GOPA
WORLDWIDE CONSULTANTS

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)

Service Contract Number – 2021.0400

Deliverable 2.2: Updated version of technical documentation for scenarios, requirements, use cases and methods, and high-level architecture

Volume III – Methods and data objects

In association with:

Istat

positium

NOMMON

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)

Service Contract Number – 2021.0400

Deliverable 2.2: Updated version of technical documentation for scenarios, requirements, use cases and methods, and high-level architecture

Volume III – Methods and data objects

Version: final

Date: 20 November 2024

The information and views set out in this publication are those of the authors and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

Cover Photo by wd toroMC from Pexels

ABSTRACT

The Multi-MNO project aims to **develop, implement and demonstrate a proposal for a reference standard processing pipeline for the future production of official statistics in Europe based on MNO data from multiple operators**. If successful, the proposal developed by the project may be endorsed as European Statistical System (ESS) standard by the relevant ESS bodies. The term "processing pipeline" refers to the combination of a methodological framework and a reference open-source software adhering to such a framework. The processing pipeline developed in this project will cover an initial set of use cases; nonetheless, it will be designed to be general enough to provide the flexibility and growth capability required to cover other future use cases. The pipeline will be demonstrated and evaluated on real data from multiple MNOs in various EU countries.

This report specifies the different methods and data objects associated to each functional block of the pipeline described in Volume I of this deliverable. We differentiate between methods and data objects common to all use cases, and methods and data objects use case specific. This document covers the full set of methods and data objects for a selection of four use cases, namely: present population estimation, M-Usual Environment, M-Home Location indicators and internal migration. The selection of the use cases is motivated by the different perspective and analyses these allow for. The present population estimation use case builds a snapshot statistical estimation of mobile devices in a specific geographical area at a given moment in time. The M-Usual Environment indicators use case offers the possibility and builds along the concept of longitudinal analyses across time and space. It also allows to build, as one of its specific products, indicators on M-Home Location (i.e. the 'de facto' population counts) and estimates on Internal Migration.

This report complements the methodological framework proposed by the Multi-MNO project for the processing of multiple MNO data for official statistics (described in Volume I of this deliverable), as well as the definition of use cases (introduced in Volume II). This project deliverable focuses exclusively on the conceptual and methodological aspects. The technical specifications, the detailed architecture and the software design are defined in other project deliverables. Nevertheless, the approach followed for defining all methodological details, from high-level architecture to the definition of use cases and detailed methods description for core use cases, gives solid evidence of the feasibility in practice and relevance of the proposed methodologies.

Authors: The report was prepared under the technical coordination of Tiziana Tuoto (Senior Researcher, ISTAT). Further contributions, expertise and research support were provided by: (in alphabetical order) Gabriele Ascari, Roberto Cenciotti, Erika Cerasti, Loredana Di Consiglio, Cristina Faricelli, Ricardo Herranz, Paolo Mattera, Miguel Picornell, Roberta Radini, Giorgia Simeoni, Margus Tiru, Villem Tonnison, Luca Valentino and Ivan Vasilyev. Specific contributions for the present population use case and the geolocation and continuous time segmentation methods (in Volumes II and III) were provided by: Edwin de Jonge, Jan van der Laan, Matthias Offermans and Martijn Tennekes.

Acknowledgments: The authors express their gratitude for the technical review and coordination ensured by Fabio Ricciato, responsible for steering the project on behalf of Eurostat.

This report was prepared in the context of the service contract ref. 2021.0400 awarded by Eurostat to the consortium led by GOPA (Germany), in collaboration with the industry partners NOMMON (Spain) and POSITIUM (Estonia), and the National Statistical Institutes ISTAT (Italy) and CBS (Netherlands).

On behalf of the contractor, project management is ensured by Florabela Carausu (GOPA).

DOCUMENT VERSION STATUS AND FUTURE UPDATES:

The document is a work-in-progress updated version of the use cases introduced in the interim version of the methodological framework. This version addresses the feedback and comments formulated by the project Advisory Board and other groups of stakeholders on the first interim version of the document, and as well extends the use cases detailed to new ones. Nevertheless, its content may change in the final version. This document and any future updates will be publicly disseminated on the Multi-MNO project webpage: https://cros.ec.europa.eu/multi-mno-project

Readers are invited to submit comments and corrections or share their views via email to multimno-project@gopa.de

Abbreviations

AB	Advisory Board
BREAL	Big Data Reference Architecture and Layers
CDRs	Call Detail Records
CF	Cell Footprint
CGI	Cell Global Identity
EC	European Commission
ESS	European Statistical System
EU	European Union
FUA	Functional Urban Area
GDPR	General Data Protection Regulation
GPS	Global Positioning System
GSBPM	Generic Statistical Business Process Model
GSIM	Generic Statistical Information Model
ID	identifier
IMEI	International Mobile Equipment Identity
IMSI	International Mobile Subscriber Identifier
IoT	Internet of Things
LAU	Local Administrative Unit
M2M	Machine to Machine
MLE	Maximum Likelihood Estimator
MCC	Mobile Country Code
MNC	Mobile Network Code
MND	Mobile Network Data
MNO	Mobile Network Operator
MSIN	Mobile Subscription Identification Number
MS	Member State
NSI	National Statistical Institute
NSS	National Statistical System
NUTS	Nomenclature of territorial units for statistics
ONA	Other National Authority
PET	Privacy Enhancing Technologies
SD	standard deviation
SDC	Statistical Disclosure Control
SDG	Sustainable Development Goals
SIM	Subscriber Identity Module
SIMS	Single Integrated Metadata Structure
TFMNO	ESS Task Force on the Use of MNO data for Official Statistics
UC	Use Case
UE	Usual Environment
WGS 84	World Geodetic System 1984

Contents

1 Introduction 1
1.1 Scope and objectives of the document 1
1.2 Document structure 1
1.3 High-level pipeline definItion 2
1.4 Overview of methods and data objects 4
2 Module/Method 1: MNO Network Topology Data Cleaning – Syntactic Checks 11
2.1 Objective 11
2.2 Parameters 11
2.3 Input Data 12
2.4 Output Data 12
2.5 Methodology 12
3 Module/Method 2: MNO Network Topology - Quality Warnings 14
3.1 Objective 14
3.2 Parameters 14
3.3 Input data 15
3.4 Output data 16
3.5 Methodology 16
4 Module 3: Estimation of Signal Strength 23
4.1 Method 1: Estimation of Signal Strength from MNO Network Topology Data with Physical Properties 23
4.2 Method 2: Estimation of Signal Strength from MNO Signal Strength Data 26
5 Module/Method 4: Cell Footprint Estimation 29
5.1 Objective 29
5.2 Parameters 29
5.3 Input data 29
5.4 Output data 29
5.5 Methodology 30
6 Module/Method 5: Cell Connection Probability Estimation 32
6.1 Objective 32
6.2 Parameters 32
6.3 Input data 32
6.4 Output data 32
6.5 Methodology 32
7 Module/Method 6: Posterior Probability Estimation 33
7.1 Objective 33
7.2 Parameters 33
7.3 Input data 33
7.4 Output data 33
7.5 Methodology 33
7.6 Pseudo-code 34
8 Module 7: MNO Event Data Cleaning – Syntactic Checks 35
8.1 Method 1: MNO Event Data Cleaning – Syntactic Checks 35
8.2 Method 2: MNO Event– General Statistics 36
9 Module/Method 8: MNO Event Data – Syntactic Quality Warnings 39
9.1 Objective 39
9.2 Parameters 39
9.3 Input data 41
9.4 Output data 41

9.5 Methodology ... 41
10 Module/Method 9: Device Demultiplex ... 47
10.1 Objective ... 47
10.2 Parameters ... 47
10.3 Input data ... 47
10.4 Output data ... 47
10.5 Methodology ... 47
11 Module/Method 10: Event Cleaning at Device Level – Semantic Checks ... 48
11.1 Objective ... 48
11.2 Parameters ... 48
11.3 Input data ... 48
11.4 Output data ... 49
11.5 Methodology ... 49
11.6 Pseudo-code ... 50
11.7 Examples ... 51
12 Module/Method 11: Device Activity Statistics ... 63
12.1 Objective ... 63
12.2 Parameters ... 63
12.3 Input data ... 63
12.4 Output data ... 63
12.5 Methodology ... 63
13 Module/Method 12: MNO Event Data at Device Level – Semantic Quality Warnings ... 66
13.1 Objective ... 66
13.2 Input data ... 66
13.3 Output data ... 66
13.4 Parameters ... 66
13.5 Methodology ... 67
13.6 Pseudo-code ... 67
13.7 Examples ... 67
14 Module 13: Daily Processing Module ... 72
14.1 Method 1: Present Population Estimation ... 72
14.2 Method 2: Daily Permanence Score Estimation ... 75
14.3 Method 3: Continuous Time Segmentation ... 83
15 Module 14: Mid-Term Processing ... 91
15.1 Method 1: Mid-Term Permanence Analysis ... 91
16 Module 15: Long-Term Processing ... 95
16.1 Method 1: Long-Term Permanence Analysis ... 95
16.2 Method 2: Long-Term Home Location Method ... 102
17 Module 16: Device Filtering & single MNO Data Aggregation ... 106
17.1 Device Filtering and Aggregation for Usual Environment ... 106
18 Module 17: Merge single MNO Aggregates in Multi-MNO Aggregates ... 110
18.1 Merge single MNO Aggregates in Multi-MNO Aggregates for Usual Environment ... 110
19 Module 18: Projection of Multi-MNO relevant aggregates for the Use Case ... 112
19.1 Projection of the Multi-MNO Aggregates from the finest level to the geographic unit systems for the Usual environment ... 112
19.2 Spatial projection for the M-Home Location ... 115
20 Module 19: Estimation ... 116
20.1 Objective ... 116
20.2 Parameters ... 116
20.3 Input data ... 116

20.4 Output data ... 116
20.5 Methodology ... 116
20.6 Example of the Estimation Procedure in the M-Usual Environment indicators Use Case ... 117
21 Home Location Changes Detection Methods ... 118
21.1 Objective ... 118
21.2 Long-Term Home Location changes detection method ... 118
21.3 Parameters ... 118
21.4 Input data ... 118
21.5 Output data ... 118
21.6 Methodology ... 119
22 Data Objects ... 126
22.1 Input Data ... 126
22.2 Intermediate Results ... 135
Annex I – SDC for demonstrator scenario/Disclosure ... 171
Objective ... 171
Parameters ... 171
Input Data ... 171
Output Data ... 171
Methodology ... 171
Example ... 172
Annex II – Present Population Estimation: Variant 2 ... 176
Annex III - Pipeline application to the M-Usual Environment indicators use case ... 185

Index of Figures

Figure 1: High-level view of the pipeline ... 3
Figure 2: Radiation patterns for the azimuth and elevation planes ... 25
Figure 3: Signal strength at ground level from above (for a specific cell) ... 26
Figure 4: From coverage area to cell footprint ... 26
Figure 5: Transformation of information from MNO grid based data to project grid tiles ... 28
Figure 6: Logistic function using S_{steep} = -92.5 and S_{mid} = 0.2 ... 30
Figure 7: Present population in a cross-border region ... 72
Figure 8: Present population per country ... 73
Figure 9: Time stamps associated to events ... 78
Figure 10: Minimum distance between the two closest vertices of two cells' footprints ... 79
Figure 11: Visual representation of overview formulas ... 80
Figure 12: Simplified example of overlapping coverage areas for cells ... 87
Figure 13: Example of a possible sequence of events. ... 87
Figure 14: Example of administrative unit borders superimposed to the grid tessellation ... 114
Figure 15: Representation of the groups of tiles composing HL1 (left) and HL2 (right) ... 120
Figure 16: Representation of the overlap between the groups of tiles composing HL1 (light blue) and HL2 (grey): the common tile is highlighted in purple ... 121
Figure 17: An outline showing the superposition of area and borders of municipality A (light yellow) and municipality B (light blue) on the area covered by the HL1 group of tiles ... 122
Figure 18: Outline showing: the superposition of municipality A (light yellow) and municipality B (light blue) areas on the area covered by the HL1 groups of tiles (left); the superposition of municipality C (orange) and municipality D (green) areas on the area covered by the HL2 groups of tiles (right) ... 123
Figure 19: Outline showing the superposition of municipality A (light yellow) and municipality B (light blue) areas on the area covered by BOTH the HL1 group of tiles (left) and by the HL2 group of tiles (right) ... 124
Figure 20: Sparse matrix for the organisation of the daily summaries for the permanence score ... 188
Figure 21: Monthly (mid-term) processing module diagram ... 190
Figure 22: Long-term processing module diagram ... 191

Index of Tables

Table 1: Overview of methods and data objects 5
Table 2: Example of the daily summary table 74
Table 3: Example of the counts table 75
Table 4: Example of the daily summary table 178
Table 5: Example of inbound roaming devices summary table 178
Table 6: Example of the daily summary table of outbound roaming devices 179
Table 7: Example of the weighted counts table 181
Table 8: Example of the present population estimates on grid tile level 184
Table 9: Example of the present population estimates on regional level 184

1 INTRODUCTION

This report is the third of a set of three separate volumes that form altogether Deliverable D2.2 of the Multi-MNO project. For a better understanding of the content of this volume, we invite the readers to familiarise themselves with the content of Volume I – Detailed scope, requirements and methodological framework and Volume II – Use cases.

1.1 SCOPE AND OBJECTIVES OF THE DOCUMENT

This report specifies the different methods and data objects associated to each functional block of the pipeline described in Volume I of this deliverable. We differentiate between methods and data objects common to all use cases, and methods and data objects use case specific. This document covers the full set of methods and data objects for a selection of four use cases, namely: Present Population Estimation, M-Usual Environment indicators, M-Home Location indicators and Internal Migration.

1.2 DOCUMENT STRUCTURE

The introductory section recalls the project's background and objectives and the high-level pipeline definition. In addition, it clarifies the scope of the document and its structure and provides an overview of the methods and data objects described in detail in the forthcoming sections. These represent the two main blocks of information developed in this document, as follows:

- **Chapters 2 to 21** introduce one by one each of the modules and/or methods the pipeline works with. These are either core/common modules and/or methods specific to the four use cases covered in the document.
- **Chapter 22** introduces the data objects the different module and methods use as input or produce as output. To provide a clear linear logic of the processing done by the pipeline, we differentiate between input data and intermediate results and output data/statistics.
- **Annex 1 'SDC Method for the demonstrator scenario'**
- **Annex 2 'Device filtering and aggregation for Present Population: variant using population totals'**
- **Annex 2 'Exemplification of the pipeline's application to the M-Usual Environment indicators use case, including pseudo-code'**

1.3 HIGH-LEVEL PIPELINE DEFINITION

Herewith, we provide a brief overview of the high-level definition of the pipeline, which is introduced in detail in Volume I of this deliverable. **FIGURE 1** displays the different elements of the pipeline. We distinguish three main workflows within the processing pipeline, namely:

- Advanced geolocation (based on previous results of work developed in the context of the ESSnet Big Data II, Work Package I);
- Longitudinal analysis of individual device data (increased temporal scope from daily to mid-term and, finally, to long-term);
- Aggregation, estimation methods (designed to embed methods from the on-going ESSnet project MNO-MINDS)

We also distinguish between the processing of the two types of MNO input data (i.e. network topology and event data), which jointly build the common workflows of the pipeline; i.e. they are not use case specific. On the opposite, the longitudinal analysis of individual device data, and the aggregation and estimation workflows are use case specific.

Along the pipeline architecture quality checks and quality warnings go along with the workflows to assure the quality of the input data, of the intermediate results and, finally, of the statistical outputs.[1] Each module and/or method uses as input either raw MNO data, eventually combined with other reference data (e.g. calendar info, etc.) or intermediate results produced by previous modules and/or methods in the pipeline. This architecture ensures the requirements of modularity, flexibility and evolvability of the pipeline.

Each of the modules and/or methods the pipeline works with, as well as their input and output data are specified in detail in the next chapters of this document. The workflows that are use case specific are detailed for four use cases, namely: present population, M-Usual Environment indicators, M-Home Location indicators and internal migration. Nonetheless, since this project is not dedicated to investigating methods for integrating MNO data with other non-MNO data sources, the integration and estimation methods are limited to basic considerations and/or cases to demonstrate this integration module in the pipeline.

[1] See, as well, for further reference the project deliverable D3.

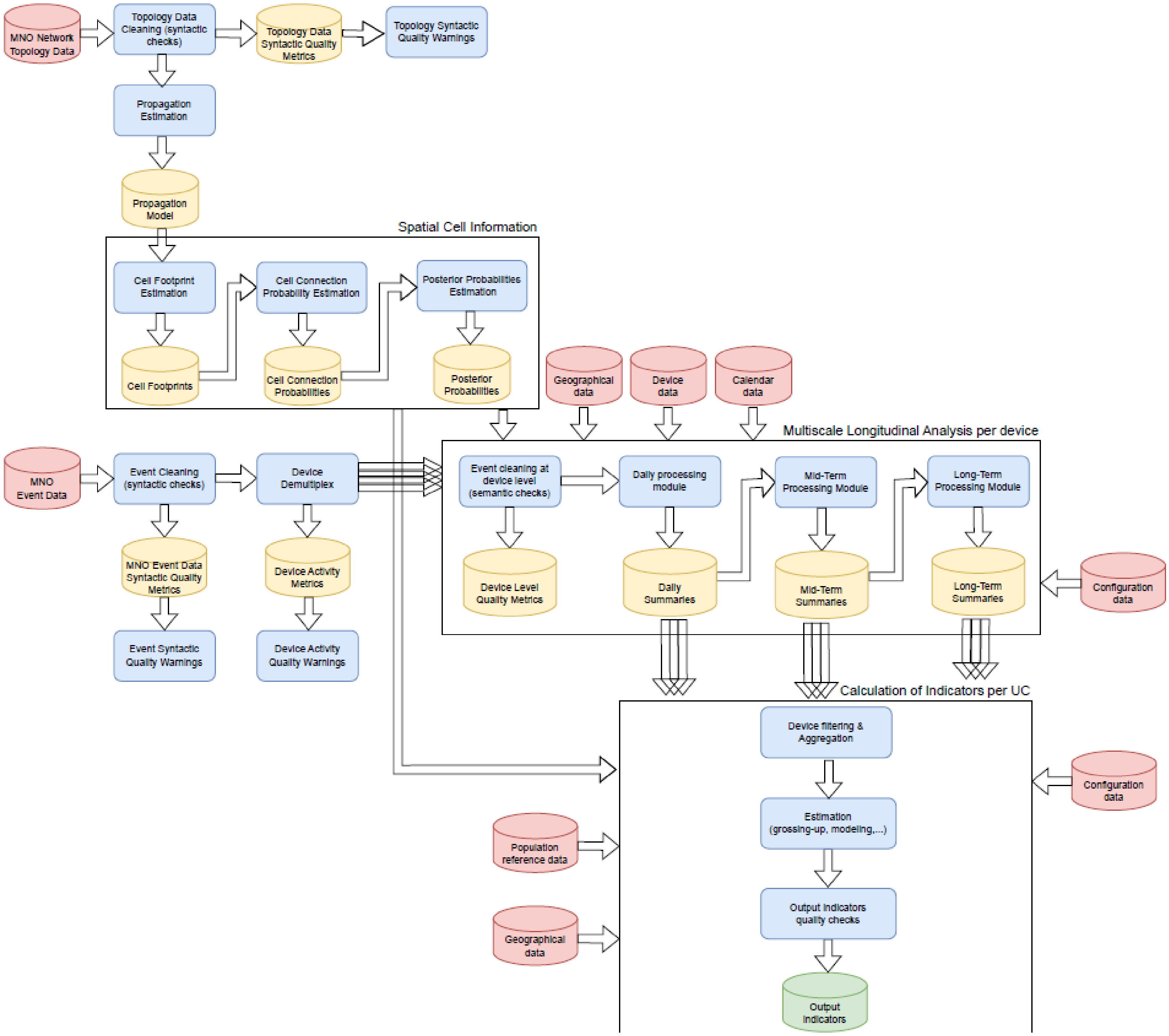

MNO Network Topology Data
Topology Data Cleaning (syntactic checks)
Topology Data Syntactic Quality Metrics
Topology Syntactic Quality Warnings
Propagation Estimation
Propagation Model
Spatial Cell Information
Cell Footprint Estimation
Cell Footprints
Cell Connection Probability Estimation
Cell Connection Probabilities
Posterior Probabilities Estimation
Posterior Probabilities
Geographical data
Device data
Calendar data
Multiscale Longitudinal Analysis per device
MNO Event Data
Event Cleaning (syntactic checks)
Device Demultiplex
MNO Event Data Syntactic Quality Metrics
Device Activity Metrics
Event Syntactic Quality Warnings
Device Activity Quality Warnings
Event cleaning at device level (semantic checks)
Device Level Quality Metrics
Daily processing module
Daily Summaries
Mid-Term Processing Module
Mid-Term Summaries
Long-Term Processing Module
Long-Term Summaries
Configuration data
Calculation of Indicators per UC
Device filtering & Aggregation
Estimation (grossing-up, modeling,...)
Output Indicators quality checks
Output Indicators
Configuration data
Population reference data
Geographical data

Figure 1: High-level view of the pipeline

1.4 OVERVIEW OF METHODS AND DATA OBJECTS

The overall purpose of this volume of deliverable D2.2 is to specify the full set of different methods and data objects included in the pipeline described in Volume I for a selection of four use cases (UCs), from the ones defined in Volume II. The selected UCs are: Present Population Estimation, M-Usual Environment indicators, M-Home Location indicators and Internal Migration. The selection of the UCs is motivated by the different perspective and analyses these allow for. The present population estimation UC builds a snapshot statistical estimation of mobile devices in a specific geographical area at a given moment in time. The M-UE indicators UC offers the possibility, and builds along the concept of, longitudinal analyses across time and space. The M-HL indicators UC provides an MNO-based instantiation of the official statistics concept of resident population (i.e. a measure for the '*de facto*' population concept). The Internal Migration UC produces estimated counts of people changing 'home location' in a given reference time period.

Before introducing the list of methods and related data objects, we clarify the following:

- Each functional block included in the pipeline may include one or several methods and generate one or more data objects.
- For each method, we specify the input data objects, the output data objects, and the data transformation, data fusion and data analysis processes applied to transform the input data objects into the output data objects.
- For each data object, we specify the structure and contents of the data object.

The table below provides an overview of the methods and data objects associated to each functional block of the pipeline and defined in this document.

Table 1: Overview of methods and data objects

FUNCTIONAL BLOCKS (MODULES)	METHODS	INPUT DATA OBJECTS	OUTPUT DATA OBJECTS *(INTERMEDIATE RESULTS OR OUTPUT DATA)*
1. MNO Network Topology Data Cleaning - Syntactic checks	1.1 MNO Network Topology Data Cleaning - Syntactic checks	✓ **MNO Network Topology Data – Raw** *One of the following two data objects should be used:* • Cell Locations with Physical Properties [INPUT] • Cell Footprint with Differentiated Signal Strength Coverage Areas [INPUT]	✓ **Clean MNO Network Topology Data [INTERMEDIATE RESULTS]** *The same data object schema is used for the input MNO Network Topology Data:* • Cell Locations with Physical Properties • Cell Footprint with Differentiated Signal Strength Coverage Areas ✓ **MNO Network Topology Data Quality Metrics [INTERMEDIATE RESULTS]**
2. MNO Network Topology Quality Warnings	2.1 MNO Network Topology Syntactic Quality Warnings	✓ **MNO Network Topology Data Quality Metrics [INTERMEDIATE RESULTS]** ✓ previous period **MNO Network Topology Data Quality Metrics [INTERMEDIATE RESULTS]**	✓ **MNO Network Topology Data Quality Warnings [INTERMEDIATE RESULTS]**
3. Estimation of Signal Strength	3.1 Estimation of Signal Strength from MNO Network Topology Data with Physical Properties	✓ **Cell Locations with Physical Properties [INPUT]**	✓ **Cell Signal Strengths [INTERMEDIATE RESULTS]**
	3.2 Estimation of Signal Strength from MNO Signal Strength Data	✓ **Cell Footprint with Differentiated Signal Strength Coverage Areas [INPUT]**	✓ **Cell Signal Strengths [INTERMEDIATE RESULTS]**
4. Cell Footprint Estimation	4.1 Cell Footprint Estimation	✓ **Cell Signal Strengths [INTERMEDIATE RESULTS]**	✓ **Cell Footprint Values [INTERMEDIATE RESULTS]**
5. Cell Connection Probability Estimation	5.1 Cell Connection Probability Estimation	✓ **Cell Footprint Values [INTERMEDIATE RESULTS]**	✓ **Cell Connection Probabilities [INTERMEDIATE RESULTS]**
6. Posterior Probability Estimation	6.1 Posterior Probability Estimation Module	✓ **Cell Connection Probabilities [INTERMEDIATE RESULTS]** ✓ **Grid Prior Land Use Probabilities [REFERENCE INPUT DATA]**	✓ **Posterior Probabilities Values [INTERMEDIATE RESULTS]**

FUNCTIONAL BLOCKS (MODULES)	METHODS	INPUT DATA OBJECTS	OUTPUT DATA OBJECTS *(INTERMEDIATE RESULTS OR OUTPUT DATA)*
7. MNO Event Data Cleaning - Syntactic checks	7.1 MNO Event Data Cleaning - Syntactic Checks	✓ **MNO Event Data – Raw [INPUT]**	✓ **Clean MNO Event Data [INTERMEDIATE RESULTS]** ✓ **MNO Event Data Syntactic Quality Metrics [INTERMEDIATE RESULTS]**
	7.2 MNO Event General Statistics	✓ **Clean MNO Event Data [INTERMEDIATE RESULTS]**	✓ **General Event Statistics Metrics [INTERMEDIATE RESULTS]**
8. MNO Event Data - Syntactic Quality Warnings	8.1 MNO Event Data Syntactic Quality Warnings	✓ **MNO Event Data Syntactic Quality Metrics [INTERMEDIATE RESULTS]** ✓ previous period **MNO Event Data Syntactic Quality Metrics [INTERMEDIATE RESULTS]**	✓ **MNO Event Data Quality Warnings [INTERMEDIATE RESULTS]**
9. Device Demultiplex	9.1 Device Demultiplex	✓ **Clean MNO Event Data [INTERMEDIATE RESULTS]**	✓ **Event Data at Device level [INTERMEDIATE RESULTS]**
10. Event Cleaning at Device Level - Semantic Checks	10.1 Event Cleaning at Device Level - Semantic Checks	✓ **Event Data at Device level [INTERMEDIATE RESULTS]** ✓ **MNO Network Topology Data – Raw** *One of the following two data objects should be used:* • Cell Locations with Physical Properties [INPUT] • Cell Footprint with Differentiated Signal Strength Coverage Areas [INPUT]	✓ **Semantically Cleaned Event Data at Device Level [INTERMEDIATE RESULTS]** ✓ **Device Semantic Quality Metrics [INTERMEDIATE RESULTS]**
11. Device Activity Statistics	11.1 Device Activity Statistics	✓ **Semantically Cleaned Event Data at Device Level [INTERMEDIATE RESULTS]** ✓ **Clean MNO Network Topology Data [INTERMEDIATE RESULTS]**	✓ **Device Activity Statistics [INTERMEDIATE RESULTS]**
12. MNO Event Data at Device Level - Semantic Quality Warnings	12.1 MNO Event Data at Device Level Semantic Quality Warnings	✓ **Device Semantic Quality Metrics [INTERMEDIATE RESULTS]** ✓ previous period **Device Semantic Quality Metrics [INTERMEDIATE RESULTS]**	✓ **MNO Event Data at Device Level Semantic Quality Warnings [INTERMEDIATE RESULTS]**

FUNCTIONAL BLOCKS (MODULES)	METHODS	INPUT DATA OBJECTS	OUTPUT DATA OBJECTS *(INTERMEDIATE RESULTS OR OUTPUT DATA)*
13. Daily Processing Module	13.1 Present Population Estimation[2]	✓ **Semantically Cleaned Event Data at Device Level [INTERMEDIATE RESULTS]** ✓ **Cell Connection Probabilities [INTERMEDIATE RESULTS]**	**OUTPUT DATA:** ✓ **Presence at a given time – cell level**
	13.2 Daily Permanence Score Estimation	✓ **Semantically Cleaned Event Data at Device Level [INTERMEDIATE RESULTS]** ✓ **Cell Footprint Values [INTERMEDIATE RESULTS]**	✓ **Daily Permanence Score**
	13.3 Continuous Time Segmentation	✓ **MNO Event Data – Raw [INPUT]** ✓ **Cell Footprint Values [INTERMEDIATE RESULTS]**	✓ **Daily Continuous Time Segmentation - Cell Level**
14. Mid-Term Processing Module	14.1 Mid-Term Permanence Analysis	✓ **Daily Permanence Score** ✓ **Calendar Info [REFERENCE INPUT DATA]**	**OUTPUT DATA:** ✓ **Mid-Term Permanence Score per daily, sub-daily and sub-monthly periods** ✓ **Mid-Tern Frequency Count per daily, sub-daily and sub-monthly periods** ✓ **Mid-Term Regularity Indices per daily, sub-daily and sub-monthly periods**
15. Long-Term Processing Module	15.1 Long-Term Permanence Analysis	✓ **Mid-Term Permanence Score per daily, sub-daily and sub-monthly periods / OUTPUT DATA** ✓ **Mid-Term Frequency Count per daily, sub-daily and sub-monthly periods / OUTPUT DATA** ✓ **Mid-Term Regularity Indices per daily, sub-daily and sub-monthly periods / OUTPUT DATA** ✓ **Calendar Info [REFERENCE INPUT DATA]**	**OUTPUT DATA:** ✓ **Long-Term Permanence Score of the period of reference (6 months) per sub-yearly, per sub-monthly and sub-daily periods** ✓ **Long-Term Frequency Count of the period of reference (6 months) per sub-yearly, per sub-monthly and sub-daily periods** ✓ **Long-Term Regularity Indices of the period of reference (6 months) per sub-yearly, per sub-monthly and sub-daily periods**
	15.2 Long-Term Home Location Method		

[2] The method description is presented as a full application of the pipeline to the UC, since it is a static/snapshot analysis. Therefore, the method description includes as well the device filtering and aggregation steps for the preferred variant, i.e. basic counting.

FUNCTIONAL BLOCKS (MODULES)	METHODS	INPUT DATA OBJECTS	OUTPUT DATA OBJECTS *(INTERMEDIATE RESULTS OR OUTPUT DATA)*
			SUB-FUNCTION OUTPUT DATA: ✓ **Usual Environment labels** ✓ **Home location labels** ✓ **Work labels** ✓ **Second home labels *(not to implement)***
16. Device Filtering & Single MNO Data Aggregation	16.1 Device Filtering and Aggregation for Usual Environment	✓ **Long-Term Labels / SUB-FUNCTION OUTPUT DATA** ✓ **Grid Prior Land Use Probabilities [REFERENCE INPUT DATA]**	**OUTPUT DATA:** ✓ **Single-MNO counts per tile representing devices with Usual Environment in that tile** ✓ **Single-MNO counts per tile representing devices with Home Location in that tile** ✓ **Single-MNO counts per tile representing devices with Work Location in that tile** ✓ **Weights value for tiles in the devices sharing**
17. Merge Single MNO Aggregates in Multi-MNO aggregates	17.2 Merge Single MNO Aggregates in Multi-MNO Aggregates for Usual Environment	✓ **Single-MNO Aggregates / OUTPUT DATA** ✓ **Geographical Areas (e.g. administrative units)** for which the deduplication factors are provided and their correspondence with the reference processing grid	**OUTPUT DATA FOR THE DEVICE:** ✓ **Multi-MNO aggregated number of devices per tile and Usual Environment, Home and Work labels (counts of deduplicated devices having Usual Environment, Home Location and Work Location label in the tile)**
18. Projection of Multi-MNO Aggregates from the finest level to the geographic unit systems relevant for the use case	18.1 Projection/Mapping Multi-MNO Counts to relevant output geography for the Usual Environment	✓ **Multi-MNO Aggregation for Usual Environment / OUTPUT DATA FOR THE DEVICE** ✓ **Geographical system used for the reference grid tiles / default INSPIRE grid with LAEA projection** ✓ **Grid Prior Land Use Probabilities [REFERENCE INPUT DATA]**	**OUTPUT DATA:** ✓ **Number of devices per Usual Environment Labelled geographical areas (Multi-MNO aggregated counts per Usual Environment label at the geographical area level)** ✓ **The factor values (for quality measures)**
19. Estimation	19.1 Estimation	✓ **Multi-MNO aggregates at the relevant geographical system** ✓ **Input geographical system**	**OUTPUT DATA:** ✓ **Weighted counts at the geographical level required by the output**

FUNCTIONAL BLOCKS (MODULES)	METHODS	INPUT DATA OBJECTS	OUTPUT DATA OBJECTS *(INTERMEDIATE RESULTS OR OUTPUT DATA)*
		✓ **Output geographical areas** ✓ **Eventually other additional external sources**	
20. Home Location changes detection method	20.1 Home Location changes detection method	✓ **Long-term Home Location labels (groups of tiles representing the HL of the individual device in the analysis period) / SUB-FUNCTION OUTPUT DATA** ✓ **Administrative areas of interest (defining the spatial resolution of the use case output)**	**OUTPUT DATA FOR THE DEVICE:** ✓ **Migration table for the individual device**

Note: In this table and in the methods description below, the term UE is used to indicate the intermediate data objects resulting from the processing of the data, since we don't reach the final stage of building the indicators from the M-Usual Environment indicators UC.

Methods for the quality evaluation of the output indicators will be developed for the final version of this document (deliverable D2.3), in coordination with the work under the project's Task 3 – Quality Framework and Business Process Model. Is to be clarified that SDC methods for dissemination are not covered by the project. Nevertheless, the SDC method developed for the testing on real data during the project is included in this document as Annex I – SDC for demonstrator scenario/Disclosure.

In the remainder of this document, we introduce and discuss the methods proposed one by one. The method that implements the estimation of the present devices in a specific area at a fixed time (i.e. the present population use case) is described in detail in Section 14.1 Method 1: Present Population Estimation (as the first method for the Module 13: Daily Processing Module). In this section, the method is described to reflect the pipeline application to the use case, since it doesn't require/perform further longitudinal analyses. For the M-Usual Environment indicators use case, which is the second one covered in this document, the description of the corresponding methods is introduced linearly (as they shall be executed along the pipeline) starting with Section 14.2 Method 2: Daily Permanence Score Estimation. In addition, the exemplification of the pipeline's application to the use case is introduced in Annex III - Pipeline application to the Usual Environment use case. Methods developed for the M-Home Location indicators are described starting from Section 16.2 Long-Term Home Location Method. The Home Location actually represents a specific labelling of the long-term Usual Environment module with its subsequent elaborations, inheriting the previous pipeline methods from the Usual Environment elaborations. In what concerns the-Internal Migration indicators, these rely on the M-Home Location labelling, as already anticipated. Hence, coherently the corresponding method for home location changes detection takes as input the long-term Home Location module, which in turn relies on Usual Environment for the previous methods. To preserve the logic of the Internal Migration indicators' production, its component modules are reported all together at the end of the pipeline methods' description, in Section 21 Home Location Changes Detection Methods. At the same time, to facilitate the recognition of the functional steps of the pipeline for this specific UC, the corresponding methods are indicated in the heading of its subparagraphs.

For the description of each method, we follow a standard format which introduces the following details:

- Objective of the method
- List of parameters
- Input data
- Output data
- Methodology
- Pseudo-code / check if note to be added in case is missing in some of the methods
- Examples (i.e. toy examples of the application of the method to facilitate its interpretation)

The methods are grouped in chapters by functional blocks/modules, as in the classification followed in the Overview of methods and data objects in the table above.

2 MODULE/METHOD 1: MNO NETWORK TOPOLOGY DATA CLEANING – SYNTACTIC CHECKS

2.1 OBJECTIVE

This module is responsible for performing syntactic checks on Network Topology Data, in order to remove erroneous entries and to produce corresponding syntactic quality metrics. It can include simple transformations to harmonise data formats, if needed. However, we expect that the input data follows the requirements defined in MNO Network Topology Data – Raw (one of the options: Cell Locations with Physical Properties [INPUT] or Cell Footprint with Differentiated Signal Strength Coverage Areas [INPUT].

Syntactic checking is the first step in data processing to ensure that data is suitable for further operations. The main task is ensuring that all required fields are present and their values follow the specified conditions. The syntactic checks module can also include checks of the general dataset's properties specified by parameters. Syntactic checks are performed on a single-entry level; there is no cross-referencing to other data collections and there are no aggregation operations, except for the computation of quality metrics. Such semantic checks are performed in subsequent modules of the pipeline.

Mapping with other standards*: the syntactic checks defined in this method can be mapped with the ones included in the sub-process 5.3 Review and Validate of GSBPM and to Validation levels 0 and 1 as defined in the ESS handbook Methodology for data validation 1.1 (Revised edition 2018)*[3]*.*

Correspondence with Deliverable D3.1*: in Table 2 of Chapter 6, this method corresponds to Quality Module QM.TopSynt.*

2.2 PARAMETERS

- **bounding_box**: Min and max values in WGS84 indicating expected country/area bounds where the cell must be located in (e.g.`'longitude_min'=-180,'longitude_max'=180,'latitude_min'=-90,'latitude_max'=90`). The geometry of the country could be substituted to the bounding box for future developments if computationally feasible.
- **technology_options**: accepted values in the technology field. Other values will be treated as out of bounds/range. E.g. `5G, LTE, UMTS, GSM`.
- **data_period_start** and **data_period_end**: the start and end date (included) for which data is to be processed.
- **timestamp_format**: the timestamp format that is expected to be in the input network data and that will be parsed. E.g. `yyyy-MM-dd'T'HH:mm:ss`
- **cell_type_options**: accepted values in the `cell_type` field. Other values will be treated as out of bounds/range. E.g. `macrocell, microcell, picocell`.

[3] Available at: https://cros-legacy.ec.europa.eu/system/files/ess_handbook_-_methodology_for_data_validation_v1.1_-_rev2018_0.pdf

- **k**: absolute value or percentage used to select the top k invalid values found during the syntactic cleaning process. This value can be absolute, such as k=10, which would result in the saving of the top 10 most frequent invalid values found (or top most frequent invalid values found that cover k% of all invalid errors, starting from the most frequent ones).
- **frequent_error_criterion** which can take two values:
  - `absolute` if one wants to find the top *k* most frequent errors (e.g. k=10); or
  - `percentage` if one wants to find the most frequent errors that represent k percentage of all errors found.

2.3 INPUT DATA

- MNO Network Topology Data - Raw [INPUT]. One of these two data objects should be used:
  - Cell Locations with Physical Properties [INPUT]
  - Cell Footprint with Differentiated Signal Strength Coverage Areas [INPUT]

Our preferred option is to use Cell Footprint with Differentiated Signal Strength Coverage Areas [INPUT]*, whenever available. If both are available we could optionally use Cell Locations with Physical Properties [INPUT] for additional quality checks (for example to compare cell footprint results).*

2.4 OUTPUT DATA

- Clean MNO Network Topology Data [INTERMEDIATE RESULTS] following the same data object format as for the input Network Topology Data:
  - Cell Locations with Physical Properties [INPUT]
  - Cell Footprint with Differentiated Signal Strength Coverage Areas [INPUT]
- MNO Network Topology Data Quality Metrics [INTERMEDIATE RESULTS]

2.5 METHODOLOGY

We aim to apply a series of checks and transformations to ensure that the MNO Network Topology Data is converted into a format that is suitable for subsequent processes. Whenever an entry fails a check or cannot be parsed or converted, it is marked as an error entry and is removed from the main dataset.

The total number of input entries should be calculated as a quality metric.

For each entry, the following checks should be performed:

- **cell_id** is parsed as numeric CGI or eCGI (see https://arimas.com/2016/10/24/cgi-ecgi/)
  - The value is not null
  - The value should be a 14-digit or 15-digit numeric code
- **valid_date_start** is parsed as a timestamp (start of validity range)
  - The value is not null
  - If input is in date or string format, attempt to convert it into timestamp.
  - Check if timestamp is coherent with the **data_period_start** and **data_period_end**
- **valid_date_end** is parsed as a timestamp (end of validity range)
  - Value can be null if end date is not known
  - If input is in date or string format, attempt to convert it into timestamp.
  - Check if timestamp is coherent with the **data_period_start** and **data_period_end**
  - Set timezone as **input_time_zone** if timezone is not included in the timestamp value.
- If the input is Cell Footprint with Differentiated Signal Strength Coverage Areas [INPUT]:
  - **geometry** is parsed as geometry type

        - The value is not null
        - Check if the data type is WKT
    - Check if **geometry** is within **bounding_box**
        - Check if geometry is within bounding box
    - **signal strength** is not null and is in the interval (0,1]
- If the input is Cell Locations with Physical Properties [INPUT]:
    - **latitude** and **longitude** are parsed as WGS84 coordinates
        - The values are not null
    - Check if **latitude and longitude** are within country **bounding_box** (or to geometry of the Country for future developments)
        - The values are between corresponding bounding box min and max values
    - For each of the following fields, check if value is present, if value is in correct format, if value is within the set of admissible values.
        - **altitude**
        - **antenna_height**
        - **directionality**
        - **azimuth_angle**
        - **elevation_angle**
        - **power**
        - **horizontal_beam_width**
        - **vertical_beam_width**
        - **frequency**
        - **technology**
        - **cell_type**

Statistics about the number and type of error entries are collected between process steps to determine the number of entries removed in each step of the process. The statistics are output following the format of MNO Network Topology Data Quality Metrics [INTERMEDIATE RESULTS].

In addition, a metric reporting the *k* most frequent errors is produced, where *k* is a number chosen by the user which can be either an absolute value or a percentage of all errors. Such metric allows the observation of recurring errors in a dataset but also the identification of potential common patterns and characteristics in the errors.

Entries with missing or invalid mandatory fields are discarded. Invalid values on optional fields are counted in quality metrics and set to null.

The main Clean MNO Network Topology Data output contains entries which are free of syntactic errors. The columns present are the same as are defined in the input data object. Column types should have been converted to types that are suitable for subsequent processing in the pipeline.

3 MODULE/METHOD 2: MNO NETWORK TOPOLOGY - QUALITY WARNINGS

3.1 OBJECTIVE

This module analyses the MNO Network Topology Data Quality Metrics [INTERMEDIATE RESULTS] produced by the Module/Method I: MNO Network Topology Data Cleaning – Syntactic Checks to identify anomalous situations that need to be further investigated. The output of the method is the MNO Network Topology Data Quality Warnings [INTERMEDIATE RESULTS] which displays: plots of the metrics over time, the anomalous data, the related warning and (whenever possible) suggestions on additional investigations.

Quality warnings from syntactic checks are linked to the following characteristics or anomalies of the input data MNO Network Topology Data - Raw:

- Size
- Missing values
- Wrong data types or formats
- Out of range values: format is correct but value is not acceptable
- Transformations performed to standardise the input

Anomalous quality metrics values should always launch a warning; nonetheless, not necessarily each warning has to correspond to an error. A warning means that something is to be checked, even if afterwards it is understood that the anomaly was not an error. In most cases, warnings imply the definition of thresholds above or under which the warning will be launched. The thresholds can be identified in different ways. In general, the method will allow to set thresholds as input parameters, but will include also a default value (e.g. 90%). The default threshold could be adjusted during testing.

Quality warnings are supposed to be executed daily.

Mapping with other standards: *the quality warnings defined in this method can be mapped to those included in the Overarching process Quality management of GSBPM.*

Correspondence with Deliverable D3.1: *in Table 2 of Chapter 6, this method corresponds to Quality Module QM.TopSynt.*

3.2 PARAMETERS

- **thresholds**: values above or under which the method launches the warning. They can be:
  i. specific values to which the metric is compared directly (e.g. the error rate is above 30%), or
  ii. defined in relative terms based on the average of the metrics in previous periods (e.g. the missing value rate for the variable cell_id in the data object is 50% higher than the average missing value rate for the variable cell_id in the last month data objects), or

iii. dynamically defined based also on the variability of the metrics in previous period (e.g. the size of the data object is smaller than the average size of the last week data objects – 2 times the standard deviations).

The **thresholds** defined are:

  - SIZE_RAW_DATA_OVER_AVERAGE: “X%” above the average
  - SIZE_RAW_DATA_UNDER_AVERAGE: “X%” under the average
  - SIZE_RAW_DATA_VARIABILITY: “X” times the standard deviation of raw data size to calculate upper and lower control limits
  - SIZE_RAW_DATA_ABS_VALUE_UPPER_LIMIT
  - SIZE_RAW_DATA_ABS_VALUE_LOWER_LIMIT
  - SIZE_CLEAN_DATA_OVER_AVERAGE: “X%” above the average
  - SIZE_CLEAN_DATA_UNDER_AVERAGE: “X%” under the average
  - SIZE_CLEAN_DATA_VARIABILITY: “X” times the standard deviation of raw data size to calculate upper and lower control limits
  - SIZE_CLEAN_DATA_ABS_VALUE_UPPER_LIMIT
  - SIZE_CLEAN_DATA_ABS_VALUE_LOWER_LIMIT
  - TOTAL_ERROR_RATE_OVER_AVERAGE: “X%” above the average
  - TOTAL_ERROR_RATE_VARIABILITY: “X” times the standard deviation of raw data size to calculate upper control limit
  - TOTAL_ERROR_RATE_ABS_VALUE_UPPER_LIMIT
  - Missing_value_RATE_field_name_AVERAGE: “X%” above the average
  - Missing_value_RATE_field_name_VARIABILITY: “X” times the standard deviation of raw data size to calculate upper control limit
  - Missing_value_RATE_field_name_ABS_VALUE_UPPER_LIMIT
  - Wrongtype/Format_RATE_field_name_AVERAGE: “X%” over the average
  - Wrongtype/Format_RATE_field_name_VARIABILITY: “X” times the standard deviation of raw data size to calculate upper control limit
  - Wrongtype/Format_RATE_field_name_ABS_VALUE_UPPER_LIMIT
  - Out_of_range_RATE_field_name_AVERAGE: “X%” above the average
  - Out_of_range_RATE_field_name_VARIABILITY: “X” times the standard deviation of raw data size to calculate upper control limit
  - Out_of_range_RATE_field_name_ABS_VALUE_UPPER_LIMIT
  - Parsing_error_RATE_field_name_AVERAGE: “X%” above the average
  - Parsing_error_RATE_field_name_VARIABILITY: “X” times the standard deviation of raw data size to calculate upper control limit
  - Parsing_error_RATE_field_name_ABS_VALUE_UPPER_LIMIT

- **period**: previous period for which the plot is created and the average and the standard deviations of the quality metric(s) are calculated (it can be week, month, quarter)

3.3 INPUT DATA

- MNO Network Topology Data Quality Metrics [INTERMEDIATE RESULTS]
- previous period MNO Network Topology Data Quality Metrics [INTERMEDIATE RESULTS]

3.4 OUTPUT DATA

- MNO Network Topology Data Quality Warnings [INTERMEDIATE RESULTS] for syntactic checks

3.5 METHODOLOGY

We list here the different methods to be applied to produce the warnings. This is to be considered as a first set of warnings that should be possible to adjust during testing and integrated further afterwards (and, similarly, for each module of the pipeline).

3.5.1 SIZE OF THE MNO NETWORK TOPOLOGY DATA - RAW [INPUT]

It would be useful to produce a plot with the time on horizontal axis and the value of the metric 'Total number of rows at the start of the method' on the vertical axis for the previous **period** selected as parameter (week, month, quarter).

For the warnings, the current day value should be compared with the thresholds. To this aim, the average and the standard deviation (SD) of the same metric over the previous period should be calculated. Afterwards,, the warnings will be launched according to the following indications:

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Value of the size of the raw data object	Size differs (greater or smaller) from the previous period average by X%.	X% (default e.g. 30%)	The number of cells is unexpectedly low/high compared to the previous period, please check if there have been issues in the network.
Value of the size of the raw data object	Size is out of control limits calculated on the basis of average and standard deviation of the distribution of the size in the previous period. Control limits = (average ± X·SD)	X (default e.g. 2)	The number of cells is unexpectedly low/high compared to the previous period, taking into account the usual variability of the cell numbers, please check if there have been issues in the network.
Value of the size of the raw data object	The size is under/above thresholds X and Y.	X = the default could be set equal to the previous lower control limit Y = the default could be set equal to the previous upper control limit	The number of cells is above/under the threshold, please check if there have been changes in the network.

3.5.2 SIZE OF THE CLEAN MNO NETWORK TOPOLOGY DATA [INTERMEDIATE RESULTS]

It would be useful to produce a plot with the time on horizontal axis and the value of the metric "Total number of rows at the end of the method" on the vertical axis for the previous **period** selected as parameter (week, month, quarter).

For the warnings, the current day value should be compared with the thresholds. To this aim, the average and the standard deviation (SD) of the same metrics over the previous period should be calculated. Afterwards, the warnings will be launched according to the following indications:

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Value of the size of the clean data object	Size differs (greater or smaller) from the previous period average by X%.	X% (default e.g. 30%)	The number of cells after syntactic checks application is unexpectedly low/high compared to the previous period.
Value of the size of the clean data object	Size is out of control limits calculated on the basis of average and standard deviation of the distribution of the size in the previous period. Control limits = (average ± X·SD)	X (default e.g. 2)	The number of cells is unexpectedly low/high compared to the previous period, taking into account the usual variability of the cell numbers, please check if there have been issues in the network.
Value of the size of the clean data object	The size is under/above thresholds X and Y.	X = Y =	The number of cells after syntactic checks application is above/under the threshold.

3.5.3 ERROR RATE

Error rate = (Total number of rows at the start of the method – Total number of rows at the end of the method) / Total number of rows at the start of the method*100

Produce a plot with the time on horizontal axis and the rate on the vertical axis for the previous **period** selected as parameter (week, month, quarter).

For the warnings, the current day value should be calculated and compared with the thresholds. To this aim, the average and the standard deviation (SD) of the same metrics over the previous period should be calculated. Afterwards, the warnings will be launched according to the following indications for each mandatory field:

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Error rate	Error rate is above the previous period average by X%.	X% (default e.g. 30%)	The error rate after syntactic checks application is unexpectedly high compared to the previous period.
Error rate	Error rate is above the upper control limit calculated on the basis of average and standard deviation of the distribution of the error rate in the previous period. Upper Control limit = (average + X·SD)	X (default e.g. 2)	The error rate after syntactic checks application is unexpectedly high compared to the previous period, taking into account its usual variability.
Error rate	The error rate is above the value X.	X% (default e.g. 20%)	The error rate after syntactic checks application is above the threshold.

The following warnings apply if the input data object is Cell Footprint with Differentiated Signal Strength Coverage Areas [INPUT]

3.5.4 WARNING FOR MISSING VALUE RATE FOR EACH MANDATORY FIELD (CELL_ID, VALID_DATE_START, VALID_DATE_END, SIGNAL_STRENGTH, GEOMETRY)

Missing value rate of *field_name* = Number of missing value for fieldname / Total number of rows at the start of the method*100.

For the warnings, the current day value should be calculated and compared with the thresholds. To this aim, the average and the standard deviation (SD) of the same metrics over the previous period should be calculated. Afterwards, the warnings will be launched according to the following indications for each mandatory field:

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Missing value rate of *field_name*	Missing value rate of *field_name* is above the previous period average by X%.	X% (default e.g. 30%)	The missing value rate of *field_name* after syntactic check application is unexpectedly high compared to the previous period.
Missing value rate of *field_name*	Missing value rate of *field_name* is above the upper control limit calculated on the basis of average and standard deviation of the distribution of the Missing value rate of *field_name* in the previous period. Upper Control limit = (average + X·SD)	X (default e.g. 2)	The missing value rate of *field_name* after syntactic check application is unexpectedly high compared to the previous period, taking into account its usual variability.
Missing value rate of *field_name*	Missing value rate of *field_name* is above the value X	X% (default e.g. 20%)	The missing value rate of *field_name* after syntactic checks application is above the threshold.

3.5.5 WARNING FOR ERRORS IN DATA TYPES/FORMAT FOR EACH MANDATORY FIELD (CELL_ID, VALID_DATE_START, VALID_DATE_END, SIGNAL_STRENGTH, GEOMETRY)

Wrong type/format rate of *field_name* = Number of values with errors in data types/format for *field_name/* Total number of rows at the start of the method*100.

For the warnings, the current daily value should be calculated and compared with the thresholds. To this aim the average and the standard deviation (SD) of the same metrics over the previous period should be calculated. Afterwards, the warnings will be launched according to the following indications for each mandatory field:

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Wrong type/format rate of *field_name*	Wrong type/format rate of *field_name* is above the previous period average by X%.	X% (default e.g. 30%)	The wrong type/format rate of *field_name* after syntactic check application is unexpectedly high compared to the previous period.
Wrong type/format rate of *field_name*	Wrong type/format rate of *field_name* is above the upper control limit calculated on the basis of average and standard deviation of the distribution of the Wrong type/format rate of *field_name* in the previous period. Upper Control limit = (average + X·SD)	X (default e.g. 2)	The wrong type/format rate of *field_name* after syntactic check application is unexpectedly high compared to the previous period, taking into account its usual variability.
Wrong type/format rate of *field_name*	The error rate is above the value X.	X% (default e.g. 20%)	The wrong type/format rate of *field_name* after syntactic checks application is above the threshold.

3.5.6 WARNING FOR OUT OF RANGE RATE (VALUE IS NOT WITHIN THE SET OF ACCEPTED VALUES) FOR EACH RELEVANT MANDATORY FIELD (VALID_DATE_START, VALID_DATE_END, SIGNAL_STRENGTH, GEOMETRY)

Out of range rate of *field_name* = Number of out of range values for *field_name* / Total number of rows at the start of the method*100.

For the warnings, the current day value should be calculated and compared with the thresholds. To this aim, the average and the standard deviation (SD) of the same metrics over the previous period should be calculated. Afterwards, the warnings will be launched according to the following indications for each mandatory field:

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Out of range rate of *field_name*	Out of range rate of *field_name* is above the previous period average by X%.	X% (default e.g. 30%)	The out of range rate of *field_name* after syntactic check application is unexpectedly high compared to the previous period
Out of range rate of *field_name*	Out of range rate of *field_name* is above the upper control limit calculated on the basis of average and standard deviation of the distribution of the Out of range rate of *field_name* in the previous period. Upper Control limit = (average + X·SD)	X (default e.g. 2)	The out of range rate of *field_name* after syntactic check application is unexpectedly high compared to the previous period taking into account its usual variability.
Out of range rate of *field_name*	Out of range rate of *field_name* is above the value X.	X% (default e.g. 20%)	The out of range rate of *field_name* after syntactic checks application is above the threshold.

3.5.7 WARNING FOR PARSING ERROR RATE FOR EACH MANDATORY FIELD SUBJECTED TO A PARSING PROCEDURE (VALID_DATE_START, VALID_DATE_END, GEOMETRY)

Parsing error rate of *field_name* = Number of values with parsing errors for *field_name* / Total number of rows at the start of the method*100.

For the warnings, the current day value should be calculated and compared with the thresholds. To this aim, the average and the standard deviation (SD) of the same metrics over the previous period should be calculated. Afterwards, the warnings will be launched according to the following indications for each mandatory field:

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Parsing error rate of *field_name*	Parsing error rate of *field_name* is above the previous period average by X%	X% (default e.g. 30%)	The parsing error rate of *field_name* after syntactic check application is unexpectedly high with respect to the previous period
Parsing error rate of *field_name*	Parsing error rate of *field_name* is over the upper control limit calculated on the basis of average and standard deviation of the distribution of the Parsing error rate of *field_name* in previous period. Upper Control limit = (average + X·SD)	X (default e.g. 2)	The parsing error rate of *field_name* after syntactic check application is unexpectedly high with respect to the previous period taking into account its usual variability.
Parsing error rate of *field_name*	Parsing error rate of *field_name* is above the value X.	X% (default e.g. 20%)	The parsing error rate of *field_name* after syntactic checks application is above the threshold.

3.5.8 COMPOSITION OF ERRORS FOR EACH MANDATORY FIELD (CELL_ID, VALID_DATE_START, VALID_DATE_END, SIGNAL_STRENGTH, GEOMETRY)

For each mandatory field, calculate the % of each type of error (missing value, wrong/format, out of range, parsing errors) on the total of errors for the same field. Build a pie chart with this % for each variable.

The following methods apply if the input data object is Cell Locations with Physical Properties [INPUT]

3.5.9 WARNING FOR MISSING VALUE RATE FOR EACH MANDATORY FIELD (CELL_ID, VALID_DATE_START, VALID_DATE_END, LATITUDE AND LONGITUDE) AND FOR EACH OPTIONAL FIELD (ANTENNA_HEIGHT, DIRECTIONALITY, AZIMUTH_ANGLE, ELEVATION_ANGLE, POWER, FREQUENCY, HORIZONTAL BEAM WIDTH, VERTICAL BEAM WIDTH, TECHNOLOGY, CELL_TYPE)

Missing value rate of *field_name* = Number of missing value for *field_name* / Total number of rows at the start of the method*100.

For the warnings, the current day value should be calculated and compared with the thresholds. To this aim, the average and the standard deviation (SD) of the same metrics over the previous period should be calculated. Afterwards, the warnings will be launched according to the following indications for each mandatory (and possibly for each optional) field:

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Missing value rate of *field_name*	Missing value rate of *field_name* is above the previous period average by X%.	X% (default e.g. 30% for mandatory fields, e.g.60% for optional fields)	The missing value rate of *field_name* after syntactic check application is unexpectedly high compared to the previous period.
Missing value rate of *field_name*	Missing value rate of *field_name* is above the upper control limit calculated on the basis of average and standard deviation of the distribution of the Missing value rate of *field_name* in the previous period. Upper Control limit = (average + X·SD)	X (default e.g. 2 for mandatory fields, e.g. 3 for optional fields)	The missing value rate of *field_name* after syntactic check application is unexpectedly high compared to the previous period taking into account its usual variability.
Missing value rate of *field_name*	Missing value rate of *field_name* is above the value X.	X% (default e.g. 20% for mandatory fileds, e.g. 50% for optional fields)	The missing value rate of *field_name* after syntactic checks application is above the threshold.

3.5.10 WARNING FOR ERRORS IN DATA TYPES/FORMAT FOR EACH MANDATORY FIELD (CELL_ID, VALID_DATE_START, VALID_DATE_END, LATITUDE AND LONGITUDE) AND FOR EACH OPTIONAL FIELD (ANTENNA_HEIGHT, DIRECTIONALITY, AZIMUTH_ANGLE, ELEVATION_ANGLE, POWER, FREQUENCY, HORIZONTAL BEAM WIDTH, VERTICAL BEAM WIDTH, TECHNOLOGY, CELL_TYPE)

Wrong type/format rate of *field_name* = Number of values with errors in data types/format for *field_name*/ Total number of rows at the start of the method*100.

For the warnings, the current day value should be calculated and compared with the thresholds. To this aim, the average and the standard deviation (SD) of the same metrics over the previous period should be calculated. Afterwards, the warnings will be launched according to the following indications for each mandatory (and possibly for each optional) field:

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Wrong type/format rate of *field_name*	Wrong type/format rate of *field_name* is above the previous period average by X%.	X% (default e.g. 30% for mandatory fields, e.g.60% for optional fields)	The wrong type/format rate of *field_name* after syntactic check application is unexpectedly high compared to the previous period.
Wrong type/format rate of *field_name*	Wrong type/format rate of *field_name* is above the upper control limit calculated on the basis of average and standard deviation of the distribution of the Wrong type/format rate of *field_name* in the previous period. Upper Control limit = (average + X·SD)	X (default e.g. 2 for mandatory fields, e.g. 3 for optional fields)	The wrong type/format rate of *field_name* after syntactic check application is unexpectedly high compared to the previous period, taking into account its usual variability.
Wrong type/format rate of *field_name*	The error rate is above the value X.	X% (default e.g. 20% for mandatory fileds, e.g. 50% for optional fields)	The wrong type/format rate of *field_name* after syntactic checks application is above the threshold.

3.5.11 WARNING FOR OUT OF RANGE RATE (VALUE IS NOT WITHIN THE SET OF ACCEPTED VALUES) FOR EACH RELEVANT MANDATORY FIELD (VALIDITY_PERIOD_START, VALIDITY_PERIOD_END, LATITUDE AND LONGITUDE) AND FOR EACH OPTIONAL FIELD (ANTENNA_HEIGHT, DIRECTIONALITY, AZIMUTH_ANGLE, ELEVATION_ANGLE, POWER, FREQUENCY, HORIZONTAL BEAM WIDTH, VERTICAL BEAM WIDTH, TECHNOLOGY, CELL_TYPE)

Out of range rate of *field_name* = Number of out of range values for *field_name* / Total number of rows at the start of the method*100.

For the warnings, the current day value should be calculated and compared with the thresholds. To this aim, the average and the standard deviation (SD) of the same metrics over the previous period should be calculated. Afterwards, the warnings will be launched according to the following indications for each mandatory (and possibly for each optional) field:

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Out of range rate of *field_name*	Out of range rate of *field_name* is above the previous period average by X%.	X% (default e.g. 30% for mandatory fields, e.g.60% for optional fields)	The out of range rate of *field_name* after syntactic check application is unexpectedly high compared to the previous period.
Out of range rate of *field_name*	Out of range rate of *field_name* is above the upper control limit calculated on the basis of average and standard deviation of the distribution of the Out of range rate of *field_name* in the previous period. Upper Control limit = (average + X·SD)	X (default e.g. 2 for mandatory fields, e.g. 3 for optional fields)	The out of range rate of *field_name* after syntactic check application is unexpectedly high compared to the previous period taking into account its usual variability.
Out of range rate of *field_name*	Out of range rate of *field_name* is above the value X.	X% (default e.g. 20% for mandatory fileds, e.g. 50% for optional fields)	The out of range rate of *field_name* after syntactic checks application is above the threshold.

3.5.12 WARNING FOR PARSING ERROR RATE FOR EACH MANDATORY FIELD SUBJECTED TO A PARSING PROCEDURE (VALIDITY_PERIOD_START, VALIDITY_PERIOD_END, LATITUDE AND LONGITUDE)

Parsing error rate of *field_name* = Number of values with parsing errors for *field_name* / Total number of rows at the start of the method*100.

For the warnings, the current day value should be calculated and compared with the thresholds. To this aim, the average and the standard deviation (SD) of the same metrics over the previous period should be calculated. Afterwards, the warnings will be launched according to the following indications for each mandatory field:

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Parsing error rate of *field_name*	Parsing error rate of *field_name* is above the previous period average by X%.	X% (default e.g. 30% for mandatory fields, e.g.60% for optional fields)	The parsing error rate of *field_name* after syntactic check application is unexpectedly high compared to the previous period.
Parsing error rate of *field_name*	Parsing error rate of *field_name* is above the upper control limit calculated on the basis of average and standard deviation of the distribution of the Parsing error rate of *field_name* in the previous period. Upper Control limit = (average + X·SD)	X (default e.g. 2 for mandatory fields, e.g. 3 for optional fields)	The parsing error rate of *field_name* after syntactic check application is unexpectedly high compared to the previous period taking into account its usual variability.
Parsing error rate of *field_name*	Parsing error rate of *field_name* is above the value X.	X% (default e.g. 20% for mandatory fileds, e.g. 50% for optional fields)	The parsing error rate of *field_name* after syntactic checks application is above the threshold.

3.5.13 COMPOSITION OF ERRORS FOR EACH MANDATORY FIELD (CELL_ID, VALIDITY_PERIOD_START, VALIDITY_PERIOD_END, LATITUDE AND LONGITUDE) AND FOR EACH OPTIONAL FIELD (ANTENNA_HEIGHT, DIRECTIONALITY, AZIMUTH_ANGLE, ELEVATION_ANGLE, POWER, FREQUENCY, HORIZONTAL BEAM WIDTH, VERTICAL BEAM WIDTH, TECHNOLOGY, CELL_TYPE)

For each field, calculate the % of each type of error (missing value, wrong/format, out of range, parsing errors) on the total of errors for the same field. Build a pie chart with this % for each variable.

4 MODULE 3: ESTIMATION OF SIGNAL STRENGTH

If the MNO provides cell location with network physical properties, the radio propagation model will have to be calculated resulting in signal strength values. If the MNO provides cell footprint including signal strength values with geographical coverage areas (polygons of grids), then the footprint should be transformed into the grid used in the software (as it is probably different from the grid that MNO uses for footprint calculation).

The proposed method is based on Tennekes and Gootzen (2021)[4].

4.1 METHOD 1: ESTIMATION OF SIGNAL STRENGTH FROM MNO NETWORK TOPOLOGY DATA WITH PHYSICAL PROPERTIES

4.1.1 OBJECTIVE

Generate signal strength values for each cell based on the available MNO Network Topology Data.

4.1.2 PARAMETERS

- Radio propagation model parameters[5]

4.1.3 INPUT DATA

- Cell Locations with Physical Properties [INPUT]

4.1.4 OUTPUT DATA

- Cell Signal Strengths [INTERMEDIATE RESULTS]

4.1.5 METHODOLOGY

This section describes the estimation of the **propagation of signal strength originating from a single cell**. We distinguish two types of cells: omnidirectional and directional, resulting in two different propagation models. Omnidirectional cells have no aimed beam and their coverage area can be thought of as a circular disk. Directional cells point in a certain direction and their coverage area can be thought of as an oval with one axis of symmetry.

[4] Tennekes, M., Gootzen, Y.A.P.M., Shah, S.H. A Bayesian approach to location estimation of mobile devices from mobile network operator data. CBDS Working Paper 06-20, available at: https://www.cbs.nl/-/media/_pdf/2020/22/cbds_working_paper_location_estimation.pdf

[5] These refer mainly to default physical properties values (such as power, tilt, etc.), in case these are missing from the input data. There are few other important parameters, such as path loss exponent, etc. The complete list of parameters implemented are available in the Deliverable D4.2 – second software release and documentation.

In practice, small cells are omnidirectional and normal cells (i.e. attached to cell towers or placed on rooftops) are directional.

\ OMNIDIRECTIONAL CELLS

For omnidirectional cells, propagation of the signal strength (g, a) is modelled as

$$S(g,a) := S_0 - S_{\text{dist}}(r_{g,a})$$

where ‘S0’ is the signal strength at r0 = 1 meter distance from the cell in dBm and ‘rg,a’ is the distance between the center of grid tile ‘g' and cell ‘a’ in meters (we take into account the placement height of the cell, but assume that devices are situated at ground level). The value of ‘S0’ can be different for every cell and is assumed to be a known property. In cell plan information, it is common to list the power P of a cell in Watt, rather than the signal strength in dBm. The value of 'S0' can be calculated from P using the conversion between Watt and dBm:

$$S_0 = 30 + 10\log_{10}(P)$$

The function Sdist(r) returns the loss of signal strength as a function of distance r**:**

$$S_{\text{dist}}(r) := 10\log_{10}(r^{\gamma}) = 10\gamma\log_{10}(r)$$

where γ is the path loss exponent, which resembles the reduction of propagation due to reflection, diffraction and scattering caused by objects such as buildings and trees. In free space, γ equals 2, which is what we used, but varying values would result in a more physically accurate model.

\ DIRECTIONAL CELLS

A directional cell is a cell that is aimed at a specific angle. Along this angle, the signal strength is received at its best. However, the signal can also be strong in other directions. We specify the beam of a directional cell a by four parameters:

- The **azimuth angle** φa is the angle from the top view between the north and the direction in which the cell is pointed, such that φa ∈ [0, 360) degrees. Note that cell towers and rooftop cells often contain three cells with 120 degrees in between.
- The **elevation angle** θa is the angle between the horizon plane and the tilt of the cell. The plane that is tilt along this angle is called the elevation plane.
- The **horizontal beam width** αa specifies in which angular difference from the azimuth angle in the elevation plane the signal loss is 3 dB or less. At 3 dB, the power of the signal is halved. The angles in the elevation plane for which the signal loss is 3 dB correspond to φa ± αa/2. In practice, these angles are around 65 degrees.
- The **vertical beam width** βa specifies the angular difference from θa in the vertical plane orthogonal to φa in which the signal loss is 3 dB. The angles in which the signal loss is 3 dB correspond to θa ± βa/2. In practice, these angles are around 9 degrees.

Let δg,a be the angle in the elevation plane between the azimuth angle φa and the orthogonal projection on the elevation plane of the line between the center of cell a and the center of grid tile g. Similarly, let ε□, be the angle from the side view between the line along the elevation angle θa and the line between the center of cell a and the center of grid tile g. Note that εg,a depends on the cell property of the installation height above ground level. We model the signal strength for directional cells as:

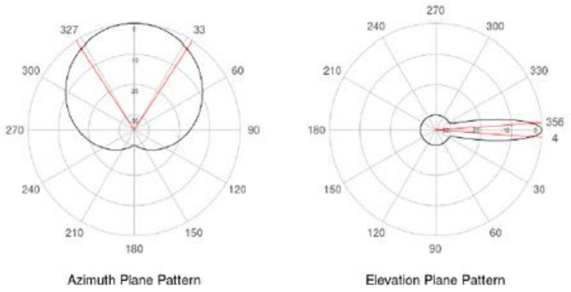
$$S(g,a) := S_0 - S_{\text{dist}}(r_{g,a}) - S_{\text{azi}}(\delta_{g,a}, \alpha_a) - S_{\text{elev}}(\varepsilon_{g,a}, \beta_a)$$

where $S0$ is the signal strength at $r0$ = 1 meter distance from the cell, in the direction of the beam so that δ = 0 and ε = 0. The signal loss due to distance to the cell, azimuth angle difference and elevation angle difference is specified by Sdist, Sazi and Selev, respectively. The definition of Sdist is similar to the omnidirectional cell. Each cell type has its own signal strength pattern for both the azimuth and elevation angles. These patterns define the relation between signal loss and the offset angles, i.e., δg,a for the azimuth and εg,a for the elevation angles. We model the radiation pattern for both Sazi and Selev by a linear transformation of the Gaussian formula, each with different values for parameters c and σ. Let

$$f(\varphi) := c - c\,\exp\left(-\frac{\varphi^2}{2\sigma^2}\right)$$

where c and σ2 are constants, whose value is determined by numerically solving equations for a set of constraints. These constraints are different for Sazi and Selev and depend on cell properties.

The resulting patterns are shown in **FIGURE 2**. The black line shows the relation between signal loss and angle in the azimuth plane (left) and elevation plane (right). The grey circles correspond to the signal loss; the outer circle means 0 dB loss (which is only achieved in the main direction), the next circle corresponds to 5 dB loss, and so forth. The red lines denote the angles corresponding to 3 dB loss. The angle between the red lines is 2αa in the azimuth plane and 2βa in the elevation plane. Although these models approximate the general curve of real radiation patterns, the radiation patterns are more complex in reality, e.g. they often contain local spikes caused by so-called side and back lobes.



Azimuth Plane Pattern

Elevation Plane Pattern

Figure 2: Radiation patterns for the azimuth and elevation planes

FIGURE 3 (top row) illustrates the signal strength at the ground level from above for a specific cell. In this case, the cell is placed at x = 0, y = 0 at 55 meters above ground level in an urban environment (γ = 4), has a power of 10 W, and is directed eastwards with an elevation angle (tilt) of 5 degrees, a horizontal beam width of 65 degrees and a vertical beam width of 9 degrees. Notice that the signal strength close to the cell, which on ground level translates to almost under the cell, is lower than at a couple of hundred meters distance. This is caused by relatively large ε angles at grid tiles nearby the cell.

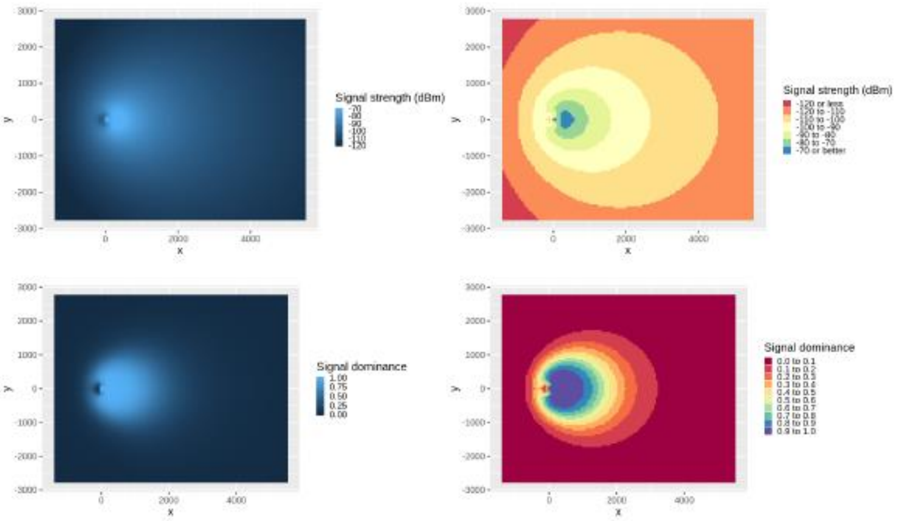


Figure 3: Signal strength at ground level from above (for a specific cell)

Once the coverage area has been estimated (containing information of signal strength), this information is 'rasterized' using the 'project_grid_cell' by defining a set of grid tiles that define the coverage area and assigning a specific signal strength (and/or other cell properties) to each grid tile following the same procedure defined in the figure below.

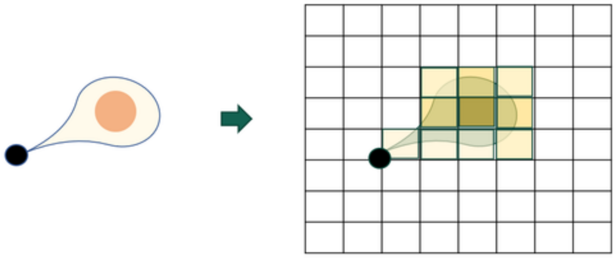


Figure 4: From coverage area to cell footprint

4.2 METHOD 2: ESTIMATION OF SIGNAL STRENGTH FROM MNO SIGNAL STRENGTH DATA

4.2.1 OBJECTIVE

Generate signal strength values for each cell based on the available MNO Network Topology Data.

4.2.2 PARAMETERS

- Radio propagation model parameters

4.2.3 INPUT DATA

- Cell Footprint with Differentiated Signal Strength Coverage Areas [INPUT]

4.2.4 OUTPUT DATA

- Cell Signal Strengths [INTERMEDIATE RESULTS]

4.2.5 METHODOLOGY

For each antenna (network cell), each grid tile of the coverage area provided by the MNO is intersected with the 'project_grid_cell'. All the intersected grids tiles from 'project_grid_cell' are considered as part of the coverage area of the antenna. The signal strength of each grid tile from 'project_grid_cell' is calculated as follows:

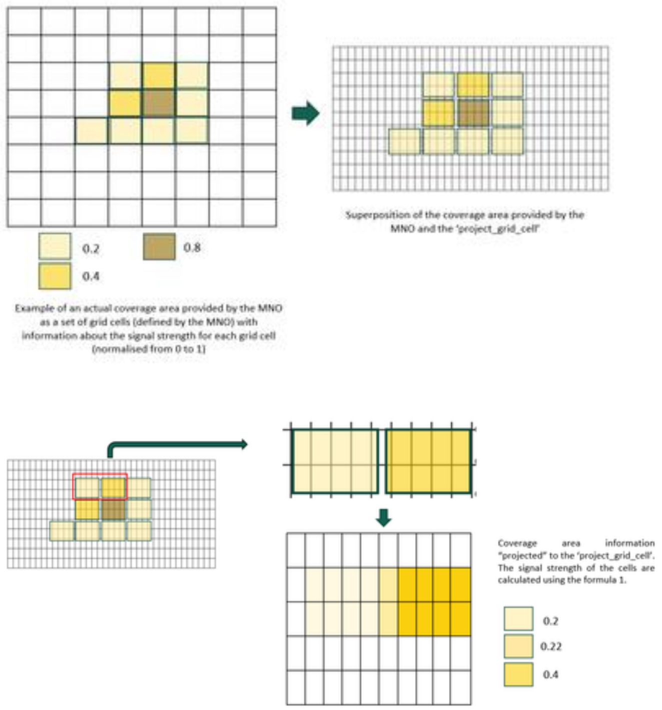
Equation 1

$$\sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}/A_T * SEi$$

Where:

- n: the number of grid tiles from 'project_grid_cell' that intersect with the MNO grid tiles
- m: the number of MNO grid tiles that intersect with the grid tile [i] from the 'project_grid_cell'.
- Aij: intersection area between the grid tile [i] and the grid tile 'j' from MNO data.
- AT: total area of the grid tile [i]
- SEi: signal strength of the MNO grid tile 'j'

Some examples of the process are shown in **FIGURE 5**.



Example of an actual coverage area provided by the MNO as a set of grid cells (defined by the MNO) with information about the signal strength for each grid cell (normalised from 0 to 1)

Superposition of the coverage area provided by the MNO and the 'project_grid_cell'

Coverage area information "projected" to the 'project_grid_cell'. The signal strength of the cells are calculated using the formula 1.

Figure 5: Transformation of information from MNO grid based data to project grid tiles

5 MODULE/METHOD 4: CELL FOOTPRINT ESTIMATION

5.1 OBJECTIVE

Generate a cell footprint based on the propagation model (signal strength). This module consists of two parts: in the first part, the signal strength values are transformed and normalised to the domain [0, 1]. In the second part, the footprint values are pruned.

The transformation and normalisation are needed because of the following. Signal strength values are provided in dBm. Their range in practice is between –70dBm (excellent connection) and –110dBm (bad connection). In this submodule, we seek a transformation function that results in values that we can use later to estimate the connection of probability in case there are multiple cells available. If for instance, there are two cells available at a specific geographic point, one with signal strength –90dBm and one with –100dBm, and all other factors, such as capacity, are out of consideration: how much more preferred is the first cell by the MNO? The footprint value should answer that: a footprint value of x is considered two times as good as a footprint value of ½ x.

The footprint values are pruned in the second submodule for computational reasons. We can expect that for each grid tile, especially in urban areas, there may be tenths or even hundreds of cells for which a signal can be perceived. However, we also expect that the signal strength is only good for a couple of them, which can be assumed to be the candidates for mobile network. By pruning the lowest footprint values, we reduce large amounts of data, which has a negligible effect of the further processing.

5.2 PARAMETERS

- Transformation function parameters:
  - For linear:
    - $\mathbf{S}_{\mathbf{min}}$ – Minimal value for the linear transformation
    - $\mathbf{S}_{\mathbf{max}}$ – Maximal value for the linear transformation
  - For logistic:
    - $\mathbf{S}_{\mathbf{steep}}$ – Parameter for calculating logistic curve
    - $\mathbf{S}_{\mathbf{mid}}$ - Parameter for calculating logistic curve
- **X**: the number of cells that are not pruned per grid tile.

5.3 INPUT DATA

- Cell Signal Strengths [INTERMEDIATE RESULTS]

5.4 OUTPUT DATA

- Cell footprint values [INTERMEDIATE RESULTS]

5.5 METHODOLOGY

The methodology of this module consists of two submodules. The first transforms the signal strength values to the 0-1 domain. The second submodule prunes these values; very low values are pruned because of computational reasons.

5.5.1 TRANSFORMATION

Let us denote S(g,a) for the signal strength value in dBm of cell a perceived in grid tile g (center). We will use s(g,a) for the output footprint value. We provide two options, a linear and a logistic function, but other any transformation can work well. Experiments need to be conducted to show which transformation is preferred.

Option 1: linear

$$s(g,a) = \max\big(0, \min(1,\ S(g,a) - S_{min})\ /(S_{max}\ -\ S_{min})\big)$$

When we set the parameters S_{min} and S_{max} to –130 and –50 respectively, all signal strength values of -50 dBm and higher will result in a footprint value of 1, and –130 dBm or lower in a footprint value of 0.

Option 2: logistic

$$s(g,a)\ =\ \frac{1}{1 + \exp\left(-S_{steep}(S(g,a) - S_{mid})\right)}$$

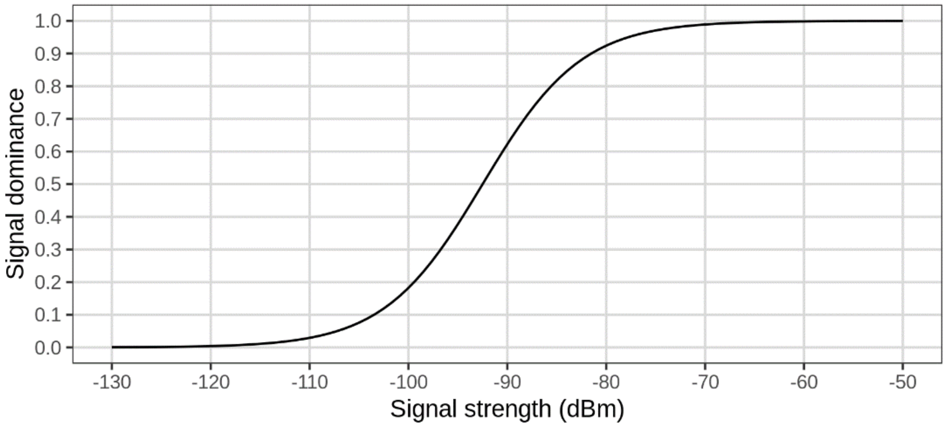


Figure 6: Logistic function using S_{steep} = -92.5 and S_{mid} = 0.2

The parameters S_{steep} and S_{mid} determine the shape of this logistic function. Figure x shows the shape when these parameters are set to –92,5 dBm and 0.2dBm respectively.

This method is introduced and described by Tennekes and Gootzen (2022). They call the output **signal dominance**.

5.5.2 PRUNING

Pruning can simply be done by removing all values lower than a fixed threshold value, say 0.01. However, a better method is to look at relative values: in grid tiles for which there are plenty of cells with good connection, all low valued cell footprint values can be pruned, whereas in grid tiles where there are only a few cells, each with low signal strength, than pruning may not be a good idea.

A better method is to use the top X (e.g., top 10) cells per grid tile. All footprint values outside this top 10 will be set to 0, and therefore not included in the output.

$$s(g,a) = s(g,a)\ if\ s(g,a)\ \in TOP\ X\ (s(g,a')\ for\ all\ a')\ 0\ otherwise$$

6 MODULE/METHOD 5: CELL CONNECTION PROBABILITY ESTIMATION

6.1 OBJECTIVE

Generate a cell connection probabilities based on footprint values. The cell connection probabilities indicate to following: what is the probability that a device is connected to a certain network cell, given that it is located in a certain grid tile?

6.2 PARAMETERS

None

6.3 INPUT DATA

- Cell footprint values [INTERMEDIATE RESULTS]

6.4 OUTPUT DATA

- Cell Connection Probabilities [INTERMEDIATE RESULTS]

6.5 METHODOLOGY

The probability that a device will use cell a when it is located in grid tile g is denoted by P(a | g), and is calculated by

$$P(a|g|) \; = \; \frac{s(g,a)}{\sum_{a' \in A}^{s}(g,a')}$$

where A is the set of all cells.

In this method, which is described by Tennekes and Gootzen (2022), we do not take load balancing and capacity into account.

Note that the P(a | g) values add up to 1 for each grid tile.

7 MODULE/METHOD 6: POSTERIOR PROBABILITY ESTIMATION

7.1 OBJECTIVE

Generate posterior probabilities based on prior probabilities and cell connection probability values. These probabilities estimate the location of a device given that it is located in a certain grid tile.[6] This location is not a single geographical point, but a spatial distribution. Note that this method does not take into account the number of devices that are connected to network cells that are in reach. This information is valuable when estimating a spatial density of multiple devices rather than the spatial distribution of a single device.

7.2 PARAMETERS

None

7.3 INPUT DATA

- Cell Connection Probabilities [INTERMEDIATE RESULTS]
- Grid Prior Land Use Probabilities [REFERENCE INPUT DATA]

7.4 OUTPUT DATA

- Posterior Probabilities Values [INTERMEDIATE RESULTS]

7.5 METHODOLOGY

We use the Bayes rule, as illustrated in Tennekes and Gootzen (2022):

$$P(g|a|) \propto P(g)P(a|g|)$$

Here, P(g | a) is the posterior probability: the probability that a device is located in grid tile g given that it is connected to cell a. P(g) are the prior probabilities. P(a | g) are the cell connection probabilities (see above).

Note that the posterior values should add up to 1 for each cell. Therefore, a normalisation is required in the implementation.

Based on the above, the following methodological processing steps can be distinguished:

1. Data ingestion
2. Joining input datasets based on grid ID
3. Multiplying the columns: prior_value and cell_connection_probability

[6] Reference work:
Tennekes, M., Gootzen, Y.A.P.M. (2022) Bayesian location estimation of mobile devices using a signal strength model, Journal of Spatial Information Science, 29-66

4. Normalising the results of the multiplication, so that it adds up to 1 for each cell

7.6 PSEUDO-CODE

```
# 1) Data ingestion

df_prior = read_prior_df()
df_conn_probs = read_conn_probs_df()

# 2) Joining connections probabilities to prior,
# so that each grid id is now associated with a cell_id, prior probability
# and a connection probability

df_joined = df_prior.join(df_conn_probs, on = "grid_id", how = "left")

# 3) Calculating the posterior as the result of multiplication

df_joined["posterior_prob"] = df_joined["prior_value"]*df_joined["CELL_CONNECTION_PROBABILITY"]

# 4) Normalizing the results so probability values sum up to 1 for each cell

summarised_posterior_df = df_joined.groupBy(["cell_id"])[["cell_id", "posterior_prob"]]\
  .agg(F.sum(F.col("posterior_prob")).alias("posterior_sum"))

df_results = df_joined.merge(summarised_posterior_df, on = "cell_id", how = "left")

# Final probability value as normalized value
# TODO deal with 0s and NaNs
df_results["posterior_probability"] = df_results["posterior_prob"]/df_results["posterior_sum"]
```

8 MODULE 7: MNO EVENT DATA CLEANING – SYNTACTIC CHECKS

8.1 METHOD 1: MNO EVENT DATA CLEANING – SYNTACTIC CHECKS

8.1.1 OBJECTIVE

The objective of this method is to perform syntactic checks on the raw MNO Event Data. Data not matching the expected syntax will be removed. Based on the removed records, quality metrics will be created.

Syntactic checking is the first step in the data processing to ensure that it is in a format suitable for subsequent operations; i.e. that the required fields are present and populated with sufficient values. Identification and removal of duplicated rows is also performed by this method. There is no cross-referencing to other data collections and there are no aggregation operations except for producing quality metrics. Such semantic checks are performed in further modules of the pipeline.

Mapping with other standards*: the syntactic checks defined in this method can be mapped with those included in the sub-process 5.3 Review and Validate of GSBPM and to Validation levels 0 and 1 as defined in the ESS handbook Methodology for data validation 1.1 (Revised edition 2018)*[7].

Correspondence with Deliverable D3.1*: in Table 2 of Chapter 6, this method corresponds to Quality Module QM.EvSynt.*

8.1.2 PARAMETERS

- **data_period_start**: Start of the data period under study; any records before this will be removed. E.g. “2023-01-01“
- **data_period_end**: End of the data period under study; any records after this will be removed. E.g. “2023-12-31“
- **bounding_box**: In case input data contains longitude/latitude data, we can filter out values that are not in the country bounding box. E.g. (60,90,110,130)
- **local_mcc**: used for comparison with the mcc value to assess if the domain of the data is domestic

8.1.3 INPUT DATA

- MNO Event Data - Raw [INPUT]

8.1.4 OUTPUT DATA

- Clean MNO Event Data [INTERMEDIATE RESULTS]

[7] Available at: https://cros-legacy.ec.europa.eu/system/files/ess_handbook_-_methodology_for_data_validation_v1.1_-_rev2018_0.pdf

- MNO Event Data Syntactic Quality Metrics [INTERMEDIATE RESULTS]

8.1.5 METHODOLOGY

1. Check if the file is readable.
2. Remove rows where a null value is in a columns **user_id** or **timestamp.** Rows with no location data will also be removed. This means that if both **cell_id** and **longitude, latitude** are null, the row is removed. If either **cell_id** or **longitude** and **latitude** exist, the row will remain.
3. Remove rows where timestamp is not in UTC format.
4. Remove rows where an invalid value for **mcc/mnc** is found. Remove rows where an invalid value for **plmn** is found where event is outbound.
5. Check if the timestamp is between **data_period_start** and **data_period_end**. Rows with timestamp before **data_period_start** or after **data_period_end** will be removed.
6. For rows where longitude/latitude data is present: Check if the coordinates fall inside the bounding box. Remove records that do not.
7. Identify duplicate raws and remove them.

Statistics about the number and type of error entries are collected between process steps to determine the number of entries removed in each step of the process. The statistics are output as MNO Event Data Syntactic Quality Metrics [INTERMEDIATE RESULTS].

8.2 METHOD 2: MNO EVENT– GENERAL STATISTICS

8.2.1 OBJECTIVE

This method computes and stores a set of statistics on the events on a daily basis. These statistics are not connected to the identification and cleaning of errors in the input data; nonetheless, these can be useful for the following:

- to identify unexpected/suspect situations in input data, that could need further investigation;
- to filter/select devices for specific use cases in the next stages in the pipeline.

The method is placed after Method 1: MNO Event Data Cleaning – Syntactic Checks in order to produce the statistics only to formally correct, complete and not duplicated events.[8]

8.2.2 PARAMETERS

- home_mno: MCCMNC corresponding to home MNO lookback_period: data lookback period of the average in specific metrics

8.2.3 INPUT DATA

- Clean MNO Event Data [INTERMEDIATE RESULTS]

[8] This method proposes additional statistics that can be useful for the quality monitoring of the pipeline. It will be developed in next software releases.

8.2.4 OUTPUT DATA

- General Event Statistics Metrics [INTERMEDIATE RESULTS]

8.2.5 METHODOLOGY

Every day, the following statistics should be computed and stored according to the structure of the General Event Statistics Metrics [INTERMEDIATE RESULTS]. The statistics are described below together with a short description of the reason why they could be useful in the pipeline.

STATISTIC	SHORT DESCRIPTION
Number of unique domestic devices per day (where MCCMNC = home_MNO)	Very high or very low values could be further investigated. The value could be compared with the expected number of users of the MNO, if available.
Number of unique inbound roaming devices per day (where MCCMNC != home_MNO)	Very high or very low values could be further investigated. The value could be compared with the expected number of inbound tourists in the country (taking also into account the market share of the MNO).
Number of unique inbound roaming devices per MCC and MNC per day (where MCCMNC != home_MNO)	Very high or very low values could be further investigated. The value could be compared with the expected number of inbound tourists from specific countries in the country (taking also into account the market share of the MNO).
Number of unique outbound roaming devices per day (where PLMN != home_MNO)	Very high or very low values could be further investigated. The value could be compared with the expected number of outbound tourists from the country (taking also into account the market share of the MNO).
Number of unique outbound roaming devices per PLMN per day (where PLMN != home MNO)	Very high or very low values could be further investigated. The value could be compared with the expected number of outbound tourists to specific countries from the country (taking also into account the market share of the MNO).
Number of domestic events per day (where MCCMNC = home MNO)	Number of events should be coherent to the data type being used (CDR/Signalling) and the number of devices.
Number of inbound roaming events per day (where MCCMNC != home MNO)	
Number of inbound roaming events per MCC and MNC per day (where MCCMNC != home MNO)	
Number of outbound roaming events per day (where PLMN != home MNO)	
Number of outbound roaming events per PLMN per day (where PLMN != home MNO)	
Number of domestic events per day per hour	Diurnal distribution of events per day follow a specific pattern (elephant chart).
Number of inbound roaming events per day per hour	
Number of outbound roaming events per day per hour	
Number of unique MCC combinations for inbound roaming data per day	How many countries in inbound roaming data?
Number of unique MCCMNC combinations for inbound roaming data per day	How many foreign MNOs in inbound roaming data?
Number of unique PLMN combinations for outbound roaming data per day	In how many countries outbound roamers are in?
Average number of domestic events per device per day	Indication on the data type (CDR/Signalling) being used: low for CDR, high for Signalling

STATISTIC	SHORT DESCRIPTION
Average number of inbound roaming events per device per day;	
Average number of outbound roaming events per device per day;	
Average (and median) number of unique cell_id's in domestic events per device per day;	Indication of how many cell_id's are devices connected to.
Average (and median) number of unique cell_id's in inbound roaming events per device per day;	Indication of how many cell_id's are devices connected to.
Average (and median) number of unique PLMN's in outbound roaming events per device per day;	Indication of how 'geographically mobile' devices are throughout the day in foreign countries (and also if they change foreign operator).
Average number of days of device ID present in days throughout the data lookback period;	Indication of continuity of device ID's. This cannot be 1 (suggests 24hrs recalculated ID hash).

9 MODULE/METHOD 8: MNO EVENT DATA – SYNTACTIC QUALITY WARNINGS

9.1 OBJECTIVE

This module analyses the MNO Event Data Syntactic Quality Metrics [INTERMEDIATE RESULTS] to identify anomalous situations that needs to be further investigated. The output of the method is the MNO Event Data Quality Warnings [INTERMEDIATE RESULTS] which displays: plots of the metrics over time, the anomalous data, the related warning and (whenever possible) suggestions on additional investigations.

Quality warnings from syntactic checks are linked to the following characteristics or anomalies of the MNO Event Data – Raw [INPUT]:

- Size
- Missing values
- Wrong data types or formats
- Out of range values: format is correct but value is not acceptable
- Transformations performed to standardise the input
- Duplications

Anomalous quality metrics values should always launch a warning; nonetheless, not necessarily each warning has to correspond to an error. A warning means that something is to be checked, even if afterwards it is understood that the anomaly was not an error. In most cases warnings imply the definition of thresholds above or under which the warning will be launched. The thresholds can be identified in different ways. In general, the method will allow to set thresholds as input parameters but will include also a default threshold (e.g. 90%). The default threshold could be adjusted during testing.

Quality warnings are supposed to be executed daily.

Mapping with other standards*: the quality warnings defined in this method can be mapped with those included in the Overarching process Quality management of GSBPM.*

Correspondence with Deliverable D3.1*: in Table 2 of Chapter 6 this method corresponds to Quality Module QM.EvSynt.*

9.2 PARAMETERS

- **thresholds**: values above or under which the method launches the warning. They can be:
  i. specific values to which the metric is compared directly (e.g. the error rate is above 30%), or
  ii. defined in relative terms based on the average of the metrics in previous period (e.g. the missing value rate for the variable cell_id in the data object is 50% higher than the average missing value rate for the variable cell_id in the last month data objects), or

iii. defined dynamically based also on the variability of the metrics in previous period (e.g. the size of the data object is smaller than the average size of the last week data objects – 2 times the standard deviations).

The average and standard deviation can be calculated also over the distribution of events by cell or by user in the same day. Thus. the following **thresholds** are to be considered:

- SIZE_RAW_DATA_BYDATE_VARIABILITY: "X" times the standard deviation of raw data size to calculate upper and lower control limits
- SIZE_RAW_DATA_BYDATE_ABS_VALUE_UPPER_LIMIT
- SIZE_RAW_DATA_BYDATE_ABS_VALUE_LOWER_LIMIT
- SIZE_CLEAN_DATA_BYDATE_VARIABILITY: "X" times the standard deviation of raw data size to calculate upper and lower control limits
- SIZE_CLEAN_DATA_BYDATE_ABS_VALUE_UPPER_LIMIT
- SIZE_CLEAN_DATA_BYDATE_ABS_VALUE_LOWER_LIMIT
- TOTAL_ERROR_RATE_BYDATE_OVER_AVERAGE: "X%" above the average
- TOTAL_ERROR_RATE_BYDATE_VARIABILITY: "X" times the standard deviation of raw data size to calculate upper control limit
- TOTAL_ERROR_RATE_BYDATE_ABS_VALUE_UPPER_LIMIT
- ERROR_RATE_BYDATE_BYCELL_OVER_AVERAGE: "X%" above the average
- ERROR_RATE_BYDATE_BYCELL_VARIABILITY: "X" times the standard deviation of raw data size to calculate upper control limit
- ERROR_RATE_BYDATE_BYCELL_ABS_VALUE_UPPER_LIMIT
- ERROR_RATE_BYDATE_BYUSER_OVER_AVERAGE: "X%" above the average
- ERROR_RATE_BYDATE_BYUSER_VARIABILITY: "X" times the standard deviation of raw data size to calculate upper control limit
- ERROR_RATE_BYDATE_BYUSER_ABS_VALUE_UPPER_LIMIT
- ERROR_RATE_BYDATE_BYCELL&USER_OVER_AVERAGE: "X%" above the average
- ERROR_RATE_BYDATE_BYCELL&USER_VARIABILITY: "X" times the standard deviation of raw data size to calculate upper control limit
- ERROR_RATE_BYDATE_BYCELL&USER_ABS_VALUE_UPPER_LIMIT
- Missing_value_RATE_BYDATE_field_name_AVERAGE: "X%" above the average
- Missing_value_RATE_BYDATE_field_name_VARIABILITY: "X" times the standard deviation of raw data size to calculate upper control limit
- Missing_value_RATE_BYDATE_field_name_ABS_VALUE_UPPER_LIMIT
- Wrongtype/Format_RATE_BYDATE_field_name_AVERAGE: "X%" above the average
- Wrongtype/Format_RATE_BYDATE_field_name_VARIABILITY: "X" times the standard deviation of raw data size to calculate upper control limit
- Wrongtype/Format_RATE_BYDATE_field_name_ABS_VALUE_UPPER_LIMIT
- Out_of_range_RATE_BYDATE_field_name_AVERAGE: "X%" above the average
- Out_of_range_RATE_BYDATE_field_name_VARIABILITY: "X" times the standard deviation of raw data size to calculate upper control limit
- Out_of_range_RATE_BYDATE_field_name_ABS_VALUE_UPPER_LIMIT
- Conversion_RATE_BYDATE_field_name_AVERAGE: "X%" above the average
- Conversion_RATE_BYDATE_field_name_VARIABILITY: "X" times the standard deviation of raw data size to calculate upper control limit
- Conversion_RATE_BYDATE_field_name_ABS_VALUE_UPPER_LIMIT
- DUPLICATION_RATE_BYDATE_OVER_AVERAGE: "X%" above the average

  - DUPLICATION_RATE_BYDATE_VARIABILITY: "X" times the standard deviation of raw data size to calculate upper control limit
  - DUPLICATION_RATE_BYDATE_ABS_VALUE_UPPER_LIMIT

- period: week, month, quarter or any previous period for which the plot is created and the average and the standard deviations of the quality metric(s) are calculated.

9.3 INPUT DATA

- MNO Event Data Syntactic Quality Metrics [INTERMEDIATE RESULTS]
- previous period MNO Event Data Syntactic Quality Metrics [INTERMEDIATE RESULTS]

9.4 OUTPUT DATA

- MNO Event Data Quality Warnings [INTERMEDIATE RESULTS] for syntactic checks

9.5 METHODOLOGY

We list here the different methods to be applied to produce the warnings. This is to be considered as a first set of warnings that should be possible to adjust during testing and integrated further afterwards (and, similarly, for each module of the pipeline).

9.5.1 SIZE OF THE MNO EVENT DATA - RAW [INPUT]

Calculate the "Total initial frequency" by summing up the "initial frequency" over cells and users and produce a plot with the value of the "Total initial frequency" over the last period (week, month, quarter) with the time on horizontal axis and the metric on the vertical axis.

For the warnings, the current day value of "Total initial frequency" should be compared with the thresholds. Since the metric is expected to be highly variable over time, the threshold should be based on its variability. To this aim, the average and the standard deviation (SD) of the same metrics over the previous period should be calculated. Afterwards, the warnings will be launched according to the following indications:

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Value of the size by date of the raw data object	Size by date is out of control limits calculated on the basis of average and standard deviation of the distribution of the size in previous period. Control limits = (average ± X·SD)	X (default e.g. 3)	The number of events in raw data is unexpectedly low/high compared to the previous period, please check if there have been issues in the network.
Value of the size by date of the raw data object	The size by date is under/above thresholds X and Y.	X = the value should be set by the user but the default could be set equal to the previous lower control limit Y = the value should be set by the user but the default could be set equal to the previous upper control limit	The number of events in raw data is above/under the threshold, please check if there have been changes in the network.

9.5.2 SIZE OF THE CLEAN MNO EVENT DATA [INTERMEDIATE RESULTS]

Calculate the "Total final frequency" by summing up the "final frequency" over cells and users and produce a plot with the value of the "Total final frequency" over the last period (week, month, quarter) with the time on horizontal axis and the metric on the vertical axis.

For the warnings, the current day value of "Total final frequency" should be compared with the thresholds. Since the metric is expected to be highly variable over time the threshold should be based on its variability. To this aim, the average and the standard deviation (SD) of the same metrics over the previous period should be calculated. Afterwards, the warnings will be launched according to the following indications:

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Value of the size by date of the raw data object	Size by date is out of control limits calculated on the basis of average and standard deviation of the distribution of the size in the previous period. Control limits = (average ± X·SD)	X (default e.g. 3)	The number of events in clean data is unexpectedly low/high compared to the previous period, please check if there have been issues in the network.
Value of the size by date of the raw data object	The size by date is under/above thresholds X and Y.	X = the default could be set equal to the previous lower control limit Y = the default could be set equal to the previous upper control limit	The number of events in clean data is above/under the threshold, please check if there have been changes in the network.

9.5.3 ERROR RATE

Total Error rate by date = (Total initial frequency - Total final frequency) / Total initial frequency*100

Produce a plot over the last period (week, month, quarter) of this rate with the time on horizontal axis and the rate on the vertical axis.

For the warnings, the current day value should be calculated and compared with the thresholds. To this aim, the average and the standard deviation (SD) of the same metrics over the previous period should be calculated. Afterwards the warnings will be launched according to the following indications:

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Error rate by date	Error rate by date is above the previous period average by X%.	X% (default e.g. 30%)	The error rate is unexpectedly high compared to the previous period.
Error rate by date	Error rate by date is above the upper control limit calculated on the basis of average and standard deviation of the distribution of the error rate in the previous period. Upper Control limit = (average + X·SD)	X (default e.g. 2)	The error rate is unexpectedly high with respect to the previous period, taking into account its usual variability
Error rate by date	The error rate by date is above the value X.	X% (default e.g. 20%)	The error rate is above the threshold.

9.5.4 ERROR RATE BY DATE FOR CELL

Error rate by date for cell_id# = (Total initial frequency for cell_id# - Total final frequency for cell_id#) / Total initial frequency for cell_id#*100

For the warnings the current daily value should be calculated and compared with the thresholds. To this aim the average and the standard deviation (SD) of the same metrics **over all the cell_id for the same day** should be calculated, than the warnings will be launched according to the following indications:

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Error rate by date for cell_id#	The error rate by date for cell_id# is above the value X.	X% (default e.g. 20%)	The error rate for cell_id# is above the threshold.
Error rate by date for cell_id#	Error rate by date for cell_id# is above the average by X%.	X% (default e.g. 30%)	The error rate for cell_id# is unexpectedly high compared to the average of the cells. Check the corresponding antenna.
Error rate by date for cell_id#	Error rate by date for cell_id# is over the upper control limit calculated on the basis of average and standard deviation of the distribution of the error rate over all the cell_id for the same day. Upper Control limit = (average + X·SD)	X (default e.g. 2)	The error rate for cell_id# is unexpectedly high compared to the average of the cells. Check the corresponding antenna.

9.5.5 ERROR RATE BY DATE FOR USER

Error rate by date for user_id# = (Total initial frequency for user_id# - Total final frequency for user_id#) / Total initial frequency for user_id#*100

For the warnings the current daily value should be calculated and compared with the thresholds. To this aim the average and the standard deviation (SD) of the same metrics **over all the user_id for the same day** should be calculated, than the warnings will be launched according to the indications in the table. In case of warning, this information will be stored in the log table for MNO as specified in the MNO Event Data Quality Warnings [INTERMEDIATE RESULTS]. The warning message to be shared with NSI will report the count of user_id that are over the thresholds per day.

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Error rate by date for user_id#	The error rate by date for user_id# is over the value X.	X% (default e.g. 20%)	The error rate for user_id# is over the threshold.
Error rate by date for user_id#	Error rate by date for user_id# is over the average by X%	X% (default e.g. 30%)	The error rate for user_id# is unexpectedly high compared to the average of the users. Check the corresponding device
Error rate by date for user_id#	Error rate by date for user_id# is over the upper control limit calculated on the basis of average and standard deviation of the distribution of the error rate over all the user_id for the same day. Upper Control limit = (average + X·SD)	X (default e.g. 2)	The error rate for user_id# is unexpectedly high compared to the average of the users. Check the corresponding device

9.5.6 ERROR RATE BY DATE FOR USER*CELL

Error rate by date for user_id#*cell_id# = (Total initial frequency for user_id#*cell_id# - Total final frequency for user_id#*cell_id#) / Total initial frequency for user_id#*cell_id# *100

For the warnings the current daily value should be calculated and compared with the thresholds. To this aim the average and the standard deviation (SD) of the same metrics **over all the user_id#*cell_id# for the same day** should be calculated, than the warnings will be launched according to the indications in the following table. In case of warning, this information will be stored in the log table for MNO MNO Event Data Quality Warnings [INTERMEDIATE RESULTS]. The warning message to be shared with NSI will report the count of user_id*cell_id that are over the thresholds per day.

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Error rate by date for user_id#*cell_id#	The error rate by date for user_id#*cell_id# is above the value X.	X% (default e.g. 20%)	The error rate for user_id#*cell_id# is above the threshold.
Error rate by date for user_id#*cell_id#	Error rate by date for user_id#*cell_id# is above the average by X%.	X% (default e.g. 30%)	The error rate for user_id#*cell_id# is unexpectedly high compared to the average. Check the interaction between the corresponding device and antenna.
Error rate by date for user_id#*cell_id#	Error rate by date for user_id#*cell_id# is above the upper control limit calculated on the basis of average and standard deviation of the distribution of the error rate over all the user_id*cell_id for the same day. Upper Control limit = (average + X·SD)	X (default e.g. 2)	The error rate for user_id#*cell_id# is unexpectedly high compared to the average. Check the interaction between the corresponding device and antenna.

9.5.7 WARNING FOR MISSING VALUE RATE FOR EACH FIELD (USER_ID, LOCATION (BOTH CELL_ID AND LATITUDE AND LONGITUDE ARE MISSING), MCC, MNC, TIMESTAMP)

Missing value rate by date of fieldname = Number of missing value for fieldname / Total initial frequency *100.

For the warnings each current daily value should be calculated and compared with the thresholds. To this aim, the average and the standard deviation (SD) of the same metrics over the **previous period** should be calculated. Afterwards, the warnings will be launched according to the following indications for each mandatory field:

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Missing value rate by date of *field_name*	Missing value rate by date of *field_name* is above the previous period average by X%.	X% (default e.g. 30%)	The missing value rate of *field_name* after syntactic check application is unexpectedly high compared to the previous period.
Missing value rate by date of *field_name*	Missing value rate by date of *field_name* is above the upper control limit calculated on the basis of average and standard deviation of the distribution of the Missing value rate of *field_name* in the previous period. Upper Control limit = (average + X·SD)	X (default e.g. 2)	The missing value rate of *field_name* after syntactic check application is unexpectedly high compared to the previous period, taking into consideration its usual variability.
Missing value rate by date of *field_name*	Missing value rate by date of *field_name* is above the value X.	X% (default e.g. 20%)	The missing value rate of *field_name* is above the threshold.

9.5.8 WARNING FOR ERRORS IN DATA TYPES/FORMAT FOR EACH FIELD (USER_ID, CELL_ID, LATITUDE, LONGITUDE, MCC, MNC, TIMESTAMP, PLMN, LOC_ERROR)

Wrong type/format rate of *field_name* = number of values with errors in data types/format for *field_name*/ Total initial frequency *100.

For the warnings, each current day value should be calculated and compared with the thresholds. To this aim, the average and the standard deviation (SD) of the same metrics over the **previous period** should be calculated. Afterwards, the warnings will be launched according to the following indications for each mandatory field:

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Wrong type/format rate by date of *field_name*	Wrong type/format rate by date of *field_name* is above the previous period average by X%.	X% (default e.g. 30%)	The wrong type/format rate of *field_name* is unexpectedly high compared to the previous period.
Wrong type/format rate by date of *field_name*	Wrong type/format rate by date of *field_name* is above the upper control limit calculated on the basis of average and standard deviation of the distribution of the Wrong type/format rate of *field_name* in the previous period. Upper Control limit = (average + X·SD)	X (default e.g. 2)	The wrong type/format rate of *field_name*is unexpectedly high compared to the previous period, taking into consideration its usual variability.
Wrong type/format rate by date of *field_name*	The wrong type/format rate by date is above the value X.	X% (default e.g. 20%)	The wrong type/format rate of *field_name* is above the threshold.

9.5.9 WARNING FOR OUT OF RANGE RATE (VALUE IS NOT WITHIN THE SET OF ACCEPTED VALUES) FOR EACH RELEVANT MANDATORY FIELD (LATITUTE, LONGITUDE, TIMESTAMP, LOC_ERROR)

Out of range rate of *field_name* = number of out of range values for *field_name* / Total initial frequency *100.

For the warnings, each current day value should be calculated and compared with the thresholds. To this aim, the average and the standard deviation (SD) of the same metrics over the previous period should be calculated. Afterwards, the warnings will be launched according to the following indications for each mandatory field:

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Out of range rate by date of *field_name*	Out of range rate by date of *field_name* is above the previous period average by X%.	X% (default e.g. 30%) Alternative:	The out of range rate of *field_name* is unexpectedly high compared to the previous period.
Out of range rate by date of *field_name*	Out of range rate by date of *field_name* is above the upper control limit calculated on the basis of average and standard deviation of the distribution of the Out of range rate of *field_name* in the previous period. Upper Control limit = (average + X·SD)	X (default e.g. 2)	The out of range rate of *field_name* is unexpectedly high compared to the previous period, taking into consideration its usual variability.
Out of range by date rate of *field_name*	Out of range rate by date of *field_name* is above the value X.	X% (default e.g. 20%)	The out of range rate of *field_name* is above the threshold.

9.5.10 WARNING FOR CONVERSION ERROR RATE FOR EACH MANDATORY FIELD SUBJECTED TO A CONVERSION PROCEDURE (TIMESTAMP)

Conversion error rate of *field_name* = number of values with conversion errors for *field_name* / Total initial frequency *100.

For the warnings, each current day value should be calculated and compared with the thresholds. To this aim, the average and the standard deviation (SD) of the same metrics over the previous period should be calculated. Afterwards, the warnings will be launched according to the following indications for each mandatory field:

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Conversion error rate by date of *field_name*	Conversion error rate by date of *field_name* is above the previous period average by X%.	X% (default e.g. 30%)	The conversion error rate of *field_name* after syntactic check application is unexpectedly high compared to the previous period.
Conversion error rate by date of *field_name*	Conversion error rate by date of *field_name* is above the upper control limit calculated on the basis of average and standard deviation of the distribution of the Conversion error rate of *field_name* in the previous period. Upper Control limit = (average + X·SD)	X (default e.g. 2)	The conversion error rate of *field_name* is unexpectedly high compared to the previous period, taking into consideration its usual variability.
Conversion error rate by date of *field_name*	Conversion error rate by date of *field_name* is above the value X.	X% (default e.g. 20%)	The conversion error rate of *field_name* is above the threshold.

9.5.11 WARNING FOR DUPLICATION RATE BY DATE

Duplication rate by date= Total number of duplication/ Total initial frequency *100

For the warnings, the current day value should be calculated and compared with the thresholds. To this aim, the average and the standard deviation (SD) of the same metrics over the previous period should be calculated. Afterwards, the warnings will be launched according to the following indications:

MEASURE DEFINITION	CONDITION FOR THE WARNING	THRESHOLD	WARNING MESSAGE
Duplication rate by date	Duplication rate by date is above the previous period average by X%.	X% (default e.g. 30%)	The duplication rate is unexpectedly high compared to the previous period.
Duplication rate by date	Duplication rate by date is above the upper control limit calculated on the basis of average and standard deviation of the distribution of the error rate in the previous period. Upper Control limit = (average + X·SD)	X (default e.g. 2)	The duplication rate is unexpectedly high compared to the previous period, taking into account its usual variability.
Duplication rate by date	Duplication rate by date is above the value X	X% (default e.g. 20%)	The duplication rate by date is above the threshold.

9.5.12 COMPOSITION OF ERRORS FOR EACH FIELD (USER_ID, CELL_ID, LATITUDE, LONGITUDE, MCC, MNC, TIMESTAMP, LOC_ERROR)

For each field calculate the % of each type of error (missing value, wrong/format, out of range) on the total of errors for the same field. Build a pie chart with this % for each variable.

9.5.13 ADDITIONAL NOTE

If the computations with the different thresholds are too demanding we can simplify (e.g. maintaining only the absolute values and the one based on variability).

10 MODULE/METHOD 9: DEVICE DEMULTIPLEX

10.1 OBJECTIVE

The objective of this method is to differentiate the cleaned data output in the Clean MNO Event Data [INTERMEDIATE RESULTS], providing a separate output for each device.

These operations are performed on a single-entry level, there is no cross-referencing to other data collections.

The output of this method is the Event Data at Device level [INTERMEDIATE RESULTS], a collection of data streams with the same structure of the input data filtered by device.

The device demultiplex is supposed to be executed daily.

10.2 PARAMETERS

The introduction of new parameters is not expected for this method.

10.3 INPUT DATA

- Clean MNO Event Data [INTERMEDIATE RESULTS]

10.4 OUTPUT DATA

- Event Data at Device level [INTERMEDIATE RESULTS]

10.5 METHODOLOGY

To produce the Event Data at Device Level, the input is processed, and a new output stream is produced for each device, differentiating by the field *user_id*. Each record of the input dataset corresponds to a MNO event, containing information about the device identifier, the timestamp of the event, the identifier of the cell to which the device is connected and the geographical coordinates of the cell. The format and the content of the original stream is maintained in the output.

11 MODULE/METHOD 10: EVENT CLEANING AT DEVICE LEVEL – SEMANTIC CHECKS

11.1 OBJECTIVE

The objective of this module is to perform checks to identify and flag semantically erroneous events of devices. The semantic checks include the following checks:

- **Valid reference to cell_id**: the cases when MNO event data includes cell_id that do not exist in the MNO topology data on the specific date.
- **Illogical change of the location of the device based on the time and distance difference:** which implies identifying and flagging geographically impossible cases where a device “jumps” from one location to another over a very short period of time which is physically impossible (e.g. 10:00:00 cell in Madrid, 10:30:00 cell in Barcelona - more than 500 km away in 30 minutes). There are at least two cases when such issues exist:
  - The location of the network cell is incorrect for the specific time of the event;
  - This may be caused by the mobile devices connected to the network by passengers on the planes (usually when taking off and landing, and only rarely during the flight). In this case they are logically correct, but the devices are not on the ground.

In future releases of the pipeline, the abovementioned checks may be enhanced and new ones may be added.

Mapping with other standards*: the semantic checks defined in this method can be mapped with those included in the GSBPM sub-process 5.3 Review and Validate.*

Correspondence with Deliverable D3.1*: in Table 2 of Chapter 6 this method corresponds to Quality Module QM.EvSem.*

11.2 PARAMETERS

- semantic_min_distance_m - Minimum distance threshold to take into account when identifying illogical location of the events (meters).
- semantic_min_speed_m_s - Minimum speed calculated based on the cell’s point-locations (meters per second).

11.3 INPUT DATA

- Event Data at Device level [INTERMEDIATE RESULTS]
- MNO Network Topology Data - Raw [INPUT]. One of these two data objects should be used:
  - Cell Locations with Physical Properties [INPUT]

  - Cell Footprint with Differentiated Signal Strength Coverage Areas [INPUT]

11.4 OUTPUT DATA

- Semantically Cleaned Event Data at Device Level [INTERMEDIATE RESULTS]
- Device Semantic Quality Metrics [INTERMEDIATE RESULTS]

11.5 METHODOLOGY

11.5.1 VALID REFERENCE TO CELL_ID

\ METHODOLOGICAL STEPS

Identify and flag the MNO events that do not have corresponding cell_id with the date on which the event happened.

1. Link event data object with network topology data object based on cell_id;
2. If event cell_id does not exist at all in the network topology set error type 1, if event cell_id exists, but shows the incorrect validity date (not the same as event date), set error type 2, increase invalid event record counter of the corresponding error type. Flag those events;
3. Store the count with proper error type in the data object Device Semantic Quality Metrics [INTERMEDIATE RESULTS];
4. Flag the event with corresponding error_flag.

11.5.2 ILLOGICAL CHANGE OF THE LOCATION OF THE DEVICE BASED ON THE TIME AND DISTANCE DIFFERENCE

\ METHODOLOGICAL STEPS

For identifying events' locations that are certainly incorrect (error type 3), identify and flag the MNO events where distance to previous and the following event are both longer than `semantic_min_distance_m` and delta distance / time (speed) to previous and the following event are both higher than `semantic_min_speed_m_s`.

For identifying potential incorrect locations of the events (error type 4) identify and flag the MNO events where distance to previous or (XOR) the following event is longer than `semantic_min_distance_m` and delta distance / time (speed) to previous or (XOR) the following event is higher than `semantic_min_speed_m_s`.

1. For each not flagged MNO event data record, get the coordinates of the cell of the event;
2. Calculate distance (in meters) and speed (in meters/second) from previous and to next event.
3. Flag events:
   A. If distance to previous and the following event location (all both above the thresholds `semantic_min_distance_m` and `semantic_min_speed_m_s`). Flag those events as error code 3 (certain error).

   B. If distance to previous or (XOR) the following event location is above the thresholds `semantic_min_distance_m` and `semantic_min_speed_m_s`. Flag those events as error code 4. This includes also if the previous or the following events don't exist (e.g. first/last events of the device).

4. Increase invalid event record counter of the corresponding error type;
5. Store the count with proper error types in the data object Device Semantic Quality Metrics [INTERMEDIATE RESULTS];
6. Store event data including error flags in the Semantically Cleaned Event Data at Device Level [INTERMEDIATE RESULTS] data object.

11.6 PSEUDO-CODE

Example provided as SQL to identify and flag events

```
WITH
    gr1 AS (
        SELECT
            a.id AS event_id
            , a.device_id
            , a.tm
            , a.comment
            , lag(a.tm) OVER (PARTITION BY a.device_id ORDER BY a.tm) AS tm_prev
            , lead(a.tm) OVER (PARTITION BY a.device_id ORDER BY a.tm) AS tm_next
            , ROUND(ST_DistanceSpheroid(b.geom, lag(b.geom) OVER (PARTITION BY a.device_id ORDER BY a.tm), 'SPHEROID["WGS 84",6378137,298.257223563]')) AS distance_m_prev
            , ROUND(ST_DistanceSpheroid(b.geom, lead(b.geom) OVER (PARTITION BY a.device_id ORDER BY a.tm), 'SPHEROID["WGS 84",6378137,298.257223563]')) AS distance_m_next
        FROM
            eurostat_speed_distance.event AS a
            , eurostat_speed_distance.cell AS b
        WHERE
            a.cell_id=b.cell_id
    )
    , gr2 AS (
        SELECT
            event_id
            , device_id
            , tm
            , tm_prev
            , tm_next
            , ROUND(EXTRACT(EPOCH FROM (tm - tm_prev))) AS d_time_prev_s
            , ROUND(EXTRACT(EPOCH FROM (tm_next - tm))) AS d_time_next_s
            , distance_m_prev
            , distance_m_next
            , ROUND((distance_m_prev / EXTRACT(EPOCH FROM (tm - tm_prev)))::numeric, 1) AS speed_prev
            , ROUND((distance_m_next / EXTRACT(EPOCH FROM (tm_next - tm)))::numeric, 1) AS speed_next
        FROM
            gr1
    )
SELECT
    event_id
```

```
          , device_id
          , tm
          , tm_prev
          , tm_next
          , d_time_prev_s
          , d_time_next_s
          , distance_m_prev
          , distance_m_next
          , speed_prev
          , speed_next
          , CASE
                    WHEN (speed_prev > 55 AND distance_m_prev > 10000 AND speed_next > 55 AND
distance_m_next > 10000) THEN 3
                    WHEN COALESCE(speed_prev > 55 AND distance_m_prev > 10000, false) OR
COALESCE(speed_next > 55 AND distance_m_next > 10000, false) THEN 4
                    ELSE 0
          END AS error_flag
FROM
          gr2
ORDER BY
          device_id
          , tm
;
```

11.7 EXAMPLES

\ VALID REFERENCE TO CELL_ID

MNO Network Topology data:

cell_id	date
21401111	2023-01-01
21401222	2023-01-01
21401444	2023-01-02

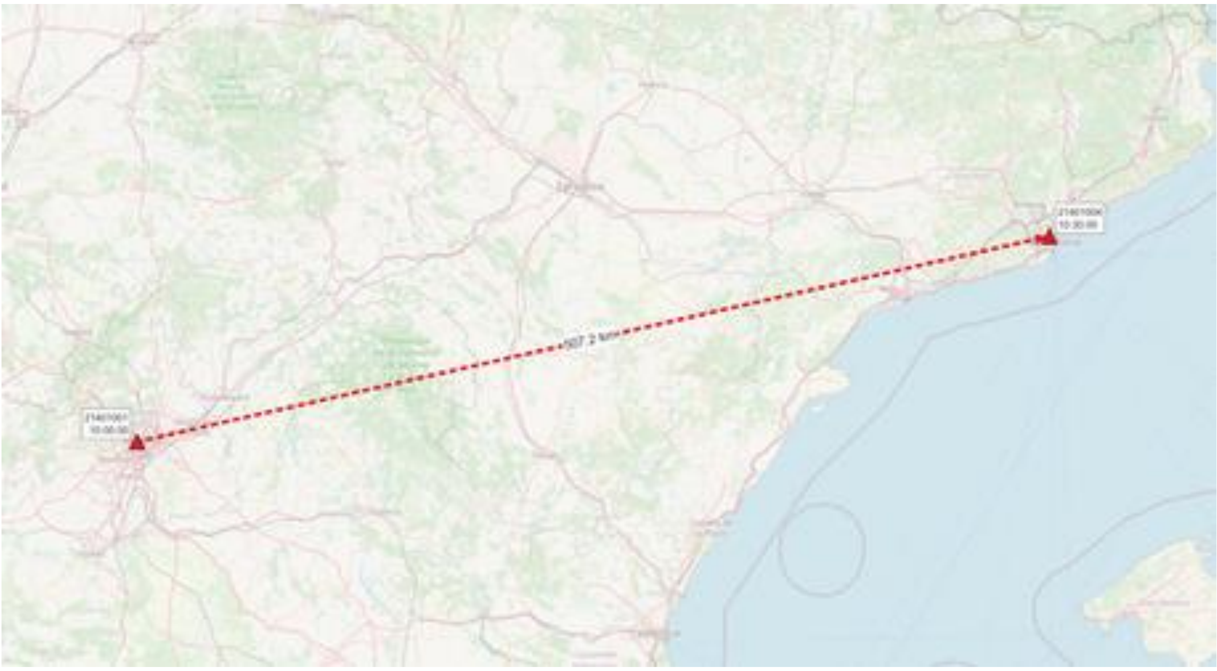
MNO Event Data table from Event Data at Device level [INTERMEDIATE RESULTS]. The third and fourth event must be flagged because third event cell_id does not exist in the MNO Network Topology data. The fourth does exist, but is not valid on the specific date of the event.

year	month	day	user_id	timestamp	mcc	mnc	cell_id	error_flag
2023	1	1	000000000000..11	00:00:00	214	67	21401111	
2023	1	1	000000000000..11	00:01:15	214	299	21401222	
2023	1	1	000000000000..11	12:05:03	214	299	**21401333**	1
2023	1	1	000000000000..11	12:05:05	214	299	**21401444**	2

\ ILLOGICAL CHANGE OF THE LOCATION OF THE DEVICE BASED ON THE TIME AND DISTANCE DIFFERENCE

Below we introduce a couple of examples of this issue that should be handled by the methodology:

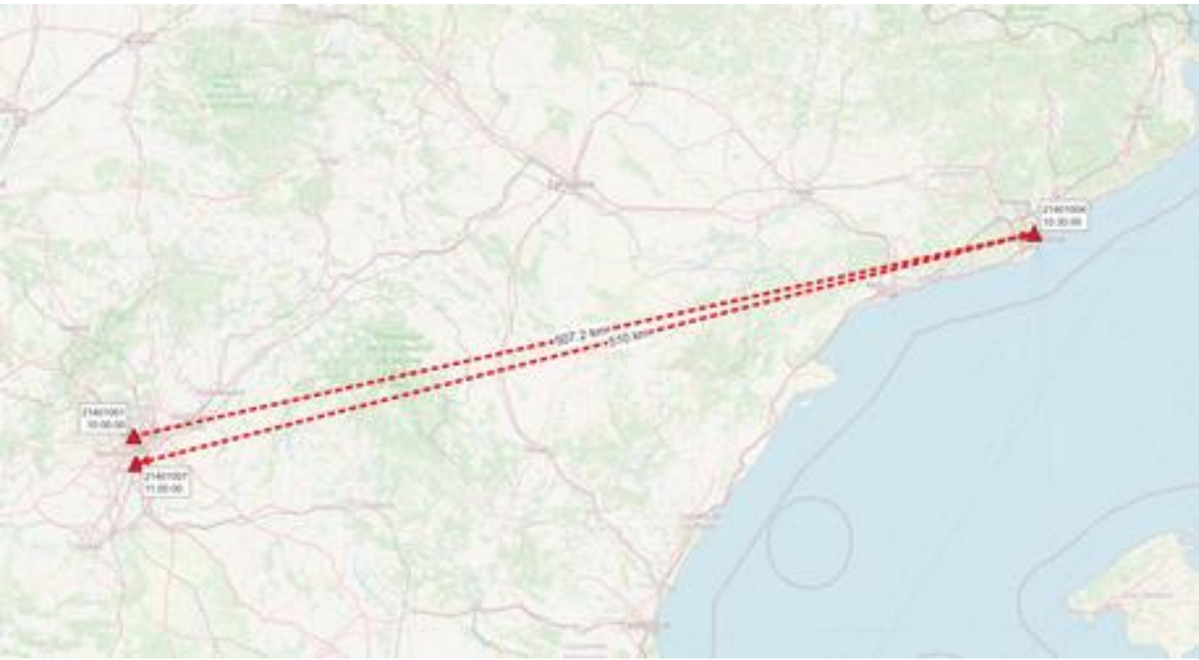
A. Two isolated events are too far away from each other to be realistic. There is no basis to identify which event is incorrect, so both events should be flagged as erroneous.



MNO events data table

event_id	device_id	time	cell_id	***Location***
1	A	2023-01-01 10:00:00	21401001	*Madrid*
2	A	2023-01-01 10:30:00	21401004	*Barcelona*

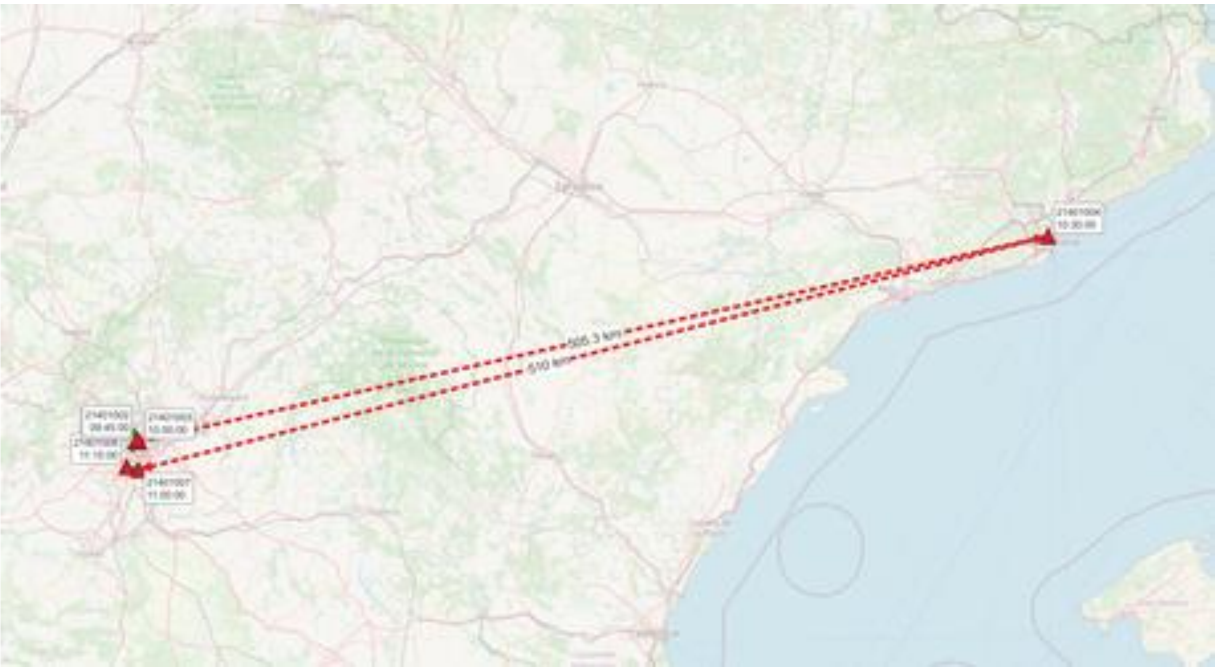
B. Consecutive events where some events are clearly erroneous probably due to the incorrect location data of the network topology (wrong coordinates of the location of the cell), as it is too far away from other events to be realistic. In this situation there is no clear indication which events are incorrect, so all should be flagged as erroneous.

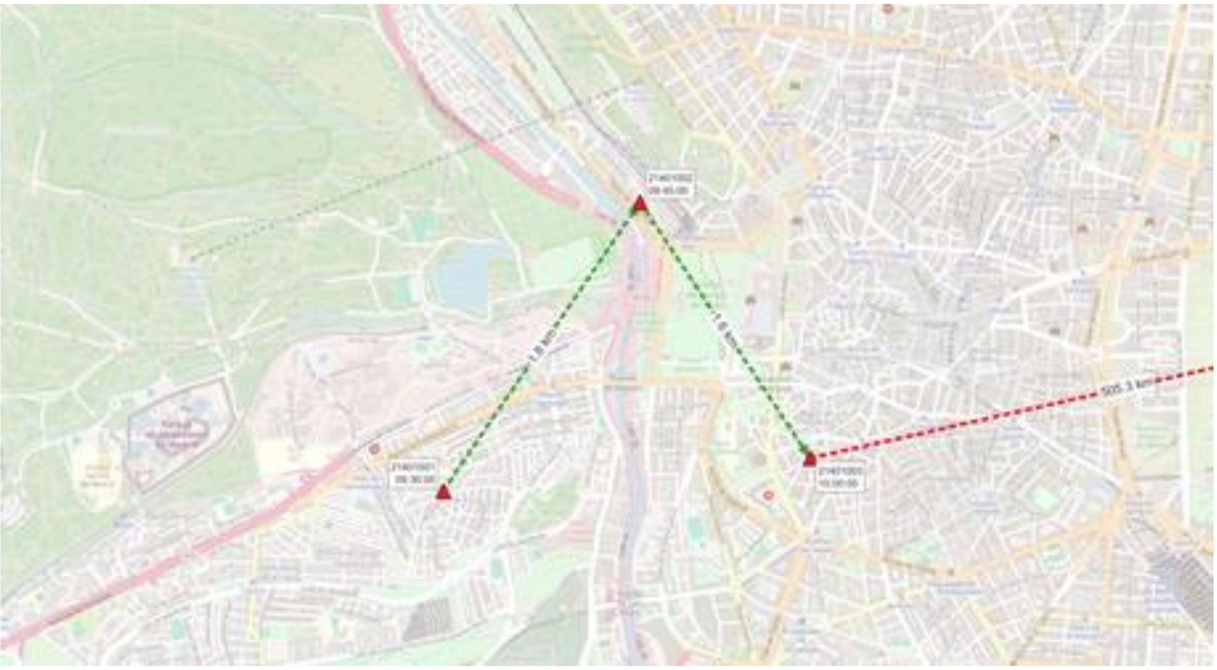


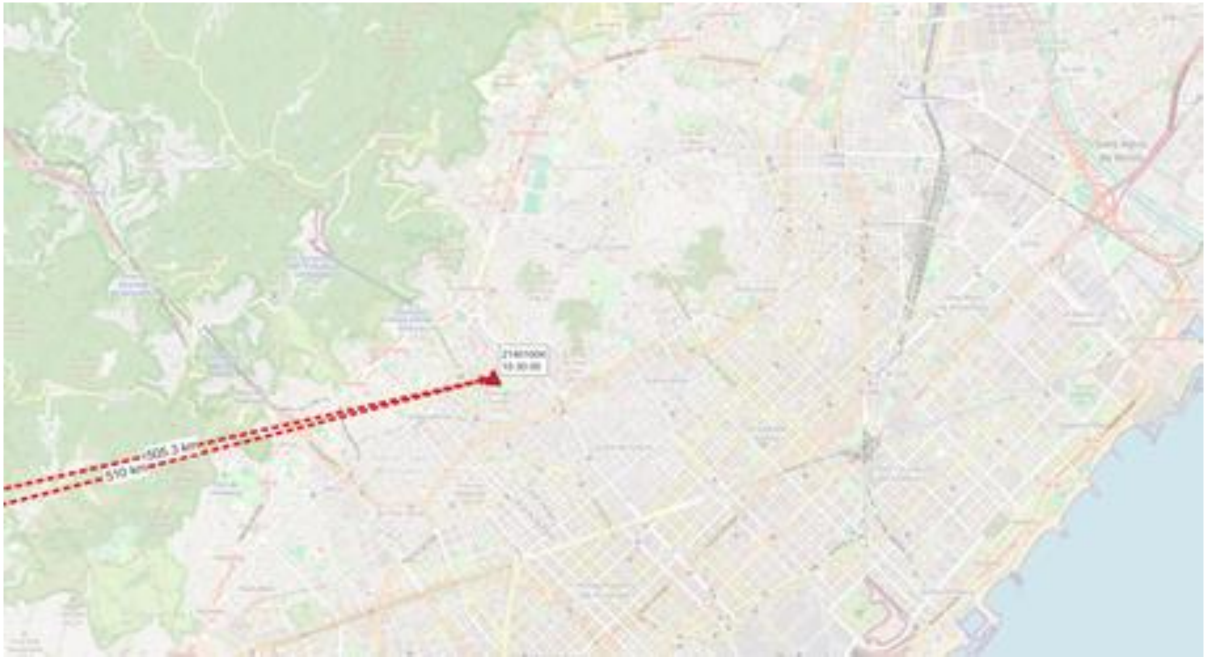
MNO events data table

event_id	device_id	time	cell_id	***Location***
3	B	2023-01-01 10:00:00	21401001	*Madrid*
4	B	2023-01-01 10:30:00	21401004	*Barcelona*
5	B	2023-01-01 11:00:00	21401007	*Madrid*

C. Consecutive events where only one event is clearly erroneous, probably due to the incorrect location data of the network topology (wrong coordinates of the location of the cell), as it is too far away from other events to be realistic. In this situation, the event at 10:30 should be flagged as erroneous, other events are not flagged.


21401002
09:45:00
21401003
10:00:00
21401008
11:10:00
21401007
11:00:00
21401004
10:30:00
505.3 km
510 km


21401002
09:45:00
1.8 km
1.6 km
21401001
09:30:00
21401003
10:00:00
505.3 km

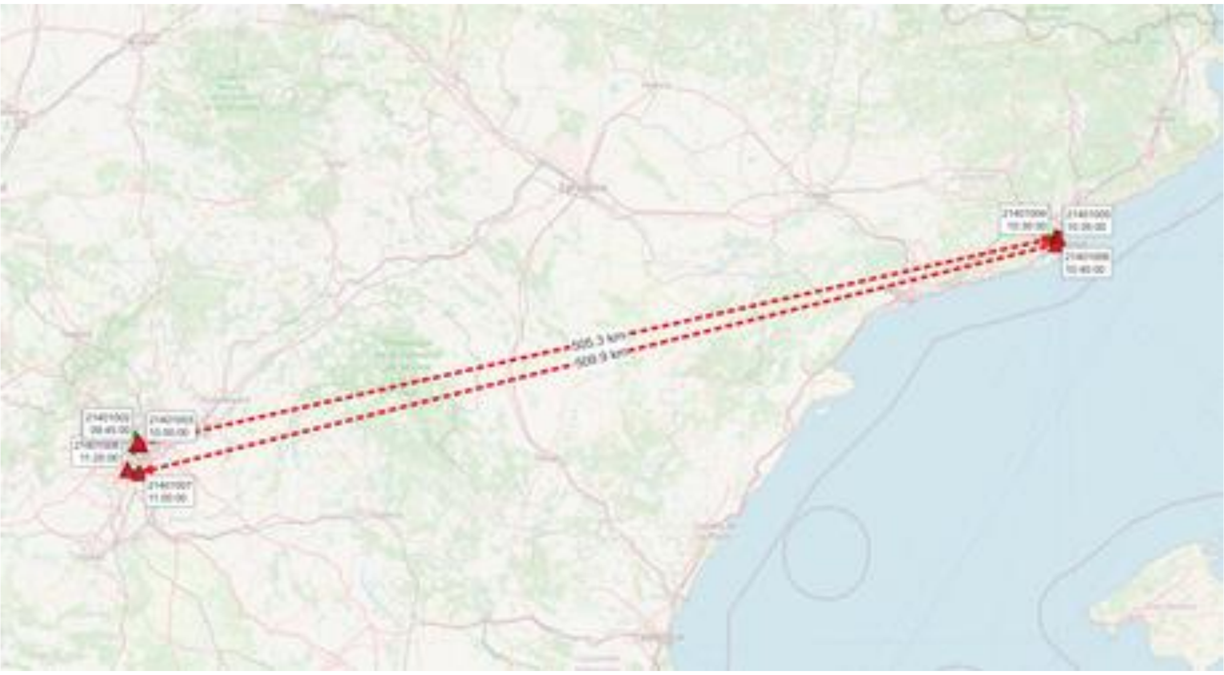

21401006
10:30:00
~505.3 km
510 km

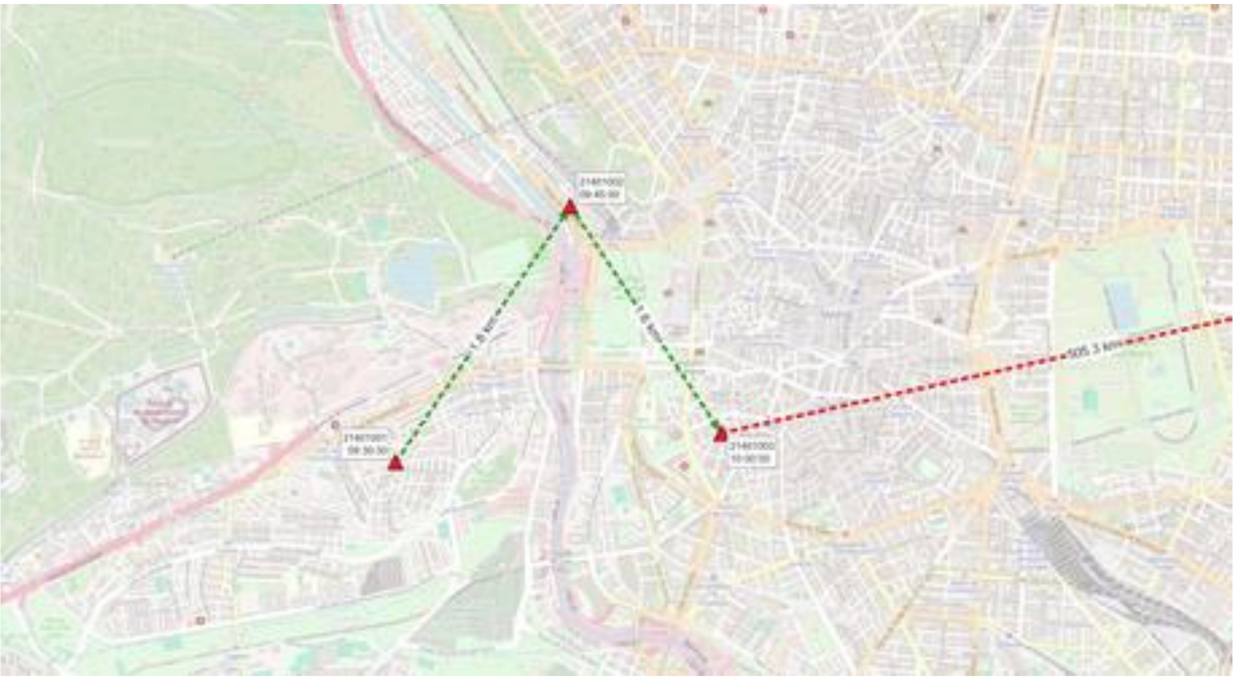
21401008
11:15:00
~6.8 km
21401007
11:00:00
~510 km

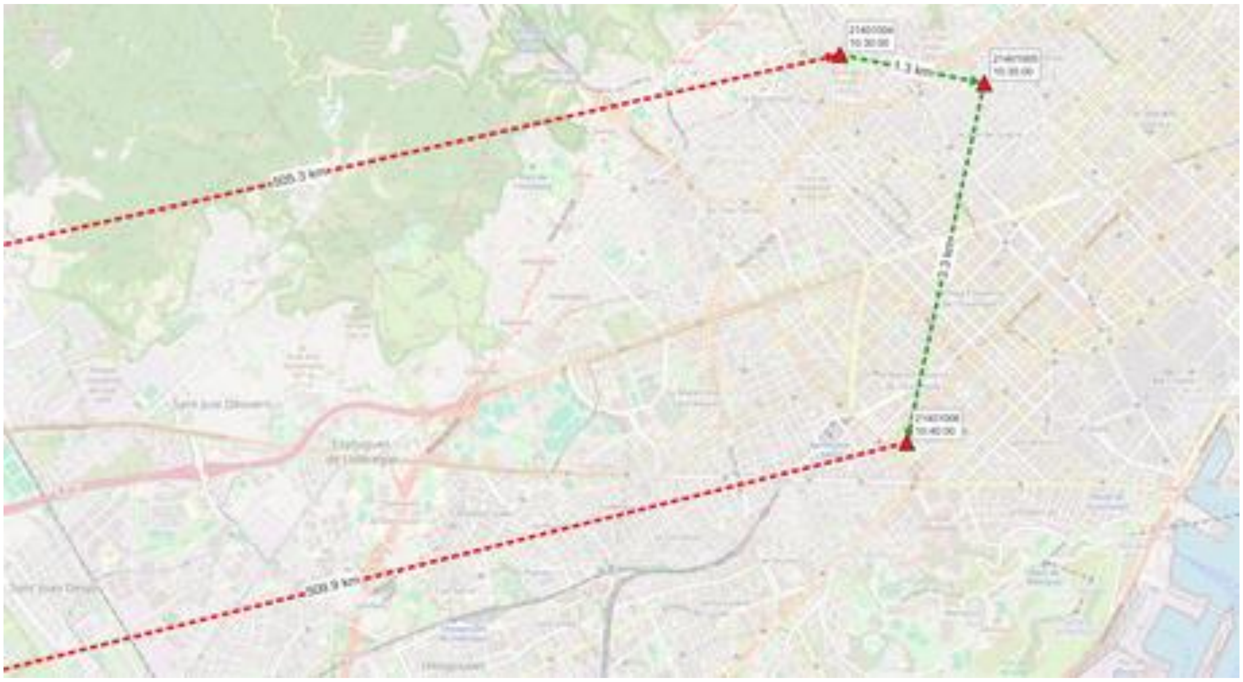
MNO events data table.

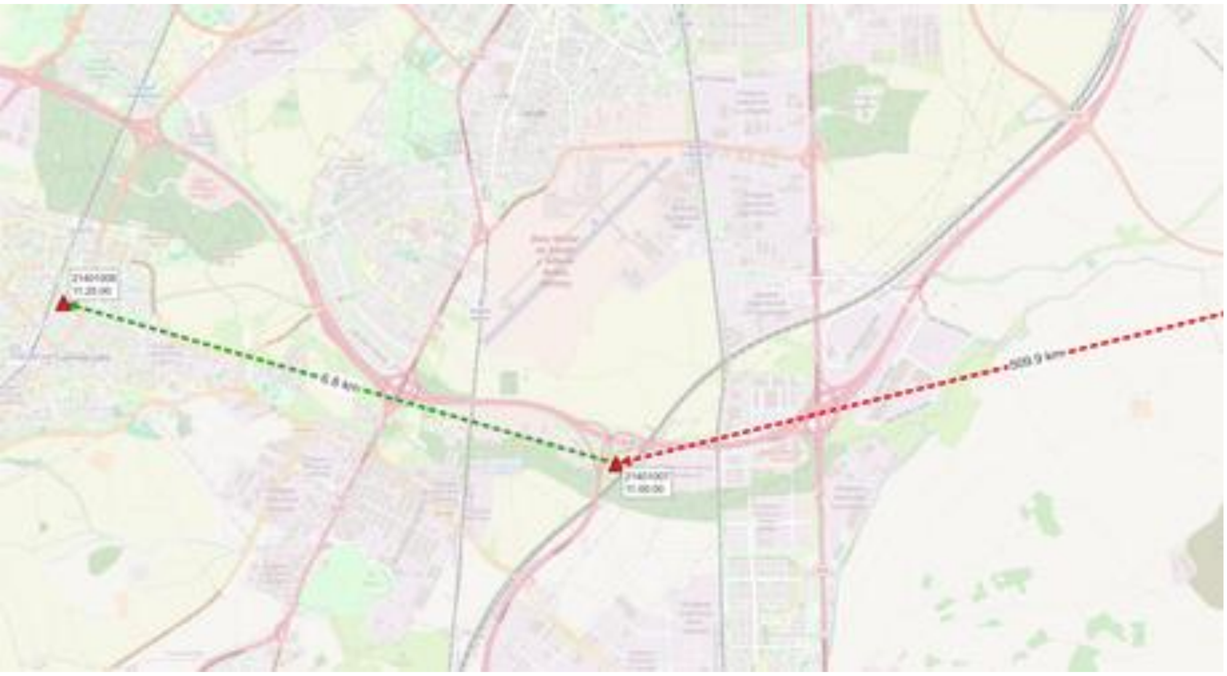
event_id	device_id	time	cell_id	*Location*
6	C	2023-01-01 09:30:00	21401001	*Madrid*
7	C	2023-01-01 09:45:00	21401002	*Madrid*
8	C	2023-01-01 10:00:00	21401003	*Madrid*
9	C	2023-01-01 10:30:00	21401004	*Barcelona*
10	C	2023-01-01 11:00:00	21401007	*Madrid*
11	C	2023-01-01 11:15:00	21401008	*Madrid*

D. Consecutive events where several events are clearly erroneous, probably due to the incorrect location data of the network topology (wrong coordinates of the location of the cells), as it is too far away from other events to be realistic. In this situation, the events at 10:30, 10:35, 10:40 should be flagged as erroneous, other events are not flagged.




21461002
09:45:00
1.6 km
1.6 km
21461001
09:30:00
21461000
10:00:00
505.3 km


10:30:00
1.3 km
10:35:00
2.3 km
10:40:00
505.3 km
509.9 km
Sant Just Desvern
Esplugues de Llobregat
Sant Joan Despí



MNO events data table:

event_id	device_id	time	cell_id	***Location***
12	D	2023-01-01 09:30:00	21401001	*Madrid*
13	D	2023-01-01 09:45:00	21401002	*Madrid*
14	D	2023-01-01 10:00:00	21401003	*Madrid*
15	D	2023-01-01 10:30:00	21401004	*Barcelona*
16	D	2023-01-01 10:35:00	21401005	*Barcelona*
17	D	2023-01-01 10:40:00	21401006	*Barcelona*
18	D	2023-01-01 11:00:00	21401007	*Madrid*
19	D	2023-01-01 11:25:00	21401008	*Madrid*

The following examples use the previous cases (A, B, C, D) as device_id's to illustrate the methodological process.

MNO Network Topology data (location here is marked for clarity purpose and is not part of the data object):

cell_id	date	lat	lon	* Location
21401001	2023-01-01	40.40896	-3.73374	*Madrid*
21401002	2023-01-01	40.42232	-3.72177	*Madrid*
21401003	2023-01-01	40.41047	-3.71145	*Madrid*
21401004	2023-01-01	41.40954	2.14124	*Barcelona*
21401005	2023-01-01	41.40724	2.15666	*Barcelona*
21401006	2023-01-01	41.37835	2.14836	*Barcelona*
21401007	2023-01-01	40.27125	-3.71865	*Madrid*
21401008	2023-01-01	40.28822	-3.79498	*Madrid*

MNO Event Data:

event_id	device_id	datetime	cell_id
1	A	2023-01-01 10:00:00	21401001
2	A	2023-01-01 10:30:00	21401004
3	B	2023-01-01 10:00:00	21401001
4	B	2023-01-01 10:30:00	21401004
5	B	2023-01-01 11:00:00	21401007
6	C	2023-01-01 09:30:00	21401001
7	C	2023-01-01 09:45:00	21401002
8	C	2023-01-01 10:00:00	21401003
9	C	2023-01-01 10:30:00	21401004
10	C	2023-01-01 11:00:00	21401007
11	C	2023-01-01 11:15:00	21401008
12	D	2023-01-01 09:30:00	21401001
13	D	2023-01-01 09:45:00	21401002
14	D	2023-01-01 10:00:00	21401003
15	D	2023-01-01 10:30:00	21401004
16	D	2023-01-01 10:35:00	21401005
17	D	2023-01-01 10:40:00	21401006
18	D	2023-01-01 11:00:00	21401007
19	D	2023-01-01 11:25:00	21401008

Joined MNO Events Data and MNO Network Topology coordinates (location here is marked for clarity purpose and is not part of the data object):

event_id	device_id	datetime	cell_id	lat	lon	* Location
1	A	2023-01-01 10:00:00	21401001	40.40896	-3.73374	*Madrid*
2	A	2023-01-01 10:30:00	21401004	41.40954	2.14124	*Barcelona*
3	B	2023-01-01 10:00:00	21401001	40.40896	-3.73374	*Madrid*
4	B	2023-01-01 10:30:00	21401004	41.40954	2.14124	*Barcelona*
5	B	2023-01-01 11:00:00	21401007	40.27125	-3.71865	*Madrid*
6	C	2023-01-01 09:30:00	21401001	40.40896	-3.73374	*Madrid*

event_id	device_id	datetime	cell_id	lat	lon	* Location
7	C	2023-01-01 09:45:00	21401002	40.42232	-3.72177	*Madrid*
8	C	2023-01-01 10:00:00	21401003	40.41047	-3.71145	*Madrid*
9	C	2023-01-01 10:30:00	21401004	41.40954	2.14124	*Barcelona*
10	C	2023-01-01 11:00:00	21401007	40.27125	-3.71865	*Madrid*
11	C	2023-01-01 11:15:00	21401008	40.28822	-3.79498	*Madrid*
12	D	2023-01-01 09:30:00	21401001	40.40896	-3.73374	*Madrid*
13	D	2023-01-01 09:45:00	21401002	40.42232	-3.72177	*Madrid*
14	D	2023-01-01 10:00:00	21401003	40.41047	-3.71145	*Madrid*
15	D	2023-01-01 10:30:00	21401004	41.40954	2.14124	*Barcelona*
16	D	2023-01-01 10:35:00	21401005	41.40724	2.15666	*Barcelona*
17	D	2023-01-01 10:40:00	21401006	41.37835	2.14836	*Barcelona*
18	D	2023-01-01 11:00:00	21401007	40.27125	-3.71865	*Madrid*
19	D	2023-01-01 11:25:00	21401008	40.28822	-3.79498	*Madrid*

After the methodology is applied, the results can be seen in the following result table with the following parameters: `semantic_min_distance_m=10000 (10 km)` and `semantic_min_speed_m_s=55` (198km/h)).

As it can be seen, for device D, no events were flagged as certain error and one highly suspicious event is no flagged at all, though, there are obviously issues with this case. Current method may be enhanced using more sophisticated methods in the future releases.

event_id	device_id	datetime	*Location	d_time_prev_s	d_time_next_s	distance_m_prev	distance_m_next	speed_prev	speed_next	error_flag
1	A	2023-01-01 10:00:00.000	Madrid		1800		505136.0		280.6	4
2	A	2023-01-01 10:30:00.000	Barcelona	1800		505136.0		280.6		4
3	B	2023-01-01 10:00:00.000	Madrid		1800		505136.0		280.6	4
4	B	2023-01-01 10:30:00.000	Barcelona	1800	1800	505136.0	515590.0	280.6	286.4	3
5	B	2023-01-01 11:00:00.000	Madrid	1800		515590.0		286.4		4
6	C	2023-01-01 09:30:00.000	Madrid		900		214.0		0.2	0
7	C	2023-01-01 09:45:00.000	Madrid	900	900	214.0	194.0	0.2	0.2	0
8	C	2023-01-01 10:00:00.000	Madrid	900	1800	194.0	505305.0	0.2	280.7	4
9	C	2023-01-01 10:30:00.000	Barcelona	1800	1800	505305.0	515590.0	280.7	286.4	3
10	C	2023-01-01 11:00:00.000	Madrid	1800	900	515590.0	142.0	286.4	0.2	4
11	C	2023-01-01 11:15:00.000	Madrid	900		142.0		0.2		0
12	D	2023-01-01 09:30:00.000	Madrid		900		214.0		0.2	0
13	D	2023-01-01 09:45:00.000	Madrid	900	900	214.0	194.0	0.2	0.2	0

event_id	device_id	datetime	*Location	d_time_prev_s	d_time_next_s	distance_m_prev	distance_m_next	speed_prev	speed_next	error_flag
14	D	2023-01-01 10:00:00.000	Madrid	900	1800	194.0	505305.0	0.2	280.7	4
15	D	2023-01-01 10:30:00.000	Barcelona	1800	300	505305.0	141.0	280.7	0.5	4
16	D	2023-01-01 10:35:00.000	Barcelona	300	300	141.0	167.0	0.5	0.6	0
17	D	2023-01-01 10:40:00.000	Barcelona	300	1200	167.0	515622.0	0.6	429.7	4
18	D	2023-01-01 11:00:00.000	Madrid	1200	1500	515622.0	142.0	429.7	0.1	4
19	D	2023-01-01 11:25:00.000	Madrid	1500		142.0		0.1		0

12 MODULE/METHOD 11: DEVICE ACTIVITY STATISTICS

12.1 OBJECTIVE

This module runs after Module/Method 10: Event Cleaning at Device Level – Semantic Checks.

This module uses the data on individual devices and produces metrics to assess the usability of these devices' data for specific procedures or use cases based on statistics on their activity. The metrics only concern the events related to the device. All metrics are computed per date. Depending on the use case, the inclusion/exclusion filter can use single or multiple dates (longer period) to assess whether the device is usable for the specific use case. This analysis will be applied in next steps of the pipeline.

Mapping with other standards: *the metrics in this method can be mapped to those in the GSBPM sub-process* 5.7 Calculate aggregates *sub-process of GSBPM.*

Correspondence with Deliverable D3.1: *in Table 2 of Chapter 6 this method corresponds to Quality Module QM.DevAct.*

12.2 PARAMETERS

None

12.3 INPUT DATA

- Semantically Cleaned Event Data at Device Level [INTERMEDIATE RESULTS] in this case avoid events with error_flags (list of flags to be avoided vs included).
    - Include error_flags: 0, 4
    - Exclude error_flags: 1,2,3
- Clean MNO Network Topology Data [INTERMEDIATE RESULTS]

12.4 OUTPUT DATA

- Device Activity Statistics [INTERMEDIATE RESULTS]

12.5 METHODOLOGY

All statistics calculated for the device are either aggregation of the events or parameters related to the events. The list of the indicators are presented in the following table along with explanations on the relevance of each indicator. Many of these indicators will be assessed over a longer period of time (combining, summarising, aggregating daily indicators) and they are use case specific. In the table below the exemplified methodology shows only how they are calculated for each date the device has been in the dataset. Each of the following metrics will be calculated by device and by date (in the local time zone).

METRIC	COMMENT	METHODOLOGY
Number of events per day	If number is too low, the device may not be usable for some use cases.	Count the number of the events of the device on specific date (NB! local time zone).
Number of unique cells per day.	If the number of unique cells is lower or equal to 1, then this suggests the device is not moving at all. If this indicator is lower or equal to 1 for many or all days, then this device may not be usable for some use cases.	Count unique (distinct) cell_id's of the events of the device on specific date (NB! local time zone).
Number of different locations per day (based on the location point of the cell).	If the number of unique locations is lower or equal to 1, then this suggests device is not moving at all. If this indicator is lower or equal to1 for many or all days, then this device may not be usable for some use cases, or may even be the basis for warning. NB! This is slightly different from the number of unique cells, as several cells may be located in one location.	Count unique (distinct) location coordinates extracted from the location of the cells of the events of the device on specific date (NB! local time zone).
Sum of the distances between the events (based on the location point of the cell).	This is a proxy indicator for the distance travelled during the day. NB! The distance is calculated based on the location of the events' cells, and does not indicate the real distances travelled. If this number is too low, in combination with the number of unique cells and/or the number of unique locations, this may suggest low-activity device or even M2M / IoT device. If the distance travelled is too long, this may be the basis for warning.	Calculate the distance between the consecutive event cells' locations of the device on specific date (NB! local time zone).
Number of unique hours in the date with events	This indicator suggests how active the device is. Normal number for signalling data and domestic devices is 24 (events are in all hours of the day). If this number is very low, this suggests the device is not active.	Extract hour value from each event, sum the number of unique hours in which there are events of the device on specific date (NB! local time zone).
Average time gap between events (in seconds)	Low value suggests the use of signalling data, high value suggests CDR data.	Calculate time gap in seconds between each event of the day of the device and calculate the average.
Standard deviation of the time gap between events (in seconds)	Combined with the average time gap, this indicator may be used to identify non-human devices.	Calculate time gap in seconds between each event of the day of the device and calculate the standard deviation.

For illustrating the metrics calculated, synthetic data files (cells and events) were used to generate the following device activity statistics table:

device_id	year	month	day	event_cnt	unique_cell_cnt	unique_location_cnt	sum_distance_m	unique_hour_cnt	mean_time_gap	stdev_time_gap
A	2023	1	1	12	10	10	45778	10	5090	2951.61
A	2023	1	2	8	2	2	7592	7	5118	3169.484
B	2023	1	1	12	10	10	45036	8	4358	3614.575
C	2023	1	1	11	1	1	0	10	5939	4039.195
C	2023	1	2	20	1	1	0	14	4173	3017.242
C	2023	1	3	12	1	1	0	10	7313	3111.024

device_id	year	month	day	event_cnt	unique_cell_cnt	unique_location_cnt	sum_distance_m	unique_hour_cnt	mean_time_gap	stdev_time_gap
C	2023	1	4	7	1	1	0	5	4062	1536.541
D	2023	1	1	112	80	80	1035035	9	276	163.491
E	2023	1	1	142	37	37	13083	2	28	17.225
F	2023	1	1	41	1	1	0	1	33	13.647
G	2023	1	1	24	13	13	51061	24	3600	0
H	2023	1	1	1	1	1	0	1	0	0

13 MODULE/METHOD 12: MNO EVENT DATA AT DEVICE LEVEL – SEMANTIC QUALITY WARNINGS

13.1 OBJECTIVE

This module analyses the Device Semantic Quality Metrics [INTERMEDIATE RESULTS] to identify anomalous situations that need to be further investigated. The output of the method is the MNO Event Data at Device Level Semantic Quality Warnings [INTERMEDIATE RESULTS] presenting plots of the metrics over time, anomalous data, the related warning and (whenever possible) the further investigation suggested.

Mapping with other standards: the semantic checks included in this method can be considered as included in the Overarching process Quality management of GSBPM.

Correspondence with Deliverable D3.1: in table 2 of chapter 6 this method corresponds to Quality Module QM.EvSem

13.2 INPUT DATA

- Device Semantic Quality Metrics [INTERMEDIATE RESULTS]
- previous periods Device Semantic Quality Metrics [INTERMEDIATE RESULTS]

13.3 OUTPUT DATA

- MNO Event Data at Device Level Semantic Quality Warnings [INTERMEDIATE RESULTS]

13.4 PARAMETERS

- error_warning_threshold_error_type_sd_lookback_d_1: number of lookback days for calculating standard deviation for error type 1 excluding the day under observation. Default should be 7 (days). This must be at least 3 (or the standard deviation is ambiguous).
- error_warning_threshold_error_type_min_sd_1: lower threshold of displaying a warning for error type 1 events in standard deviation (combined logically with error_warning_threshold_error_type_min_prc_1). Default should be 2 (2 times the standard deviation);
- error_warning_threshold_error_type_min_prc_1: percentage of error type 1 events to display warning (combined logically with error_warning_threshold_error_type_min_sd_1). Default 0 (% of the total events);
- error_warning_threshold_error_type_sd_lookback_d_2: number of lookback days for calculating standard deviation for error type 2 excluding the day under observation. Default should be 7 (days);

- error_warning_threshold_error_type_min_sd_2: lower threshold of displaying a warning for error type 2 events in standard deviation (combined logically with error_warning_threshold_error_type_min_prc_2). Default should be 2 (2 times the standard deviation);
- error_warning_threshold_error_type_min_prc_2: percentage of error type 2 events to display warning (combined logically with error_warning_threshold_error_type_min_sd_2). Default 0 (% of the total events).
- error_warning_threshold_error_type_sd_lookback_d_3: number of lookback days for calculating standard deviation for error type 3 excluding the day under observation. Default should be 7 (days);
- error_warning_threshold_error_type_min_sd_3: lower threshold of displaying a warning for error type 3 events in standard deviation (combined logically with error_warning_threshold_error_type_min_prc_3). Default should be 2 (2 times the standard deviation);
- error_warning_threshold_error_type_min_prc_3: percentage of error type 3 events to display warning (combined logically with error_warning_threshold_error_type_min_sd_3). Default 0 (% of the total events).
- error_warning_threshold_error_type_sd_lookback_d_4: number of lookback days for calculating standard deviation for error type 4 excluding the day under observation. Default should be 7 (days);
- error_warning_threshold_error_type_min_sd_4: lower threshold of displaying a warning for error type 4 events in standard deviation (combined logically with error_warning_threshold_error_type_min_prc_4). Default should be 2 (2 times the standard deviation);
- error_warning_threshold_error_type_min_prc_4: percentage of error type 4 events to display warning (combined logically with error_warning_threshold_error_type_min_sd_4). Default 0 (% of the total events).

13.5 METHODOLOGY

1. Calculate the percentage of each error type on the specific time period from the total (total events = all error type events (including error 0)). This to be displayed on a plot
2. Calculate the standard deviation over the lookback days for each error type of the percentage;
3. If the percentage of a specific type of error of the specific day is over the upper control limit (= average of percentages of errors over the lookback period + error_warning_threshold_error_type_min_sd_1 * standard deviation) then display warning for the specific date. NB! In case the data period is shorter than the lookback days and standard deviation is either not possible to calculate, or is too short (e.g., <3 days of data only available), then dismiss the standard deviation threshold and use percentage threshold only;
4. Display the list of cell_ids corresponding to events flagged as errors with the type of error.

13.6 PSEUDO-CODE

```
IF((error_value_code_X / total) > AVERAGE(lookback_d)+(STDEV(lookback_d)*2))
THEN DISPLAY WARNING
```

13.7 EXAMPLES

Example data from Device Semantic Quality Metrics [INTERMEDIATE RESULTS]

data_period_start	type_of_error	value
2023-07-10	0	139867548
2023-07-10	1	31
2023-07-10	2	18275

data_period_start	type_of_error	value
2023-07-10	3	15171
2023-07-10	4	9373
2023-07-11	0	141279294
2023-07-11	1	41
2023-07-11	2	16369
2023-07-11	3	24656
2023-07-11	4	6894
2023-07-12	0	141077339
2023-07-12	1	23
2023-07-12	2	20807
2023-07-12	3	35847
2023-07-12	4	9129
2023-07-13	0	138953746
2023-07-13	1	8
2023-07-13	2	18469
2023-07-13	3	27154
2023-07-13	4	6453
2023-07-14	0	139271472
2023-07-14	1	27
2023-07-14	2	20958
2023-07-14	3	31397
2023-07-14	4	9421
2023-07-15	0	139953901
2023-07-15	1	32
2023-07-15	2	19334
2023-07-15	3	33295
2023-07-15	4	9230
2023-07-16	0	139301690
2023-07-16	1	46
2023-07-16	2	17432
2023-07-16	3	39002
2023-07-16	4	8803
2023-07-17	0	140520438
2023-07-17	1	21
2023-07-17	2	16743
2023-07-17	3	12784
2023-07-17	4	7153
2023-07-18	0	140260429
2023-07-18	1	22
2023-07-18	2	19037
2023-07-18	3	11963
2023-07-18	4	9652
2023-07-19	0	140272168
2023-07-19	1	4
2023-07-19	2	17481
2023-07-19	3	7894
2023-07-19	4	9628

data_period_start	type_of_error	value
2023-07-20	0	139851010
2023-07-20	1	3
2023-07-20	2	18622
2023-07-20	3	14821
2023-07-20	4	9279
2023-07-21	0	98015873
2023-07-21	1	34599296
2023-07-21	2	20987
2023-07-21	3	6532564
2023-07-21	4	6938
2023-07-22	0	140310366
2023-07-22	1	34
2023-07-22	2	4683323
2023-07-22	3	16665
2023-07-22	4	5293752
2023-07-23	0	139378787
2023-07-23	1	24
2023-07-23	2	18519
2023-07-23	3	14955
2023-07-23	4	7173

Percentage of errors, standard deviation * 2 threshold (7 days lookback) and percentage of error type set to 1%:

Date	Error 1	Error 2	Error 3	Error 4	Error 1 avg + sd*2	Error 2 avg + sd*2	Error 3 avg + sd*2	Error 4 avg + sd*2	Display warning	Display warning	Display warning	Display warning
2023-07-10	0.000022%	0.013062%	0.010843%	0.006699%					FALSE	FALSE	FALSE	FALSE
2023-07-11	0.000029%	0.011582%	0.017446%	0.004878%					FALSE	FALSE	FALSE	FALSE
2023-07-12	0.000016%	0.014742%	0.025398%	0.006468%					FALSE	FALSE	FALSE	FALSE
2023-07-13	0.000006%	0.013286%	0.019534%	0.004642%					FALSE	FALSE	FALSE	FALSE
2023-07-14	0.000019%	0.015042%	0.022534%	0.006761%					FALSE	FALSE	FALSE	FALSE
2023-07-15	0.000023%	0.013808%	0.023779%	0.006592%					FALSE	FALSE	FALSE	FALSE
2023-07-16	0.000033%	0.012508%	0.027985%	0.006316%					FALSE	FALSE	FALSE	FALSE
2023-07-17	0.000015%	0.011912%	0.009095%	0.005089%	0.000038%	0.015686%	0.031660%	0.007711%	FALSE	FALSE	FALSE	FALSE
2023-07-18	0.000016%	0.013569%	0.008527%	0.006879%	0.000037%	0.015761%	0.032400%	0.007504%	FALSE	FALSE	FALSE	FALSE
2023-07-19	0.000003%	0.012459%	0.005626%	0.006862%	0.000034%	0.015630%	0.033952%	0.007731%	FALSE	FALSE	FALSE	FALSE
2023-07-20	0.000002%	0.013312%	0.010594%	0.006633%	0.000035%	0.015166%	0.033058%	0.007859%	FALSE	FALSE	FALSE	FALSE
2023-07-21	24.860163%	0.015080%	4.693755%	0.004985%	0.000036%	0.015171%	0.032098%	0.007613%	TRUE	FALSE	TRUE	FALSE
2023-07-22	0.000023%	3.115898%	0.011088%	3.522027%	20.949944%	0.015196%	3.957762%	0.007699%	FALSE	TRUE	FALSE	TRUE

Date	Error 1	Error 2	Error 3	Error 4	Error 1 avg + sd*2	Error 2 avg + sd*2	Error 3 avg + sd*2	Error 4 avg + sd*2	Display warning	Display warning	Display warning	Display warning
2023-07-23	0.000017%	0.013283%	0.010727%	0.005145%	20.949944%	2.627870%	3.957420%	2.969016%	FALSE	FALSE	FALSE	FALSE

14 MODULE 13: DAILY PROCESSING MODULE

14.1 METHOD 1: PRESENT POPULATION ESTIMATION

14.1.1 OBJECTIVE AND METHODOLOGY

The Present Population refers to the number of people present in a specific geographical area at a particular fixed moment in time. For instance, the number of people in Paris on Friday 14th of July 2023 at 2 p.m.

We focus on the production of the present population statistics for a specific country[9]. The spatial resolution of the output can either be a regular grid (e.g. INSPIRE) or a zoning system of interest (e.g. administrative divisions such as municipalities).

Special attention is needed for the country border regions. The mobile network coverage of an MNO usually crosses the borders to a certain extent (typically a few kilometres). As a result, it is expected that part of the devices registered at cells around the country borders are located abroad, and vice versa: part of the devices registered at foreign cells close to the border are expected to be in the country of interest.

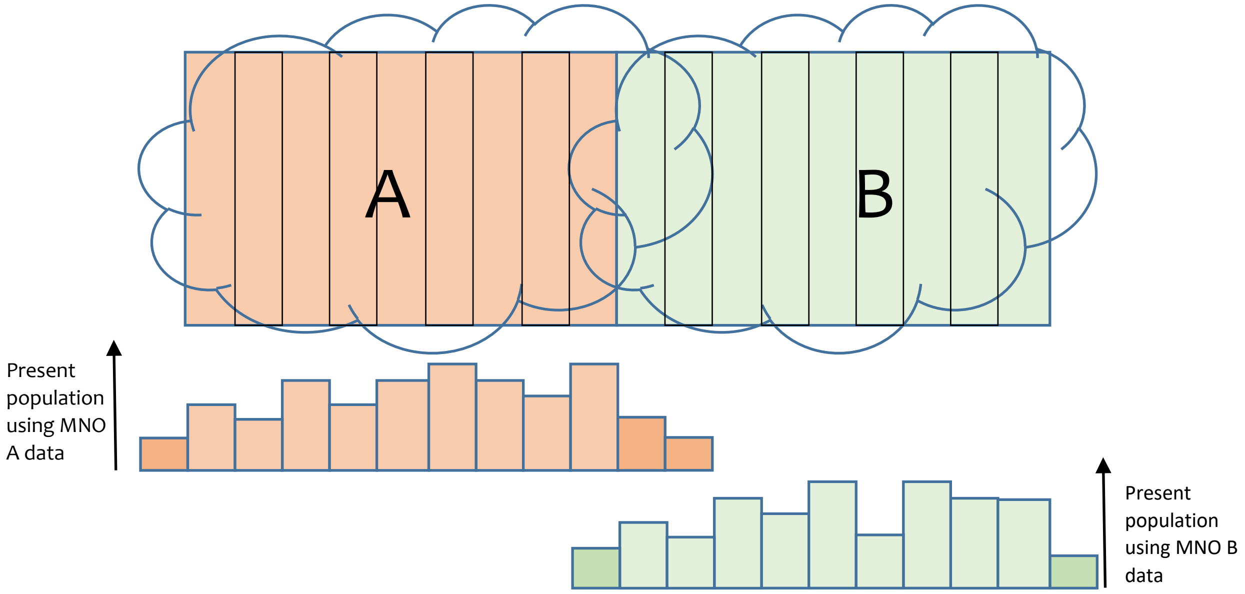

A
B
Present population using MNO A data
Present population using MNO B data

Figure 7: Present population in a cross-border region

[9] Reference work considered for this use case application include:
Laan, J van der, Jonge, E. de (2019) Maximum likelihood reconstruction of population densities from mobile signalling data. In NetMob'19, 2019.
Shepp, L. and Vardi, Y. (1982) Maximum likelihood reconstruction for emission tomography. IEEE Transactions on Medical Imaging, 1982.

The situation of two neighbouring countries, A and B, is depicted in **FIGURE 7.** For simplicity, let us assume there is only one MNO in each country. The clouds represent their network coverages. For simplicity, we use a one-dimensional consisting of 9 columns in each country. Bar charts of the present population estimates at a specific time using MNO data from countries A and B are depicted in orange and green respectively. Dark orange/green represent estimates for foreign grid tiles.

For estimating the present population of country A, not only the MNO data of A is required, but also the MNO data of B, at least in the optimal case. Otherwise, an underestimation is expected in the regions near the borders. This can be seen in **FIGURE 8**.

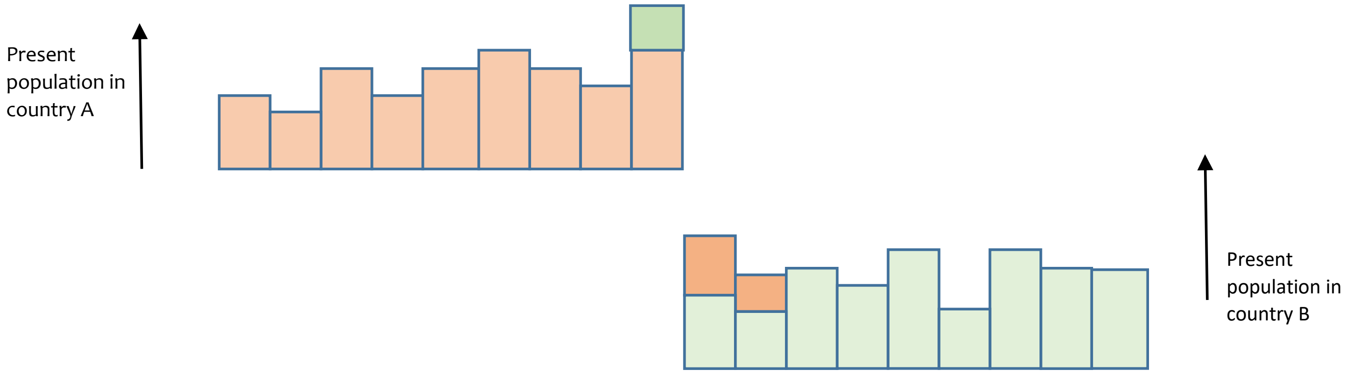


Figure 8: Present population per country

Required configuration data

- A set of grid tiles of the INSPIRE 100x100 m grid that covers the country of interest, including a large buffer (e.g. 50 km) around the borders. We call this set G.
- A zoning specification (tessellation), for which the present population is estimated. Each zone consists of a subset of grid tiles. These subsets do not overlap. For simplicity, we assume that each grid tile belongs to one zone, even though zone borders may cross the tiles. The present population may also be produced for the 100x100 m or 1 km x 1 km grid tiles rather than the zones, but the zones are required for the weighting method. We will use municipalities throughout this section.
- The date/timestamps for which the present population is estimated.

14.1.2 SPATIAL CELL INFORMATION

The MNO network topology data is processed in the standard way. The footprint and cell connection probabilities are estimated for all cells and grid tiles that are included in G.

This UC only requires the cell connection probabilities. Note that the home location estimation is required, but not included in the description of this application. For the home location estimation other components from the spatial cell information block may be needed.

14.1.3 PSEUDO-CODE

Let t be the timestamp for which the present population is estimated. We assume that timestamps are encoded as numerical variables, so t1 > t2 means that t1 is later than t2.

In addition, when y.x is used rather than y[i].x, then the applied function is iterated over all i.

Let m be the number of cells.

Let GAP be the maximum allowed time gap for which the device is included in the daily summary. In other words, the device is included in the output if there is at least one event in the time window [t – GAP, t + GAP]. More precisely, the daily summary table is filled with records such that for each device k, each cell A, and each target timestamp t a record is included if and only if at least one event of device k is included in the event data for which the connected cell is A and for which its timestamp is in the time window [t – GAP, t + GAP].

The daily summary table is shown in **TABLE 2** and the pseudocode to create this table is the following:

```
event = eventData.get(k)

event.timeDiff = ABS(event.time - t)

id = WHICH.MIN(event.timeDiff)

IF event[id].timeDiff <= GAP

  table.addRow(event[id].cell, k)

END IF
```

Table 2: Example of the daily summary table

Device ID	Cell ID	day	time
1	1	2023-07-14	2pm
2	1	2023-07-14	2pm
3	2	2023-07-14	2pm
4	3	2023-07-14	2pm
5	3	2023-07-14	2pm

14.1.4 DEVICE FILTERING, AGGREGATION AND GROSSING UP: BASIC COUNTING

There are various methods that can be applied for this block. Herewith, we present the one that is preferred for implementation, due to its consistency/alignment with subsequent methods and its suitability for application to other use cases described in Volume II. The method is the basic counting of devices. A second method, which requires home location estimates is introduced for illustrative purposes of potential variations as Annex II – Present Population Estimation: Variant 2. Note that these methods are just illustrations how to apply the pipeline to produce statistics on the present population. The pipeline will also fill more sophisticated methods.

The required input for this module consists of:

- The daily summaries as described above.

Let m be the number of cells. Let K be the set of unique devices in the event data.

The daily summary table is aggregated to a count table as follows. For each timestamp t and each cell A the number of records in the daily summary is counted. In other words, the number of unique devices is counted that are connected to A during timestamp t (or more precisely, in the time window [t – GAP, t + GAP]).

In pseudo code:

```
DECLARE counts[m]

FOR i = 1 TO m

  counts[i] = 0

END FOR

FOR k in K

 counts[k.cell] += 1

END FOR
```

Here, k.cell the cell id is for device k from the daily summary table.

The output is contained in the following table.

Table 3: Example of the counts table

cell ID	count	day	time
1	25436	2023-07-14	2pm
2	5342	2023-07-14	2pm
3	304334	2023-07-14	2pm
4	145755	2023-07-14	2pm

14.2 METHOD 2: DAILY PERMANENCE SCORE ESTIMATION

14.2.1 OBJECTIVE

The objective of this method is to integrate information in the Module/Method 10: Event Cleaning at Device Level – Semantic Checks with the Cell Footprint Data Object (Output of the Cell Footprint estimation) and to generate as output a *permanence* score for each tile of the reference grid and for each time slot.

The permanence score is a proxy of the time spent by the device, in each grid tile of the reference grid system, conveniently discretised.

The generated score is not a measure of time. In fact, for each event the permanence score is equally assigned to all grid tiles associated to the corresponding cell footprint. Hence, the total permanence score, obtained by summing up each score over all the tiles of the grid, do not correspond to the observation time. However, for a single tile, the daily permanence score cannot exceed the observation time.

The permanence score is an intensive quantity over the space (similar to a density permanence score) and it is meant to refer only to stay time.

The day is partitioned in time slots of equal duration. By default, we suggest 24 time slots of 1-hour duration; however, the user can choose a different number of time slots for the daily summary configuring the dedicated parameter.

The permanence score is meant to detect stays. As the focus is solely on stays, the method seeks to 'filter out' movements (non-stays), and this in turn requires considering the relation between geographical distance and inter-arrival time between groups of consecutive events. other use cases, a different distance function between cells could be used, developing another daily processing method or with limited revisions of this one.

The method is executed per single days and single devices.

14.2.2 PARAMETERS

- **Time_slot_number** (n in the text): number of time slots for the daily summary; each the time slot duration is 24h/n. Default value: 24 (the time slot length is 1h)
- **Max_time_thresh** (T*sd* in the text): maximum permanence threshold used to assign the permanence score in the case of successive events taking place in different places, e.g. T*sd* = 15min.
- **Max_time_thresh_day** (Tsc_1 in the text): daytime maximum permanence threshold used to assign the permanence score in the case of successive events taking place in the same cell; default value. Tsc_1 = 2 h.
- **Max_time_thresh_night** (Tsc_2 in the text): nighttime maximum permanence threshold used to assign the permanence score in the case of successive events taking place in the same cell during the night; default value Tsc_2 = 8 h.
- **Max_vel_thresh** (T*vmax* in the text): is the "velocity threshold" used to exclude movements from the permanence score estimation; default value T*vmax* = 50 km/h
- **Score_interval (optional):** this represents a potential discretization of the timeslot to assign score to a finest granular time intervals than the timeslot. The score intervals depend on the time_slot_number e.g. when the number of time slot is 24, a possible choice could be 4 equal intervals. Different choice for the time slots number will produce a different discretization, in particular, the smaller the number of time slots, the more numerous the score intervals will be. In the first release we believe this sophistication is not required and we are fine with a single interval per each time slot. So, default value=1 when n=24

14.2.3 INPUT DATA

- Semantically Cleaned Event Data at Device Level [INTERMEDIATE RESULTS]
- Cell footprint values [INTERMEDIATE RESULTS]

14.2.4 OUTPUT DATA

- Daily Permanence Score (see a possible example in Examples)

14.2.5 METHODOLOGY

The method performs the analysis on the input per single device k and per single day d. The output is the daily permanence score per grid tile and time slot. Time slots come from the discretisation of the day-time length of 24 hours in n time intervals of equal length.

The input processed by the method consists of the target day events plus the last event of the previous day and the first event of the next day.

The permanence score of a cell is a proxy of the sum of all the time intervals spent in the cell in the timeslot. By assigning the time intervals spent in a cell to all the tiles composing the cell footprint (CF), the daily permanence score is then given per tile. The procedure below represents the rationale to follow to estimate the output.

The method takes as input:

- the Cell Footprint Values [INTERMEDIATE RESULTS]
- the Semantically Cleaned Event Data at Device Level [INTERMEDIATE RESULTS]

Each timestamp t_i is linked to the cell_id and to the set of tiles composing the cell footprint in the target day. The tiles set represents all the geographical locations in which the device could be at timestamp t_i.

1. if the target event refers to a supposed stay (in a given cell) the corresponding spent time is estimated and used to calculate the permanence score associated to the event;
2. if the target event refers to a supposed fast movement the event does not contribute to the permanence score estimation;
3. the time when the device is not connected to any cell is assigned to a location defined "unknown".

The previous points are illustrated in detail in the following paragraphs.

Note: in a future evolution of the method (beyond this project), it could be helpful to consider other cell information like presence of highways, railways, stations or underground, as a further criterion to check if the event relates to fast movements.

14.2.5.1 SUPPOSED STAY

\ SUCCESSIVE EVENTS CORRESPONDING TO THE SAME CELL

If two successive events are associated to the same cell, the procedure of spent time assignment is a bit different. We want in fact to take into account cases in which the device is turned off in the same location (same cell) as it frequently occurs during the night and cases in which the device loose connection even if it remains in the same place (same cell).

When two events are associated to the same cell, the threshold corresponding to the maximum spent time is set as follows:

- during the day-time the threshold Tsc1 is set by the parameter **Max_time_thresh_day** (default value = 2h)
- during the night-time the threshold Tsc2 is set by the parameter **Max_time_thresh_night** (default value = 8h)

A quality metric that could be associated to this assignment is the percentage of time slots over a day whose value has been assigned without observation (as a sort of imputation rate).

\ SUCCESSIVE EVENTS CORRESPONDING TO DIFFERENT CELLS

When two successive events are associated to different cells, the time interval between two successive locations (cells) is estimated and distributed between the two cells. The unit of measure of the time interval is the minute.

The time spent in the cell associated to the event is given by the semi-time distance between the event timestamp and the previous one plus the semi-time distance between the event timestamp and the following one (Δt2 in the figure below):

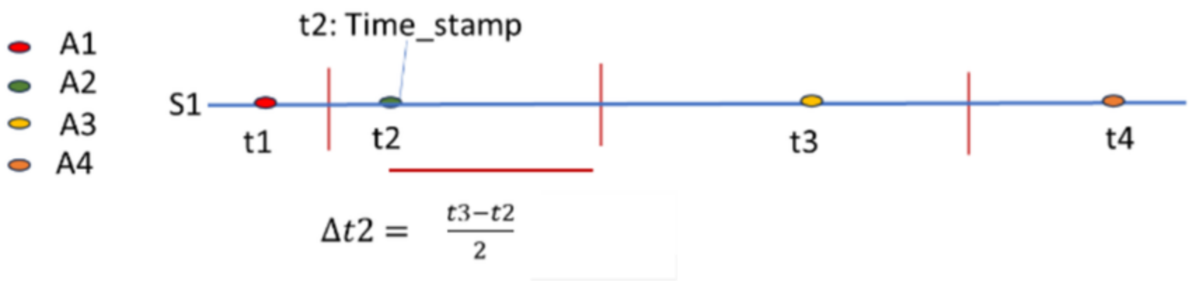


Figure 9: Time stamps associated to events

Dots in the figure represent the events (at time t1, t2...), their colours represent the associated CFs (A1, A2..., different colours for different CFs), while the red vertical lines correspond to the temporal semi-distance between two events. As an example, the spent time spent in A2 is given by Δt2 plus the Δt1.

If the time interval t3-t2 exceed a given threshold T*sd* (parameter Max_time_thresh, default value =15 min) Δt2 is assigned to be equal to the threshold T*sd*. Then the spent time assigned to a cell for a single event can be equal to 30 min (2*T*sd*) at maximum. When the threshold is exceeded, the remaining part of the time interval is assigned to the location "unknown".

After processing all the events of a day (having assigned to cells the times spent corresponding to all day events) the daily permanence score is obtained summing up all the spent times of the given cell. Once a spent time relative to an event is assigned to a cell, it is also assigned equally to all tiles composing the cell footprint. Hence the final permanence score is given per tile.

14.2.5.2 SUPPOSED FAST MOVEMENT

The space distance between two generic cells Aj and Ak, (i.e. dist(Aj,Ak)), is defined as the minimum Euclidean distance between all the Aj tiles and all the Ak tiles.

- if Aj and Ak are adjacent cells or overlapping cells the minimum distance is set to zero; cells having a tile vertex in common are considered adjacent cells).
- if Aj and Ak are not adjacent cells nor overlapping cells dist(Aj, Ak) is measured as the minimum distance between the two closest vertices of the two cells footprints.

See figure below.

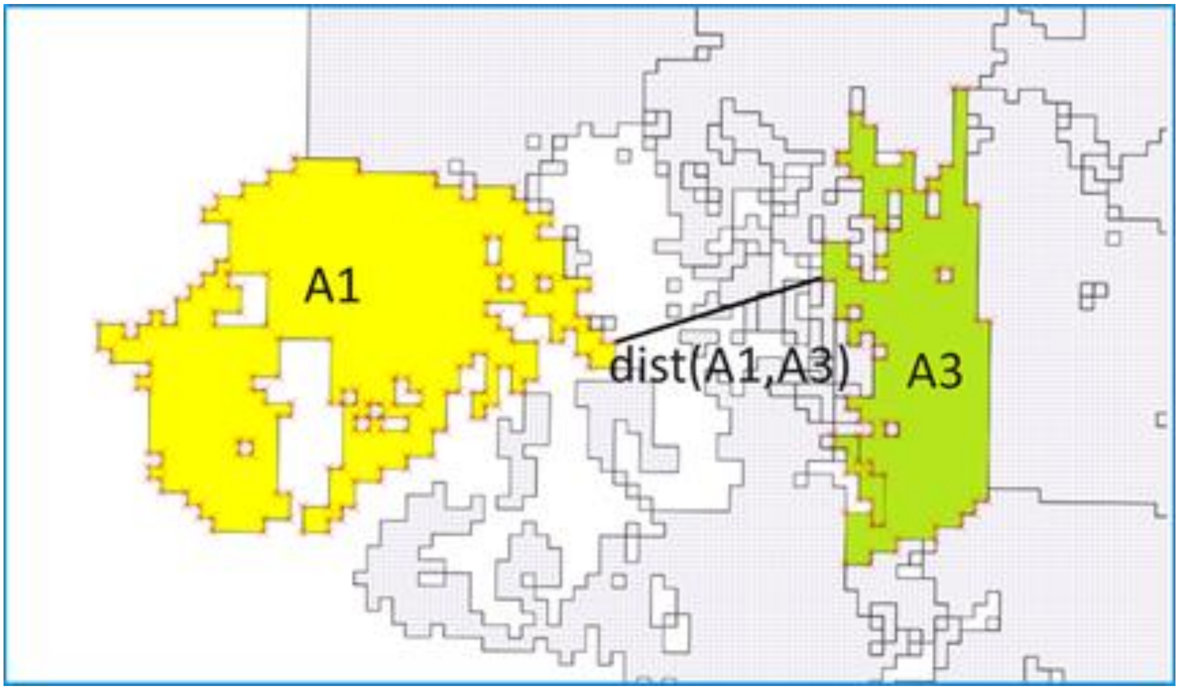


Figure 10: Minimum distance between the two closest vertices of two cells' footprints

To check if an event t2 refers to a movement, the method assesses the supposed speed the device has when the event is registered, using the previous and next event's time (t1 and t3) and location (A1 and A3). The device speed is estimated by dividing the approximated spatial distance the device travelled to go from cell A1 to cell A3 by the time interval between t1 and t3. If the estimated device speed exceeds the velocity threshold T*vmax,* then t2 event is taken as a fast movement. The suggested value for the velocity threshold is 50 km/h.

Let's call TravelDist(A1, A2, A3) the approximated travelled spatial distance between A1 and A3 passing by A2. It is calculated as the maximum value between dist(A,A3) and the sum of dist(A1, A2) and dist(A2, A3):

TravelDist(A1, A2, A3) = Max[dist(A1,A3), dist(A1,A2) + dist(A2,A3)]

In general, the travelled distance should be equal to the distance between A1 and A2, plus the distance between A2 and A3 (see figure1 below). However, since the cells are not represented as geographical points and cover instead an area, there are some cases where we need to take the distance between A1 and A3 as travelled distance: this is the case when A2 is contiguous to both A1 and A2 (see figure 2 below). dist(A1, A2) and dist(A2, A3) are in fact both zero and hence the travelled distance appear to be zero, even if the device travelled for a non-zero spatial distance. Then in this specific case we need to take the dist(A1, A3) as the approximated travelled distance.

We summarise it with the following formulas:

- in figure *a*, TravelDist(A1,A2,A3) = dist(A1,A3). dist(A1,A2) =dist(A2,A3)=0.
- in figure *b*, TravelDist(A1,A2,A3) = dist(A1,A2) + dist(A2,A3). dist(A1,A2) !=0 and dist(A2,A3)!=0

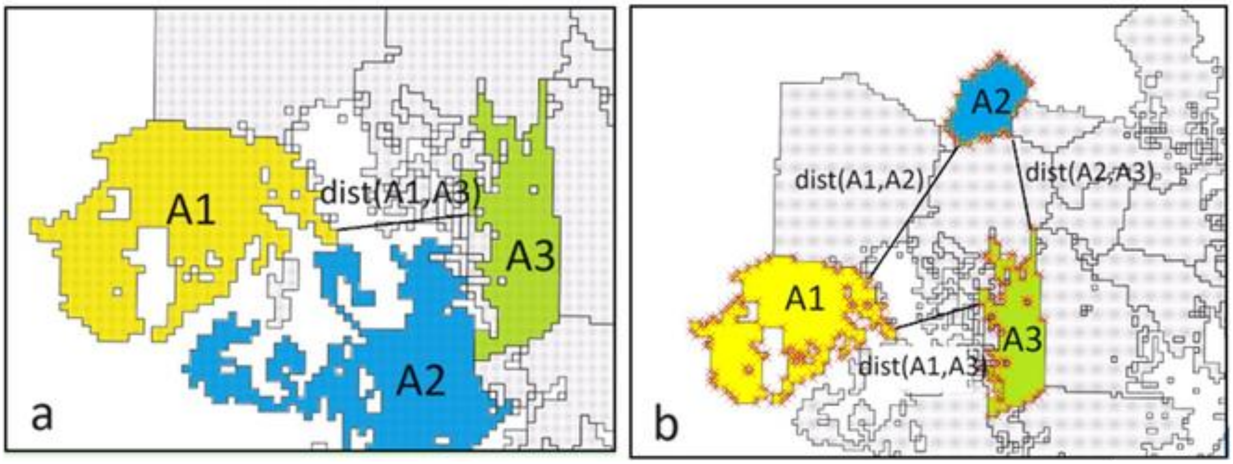


Figure 11: Visual representation of overview formulas

14.2.5.3 TIME DISCRETIZATION

The daytime length is splitted in n equal time intervals (n is set by the input parameter Time_slot_number). Hence in the method output the calculated permanence scores are given per time slot. We suggest using Time_slot_number = 24, so to have time slot of 1 hour length.

The permanence score reported in the output is indeed a score, instead of time values.

We propose to assign score 1 to all the tiles where the device is estimated to spent time for more than half the time_slot size. The estimated spent time is calculated as the sum of the estimated time spent in the analysed time slot on the basis of the events observed. When the overall estimated spent time is less than half the time_slot size, the score is 0.

If we adopt the **Score_interval** parametrization, the rule for assigning the score will change accordingly.

For example, assuming time slots of 1 hour, the **Score_interval =4,** the score can take 4 values, as follows:

- minutes in the time slot are between 1 and 15, permanence score value = 1
- minutes in the time slot are between 16 and 30, permanence score value = 2
- minutes in the time slot are between 31 and 45, permanence score value = 3
- minutes in the time slot are between 46 and 60, permanence score value = 4

14.2.6 PSEUDO-CODE

Note: The pseudocode represents part of the procedure for illustrative purposes.

For a single device k, we consider a daily partition in “n” time slots (total number of time slots n = 24) and we call “h” the index of a single time slot. Each time slot h corresponds to a time interval in the day of one hour that we call time window, defined by as follows:

time window[h]= time_end[h]- time_start[h]

Let us define the following:

- k is the device index
- i is the event index: event[i] refers to the i-th event, event[i].time refers to the timestamp of the i-th event, tile(CF) is the set of tiles included in the CF.
- event[i].cell_id is the id of the cell to which the device is connected at the event[i].
- event[i].cell_id is associated to event[i].CF through the cell footprint estimation procedure. event[i].CF is the cell footprint, that is the set of tiles corresponding to the geographical area covered by the cell represented as a grid.
- tile(event[i]) is the array of all tiles associated to event[i].CF.
- tile.permanence[k][h] is the permanence score corresponding to the total time the device k has been seen in the tile in a given daily time slot h, estimated by the function stay_permanence_intiles().
- for a given device k and a given time slot h, device[k][h].unknowntime is the total time the device k is in an unknown location.
- given two cell footprints A1 and A2 dist(A1 ,A2) is the distance between A1 and A2.
- T*sd* is defined in Methodology
- T*vmax* is the “velocity threshold” as defined in Methodology
- “+=” is used for the increment function
- “==” is used for the strict equality
    - for a single time slot h, we cycle on the day events (i = event index, m = total number of day events), using the function stay_permanence_intiles(). Given a device k and a time slot h stay_permanence_intiles() function is defined as:

```
for i=1:m

if event[i].time is in time slot[h]

if event[i].CF == event[i-1].CF

   tile(event[i].CF).permanence += event[i].time - event[i-1].time;

   i+=1;

else if  distmin( event[i-1].CF, event[i+1].CF) < (event[i+1].time - event[i-1].time) * Tvmax

    left_time = event[i].time - event[i-1].time;

    tile(event[i-1].CF).permanence += min(left_time/2, Tsd);

    tile(event[i].CF).permanence += min(left_time/2, Tsd)

    device[k].unknowntime += max( left_time -- 2*Tsd, 0)

i+=1;

else  skip event[i];
```

```
    i+=1;
endif
end
```

The function returns tile.permanence[k][h] and device[k][h].unknowntime as output.

14.2.7 EXAMPLES

Example of daily permanence score summary

Below we show a table representing an example of the method output. The column index indicates the time slots while the row index indicates the grid tiles having a non-zero permanence score. The time spent in an unknown location is also included and it is represented by the tile called "unknown" in the last row of the table.

The values in the table are the permanence score expressed as 0 and 1.

Note that the output of the method is a table per device.

	Time_slot[1]	**Time_slot[2]**	**Time_slot[3]**	**Time_slot[..]**	**Time_slot[24]**
Tile_id[310]	1	1	0	0	0
Tile_id[312]	1	1	0	0	0
Tile_id[314]	0	0	1	0	0
Tile_id[376]	1	1	1	0	0
Tile_id[388]	1	1	1	0	0
Unknown	0	0	0	1	1

Alternatively, when the Score_interval =4

The values in the table are the permanence score.

	Time_slot[1]	**Time_slot[2]**	**Time_slot[3]**	**Time_slot[..]**	**Time_slot[24]**
Tile_id[310]	1	4	3	2	0
Tile_id[312]	3	1	0	3	1
Tile_id[314]	0	3	2	2	1
Tile_id[376]	1	4	4	2	0
Tile_id[388]	1	4	2	4	3
Unknown	2	1	1	0	1

14.3 METHOD 3: CONTINUOUS TIME SEGMENTATION

14.3.1 OBJECTIVE

The goal of the 'Continuous Time Segmentation' is to determine at each time point the nearest cell of a device and to assign a state 'stay', 'move' or 'unknown' to a device. The MNO input data is event data: an event *e* registers the activity of device *d* at a time *t* at cell *c* and can be denoted with the triple (*d*, *t*, *c*). This event data does not directly support the following queries:

a) Which cell(s) device *d* was connected to at a chosen time t_i? Use cases have the need to count the number of devices at location l_i at time t_i. Since event data is a point in time and events have a much lower frequency than seconds, in general there will not be an event at chosen time t_i, e.g. 12:00:00 am, so an interpolation is needed.

b) Which cells device *d* was connected to in time interval [t_1, t_2)? Use cases have the need to count the average number of devices or the number of unique devices at location *l* in period *p*. The events measured in [t_1, t_2) underestimate the number of devices at *l*, since they represent only devices that were active on the mobile network during period *p*. Other devices present at location *l* may be switched off or active on WiFi and private networks. Interpolation helps to capture devices that are present at *l* but not active on the mobile phone network. In some networks an event is generated when a device switches from cell, which makes it more accurate to track the closeness of a device to a cell. Based on previous experience, these data are often missing, because the MNO does not permanently store those switches; i.e. the operational local network has the information, but it is ephemeral and not stored permanently in the MNO event data.

c) Was *d* at rest ('stay') or moving ('move') at time t_i*?*

d) What is the time and duration of a stay or move(ment)?

These queries are input for use cases. For example, for the present population we need to be able to determine where every device is at a certain point in time. Time Segmentation transforms discrete events into contiguous time segments. A time segment is an interpolation of event data in time, describing the where-about and movement of a device in terms of cell-ids. It has a start and end time, a set of cells associated with them determining the position of the device during the time interval and has a state e.g. 'stay', the device is not moving for a certain period of time, or a 'move', the device is moving from one location to the next. For device d it records location as a set of cell ids g in interval [begin,end) and derives a state s ('stay', 'move', 'undetermined', 'unknown') resulting in the tuple (d, [begin,end), g, s). While an event has one cell c, the set g of a time segment can contain 0, 1 or more cells, depending on the state of the segment. For example, a stay time segment typically contains more overlapping cells. When it cannot be determined whether a device is 'stay' or 'move' the state is 'undetermined'. When cell_id is not known, the state s is set to 'unknown'. Note that in many cases we will be unable to determine reliably whether a device is moving or at rest.

Possible (location, state) tuples for a time segment are:

- *(known, stay)*: the location of the device is known and the device is staying in one location for a certain period of time.
- *(known, move)*: the device is moving from one location to the next; the location of the device is somewhere in between the two locations.
- *(known, undetermined)*: the location of the device is known, but it is unclear whether or not the device is moving.
- *(unknown, undetermined):* the location of the device is unknown: there are no events for a certain (longer) period of time, but assumption is that the device is in the country.
- *(not-in-use, undetermined)*: the device is no longer in use. (implemented as unknown for now).
- *(abroad, undetermined)*: the location of the device is known to be in another country. (implemented as unknown for now).

Note that the difference in locations ‘unknown', 'not-in-use' and ‘abroad' is relevant for the UCs; they matter for estimating the number of persons, because they differ in the assumption on presence of the device (and by proxy presence of a person). They are all implemented as unknown because the difference between these states cannot be generally derived on a daily basis; e.g. a criterion for ‘not-in-use' could be that events are missing for several days. It is possible to detect devices leaving or entering the country (*abroad*) using events generated at airports and borders crossing, but for implementing the time line for a device correctly, information is needed for longer time periods. Therefore, should a UC need this information, a further refinement of the time segments derived in the current process step is needed.

In principle, UCs are free to implement their own time segmentation logic, but having a default time segmentation assures that results from UCs sharing the time segmentation have a coherent output.

14.3.2 PARAMETERS

- ‘**study_date**': date from which time segments will be computed.
- **min_time_stay**: minimum dwell time to be considered as 'stay'.
  - default value: 15 minutes (somewhere in the range of [15,30] minutes, to be determined).
- **max_time_missing_stay**: maximum time difference between events to be considered a ‘stay'. If larger, the time segment will be marked ‘unknown'.
  - default value: 12 hours, to support devices being offline at home, e.g. while sleeping, or work addresses, e.g. while working.
- **max_time_missing_move**: maximum time difference between events to be considered a ‘move'. If larger, the time segment will be marked ‘unknown'
  - default value: 2 hours: most commuting durations are less than 2 hours.
- **pad_time**: half the size of an isolated time segment: between two ‘unknown' time segments. It expands the isolated event in time, by ‘padding' from the ‘unknown' time segments on both sides.
  - default value: 5 minutes, must be less than **min_time_stay**/2, otherwise the isolated segment can just be a “stay”.

Parameters for the determination of the cell intersection groups:

- **min_signal_strength**: minimum signal strength for a coverage area to be considered.
  - default value: ? (to be discussed)
- **min_signal_fraction**: minimum signal strength fraction for a coverage area to be considered.
  - default value: 0.05, each cell in an intersection group should at least contribute for 5% to the signal strength.
- **max_distance_overlap**: maximum distance between coverage area to consider them overlapping.
  - default value: 10 meters, accounts for the accuracy error in the coverage area.

14.3.3 MODULE INPUT DATA

- Clean MNO Event Data [INTERMEDIATE RESULTS]
- Cell Footprint Values [INTERMEDIATE RESULTS]

14.3.4 MODULE OUTPUT DATA

- Daily Continuous Time Segmentation [INTERMEDIATE RESULTS]

14.3.5 METHODOLOGY

The goal of the method is to divide the time line for a given device into time segments. For each time segment, the location of the device is determined by the cells the device was connected to. Therefore, events from different cells should only be grouped into one time segment when events at these different cells likely correspond to a device that stays at one location; events are grouped where the cells have overlapping coverage areas. As the cell plan changes relatively infrequently (e.g. once per day), and the cell plan is the same for all devices, it is efficient to precalculate the overlap of the cells prior to the time segmentation. For the time segmentation it is only necessary to know if a set of cells are overlapping and that, therefore, a set of subsequent events at this set of cells can be grouped into one time segment. These sets of overlapping cells will be called *cell intersection groups*. The algorithm for calculating these groups is presented in the next section. This section will start with presenting the algorithm for the continuous time segmentation.

14.3.6 ALGORITHM FOR CONTINUOUS TIME SEGMENTATION

The continuous time segmentation module conducts the following processes per device *d*.

1. Select all events of D-1, **‘study date’** D and D+1 and order them by time. D+1 are only used for “look-ahead” and will not result in time segments.
2. IF no events in D, create a time segment ts (d, [0:00, 24:00), {}, “unknown”), i.e. with state s “unknown” and group = {} (empty), and STOP.
3. Select the last time segment of D-1: ts(d, [begin,end), *group*, state). If none create an unknown time segment for the previous day: *ts* = (d, [0:00**,** 24:00) - day, {}, “unknown”).
4. Select next event e(d,t,cell), if none STOP.
5. IF ts.state = “undetermined” AND intersection(ts.group, e.cell)) is an intersection group g (i.e. they overlap):
    1. IF e.*t* - ts.*end* < ***max_time_missing_stay*** change *ts* to *(d, [ts.begin,* ***e.t****]*, **g**, *“undetermined”)*
        1. IF ts.end - ts.begin > **min_time_stay**, set state of segment ts to “stay”: (d, [ts.begin, ts.end), ts.group, **“stay”**)
        2. GOTO 4
6. IF ts.state == “stay” AND e.cell is in ts.g (i.e. e.cell is in the current intersection group):
    1. IF e.*t* - ts.*end* < ***max_time_missing_stay*** change *ts* to *(d, [ts.begin,* ***e.t****]*, ts.group, *“stay”)*
        1. GOTO 4
7. IF ts.state = “move” AND intersection(ts.group, e.cell) is an intersection group g (i.e. they overlap):
    1. IF e.*t* - ts.*end* < ***max_time_missing_stay*** create a new *ts* to *(d, [****ts.end****,* ***e.t****]*, **g**, *“****undetermined****”)*
        1. IF ts.end - ts.begin > **min_time_stay**, set state of segment to “stay”: (d, [ts.begin, ts.end), ts.group, **“stay”**)
        2. GOTO 4
8. IF *e.t* - ts.end < **max_time_missing_move**, divide the interval [ts.end, e.t] in half with two new “move” time segments. *ts*$_1$ and *ts*$_2$:
    1. create time segment *ts*$_1$*(d, [ts.end, ts.end + (e.t - ts.end)/2) , ts.g, “move”)*
    2. *create time segment ts*$_2$*(d, [ts.end +(e.t - ts.end)/2, e.t) , e.c, “move”)*

    3. Set *ts* to ts_2
    4. GOTO 4
9. ELSE adjust ts.end = ts.end + **pad_time**

    1. create time segment ts_1(*d*, [ts.end + **pad_time**, *e.t* - **pad_time**) , {}, “unknown”)
    2. create time segment ts_2(d, [e.t-**pad_time**, e.t), e.c, “undetermined”).
    3. Set current time segment *ts* to ts_2
    4. GOTO 4.

14.3.7 DETERMINATION OF CELL INTERSECTION GROUPS

Intersections for overlapping cells:

1. Determine for day D overlapping cell intersections {*g*}: cells with overlapping effective cell footprints, with a minimal signal strength. These are combinations of cells for which the intersection of the effective footprints is not empty and the signal strength in the intersection is larger than **min_signal_strength**.
   Note that individual cells also are a cell intersection.
2. Restrict each coverage area to the **min_signal_strength**.
3. Start with a set of cell intersections in which each cell is a cell intersection, and expand each area with **max_distance_overlap**/2. In GIS terminology this is a ‘buffer’ query.
4. For each cell intersection find spatially intersecting cell intersections and add these to the set.
5. An extra relative selection criterion could be used to account for unbalanced intersections having both strong and weak signal cells: restrict the set to intersections in which each cell has at least **min_signal_fraction** (e.g. 5%) of total signal strength in the intersection.
6. Goto 4 until no new intersections can be made.

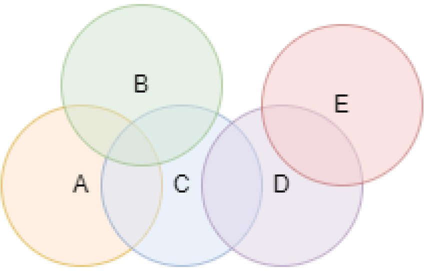
14.3.8 INTERSECTION GROUPS (EXAMPLES)

As an illustration of how intersection groups are determined, the figure below shows a simplified example of 5 cells that are partially overlapping: when the respective circles of two cells overlap, the coverage areas overlap in such a way that it is likely that a device that has events at both cells is staying at the same location. For example, if we have in a relatively short time events at cells A, B and C, it is likely that the device is staying at one location, while subsequent events at A and D indicate that the device has moved from A to D. Calculating these overlaps can be computational intensive. However, in order to determine if a set of events, such as A, B, C belong to one stay, one only needs to know whether the set off cells {A, B, C} correspond to overlapping cells. These sets can be calculated once per day (or however frequent the cells change) and reused in the determination of the time segments.

The number of intersection groups depends on the number of overlaps between *N* cells. Suppose each cell has maximally *o* overlapping cells, then an upper boundary for the number of intersections groups is $(2^{o}-1) \times N$. The number of groups will be a lot smaller, since the upper boundary assumes that all neighbors are also overlapping (which is certainly not the case) and it counts the groups multiple times for each cell.

When cells have less than 4 overlapping neighbors this results in less than 4!*N* or 24*N* number of groups. The intersection groups can be stored as a sparse matrix (for each cell, the cells with which the cell overlaps need to be stored). The amount of storage needed is therefore O(*N* x *o*).

Figure 12: Simplified example of overlapping coverage areas for cells



The example above results in the following set of cell intersection groups:

{A}	{B}	{C}	{D}	{E}	{A, B}	{A,C}
{B,C}	{C,D}	{D,E}	{A, B, C}			

14.3.9 TIME SEGMENTATION (EXAMPLES)

The figure below shows an example of a possible sequence of events.

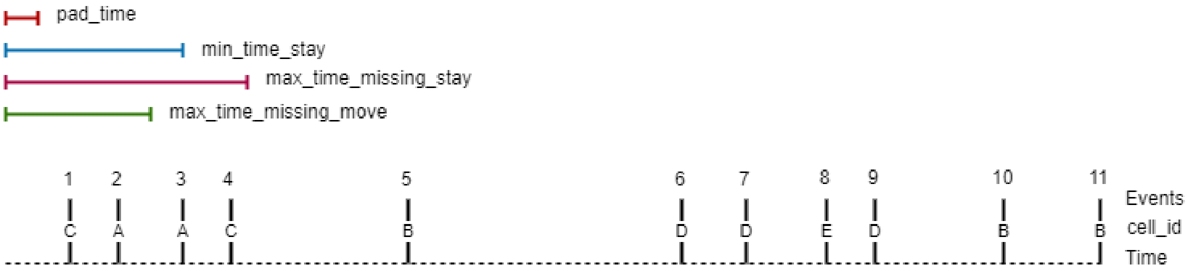
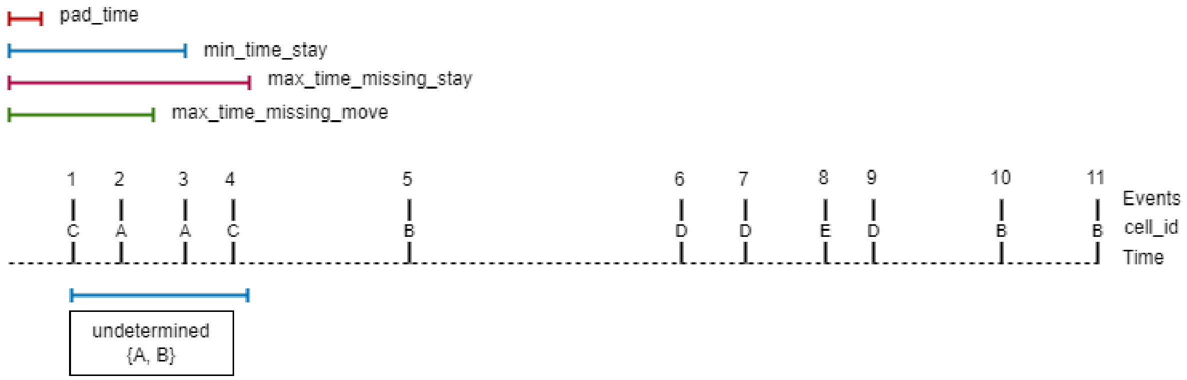


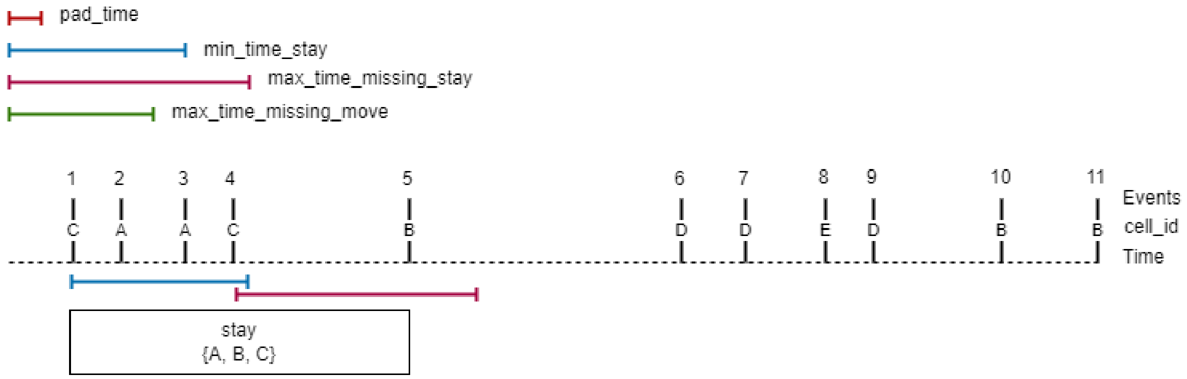
Figure 13: Example of a possible sequence of events.

The horizontal line indicates time. The lengths of the various parameters in the algorithm are indicated by the coloured horizontal bars and the events are the vertical lines with the letters indicating the cell_ids (the same as in the example of the cell intersection groups).

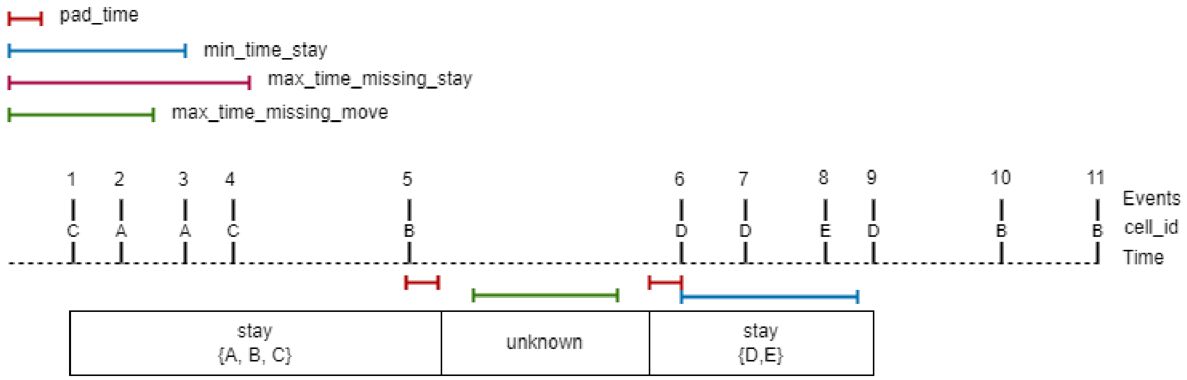
The figures below show step-by-step how the previously presented algorithm determines the time segments. First, the events 1-4 are processed on by one. As these are all within **max_time_missing_stay** of each other and all belong to a cell intersection group these are all grouped into one time segment. The length of this segment is smaller than **min_time_stay** therefore this segment is not a stay. Neither is it a move as there is no previous time segment**.**



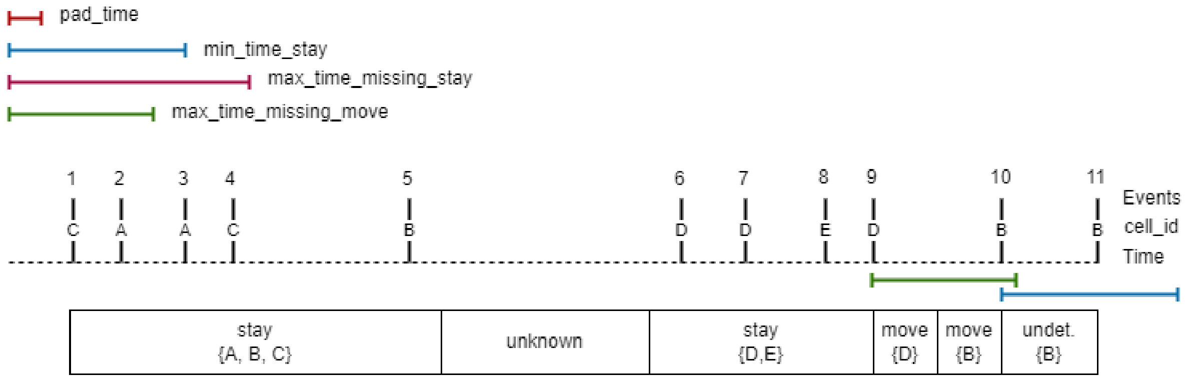
Event 5 also belongs to a cell intersection group with the current time segment. Therefore, B is added to the time segment. Since the length of the time segment is now longer than **min_time_stay**, this time segment is now labelled as a stay.



Event 6 and the current time section do not belong to a same intersection group. Therefore, a new time segment is started. The duration between event 5 and 6 is longer than **max_time_missing_move**. Therefore, we do not determine the move between 5 and 6 and the time segment between 5 and 6 is labelled as unknown. The end time of the previous time segment and the time segment to which event 6 belongs are extended with **pad_time**. Events 6-9 belong to the same cell intersection group and are therefore combined into one stay.



There is no cell intersection group to which both event 10 and the current time segment. Furthermore, the time between event 9 and 10 is smaller than **max_time_missing_move**. Therefore, moves are introduced between events 9 and 10. These moves split the time period between 9 and 10 into two time segments. Events 10 and 11 do overlap and are, therefore, combined into a time segment. The length is currently too short to label this time segment as a stay.



The following tables show the same example given the event input data and the resulting time segment records.

MNO event Data

event_id (E)	device_id	timestamp	cell_id
1	1	2023-01-01 05:15:00	C
2	1	2023-01-01 05:16:00	A
3	1	2023-01-01 05:18:00	A
4	1	2023-01-01 05:19:00	C
5	1	2023-01-02 09:30:00	B
6	1	2023-01-02 19:15:00	D
7	1	2023-01-02 19:17:00	D
8	1	2023-01-03 19:21:00	E
9	1	2023-01-03 19:22:00	D
10	1	2023-01-03 20:30:00	B
11	1	2023-01-03 20:35:00	B

The output data object of ‘Continuous Time Segmentation’ created in this process is presented in the table below.

time_segment_id	state	device_id	begin	end	cells_id
1	stay	1	2023-01-02 05:15:00	2023-01-02 09:35:00	A,B,C
2	unknown	1	2023-01-02 00:09:35	2023-01-02 19:10:00	null
3	stay	1	2023-01-02 19:10:00	2023-01-02 19:22:00	D,E
4	move	1	2023-01-02 19:22:00	2023-01-02 19:26:00	D
5	move	1	2023-01-02 19:26:00	2023-01-02 20:30:00	B
6	undetermined	1	2023-01-02 20:30:00	2023-01-02 20:35:00	B

14.3.10 POSSIBLE REFINEMENTS STAY DETECTION

An important aspect of Time Segmentation is *stay* detection, which is implemented by detecting if subsequent events have cells that are overlapping. This assumption is based on that a device at rest may switch cells, because the network of a MNO may load balance connections on the network: a so-called *handover*. Switching cells at rest (*stay*) can only happen if the device can connect to both cells. When a device is not at rest (i.e. *move*) the device connects to different cells. If subsequent events have different cells, it is thus the question if the device is staying or moving.

There are refinements possible for detecting if a cell change is a device at rest (*stay*) or on the *move*, for example:

- A probabilistic approach: given that a device is connected to cell A and using the location probabilities derived with the location algorithm, the probability of the device being there is known. For each location it is possible to determine the probability of connecting to B, which allows to derive the likelihood, that the device has moved or not. Setting a threshold on the likelihood would create a decision boundary on *stay* or *move*.
- A event driven approach: take a time window of length **w**, when there are back-and-forth switches between cells that have overlapping coverage areas, then we consider them to be in the same stay. Using this approach would require a ”windowed” processing mechanism.

15 MODULE 14: MID-TERM PROCESSING

15.1 METHOD 1: MID-TERM PERMANENCE ANALYSIS

15.1.1 OBJECTIVE

The objective of this module is to obtain indicators at a mid-term time scale, e.g. one month as default mid-term scale. Daily data are collected and analysed in order to extract information on visited locations (tiles) in terms of regularity, frequency and time spent, at a mid-term level. The discretised permanence score of all tiles, output of the previous module, is taken as input for this method.

The module is executed per single month and single device.

15.1.2 PARAMETERS

The suggested periods and their definitions are listed in the following:

- **Day:**
    - start_time = 4 a.m.
    - end_time = 4 a.m. (the day after)
- **Period of analysis**:
    - mid_start_day: first day of the month (e.g. 01/05/2022)
    - mid_end_day: last day of the month (e.g. 31/05/2022)
- **Input tables**: day tables from reg_start day to reg_end_day.
    - reg_start_day: 15 days before the month (e.g. 16/04/2022, it can be equal to mid_start_day, as latest)
    - reg_end_day: 15 days (e.g. 15/06/2022, it can be equal to mid_end_day, as latest)
- **Sub-daily periods:**
    - night: start_time = 8 p.m.; end_time = 8 a.m. (the day after)
    - working hours: start_time = 8 a.m.; end_time = 5 p.m.
    - evening hours: start_time = 4 p.m.; end_time = 10 p.m.
- **Sub-monthly periods:**
    - working days: Monday, Tuesday, Wednesday, Thursday, Friday.
    - holidays (from the calendar info object)
    - weekends:

    - weekend_start_time: Saturday 4 a.m.
    - weekend_end_time: Monday 4 a.m.

  (Note: these are parameters that can be modified by defining different start/end time, e.g. Friday 8 p.m. – Sunday 8 p.m.)

- day of the week: Monday, Tuesday, etc. (from calendar info object)

15.1.3 INPUT DATA

- 14.2 Daily Permanence Score Estimation
- Calendar Info [INPUT DATA]

15.1.4 OUTPUT DATA

- mid-term permanence score per daily, sub-daily and sub-monthly periods
- mid-term frequency count per daily, sub-daily and sub-monthly
- mid-term regularity indices per daily, sub-daily and sub-monthly

Please see the examples heading under this section: Examples of output data

15.1.5 METHODOLOGY

The method processes data relative to all days of a single month and produces output measures aggregated at monthly level. More precisely, the method takes as input the output of the 'Daily Permanence Score Method', i.e. tables of permanence scores given per tile and per time slot. The tables for all days of the reference month are collected and the monthly measures are produced. The method may take as input also tables relative to a selection of the last days of the previous month and of the first days of the successive month, to calculate the regularity indices (see reg_start_day and reg_end_day parameters).

The output of the method is then provided at monthly level. The output aggregated measures are: the *mid-term permanence score*, the *mid-term frequency count* and the *mid-term regularity measures,* all provided per grid tile and per daily, sub-daily and sub-monthly periods. The sub-daily and sub-monthly periods are defined in the parameters section (above) along with the hourly intervals to identify them:

- day
- night-time
- working hours
- evening hours
- weekends
- working days
- holidays
- day of week

Clearly, if the sub-daily periods are overlapping, the permanence score in the sub-daily periods do not need to sum up to the daily permanence score. Furthermore, the hourly intervals defining time sub-periods are given as suggested default values; therefore, they can be configurable by the NSI in order to fit different lifestyles across countries.

Daily summaries are aggregated monthly by using different methods to provide monthly indicators. The procedures below represent the rationale to estimate the output:

- the **mid-term permanence score** for each tile is calculated as the summation of the tile daily scores in the given month for the daily period and for the considered sub-daily and sub-monthly periods.
- the measure of the **mid-term frequency count** per tile is calculated by counting the number of days of the target month in which the tile has non-zero permanence score. The monthly frequency is to be calculated per daily, sub-daily and sub-monthly periods.
- measures of the **mid-term regularity indices** per tile are the mean and the standard deviation of the temporal distance in number of days between consecutive permanencies in the given tile (n.b. to calculate the temporal distance in number of days between consecutive permanencies in the given tile, we also use the last visit in the tile during the last *n* days of the previous month and the first visit in the first *m* days of the next month). The measures are to be calculated for daily, sub-daily and sub-monthly periods. To consider borders in the regularity indices calculation, we compute the mean of distances between visits. The first value to take into account is the number of days between the last visit in the *n* last days of the previous month and the first visit in the target month. Similarly, the last value is the number of days between the last visit in the target month and the first visit in the *n* first days of the next month. If visits are not present in the last *n* days of the previous month the days' distance is calculated between reg_start_day and the first visit in the target month. The same holds true for the last day's distance, or if reg_start_day is equal to mid_start_day and reg_end_day is equal to mid_end_day.

The above monthly indicators are reported in the mid-term summaries and provided per all the tiles having a non-zero value of the monthly permanence score per daily, sub-daily and sub-monthly periods (see the Examples heading, below in this section). In addition, the information on the total assigned permanence score over the month and the number of days in which the device has been observed are recorded in the last row of the tables as 'Device Observation'.

15.1.6 EXAMPLES OF OUTPUT DATA

Below we introduce examples of the method's outputs in a tabular form. Columns report the monthly indicators, while rows indicate the tiles id. The tile relative to the 'unknown location' is also included in the last row of the table.

The output tables are given per each considered sub-daily period, sub-monthly period and also per combinations of sub-daily periods and sub-monthly periods.

DAILY	**Permanence score**	**Frequency count (in number of days)**	**Mean and *STD* (distance between consecutive visits in days)**
tile_id[310]	50	6	7 (2)
tile_id[312]	120	10	1 (0)
tile_id[314]	60	8	15 (2)
tile_id[376]	20	18	5 (1)
tile_id[377]	200	20	10 (3)
Unknown	100	30	1 (0)
Device Observation	300	30	

WEEKENDS	**Permanence score**	**Frequency count (in number of days)**	**Mean and *STD* (distance between consecutive visits in days)**
tile_id[310]	x1	y1	z1 (std1)
tile_id[312]	x2	y2	z2 (std2)
tile_id[314]	x3	y3	z3 (std3)
tile_id[376]	x4	y4	z4 (std4)
tile_id[377]	x5	y5	z5 (std5)

Unknown	x6	y6	z6 (std6)
Device Observation	x7	y7	
EVENING hours	**Permanence score**	**Frequency count (in number of days)**	**Mean and *STD* (distance between consecutive visits in days)**
tile_id[310]	x1	y1	z1 (std1)
tile_id[312]	x2	y2	z2 (std2)
tile_id[314]	x3	y3	z3 (std3)
tile_id[376]	x4	y4	z4 (std4)
tile_id[377]	x5	y5	z5 (std5)
Unknown	x6	y6	z6 (std6)
Device Observation	x7	y7	

EVENINGS hours in WEEKENDS	**Permanence score**	**Frequency count (in number of days)**	**Mean and *STD* (distance between consecutive visits in days)**
tile_id[310]	x1	y1	z1 (std1)
tile_id[312]	x2	y2	z2 (std2)
tile_id[314]	x3	y3	z3 (std3)
tile_id[376]	x4	y4	z4 (std4)
tile_id[377]	x5	y5	z5 (std5)
Unknown	x6	y6	z6 (std6)
Device Observation	x7	y7	

16 MODULE 15: LONG-TERM PROCESSING

16.1 METHOD 1: LONG-TERM PERMANENCE ANALYSIS

16.1.1 OBJECTIVE

The objective of this method is to obtain measures on a large time scale (e.g. 6 months, 1 year). It aims to produce as output a proxy for the *Usual Environment* of each device and a tentative identification of *Home Location*, *Second Home* and *Work/Study* place. The 'usual environment' (UE) refers to the specific location or setting where individuals typically spend the majority of their day or where they have their primary activities and interactions[10] (see for further details the M-UE indicators use case in Volume II).

Here we define rules to operationalise the above concepts. These rules include parameters and their default values, which can be customised by NSIs in each country according to their needs. Nevertheless, the adoption of default values can facilitate the comparability among countries.

In this method, mid-term output data are collected and analysed in order to extract information on the most visited locations in terms of regularity, frequency and time spent over a long-term period (e.g. 1 year, 6 months or a lower number of months). The regularity, frequency and permanence score mid-term indicators of all visited tiles, output of the Method 1: Mid-Term Permanence Analysis, are taken as input for this method. The method is executed per single device.

16.1.2 PARAMETERS

- **Period of reference for the output:** default 6 months
    - period of analysis:
        - long_start_day: e.g. 01/01/2022
        - long_end_day: e.g. 30/06/2022
- **Sub-daily periods: inherited from mid-term method**
    - night: start_time = 8 p.m.; end_time = 8 a.m. the day after
    - working hours: start_time = 8 a.m.; end_time = 5 p.m.
    - evening hours: start_time = 4 p.m.; end_time = 10 p.m.
- **Sub-monthly periods: inherited from mid-term method**
    - working days (from calendar info)
    - holidays (from calendar info)

[10] See: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Usual_environment

  - weekends (from calendar info)
- **Sub-yearly periods:**
  - season (from calendar info)
- **Gap_ps_thresh**
- **Tot_ps_thresh:** 300 (average PS=5 per day in 60 days)
- **Freq_days_tresh:** 30% of number of days in the reference period (36 days for the 6-month period of reference)
- **UE_gap_thres:** 20% of the PSmax
- **UE_ps_thres:** 70% for daily period (*to be considered the need to set different default values on sub-periods indices*)
- **UE_ndays_thres:** 70% for daily period measures (*to be considered the need to set different default values on sub-periods indices*)
- **Home_ps_thres:** 80% for daily period measures (*to be considered the need to set different default values on sub-periods indices*)
- **Home_ndays_thres:** 80% for night-time sub-period measures (*to be considered the need to set different default values on other sub-periods indices*)
- **Work_ps_thresh:** 80% for workingdays&daytime sub-period measures (*to be considered the need to set different default values on other sub-periods indices*)
- **Work_ndays_thres:** 80% for workingdays&daytime sub-period measures (*to be considered the need to set different default values on other sub-periods indices*)

Thresholds, very important note:

Thresholds play a crucial role in this method. At the time of drafting this document, there was only limited knowledge about the distribution of the permanence score (PS), frequency and regularity index, especially regarding observations based on signalling data. Hence, the proposed default values should be considered as purely tentative and revised values will be proposed after the analysis of the above-mentioned distributions resulting from the first testing round in project's Task 5.

Potential parameters for subsequent implementation:

- **UE_meanfreq_thresh**
- **UE_stdfreq_thresh**
- **UE_meanreg_thresh**
- **UE_stdreg_thresh**

Similar threshold parameters for *Home Location, Work and Second Home* shall be defined (see below Long-Term Home Location Method).

Note: *the parameters' values should be coherent with the values of the same parameters used in the previous methods. These should be taken from parameters defined in the daily permanence score method and the mid-term* *permanence method.* ***The default values of all the methods' parameters can be revised after the testing phase according to the results of the analysis on real data.***

16.1.3 INPUT DATA

- Mid-Term Permanence Analysis – Output Data; see: Examples of output data
- Calendar Info [REFERENCE INPUT DATA]

16.1.4 OUTPUT DATA

- Long-term permanence score of the reference period (6 months) per sub-yearly, per sub-monthly and sub-daily periods (see the examples heading under this section).
- Long-term frequency count of the period of reference (6 months) per sub-yearly, sub-monthly and sub-daily periods.
- Long-term regularity indices of the period of reference (6 months) per sub-yearly, per sub-monthly and sub-daily periods.

16.1.5 SUB-FUNCTION OUTPUT DATA

- *Usual Environment* labels
- *Home Location* labels
- *Work* labels
- *Second Home* labels (not developed in the current description)

16.1.6 METHODOLOGY

The method processes indicators relative to all months in the reference period and returns the above listed output. More precisely, the method takes as input the output of the Method 1: Mid-Term Permanence Analysis, i.e. tables of permanence scores, frequency and regularity measures given per tile for different sub-daily and sub-monthly periods. Tables for all the months of the analysed period are collected and the long-term measures are provided.

The sub-daily, sub-monthly and sub-yearly periods are defined in the Parameters section above, and they are the following:

- night-time
- working hours
- weekends
- working days
- season
- holidays

The sub-yearly periods are given as suggested values; therefore, they can be configurable by the NSI in order to fit different lifestyles across countries.

Monthly summaries are processed by using the following procedures to provide long-term indicators. The procedures below represent the rationale to estimate the output:

1. the **long-term permanence score** for each tile is calculated as the sum of the tile monthly scores in the given period and for the considered sub-daily, sub-monthly and sub-yearly periods.

2. the **long-term frequency count** per tile is calculated by counting the number of days of the reference periods in which the tile has non-zero permanence score (in the mid-term summaries). Mean and standard deviation of monthly frequencies are also computed. All measures are calculated also for sub-daily, sub-monthly and sub-yearly periods.
3. **long-term regularity indices** per tile are calculated by taking the mid-term monthly mean distances between consecutive permanencies in the given tile (unit of measure is 'day') and by computing the mean and the standard deviation. The measures are to be calculated for sub-daily, sub-monthly and sub-yearly periods. In the following we refer to the monthly mean distances between consecutive visits in a tile as "monthly regularity indices".

The above indicators are reported in the long-term summaries and provided per all tiles having a non-zero value of the mid-term measures and per sub-daily, sub-monthly and sub-yearly periods and combinations of sub-periods (see the heading 'Examples under this section). In addition, the total assigned permanence score in the period of reference and the number of days in which the device has been observed are recorded in the last row of the tables as 'Device Observation'. They are calculated by summing up the monthly Device Observation values, which are denoted by **Tot_assigned_PS** and **Tot_observed_ndays**. They are calculated by summing up the monthly Device Observation values.

16.1.7 SUB-FUNCTION LABELS ASSIGNATION

16.1.7.1 FILTERS

Usual Environment and other labels are not assigned if:

- the ***device total assigned permanence score*** in the period of reference is lower than the given threshold *Tot_ps_thresh* (the device is denoted as '*rarely observed*').
- the ***device total frequency count on all tiles*** is lower than *Freq_days_thresh* (the device is denoted as '*discountinously observed*').

Note**: a condition to filter out 'highly moving' devices may be added later on, after the results of the first round of the testing phase in the project's Task 5.**

16.1.7.2 PRE-PROCESSING

***Note:** the parametrisation in the module allows to produce results for labels relative to different periods of reference. For example, the Home Location (according to the house of residence statistical definition) can be assigned for or based on a 6-month period and the Usual Environment on 1-year period basis, if needed.*

Pre-processing procedure:

Create a descending ordered list of tiles in terms of permanence scores (PSs). Calculate the difference of permanence score between consecutive tiles in the list. The first tile or group of tiles in the list are the tiles having the maximum PS value.

More precisely, we denote as 'gap' the first point in the descending list where the difference in PS exceeds a given threshold *Gap_ps_thresh.* We select all the tiles before the gap in the descending ordered list. If we want to select only the tile or the tiles with the maximum PS value, then the *Gap_ps_thresh* value needs to be set to 1.

All the procedures can be implemented also for measures in the sub-periods' tables. Therefore, in the same way, we can select first tiles in the permanence scores descending list for the different sub-periods (for example "night-time").

In the label assignation procedures below, relative values for the device activity are computed using the 'Device observation' values of the long-term indicator (see the tables in the 'Examples' heading under this section).

16.1.7.3 USUAL ENVIRONMENT LABELS

Let's define:

> **PSmax** = the maximum value of the device's PS in a tile (first row of the ordered list table, see the 'Examples' section).

Set the *Gap_ps_thresh* parameter equal to the *UE_gap_thres* value (default value = 20% of the **PSmax**) in the pre-processing step and identify the first groups of tiles in the list. By varying *UE_gap_thres* value it is possible to implement a more or less broad version of the UE.

Once detected the first group of tiles, assign the labels following the procedures described below.

***Note**: the Usual Environment output is given as a table per device and provided for a specific time period (default 6 months), as in the example shown below. A set of outputs relative to overlapping periods of 6 months (rolling window with one month lag) will give the dynamic identification of the device's Usual Environment.*

Implementation: The rules described below represent a tentative procedure to assign the *UE* labels and they should be considered as a first attempt to formalise the process. Hence, for the **first implementation** of this sub-function we propose a simplified version of the rules. Nonetheless, it may be useful to have the whole procedure in mind. **The procedure for the final implementation will be refined and confirmed after the analysis of the results of the first testing phase in the project's Task 5 and, if it will be the case, it can be modified according to a probabilistic approach.**

Quality measure: In the codes implementation, counters need to be inserted, one per each implemented rule. They allow to count how many cases are detected using the different procedure's rules and asses the utility and relevance of each rule. Moreover, they can help with the interpretation of the results of the first testing phase from Task 5 and are important to refine or change the label assignment procedure. In the labels output table there should be a column indicating, for each labelled tile, the assignation rule used (see the 'Examples' in this section).

The Usual Environment output is given as a table per device and it is provided for a specific time period (default 6 months, as in the example shown in the 'Examples' below). A set of outputs relative to overlapping periods of 6 months (rolling window with one month lag) will give the dynamic identification of the device Usual Environment.

16.1.7.4 USUAL ENVIRONMENT LABELS ASSIGNMENT – TENTATIVE PROCEDURE

Apply the following rules to each tile of the first groups of tiles in order to identify UE tiles:

StepA - rules on permanence score:

1. if the device was in the tile at least *UE_ps_thres* (default value 70%) of the total assigned PS (**Tot_assigned_ps**) the tile is labeled as *Usual Environment*. Rule code: a.1
2. check if the condition 1 is fulfilled for sub-periods and combination of sub-periods, with *UE_ps_thres* equal to the default value = 70% (possibly different default *UE_ps_thres* values for different sub-periods). In this case the tile is labelled as *Usual Environment*. Rule code: a.2

StepB - rules on frequency count and regularity indices:

1. if the device was in the tile at least *UE_ndays_thres* of the total number of days of observation (**Tot_observed_ndays**), the tile is labeled as *Usual Environment*. Rule code: b.1
2. check if the condition 1 is fulfilled for the sub-periods and combination of sub-periods (different thresholds applied). In this case the tile is labelled as *Usual Environment*. Rule code: b.2

3. if the mean monthly number of visits (frequency count) of the device in the tile is at least *UE_meanfreq_thres* (in the period of reference) and the standard deviation of the frequency count is lower than *UE_stdfreq_thres*, the tile is labeled as *Usual Environment*. Rule code: b.3
4. if the previous condition is not fulfilled check if the mean of the regularity index is lower than *UE_meanreg_thres* and the standard deviation of the regularity index is lower than *UE_stdreg_thres*. If it is the case the tile is labeled as *Usual Environment*. Rule code: b.4

UE labels assignment - Simplified rules version

Apply the following rules and parameter values to the first groups of tiles in order to identify UE tiles:

- StepA rule number 1 (Rule code: a.1)
- StepA rule number 2 (Rule code: a.2)
- if no tiles fulfill the conditions *UE* label is not assigned.

UE_gap_thres = 20% of the PSmax.

UE_ps_thres = 70%

UE_ndays_thres =70%

16.1.8 EXAMPLES

16.1.8.1 LONG-TERM MEASURES

Below we show examples of the output measures of the long-term permanence analysis method, displayed in a tabular form. Columns report the long-term indicators, while the row index indicates the tiles ids. The tile relative to the 'unknown location' is also included in the last row of the table.

The output tables are given per each considered sub-daily, sub-monthly and sub-yearly period and combinations of them.

6 months	Permanence score	Total frequency count	Monthly mean and *STD* of the frequency count	Regularity (mean of monthly regularity indices and *STD*)
tile_id[310]	50	11	7 (2)	4 (1)
tile_id[312]	120	40	1 (0)	15 (3)
tile_id[314]	60	18	15 (2)	2 (1)
tile_id[376]	20	11	5 (1)	9 (4)
tile_id[377]	200	70	10 (3)	5 (3)
Unknown	100	55	1 (0)	38 (4)
Device Observation	800	90		

6 months / weekends	Permanence score	Total frequency count	Monthly mean and *STD* of the frequency count	Regularity (mean of monthly regularity indices and *STD*)
tile_id[310]	20	6	3 (0)	
tile_id[312]	60	4	1 (0)	
tile_id[314]	25	8	4 (2)	
tile_id[376]	10	2	1 (1)	
tile_id[377]	30	2	2 (3)	

6 months / weekends	Permanence score	Total frequency count	Monthly mean and *STD* of the frequency count	Regularity (mean of monthly regularity indices and *STD*)
Unknown	100	10	4 (0)	
Device Observation	250	10		

16.1.8.2 LONG-TERM LABELS

Below we show an example of ordered lists we use to assign labels.

Ordered list - 6 months	PS (absolute value)	Frequency count
tile_id[377]	400	70
tile_id[312]	218	40
tile_id[314]	98	18
tile_id[310]	10	11
tile_id[376]	10	11
Unknown	100	55
Device Observation	534	116

Ordered list - 6 months	PS (relative value)	Frequency count (relative value)
tile_id[377]	75%	60%
tile_id[312]	40%	34%
tile_id[314]	18%	15%
tile_id[310]	2%	9%
tile_id[376]	2%	9%
Unknown	2%	13%
Device Observation	534	116

Ordered list - 6 months/weekends	PS (absolute value)	Frequency count
tile_id[377]	167	44
tile_id[312]	79	32
tile_id[314]	55	22
tile_id[310]	30	19
tile_id[376]	13	11
Unknown	82	40
Device Observation	100	35

16.2 METHOD 2: LONG-TERM HOME LOCATION METHOD

Note: For the sake of completeness, and to facilitate the reading and understanding of this method, several of the aspects already introduced in the description of Method 1: Long-Term Permanence Analysis are also kept in the description of this method, although up to a certain point this may be repetitive.

16.2.1 OBJECTIVE

The objective of this method is to obtain measures on a large time scale (e.g. 6 months, 1 year). It aims to produce as output a proxy for the *Usual Environment* of each device and a tentative identification of *Home Location*, and *Work/Study* place. The 'usual environment' (UE) refers to the specific location or setting where individuals typically spend the majority of their day or where they have their primary activities and interactions[11] (see for further details the M-UE indicators use case in Volume II).

Here we define rules to operationalise the above concepts. These rules include parameters and their default values, which can be customised by NSIs in each country according to their needs. Nevertheless, the adoption of default values can facilitate the comparability among countries.

In this method, mid-term output data are collected and analysed in order to extract information on the most visited locations in terms of regularity, frequency and time spent over a long-term period (e.g. 1 year, 6 months or a lower number of months). The regularity, frequency and permanence score mid-term indicators of all visited tiles, output of the Method 1: Mid-Term Permanence Analysis, are taken as input for this method. The method is executed per single device.

16.2.2 PARAMETERS

***Note:** the parameter's values need to be coherent with the values of the same parameters used in the previous methods, they need to be taken from parameters defined in the daily method and the mid-term method. The default values of all the methods parameters can be revised after the testing phase according to the analysis results on real data.*

- **Period of reference for the output**: default 6 months
- **Sub-daily periods**:
  - night: start_time = 8pm; end_time = 8am (on the next day)
  - working hours: start_time = 8am; end_time = 5pm
  - evening hours: start_time = 4pm; end_time = 10pm
- **Sub-monthly periods**:
  - working days (from calendar info)
  - holidays (from calendar info)
  - weekends (from calendar info)
- **Sub-yearly periods:**
  - season (from calendar info)
- **Tot_ps_thresh:** average PS per tile (average PS=6h per day in 80 days)

[11] See: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Usual_environment

- **Freq_days_tresh:** 20% of number of days in the reference period (36 days for a 6-months period of reference)
- **Gap_ps_thresh:** 1
- **home_perc_thres:** 80%
- **work_perc_thresh:** 80%
- **Radius_area:** 1km

Thresholds, very important note:

Thresholds play a crucial role in this method. At the time of drafting this document, there was only limited knowledge about the distribution of the permanence score (PS), frequency and regularity index, especially regarding observations based on signalling data. Hence, the proposed default values should be considered as purely tentative and revised values will be proposed after the analysis of the above-mentioned distributions resulting from the first testing round in project's Task 5.

16.2.3 INPUT DATA

- Mid-Term Permanence Analysis – Output Data; see: Examples of output data
- Calendar Info [REFERENCE INPUT DATA]

16.2.4 OUTPUT DATA

- Long-term permanence score of the period of reference (6 months) per sub-yearly, per sub-monthly and sub-daily periods (see a possible example in the example section).
- Long-term frequency count of the period of reference (6 months) per sub-yearly, sub-monthly and sub-daily periods.
- Long-term regularity indices of the period of reference (6 months) per sub-yearly, per sub-monthly and sub-daily periods.
- *Usual Environment* as group of tiles
- *Home Location* label
- *Work/Study* label

16.2.5 METHODOLOGY

The method processes indicators relative to all months in the reference period and returns the above listed output. More precisely, the method takes as input the output of the Method 1: Mid-Term Permanence Analysis, i.e. tables of permanence scores, frequency and regularity measures given per tile for different sub-daily and sub-monthly periods. Tables for all the months of the analysed period are collected and the long-term measures are provided.

The sub-daily, sub-monthly and sub-yearly periods are defined in the parameter section, and they are the following:

- night-time
- working hours
- weekends
- working days
- season
- holydays

The sub-yearly periods are given as suggested values; therefore, they can be configurable by the NSI in order to fit different lifestyles across countries.

Monthly summaries are processed by using the following procedures to provide long-term indicators. The procedures below represent the rationale to estimate the output:

1. the **long-term permanence score** for each tile is calculated as the sum of the tile monthly scores in the given period and for the considered sub-daily, sub-monthly and sub-yearly periods.
2. the **long-term frequency count** per tile is calculated by counting the number of days of the reference periods in which the tile has non-zero permanence score (in the mid-term summaries). Mean and standard deviation of monthly frequencies are also computed. All measures are calculated also for sub-daily, sub-monthly and sub-yearly periods.
3. **long-term regularity indices** per tile are calculated by taking the mid-term monthly mean distances between consecutive permanencies in the given tile (unit of measure is 'day') and by computing the mean and the standard deviation. The measures are to be calculated for sub-daily, sub-monthly and sub-yearly periods. In the following we refer to the monthly mean distances between consecutive visits in a tile as "monthly regularity indices".

The above indicators are reported in the long-term summaries and provided per all tiles having a non-zero value of the mid-term measures and per sub-daily, sub-monthly and sub-yearly periods and combinations of sub-periods (see the heading 'Examples under this section).

16.2.6 LABELS ASSIGNATION

16.2.6.1 FILTERS

Usual Environment and other labels are not assigned if:

- the ***device total assigned permanence score*** in the period of reference is lower than a given threshold *Tot_ps_thresh* (the device is denoted as '*rarely observed*').
- the ***device total frequency count on all tiles*** is lower than *Freq_days_thresh.*

Note**: a condition to filter out 'highly moving' devices may be added later on, after the results of the first round of the testing phase in the project's Task 5.**

16.2.6.2 PRE-PROCESSING

***Note:** the parametrisation in the module allows to produce results for labels relative to different periods of reference. For example, the Home Location (according to the house of residence statistical definition) can be assigned for or based on a 6-month period and the Usual Environment on 1-year period basis, if needed.*

Pre-processing procedure:

Create a descending ordered list of tiles in terms of permanence scores (PSs). Calculate the difference of permanence score between consecutive tiles in the list. The first tile or group of tiles in the list are the tiles having the maximum PS value.

More precisely, we denote as 'gap' the first point in the descending list where the difference in PS exceeds a given threshold *Gap_ps_thresh.* We select all the tiles before the gap in the descending ordered list. If we want to select only the tile or the tiles with the maximum PS value, then the *Gap_ps_thresh* value needs to be set to 1.

All the procedures can be implemented also for measures in the sub-periods' tables. Therefore, in the same way, we can select first tiles in the permanence scores descending list for the different sub-periods (for example 'night-time').

Computing the total time of observation (from the total assigned PS) of the device and the number of days of observation: collect the two quantities from the mid-term summaries and aggregate the values by summing each measure over the period of reference.

16.2.6.3 HOME AND WORK LABELS

Here we propose two tentative preliminary procedures to assign the Home Location label and the Work label. However, the procedure should be considered as a first attempt to formalise the process, since, as already stated in the paragraph on 'Thresholds', the procedures need to be completed and refined after the analysis of the preliminary test phase results of Task 5.

Set the parameter *Gap_ps_thresh* =1 in the pre-processing and identify the first tile or group of tiles in the list. Once detected the first tiles, divide the absolute values of permanence scores by the device's total assigned PS; in this way we get the tiles' **relative** permanence scores values. Then assign the labels following the procedures described below.

Home location label
Step1 - rules (on permanence score):

1. if the device was in the first tiles at least 80% (*home_perc_thres*) of the total time of observation (percentage of the total assigned PS) in the period of reference the tiles are labeled as *Home Location*.

Work label
Step1 - rules (on ps):

1. if the device was in the first tile at least 80% (*work_perc_thresh*) of the of the total time of observation in the workingdays&daytime sub-period the tiles are labeled as *Work*.

Note: *it is worth noting that the Work label assigned here cannot be used in the UC 4B 'Commuting'. For statistics on commuting, in fact, the requirements to define destinations are stricter than the ones used here. In particular, it is highly relevant to note that in the UE method the first step of analysis is the detection of 'stays' and then the labelling is applied to all stay locations. Differently, in the UC 4B 'Commuting', we focus on the origin-destination pairs, for which the observation unit is the displacement.*

16.2.7 EXAMPLE

16.2.7.1 OUTPUT: ASSIGNED LABELS

Labels	UE	Label
tile_id[377]	1	Home
tile_id[312]	1	Second home
tile_id[314]	1	Work
tile_id[310]	1	0
tile_id[376]	1	0

17 MODULE 16: DEVICE FILTERING & SINGLE MNO DATA AGGREGATION

17.1 DEVICE FILTERING AND AGGREGATION FOR USUAL ENVIRONMENT

17.1.1 OBJECTIVE

The objective of this method is to produce aggregated measures from the longitudinal analysis performed by the single MNO at the level of the single device. The method takes as input all the single device long-term outputs and produces aggregated figures within a single MNO.

Eventually, the method also disregards some devices (filtering) if, based on the results of the longitudinal analysis, they are not relevant for the calculation of the desired outputs. Hence, the filtering at this stage is UC specific.

The method processes data on the 100 x 100m spatial reference grid.

17.1.2 PARAMETERS

- **Filtering parameters:** they will be added as far as new UCs are developed (the filtering parameters are UC specific; in the UE UC filtering is not required).
- **Weighting parameters:** they depend on the available additional data (the weighting parameters are UC specific). Default: **weight_td** =1/n (tile t, device d, n being the number of tiles with UE, Home or Work label for the device d).

17.1.3 INPUT DATA

- Long-Term Labels, see: Long-term labels
- Grid Prior Land Use Probabilities [REFERENCE INPUT DATA] (Land use information optional)

17.1.4 OUTPUT DATA

- Single MNO counts per tile representing devices with Usual Environment in that tile
- Single MNO counts per tile representing devices with Home location in that tile
- Single MNO counts per tile representing devices with Work location in that tile
- Weights value for tiles in the devices sharing.

17.1.5 METHODOLOGY

This method produces the first aggregated measures of the processing pipeline, after the longitudinal analyses steps. The first aggregation performed in the workflow is the aggregation over different devices. The aggregation over the spatial dimension (over different tiles), to reach the requested spatial resolution for the resulting indicators, is supposed to be applied as one of the last steps (see Projection of the Multi-MNO Aggregates from the finest level to the geographic unit systems for the Usual environment).

The method process is composed by three steps: (1) an optional filter to keep only the devices relevant for the output indicators; (2) a procedure that weights the contribution of different tiles, in the number of devices counting; and (3) the summation over different devices. The steps are described in details in the following sub-sections.

17.1.5.1 STEP 1. FILTERING (OPTIONAL)

Devices are filtered according to UC specific processes, whenever needed (n.b. not needed for UE UC). If the filtering step is carried out, the number of disregarded devices is recorded as a quality metric.

17.1.5.2 STEP 2. WEIGHTING

The step implements the weighting of the share of different tiles for the aggregation. The choice of the weighting values can be done by exploiting information from other reference data available integrated in the workflow. An example, described in the option B, can be to integrate information on land use that helps limiting the error of the devices estimated per tile.

According to the available additional data (and according to the UC) different parameters can be used to define the weights. Here we describe two possible cases:

Option A - (no additional data)

A single device contributes equally in the devices aggregation per tile:

- if a tile t is labeled as UC for a device d, and
- if the device d has a total of n tiles labelled as UE

then the tile t weight of the device d is **weigth_td** =1/n. Hence, the device d contribution is equally divided among all tiles labeled as device d UE.

Option B - (with land use data):

A single device does not contribute equally in the devices aggregation per tile:

- if a tile t is labeled as UC for a device d, and
- if the device d has a total of n tiles labelled as UE, and
- if the tile t has other features taken from additional data

then the tile t weight of the device d is dependent of the additional data information. Here we take into account the availability of land use data; in this case: ***weight_td*** =1/n

- if the tile t corresponds to a residential area the ***weight_td*** =1/n
- if the tile t corresponds to a rural area the ***weight_td*** <1/n.

The tiles weights associated to a single device d must sum 1, hence a rescaling functions need to be applied.

For land use information, the use of CORINE Land Cover is encouraged. Data are available at a resolution up to 100 x 100m in the same grid system as used in this project (INSPIRE grid). This ensures that each grid tile corresponds to one and only one land use class.

Clearly, option A is a special case of option B.

17.1.5.3 STEP 3. AGGREGATION

In each tile, the weight corresponding to different devices are aggregated via summation, to produce an aggregated measure per each tile. For the above tile t, the final measure is calculated by summing all the tile weight_tdi.

17.1.6 EXAMPLE

Below we show examples of the outputs of the method, displayed in a tabular form.

Selected devices (if any) are reported in the last row of the tables.

UE table	**Weighted number of devices**
tile_id[310]	50.95
tile_id[312]	120.33
tile_id[314]	60.09
tile_id[376]	20.65
tile_id[377]	200.76
…	…
Filtered devices	100

Home Location table	Weighted number of devices
tile_id[310]	25.25
tile_id[312]	73.35
tile_id[314]	30.09
tile_id[376]	9.82
tile_id[377]	144.36
…	…
Filtered devices	100

Work location table	Weighted number of devices
tile_id[310]	30.76
tile_id[312]	80.33
tile_id[314]	16.09
tile_id[376]	22.65
tile_id[377]	177.76
…	…
Filtered devices	100

As further output, a table with tiles weight is provided:

Tiles weights	Weighting factors
tile_id[310]	1
tile_id[312]	1
tile_id[314]	0.5
tile_id[376]	0.3
tile_id[377]	0.7

18 MODULE 17: MERGE SINGLE MNO AGGREGATES IN MULTI-MNO AGGREGATES

18.1 MERGE SINGLE MNO AGGREGATES IN MULTI-MNO AGGREGATES FOR USUAL ENVIRONMENT

18.1.1 OBJECTIVE

The objective of this method is to combine data aggregated by devices deriving from different MNOs. This method produces multi-MNO aggregated measures processing the single-MNO aggregated measures at the level of one single tile.

18.1.2 PARAMETERS

- **α_MNOg**: intra-MNO deduplication factor to be used to deduplicate the devices' aggregates within the single MNO; it is provided per territorial area g.
- **α_g**: inter-MNO deduplication factor to be used to deduplicate the devices' aggregates between different MNOs; it is provided per territorial area g.

18.1.3 INPUT DATA

- Single MNO Aggregates
- Geographical areas (e.g. administrative units) for which the deduplication factors are provided including their correspondence with the reference processing grid.

18.1.4 OUTPUT DATA FOR THE DEVICE

- Multi-MNO aggregated number of devices per tile and UE, Home and Work labels (counts of deduplicated devices having UE label, Home Location label and Work Location label in the tile).

18.1.5 METHODOLOGY

18.1.5.1 STEP 1. INTRA-MNO DEDUPLICATION

Devices' aggregates for each single MNO produced by Device Filtering and Aggregation for Usual Environment method are taken as input and deflated according to the inter-MNO deduplication factor *α_MNOg*, specific per territorial areas g. This step produces deduplicated individual MNO aggregates per tile:

Deduplicated Single MNO aggregates in tile t = Single-MNO aggregates in tile t * **α_MNOg** (tile t contained in area g).

18.1.5.2 STEP 2. SUMMATION OVER MULTIPLE MNOS

Per each tile, deduplicated Single MNO aggregates are summed up.

For tile t:

Multi-MNO aggregate in tile t = Sum of Deduplicated Single MNO aggregates in tile t for all the MNOs having aggregates in tile t.

18.1.5.3 STEP 3. INTER-MNO DEDUPLICATION

Per each tile the total Multi-MNO aggregate is rescaled using the factor *α_g*.

Deduplicated Multi-MNO aggregate in tile t = Multi-MNO aggregate in tile t * **α_g** (tile t contained in area g).

18.1.6 EXAMPLES

Below we show an example of the output of the method, displayed in a tabular form. Columns represent the tile ids and the number of devices for UE, Home and Work labels obtained as aggregation of all the MNOs contributing to the analysis.

	Number of devices - UE label	**Number of devices - Home location label**	**Number of devices - Work location label**
tile_id[310]	150.95	45.25	15.95
tile_id[312]	120.33	65.38	70.50
tile_id[314]	190.11	360.09	244.09
tile_id[376]	420.23	175.65	320.65
tile_id[377]	430.71	202.62	500.76
…	…	…	…

19 MODULE 18: PROJECTION OF MULTI-MNO RELEVANT AGGREGATES FOR THE USE CASE

19.1 PROJECTION OF THE MULTI-MNO AGGREGATES FROM THE FINEST LEVEL TO THE GEOGRAPHIC UNIT SYSTEMS FOR THE USUAL ENVIRONMENT

19.1.1 OBJECTIVE

The objective of this method is to process multi-MNO measures at the grid tile level to obtain the multi-MNO measure at a geographical scale relevant for the UC. The method takes as input the measures provided by the output of Merge Individual MNO Aggregates in Multi-MNO Aggregates for Usual Environment (the multi-MNO aggregated counts per UE UC label), featuring the spatial resolution of the reference 100 x 100m grid. The spatial resolution of the output, that is the geographical scale of the output indicators, can be chosen by the user according to the NSI's needs for the specific official statistics. It can be a spatial subdivision of whatever type, such as administrative units at different levels, environmental classification of areas, geographical zones specified by their area size, etc.

19.1.2 PARAMETERS

- **Overlap_funselection:** integer identifying the selected sub-function to calculate tiles factors out of the overlap procedure 'Case1' (overlap between reference grid and output spatial units). Default value 1
- **Weight_per_tile:** percentage of the tile area in the geographical output unit
- **Radius_area:** default 1km (**it can be variable:** instead of a fixed value it can be dependent on the land cover feature of the area)

***Note**: Land cover input can give information to implement circle areas with different radius value. Circle areas corresponding to a rural area, in fact, may need to be larger than the ones corresponding to a residential area.*

19.1.3 INPUT DATA

- Multi-MNO aggregated number of devices
- Geographical system used for the reference grid tiles. Default: INSPIRE grid with LAEA projection.
- Geographical scale of the output (administrative units at different levels, environmental classification of areas, geographical zones specified by their area size, etc.), i.e. Geographical areas of interest. Default: NUTS5
- Grid Prior Land Use Probabilities [REFERENCE INPUT DATA] (Land use information optional)

19.1.4 OUTPUT DATA

- Number of devices per UE labelled geographical areas (multi-MNO aggregated counts per UE label at the geographical area level)
- The factors values (for quality measures)

19.1.5 METHODOLOGY

The method processes the multi-MNO aggregated counts at the single tile level to obtain the number of devices per UE labelled geographical areas (n.b. UE UC labels include UE label and Home/Work labels). Hence, the method performs a spatial aggregation of the input according to the described procedure.

The spatial aggregation procedure can be divided in two classes of procedure, according to the type of geographical areas chosen for the output indicators:

- Class 1: the output geographical areas are given (by shape file types of subdivision);
- Class 2: the output geographical areas are associated to 'points of interest' rather than given by already specified polygons, as administrative classification, land use zones, pre-determined grids. In this case the areas are taken as circle or square area around the point of interest.

19.1.5.1 CLASS 1 CASE

Create the overlap between the administrative boundaries and the INSPIRE grid geographical system. For each geographical unit belonging to the chosen output geographical system perform the weighted sum of the multi-MNO aggregated counts, per UE UC label, corresponding to all the tiles contained in the geographical unit. The weight of the multi-MNO aggregated count of each tile is defined as a factor calculated by one of the sub-functions described. Two alternative sub-functions are provided for this purpose. The input binary parameter is used to select the desired sub-function.

***Note**: The factors values per tiles are provided as a further output of the method for quality measures.*

\ SUB-FUNCTION 1

Parameter *Overlap_funselection*=1. The factor representing the tile weight is proportional to the area of the grid tile contained in the geographical unit of the output spatial scale. We provide an example in the following lines:

> We chose the municipality as the geographical scale for the output. Tiles completely inside the municipality boundaries have a weighting factor of 1, while tiles at the edges of boundaries of two municipality have a weighting factor equal to the fraction of the single tile area in each considered municipality.

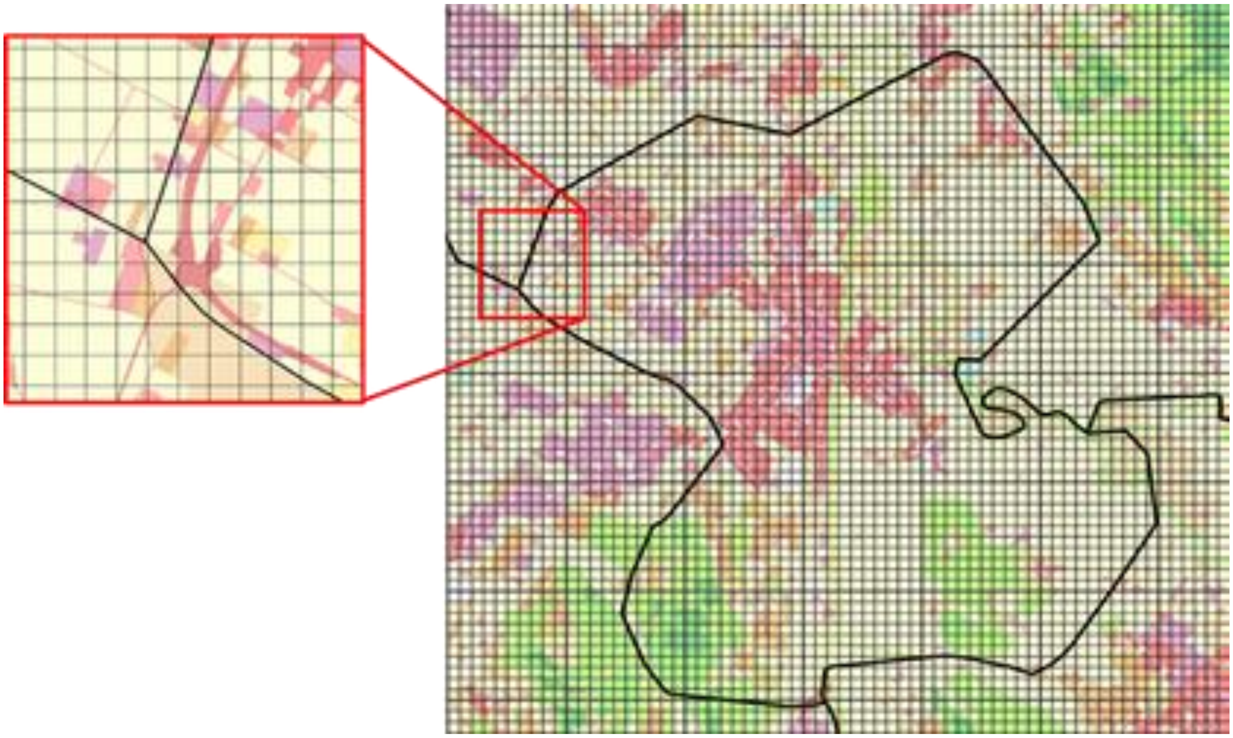


Figure 14: Example of administrative unit borders superimposed to the grid tessellation

The figure shows an example of administrative unit borders (in black) superimposed to the grid tessellation. The zoomed area highlights the intersection of both system boundaries, showing tiles fractions belonging to different municipalities.

\ SUB-FUNCTION 2

If the grid tiles have a 100 x100m size and, more in general, if the spatial resolution of the reference grid is much higher than the spatial resolution of the desired output (administrative units or areas radius), factors can be calculated in a simplified way with a negligible error. However, the choice is not automatic and an input parameter is used to select this sub-function when the user thinks it is the case.

Parameter *Overlap_funselection*=2. The factor representing the tile weight is equal to 1 if the centroids of the tile is contained in the geographical unit of the output spatial scale, otherwise it is equal to 0.

Note*: The calculation of the fractions serving as weights in sub-function 1 or the tiles centroids in sub-function 2 can be done once for all, independently from the rest of the processing, since the grid tiles system and the output geographical system are static.*

19.1.5.2 CLASS 2 CASE

Take as input the geographical points of interest (e.g. the center of each country province) and for each point detect the circle area around it, defined by a radius value *Radius_area* (default 1 km) around the labelled tiles as the *Home/Work location* area. For each circle area around the point, perform the sum of the multi-MNO aggregated counts, per UE label, corresponding to all the tiles contained in the circle unit. If circle areas corresponding to two different point of interest overlap, divide the counts corresponding to the overlapping areas equally between the two circle areas.

As in the previous case, the circle boundaries that lay inside a single tile take a fraction of the tile count equal to a weight factor. As well, the default for the weight is equal to the fraction of the tile area contained in the considered circle area.

19.1.6 EXAMPLES

Example of output tables:

Geographic unit	num Devices UE label	num Devices Home label	num Devices Work label
Id1	60	20	10
Id2	150	100	40
Id3	70	25	5
Id4	40	10	5

Tile	Geographical unit	Factor value
T1	Id4	1
T2	Id4	1
T3	Id4	0.8
T3	Id5	0.2
T4	Id5	1
T5	Id5	1

19.2 SPATIAL PROJECTION FOR THE M-HOME LOCATION

FROM TILES TO GEOGRAPHICAL AREAS OF INTEREST (E.G. ADMINISTRATIVE AREAS)

Once identified the labelled tiles we aggregate them to get locations with a lower spatial resolution, in order to have results at the desired spatial scale (e.g. NIL, 1km x 1km areas, NUTS, municipalities). The default for the spatial scale is NUTS5. More precisely:

1. **if the resulting geographical areas need to correspond to administrative areas**:

   Create the overlap between the administrative boundaries and the INSPIRE grid. Then point out the *Home/Work* labelled tiles and select the corresponding administrative areas as follows:

   - if the *Home/Work* labelled tile is contained in only one administrative area, take area as the *Home/Work location* administrative area.

   Note: the method to compute the overlap between the administrative boundaries and the INSPIRE grid and to deal with labelled tiles not completely contained in one administrative area will be added later on.

2. **if the resulting geographical areas do not need to correspond to administrative areas:**

   Take the circle area with *Radius_area* radius value (default 1km) around the labelled tiles as the *Home/Work location* area.

20 MODULE 19: ESTIMATION

General module, applicable to all longitudinal perspective UCs when MNO data is used as primary source for the statistical indicators. For an exemplification of the UE UC, please see Examples sub-section.

This method is limited to basic considerations and/or cases to demonstrate its integration in the pipeline. While its scope is sufficient for the goal of the Multi-MNO project, the method may be subject for refinement following the results of the MNO-MINDS research project.

20.1 OBJECTIVE

The objective of this module is to produce the final indicator, eventually by integrating MNO data with other external sources.

MNO data can be used as primary source or as auxiliary information in a statistical model. Only in the former case, these need to be weighted to represent the target population. Weighting is not needed for the use of MNO in statistical models, where MNO data can be processed directly in different model frameworks, which depend on the target indicators and the other available sources.

Here we describe the general weighting procedure for MNO data uses as primary source.

20.2 PARAMETERS

- Adjustment factor for under-coverage of the MNO data: w_d

20.3 INPUT DATA

- Multi-MNO aggregates at the relevant geographical system
- Input geographical system
- Output geographical areas
- Eventual external sources

20.4 OUTPUT DATA

- Weighted counts at geographical level required by the output

20.5 METHODOLOGY

The method processes the multi-MNO aggregated counts to obtain the target indicator. The method consists in multiplying the multi-MNO aggregated counts by an adjustment factor that might be different depending on the features of the territorial areas. The adjustment factor(s) represents the conversion factor from the device population to the target population, usually the resident population in the given country. It could be a generic one for the whole country or differentiated by geographic areas or other specific domains.

The assessment of this conversion factor requires some external information, compared to the MNO data.

20.6 EXAMPLE OF THE ESTIMATION PROCEDURE IN THE M-USUAL ENVIRONMENT INDICATORS USE CASE

From Projection of the Multi-MNO Aggregates from the finest level to the geographic unit systems for the Usual environment we have the aggregates of devices whose UE is in the selected geographic areas. Those counts are multiplied by the factor(s) **wd** where d denotes the geographic area.

In the following example, **wd** varies with the geographical area. However, we can also have a single **wd** constant across the areas.

Geographic area	Aggregates of devices	w_d	Aggregates of target population
1	60	1.1	Round(60* 1.1)=66
d	100	1.05	Round(100* 1.05)=1’5
D	300	1	Round(300*1)=300

21 HOME LOCATION CHANGES DETECTION METHODS

21.1 OBJECTIVE

The objective of this section is to detect a change in the home location of the target device in a given period of observation. This section describes the methods required for the construction of an indicator representing the migration flow of peoples in a given country between different administrative units. The default period of the analysis is 12 months.

21.2 LONG-TERM HOME LOCATION CHANGES DETECTION METHOD

The aim of the method is to check whether the device home location relative to the first 6 months is different from the device home location relative to the second 6 months for the given device. Hence, it takes as input the output of the Long-Term Home Location method, that provides the main device Home Location in a given period, in terms of a group of tiles. Then a procedure is implemented to compare the two home locations, i.e. the two groups of tiles, and to assess if a migration occurred in the period of analysis.

The module ensures flexibility, and all parameters can be adjusted. The method is executed per single device.

21.3 PARAMETERS

- **Start_date**: date corresponding to the start of the analysis period.
- **Obs_period**: integer number of months corresponding to the analysis period. Default: 12 months.
- **Input_num_month**: time frame which the HL input data object is defined on, expressed by an integer number of months. Default: 6 months.
- **Num_HLinput**: number of HL data objects taken as module input. Default = 2.
- **Mig_thresh**: threshold value for the home location change detection. Default = 0.5

21.4 INPUT DATA

- Home Location data objects, output of the Long-Term Home Location method (groups of tiles representing the HL of the individual device in the analysis period **obs_period**).
- Administrative areas of interest (defining the spatial resolution of the UC output).

21.5 OUTPUT DATA

- Migration table for the individual device

21.6 METHODOLOGY

The input of the module are the data objects produced by the Long-Term Home Location method. Each input data object is composed by a group of tiles indicating the device Home Location (HL) detected in a specific time frame (expressed in months) by the Long-Term Home Location method. The integer number of months of the time frame where the device HL is defined on, is indicated by the parameter *input_num_months* in this method.

For the given device at hand, the module takes in input multiple HL data objects. Specifically, the number of HL data objects taken as input is indicated by the parameter *num_HLinput.* The input HL data object corresponds to consecutive periods, spanning the whole analysis period obs_period. The time length of the periods is indicated by the parameter *input_num_months* and is expressed in number of months.

The default values of the parameters, listed below, are relative to the case of two HL input objects corresponding to two consecutive periods of 6 months each (period T1 and period T2), spanning a 12-month observation period:

- num_HLinput = 2
- input_num_month = 6 months
- obs_period = 12 months

The parameters values need to be set in order to fulfil the following constraints:

- *num_HLinput* = *obs_period/ input_num_months* is an integer (default *num_HLinput*= 12/6 months = 2).
- *num_HLinput* > 2

Below we describe the **steps of the procedure to detect the device migration.**

21.6.1 STEP 1. ASSIGNMENT OF WEIGHTS TO THE HOME LOCATIONS TILES

The input is given by two home locations (*num_HLinput*=2): one home location HL1 relative to the first 6 months (period T1) and one home location HL2 relative to the next 6 months (period T2). Each HL is associated to a group of tiles. Let *HL1_num_tiles* be the number of tiles of the HL1 group and *HL2_num_tiles* the number of tiles of the HL2 group, we assign a weight to each tile of the two groups as follow:

- *HL1_tiles_weight* = 1/HL1_num_tiles
- *HL2_tiles_weight* = 1/HL2_num_tiles

The tile weight represents the probability that the device HL is, in fact, located in the tile geographical area. The above weights' form denote that the uncertainty is equally distributed among all the tiles belonging to the HL group. A different tile weighting criterion may be chosen, if needed.

ILLUSTRATIVE CASE

For the handled device, we get two groups of tiles: one group defining the device HL relative to the first 6 months of the analysis period (HL1 period T1) and one group defining the device HL relative to the next 6 months of the analysis period (HL2 period T2); see **FIGURE 15**.

The HL1 group is composed by 5 tiles; therefore, we associate the weight *HL1_tiles_weigth* = 0.2 to each tile belonging to the HL1 group.

The HL2 group is composed by 4 tiles; therefore, we associate the weight *HL2_tiles_weigth* = 0.25 to each tile belonging to the HL2 group.

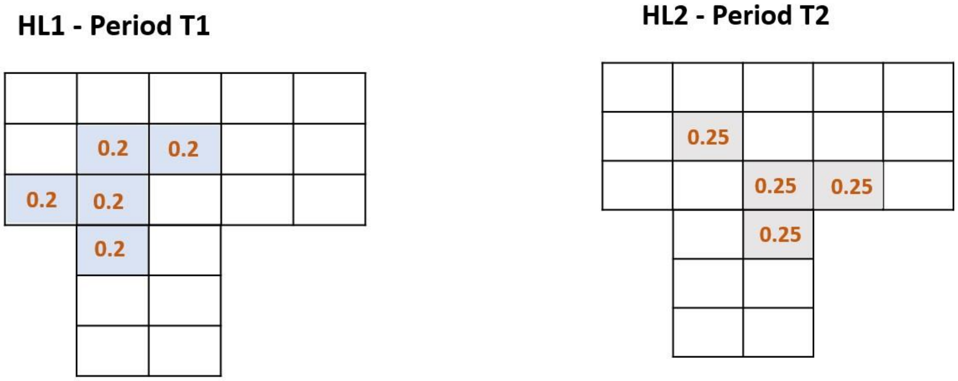


Figure 15: Representation of the groups of tiles composing HL1 (left) and HL2 (right)

Each tile of the HL1 group bears a probability of 0.2 to be the site of the device HL in the period T1 (left panel of the figure); each tile of the HL2 group bears a probability of 0.25 to be the site of the device HL in the period T2 (right panel of the figure).

21.6.2 STEP 2: FILTERING

Knowing the group of tiles representing the HLs for the device in the two consecutive periods, we can compare them and assess the magnitude of the difference. If the two groups are not identical, the magnitude of the difference with respect to a threshold value *Mig_thresh* is used to assess whether a possible change in HL (migration) occurred or the differences are an effect of the temporal fluctuations in the MNO's network coverage between the two 6-month periods.

Starting from the HL1 and HL2 groups of tiles, we compute the percentage of overlap between the two groups:

$$OVL\_perc = (2 \times share\_num\_tiles)/(HL1\_num\_tiles + HL2\_num\_tiles)$$

where *share_num_tiles* is the number of tiles belonging to both HL1 and HL2 groups of tiles.

If *OVL_perc* is

- equal or greater than *Mig_thresh:* no change occurred. The device is not considered as a migrating device and the next steps of the module are skipped.
- lower than *Mig_thresh:* a change in the HL occurred. The device is considered as a possible migrating device and the next steps of the module are computed.

The value of the *Mig_thresh* parameter can take values from 0 to 1 and it can be tuned according to NSIs' preferences.

If the *Mig_thresh* is set to 1, the next steps are computed, unless the two group of tiles are strictly identical. On the contrary, the lower the *Mig_thresh* value, the higher the number of no overlapping tiles needed to compute the next steps in the module.

ILLUSTRATIVE CASE

In our illustrative case HL1 and HL2 groups of tiles have 1 tile in common:

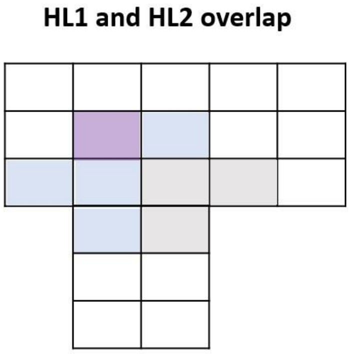


Figure 16: Representation of the overlap between the groups of tiles composing HL1 (light blue) and HL2 (grey): the common tile is highlighted in purple

We have the following value for the overlap percentage:

- *HL1_num_tiles* = 5
- *HL2_num_tiles* = 4
- *share_num_tiles* = 1
- *OVL_perc* = 2/9

21.6.3 STEP 3: SPATIAL AGGREGATION TO THE RESOLUTION OF INTEREST (METHOD 18.3 FOR INTERNAL MIGRATION UC)

In this step, we take into account a second input of the module: the geographical areas of the administrative units of interest for the internal migration statistics. This input defines the spatial resolution of the Internal Migration UC output, which is the geographical level of the internal migration statistical indicators. Examples of administrative units are regions, provinces, municipalities, etc.

The administrative units' geographical areas have a larger spatial scale with respect to the tiles. This step implements a spatial aggregation on the tiles' weights of the device HLs groups HL1 and HL2 in order to get values at the administrative unit level.

The method can be executed with administrative units input at different geographical level, allowing to obtain indicators at multiple spatial scales.

***Remark**: For a given device, the module provides an output only if migration between at least two administrative areas has been identified.*

\ CASE 1: ALL TILES OF HL1 AND HL2 FALL IN ONE ADMINISTRATIVE UNIT AREA

The case when all the tiles of an HL group fall in the same administrative unit area is trivial: the device HL is taken to be located in that administrative area with probability = 1.

If this is the case for both HL1 (all H1 tiles falling in the administrative area A) and HL2 (all H2 tiles falling in the administrative area C) the device is taken as a device migrated from administrative area A to administrative area C.

\ CASE 2: THE TILES OF HL1 FALL IN TWO ADMINISTRATIVE UNIT AREAS

When the tiles of HL1 group fall in two different administrative unit areas, administrative area A and administrative area B, the device HL1 is taken to be located in the administrative area A with a probability P_A equal to the sum of the weights of the HL1 tiles falling in the area A. The probability for the device HL1 to be located in the administrative area B, instead, is P_B = 1- P_A, and equal to the sum of the weights of the HL1 tiles falling in the area B.

The probabilistic approach leads to an output that provides the probabilities for the single device to have migrated from and to different administrative areas.

ILLUSTRATIVE CASE

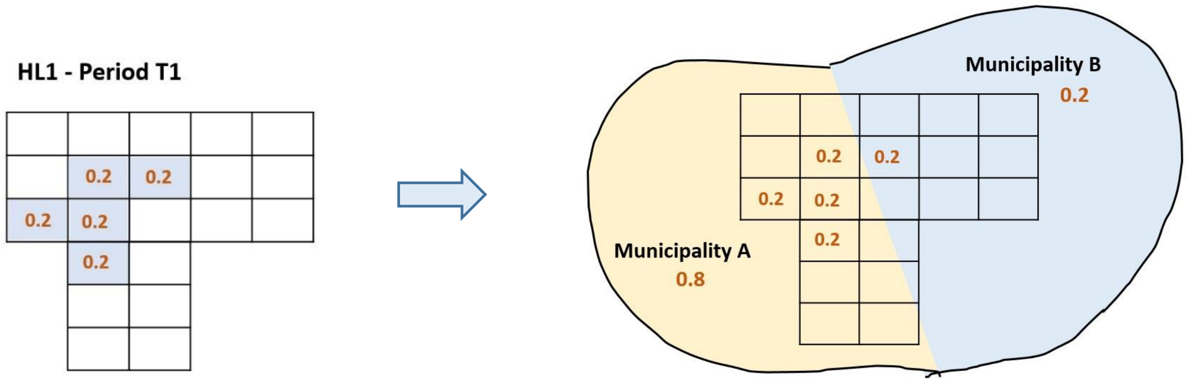


Figure 17: An outline showing the superposition of area and borders of municipality A (light yellow) and municipality B (light blue) on the area covered by the HL1 group of tiles

The group of tiles corresponding to the device home location HL1 falls in two different administrative unit areas, Municipality A and Municipality B. The described method computes the probability for the device home location HL1 to be located in the Municipality A as the sum of the weights of the tiles falling in the Municipality A: P_A = 0.2+0.2+0.2+0.2 = 0.8. In the same way, the probability the device home location HL1 is located in the Municipality B is P_B = 0.2.

\ CASE 3 - FROM HL1 TO HL2: THE TILES OF HL1 AND HL2 FALL IN TWO ADMINISTRATIVE UNIT AREAS

Let us consider the case where the tiles of HL1 group fall in two different administrative unit areas, Administrative area A and Administrative area B, and the tiles of HL2 group fall in two different administrative unit areas as well, Administrative area C and Administrative area D.

Similarly to as detailed before in Case 2, the device home location HL2 is taken to be located in the Administrative area C with a probability P_C equal to the sum of the weights of the HL2 tiles falling in the area C. The probability for the device home location HL2 to be located in the Administrative area D instead, is P_D = 1- P_C, and equal to the sum of the weights of the HL2 tiles falling in the area D.

After computing all the probabilities for both home locations to belong to the administrative areas A, B, C and D, the module calculates the migration flows probabilities for the handled device.

Given the uncertainty on the HLs administrative area of belonging, the possible migrations to be considered are:

- from A to C, with probability $P_{AC} = P_A \times P_C$
- from B to C, with probability $P_{BC} = P_B \times P_C$
- from A to D, with probability $P_{AD} = P_A \times P_D$
- from B to D, with probability $P_{BD} = P_B \times P_D$

The same holds true also if both HL1 and HL2 fall in the same pair of administrative units (HL1 in the Administrative area A and B e HL2 in the Administrative area A and B) (see the note at the end of the Illustrative case below).

ILLUSTRATIVE CASE

The following figure shows the device migration from HL1 to HL2. As shown before, the tiles of HL1 group fall in the Municipality A and B with probability 0.8 and 0.2 respectively. The tiles of HL2 group fall in the Municipality C and D with probability 0.6 and 0.4 respectively.

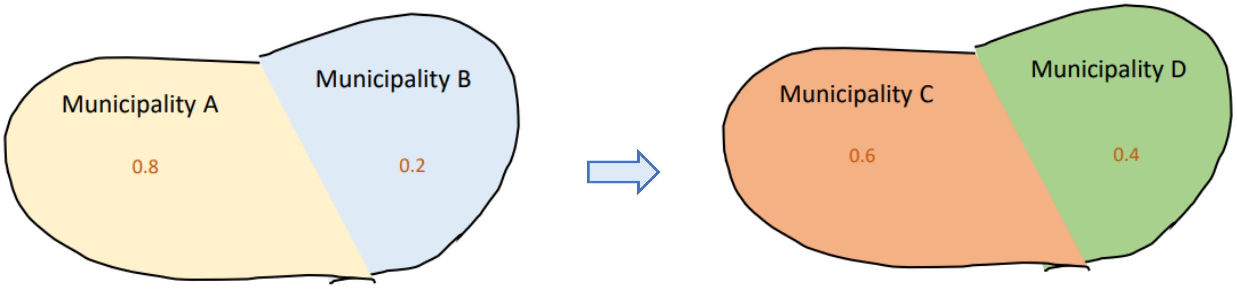
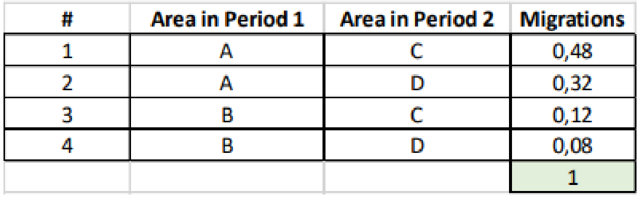


Figure 18: Outline showing: the superposition of municipality A (light yellow) and municipality B (light blue) areas on the area covered by the HL1 groups of tiles (left); the superposition of municipality C (orange) and municipality D (green) areas on the area covered by the HL2 groups of tiles (right)

According to the described method, in this case the migration probabilities are the one shown in the following table:



#	Area in Period 1	Area in Period 2	Migrations
1	A	C	0,48
2	A	D	0,32
3	B	C	0,12
4	B	D	0,08
			1

NOTE

Let us see the case in which both HL1 and HL2 fall in the same two administrative areas, like Municipality A and B. This case can occur when the home locations of the device in the two period T1 and T2 is in the same area covered by two adjacent administrative units. There are two possibilities:

1. The number of overlapping tiles between HL1 and HL2 is such that *OVL_perc < Mig_thresh*

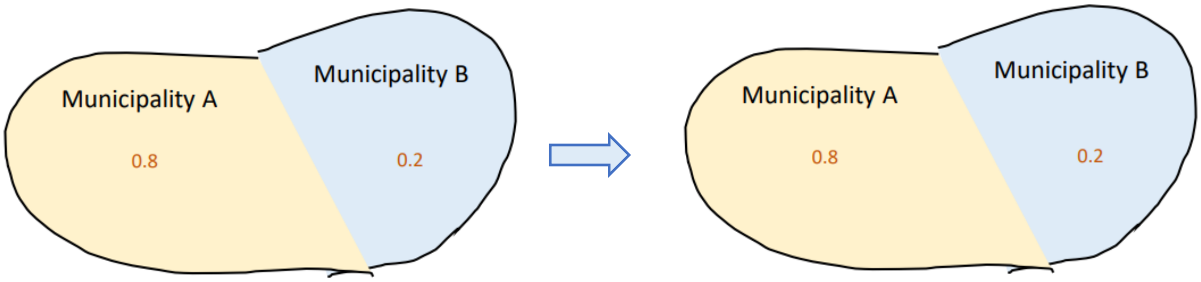
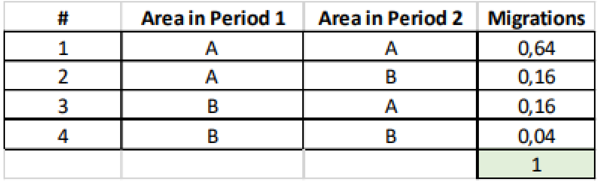


Figure 19: Outline showing the superposition of municipality A (light yellow) and municipality B (light blue) areas on the area covered by BOTH the HL1 group of tiles (left) and by the HL2 group of tiles (right)

In this case, given the uncertainty on the HLs administrative area of belonging, we need to take into account all the four directions of migration as in the table below. The output of this module, in this case, is the number of migrating devices in the period of analysis per each pair of administrative units.



#	Area in Period 1	Area in Period 2	Migrations
1	A	A	0,64
2	A	B	0,16
3	B	A	0,16
4	B	B	0,04
			1

2. The number of overlapping tiles between HL1 and HL2 is such that *OVL_perc ≥ Mig_thresh*

As explained before, if in this case the device is taken as not migrating, the aggregation steps are not computed and the output of the module is only the non-migrating flag for the device.

21.6.4 STEP 4: DEVICE AGGREGATION MODULE

The module takes as input the output of the previous step for all the devices in the processing. ***Note***: *The devices whose HL did not change in the analysis period are disregarded*.

For a single device, the module input is a table of migration probabilities between two or more administrative units.

After collecting all the input, the module sums up all the migration probabilities corresponding to each pair of administrative units. Each device, in fact, contributes to the count with its own migration probabilities, that reduce to 1 when the device migration involves only one migration route (Case 1 above).

The output table indicates the total number of devices migrating from an administrative unit to another, for each pair of administrative units.

OUTPUT

The output of this module is the number of migrating devices in the period of analysis per each pair of administrative units.

EXAMPLE

Number of migrating devices per each pair of administrative units

Administrative area pairs	num_ devices
Id_AdmArea[1]-Id_AdmArea[2]	23.7
Id_AdmArea[3]-Id_AdmArea[1]	102.2
Id_AdmArea[4]-Id_AdmArea[2]	15.0

22 DATA OBJECTS

22.1 INPUT DATA

22.1.1 CELL LOCATIONS WITH PHYSICAL PROPERTIES [INPUT]

CELL LOCATIONS WITH PHYSICAL PROPERTIES	
Description	Contains information about the location and physical properties of network cells for a specific day. Data updated along with MNO event data representing the network parameters for all active cells for a specific date.
Owner/Holder	MNO
Mandatory/Optional	Mandatory (in case of choosing the option of providing cell with physical properties)
Object/Unit/Record	Characteristic of a specific cell
Contents	**Mandatory fields**: • **cell_id**: ○ Type: String ○ Requirements: 14-digit or 15-digit numeric code following CGI and eCGI standards ○ Description: Code uniquely identifying one cell. • **latitude:** ○ Type: Float ○ Requirements: Latitude value in WGS84 system. Value has to be within WGS84 bounds. ○ Description: Latitude of cell location (location of the antenna). • **longitude:** ○ Type: Float ○ Requirements: Longitude in WGS84 system. Value has to be within WGS84 bounds. ○ Description: Longitude of cell location (location of the antenna). **Optional fields**: • **altitude:** ○ Type: Float ○ Requirements: ○ Description: Altitude (meters) of the antenna base from the sea level. • **antenna_height:** ○ Type: Float ○ Requirements: Positive value ○ Description: Height of the antenna in meters from ground • **directionality:** ○ Type: Integer ○ Requirements: value is either 0 or 1 ○ Description: 0 for omnidirectional antennas and 1 for directional antenas. • **azimuth_angle:** ○ Type: Float, nullable ○ Requirements: value between 0 and 360 if 'directionality' equal to 1, null otherwise.

CELL LOCATIONS WITH PHYSICAL PROPERTIES

  - Description: angle in degrees of the main propagation direction with respect to the North clockwise; for directional cells only.
- **elevation_angle:**
  - Type: Float
  - Requirements: value between -90 and 90
  - Description: Antenna placement angle; also known as tilt
- **horizontal_beam_width:**
  - Type: Float
  - Requirements: value between 0 and 360
  - Description: The angular extent of the cell beam in the horizontal plane
- **vertical_beam_width:**
  - Type: Float
  - Requirements: value between 0 and 360
  - Description: The angular extent of the cell beam in the vertical plane
- **power:**
  - Type: Float
  - Requirements: Positive value
  - Description: W
- **frequency:**
  - Type: Integer
  - Requirements: Positive value
  - Description: MHz
- **technology:**
  - Type: String
  - Requirements: One of the accepted values according to technology options parameter. E,g,: 5G, LTE, UMTS, GSM.
  - Description: Technology of the cell.
- **valid_date_start:**
  - Type: String
  - Requirements: String with date and time following ISO:8601 format: YYYY-MM-DDThh:mm.ss. Has to be earlier than **valid_date_end**.
  - Description: Start of time window in which the antenna is operational in this location. Period start timestamp is *included* within the time window.
- **valid_date_end:**
  - Type: String, nullable
  - Requirements: String with date and time following ISO:8601 format: YYYY-MM-DDThh:mm.ss. Has to be later than **valid_date_start**. It shall be set to null if it still operational.
  - Description: End of time window in which the antenna is operational in this location. Period end timestamp is *excluded* from the time window.
- **cell_type:**
  - Type: String
  - Requirements: One of the accepted values according to cell_type_options parameter.
  - Description: picocell, femtocell, etc.

Example:

cell_id	latitude	longitude	altitude	antenna_height	directionality	azimuth_angle	elevation_angle	horizontal_beam_width	vertical_beam_width	power	frequency	technology	valid_date_start	valid_date_end	cell_type
214030412038931	-3.62958	40.51873	20.0	42	1	90	4	65	9	3	3500	LTE	2023-07-20T10:00:00	2023-12-31T23:30:00	TBD
214035484123541	-3.8245	40.8952	30.5	12	0	null	5	42	9	7	1800	LTE	2023-07-20	null	TBD

22.1.2 CELL FOOTPRINT WITH DIFFERENTIATED SIGNAL STRENGTH COVERAGE AREAS [INPUT]

CELL FOOTPRINT WITH DIFFERENTIATED SIGNAL STRENGTH COVERAGE AREAS	
Description	Contains information about the area of each cell and spatial distribution of signal strength represented as geographical polygons. Data updated along with MNO event data representing the network parameters for all active cells for a specific date.
Owner/Holder	MNO
Mandatory/optional	Mandatory (if available, replaces Cell Locations with Physical Properties)
Object/Unit/Record	(Multi-)polygons of coverage areas of homogenous signal strength per one cell
Contents	**Mandatory Fields:** • **cell_id**: ○ Type: String or 64-bit integer ○ Requirements: 14-digit or 15-digit numeric code following CGI and eCGI standards ○ Description: Code uniquely identifying one cell. • **valid_date_start:** ○ Type: String ○ Requirements: String with date and time following ISO:8601 format: YYYY-MM-DDThh:mm.ss. Has to be earlier than **valid_date_end**. ○ Description: Start of time window in which the antenna is operational in this location. Period start timestamp is *included* within the time window. • **valid_date_end:** ○ Type: String, nullable ○ Requirements: String with date and time following ISO:8601 format: YYYY-MM-DDThh:mm.ss. Has to be later than **valid_date_start**. It shall be set to null if it still operational. ○ Description: End of time window in which the antenna is operational in this location. Period end timestamp is *excluded* from the time window. • **signal strength:** ○ Type: Decimal ○ Requirements: Value within interval (0,1] ○ Description: Normalized signal strength. • **geometry:** ○ Type: Geometry (WKT String) ○ Requirements: Valid (multi)-polygon geometry. TBD ○ Description: Polygonal coverage area of the cell.

Example:

cell_id	signal strength	geometry
214030412038931	0.8	[multipolygon]
214030412038931	0.2	[multipolygon]
214035484123541	0.5	[multipolygon]
214035484123541	0.3	[multipolygon]
214035484123541	0.1	[multipolygon]

22.1.3 MNO EVENT DATA – RAW [INPUT]

MNO EVENT DATA – RAW [INPUT]	
Description	'*MNO Event Data*' contains geolocation data from MNO subscribers. Data shall be created using at least one of the following data sources: (i) **CDRs** and/or (ii) **signalling data**. Additional information from MNO Apps can be added in order to improve the quality of the dataset, but this information is not mandatory. CDRs information shall contain all the information coming from voice, messages, internet connections, etc. CDRs shall also include roaming-in and roaming-out data. Each record of the dataset corresponds to a **MND event**, containing at least information about the identifier of the user, the timestamp of the event and the identifier of the cell to which the user is connected. When location information is estimated at point level (e.g. through signal triangulation or GPS data) information can be also be provided. This dataset shall only contain information about **personal mobile devices**. IoT, M2M and other related devices not associated to people shall not be included in the dataset.
Owner/Holder	MNO
Mandatory/Optional	Mandatory
Object/Unit/Record	Mobile network event associated to a specific subscriber
Contents	**Mandatory fields**: • **user_id:** ○ Type: Binary ○ Requirements: 32 bytes(256 bits) field. ○ Description: Unique pseudonymized identifier of the device, generated by hashing the user's IMSI using the SHA-256 function. • **timestamp:** ○ Type: String ○ Requirements: String with date and time in UTC following ISO:8601 format: YYYY-MM-DDThh:mm.ss ○ Description: Point in time where the event took place. • **mcc:** ○ Type: Integer ○ Requirement: 3 digits code ○ Description: Mobile Country Code derived from the user's IMSI. • **mnc:** ○ Type: String ○ Requirement: 2 or 3 digits code ○ Description: Mobile Network Code, a code of a home operator. It might help to assess the selectivity bias that is in place due to preferential roaming agreements between MNOs. This must be string, as it can start with 0 digit. Possible options can also be 01 or 001, so it cannot be integer. • **plmn:** ○ Type: Integer ○ Requirement: 5 or 6 digits code. Mandatory only for outbound data ○ Description: Network identifier of the foreign roaming partner MNO, consists of PLMN=MCC+MNC. • **cell_id:** ○ Type: String ○ Requirements: 14 or 15 character length string. All characters must be numbers. Optional if "latitude" and "longitude" are not null. ○ Description: Identifier of the cell following CGI and eCGI standards. • **latitude:** ○ Type: Float ○ Requirements: Latitude value in WGS84 system. Value has to be within WGS84 bounds. Optional if "cell_id" is not null. ○ Description: Latitude value of the location of the event.

MNO EVENT DATA – RAW [INPUT]

- **longitude**:
  - Type: Float
  - Requirements: Longitude value in WGS84 system. Value has to be within WGS84 bounds. Optional if “cell_id” is not null.
  - Description: Longitude value of the location of the event.

Optional fields:

- **loc_error**:
  - Type: Float
  - Requirements: Positive value. If “latitude” and “longitude” are null, this field shall be set to null.
  - Description: Location error in meters.

Example:

user_id	timestamp	mcc	mnc	plmn	cell_id	latitude	longitude	loc_error
000000000000..01	2023-01-01T00:00:00	214	01	null	214030412038931	-3.62958	40.51873	100.0
000000000000..10	2023-01-01T00:01:15	214	01	null	214030412038931	-3.62952	40.51871	100.0
000000000000..11	2023-01-01T12:05:03	214	01	null	214035484123541	null	null	null

22.1.4 GRID MODEL [REFERENCE INPUT DATA]

GRID MODEL [REFERENCE INPUT DATA]	
Description	INSPIRE grid geometry with additional information
Owner/Holder	Pipeline
Mandatory/Optional	Mandatory
Object/Unit/Record	grid centroid geometry with additional information
Contents	**Mandatory fields**: • **grid_id**: ○ Type: String ○ Requirements: string following INSPIRE specification format ○ Description: Code uniquely identifying one grid tile. • **geometry:** ○ Type: Binary ○ Requirements: ETRS89 Lambert Azimuthal Equal Area coordinate reference system (EPSG:3035) ○ Description: grid centroids point geometry **Optional fields**: • **elevation:** ○ Type: Float ○ Requirements: ○ Description: Elevation of a grid centroids • **land_use_main** ○ Type: string ○ Main land use category • **prior_probabilty_value** ○ Type: float ○ Prior probability value.

Example:

grid_id	elevaion	land_use_main	prior_probabilty_value	geometry
100mN4056000E5275300	12.1	RURAL	0.00	POINT()
100mN4056000E5275400	11.9	URBAN	0.70	POINT()

22.1.5 GRID PRIOR LAND USE PROBABILITIES [REFERENCE INPUT DATA]

GRID PRIOR LAND USE PROBABILITIES [REFERENCE INPUT DATA]	
Description	The prior probabilities per grid tile with land use values.
Mandatory/Optional	Mandatory
Object/Unit/Record	Grid tile
Contents	**Mandatory fields:** • **valid_date_range** o Type: Array o The date to which the properties apply. • **grid_id** o Type: o Unique ID of grid tile • **land_use_main** o Type: string o Main land use category • **prior_value** o Type: float o Prior probability value.

Example:

grid_id	valid_date_range	land_use_main	prior_value
123231342131341	[2023-07-20, 2025-07-20]	RURAL	0.00
123231342131342	[2023-07-20, 2025-07-20]	URBAN	0.70
123231342131343	[2023-07-20, 2025-07-20]	FOREST	0.10
123231342131344	[2023-07-20, 2025-07-20]	WATER	0.05

22.1.6 CALENDAR INFO [REFERENCE INPUT DATA]

CALENDAR INFO [REFERENCE INPUT DATA]	
Description	This data object includes the classification of calendar days into national holidays or regular days. It is a data object that does not take input from other data objects in the pipeline. One object per year.
Owner/Holder	Open data, which can be collected by the 'author' implementing the first module of the pipeline where the object is used.
Mandatory/Optional	Mandatory
Object/Unit/Record	information on type of days for all years covered (one data object per year)
Contents	• **'day'**: date of all the days in the year covered. Type: datetime • **'Italy'**: 0 if the day is a regular day, 1 if it is a holyday for country Italy. Type: binary. • **'Spain'**: 0 if the day is a regular day, 1 if it is a holyday for country Spain. Type: binary. • **'day_type'**: Type: character. String describing the weekday type: "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday". • + one column per country

Example:

day	**Italy**	**Spain**	**weekday**
01/01/2022	1	1	Saturday
01/02/2022	0	0	Sunday
01/03/2022	0	0	Monday
01/04/2022	0	0	Tuesday
01/05/2022	0	0	Wednesday
01/06/2022	1	1	Thursday

22.2 INTERMEDIATE RESULTS

22.2.1 CLEAN MNO NETWORK TOPOLOGY DATA [INTERMEDIATE RESULTS]

This data object follows the same format as the input Network Topology Data used, respectively:

- See data object format for Cell Locations with Physical Properties [INPUT]
- Cell Footprint with Differentiated Signal Strength Coverage Areas [INPUT]

22.2.2 MNO NETWORK TOPOLOGY DATA QUALITY METRICS [INTERMEDIATE RESULTS]

MNO NETWORK TOPOLOGY DATA QUALITY METRICS [INTERMEDIATE RESULTS]	
Description	Quality metrics produced by Module/Method I: MNO Network Topology Data Cleaning – Syntactic Checks
Mandatory/Optional	Mandatory
Object/Unit/Record	Quality metrics for each collection of Network Topology Data
Contents	Description of all the fields and values using bullet points: • **result_timestamp**: ○ Type: Timestamp ○ Requirements: - ○ Description: Timestamp of the start of the process when the metrics were produced. One process can generate multiple metrics. This can also be saved as the metadata for this data object. • **date**: ○ Type: Date ○ Requirements: Same value as provided by the process input parameter *date* ○ Description: Date of the dataset to which the quality metrics refer (not from topology data but from parameters) • **field_name**: ○ Type: String ○ Requirements: Either null or same as the name of a column present in input data ○ Description: Name of the field to which the metric refers to. Value is null if the metric refers to multiple fields. • **type_code**: ○ Type: Integer ○ Requirements: One value from the type codes (see table below). ○ Description: Numeric code indicating the type of the metric. See table below. • **accumulated_percentage:** ○ Type: Float ○ Description: Accumulated percentage with respect to the total number of invalid values, accumulated from the most frequent error up to this one, included. • **value**: ○ Type: Integer ○ Requirements: - ○ Description: Numeric value of the metric.

Codes types:

Code	Short description
0	no errors
1	value is null
2	value is not within the set of accepted values
3	unsupported input data type
4	unable to parse correctly
100	total rows at the start of method
101	total rows at the end of method

Example:

date	field_name	type_code	value
01-01-2023	cell_id	0	1900
01-01-2023	cell_id	1	95
01-01-2023	cell_id	2	5
01-01-2023	-	100	2000
01-01-2023	-	101	1900

Example Top Frequent Errors:

result_timestamp	field_name	type_code	error_value	error_count	accumulated_percentage	year	month	day
01-01-2023	cell_id	2	000000	400	40.0	2023	1	1
01-01-2023	cell_id	1	null	300	70.0	2023	1	1
01-01-2023	cell_id	2	123456789	200	90.0	2023	1	1
01-01-2023	cell_id	2	xxx123	50	95.0	2023	1	1
01-01-2023	cell_id	2	AVSADD	50	100.0	2023	1	1

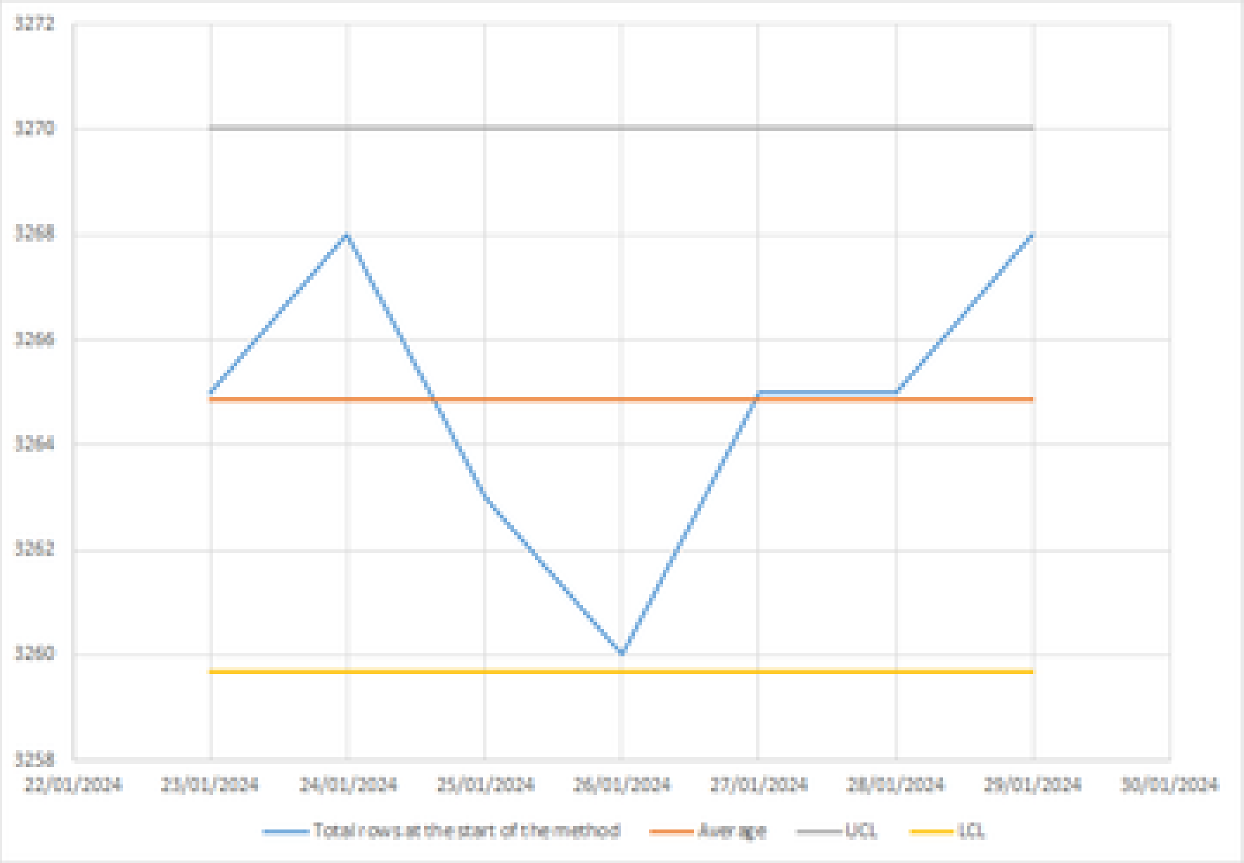
22.2.3 MNO NETWORK TOPOLOGY DATA QUALITY WARNINGS [INTERMEDIATE RESULTS]

MNO NETWORK TOPOLOGY DATA QUALITY WARNINGS [INTERMEDIATE RESULTS]	
Description	Summary report produced by Module/Method 2: MNO Network Topology - Quality Warnings presenting plots of the main metrics and quality warnings in the case the quality metrics is over one of the threshold. Cases that generate warnings should be also stored in a log table. For the implementation of the pipeline in the test environment can be a summary report, in further developments it can become a dynamic dashboard. The analysis of the output can give insights on relevant errors in the input topology data that can affect the quality of statistical output and can also provide hints on the errors causes that could be removed in data for next periods.
Mandatory/Optional	Mandatory
Object/Unit/Record	Quality warnings
Contents	The report will have the title "MNO Network Topology Data Quality Warnings" + the date to which it is referred. It will always present 3 plots: • plot 1. Size of the MNO Network Topology Data, either one of the options Cell Locations with Physical Properties [INPUT] or Cell Footprint with Differentiated Signal Strength Coverage Areas [INPUT] Days on horizontal axis and the value of 'Total rows at the start of the method' from the MNO Network Topology Data Quality Metrics [INTERMEDIATE RESULTS] on the vertical axis. The period on the horizontal axis is set as a parameter of the method. It would be useful to have in the same plot in the vertical axis also the average of the metrics and the control limits • plot 2. Size of the Clean MNO Network Topology Data [INTERMEDIATE RESULTS] Days on horizontal axis and the value of 'Total rows at the end of the method' from the MNO Network Topology Data Quality Metrics [INTERMEDIATE RESULTS] on the vertical axis. The period on the horizontal axis is set as a parameter of the method. It would be useful to have in the same plot in the vertical axis also the average of the metric and the control limits • plot 3. Error rate Days on horizontal axis and the value of error rate calculated as specified in Module/Method 2: MNO Network Topology - Quality Warnings on the vertical axis. The period on the horizontal axis is set as a parameter of the method. It would be useful to have in the same plot in the vertical axis also the average of the metric and the upper control limit Then the report will include warnings only in the case the daily value of the metrics is out of the thresholds indicated in Module/Method 2: MNO Network Topology - Quality Warnings. In such cases the following information should be reported: **ATTENTION WARNING** <Measure definition> = Daily Value - <Condition for the warning>= value of the limit **<Warning message>** e.g. Error rate = 25% - Error rate is over the upper control limit calculated on the basis of average and standard deviation of the distribution of the error rate in previous **week**. Upper Control limit = **10%** **The error rate after syntactic checks application is unexpectedly high with respect to previous period, taking into account its usual variability** Finally the Composition of errors for each field should be reported with a pie chart. The warnings should also be stored in a log table with the same information, the title and the date. The same log table can be used for all quality warnings. It can be agreed a certain time after which the log table can be cleaned.

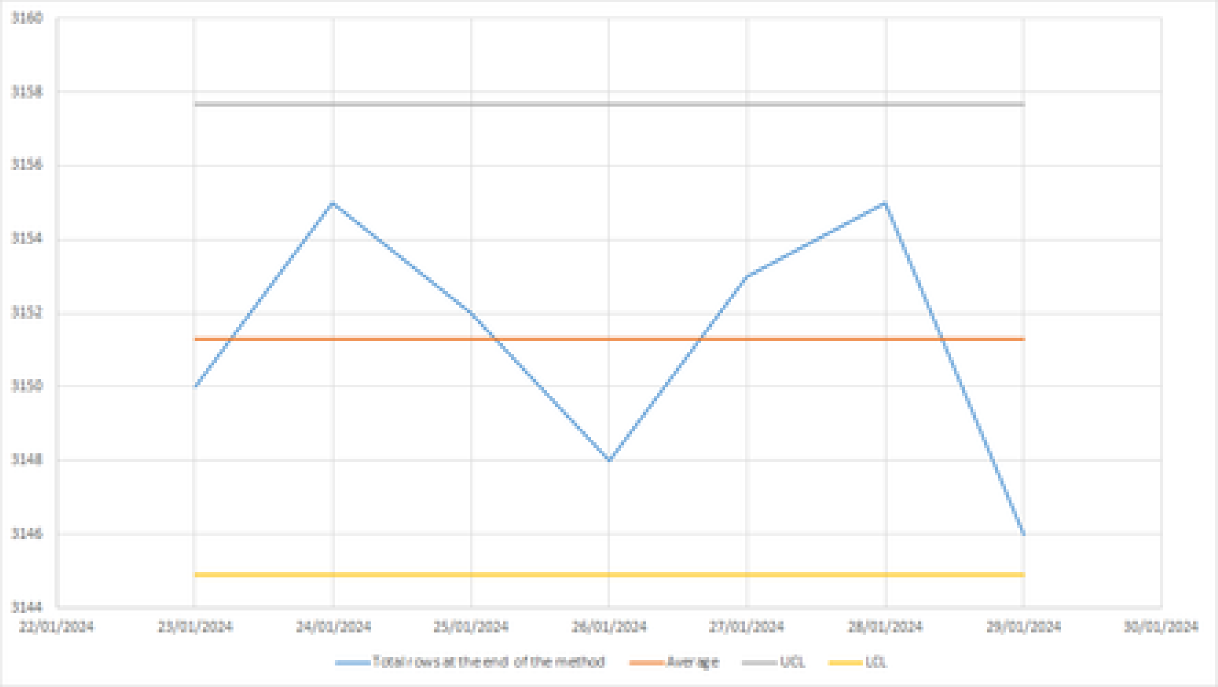
Examples:

MNO Network Topology Data Quality Warnings - 29/01/2024

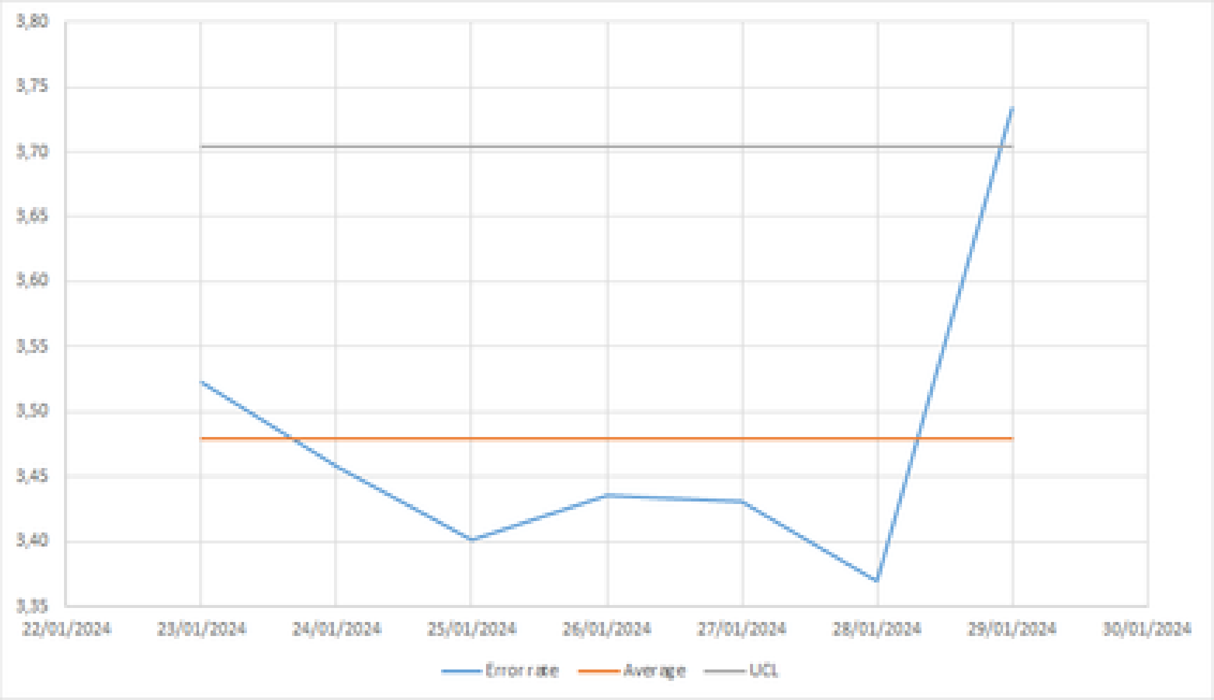
1. Size of the MNO Network Topology Data, either one of the options Cell Locations with Physical Properties [INPUT] or Cell Footprint with Differentiated Signal Strength Coverage Areas [INPUT]



2. Size of the Clean MNO Network Topology Data [INTERMEDIATE RESULTS]



3. Error rate



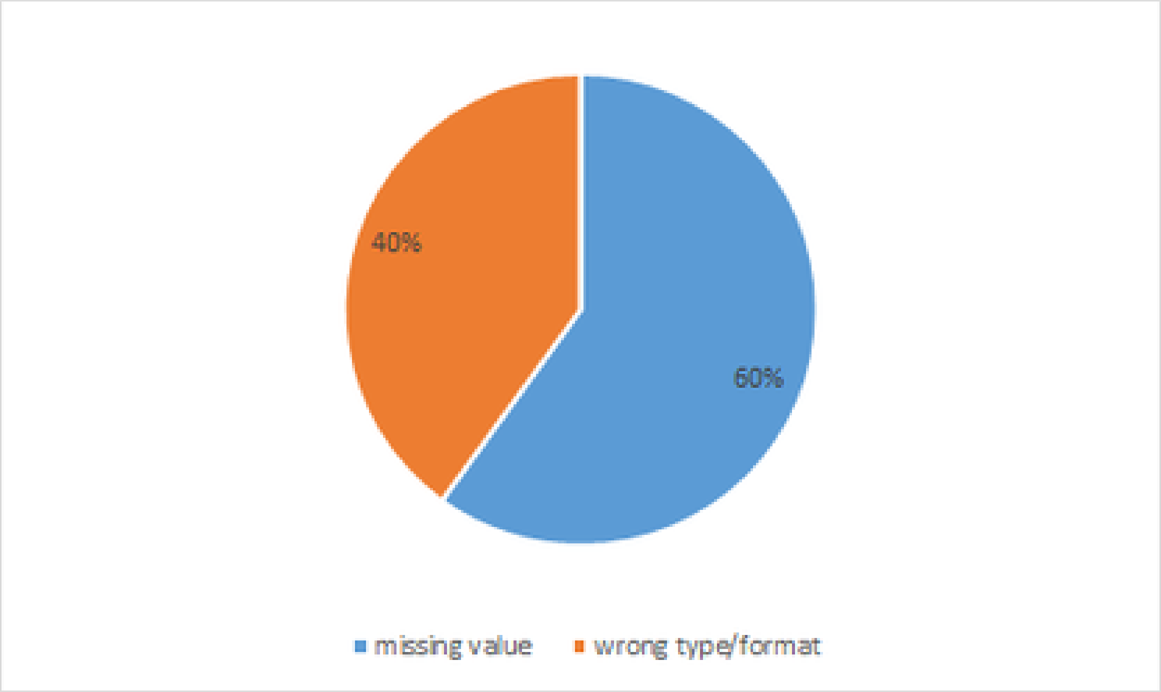
ATTENTION WARNING

Error rate = 3.73% - Error rate is over the upper control limit calculated on the basis of average and standard deviation of the distribution of the error rate in previous week. Upper Control limit = **3.70%**

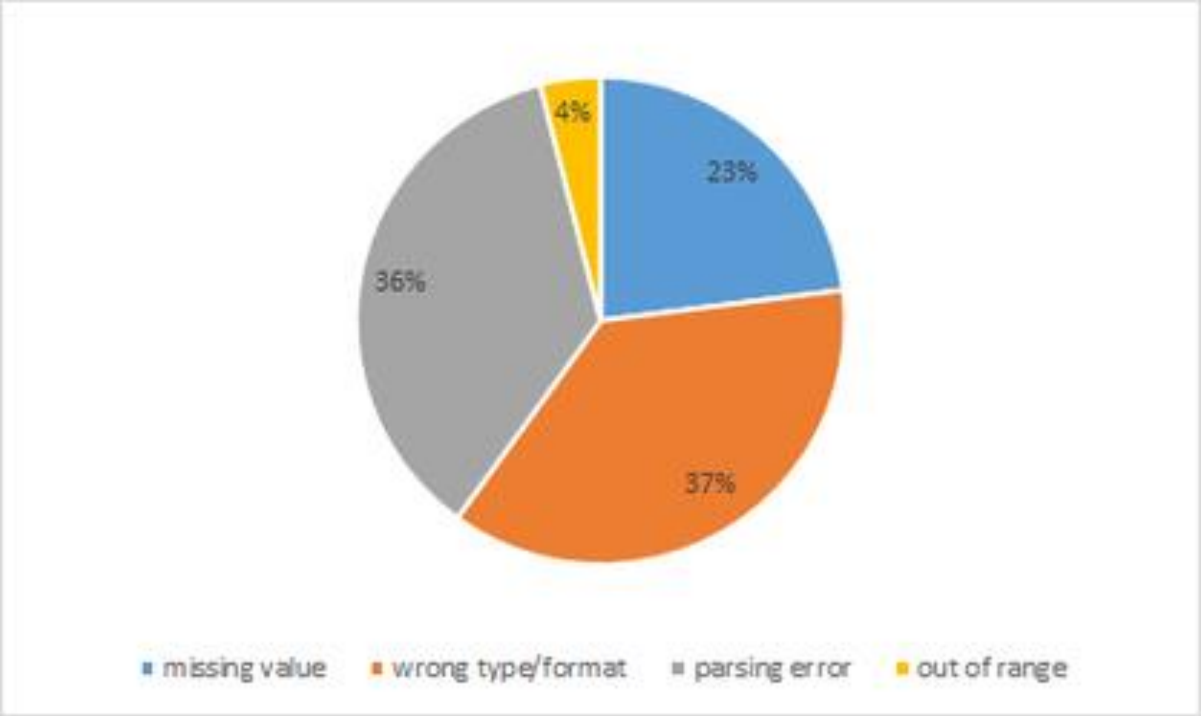
The error rate after syntactic checks application is unexpectedly high with respect to previous period, taking into account its usual variability

Composition of errors:

1. cell_id



2. geometry


4%
23%
36%
37%
■ missing value
■ wrong type/format
■ parsing error
■ out of range

[…]

Log table example:

Title	Date	Measure_definition	Daily_value	Condition	Parameter_time	Value_condition	warning_text
MNO Network Topology Data Quality Warnings	2024-01-29 00:00:00	Error rate	3.73	Error rate is over the upper control limit calculated on the basis of average and standard deviation of the distribution of the error rate in previous	week	**3.70**	**The error rate after syntactic checks application is unexpectedly high with respect to previous period, taking into account its usual variability**

22.2.4 CELL SIGNAL STRENGTHS [INTERMEDIATE RESULTS]

CELL SIGNAL STRENGTHS [INTERMEDIATE RESULTS]	
Description	The signal strength values per cell per grid tile
Mandatory/Optional	Mandatory
Object/Unit/Record	Cell / grid tile combination
Contents	Mandatory fields: • **cell_id** ○ Unique ID of cell. • **grid_id** ○ Unique ID of grid tile • **valid_date_start** ○ Start date of validity period • **valid_date_end** ○ End date of validity period • **signal_strength** ○ Signal strength in dBm Optional fields: • **distance_to_cell** ○ Distance of grid tile to cell location may be necessary for some calculation during the Location Assignation Module (e.g., taking into account the Timing Advance parameter of the MNO event data).

Example:

cell_id	grid_id	valid_date_start	valid_date_end	signal_strength	distance_to_cell
214030412038931	123231342131341	2023-01-01	2025-01-01	0.5405	4623
214030412038931	123231342131342	2023-01-01	2025-01-01	0.4193	4627
214030412038932	123231342131341	2023-02-01	2025-01-01	0.9744	4629

22.2.5 CELL FOOTPRINT VALUES [INTERMEDIATE RESULTS]

CELL FOOTPRINT VALUES [INTERMEDIATE RESULTS]	
Description	The cell footprint values per grid tile
Mandatory/Optional	TBD
Object/Unit/Record	Cell / grid tile combination
Contents	**Mandatory fields:** • **cell_id** ○ Unique ID of cell • **grid_id** ○ Unique ID of grid tile • **valid_date_start** ○ Start date of validity period (inclusive) • **valid_date_end** ○ End date of validity period (exclusive) • **footprint** ○ Cell footprint value (0 to 1)

Example:

cell_id	grid_id	valid_date_start	valid_date_end	footprint
123456789101112	123231342131341	2023-01-01	2023-01-02	0.5405
123456789101112	123231342131342	2023-01-01	2023-01-02	0.4193
123456789101112	123231342131343	2023-01-01	2023-01-02	0.9744
123456789101112	123231342131344	2023-01-01	2023-01-02	0.1633

22.2.6 CELL CONNECTION PROBABILITIES [INTERMEDIATE RESULTS]

	CELL CONNECTION PROBABILITIES [INTERMEDIATE RESULTS]
Description	The cell connection probabilities per grid tile. Given that a device is located in a certain grid tile, then how likely is it to be connected to a certain network cell.
Mandatory/Optional	TODO If the data object is mandatory or optional
Object/Unit/Record	grid_id + cell_id with probability for a given date
Contents	**Mandatory fields**: • **cell_id:** ○ Type: String ○ Description: Unique ID of cell. • **grid_id:** ○ Type: String ○ Description: Unique ID of grid tile. • **valid_date_start** ○ Start date of validity period • **valid_date_end** ○ End date of validity period • **cell_connection_probability:** ○ Type: 64-bit float ○ Requirements: Within range [0,1] ○ Description: cell connection probability value.

Example:

CELL_ID	GRID_ID	VALID_DATE	CELL_CONNECTION_PROBABILITY
34528375	123231342131341	2023-07-20	0.5405
34528375	123231342131342	2023-07-20	0.4193
34528375	123231342131343	2023-07-20	0.9744
34528375	123231342131344	2023-07-20	0.1633
34528375	123231342131411	2023-07-20	0.8369
34528375	123231342131412	2023-07-20	0.0963
34528375	123231342131413	2023-07-20	0.8667
95658724	123231342131341	2023-07-20	0.6977

22.2.7 POSTERIOR PROBABILITIES VALUES [INTERMEDIATE RESULTS]

POSTERIOR PROBABILITIES VALUES [INTERMEDIATE RESULTS]	
Description	Cell connection probability values per grid tile along with posterior probabilities. Cell connection probabilities indicate, that given that a device is located in a certain grid tile, how likely is it to be connected to a certain network cell. Posterior probabilities are the posterior of cell connection probabilities, given the use of prior probabilities.
Mandatory/Optional	Optional
Object/Unit/Record	Grid tile, cell id and date combination
Contents	**Mandatory fields:** • **cell_id** ○ Type: String ○ Description: Unique ID of cell. • **grid_id** ○ Type: String ○ Requirements: string following INSPIRE specification format ○ Description: Code uniquely identifying one grid tile. • **valid_date_start** ○ Start date of validity period from cell connection probabilities ○ End date of validity period from cell connection probabilities • **cell_connection_probability** ○ Type: 64-bit float ○ Requirements: Within range [0,1] ○ Description: cell connection probability value. • **posterior_probability** ○ Type: float ○ Posterior probability value

Example:

cell_id	grid_id	cell_connection_probability	posterior_probability
214030412038931	123231342131	0.5405	0.21
214030412038931	123231342131	0.4193	0.70
214030412038931	123231342131	0.9744	0.00
214030412038931	123231342131	0.1633	0.70
214030412038931	123231342131	0.8369	0.70
214030412038931	123231342131	0.0963	0.70
214030412038931	123231342131	0.8667	0.52
214030412038931	123231342131	0.6977	0.00

22.2.8 CLEAN MNO EVENT DATA [INTERMEDIATE RESULTS]

CLEAN MNO EVENT DATA [INTERMEDIATE RESULTS]	
Description	This data is basically the same as MNO Event Data – Raw [INPUT]. The only difference is that the events with syntactic errors have been removed.
Mandatory/Optional	Mandatory
Object/Unit/Record	Mobile network event associated to a specific subscriber
Contents	**Mandatory fields**: • **year:** ○ Type: Integer 16 ○ Requirements: Integer of 16 bits. ○ Description: Year the event took place. • **month:** ○ Type: Integer 8 ○ Requirements: Integer of 8 bits. ○ Description: Month the event took place. • **day:** ○ Type: Integer 8 ○ Requirements: Integer of 8 bits. ○ Description: Day the event took place. • **user_id:** ○ Type: Binary ○ Requirements: 32 bytes (256 bits) field. ○ Description: Unique pseudonymised identifier of the device. • **user_id_modulo:** ○ Type: Integer ○ Requirements: Integer of 8 bits. ○ Description: Modulo division result, as applied to the integer part of the user_id column. • **timestamp:** ○ Type: Time ○ Requirements: Parquet time type in hour, minutes and seconds. ○ Description: Point in time where the event took place. • **mcc:** ○ Type: Integer ○ Requirement: 3 digits code ○ Description: Mobile Country Code derived from the user's IMSI. • **mnc:** ○ Type: String ○ Requirement: 2 or 3 digits code ○ Description: Mobile Network Code, a code of a home operator. It might help to assess the selectivity bias that is in place due to preferential roaming agreements between MNOs. This must be string, as it can start with 0 digit. Possible options can also be 01 or 001, so it cannot be integer. • **plmn:** ○ Type: Integer ○ Requirement: 5 or 6 digits code. Mandatory only for outbound data ○ Description: Network identifier of the foreign roaming partner MNO, consists of PLMN=MCC+MNC. • **cell_id:** ○ Type: String ○ Requirements: 14 or 15 character length string. All characters must be numbers. Optional if “latitude” and ”longitude” are not null. ○ Description: Identifier of the cell following CGI and eCGI standards.

CLEAN MNO EVENT DATA [INTERMEDIATE RESULTS]

- **latitude:**
  - Type: Float
  - Requirements: Latitude value in WGS84 system. Value has to be within WGS84 bounds. Optional if "cell_id" is not null.
  - Description: Latitude value of the location of the event.
- **longitude**:
  - Type: Float
  - Requirements: Longitude value in WGS84 system. Value has to be within WGS84 bounds. Optional if "cell_id" is not null.
  - Description: Longitude value of the location of the event.

Optional fields:

- **loc_error**:
  - Type: Integer
  - Requirements: Positive value
  - Description: Location error in meters.

Example:

year	month	day	user_id	timestamp	mcc	mnc	plmn	cell_id	lon	lat	loc_error
2023	01	01	000000000000..01	00:00:00	214	01	null	214030412038931	40.51873	-3.62958	100
2023	01	01	000000000000..01	00:01:15	214	01	null	214030412038931	40.51871	-3.62952	100
2023	01	01	000000000000..10	12:05:03	214	01	null	214035484123541	null	null	null

22.2.9 MNO EVENT DATA SYNTACTIC QUALITY METRICS [INTERMEDIATE RESULTS]

MNO EVENT DATA SYNTACTIC QUALITY METRICS [INTERMEDIATE RESULTS]	
Description	Quality metrics produced by Module/Method 7: MNO Event Data Cleaning – Syntactic Checks. It includes counts of records removed or labelled by variable and by type of error. This data object also includes table to show distribution of records before and after the application of the method by **user_id** and **cell_id**.
Mandatory/Optional	Mandatory
Object/Unit/Record	Quality metrics
Contents	Description of all the fields and values using bullet points: • **result_timestamp**: timestamp of the start of the process when the metrics were produced. One process can generate multiple metrics. • **date** : ○ Type: DateType ○ Requirements: The date that the data was about. ○ Description: The date for which the quality metrics were produced. • **variable: name of the field to which the metric refers to. It could be null if the error refer to more than one variable.** • **type of error** code indicating. ○ 1. missing value ○ 2. not right syntactic format ○ 3. out of admissible values ○ 4. inconsistency between variables. ○ 9. no errors ○ 10 different location duplicate ○ 11 same location duplicate ○ 0 the file is not readable ○ other types of errors can be added • **type of transformation** code indicating ○ 1. Converted timestamp ○ 2. Other conversion ○ 9. No transformation ○ other types of transformation can be added • **value**: count of records with the characteristics in the previous field. If the file is not readable the value of the readability error should be 0. Also frequency distributions by cell_id and by user and by day at the beginning and at the end of the module should be produced (NB these frequency distributions should not be provided to NSI, but can be useful for MNO to investigate in case of specific quality warnings)

Example:

type_of_error	error_type_description
1	Missing value
2	Not right syntactic format
3	Out of admissible values
4	Inconsistency between variables
5	No location (no cell_id and no latitude&longitude), for that type or error there is None for variable column
0	the file is not readable
9	No error

type_of_transformation	error_type_description
1	Converted timestamps
2	Other conversion
9	No transformation

result_timestamp	date	variable	type_of_error	type of transformation	value
		cell_id	1	-	1000
		cell_id	2	-	20
		cell_id	9	-	10000
		timestamp	-	1	1

cell_id	user_id	date	initial_frequency
214030412038931	01	2023-07-20	200
214030412038931	02	2023-07-21	600

cell_id	user_id	date	final frequency
214030412038931	01	2023-07-20	10
214030412038931	02	2023-07-20	600

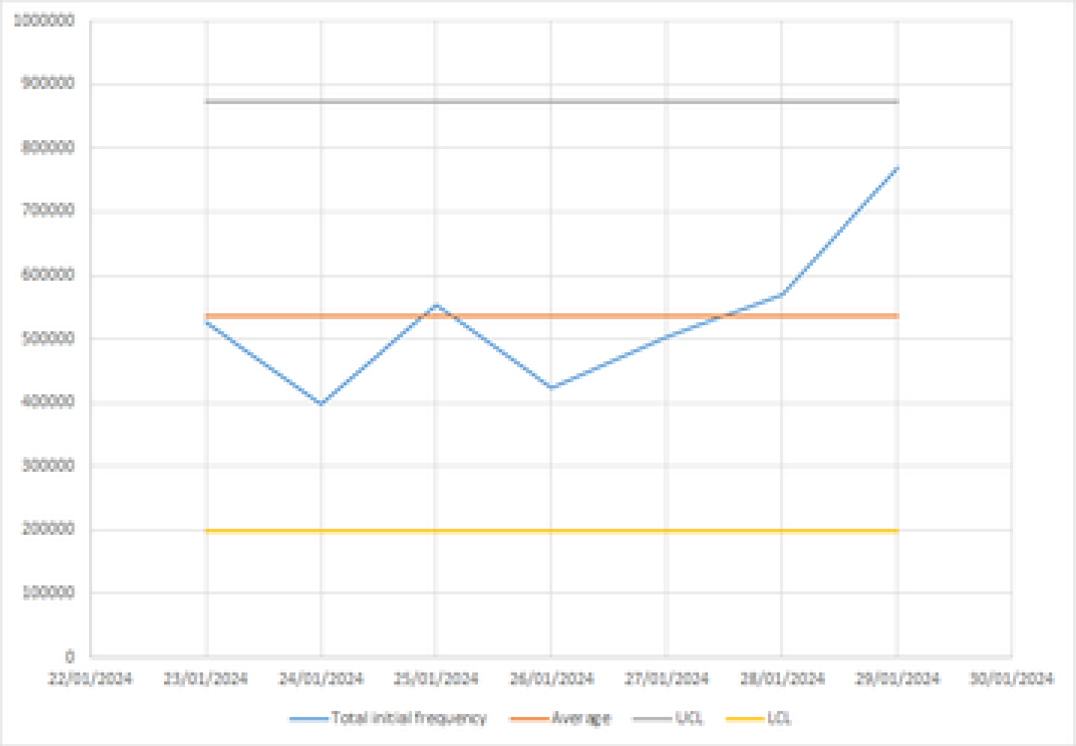
22.2.10 MNO EVENT DATA QUALITY WARNINGS [INTERMEDIATE RESULTS]

MNO EVENT DATA QUALITY WARNINGS [INTERMEDIATE RESULTS]	
Description	Summary report produced by Module/Method 8: MNO Event Data – Syntactic Quality Warnings presenting plots of the main metrics and quality warnings, in case the quality metrics are outside the Event Data Quality Warnings thresholds. Cases that generate warnings should be also stored in a log table. For the implementation of the pipeline in the test environment can be a summary report, in further developments it can become a dynamic dashboard. The analysis of the output can give insights on relevant errors in the input event data that can affect the quality of statistical output and can also provide hints on the errors causes that could be removed in data for next periods.
Mandatory/Optional	Mandatory
Object/Unit/Record	Quality warnings
Contents	The report will have the title "MNO Event Data Quality Warnings" + the date to which it is referred It will always present 3 plots: • plot 1. Size of the MNO Event Data – Raw [INPUT] Days on horizontal axis and the value of 'Total initial frequency' from the MNO Event Data Quality Metrics [INTERMEDIATE RESULTS] on the vertical axis. The period on the horizontal axis is set as a parameter of the method. It would be useful to have in the same plot in the vertical axis also the average of the metrics and the control limits • plot 2. Size of the Clean MNO Event Data [INTERMEDIATE RESULTS] Days on horizontal axis and the value of "Total final frequency" from the MNO Event Data Quality Metrics [INTERMEDIATE RESULTS] on the vertical axis. The period on the horizontal axis is set as a parameter of the method. It would be useful to have in the same plot in the vertical axis also the average of the metric and the control limits • plot 3. Error rate Days on horizontal axis and the value of error rate calculated as specified in Module/Method 8: MNO Event Data – Syntactic Quality Warnings on the vertical axis. The period on the horizontal axis is set as a parameter of the method. It would be useful to have in the same plot in the vertical axis also the average of the metric and the upper control limit Then the report will include warnings only in the case the daily value of the metrics is out of the thresholds indicated in the Module/Method 8: MNO Event Data – Syntactic Quality Warnings. In such cases the following information should be reported: **ATTENTION WARNING** <Measure definition> = Daily Value - <Condition for the warning>= value of the limit **<Warning message>** e.g. Error rate = 25% - Error rate is over the upper control limit calculated on the basis of average and standard deviation of the distribution of the error rate in previous **week**. Upper Control limit = **20%** **The error rate after syntactic checks application is unexpectedly high with respect to previous period, taking into account its usual variability** In the case the warning includes the user_id only the count of the cases that were out of the thresholds should be reported in the warning. e.g. **The error rate by user_id after syntactic checks application is unexpectedly high compared to the average of the users in X cases.** Finally the Composition of errors for each field should be reported with a pie chart. The warnings should also be stored in a log table with detailed information, the title and the date. The same log table can be used for all quality warnings. It can be agreed a certain time after which the log table can be cleaned. The log table is supposed to be used by MNO for further investigation in case of need)

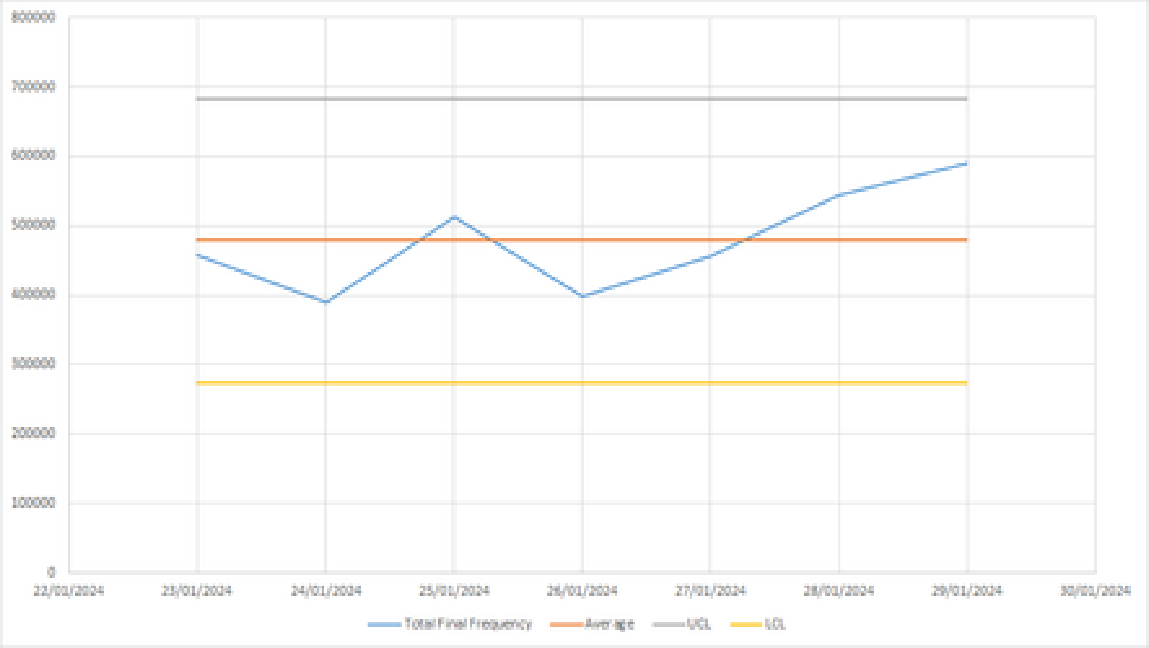
Example:

MNO Event Data Quality Warnings - 29/01/2024

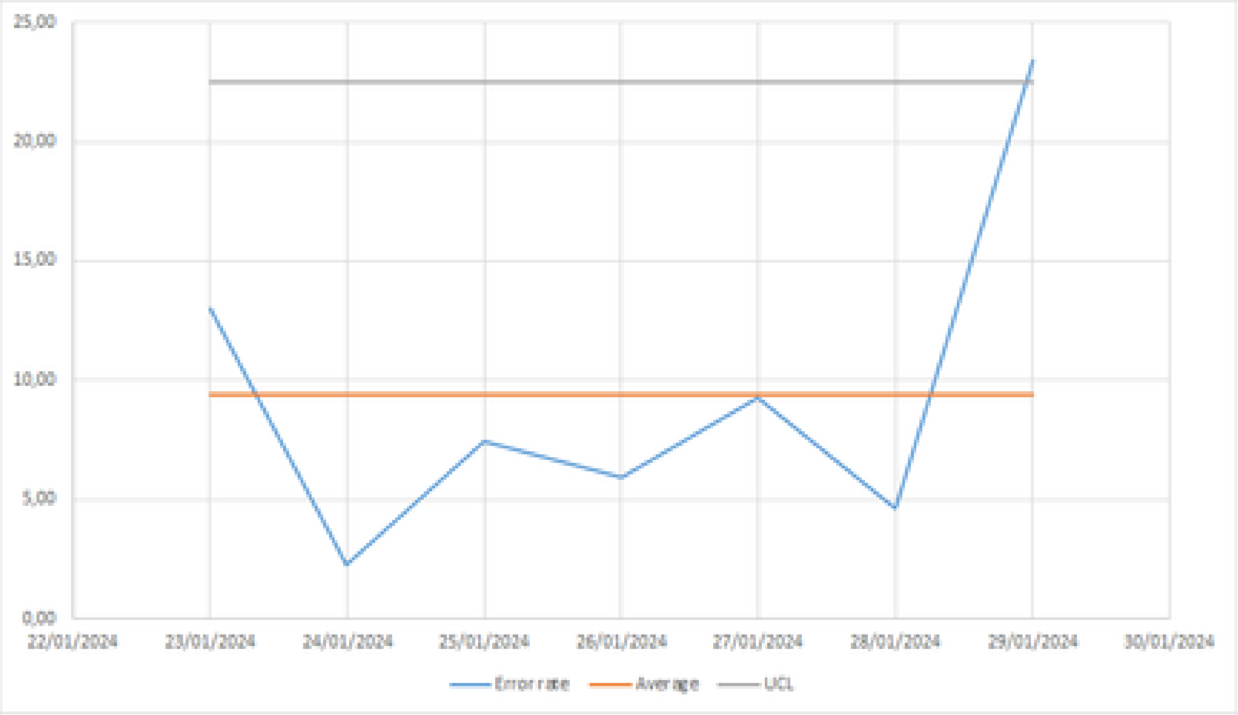
1. Size of the MNO Event Data – Raw [INPUT]



2. Size of the Clean MNO Event Data [INTERMEDIATE RESULTS]



3. Error rate



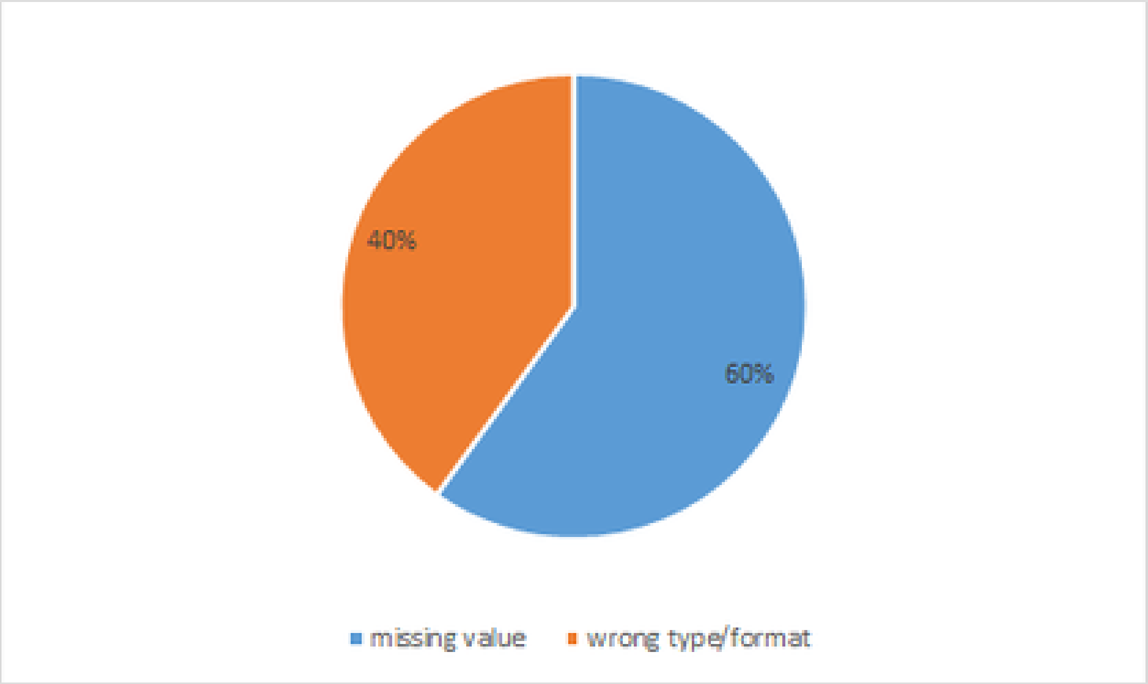
ATTENTION WARNING

Error rate =23.41% - Error rate is over the upper control limit calculated on the basis of average and standard deviation of the distribution of the error rate in previous week. Upper Control limit = **22.48%**

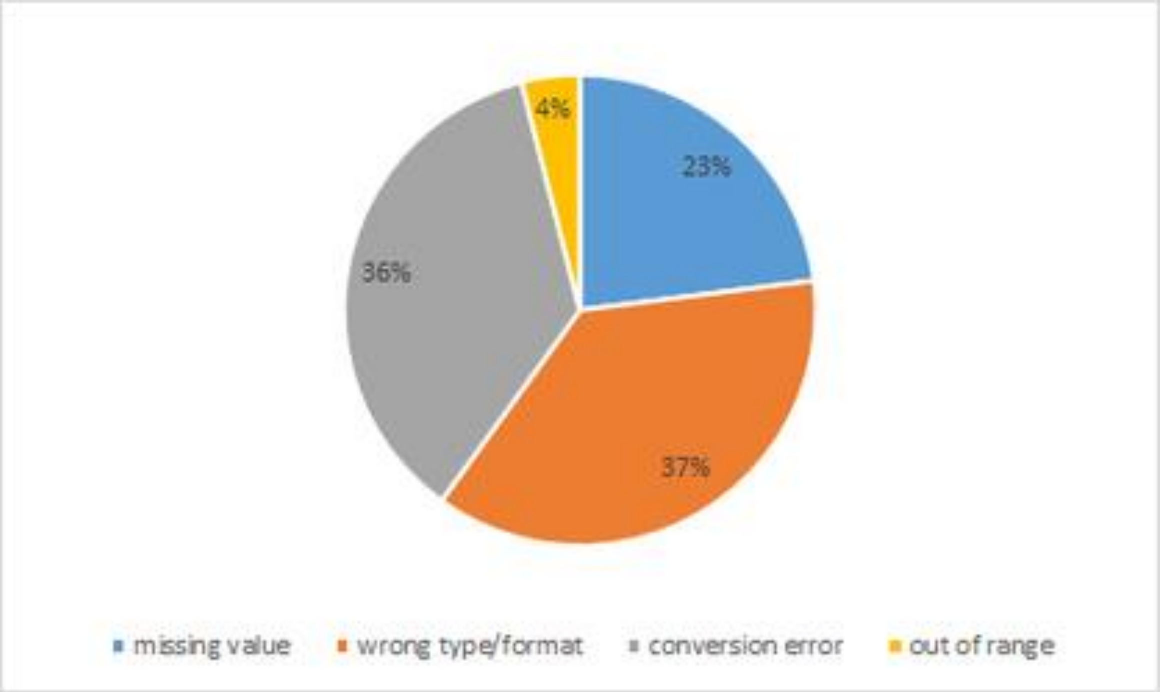
The error rate after syntactic checks application is unexpectedly high with respect to previous period, taking into account its usual variability

Composition of errors

1. user_id



2. timestamp



[...]

Log table example:

Title	Date	Measure_definition	Daily_value	Condition	Parameter_time	Value_condition	warning_text
MNO Event Data Quality Warnings	2024-01-29 00:00:00	Error rate	23.41	Error rate is over the upper control limit calculated on the basis of average and standard deviation of the distribution of the error rate in previous	week	**22.48**	**The error rate after syntactic checks application is unexpectedly high with respect to previous period, taking into account its usual variability**

22.2.11 EVENT DATA AT DEVICE LEVEL [INTERMEDIATE RESULTS]

EVENT DATA AT DEVICE LEVEL [INTERMEDIATE RESULTS]	
Description	The structure of these data is the same as Clean MNO Event Data [INTERMEDIATE RESULTS], but provides a separate output for each device (the user_id field is fixed)
Mandatory/Optional	Mandatory
Object/Unit/Record	Mobile network event associated to a specific subscriber
Contents	**Mandatory fields**: • **year:** o Type: Integer 16 o Requirements: Integer of 16 bits. o Description: Year the event took place. • **month:** o Type: Integer 8 o Requirements: Integer of 8 bits. o Description: Month the event took place. • **day:** o Type: Integer 8 o Requirements: Integer of 8 bits. o Description: Day the event took place. • **user_id:** o Type: Binary o Requirements: 32 bytes (256 bits) field. o Description: Unique pseudonymized identifier of the device. • **timestamp:** o Type: Time o Requirements: Parquet time type in hour, minutes and seconds. o Description: Point in time where the event took place. • **mcc:** o Type: Integer o Requirement: 3 digits code o Description: Mobile Country Code derived from the user's IMSI. • **mnc:** o Type: String o Requirement: 2 or 3 digits code o Description: Mobile Network Code, a code of a home operator. It might help to assess the selectivity bias that is in place due to preferential roaming agreements between MNOs. This must be string, as it can start with 0 digit. Possible options can also be 01 or 001, so it cannot be integer. • **plmn:** o Type: Integer o Requirement: 5 or 6 digits code. Mandatory only for outbound data o Description: Network identifier of the foreign roaming partner MNO, consists of PLMN=MCC+MNC. • **cell_id:** o Type: String o Requirements: 14 or 15 character length string. All characters must be numbers. Optional if "latitude" and "longitude" are not null. o Description: Identifier of the cell following CGI and eCGI standards. • **latitude:** o Type: Float o Requirements: Latitude value in WGS84 system. Value has to be within WGS84 bounds. Optional if "cell_id" is not null. o Description: Latitude value of the location of the event.

EVENT DATA AT DEVICE LEVEL [INTERMEDIATE RESULTS]

- **longitude**:
  - Type: Float
  - Requirements: Longitude value in WGS84 system. Value has to be within WGS84 bounds. Optional if “cell_id” is not null.
  - Description: Longitude value of the location of the event.

Optional fields:

- **loc_error**:
  - Type: Integer
  - Requirements: Positive value
  - Description: Location error in meters.

Example:

year	month	day	user_id	timestamp	mcc	mnc	plmn	cell_id	lon
2023	01	01	000000000000..05	00:00:00	214	67	null	214030412038931	40.51873
2023	01	01	000000000000..05	00:01:15	214	299	null	214030412038931	40.51871
2023	01	01	000000000000..05	12:05:03	214	299	null	214035484123541	null

[…]

year	month	day	user_id	timestamp	mcc	mnc	plmn	cell_id	lon	lat	loc_error
2023	01	01	000000000000..11	00:00:00	214	67	null	214030412038931	40.51873	-3.62958	100
2023	01	01	000000000000..11	00:01:15	214	299	null	214030412038931	40.51871	-3.62952	100
2023	01	01	000000000000..11	12:05:03	214	299	null	214035484123541	null	null	null

22.2.12 SEMANTICALLY CLEANED EVENT DATA AT DEVICE LEVEL [INTERMEDIATE RESULTS]

SEMANTICALLY CLEANED EVENT DATA AT DEVICE LEVEL [INTERMEDIATE RESULTS]	
Description	The structure of these data is the same as Event Data at Device level [INTERMEDIATE RESULTS]
Mandatory/Optional	Mandatory
Object/Unit/Record	Mobile network event data associated to a specific subscriber, after Module/Method 10: Event Cleaning at Device Level – Semantic Checks has been completed.
Contents	**Mandatory fields**: • **year:** o Type: Integer 16 o Requirements: Integer of 16 bits. o Description: Year the event took place. • **month:** o Type: Integer 8 o Requirements: Integer of 8 bits. o Description: Month the event took place. • **day:** o Type: Integer 8 o Requirements: Integer of 8 bits. o Description: Day the event took place. • **user_id:** o Type: Binary o Requirements: 32 bytes (256 bits) field. o Description: Unique pseudonymised identifier of the device. • **timestamp:** o Type: Time o Requirements: Parquet time type in hour, minutes and seconds. o Description: Point in time where the event took place. • **mcc:** o Type: Integer o Requirement: 3 digits code o Description: Mobile Country Code derived from the user's IMSI. • **mnc:** o Type: String o Requirement: 2 or 3 digits code o Description: Mobile Network Code, a code of a home operator. It might help to assess the selectivity bias that is in place due to preferential roaming agreements between MNOs. This must be string, as it can start with 0 digit. Possible options can also be 01 or 001, so it cannot be integer. • **plmn:** o Type: Integer o Requirement: 5 or 6 digits code. Mandatory only for outbound data o Description: Network identifier of the foreign roaming partner MNO, consists of PLMN=MCC+MNC. • **cell_id:** o Type: String o Requirements: 14 or 15 character length string. All characters must be numbers. Optional if “latitude” and “longitude” are not null. o Description: Identifier of the cell following CGI and eCGI standards. • **latitude:** o Type: Float o Requirements: Latitude value in WGS84 system. Value has to be within WGS84 bounds. Optional if “cell_id” is not null. o Description: Latitude value of the location of the event.

SEMANTICALLY CLEANED EVENT DATA AT DEVICE LEVEL [INTERMEDIATE RESULTS]

- **longitude**:
  - Type: Float
  - Requirements: Longitude value in WGS84 system. Value has to be within WGS84 bounds. Optional if "cell_id" is not null.
  - Description: Longitude value of the location of the event.
- **error_flag**:
  - Type: Integer, referring to global error type code
  - Requirements: Must be one of the global error type codes
  - Description: Error flag referring to an error type code of the specific identified error

Optional fields:

- **loc_error**:
  - Type: Integer
  - Requirements: Positive value
  - Description: Location error in meters.

Example:

year	month	day	user_id	timestamp	mcc	mnc	plmn	cell_id	lon	lat	loc_error	error_flag
2023	01	01	000000000000..11	00:00:00	214	67	null	214030412038931	40.51873	-3.62958	100	0
2023	01	01	000000000000..11	00:01:15	214	299	null	214030412038931	40.51871	-3.62952	100	1
2023	01	01	000000000000..11	12:05:03	214	299	null	214035484123541	null	null	null	3
2023	01	01	000000000000..11	12:06:15	214	299	null	214035484123549	null	null	null	2
2023	01	01	000000000000..11	12:08:03	214	299	null	214035484123538	null	null	null	4

22.2.13 DEVICE SEMANTIC QUALITY METRICS [INTERMEDIATE RESULTS]

DEVICE SEMANTIC QUALITY METRICS [INTERMEDIATE RESULTS]	
Description	Quality metrics produced by Module/Method 10: Event Cleaning at Device Level – Semantic Checks It includes counts semantical errors in the MNO event data.
Mandatory/Optional	Mandatory
Object/Unit/Record	Quality metrics
Contents	Description of all the fields and values using bullet points: • **result_timestamp**: timestamp (UTC) of the start of the process when the metrics were produced. One process can generate multiple metrics. • **data_period_start**: Start of the data period (UTC) under study, any records before this will be removed. Example "2023-01-01 00:00:00". • **data_period_end**: End of data period (UTC) under study, any records after this will be removed. Example "2023-01-01 23:59:59" • **variable:** name of the field to which the metric refers to. It could be null if the error refer to more than a variable. • **type of error (NB! MUST BE UNIFORMED TO GLOBAL ERROR TYPES AND ERROR CODES)**: ○ 0 - no error; ○ 1 - cell_id does not exist (no corresponding event cell_id in network topology); ○ 2 - cell_id not valid (cell_id exists in network topology but with different date than of event date); ○ 3 - event location incorrect (single event); ○ 4 - event location probably incorrect (suspicion); • **value**: count of erroneous events

Example:

result_timestamp	data_period_start	data_period_end	variable	type_of_error	value
2023-07-21 07:15:57	2023-07-20 00:00:00	2023-07-20 23:59:59	cell_id	1	21
2023-07-21 07:15:57	2023-07-20 00:00:00	2023-07-20 23:59:59	cell_id	2	53246
2023-07-21 07:15:57	2023-07-20 00:00:00	2023-07-20 23:59:59	cell_id	3	23513
2023-07-21 07:15:57	2023-07-20 00:00:00	2023-07-20 23:59:59	cell_id	4	1153
2023-07-21 07:15:57	2023-07-20 00:00:00	2023-07-20 23:59:59	cell_id	0	139851010
2023-07-22 07:16:12	2023-07-21 00:00:00	2023-07-21 23:59:59	cell_id	1	34599296
2023-07-22 07:16:12	2023-07-21 00:00:00	2023-07-21 23:59:59	cell_id	2	6532564
2023-07-22 07:16:12	2023-07-21 00:00:00	2023-07-21 23:59:59	cell_id	3	64724
2023-07-22 07:16:12	2023-07-21 00:00:00	2023-07-21 23:59:59	cell_id	4	62261
2023-07-22 07:16:12	2023-07-21 00:00:00	2023-07-21 23:59:59	cell_id	0	98015873

22.2.14 DEVICE ACTIVITY STATISTICS [INTERMEDIATE RESULTS]

DEVICE ACTIVITY STATISTICS [INTERMEDIATE RESULTS]	
Description	Metrics produced by Module/Method 11: Device Activity Statistics
Mandatory / Optional	Mandatory
Object / Unit / Record	Metrics / statistics
Contents	Description of the metrics computed for the specific device along with their values and the choice of period parameter (if needed). ● **user_id:** ○ Type: Binary ○ Requirements: 32 bytes (256 bits) field. ○ Description: Unique pseudonymized identifier of the device. ● **year:** ○ Type: Integer 16 ○ Requirements: Integer of 16 bits. ○ Description: Year the metrics refer to in the local timezone. ● **month:** ○ Type: Integer 8 ○ Requirements: Integer of 8 bits. ○ Description: Month the metrics refer to in the local timezone. ● **day:** ○ Type: Integer 8 ○ Requirements: Integer of 8 bits. ○ Description: Day the metrics refer to in the local timezone. ● **event_cnt:** ○ Type: Integer 32 ○ Requirements: Integer of 32 bits. ○ Description: Number of events per day. ● **unique_cell_cnt:** ○ Type: Integer 16 ○ Requirements: Integer of 16 bits. ○ Description: Number of unique cells per day. ● **unique_location_cnt:** ○ Type: Integer 16 ○ Requirements: Integer of 16 bits. ○ Description: Number of different locations per day (based on the location point of the cell). ● **sum_distance_m:** ○ Type: Integer 32 ○ Requirements: Integer of 32 bits. ○ Description: Sum of the distances between the events (based on the location point of the cell). ● **unique_hour_cnt:** ○ Type: Integer 8 ○ Requirements: Integer of 8 bits. Up to 24. ○ Description: Number of unique hours in the date with events. ● **mean_time_gap:** ○ Type: Integer 32 ○ Requirements: Integer of 32 bits. ○ Description: Average time gap between events (in seconds). ● **stdev_time_gap:** ○ Type: Float ○ Requirements: Float ○ Description: Standard deviation of the time gap between events (in seconds).

DEVICE ACTIVITY STATISTICS [INTERMEDIATE RESULTS]	
	Notes • All the indicators are calculated per device per day. When longer period assessment of the device activity is needed (e.g., for specific use case), then this must be done by combining the metrics for different dates that are inside the necessary period. For simplicity and optimisation reasons, this longer-period aggregates are not stored in this data object. This is also necessary due to the requirement of periodical deletion of historical device-level data that can be successfully done using the "date" here, but could not be done very well with longer periods.

Example:

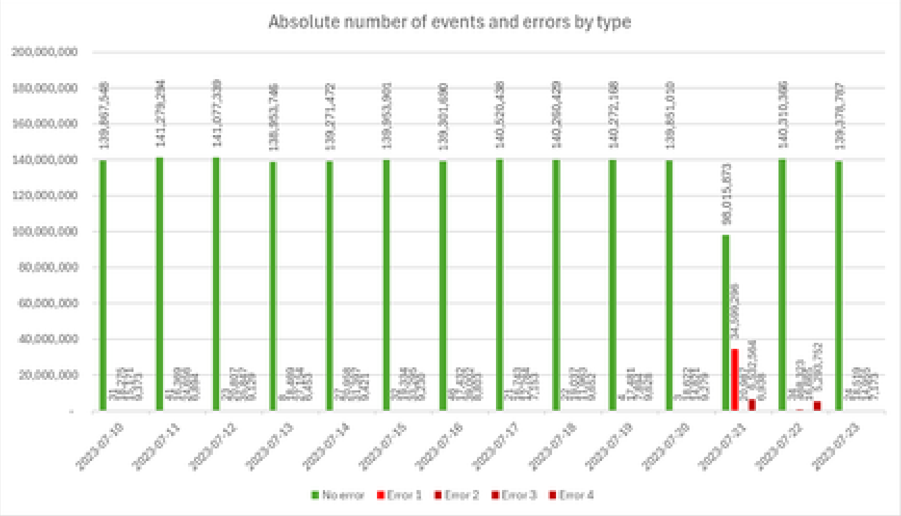
device_id	year	month	day	event_cnt	unique_cell_cnt	unique_location_cnt	sum_distance_m	unique_hour_cnt	mean_time_gap	stdev_time_gap
A	2023	1	1	12	10	10	45778	10	5090	2951.61
A	2023	1	2	8	2	2	7592	7	5118	3169.484
B	2023	1	1	12	10	10	45036	8	4358	3614.575
C	2023	1	1	11	1	1	0	10	5939	4039.195
C	2023	1	2	20	1	1	0	14	4173	3017.242
C	2023	1	3	12	1	1	0	10	7313	3111.024
C	2023	1	4	7	1	1	0	5	4062	1536.541
D	2023	1	1	112	80	80	1035035	9	276	163.491
E	2023	1	1	142	37	37	13083	2	28	17.225
F	2023	1	1	41	1	1	0	1	33	13.647
G	2023	1	1	24	13	13	51061	24	3600	0
H	2023	1	1	1	1	1	0	1	0	0

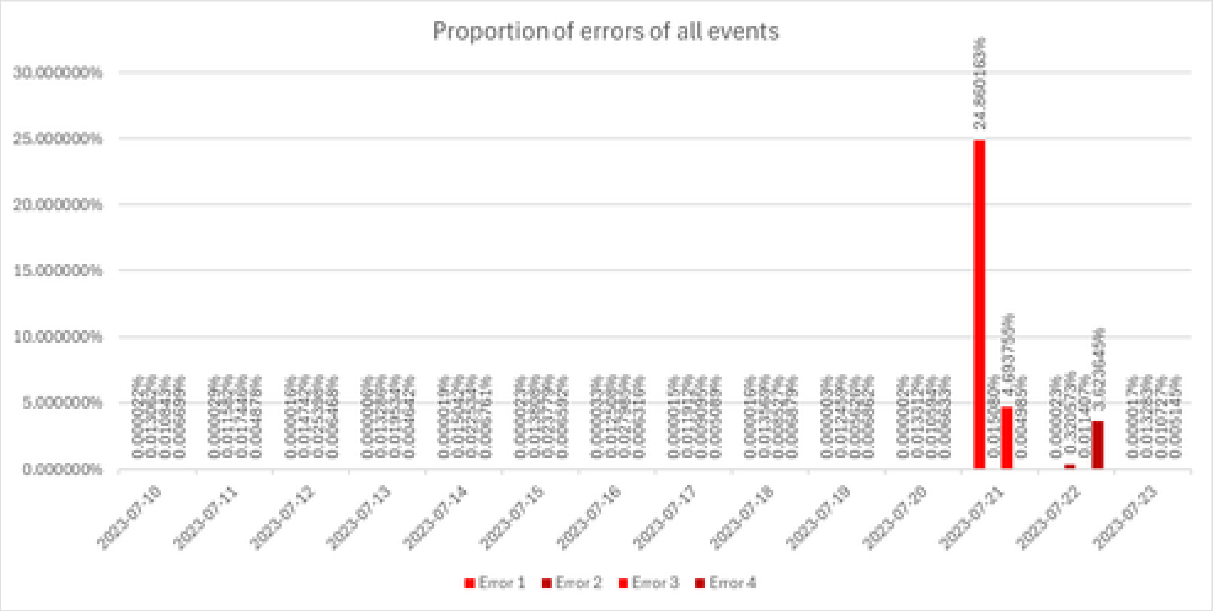
22.2.15 MNO EVENT DATA AT DEVICE LEVEL SEMANTIC QUALITY WARNINGS [INTERMEDIATE RESULTS]

MNO EVENT DATA AT DEVICE LEVEL SEMANTIC QUALITY WARNINGS [INTERMEDIATE RESULTS]	
Description	Summary report produced by Module/Method 12: MNO Event Data at Device Level – Semantic Quality Warnings presenting plots of the main metrics and quality warnings in the case the quality metrics are out of the warning thresholds. Cases that generate warnings should be also stored in a log table. For the implementation of the pipeline in the test environment can be a summary report, in further developments it can become a dynamic dashboard. The analysis of the output can give insights on relevant errors in the input event data that can affect the quality of statistical output and can also provide hints on the errors causes that could be removed in data for next periods.
Mandatory/Optional	Mandatory
Object/Unit/Record	Quality warnings
Contents	The report will have the title "MNO Event Data at Device Level Semantic Quality Warnings" + the date to which it is referred It will always present 2 plots: 1. "Absolute number of events and errors by type"; 2. "Proportion of errors of all events". Then the report will include warnings only in the case the daily value of the metrics is out of the thresholds indicated in the Module/Method 12: MNO Event Data at Device Level – Semantic Quality Warnings. In such cases the following information should be reported: **ATTENTION WARNING** YYYY-MM-DD: Error rate for error type X =xx.xx% - Error rate is over the set error threshold. The warnings should also be stored in a log table with the same information, the title and the date. The same log table can be used for all quality warnings. It can be agreed a certain time after which the log table can be cleaned. Data object will also include the table of cell_id's with appropriate date, error type and a count of events for this error and this cell. This is based on Semantically Cleaned Event Data at Device Level [INTERMEDIATE RESULTS] events table where error_flag is to be used as error type.

Example:

The chart displays the absolute number of events and errors by type.





ATTENTION WARNING

2023-07-21: Error rate for error type 1 =24.86% - Error rate is over the set error threshold.

2023-07-21: Error rate for error type 3 =4.69% - Error rate is over the set error threshold.

2023-07-22: Error rate for error type 2 =3.12% - Error rate is over the set error threshold.

2023-07-22: Error rate for error type 4 =3.52% - Error rate is over the set error threshold.

cell_id	date	error_type	count
21403468392467	2023-07-21	1	542
21403682372813	2023-07-21	2	323
21403293746273	2023-07-22	3	51142
21403235523753	2023-07-22	4	11

22.2.16 DAILY CONTINUOUS TIME SEGMENTATION [INTERMEDIATE RESULTS]

	DAILY CONTINUOUS TIME SEGMENTATION [INTERMEDIATE RESULTS]
Description	Daily time segments of a specific user covering the 24 hours of a specific date under study. The individual MND events are grouped into time segments. Four categories are supported: (i) **stay** (set of events that are close in time and space during a minimum dwell time), (ii) **unknown** (gap of information), (iii) **undetermined** (punctual events that are not possible to classify either as stay or move) and (iv) **move** (rest of the day that it is not classified as any other group).
Object/Unit/Record	Time segment
Contents	• **'time_segment_id'**: unique identifier of the time segment associated to a specific user • **'time_segment_type'**: type of time segment (stay, move, undetermine or unknown) • **'device_id'**: unique pseudonymised identifier of the device. • **'initial_timestamp':** timestamp ('YYYY-MM-DD hh:mm:ss') in UTC standard indicating the date and time of the first event of the time segment • **'final_timestamp':** timestamp ('YYYY-MM-DD hh:mm:ss') in UTC standard indicating the date and time of the last event of the time segment. 'Null' in case of undetermine events. • **'cells_id':** set of cells identifiers associated to the time segment. For 'stay' it represents a set of unique cell identifiers regardless of the numer of events associated to each cell. For 'move' it represents the complete list of events ordered by time. 'Null' for 'unknown' time segments. • **'move_timestamps':** timestamp associated to the move events. 'Null' for time segments different from 'move'

Example:

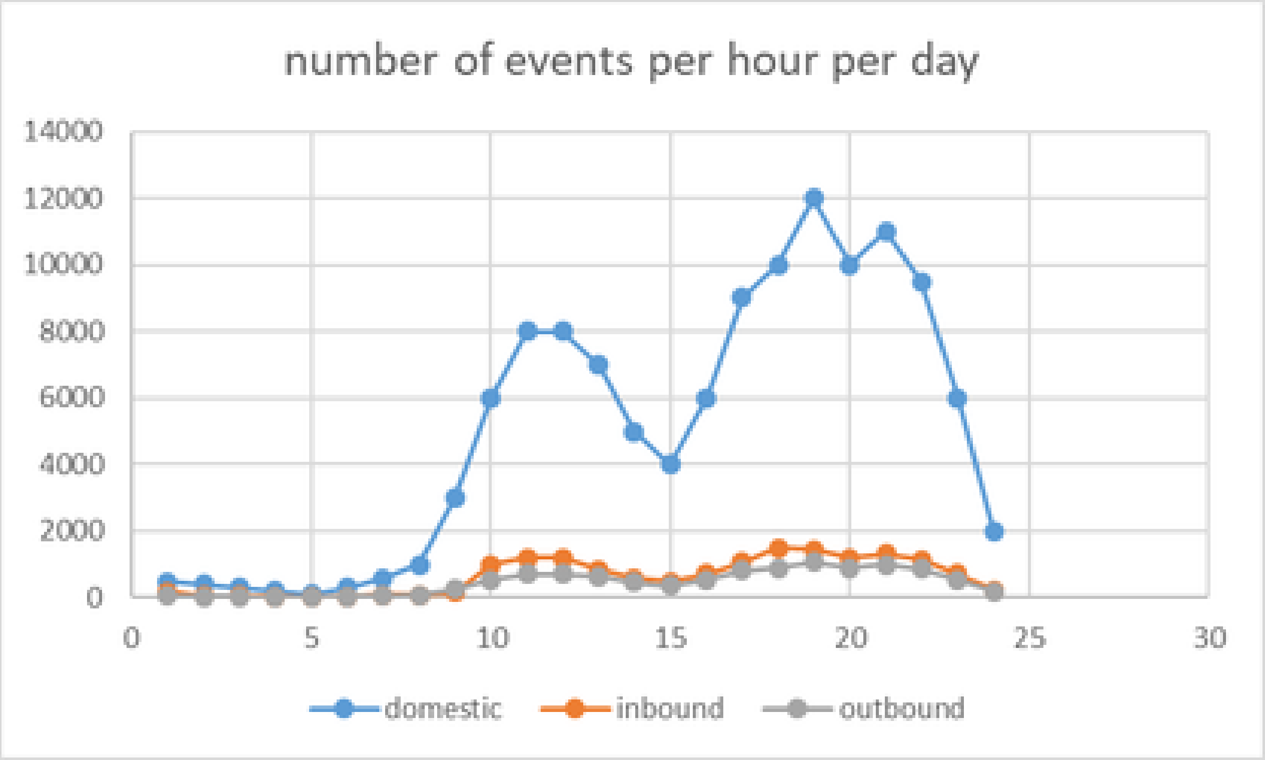
time_segment_id	time_segment_type	device_id	initial_timestamp	final_timestamp	cells_id	move_timestamps
1	stay	1	2023-01-01 00:00:00	2023-01-01 06:45:01	214030412038931, 214030412038932, 214030412038935, 214030412038938	null
2	move	1	2023-01-01 06:45:01	2023-01-01 07:16:21	214030412038940, 214035484123541, 214035484123544	2023-01-01 07:01:10, 2023-01-01 07:12:03,2023-01-01 07:16:21
3	unknown	1	2023-01-01 07:16:21	2023-01-01 22:16:15	null	null
4	stay	1	2023-01-01 22:16:15	2023-01-01 23:59:59	214030412038931, 214030412038932	null
1	stay	2	2023-01-01 00:00:00	2023-01-01 11:49:35	214030412038964, 214030412038965	null
…	…	…	…	…	…	…

22.2.17 GENERAL EVENT STATISTICS METRICS [INTERMEDIATE RESULTS]

GENERAL EVENT STATISTICS METRICS [INTERMEDIATE RESULTS]	
Description	Quality metrics produced by MNO Event– General Statistics It stores daily general statistics on MNO event data that can be used in the next modules of the pipeline.
Mandatory/Optional	Mandatory
Object/Unit/Record	Metrics / statistics
Contents	• **year:** o Type: Integer 16 o Requirements: Integer of 16 bits. o Description: Year the metrics refer to in the local timezone. • **month:** o Type: Integer 8 o Requirements: Integer of 8 bits. o Description: Month the metrics refer to in the local timezone. • **day:** o Type: Integer 8 o Requirements: Integer of 8 bits. o Description: Day the metrics refer to in the local timezone. • **size_un_domestic_devices:** Number of unique domestic devices per day (where MCCMNC = home_MNO); • **size_un_inbound_devices:** Number of unique inbound roaming devices per day (where MCCMNC != home_MNO); • **size_un_outbound_devices:** Number of unique outbound roaming devices per day (where PLMN != home_MNO); • **size_domestic_events:** Number of domestic events per day (where MCCMNC = home MNO); • **size_inbound_events:** Number of inbound roaming events per day (where MCCMNC != home MNO); • **size_outbound_events:** Number of outbound roaming events per day (where PLMN != home MNO); • **n_un_MCC_inbound:** Number of unique MCC combinations for inbound roaming data per day; • **n_un_MCCMNC_inbound:** Number of unique MCCMNC combinations for inbound roaming data per day; • **n_un_PLMN_outbound:** Number of unique PLMN combinations for outbound roaming data per day; • **ave_domestic_events_perdevice:** Average number of domestic events per device per day; • **ave_inbound_events_perdevice:** Average number of inbound roaming events per device per day; • **ave_outbound_events_perdevice:** Average number of inbound roaming events per device per day; • **ave_un_cell_id_domestic_perdevice:** Average number of unique cell_id's in domestic events per device per day; • **med_un_cell_id_domestic_perdevice:** Median number of unique cell_id's in domestic events per device per day; • **ave_un_cell_id_inbound_perdevice:** Average number of unique cell_id's in inbound roaming events per device per day; • **med_un_cell_id_inbound_perdevice:** Median number of unique cell_id's in inbound roaming events per device per day; • **ave_un_PLMN_outbound_perdevice:** Average number of unique PLMN in outbound roaming events per device per day;

GENERAL EVENT STATISTICS METRICS [INTERMEDIATE RESULTS]	
	- **med_un_PLMN_outbound_perdevice:** Median number of unique PLMN in outbound roaming events per device per day; - **ave_days_device**: Average number of days of device ID present in days throughout the data lookback period; - **lookback_period**: parameter considered in previous statistics **Plot**: Number of domestic events per day per hour; Number of inbound roaming events per day per hour; Number of outbound roaming events per day per hour; In addition, the following **tables** should be produced: - Number of unique inbound roaming devices per MCC and MNC per day (where MCCMNC != home_MNO) - Number of unique outbound roaming devices per PLMN per day (where PLMN != home MNO); - Number of inbound roaming events per MCC and MNC per day (where MCCMNC != home MNO); - Number of outbound roaming events per PLMN per day (where PLMN != home MNO);

Example:



year	month	day	MNC	MCC	Number of unique inbound roaming devices	Number of inbound roaming events
2024	07	01	10	222	2500	148000
2024	07	01	01	222	3000	200000
2024	07	01	65	472	62	58001

year	month	day	PLMN	Number of unique outbound devices	Number of outbound roaming events
2024	07	01	10	2500	148000
2024	07	01	01	3000	200000
2024	07	01	65	62	58001

ANNEX I – SDC FOR DEMONSTRATOR SCENARIO/DISCLOSURE

OBJECTIVE

The objective of this method is to apply a disclosure limitation method in the testing phase, before MNOs share the aggregates to NSIs.

Disclosure Risks assessment and SDL application is given for the application to UCs where only one dimension-geography is present. For example, we refer to the UCs Present Population and Usual Environment.

PARAMETERS

- A parameter k for the k-anonymity assessment: set the default value equal to 5 (to be agreed with MNOs);
- Percentage (see below) to choose between two different anonymisation methods: suppression (method A) or aggregation of areas (geographical recoding) (method B), default value =5%;
- Maximum level of aggregation (needed for method B) (optional): MAX.

INPUT DATA

- (Weighted) Counts at geographical level specified by the UC;
- A nested space subdivision: territory is subdivided in different nested levels; (for method B), e.g. contiguous grid cells into bigger polygons;
- Desired output geographical levels: for example, INSPIRE 1000x1000 grid for the present population UC, and/or administrative divisions as for the Home location UC.

OUTPUT DATA

- (Weighted) Counts at the input geographical level, with zero values for those geographical areas with confidential data (when method A is applied) and/or possibly aggregated counts at a coarser geographical level for those areas with disclosure risk (when Method B is applied).

METHODOLOGY

In the testing scenario, the method processes the output MNO counts before the data are shared with the NSI

Let 1...A be the areas at the UC desired hierarchical geographical level, L, e.g. the tiles of the INSPIRE 1000x1000 grid for the Present population UC.

Consider also the higher geographical level L+1 and so on; for example in the Present population UC: the 2000 x2000 INSPIRE grid and so on (4000 x4000, 8000 x8000, 16000 x16000, only nested hierarchy is considered in the experiments).

Step 1 *Check if the input counts comply with k-anonymity, that is the count for area a is greater than k or equal to 0 for all a=1... A,*

That is zero counts are not considered confidential.

Step 2 *Evaluate the percentage of the amount of counts for the areas with sensitive values over the total count of the coarser area, e.g. if a (1000 x1000) cell contain a sensitive value consider the ratio of the counts for all the sensitive cells over the total count for the higher level in the hierarchy, 2000 x2000 cell.*

Step 3 *Compare the percentage to P (default 5%)*

-if less the percentage is less than P then

Step 3.1*. Replace the observed values less than the threshold k with zero*

Step 3.1.1 *Record the total counts for cells replaced with zero and produce indicator of the amount of suppression over the total count.*

Otherwise

Step 3.2 *aggregate all the areas counts belonging to the same coarser area at the coarser geographical level L+1; e.g aggregate the 1000x1000 tiles belonging to the same 2000x2000 tile at the coarser level, this will be the new output unit.*

Step 4 *Check for the new areas k-anonymity compliance. Stop if it is obtained or repeat step 2 on the new coarser areas until k-anonymity is satisfied (or until maximum level of aggregation is reached, in this case go to method A.)*

Note that the aggregation process depends on the a priori chosen grid cells at the different levels

EXAMPLE

Example 1

From Module 18: Projection of Multi-MNO relevant for the Use Case, the counts at desired geography level are produced:

Coarser geographical level	Geographical level, L Area	**Aggregates of target population**	**K-anonymity compliance (yes/no)**
	1	66	yes
	2	50	yes
	3	18	yes

Coarser geographical level	Geographical level, L Area	**Aggregates of target population**	**K-anonymity compliance (yes/no)**
	a	15	yes
M	A-2	20	yes
M	A-1	24	yes
M	A	2	no

Area A does not comply k-anonymity. 2< k=5,

Sum the counts over the coarser geographical level 20+24+2, the ratio

2/(20+24+2) <P=0.05 then use method A in step3.1

The final output table is:

Geographical level, L	**Aggregates of target population**
1	66
2	50
3	18
a	15
A-2	5
A-1	10
A	0

Example 2

Coarser geographical level, L+1	Geographical level, L /INSPIRE 100x100 grid Area	**Aggregates of target population**	**K-anonimity compliance (yes/no)**
1	1	66	yes
1	2	50	yes
1	3	18	yes
	a	15	yes
M-1	A-3	6	yes
M	A-2	5	yes
M	A-1	10	yes
M	A	3	no

Area A-2 and A do not comply k-anonymity.

Sum the counts over the coarser geographical level 5+10+3, the ratio

3/(20+24+2) >P=0.05 then use method B in step 3.2.

That is, sum the counts over all the areas belonging to the same coarser level M and this new area is given in the output; the other counts and areas are not modified.

Note that the released data will have a mixed geographical disaggregation:

Area	**Aggregates of target population**	**K-anonymity compliance (yes/no)**
1 geographical level L	66	yes
2 geographical level L L	50	yes
3 geographical level L	18	yes

Area	**Aggregates of target population**	**K-anonymity compliance (yes/no)**
a geographical level L	15	yes
A-3 geographical level L	6	yes
M geographical level L+1	18	yes

No further step is needed.

ANNEX II – PRESENT POPULATION ESTIMATION: VARIANT 2

METHODOLOGY FOR VARIANT 2: USING POPULATION TOTALS

The required input for this module consists of:

- The daily summaries as described above.
- The market share of the MNO per municipality. In case the market share is only known for larger (aggregated) regions, or even only on country level, this value can be used for all municipalities.
- A weight for each foreign country of the inbound roaming devices. These weights can be determined by taking the market share of the foreign MNOs into account that use the MNO or interest as preferred one.
- The market share of the MNO per foreign country, defined as a fraction between 0 and 1.
- Population counts from administrative data per municipality.

We require the home location estimates, which can be achieved by applying the modules for daily-, mid-term and long-term as specified for the home location UC. The home location estimates are required in order to gross up the count of mobile device to the target population of individuals.[12] For the present population UC the home locations should be determined in the same zoning system for which administrative population data is available, which usually is also the zoning system of the output. In this example, we assume the home locations are determined on municipality level, but it could also be the regular 1 km x 1 km grid since recently many census-like outputs are available at this level of disaggregation.

We use the same notation as in Section 6.1 Usual Environment use case, unless mentioned otherwise.

Let t be the timestamp for which the present population is estimated. We assume that timestamps are encoded as numerical variables, so t1 > t2 means that t1 is later than t2.

In addition, when y.x is used rather than y[i].x, then the applied function is iterated over all i.

Let GAP be the maximum allowed time gap for which the device is included in the daily summary. In other words, the device is included in the output if there is at least one event in the time window [t – GAP, t + GAP].

Let m be the number of cells. Let z be the number of zones (municipalities). Let K be the set of unique devices in the event data.

[12] Suarez Castillo, Milena, Sémécurbe, Francois, Ziemlicki, Cezary, Tao, Haixuan Xavier and Seimandi, Tom. "Temporally Consistent Present Population from Mobile Network Signaling Data for Official Statistics" Journal of Official Statistics, vol.39, no.4, 2023, pp.535-570. https://doi.org/10.2478/jos-2023-0025

```
// create an empty array of number of cells (m) by the number of home
municipalities(z)

DECLARE counts[m][z]

FOR i = 1 TO m

  FOR j = 1 TO z

    counts[i][j] = 0

  END FOR

END FOR



FOR k in K

 home = longTermSummary.home(k)

 IF missing(home) NEXT

 IF k in table.device

   counts[table.cell[table.device == k]] += 1

 END IF

END FOR
```

The output counts is a two-dimensional array. When pivoting to a long format the output table will be of the following format, where the day and time columns have been added for completeness.

Table 4: Example of the daily summary table

cell ID	home municipality	count	day	time
1	A	5436	2023-07-14	2pm
1	C	3213	2023-07-14	2pm
2	A	342	2023-07-14	2pm
3	A	53434	2023-07-14	2pm
3	B	325	2023-07-14	2pm
3	C	23454	2023-07-14	2pm
4	A	34545	2023-07-14	2pm

For **inbound roaming devices** we can apply the same method, obtaining the following output table:

Table 5: Example of inbound roaming devices summary table

cell ID	country	count	day	time
1	X	564	2023-07-14	2pm
1	Y	234	2023-07-14	2pm
2	X	12	2023-07-14	2pm
3	X	6543	2023-07-14	2pm
3	Y	275	2023-07-14	2pm
4	Y	1645	2023-07-14	2pm

Note that for the calculation of the weights (in the next module) it may also be relevant not only from which country the roaming devices come from, but also from which MNOs.

For **outbound roaming devices** a similar table should be constructed. Instead of the cell ID column we have the country code. Therefore, this table should be in this format:

Table 6: Example of the daily summary table of outbound roaming devices

country	home municipality	count	day	time
X	A	4334	2023-07-14	2pm
X	B	136	2023-07-14	2pm
X	C	6458	2023-07-14	2pm
Y	A	5346	2023-07-14	2pm
Y	C	1654	2023-07-14	2pm

```
DECLARE weightedCounts[m]
weightedCounts = 0

// loop over home regions
FOR j in 1 TO z
  total = pop[j] * marketShare[j]

  //sum counts with j as home
  countSum = 0
  FOR i in 1 TO m
     countSum += counts[i][j]
  END FOR

  //add outbound roaming totals
  countSum += outRoamingCounts[j]

  // calculate weight for the MNO for region j
  weight = total / countSum
  // weight (gross up) the counts
  FOR i in 1 TO m
    weightedCounts[i] += counts[i][j] * weight
  END FOR
END FOR

// loop over home countries (inbound roaming)
// popCountry are the population totals per country
FOR j in 1 TO q
  // estimate total number of foreign people from country j
  total = popCountry[j] * marketShareCountry[j]

  //sum counts with j as home
  countSum = 0
  FOR i in 1 TO m
     countSum += outRoamingCounts[i][j]
  END FOR

  // weight for foreign people from country j
  weight = total / countSum

  // weight the counts
  // countsInRoaming taken from Table 6.2.2.
  FOR i in 1 TO m

    weightedCounts[i] +=  countsInRoaming[i][j] * weight

  END FOR

END FOR
```

The output, weightedCounts is a one-dimensional array with weighted counts per cell id:

Table 7: Example of the weighted counts table

cell ID	weighted count	day	time
1	25436	2023-07-14	2pm
2	5342	2023-07-14	2pm
3	304334	2023-07-14	2pm
4	145755	2023-07-14	2pm

The sum of weighted counts is the estimated total number of people that use the MNO used in this case study.

ESTIMATION

In this module, this total is spatially distributed over the country. As noted in Section 6.2.1 Concepts and requirements for the present population use case, a part of this distribution will be placed across the border, in case the MNO has coverage there and, by the same reasoning, it is expected that people in the country of interest are still registered on foreign MNOs that still have coverage. We will only take the former into account, because for the latter, we require the event and network topology data of the foreign MNOs.

This module requires:

- The weighted counts from the previous module.
- The cell connection probabilities from the box 'Spatial Cell Information'.

We apply the Maximum Likelihood Estimator (MLE) to estimate the spatial density over the INSPIRE 100x100m grid. This can be done by applying the Expectation-Maximization (EM) algorithm to compute the MLE as proposed by Shepp and Vardi (1982) and applied by Laan and Jonge (2019). However, we describe another calculation method, using an iterative Bayesian procedure.

Let c be the vector with weighted cell counts from the previous table, so c[i] denotes the count (either the device counts from ERROR! REFERENCE SOURCE NOT FOUND. or the weighted counts from ERROR! REFERENCE SOURCE NOT FOUND.) for cell i. We define N to be the population total, so

```
N = SUM(c)
```

Let dens be the spatial density distribution, which is a vector over all grid tiles (the total number of grid tiles is denoted by g). As initial step, we allocate N/g to each grid tile, and use this as uniform prior.

```
DECLARE dens[g]
```

```
FOR i in 1 TO g //grid tiles

  dens[i] = N/g

END FOR
```

Let connProb be the array that contains the cell connection probabilities. connProb[j][i] is the probability that a device is connected to cell i given that it is located in grid tile j.

The following iterative procedure converges to the MLE[13]. The prior dens is multiplied with connProb. The obtained posterior is used in the next iteration as prior. This process continues until convergence has been achieved. To guide this process, we will set to parameters: maxIter, the maximum number of iterations; and diffThreshold, the minimum difference threshold value between the prior and posterior distributions. For simplicity, we sum the absolute differences per grid tile, but more sophisticated methods can be used here.

```
iter = 0 // keep track of number of iterations

DO

  //temporary array of g x n (grid tiles times cells)

  DECLARE a[g][n]

  FOR j in 1 TO m //number of cells

    FOR i in 1 TO g

      a[j][i] = dens[j]

    END FOR

  END FOR



  // Bayesian iteration, taking a as prior

  a[j][i] = a[j][i] * connProb[j][i]



  // Create posterior: normalize array a

  FOR j in 1 TO m

    a[j][i] = a[j][i] / SUM(a[j]
```

[13] F. Ricciato and A. Coluccia (2023), On the Estimation of Spatial Density From Mobile Network Operator Data, IEEE Transactions on Mobile Computing, vol. 22, no. 6, pp. 3541-3557

```
  END FOR


  // Update spatial density estimation

  // Multiply the posterior by the cell counts c

  FOR i in 1 TO g

    newDens[i] = 0

    FOR j in 1 TO m

      newDens[i] += a[j][i] * c[j]

    END FOR

  END FOR


  // Calculate difference with dens (the old spatial density)

  df = 0

  FOR i in 1 TO g

    df += ABS(newDens[i] - dens[i])

  END FOR


  dens = newDens

  iter += 1

  // continue the iterative procedure while the difference is large enough

  // and the number of iterations smaller than maxIter

WHILE df > diffThreshold AND iter < maxIter
```

The output that we obtained dens, can be written in the following output table:

Table 8: Example of the present population estimates on grid tile level

grid tile	population	day	time
1	654	2023-07-14	2pm
2	234	2023-07-14	2pm
3	1654	2023-07-14	2pm

In case the present population should be estimated for a specific zoning system, say municipalities, then the population can be aggregated accordingly:

Table 9: Example of the present population estimates on regional level

municipality	population	day	time
A	563465	2023-07-14	2pm
B	125435	2023-07-14	2pm
C	23654	2023-07-14	2pm

Note that these are the estimates for the part of the population that are subscribers of the single MNO in this UC. The same process should ideally be run for all the other MNOs in the country of interest. The estimates can simply be summed. In case data from at least one MNO is missing, the market shares of the MNOs for which data is available can be normalised to 1. In this way, we pretend that the MNOs for which data is missing are non-existent and the other MNOs share the market among them. The present population will therefore be estimated using the event data from those MNOs as described above. This is suboptimal, because the main assumption is made that the mobility behaviour of costumers of the used MNOs represent those of the costumers of other MNOs.

ANNEX III - PIPELINE APPLICATION TO THE M-USUAL ENVIRONMENT INDICATORS USE CASE

In this annex, we describe the application of the high-level pipeline to the use case on M-Usual Environment indicators, providing input to the specification of the methods for the daily, mid-term and long-term processing modules. We also provide descriptions of the summaries we propose to compute along the longitudinal analysis.

Furthermore, we provide a pseudo-code for the elaboration described in the daily processing module.

CONCEPTS AND REQUIREMENTS FOR THE USUAL ENVIRONMENT USE CASE

According to the Eurostat glossary[14], the Usual Environment (UE) refers to the geographical area in which an individual carries out their regular daily activities. The areas that make up the UE do not need to be contiguous. In order to determine an individual's UE, the following criteria must be taken into account:

- duration of the visit;
- frequency of the visit;
- purpose of the visit.

However, since the 'purpose' cannot be observed with MNO data, the concept of UE must rely exclusively on frequency and duration. The concept plays a major role in tourism statistics; to be considered a tourism trip, the trip must take the traveller outside their UE. It is important to note that the UE can also be a set of non-adjacent locations; i.e. a cluster of non-contiguous geographical areas.

In this first version of the UC, we focus on determining the UE and assigning the labels of Home Location, Workplace and Second Home whenever possible. Such location labels will be assigned in the long-term module.

An extended and more advanced UE use case will be further developed during the project.

For some mobile devices, it will not be possible to determine the UE with the available information, e.g. in the following cases:

1. 'Rarely observed': devices that are detected for an insufficient amount of time (below a certain threshold *Tobs)*.
2. 'Highly moving': devices with no places reaching a given visit duration/frequency threshold *Tprev* even if passing the threshold *Tobs*. The threshold *Tprev* is a parameter whose value must be set according to the definition of the UE (e.g. one third of the total observation time in one year and the presence of regularity in the frequency of visits)

The identification of these devices will be implemented in the mid-term or in the long-term processing module.

[14] See: Glossary:Usual environment - Statistics Explained (europa.eu)

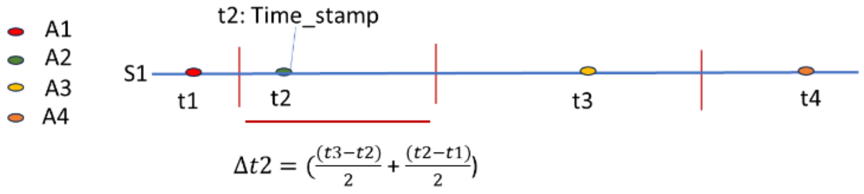
DAILY PROCESSING MODULE

The first steps in the pipeline (including the data cleaning, the calculation of quality metrics and the device demultiplex) are pre-processing steps common to all the UCs and are therefore not reported in this section. This description starts with the device demultiplex; it logically applies to the single device k.

The first step specific to the current UC is detailed by the function *"Assignment of daily permanence scores"*. This function integrates information from the Event Data Flow and from the Cell Footprint Data Object (Output of the Cell Footprint estimation) and gives as output a "*permanence score*" for each tile of the reference grid and for each time slot associated with the temporal resolution we have chosen (e.g. one hour). The permanence score is a proxy of the time spent by the device in all tiles of the Cell Footprint (CF) associated with the cell_id to which the device is connected.

To estimate the time spent in a CF, we can proceed as follows:

1. Cell ids in the Event Data Flow by device are associated with the corresponding CF from the Cell Footprint data object;
    a. In case of contiguous events in the same CF, the elapsed time between the events is assigned to the CF if it is below a given threshold. This threshold could also be differentiated for nights and days because many devices are switched off during the night. For example, we can fix the night threshold to about 8 hours and the daily one to about 2 hours. In case the elapsed time is over the threshold, it will be associated with an "unknown place" rather than to a CF.
    b. In the case of two contiguous events occurring in different CFs, the semi-distance in time between each pair of events is calculated and the time spent in the two CFs is calculated as the time interval (Δt) between the two semi-distances before and after the event. An example is provided below:



In the figure above, the dots represent events (at time t1, t2...), their colours represent the associated CFs (A1, A2..., different colours for different CFs), while the red vertical lines correspond to the temporal semi-distances between two events. Δt2, with the explicit formula, represents the time spent in the CF A2.

If the temporal distance between two events occurring in two disjoint CFs is too long, we may not be sure of the position of the device in that time window, so we use a time threshold (T*sd*) for the semi-distance assignment, as follows:

2. When the calculated semi-distance between two events is higher than a threshold value (T*sd*) we assign the threshold value T*sd* instead of the semi-distance. As an example, we may take T*sd* = 15 min. The remaining time periods not covered by the threshold (i.e. the portion of the semi-distance exceeding T*sd*) are associated to an "unknown place" rather than to a CF. In formulas:

$$CFP = \min(\Delta t2, Tsd)$$
$$UNP = \max(\Delta t2 - Tsd, 0)$$

Where CFP is the Cell Footprint Permanence and UNP is the unknown permanence

Remark

Time periods not covered by events may be due to several reasons related to network dynamics or individual behaviour. Cases in which a device is switched off and then turned on in the same cell pose a problem in this use case, as these cases may indicate the overnight sleeping behaviour of individuals that we are interested in detecting. We deal with these situations in step 2, assuming that devices "appearing" and "disappearing" in the same CF are not moving out of that CF during the night. A different rationale and consequently different choices may be applied in other use cases.

Since the purpose of the UC is to identify UEs, we are mainly interested in finding 'stays/permanence' in the process, while we might disregard 'moves'. When calculating the time spent per CF, we will discard events corresponding to movements as described below.

3. In the case of subsequent events (t1, t2, and t3 in figure 1) taking place in 3 different CFs (A1, A2, and A3 in the figure), the intermediate event can be disregarded if the calculated device speed is above an assigned threshold. We define a space distance between CF A1 and CF A3 as the minimum distance between CF A1 and CF A3 (if A1 and A3 are adjacent, or even overlapping, the minimum distance is zero). The lower bound for the speed of the device can be approximated by the ratio between the minimum space distance calculated and the time interval between the events associated with CF A1 and CF A3. If the device speed calculated to reach CF A3 from CF A1 is greater than an assigned Speed threshold (T*vmax*), the CF in between is assigned a zero-permanence score, i.e. CF A2 (equivalent to discarding the associated event). T*vmax* is a parameter to be chosen and agreed upon. The pseudo-code for this processing is provided in Annex 2 - Pseudo-code for the daily processing module of the usual environment use case. In this processing, it could be helpful to consider cell information associated with, for example, highways, railways, stations or underground as a further criterion to discard the event in these cells for this UC.

Important remark

The threshold defined in Step 4 and the dedicated function to calculate the "speed" of the device in Step 5 are defined and set for the specific use case in the usual environment. Another use case, e.g. the one on Mobility, will require a more refined function to identify the speed of the device compared to the one proposed in this use case. The pipeline allows for such flexibility, preserving some easy-to-set and easy-to-explain choices. We will also discuss these choices from the perspective of a time segmentation/user diary approach.

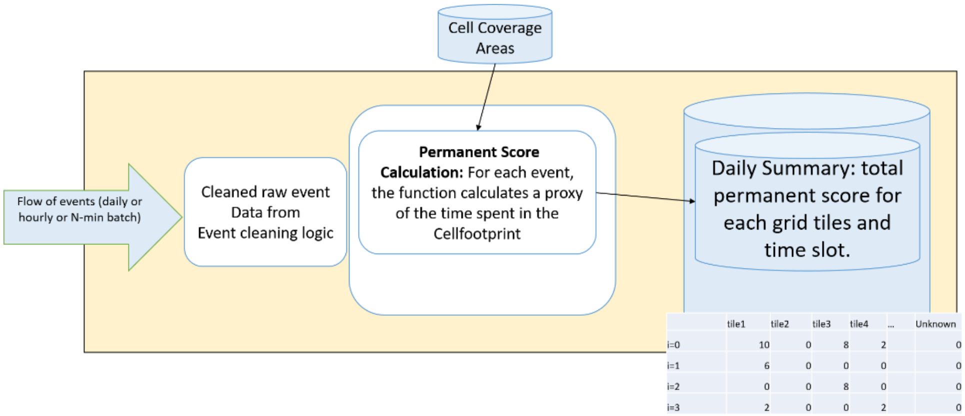
4. As final step, we use the time spent (Δt) in a given CF for a given event to assign a permanence score to each tile belonging to the CF. The score assignment can be done in different ways depending on the use case. For the purpose of the UE UC, it seems reasonable to give the same permanence score, equal to the whole time spent, to all the tiles covered by the corresponding CF.

The resulting (intermediate) output contains the permanence score per tile per event. It is worth noting that, if we use the equal assignment described above of Δt to each tile belonging to a CF, the permanence score is an intensity measure, so we cannot sum the permanence scores of different tiles.

DAILY SUMMARIES

The output of this process step is the daily permanence score for each tile and time slot, calculated as the incremental sum of the event-by-event permanence scores, conveniently discretised and normalised. This approach is advantageous because it allows more detailed information to be maintained, which could be used in more advanced versions of the use case without excessive computational effort.

For the sake of illustration, the daily summaries for the permanence score can be organised in the following sparse matrix, where we report the daily timeslot in rows and the tiles, including the 'unknown' tile, in columns. This representation is for illustration only and the actual data structure will be defined in the implementation phase.


Cell Coverage Areas
Flow of events (daily or hourly or N-min batch)
Cleaned raw event Data from Event cleaning logic
Permanent Score Calculation: For each event, the function calculates a proxy of the time spent in the Cellfootprint
Daily Summary: total permanent score for each grid tiles and time slot.

	tile1	tile2	tile3	tile4	...	Unknown
i=0	10	0	8	2		0
i=1	6	0	0	0		0
i=2	0	0	8	0		0
i=3	2	0	0	2		0

Figure 20: Sparse matrix for the organisation of the daily summaries for the permanence score

It could be useful to evaluate some quality metrics during this process. For example, convenient indicators can be:

- Number of tiles with a non-zero permanence score
- Maximum permanence score
- Median permanence score
- Minimum permanence score
- Daily tile frequency

These indicators can be used to identify potential issues in the mobile network and unusual behaviours of the device. Quality indicators also provide alerts for unexpected events.

MID-TERM PROCESSING MODULE

Starting from daily summaries, mid-term (e.g. monthly) behaviours and information can be extracted through aggregation procedures. For the purpose of the UE UC, it may be convenient to differentiate between the permanence score in specific sub-daily periods for each tile, e.g. to record the permanence score relative to the night time (8pm-8am), working time (8am-5pm) and evening time (4pm-10pm). Obviously, the permanence score in the sub-daily periods will not sum up to the daily permanence score if the sub-daily periods overlap. Furthermore, the above hourly intervals defining time sub-periods are given as recommended sub-daily periods that overlap when the corresponding daily life activities also overlap (e.g. dinnertime and night time). However, they will be configurable by NSIs in order to suit different lifestyles across countries. The recommended sub-daily periods as well as the pros and cons of overlapping daily periods, will be discussed with the interested stakeholders (TFMNO, Advisory Board, etc.) and the result of the discussions will be reported in the final version of this document.

Daily summaries are aggregated monthly by using different methods providing monthly indicators, for example:

- The monthly permanence score for each tile S can be calculated as a summation of the tile daily scores for the daily period and for all the considered sub-daily periods.
- The summation can also be performed by differentiating specific types of days, hence summing per different sub-monthly periods, such as working days, weekends and holidays; This operation may require the use of information collected in the data object 'Calendar info';
- A measure of the monthly permanency's *frequency* per tile can be calculated counting the number of days the tile has a non-zero permanence score; the monthly frequency is calculated for the daily and sub-daily periods, as well as during the working days, weekends and holidays;
- Measures of the "*regularity*" of permanencies per tile are calculated using at least the mean and the median values of the distribution of the number of days between consecutive permanencies in each given tile for an individual. It is recommended to also calculate these measures for daily and sub-daily periods, as well as for working days, weekends and holidays; additional measures or more sophisticated functions to identify regularity, as well as values for the aforementioned parameters, will be discussed with the TFMNO, the Advisory Board, the MNOs and will be reported in the final version of this document;
- Prevalent tiles represent the device's usual environments and can be identified by applying some thresholds to previously computed indicators (details on recommended threshold values, e.g. for the monthly permanence score S, will be discussed in the final version of the document). According to the definition of the usual environment concept, the threshold values should be based on both *regularity* and *frequency*. Prevalence tiles can be calculated per daily and sub-daily periods, as well as per working days, weekends and holidays.

MID-TERM SUMMARIES

The results of the functions described above can be reported in the mid-term summaries. Regarding the measure of tiles prevalence, the resulting prevalent tiles can be organised in an ordered list after applying the above-mentioned thresholds.

Considering the spatial resolution of interest, for example the administrative territory such as: municipal, district, and other, it is necessary to transform the territory from tiles to administrative territory and to recalculate the mid-term summaries.

As with the daily summaries, we can evaluate some quality metrics for the monthly analysis. They can be:

- Number of tiles with non-zero permanence score per month;
- Maximum permanence score per month (this can be used to eliminate devices with a maximum permanence score below a given threshold);
- Median permanence score per month;
- Minimum permanence score per month.

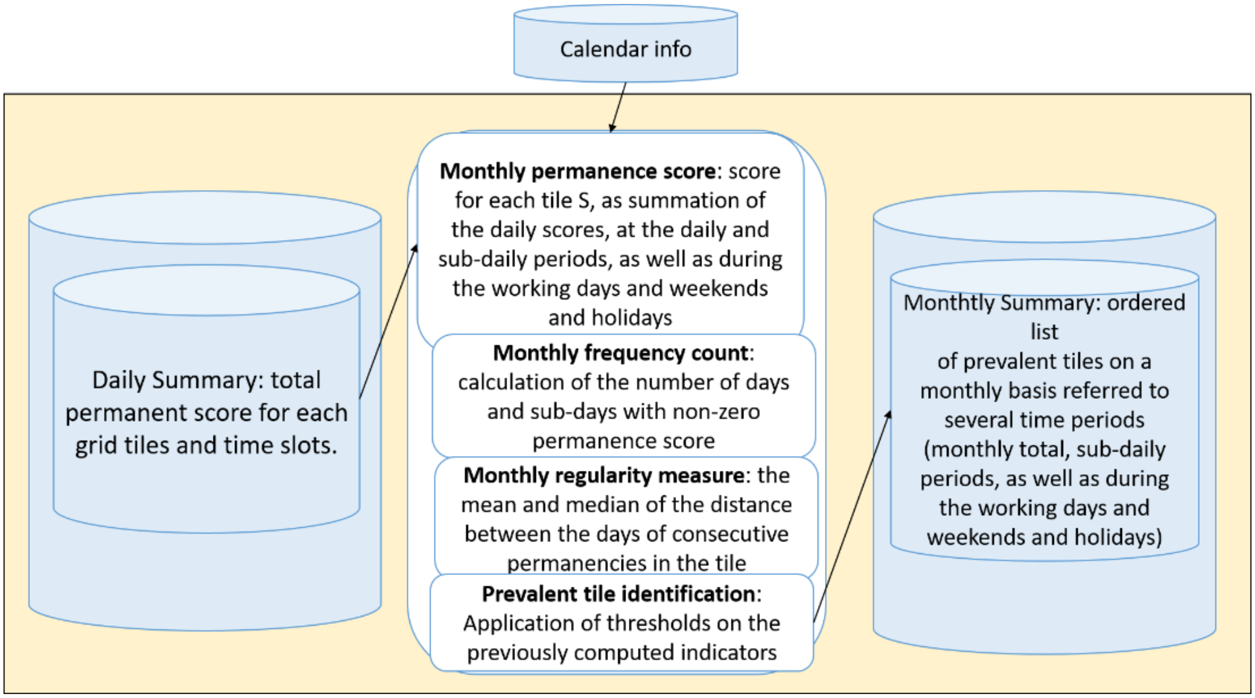

Calendar info
Monthly permanence score: score for each tile S, as summation of the daily scores, at the daily and sub-daily periods, as well as during the working days and weekends and holidays
Monthly frequency count: calculation of the number of days and sub-days with non-zero permanence score
Monthly regularity measure: the mean and median of the distance between the days of consecutive permanencies in the tile
Prevalent tile identification: Application of thresholds on the previously computed indicators
Daily Summary: total permanent score for each grid tiles and time slots.
Monthtly Summary: ordered list of prevalent tiles on a monthly basis referred to several time periods (monthly total, sub-daily periods, as well as during the working days and weekends and holidays)

Figure 21: Monthly (mid-term) processing module diagram

LONG-TERM PROCESSING MODULE

The long-term processing step provides the final output at device level; i.e. the UE identification.

'Usual Environment':
Starting from the mid-term summaries we consider the first monthly prevalent (clusters of) tiles in terms of total permanence score and frequency. With regard to the total permanence score, the number of prevalent clusters of tiles to be considered could be identified in a number of ways, for example dynamically, taking into account the percentage of the total time the device is observed. The output of the procedure is the collection of areas corresponding to the union of the n prevalent clusters of tiles in terms of permanence scores and the m prevalent ones in terms of frequency that do not overlap with previous ones.

Starting from the mid-term summaries, information can be extracted to identify a list of places of interest. Each place of interest comes with a number of characteristics on which the heuristics for its identification are based. A possible procedure is provided below:

1. *'Home location'*: In the simplest case the home location would be the single tile (or single cluster of tiles) whose permanence score summed on all months is the highest, with a gap from the second highest above a certain threshold value (Gap). If there are two or more tiles (or clusters of tiles) with the total permanence score which have a difference that does not exceed the threshold Gap, we proceed as follows:

We take into account the corresponding permanence scores in the night periods (to exclude work places). We take the home location as the one with the highest night permanence score. If even the difference between the night permanence scores is under a defined threshold, we take all as possible home locations for the device (and in the next steps, we will distinguish, if possible, the 'second home'). Otherwise, the tile (or cluster of tiles) with the higher total permanence score can be labelled as the probable workplace and used in the workplace analysis.

2. *'Second home'*: can be identified as the second most prevalent tile (or cluster of tiles) resulting from night permanence scores. If we were not able to assign a single home location in step 1, here we analyse the permanence scores in weekends, holidays and summertime to distinguish the second home. Additional heuristics can be defined to distinguish between a workplace home (during working days) and a family home (during weekends and holidays).
3. *'Workplace'*: can be identified as the tile (or cluster of tiles) with the highest permanence score in the daytime. Cases of individuals working mainly at night are not detected by this procedure and therefore need a specific analysis of additional information (e.g. information about the cell in specific places like hospital, airports, places needing night surveillance, etc.).
4. *'Unlabelled UE':* cases where, despite the presence of prevalent tiles, it is not possible to distinguish between Home, Second Home and Workplace.

LONG-TERM SUMMARIES

The results of the functions described above can be reported in the long-term summaries. In addition to the identification of the usual environment tiles specifically for 'Home location', 'Workplace' and 'Second home', an output of the procedure will be the information that no usual environment has been identified.

As with the mid-term summaries, some quality metrics could be defined for the long-term analysis.

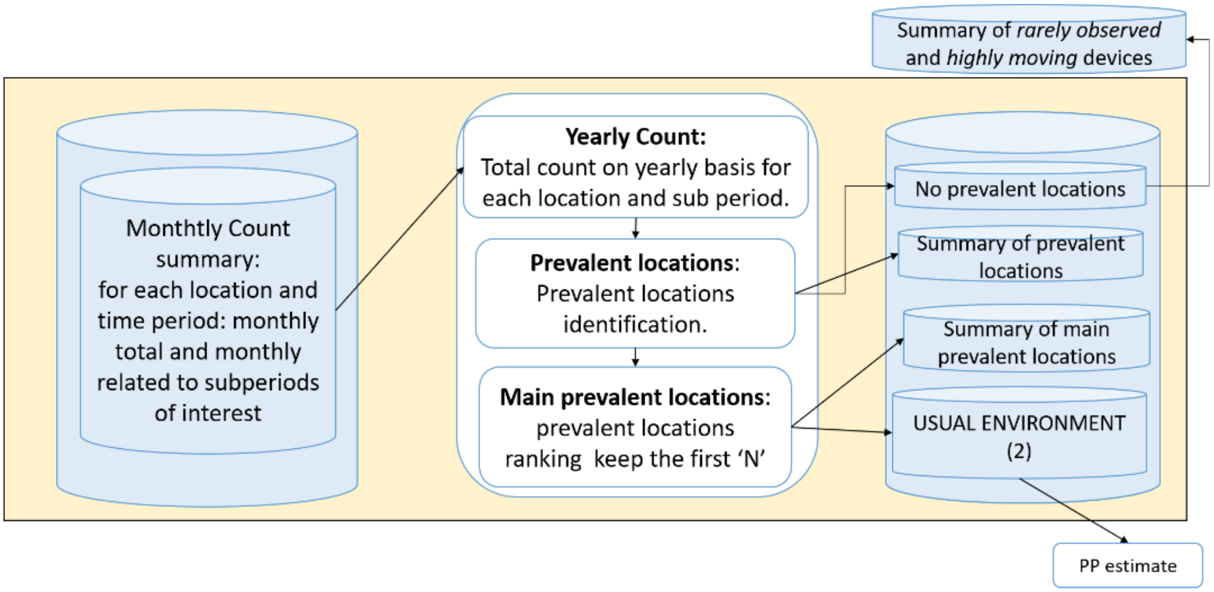

Summary of *rarely observed* and *highly moving* devices
Monthtly Count summary: for each location and time period: monthly total and monthly related to subperiods of interest
Yearly Count: Total count on yearly basis for each location and sub period.
Prevalent locations: Prevalent locations identification.
Main prevalent locations: prevalent locations ranking keep the first 'N'
No prevalent locations
Summary of prevalent locations
Summary of main prevalent locations
USUAL ENVIRONMENT (2)
PP estimate

Figure 22: Long-term processing module diagram

CALCULATION OF INDICATORS

Long-term summaries elaborated for the different devices can then be aggregated to obtain the population estimates required by the use case. Such aggregation and the estimation procedures are not reported in the present document, they will be detailed in deliverable D2.2.

PSEUDO-CODE

In this annex, we provide the pseudo-code for the daily processing module of the Usual Environment use case, described in human language in Method 2: Daily Permanence Score Estimation, to calculate the 'stays/permanence' for each tile for each device.

For a single device k, we consider a daily partition in "n" time slots (n can be 6, 12, 24) and we call "h" the index of a single time slot. Each time slot h corresponds to a time interval in the day that we call time window, defined by as follows:

```
time window[h]= time_end[h]- time_start[h]
```

Let us define the following:

- k is the device index
- i is the event index: event[i] refers to the i-th event, event[i].time refers to the timestamp of the i-th event, tile(CF) is the set of tiles included in the CF.
- event[i].cell_ID is the id of the cell to which the device is connected at the event[i].
- event[i].cell_ID is associated to event[i].CF through the cell footprint estimation procedure. event[i].CF is the cell footprint, that is the set of tiles corresponding to the geographical area covered by the cell represented as a grid (see Section 7.2)
- tile(event[i]) is the array of all tiles associated to event[i].CF.
- tile.permanence[k][h] is the total time the device k has been seen in the tile in a given daily time slot h, estimated by the function stay_permanence_intiles().
- tile.frequency[k][h] is the daily frequency the device has been seen in the tile in a given daily time slot h, estimated by the function stay_permanence_intiles()(it can be used as a quality metric).
- for a given device k and a given time slot h, device[k][h].unknowntime is the total time the device k is in an unknown location.
- given two cell footprints CF1 and Cf2, distmin(CF1,CF2) is the minimum distance between CF1 and CF2 (the specific distance function to adopt is to be defined). If CF1 and CF2 are adjacent cell, distmin(CF1,CF2)=0
- T*sd* is the maximum permanence threshold assigned in case of sparse event observations, e.g. T*sd* = 15min, the value to be decided together with the partners
- T*vmax* is the "velocity threshold", e.g. T*vmax* = 50 km/h, to be decided together with the partners. It is used to disregard events corresponding to "moves" as opposed to "stays/permanence" according to the procedure described below (the event/record is flagged for excessive speed):
- "+=" is used for the increment function
- "==" is used for the strict equality

For a single time slot h, we cycle on the events train (event index i), using the function stay_permanence_intiles(). Given a device k and a time slot h stay_permanence_intiles() function is defined as:

```
for i=1:m

if event[i].time is in time window[h]

if event[i].CF == event[i-1].CF

    tile(event[i].CF).permanence += event[i].time - event[i-1].time;

    tile(event[i].CF).frequency +=1

    i+=1;

else if  distmin( event[i-1].CF, event[i+1].CF) < (event[i+1].time - event[i-1].time) * Tvmax

     left_time = event[i].time - event[i-1].time;

     tile(event[i-1].CF).permanence += min(left_time/2, Tsd);

     tile(event[i].CF).permanence += min(left_time/2, Tsd)

     device[k].unknowntime += max( left_time -- 2*Tsd, 0)

      tile(event[i].CF).frequency +=1

      i+=1;

else  skip event[i];

       i+=1;

endif

end
```

The function stay_permanence_intiles() returns tile.permanence[k][h], tile.frequency[k][h] and device[k][h].unknowntime as output.

In order to get a permanence score per tile we do the following:

```
            tile.permanence_score[k][h] = tile.permanence[k][h] / (time_end[h]- time_start[h])
```

We perform this normalisation in order to get values representing a density of permanence per tile. The permanence values obtained for each tile can be conveniently discretised.