the Commuting O/D Matrix for the Italian population. In fact, the definitions of commuter that can be obtained from the population census survey cannot be reproduced with the MND, since the reason of the movement is unknown in the MND.

The creation of the O/D Matrix of interest to Istat, which from now on we will define the O/D Matrix of Commuting, required the collection of the production requirements (**Specify Needs**) and once

the product of interest was identified, the definitions of the statistical units and variables that characterize the product.

4.3 New data sources

In accordance with the Italian implementation of the GDPR, the MNOs are enabled to archive and use data from the network for no more than 15 months, therefore the "**Acquisition and Recording**" service is carried out by the MNO and with it all services that process network raw microdata.

The activities of analysis and study of the new source are an implementation of the "**New data sources exploration**" function, and they were carried out independently by the NSI, and then enriched by a phase of study of the source jointly with the MNO. This study anticipated the processing phase, it was supported by visual analyses. For instance, in this phase we concentrated on the different coverage of antennas by the territorial distribution.

Depending on the interesting analysis, it was decided to use the probe data coming from the 4G network. Furthermore, in this first analysis, a sampling step was introduced, i.e., the data were collected from the network systems at each time T.

Since this application was intended with exploratory purposes, the geographical area of reference for the output was only an Italian region (Lombardy). This choice was established in the "**Data sources exploration**" phase. In fact, Lombardy has a high variety of the territory from several viewpoints:

- the morphology: Lombardy includes mountains, flat land, and pre-alpine area of the Lakes;
- the urbanization: Lombardy includes several large cities, i.e. Milan;
- the functionality and network: in this territory we have many firms and it is a crucial point in the communication network connecting the large cities of the northern and the southern area.

In addition to the territorial selection, along the year we choose 5 weeks that include the reference period of the Population Census. The choice of 5 weeks was considered sufficient to observe a habitual behavior and to discriminate travel for work and study that require a certain regularity from those for tourism and leisure. Furthermore, the period close to the reference date of the Yearly Population Census will make the MND comparable with those of the Census questionnaire which detects the commuting phenomenon with specific questions.

The validation phase was based on Istat data , i.e. the data from the Population and Labor Registers, and the census sample survey; it was necessary to structure the data so that they could be comparable; in particular the MND should be reported to territorial units consistent with Istat data, i.e. Municipalities, ACE (census area) and NIL (Local Identity Nuclei: NIL represent a territorial atlas, a verification and consultation tool for planning services, but above all for knowledge of the neighborhoods that make up the different local realities, highlighting unique and different characteristics for each nucleus). These activities correspond to the "**Data Representation**" function.

At this point, the raw microdata was filtered, deduplicated and the data were transformed according to established formats and coding jointly defined. This corresponds to the "**Data Wrangling**" function.

4.4 Modeling and outputs

On this basis, the modeling phase started. According to the definitions of commuters, the algorithm was developed on the basis of some parameters that would allow to label the trips and places visited.

This activity required to cycle the activities, as described in the "**Big Data Life Cycle**", implementing the "**Development and Information Discovery**" process through the functions of:

- Acquisition and Recording;
- Data Wrangling – Couples three business processes: extraction, cleaning and annotation
- Modelling and Interpretation – Using algorithms and models specific to Big Data

At each cycle, the impacts of the different values of the parameters and of the choices made on the localization techniques were analyzed and assessed (Review and Validate) through "**Visual Analysis**" made with reports and dashboards with territorial maps.

We moved on to the final modeling phase of the algorithm and the provision of aggregated data for:

1. the O / D Matrix of Commuting;
2. the usual resident population (calculated as the population usually present in a municipality at night);
3. the O / D Displacement Matrix, i.e. all those O-D displacements that do not foresee the return to the Origin during the day.

The first output corresponds to the desired and expected one. For this output we start an integration process with the corresponding data of the population census and with the data of the integrated register system, (**Integrate Survey & Register Data function**). Also, in this case we will make use of "**Visual Analysis**" functions to support the evaluation functions of the outputs and uses that can be implemented to support census production and more.

The other outputs were provided to support the validation that have to be carried out with respect to:

- the inference choices applied to link the information on SIM aggregates to that of population aggregates;
- the criteria for implementing commuting measures according to the definition given;
- the possible distortions due to localization techniques.

The evaluation of the outputs is still in progress; however, it has already highlighted that the localization introduces considerable distortion, in very small areas with mountain terrain orography, therefore one of the proposed solutions is to aggregate the territory by carefully studying the coverage structure of the antennas, with a tradeoff between granularity and a quality of output. An unresolved issue is the "**output privacy**" constraints of the MNO: at the moment MNO cannot provide aggregate data below a certain threshold and this constraint introduces distortion in the final outputs. For this issue, a possible solution requires actions with the authority for data protection, for defining a joint privacy management for the MNO and the NSI.

4.5 Description of the process and the Business functions

The specific modules identified in the realization of O/D Matrix of Commuting can be organized according to the BREAL on the Life Cycle for official statistics introduced in chapter 2 and illustrated in Figure 1.

The process Development and Information Discovery, illustrated in Figure 5, starts on the results of the function: **Specifying Needs**, **New Data Sources Exploration, Metadata** and **Trust Management** and in this case corresponds to: the Product Specifications of O/D Matrix of Commuting, the experience on CDR and Probe Data, the deliverable on Metadata and the Glossary, the definition of the agreement with the MNO. The process is composed into some functions of BREAL model that use the specialization of application service of model as follows:

- the "*Acquisition and registration*" function: the data analytics MNO team acquires from the MND for the analyzes agreed with NSI; the NSI team acquires the aggregated data prepared according to the algorithms defined in the collaboration. In addition to the aggregated data, the metadata that feed the glossary are also acquired.
- the Business Function "*Data Wrangling*" is serviced by the application service "*Data preparation, filtering and deduplication*" that specializes in:
    - "*filter Non-residential SIM in the chosen region and SIM used for IoT or M2M*" ,
    - *"Selection of MND in the region and sampling with time interval T",*
    - *"Link MND with BSA and cell centroid".*

  Moreover, it is serviced by the application service "*Data aggregation*" that specializes in "*Count the time spent in a BSA*";
- The Business Function "*Data Representation*" is serviced by the application service "*Derivation of the position on the territory (Municipality, NIL, ACE) using the centroid of BSA*" that is specialization of application service "*Data derivation*";
- The Business Function "*Modeling and Interpretation*" is serviced by:
    - the application service "*Derivation of the position on the territory (Municipality, NIL, ACE) using the centroid of BSA* "that is specialization of application service the "*Data linking & enriching* ". This solution is not the preferred one, in fact it has been evaluated that it has a good quality in the inhabited areas but it creates problems on the borders of the municipalities and leaves many small municipalities uncovered; on the other hand, it is quick to implement and has allowed to explore the other phases of the project;
    - the application services: "*Data elaboration and estimation of usual resident population*", "*Data elaboration and estimation of daily Origin and Destinations per each SIM*" are the specialization of application service "*Methodology application*". The service *estimation of daily Origin and Destinations* uses the *parameter of length of stay in a place and parameter of frequency of visit to a place* to choose the place of Destination, the algorithm decides the most popular place or where SIM spent the most time during the day;
    - the application services: "*Data estimation of population density by number of SIM*", "*Data elaboration and estimation of daily Origin and Destinations per each SIM*",

"*Estimation of usual Origin and Destinations data per SIM"* and "*Classification of SIM: Commuters or NOT*" are the specialization of application service "*Methodology application*". The service *estimation of Origin and Destinations* uses the *parameter of length of stay in a place and of frequency of visit to a place* to choose both the place of daily and usual destination; the algorithm decides the most popular place or where SIM spent the most time during the day using the parameters defined in a pre-analysis and then the classification of the SIMs for which Origin-Destination areas have been assigned is carried out;

- the application services: *"Data estimation of population and O/D matrix commuting density by number of SIM"* is the specialization of application service "*Statistical aggregation*". This estimation uses the MNO Market share referred to the territory.

- The Business Function *"Integrate Survey & Register Data"* is serviced by the application service *"Integration with other source: Population Census, Labor Register and Individual "Register"* that is specialization of application service the "*Match and Integrate data*";

The validation phase is still ongoing, with the use of benchmark sources (Figure 4).

The final output was chosen as information to support the dissemination of the products of the population census, but in reality, this could become a new output of the statistical production, perhaps initially published in the experimental statistics.

Also, in this case, an appropriate method of publication and information must be developed for the citizen so that he/she does not feel spied on and is appropriately informed about the privacy and security policies adopted by NSI.

Figure 4: BREAL Business Functions and application services for the O/D Matrix of Commuting

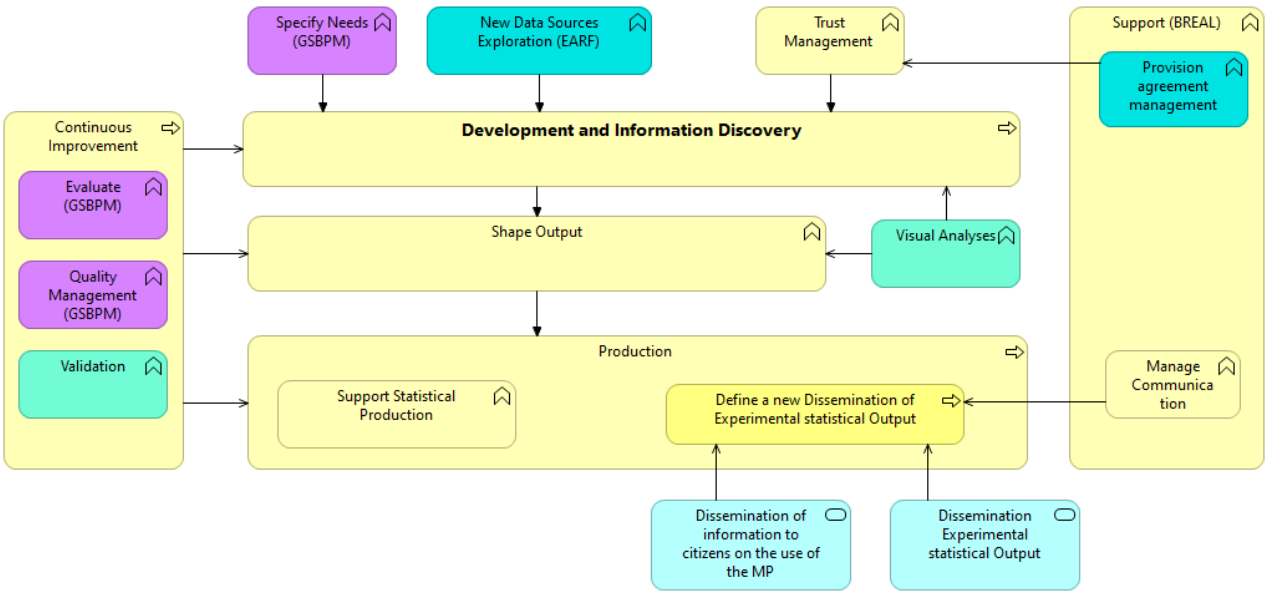
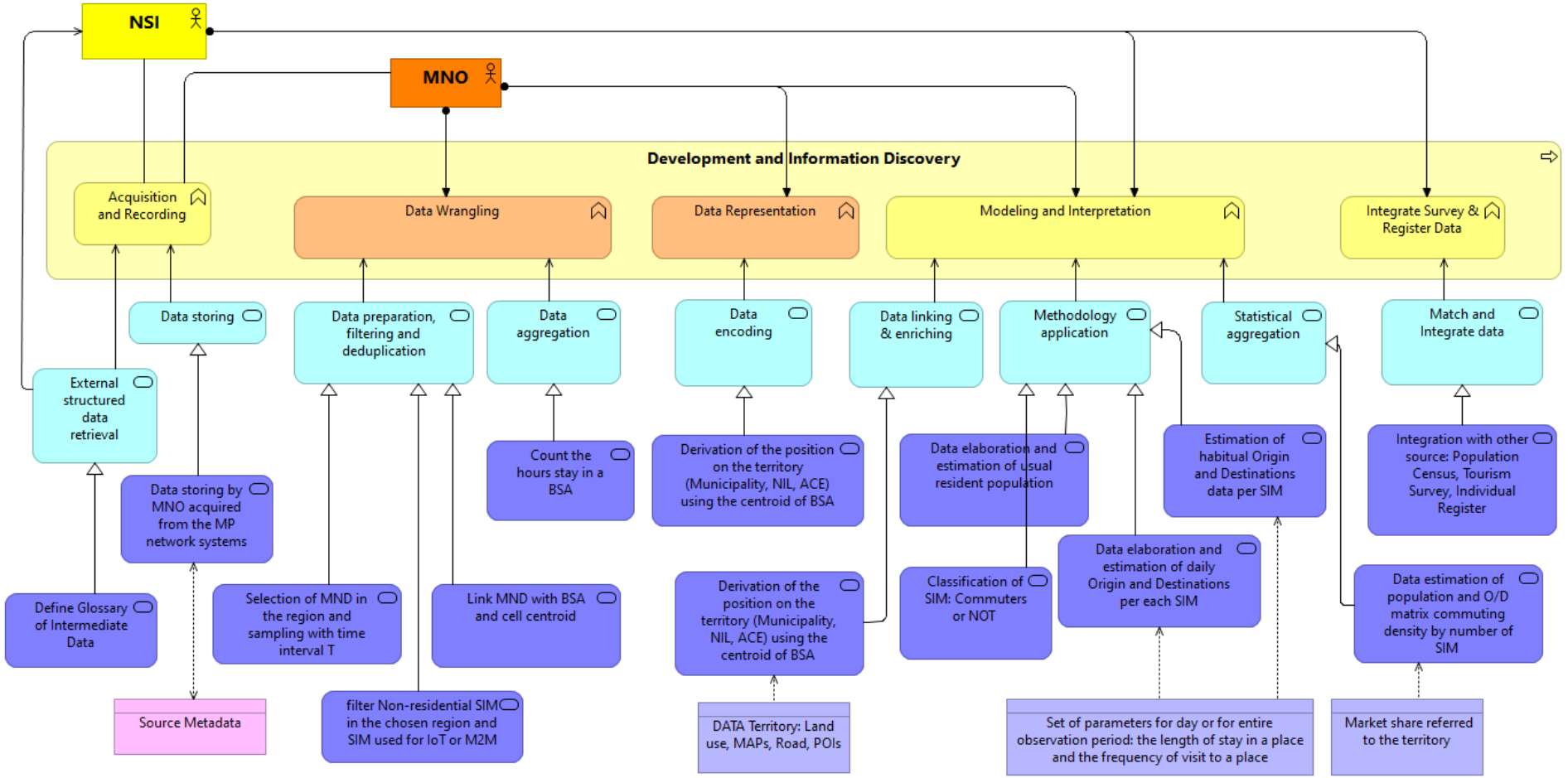


Figure 5: BREAL Business Functions and application services for the O/D Matrix of commuters



5. Generic Information Architecture

In the Deliverable I.5 "First proposed standards and metadata for the production of official statistics with mobile network data" [5], the business information level was described with particular attention to the Data Layer, which represents the raw data provided, by the MNOs and data of support used by the various production processes.

Also in this case the three layers of the metadata (and therefore also of the data) have been represented as:

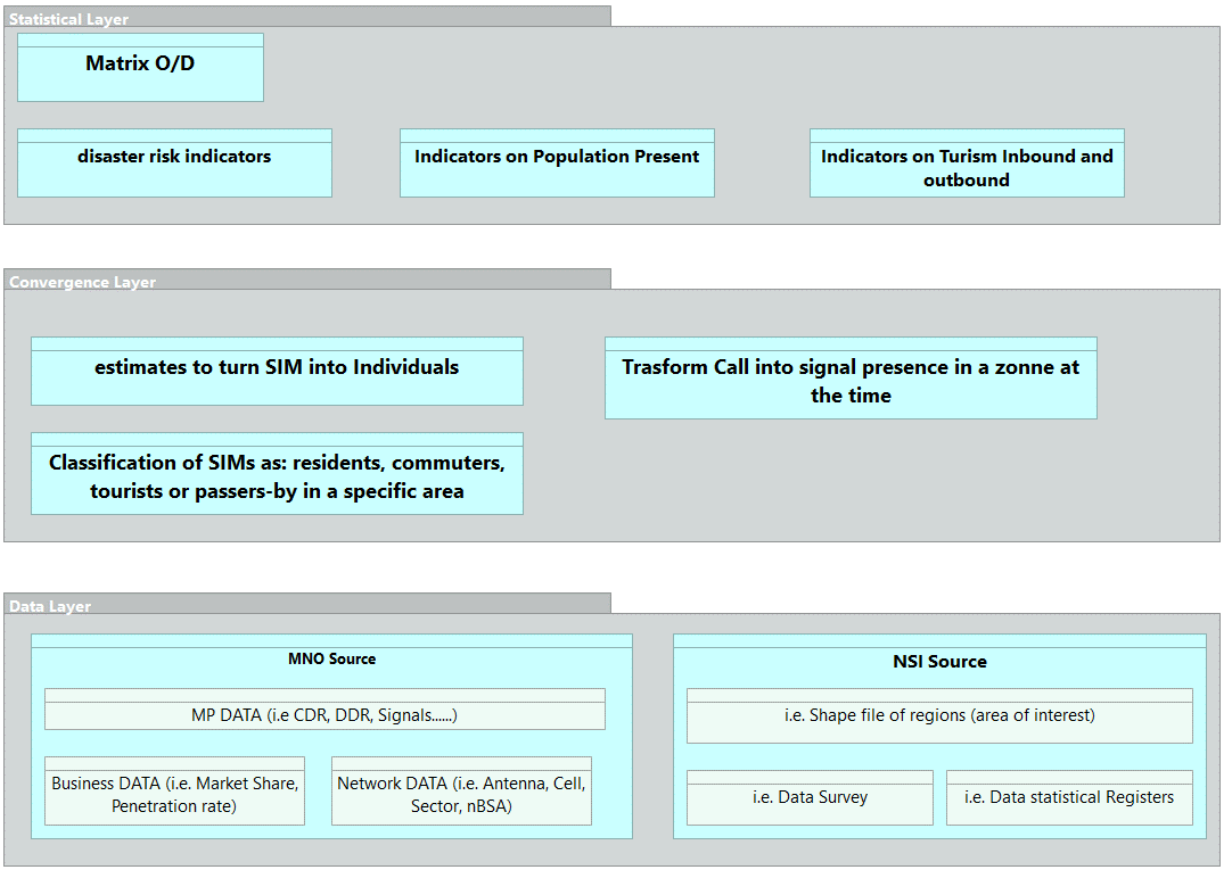
1. Data layer: input data, which mainly represent MNO data and NSI data and they are acquired and stored by the BREAL “Acquisition and Recording” business function [7];

2. Convergence layer: throughput data representing the filtered, processed and transformed data, also called Intermediate data, and are produced by the functions: “Data Wrangling” and “Data Representation”;

3. Statistical layer: output data, which represents the production data of statistical outputs, and they are produced by “Modeling and Interpretation”, “Integrate Survey and Register Data”, “Enrich Statistical Registers” and “Shape Output”.

This representation also complies with the ESS Reference Methodological Framework (ESS RMF) proposed for the standard statistical production of mobile network data [6].

Figure 6: BREAL Information Architecture with MND



6. Support Application Services

The section dedicated to the Support Application Services provides the definition of a series of support functions relevant for the management of Big Data, such as the IT management functions, identifying the corporate functions of data organization and infrastructure, network and IT with associated services. These will become clearer when MND processing goes into an implementation phase of a production process.

As mentioned in the previous paragraph, the processing of MND requires a careful assessment of the risk of privacy violation and with it the preparation of IT, organizational measures to mitigate and manage the risk as well as a measurement of this with particular attention to the risk of re-identification.

This set of services: **Manage Communication, Provision Agreement Management, Method and Tool Management and Legislative Work Participation,** also includes services relating to communication to citizens, agreements with providers and sharing within the scientific community and not only of knowledge on MND and also of data processing algorithms.

Bibliography

[1] UNECE (2019). GSBPM v5.1. https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1.

[2] UNECE (2019). GSBPM (version history) https://statswiki.unece.org/display/GSBPM/Old+Versions+of+the+GSBPM.

[3] **Eurostat (2015). ESS EARF. https://ec.europa.eu/eurostat/cros/content/ess-enterprise-architecture-reference-framework_en**.

[4] WPF (2020). BREAL: Big Data Reference Architecture and Layers. Business Layer. European project ESSnet on Big Data II. WPF_Deliverable_F1_BREAL_Big_Data_REference_Architecture_and_Layers_v.03012020.pdf.

[5] WPI (2020). First proposed standards and metadata for the production of official statistics with mobile network data. European project ESSnet on Big Data II. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/63/WPI_Deliverable_I5_First_proposed_standards_and_metadata_for_the_production_of_official_statistics_with_mobile_network_data_2020_05_28_draft.pdf

[6] WPI (2020). A proposed production framework with mobile network data. European project ESSnet on Big Data II. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/f/fb/WPI_Deliverable_I3_A_proposed_production_framework_with_mobile_network_data_2020_05_31_draft.pdf.

[7] WPF Template WPF Deliverable F2 noAnnex.pdf

[8] WP4 (2019). Create and communicate success stories. European project "Implementing Shared Statistical Services". https://ec.europa.eu/eurostat/cros/content/wp4-create-and-communicate-success-stories_en

[9] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/d/df/WPK_Deliverable_K5_First_draft_of_the_methodological_report_2020_06_17_final.pdf

[10] Zhang, Y.,Qin, X.,Dong, S.,Ran, B., "Daily O-D Matrix Estimation Using Cellular Probe Data", Transportation Research Board 89th Annual Meeting, 2010

[11] Bonnel,P., Fekih,M., Smoreda, Z.,"Origin-Destination estimation using mobile network probe data", Transportation Research Procedia, 2018

[12] Colak,S.,Alexander,L.P., Alvim,B.,Mehndiratta,S.R., "Analyzing Cell Phone Location Data for Urban Travel", in Transportation Research Record Journal of the Transportation Research Board 2526:126-135, 2015

[13] Toole JL, Colak S, Sturt B, Alexander LP, Evsukoff A, Gonzalez MC (2015), The path most traveled: Travel demand estimation using big data resources. Transportation Research Part C, http://dx.doi.org/10.1016/j.trc.2015.04.022

[14] Glossary: Business functions. Eurostat. Statistics explained. https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Business_functions