Trusted Smart Statistics: methodological developments based on new data sources

2022-IT-TSS-METH-TOO
Project n. 101132744

Work Package 2
Landscaping analysis of the non-MNO data sources

Deliverable 2.3

Consolidated analysis of non-MNO sources and report on their availability and their connection with potential target statistics

August 2025

Partner in charge: Destatis (Germany) – *Gloria Deetjen*

Authors[1]: CBS (The Netherlands) – *Yvonne Gootzen, Johan van der Valk*
DESTATIS (Germany) – *Gloria Deetjen, Natalie Rosenski*
INSEE (France) – *Mélina Hillion, Marie-Pierre Joubert, Chloé Breton, Julien Pramil*
INE-PT (Portugal) – *Sónia Quaresma, Antonio Portugal, Pedro Cunha*
INS (Romania) – *Marian Necula, Rares Cartuta, Bogdan Oancea*
SCB (Sweden) – *Remy Kamali, Wictoria Widén, Pär Hammarström*
ISTAT (Italy) - *Giorgia Simeoni, Gabriele Ascari*

MNO-MINDS | Eurostat CROS (europa.eu)

Co-funded by the European Union

[1]Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Acknowledgements

We thank all respondents of the survey on data sources for their valuable contributions, as well as the participants of the SPRINT and ESTP course for their active engagement and inspiring discussions. We extend special thanks to M. Francesca D'Ambrogio and Immacolata Fera (Istat) for their assistance with English language, to Stefan Svanström (SCB) for sharing his expertise in GIS, and to Daniel Knapp (Destatis) for his advice during the development of the short survey. We are also grateful to the tourism experts at INE and Destatis for their insightful input on data sources, and to Bruno Zamengo (Motion Analytica) for providing information on event ticket data.

Deliverable 2.3
Consolidated analysis of non-MNO sources and report on their availability and their connection with potential target statistics

Summary

In most applications of Mobile Network Operator (MNO) data, such data are used together with other sources to calibrate, stabilise, or correct results. To explore the possibilities and limitations of MNO data for official statistics, it is therefore essential to analyse potential complementary sources and their relevance to target statistics.

The objective of Work Package (WP) 2 in the ESSnet project MNO-MINDS is to identify and assess data sources that can be integrated with MNO data. This was carried out through a landscaping analysis, evaluating each source's potential for integration—here referred to as *non-MNO data sources*. The assessment considered both the cost of access and handling, and the relevance for integration. Each source was scored and categorised as *most promising*, *promising*, or *less promising*. While these categories provide guidance on costs and benefits, the final choice of data sources for specific applications remains with the reader or the National Statistical Institutes (NSIs).

The first two WP2 deliverables presented initial landscaping results through an assessment matrix, detailed written analysis, and data scoring of most non-MNO sources considered. This third and final deliverable consolidates the complete landscape analysis. It further includes: (i) results from a short survey among NSIs within the European Statistical System (ESS) on data availability, (ii) links to potential target statistics through the data scoring, and (iii) an example of a commuter application scenario with a metadata-analysis, presented in a separate chapter.

Altogether, 18 data sources were assessed. The *most promising* category includes only official sources such as the census, total population registers, national travel surveys, land use and land cover registers, and tourism accommodation statistics. The *promising* group covers a broader range of sources, including sensors, pollution and satellite data, electronic invoices, tourism household and border surveys, tourism platform data, and credit card transaction data. The *less promising* group comprises vessel and boat traffic data, event ticket data, Google Maps popular time, smart meters, connected vehicles, and social media.

Index

List of Tables

List of Figures

1. Introduction

Society's increasing digitalization has led both to the emergence of new data sources (e.g. smart sensors, MNO data, transaction data, remote sensing data, …) and to an increasing need for more detailed and frequent official statistics, notably to combat fake news. Moreover, the acceleration of global warming and the growing awareness of how important monitoring sustainable development goals are, have highlighted the need for more accurate statistical indicators, capable of capturing the complexity of the phenomena. New indicators and methodologies describing society in the Information Age have been identified among others by the European Statistical Advisory Committee (ESAC, 2018).

Many new data sources have already proven their potential for informing public policies. Mobile Network Operator (MNO) data has been widely utilised by National Statistical Institutes (NSI). During the COVID-19 crisis, MNO data were used in France and Germany to document population movements during lockdowns, allowing for better calibration of public services (Coudin, Poulhes, & Suarez-Castillo, 2021). These data have also provided short term conjunctural indicators to estimate the crisis impact on the economy in a shorter period than with traditional indicators. Outside the COVID crisis period, many European NSIs have access to aggregated and anonymised MNO data. These data are typically bought by NSIs or obtained thanks to specific research funding. They are applied in pilot projects studying day-time population, population mobility or tourist attendance. The ESS published in 2021 a position paper on the future Data Act proposal stating that: Access to privately held data is urgently needed for producing new, faster, more detailed official statistics (ESS, 2021). The future evolution of the legislation regarding access to private data might allow NSIs to gain access to more detailed data. Regulation (EC) N. 223 (No223, 2024) goes in this direction by stating that "access to new data sources in general, and in particular to privately held data, for the development and production of European official statistics on a sustainable basis and according to fair, clear, predictable and proportionate rules, in line with the Union's fundamental rights framework, should be ensured", yet discussions have to be held regarding its concrete application in every EU country.

MNO data used by NSIs are mostly signaling data. These data are initially recorded by the operator for technical reasons (e.g. network maintenance, mobile service delivery, damage detection). This is both an asset, since they capture a large spectrum of information, and a challenge because they do not necessarily meet official statistics' quality criteria. Combining information from multiple mobile operators is a key element to improve the representativeness and therefore the quality of the statistics. Eurostat's service contract "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)" is precisely aimed for this purpose. Its goal is to release an open-source software allowing to implement the methodology. Yet this project targets the combination of multiple MNO data. The current ESSnet MNO-MINDS aims at combining MNO data with other data to obtain better official statistics, which are more accurate and cover a larger spectrum of issues. This Work Package focuses on the landscaping and scoring of the most promising data sources to be combined with MNO data. These sources are also referred to as *non-MNO data*.

This deliverable aims to provide an overview on several non-MNO sources that can potentially combined with MNO data. It classifies them into three groups:

Most promising: Data sources in this category are widely accessible for European NSIs, they mostly or fully comply with quality criteria from the European Statistics Code of Practice (CoP), and there is a high range of application scenarios feasible. The analysis for these sources is provided in detail with examples for potential applications in official statistics.

Promising: These sources have a high potential to be applied together with MNO data in official statistics but would need to overcome some methodological, technical, legal obstacles. For example, they might be difficult to access but with a perspective at middle term. Such obstacles are highlighted in the analysis and scoring parts for each source individually.

Less promising: This category describes data sources which are in a middle term perspective (approx. 5 years) less relevant to be applied together with MNO data. The reasoning for each source may vary significantly and is indicated in the dedicated sections. Whereas some sources require substantial work to be applied alone in official statistics or whose usage would even raise considerable problems of social acceptability, other sources are indeed promising for official statistics but currently lack relevant applications together with MNO data.

The first part of this deliverable delves into the reasons for combining MNO and non-MNO data which is indeed an essential step for identifying accurate non-MNO data. In addition, the main types of non-MNO data which are considered in the analysis are described as well as the whole process of the landscaping analysis. Next, the criteria for data scoring are introduced in chapter 4 before the detailed scoring for each source is portrayed in chapter 5.

One crucial element to assess the relevance of data sources for the integration with MNO data is whether they are available or potentially available to NSIs. Especially for privately held new data sources it is hard to gain an overview how many NSIs have current access to which source. Therefore, a short survey was conducted and the main results are presented in chapter 6.

Further topics that regard data integration not for each source individually but as whole are covered in chapter 7: It draws the connection to WP3 by referring to potential target statistics and to WP4 by outlining questions for a potential survey that may enable a better linkage of data sources. In addition, the relevance of metadata analysis are elaborated with the example of a commuter scenario.

The consolidated analysis in chapter 8 provides a broad overview on general results and refers especially to data sources for tourism statistics as there are sources in different scoring categories considered in the landscaping analysis. Finally, the conclusion summarises the main results.

2. Reasons to combine MNO and non-MNO data

The potential of Mobile Network Operator (MNO) data to contribute to official statistics has been extensively documented. Building on the significant methodological advancements and open-source tools developed under ESSnet Big Data 1 and 2, the ESS Task Force on the use of MNO data for Official Statistics published in September 2023 a position paper entitled “Reusing Mobile Network Operator data for Official Statistics: the case for a common methodological framework for the European Statistical System” (MNO, 2023). This document underlines, among its key high-level requirements, the necessity of establishing a framework capable of systematically integrating MNO data with complementary non-MNO data sources.

Although a range of valuable applications of MNO data—such as measuring population presence and analysing mobility patterns—has been identified, the limitations of using these data in isolation are well recognised and have been highlighted in several publications by National Statistical Institutes (Sakarovitch, Bellefon, Givord, & Vanhoof, 2019). In light of these constraints, this report will not revisit the detailed shortcomings but will instead concentrate on two primary avenues through which the combined use of MNO and non-MNO data can substantially enhance the quality and reliability of official statistics. This focus will, in turn, serve to identify the most relevant non-MNO data sources to complement and strengthen the value derived from MNO data.

2.1. Improving population statistics

A fully exhaustive database covering the entire population is unlikely to exist. Even administrative registers have limitations, as most National Statistical Institutes (NSIs) primarily maintain databases focused on the residential population and regular mobility patterns (e.g., commuting). The official statistical system thus lacks comprehensive information on daytime population and mobility behaviours at fine geographical and temporal scales.

New data sources, such as MNO data, traffic loop sensors, or transaction data, can address these gaps but also present challenges. As they are not initially collected for statistical purposes, their representativeness is often biased due to market share variations or technical factors (e.g., users switching off their phones).

The primary objective of combining MNO and non-MNO data is to improve both the representativeness and the spatial-temporal precision of population statistics. This approach must address potential inconsistencies between sources, such as divergent definitions of fundamental units or reference periods. Regular data access is also essential for producing reliable official statistics in this area. In line with the concepts introduced in WP3, MNO data would serve as a proxy for the target measure, offering unique insights into daytime population with a level of precision unmatched by other sources. Non-MNO sources would act as auxiliary data, strengthening the robustness of the analysis. As noted in the position paper (MNO, 2023), the observed population of mobile devices does not perfectly align with the target human population, and the gap between the two is dynamic. Therefore, combining MNO and non-MNO data enables the calibration, correction, and stabilisation of final statistics.

Several applications can be envisaged. Daytime population estimates could enhance tourism statistics, particularly in countries without border surveys. Mobility statistics represent another key area: once geographical areas are categorised by function, MNO data can provide detailed information on the use of these areas based on individuals' place of residence. For example, combining MNO data with tax register data containing socio-demographic and income variables would enable a more comprehensive analysis of mobility patterns by population characteristics.

2.2. Broadening the scope of issues covered by official statistics

Combining MNO and non-MNO data sources allows official statistics to address emerging topics and remain relevant in the context of rapid societal and environmental changes. In this framework, MNO data provide detailed insights into presence and mobility patterns, while non-MNO data continue to serve as the primary basis for information.

Potential application areas include the analysis of car-sharing patterns, social media usage, and factors contributing to local spikes in air pollution. When evaluating candidate sources for integration with MNO data, priority will be given to applications focusing on daytime population, tourism, and mobility - areas collectively identified by ESSnet MNO-MINDS as the most critical.

Further details on the potential application scenarios of MNO data for official statistics, as well as the related methodological and operational challenges, are currently being examined in Eurostat's Multi-MNO project.

3. Landscaping process

Before addressing the theoretical and analytical parts of this deliverable, a short overview of the landscaping process helps to clarify how non-MNO data sources were identified, analysed, and scored, and how the results were derived.

The process began with insights from past and ongoing experimental studies using MNO data in combination with other sources, which served as a basis for identifying potential non-MNO data sources. This initial list was further enriched by existing literature and brainstorming contributions from the SPRINT event.

Based on selection criteria and data-type categories, an assessment matrix was developed as the foundation of the analysis. Results from this first stage were presented in Deliverable D2.1 of WP2. The analysis was then deepened for all sources and a scoring scheme was introduced. Scoring comprised:

- **Primary scoring**, which summarised the main categories from the assessment matrix.
- **Secondary scoring**, which applied principles from the European Statistics Code of Practice.

Further details on selection criteria and scoring are provided in Chapter 4.

Over nearly two years, the analysis was refined through additional literature reviews, extended assessments, and expert interviews for specific data sources. The full results and final list of sources are presented in Chapter 5. Each data source was assigned to a WP2 partner, which occasionally led to slight differences in analytical style despite efforts to harmonise. Partners also included specific examples of how non-MNO data could be integrated with MNO data; however, these examples are illustrative only and do not limit the applicability of the results.

Figure 1: Landscaping process

A short survey was subsequently conducted among National Statistical Institutes (NSIs) to assess the availability of data sources and to complement the earlier scoring. Survey results are presented in Chapter 6.

Finally, Chapter 7 links the landscaping results to potential target statistics from WP3 and to a dedicated MNO survey from WP4. It also supports the integration of diverse datasets with MNO data through a metadata analysis. Chapter 8 consolidates the findings, highlighting the main advantages and limitations, while the final list and scores for each source are included in Chapter 5.

4. Criteria for analysing the relevance of non-MNO data

This chapter describes the criteria for analysing and scoring data sources for the integration with MNO data. Data can be categorized by data types and whether they are traditional or new data sources. Scoring consists of primary scoring dimensions from the assessment matrix and secondary scoring dimensions derived from the European Statistics Code of Practice (CoP).

4.1. Types of non-MNO Data

To assess the relevance of various non-MNO data sources, this report builds on the work of ESSnet WP3, the Norwegian NSI (Zhang, Haraldsen, Pekarskaya, & Hole, 2018) and ESSnets Big Data 1 and 2 (Kowarik & members, 2020).

The resulting categories of non-MNO data sources are therefore as follows:

Data Type
a. **Survey** (national travel survey, labour force survey (LFS), SILC, …)
b. **Register and administrative data** (employment register, diagnoses, wage, income tax, welfare payments, …)
c. **Transaction** (scanner data, point-of-sales receipt, bankcard or giro payment, P2B or B2B invoice, P2P (e.g., Paypal), property sales contracts, ownership registration, …)
d. **Static detection of connected objects and environmental phenomena** (smart meters readings, weather station readings, traffic loop signals, lorry tracking signals, …)
e. **Mobile airborne sensing** (satellite images, drone images, airborne laser scanning, …)
f. **Internet** (web pages, social media posts, …)

Table 1: List of Data Types

Some non-MNO data sources, such as survey-based censuses, are produced directly by official statistics and are designed from the outset to meet NSIs' quality standards and variable requirements. Others, though originally collected for administrative purposes, are traditionally used by NSIs. These data typically require specific processing to meet quality standards but are generally well structured and aim to provide exhaustive population coverage. One data type is not explicitly mentioned in Table 1 but emerged especially for conducting censuses — the combination of survey and register data — to reflect its growing use by NSIs and its distinct characteristics compared to survey or administrative data alone.

A third category includes data produced for purposes unrelated to official statistics, by either private or public entities. These sources are often less structured, may lack documentation on variables relevant to NSIs, and frequently present quality issues related to the data collection process.

In addition to distinguishing data types, it is useful to categorise non-MNO data into *traditional data sources* — official data such as those produced or used by NSIs — and *new data sources*. These two categories generally differ in terms of access conditions, characteristics, limitations, and potential benefits.

Traditional Data Sources	*New Digital Data Sources*
• *Survey* • *Combination of Survey and Register* • *Register and administrative data*	• *Transaction* • *Static detection of connected objects and environmental phenomena* • *Mobile airbone sensing* • *Internet*

Figure 2: Categorisation of Data Types

Traditional data often face challenges such as limited timeliness and granularity, non-response, and the need to balance low burden for respondents with sufficient data collection. New data sources, by contrast, typically suffer from non-representativity, large and not harmonised datasets, access might be limited or it may be difficult for NSIs to gain access, plus there might be potentially high costs.

Given these differences, it is essential to consider the data category of each source when analysing their level of relevance for the integration with MNO data in the following sections.

4.2. Primary scoring dimensions

To evaluate the potential of non-MNO data sources, WP2 adapted the *Big Data Classification Matrix* from ESSnet Big Data 2 (Kowarik & members, 2020) to the specific context of combining MNO and non-MNO data (Figure 3). The key scoring dimensions are outlined below, with particular focus on new data sources, as traditional sources already used by NSIs typically meet these criteria.

Data type: Non-MNO sources vary significantly in size and complexity. Some, such as air pollution data at the municipal level, can be processed using standard NSI tools, while others (e.g. card transaction data, social media posts) require big data infrastructure, advanced software, and specialised data science skills. This factor is crucial for assessing the cost-benefit ratio of each source.

Data access: Access is influenced by the data owner's status (public or private), the number of owners, and the degree of harmonisation. Multiple owners or disparate formats (e.g. transport ticketing data produced independently by cities) can create major integration challenges. Access stability — whether one-off or continuous — is another critical consideration. Costs must be distinguished between minimal compensation for data extraction and financial benefits to the operator. Regulation 223 stipulates that access for official statistics should be free of charge, with any compensation limited to processing services; however, these costs can still be significant and must be factored into the scoring.

A short survey was conducted to better understand access conditions, covering both current and prospective availability. Results of the scoring for each data source are presented in Chapter 5, while access conditions are detailed in Chapter 6. Data owners' commercial interests can also impose

restrictions on topics covered or publication timeliness to avoid competition or sensitive issues, and these constraints must be considered.

Data aggregation level: The level of aggregation affects the methods available for combining data. WP3 distinguishes between M-enabler methods, which require access to micro-data (pseudonymised individual-level data with geolocation and timestamps), and M-executor methods, which work with macro-data (aggregated at broader temporal or geographical levels). Given regulatory and confidentiality challenges, micro-data access is often difficult, so most current methods rely on macro-data. Nonetheless, the scoring matrix also accounts for potential M-enabler opportunities.

Metadata: The availability and quality of metadata are essential for data integration. Reference periods must be aligned as closely as possible, particularly since MNO data offer high temporal frequency. For instance, when combining with census data, observation years should match closely to maximise accuracy.

CATEGORY OF ANALYSIS	SCORING DIMENSION (low/moderate/high)
Data Type	Technical cost of handling the dataset
Do you know the size of the data set? Will it be a problem to treat it at once? Will you split it for processing?	
Do you know the structure of the dataset? Are many different files considered a collection? A. Do you have to relate several files to have the entire dataset? B. Are the variables that enable linking of the data already known? If not do you have already a proposal to test the linkability?	
Access	
Who owns the data? Public administration, one company, several companies? **Could the multiplicity of actors lead to multiple data formats and therefore potential integration and harmonisation problems**?	Ease of access to a temporally and geographically harmonized data source
Is it possible to get access with a certain **stability?** Does it have to be paid?	
Are there limitations to the amount **or aggregation level** of data that can be accessed? A. What is the nature of this limitation? Legal, technical, financial, other?	Ease of access to detailed data
Is there a possibility to access the data to study its relevance?	
Are there potentially competing uses or specific restrictions in the application scenarios (operator publishing similar statistics, etc.)?	Range of possible application scenarios
Is this data available in all EU countries?	EU availability of these data
Metadata	
Is the definition of the population accessible? If not do you already have a method to address this issue?	Accuracy and robustness of available metadata
Is the **reference period** of the data available?	
Is the detailed methodology used to build these data available?	
Is the base unit of the dataset accessible?	
Do the units have an identifier?	

Do you have the necessary variables to reach the relevant granularity level for the statistical unit?	
Is there background information that you need to link the base units of the data set to the statistical unit, but that doesn't have the base units of the data set?	
Is there auxiliary information to make the data set useful with auxiliary data (NSI or another source)?	
Does the data contain sensitive variable? (Meaning legal or ethical issues related to its use)	Non-sensitivity of these data

Figure 3: Assessment Matrix - first part

4.3. Secondary scoring dimension

An essential characteristic of ESSnet's MNO-MINDS is its focus on producing official, rather than solely experimental, statistics. The short list of promising data sources will therefore consider (i) their availability across all NSIs and (ii) their alignment with the European Statistics Code of Practice (CoP, 2017). While non-MNO sources already used in official statistics naturally comply with CoP, many new data sources may not. This compliance should not be a prerequisite for a source to be considered promising for integration with MNO data; however, assessing its adequacy against CoP principles provides a valuable estimate of the costs and complexities involved in official statistics production. As such, this will be treated as a *secondary scoring criterion.*

The CoP encompasses 16 principles across three areas: *Institutional environment, Statistical process*, and *Statistical output.* Each principle includes indicators reflecting best practices. Not all are relevant for evaluating non-MNO sources; the selection of applicable principles and indicators is outlined in Figure 4.

Institutional environment: Only principles directly affecting data usability and not already assessed in the primary matrix were retained:

- Professional independence (P1): General in scope; not source-specific and thus excluded.
- Mandate for data collection (P2): Already assessed in the primary matrix.
- Adequacy of resources (Indicator 3.2): Cost–benefit analysis is crucial, particularly regarding financial compensation to data providers.
- Quality monitoring (Indicator 4.3): Essential for sources not designed for official statistics.
- Error notification (Indicator 6.3): Requires data providers to inform NSIs of detected errors.
- Data security (Indicator 5.5): Secure data transmission is mandatory.

Meeting these CoP requirements, particularly for non-traditional sources, often entails significant technical and organizational costs.

Statistical process: The following indicators were selected:

- Conceptual coherence (Indicator 8.1): Ensuring strong conceptual alignment with NSI processes.
- Cooperation with data producers (Indicator 8.7): Builds trust and facilitates methodological transparency.
- Respondent burden (P9): New passive data sources can reduce respondent burden; linking across data sources (Indicator 9.6) is a key consideration.
- Cost effectiveness (P10): Broadly addresses the challenges and costs of new sources. Among its indicators, *standardization (10.4)* is crucial.

Statistical output:

- Relevance (P11): Already included in the primary matrix.
- Accuracy and Reliability (P12): Detailed monitoring of errors is important for selected sources.
- Timeliness and Punctuality (P13): Aligning source frequency with official statistical needs is a scoring criterion.
- Coherence and Comparability (Indicators 14.1 & 14.2): Ensuring regular, consistent, and geographically comparable access is essential, especially for privately held data. Cross-national comparability remains a particular challenge.
- Accessibility (Indicator 15.4): While less critical, granting researchers access to microdata improves transparency and reproducibility. Article 23 of Regulation 223 encourages such access, supporting the public good nature of NSI work.

While the aim of this work package is to provide a list of data sources to be integrated with MNO data categorised by different levels of qualification, it is important to highlight that this list which is resulting from the primary and secondary scoring does not aim to limit the reader's decision on which sources to utilise. The goal is to provide starting points and to outline the possibilities and limitations as well as to show which aspects are crucial to take into consideration.

Therefore, the authors refrain from providing a detailed secondary scoring for each source e.g. by including weights to each indicator but to outline the main cost and gains instead. For this, the short survey on availability of data sources included abovementioned principles to collect information on the current situation for new data sources that national statistical institutes indicated having current access to. Results are presented in chapter 6. Official data sources are expected to comply with quality criteria from the Code of Practice.

Quality requirements (derived from the European Statistics Code of Practice)
PRINCIPLE 3 Adequacy of Resources
Indicator 3.2: Is the scope, detail and cost of this source commensurate with needs?
PRINCIPLE 4 Commitment to Quality
Indicator 4.3: Is it possible to regularly monitor output quality?
PRINCIPLE 5 Statistical Confidentiality and Data Protection
Indicator 5.5: Is it possible to put in place the necessary regulatory, administrative, technical and organisational measures to protect the security and integrity of statistical data and their transmission?
PRINCIPLE 6 Impartiality and Objectivity
Indicator 6.3: Is there a process to inform the NSI in case an error is discovered?
Indicator 6.4: Are information on data sources, methods and procedures publicly available?
PRINCIPLE 8 Appropriate Statistical Procedures
Indicator 8.1: Are the definitions and concepts used in this source a good approximation of the concepts required for statistical purposes?
Indicator 8.7: Do the data holders collaborate with the NSI in improving data quality (consider feedback…)
PRINCIPLE 9 Non-excessive Burden on Respondents
Indicator 9.6: Statistical authorities promote measures that enable the linking of data sources in order to minimise response burden.
PRINCIPLE 10 Cost Effectiveness
Indicator 10.4: Statistical authorities promote, share and implement standardised solutions that increase effectiveness and efficiency
PRINCIPLE 12 Accuracy and Reliability
Indicator 12.1: Are source data, integrated data, intermediate results and statistical outputs regularly assessed and validated?
Indicator 12.2: Are sampling errors and non-sampling errors measured and systematically documented?
Indicator 12.3: Are revisions regularly analysed in order to improve source data, statistical processes and outputs?
PRINCIPLE 13 Timeliness and Punctuality
The periodicity, timeliness and punctuality of data source meets the needs of official statistics production
PRINCIPLE 14 Coherence and Comparability
Indicator 14.1 and 14.2: Are statistics based on this source coherent, consistent and comparable over a reasonable period of time?
Indicator 14.5: Is cross-national comparability of statistics based on this source possible?
PRINCIPLE 15 Accessibility and Clarity
Indicator 15.4: Is access to microdata allowed for research purposes?

Figure 4: Assessment Matrix, second part - Official Statistics

5. Source scoring

This chapter presents the detailed source scoring by first introducing the full list of data sources and their ranking in the three categories *most promising, promising, and less promising* and by then going through the analysis of each source.

5.1. Source scoring in a nutshell

The full list of 18 data sources considered for the landscaping analysis is presented in Table 7. All sources were scored in one of the three categories resulting from primary and secondary scoring criteria. *Most promising* comprises five sources that have proven their relevance and applicability in application scenarios or experimental studies and most are widely available in European NSIs. These sources are official data and are expected to mostly or fully comply with quality criteria from the CoP.

There are seven sources scored as *promising* and many of them are new data sources which are not fully accessible yet or would require some work to be applied in official statistics. However, there are potential applications and clear additional value identified to integrating them with MNO data.

Less promising sources consists of six sources that either require extensive efforts to be applied in official statistics or for which no adequate benefit of integration with MNO data could be identified yet. As a result, costs and gains are currently not reasonable however this may change in the near future.

Most promising
Census
Population register
National travel surveys
Land use and land cover register
Tourism accommodation statistics
Promising
Vehicle, bicycle and pedestrian sensors
Pollution data
Satellite data
Electronic invoices
Tourism surveys
Tourism platform data
Credit card transaction data
Less promising
Google Maps popular time
Vessel (boat) traffic data
Smart meters
Event (ticket) data
Social media
Connected vehicles

Table 2: Full list of data sources for the integration with MNO data

The following analysis builds on the evaluations conducted by participating countries. It is based on data sources to which they currently have access or have examined theoretically, even without direct access. Further, country-specific examples illustrate important aspects however the data analysis and scoring still considers the source in general.

5.2. Most promising sources

The sources identified as the most promising—those demonstrating the best cost–benefit potential for integration with MNO data—are those traditionally used by National Statistical Institutes (NSIs). These sources meet nearly all the requirements of the assessment matrix and align closely with the defined application scenarios, particularly those aimed at improving population coverage.

A summary of the analysis of these sources is presented below:

5.2.1 Census

Data description

A census determines the official population number. Whether it is conducted purely by a survey or register-assisted with supplementary surveys, results are derived eventually from extrapolation (unless it is a full survey). In this analysis, three sources are considered:

- German Census 2022

German Census is conducted every ten years. It is register-assisted and integrates data from administrative registers with supplementary surveys. It covers the following topics: population in brief, households, buildings, population: education and employment, families, dwellings.

- French Census

The population census is an annual survey, exhaustive in municipalities with fewer than 10,000 inhabitants, and covering 40% of dwellings in municipalities with more than 10,000 inhabitants. The population census is used to determine the legal population of France and its administrative districts. National results use 5 annual surveys.

- German Microcensus (SILC, LFS)

The "Mikrozensus" (microcensus) is a yearly survey of about 1% of the population (ca. 810 000 persons). It incorporates mandated EU-Surveys like SILC and LFS and some other national questions.

For the following analysis and scoring of census data as a potential data source to be integrated with MNO data, the evaluated aspects are harmonised as much as possible, however country-specific or source-specific differences are pointed out whenever necessary. The variables for each census are listed in annex 11.1.

For the German register-assisted census 2022, copies of administrative registers served as base data. Additionally, less than 10% of the population were surveyed in order to correct inaccuracies from the registers. With the surveys, further information was collected, e.g. information on education and employment. The information on buildings and housing came from the owners who were questioned by mail. In addition, a further survey is conducted in residential establishments and collective living quarters. For the time between censuses, an intercensal population figure is estimated annually based on the latest census, administrative information and statistics. Conducting a census is complex and costly (in terms of time, resources, etc.).

The German and the French census are determining the official population number. Both are providing additional information on certain variables listed in the appendix and both include information on housing and buildings. The difference to the German microcensus is that microcensus does not aim to produce the official population number but to provide more details on a variety of topics. Microcensus is conducted more frequently which allows for analysis over time. Because it comprises e.g. SILC and LFS questions, it is highly comparable EU-wide. Microcensus is a relevant tool for extrapolating but has more limitations at finer geographical levels. Here, the German census provides numbers also on small geographical levels.

None of the three censuses are full surveys, even if there might be fully exhaustive sub-parts. Therefore, the result is achieved by extrapolating which represents the main difference, e.g. in comparison to a total population register. Census results represent a crucial source for further analysis and provides a data base for science, politics and society.

Challenges associated with these databases and ways of handling them

Statisticians at the NSIs can be expected to be familiar with census data because census are typically conducted by NSIs themselves. Therefore, NSIs have access to micro data, however linkage at finer levels than the published results require a legal check which represents the first challenge. But already with the publicly available results (e.g. 100mx100m grid cells in Germany), there are lots of (potential) applications with MNO data possible. Another challenge is the time aspect: census provides the most accurate population figure but is conducted once in every few years. On the one hand, this is an argument for linking it with MNO data but on the other hand, it can represent a challenge especially with increasing time span between census data and MNO data. As mentioned above, micro census allows for analysis over time and with broader information but lacks geographical granularity and representativeness which census do provide.

Legal challenges can be handled by a legal check and if it does not allow for linkage at finer levels, analysis should focus on publicly available data until the required laws are in place. Challenges in terms of time will remain to some extend and ways to improve can be investigated. On the other hand, cases for which MNO data and census indeed are available for the same point in time (or very close to it), will show valuable insights.

Accuracy for Application Scenarios with MNO data

There have been experimental studies conducted with total population figures and/or further data from Total Population Register and Census combined with MNO data. For many MNO application scenarios, there is some kind of resident or de facto population or socio-demographic information needed, even when there is a different research focus.

The following potential application scenarios have been identified:

*De-jure/resident/night-time population
*De-facto population
*estimating non-registered persons
*improving temporal and spatial granularity of population figures

For most application scenarios and target statistics, census data is used as an auxiliary variable when integrated with MNO data. For fewer cases, there might be scenarios for which census data serves as a proxy (target variable). Census results are complementary to MNO data because they provide a base for the total population size whenever the available MNO data does not comprise data from all MNOs (extrapolation). Census data allow to improve MNO representativeness, especially with night time comparisons (to ease concept reconciliation with residential population). Census data also bring multiple socio-demographic pieces of information.

The value of linking census data with MNO data lies not only in these direct examples above but also that the combined census and MNO data can be further used for many applications with linkage to further data sources.

Data analysis through the assessment matrix

In general, census data sources comply with the requirement of the assessment matrix. For German census, data is available in the database and can be accessed there (https://ergebnisse.zensus2022.de/datenbank/online). Access is possible also via API and software packages, e.g. {restatis}. Linkage with MNO data can be at the geographical level, e.g. usually the 1x1km INSPIRE grid is suitable.
For the French census, data structure geolocation variables are known. Sub-municipal geolocation is only available for the 2017 census results. Multiplicity of actors is not expected to result in multiple data formats, especially if linkage is realised at the geolocated level (e.g. 1x1km grid).

There are no legal or ethical problems to access the data, yet for micro census, linkage with other data sources requires a legal check. Data can be accessed at a detailed level, limitations are on the individual data level and on very fine geographical levels for confidentiality reasons. Not every EU country conducts a census. The remaining countries will have an alternative data source with comparable figures (e.g. Total Population Register). Microcensus results are available for all EU countries that conduct a micro census. EU-wide surveys (Silc, LFS, ICT) are available for all EU-Countries.

Data Scoring for the combination with MNO data

The costs to treat total population numbers and information from Censuses are relatively moderate in the sense that these sources are anyway produced by NSIs and statistical officers are used to work with this data source. The infrastructure, quality standards, and knowledge about the data source is already in place. Depending on the desired use cases / target statistics, there may be legal checks required to clarify which linkage on micro data is or is not allowed. For published results there should not be issues, although one should be cautious when combining multiple data sources.

Regarding the scoring, EU availability has been scored “high” even though some countries might not conduct a census, however these countries will have comparable information, e.g. in the form of a total population register. However, as register and census differ (as explained above) they are not analysed in the same chapter. For most aspects, scoring was considered for publicly available census results.

CENSUS			
Ease of handling the dataset			**high**
Ease of access to a temporally and geographically harmonized data source.			**high**
Ease of access to detailed data			**high**
Range of possible application scenarios			**high**
EU availability of these data			**high**
Accuracy and robustness of the information available on these data			**high**
Non-sensitivity of these data		**moderate**	

5.2.2 Total Population Register

Data description

Swedish Total Population Register

The Swedish Total Population Register is a cornerstone of population data in Sweden. It serves as the foundation for official population statistics and provides critical data for many of Statistics Sweden's activities. The register includes information on the population and its changes, closely mirroring the population register maintained by the Swedish Tax Agency. A list of variables can be found in annex 7.1.

Challenges associated with this database and ways of handling them

Key Challenge: Under-Coverage of Non-Registered Individuals

A significant limitation of the Total Population Register (TPR) is its under-coverage of individuals not officially registered in Sweden. Despite this, the TPR remains the primary source for official statistics on population and households. These statistics cover population distributions by sex, age, marital status, and geographic areas (e.g., counties and municipalities). They also include vital events such as internal migration, births, deaths, marriages, divorces, immigration, and emigration. These figures are updated multiple times a year.

Role in Sampling and Surveys

The TPR is extensively used as a sampling frame for various surveys conducted by Statistics Sweden. These surveys, commissioned for different purposes, rely on TPR data to design samples for questionnaire and interview-based studies, such as:

- **Labour Force Surveys**
- **Political Party Preference Surveys**
- **Surveys on Living Conditions**

Supplementary Data for Registers and Surveys

The TPR is also a valuable supplementary data source for other registers and surveys, reducing the number of questions posed to respondents and easing respondent burden. For instance, the TPR provides:

- Background data for registers related to labour market statistics, economic welfare statistics, and education statistics, ensuring comprehensive socioeconomic coverage.
- Coordination for register populations and surveys involving individual-level statistics.
- Data essential for statistical packages and population projections.
- Updated data for ongoing questionnaire and interview surveys.
- Personal identity number supplementation for various datasets.
- Authentication support for register extracts, adhering to the Personal Data Act.

Metadata are available here: https://metadata.scb.se/mikrodataregister.aspx?produkt=BE0102

The application domains and quality issues associated with this data source are broadly comparable to those identified for the census. The EU-wide availability survey, further elaborated in Chapter 6, indicates a high level of availability of this type of data source across EU Member States. For additional details, please refer to Chapter 6.

Data Scoring for the combination with MNO data

The costs to treat total population numbers and information from Population Register are similar to those exposed for Censuses. The main difference lies in the EU availability and non-sensitivity because TPR allows linkage on finer levels than (publicly available) census results.

Total Population Register			
Ease of handling the dataset			**high**
Ease of access to a temporally and geographically harmonized data source.			**high**
Ease of access to detailed data			**high**
Range of possible application scenarios			**high**
EU availability of these data			**high**
Accuracy and robustness of the information available on these data			**high**
Non-sensitivity of these data	**low**		

5.2.3 National Travel Surveys

Data description

The National Travel Survey (RVU Sweden)

Transport Analysis, a government-appointed authority in Sweden, oversees official statistics on transport and communications. Since 2019, the **National Travel Survey (RVU Sweden)** has been conducted annually, covering Sweden's population aged 6–84. The survey employs a combined paper and online format, with a stratified sample by county, age group, and gender. For 2023, approximately 12,200 individuals were surveyed. Organizations and regions can purchase additional samples to focus on specific interest groups.

Purpose and Usage

The survey provides critical insights for:

- National and regional transport policy development.
- Infrastructure and transport planning.
- Road safety initiatives.
- Research on travel and communication patterns.

Variables Collected

The RVU database captures detailed information across several domains:

Overview of variables in the Travel Survey database:

Area	Description
Individual and Household	Gender, age, driver's license status, access to bicycles, disabilities, household composition, etc.
Cars and Parking	Household car ownership, type, usage, and parking facilities.
Travel Cards	Public transport season tickets.
Method	Response method (paper or web).
Daily Movements	Mode of travel, route, purpose, addresses (start, end, target), travel times, and durations.

Table 3: National travel survey - variables

The survey's published statistics are broken down by sex, age, municipality, mode of transport, trip purpose, and travel distance.

Breakdown of Modes of Transport:

- On foot
- Bicycle
- Car
- Public transport (bus, train, tram, metro)
- Other (taxi, air, sea, transport services)

Trip Purposes:

- Work, business, school
- Services and shopping
- Leisure
- Other

Challenges associated with this database and ways of handling them

Survey Limitations

The RVU faces several challenges:

- **Non-Response and Coverage Issues:** The overall response rate is approximately 17%, with lower participation among men and younger age groups (15–44 years).
- **Sample Size Constraints:** Limited sample size restricts the ability to analyze small geographical areas, less common travel modes, or infrequent trip purposes.

Addressing Challenges

Transport Analysis acknowledges these limitations and actively engages with stakeholders to address gaps. Customized data extracts and expanded sampling are available to enhance usability for specific needs. Efforts to increase response rates and improve survey relevance are ongoing.

Integration with Mobile Network Operator (MNO) Data

Combining survey data with MNO data enhances population representativeness and fills critical gaps:

- **Complementary Strengths:** The RVU provides demographic details (e.g., age, gender), travel purposes, and modes of transport, which are difficult to infer from MNO data.
- **Improved Precision:** MNO data offers granular insights into travel flows, time periods (<1 year), and emerging patterns in new residential or business areas.
- **Enhanced Relevance:** Integration can provide more comprehensive, timely, and geographically detailed mobility statistics, especially during dynamic events like pandemics.

Potential for Improvement

Transport Analysis foresees no significant conflicts between the two data sources and considers integration an opportunity for enhancement. Adjustments to survey questionnaires to align with MNO data could improve both sources. Enhanced statistics, enabled by MNO data, include:

- Total travel volumes.
- Detailed mobility patterns.
- Finer geographical granularity.
- Faster updates to reflect changing travel behaviors.

Leveraging the combined strengths of RVU and MNO data, Transport Analysis can deliver more accurate, relevant, and timely insights, effectively meeting the evolving needs of stakeholders while mitigating the limitations inherent in traditional survey methodologies.

Data analysis through the assessment matrix

The data is stored in an accessible database, enabling straightforward queries, processing, and summarization. Users can easily download datasets in simple Excel format directly from the agency's website. The data structure is intuitive, and variables required for linking datasets are clearly defined.

Ownership of the data resides with **Transport Analysis**, and it is classified as a "public good," ensuring free access without competitive use restrictions or specific legal issues.

Availability Across EU Countries
The EU-wide availability survey for this data source indicates a moderate level of coverage. For further details, please refer to Chapter 6.

Target Population
The target population for the survey is clearly defined: all individuals registered in Sweden aged 6–84 years. The survey employs a stratified sampling approach based on data from the **Total Population Register (TPR)** maintained by Statistics Sweden. The sample is stratified by:

- **County**
- **Age group**
- **Gender**

For 2023, the sample includes approximately 12,200 respondents. Each participant answers questions related to a specific measurement day, collectively covering the entire survey reference period. The dataset is based on an anonymized unique personal identifier for each respondent.

Survey Methodology
Since 2019, the survey has been conducted using a combined paper and online format, allowing comparability across the 2019–2023 period. However, surveys conducted before 2019 relied on telephone interviews. Due to this methodological change, comparisons between surveys conducted before and after 2019 should be approached with caution, as differences may partly reflect variations in methodology rather than genuine changes in travel behavior.

Comparability and Categorization
The survey data supports robust comparability across groups. For example:

- **Municipal Group Division:** Data is often reported by functional regions, a commonly used classification in Sweden.
- **Categorization Options:** The travel habits database enables detailed categorizations by age, mode of transport, and trip purpose (errand).
-

These categorizations facilitate nuanced analysis and ensure the data can be used effectively to support research, policy development, and planning initiatives.

Accuracy for Application Scenarios with MNO data

Illustration of a concrete application scenario with National travel survey and MNO data: Collaboration between Statistics Sweden and the Southeast Region

Background

Sweden's decentralized governance model grants Regional Authorities significant mandates. Within this framework, three large regions and 13 municipalities — collectively known as the *Southeast Triangle* — have launched a joint project to strengthen cross-border collaboration and stimulate economic growth.

Identified Data Gap

During the initiative, the regions discovered a *critical lack of data on cross-county travel.* Existing sources, such as the Swedish Travel Habits Survey and registers on housing, education, and workplace statistics, were insufficient for understanding regional mobility. Specifically, they lacked detail on:

- Leisure trips and visitor flows
- Inter-regional work commutes
- Population presence dynamics (e.g., individuals remaining in one location despite workplaces elsewhere)

Unlike most other data sources used by state, municipalities, and regions, **mobility data transcend administrative boundaries**. This makes them uniquely capable of depicting functional geographies — patterns of movement and interaction that extend across urban borders, municipal lines, regional divisions, NUTS areas, and other government-defined boundaries. It is precisely these administrative divisions that have caused complications in the Southeast.

MNO Data as a solution

To address this gap, the regions identified mobile network operator (MNO) data as a suitable complementary source and approached Statistics Sweden (SCB) to establish a partnership. The collaboration aims to provide access to MNO data for regional mobility analyses. The project is scheduled to be formally launched on 1 January 2026.

Project Goals

The initiative will focus on answering five key questions:

1. **Urban flow analysis**: What do travel flows within urban areas look like?
2. **Regional exchange**: How does movement occur between regional locations?
3. **Temporal patterns**: How does travel vary by time of day, weekday/weekend, and holidays (e.g., leisure trips)?
4. **Population monitoring**: How many people are present in a town at different times (day, evening, night, weekend) to inform municipal service planning?
5. **Seasonal variations**: How does the day/night population fluctuate across the year?

Scenario planning and route modeling

Beyond descriptive analysis, the project will also explore *scenario modeling* using route selection engines such as *OpenTripPlanner*. Different travel scenarios will be tested, including:

- **Baseline travel**: Patterns predicted by the Travel Habits Survey.
- **Preferred travel patterns** (prioritizing sustainability):
    1. Walking
    2. Cycling

3. Public transport
4. Car use

Broader significance

The current initiative demonstrates how regional authorities can collaborate with national agencies to leverage innovative data sources for improved decision-making, planning, and policy development. It highlights the potential of MNO data to overcome traditional administrative barriers and support evidence-based regional strategies.

Data Scoring for the combination with MNO data

National Travel Surveys			
Ease of handling the dataset			**high**
Ease of access to a temporally and geographically harmonized data source.		**moderate**	
Ease of access to detailed data			**high**
Range of possible application scenarios			**high**
EU availability of these data		**moderate**	
Accuracy and robustness of the information available on these data			**high**
Non-sensitivity of these data			**high**

5.2.4 Land Use and Land Cover register

Data description

Land use and land cover (LULC) registers provide systematic information on Earth's surface coverage and purpose, such as urban areas, agricultural land, and forests. For example, urban areas might include residential zones, commercial districts, and industrial parks, while agricultural land could encompass crop fields, orchards, and pastures. Forests might be categorized into deciduous, coniferous, or mixed types. These registers, like the CORINE LULC database, are standardized and widely used across EU countries. They are usually stored in GIS databases, which typically consist of multiple tables, the design of such databases also enables linking. For instance, one table might contain land cover classifications (e.g., forest, water bodies), while another table could include land use details (e.g., recreational areas, transportation networks). LULC data are owned by national institutions, including NSIs, and are freely available for state use, though not publicly accessible. The data are highly standardized, with clear definitions, base units, and identifiers, making them suitable for statistical purposes, fulfilling all the necessary standards.

Challenges associated with these databases and ways of handling them

When choosing to work with LULC registers, there always is the risk of outdated information, which can potentially affect the accuracy of analyses. Regular/constant updates are required to maintain the relevance of the data. Another aspect rising concerns is that accessing comprehensive LULC data for a country may require partnerships and agreements between multiple data holders at a national level, as database systems and designs can vary. Lastly, statistical disclosure control needs to be conducted to address potential disclosure risks, since registers could contain sensitive information (e.g., household locations).

To address these challenges, several treatments can be implemented. First, establishing a standardized update schedule for LULC registers, coordinated at the national or EU level, can mitigate the risk of outdated information. This could involve setting up automated systems for data collection and validation, ensuring consistent updates. Second, encouraging collaboration between national institutions and data holders through centralized agreements or frameworks can streamline access to comprehensive LULC data, reducing administrative hurdles. For instance, creating a unified data-sharing platform or adopting common database standards across countries could simultaneously cut the time spent on pre-processing data and improve the flow of information among NSIs and other LULC providers. Finally, for the management of disclosure risks, a great amount of care should be put into properly applying statistical disclosure control methods, such as data anonymization, aggregation, or perturbation techniques, in order to safeguard sensitive information while also maintaining the utility of the data for its intended statistical purposes.

Accuracy for Application Scenarios with MNO data

LULC registers complement MNO data by enhancing the geographical precision of analyses, two potential use cases being accurately pinpointing home locations to produce mobility indicators or assessing urban green space usage. By integrating LULC registers with MNO data, the scope of official statistics can be expanded, enabling new and more detailed analyses that leverage the strengths of both data sources.

Among the potential use cases, there is the identification of home locations for mobility indicators. LULC data can refine the process due to its capacity to distinguish residential areas from other land uses, such as commercial or industrial zones. This can prove itself valuable for creating accurate mobility indicators (e.g., daily commuting patterns, population displacement during events like natural disasters). However, outdated LULC can give rise to several issues. For instance, if a new housing development is not yet reflected in the LULC register, MNO data might incorrectly associate mobile activity in that area with a nearby commercial zone, skewing the accuracy of the analyses.

Another application centers on evaluating urban green space usage. LULC data is used to define spatial boundaries and categorizes green areas, while MNO data can track the number of people visiting those areas, providing the necessary details for a proper analysis of how these spaces are used. Although, it is necessary to mention that if LULC datasets are not updated frequently, shifts in

land use (e.g., a park transformed into housing) might go undetected, leading to unreliable insights about actual green space demand and accessibility.

In disaster response scenarios, LULC data can help pinpoint impacted regions, such as residential zones affected by flooding, while MNO data would monitor population displacement. Together, these datasets offer real-time visibility into evacuation routes and the return of residents to disaster-stricken areas. Nevertheless, outdated LULC records may overlook critical post-disaster changes, such as temporary relief camps or unsafe zones, compromising response efforts. Similarly, transportation planning benefits from pairing LULC classifications (urban, rural, industrial) with MNO-derived mobility patterns. Identifying high-traffic corridors, for instance, could guide decisions on new transit routes. However, delayed updates to LULC data risk misclassifying regions undergoing rapid development. A rural area transformed into a bustling suburb might retain its original classification, resulting in infrastructure plans that fail to address actual traffic demands.

Data analysis through the assessment matrix

The LULC data is State owned and freely (no financial cost) available, but not publicly available.

To our current knowledge no institution or private entity provides statistics recurrent and timely statistics based on fusion between MNO and land use and land cover data. Even if we run on the assumption that some statistics are produced by combining these two types of datasets, we can safely consider that the cost of matching the quality criteria used in official statistics are too restrictive or impose a set of constraints which makes it prohibitive (in terms of scope and trusted statistics) to compete with products delivered by official statistics.

Because the dataset scales with spatio-temporal coverage, processing it in one go might not always be feasible, meaning it could need to be broken down into smaller subsets. Identifiers are included and Microdata access is available for research under certain conditions.

Data Scoring for the Combination with MNO Data

Stored in standardized GIS databases equipped with clear linking mechanisms, LULC data are generally straightforward to manage for statistical work. Though harmonized across EU countries, obtaining access often involves navigating partnerships and agreements between multiple data holders, a process that can create delays or administrative hurdles. While detailed LULC datasets exist, they're typically accessible only to authorized bodies like NSIs, not the broader public. The data's versatility supports diverse applications, from urban planning and environmental monitoring to refining analyses of MNO data. Mandated for use in all EU member states, LULC registers guarantee broad availability, bolstered by their high standardization and frequent updates, which ensure accuracy and reliability for statistical purposes. That being said, while most LULC information is non-sensitive, certain registers might include details like household locations, requiring statistical disclosure controls to mitigate privacy risks.

Land Use and Land Cover register			
Ease of handling the dataset			**high**
Ease of access to a temporally and geographically harmonized data source.		**moderate**	
Ease of access to detailed data			**high**
Range of possible use cases			**high**
EU availability of these data			**high**
Accuracy and robustness of the information available on these data			**high**
Non-sensitivity of these data		**moderate**	

5.2.5 Tourism Accommodation statistics

Data description

Monthly Tourism Statistics comprise statistics on capacity and occupancy of tourist accommodation. In most ESS member countries, these numbers are collected via surveys that are filled in by accommodation establishments and then reported to the NSI. The resulting data type of the underlying dataset is consequently “register and administrative data”.

In Germany, all accommodation establishments with more than ten beds or camping pitches and belonging to NACE 55.1, 55.2, 55.3, training homes, and preventive care and rehabilitation properties are obliged to report the following characteristics on a monthly basis:

*arrivals

*nights spent

*length of stay

*number of accommodating companies

*number of beds/sleeping accommodation

*(guests') country of residence

*mean of accommodation (hotel, holiday apartment, hostel, ...)

Some information is reported annually, e.g. the business' total number of beds/sleeping accommodation.

Challenges associated with this database and ways of handling them

Data does not cover accommodation establishments with less than ten beds or camping pitches and is provided relatively timely but with MNO data, timeliness and granularity could be improved. For the integration, it might be useful to include additional data sources to either improve precision or extent the scope of flash estimates, however this may result in challenges e.g. in case data sources have different data formats.

Accuracy for Application Scenarios with MNO data

The combination of data from tourism accommodation statistics and MNO data can potentially enable flash estimates for tourism statistics as outlined in the “nights spent” application scenario in WP3 deliverable 3.2. Potential applications for tourism statistics may also include other variables that are listed above.

Data analysis through the assessment matrix

Aggregated data is available at databases provided by NSIs and by Eurostat. Access is free of charge. Data is available at NUTS 0-3 and possibly at finer geographical units. Additionally, data is available for pre-defined "holiday regions" (total population can vary with time). Data is reported per month and per establishment. Therefore, NSIs hold data on a smaller geographical unit as the published data.

Data Scoring for the Combination with MNO Data

Tourism Accommodation Statistics			
Ease of handling the dataset			**high**
Ease of access to a temporally and geographically harmonized data source.			**high**
Ease of access to detailed data			**high**
Range of possible use cases		**moderate**	
EU availability of these data			**high**
Accuracy and robustness of the information available on these data			**high**
Non-sensitivity of these data		**moderate**	

5.3. Promising sources

Promising sources are defined as sources which would require additional efforts, or which are not yet fully accessible. But there are clear potential applications and additional value for official statistics identified.

5.3.1 Vehicle, bicycle and pedestrian sensor data

Data description

Sensor-based data offers valuable insights, particularly in the realms of transportation and environmental management. In Sweden, these data sources increasingly inform analyses and policy decisions. However, their utility is constrained by two significant limitations:

1. **Geographic Coverage**: Sensors are unevenly distributed, with dense coverage in some urban areas (e.g., Stockholm and Gothenburg) and negligible presence elsewhere.
2. **Data Fragmentation**: Sensor networks vary widely in purpose and design, resulting in inconsistent datasets.

Two predominant sources in Sweden—likely representative of trends in other EU nations—are:

- **Vehicle sensors** used for congestion tax systems in Stockholm and Gothenburg.
- **Bicycle and pedestrian sensors**, primarily implemented in Stockholm for environmental and urban planning purposes.

This document focuses on these two sources while exploring potential synergies with Mobile Network Operator (MNO) data to enhance statistical analyses.

Vehicle, bicycle and pedestrian sensors in the city of Stockholm

Vehicle sensor locations for congestion tax – Stockholm and Gothenburg

Description: Sweden operates a congestion tax system in Stockholm and Gothenburg to reduce traffic congestion. This tax applies to both Swedish and foreign-registered vehicles. The system, managed by the Swedish Transport Agency, provides data on vehicle passages recorded at toll stations during weekdays from 06:00 to 18:29.

Key insights available from this data include:

- Monthly time series of vehicle passages.
- Passage counts at specific toll stations, including untaxed proportions.

The map below illustrates the toll stations within Stockholm municipality, where vehicle passages are automatically recorded during taxable hours.

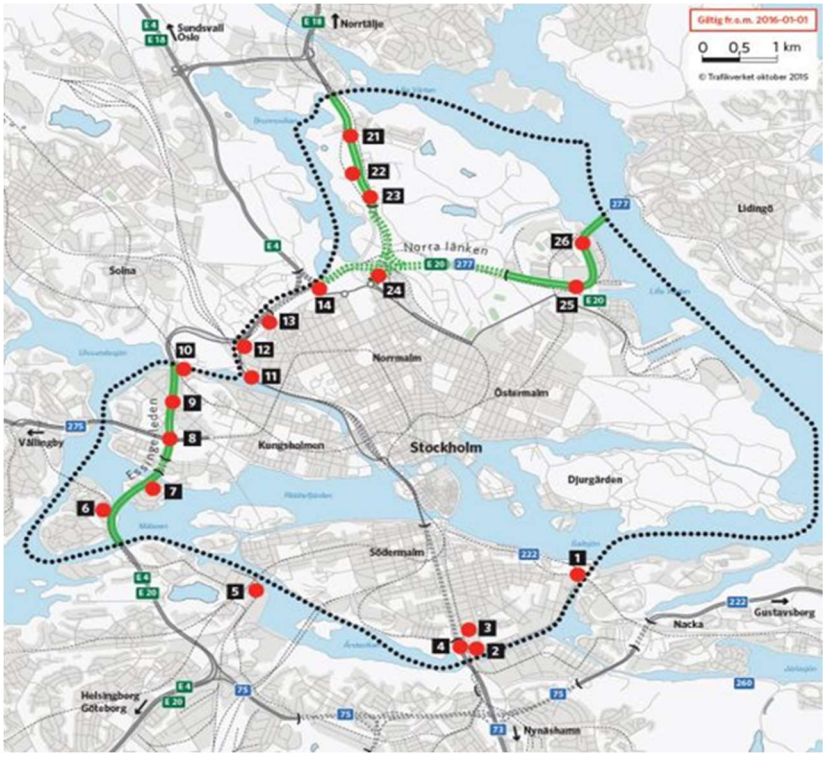


Figure 5: Toll stations within Stockholm municipality

(Source: https://www.transportstyrelsen.se/globalassets/global/bilder/vag/trangselskatt/karta_trangselskatt_sthlm.pdf)

Challenges associated with these databases and ways of handling them

1. **Limited Geographic Coverage**: Sensors are concentrated in the inner cities of Stockholm and Gothenburg, with no data available for other regions.
2. **Mismatch with Target Measures**: While vehicle sensors measure passages, they do not directly capture person-crossings, which limits their application in broader mobility statistics.

Combining sensor data with MNO data addresses some limitations. MNO data can provide estimates of person-crossings by tracking SIM movements, although it cannot fully capture all persons or crossings. Sensor data serves as a proxy for calibration, enhancing the reliability of combined datasets.

For instance, vehicle sensors measure crossings by vehicle type at specific locations and times, which can complement MNO data to approximate person-crossings. Such integration creates valuable datasets for public use and offers mutual benefits to MNOs.

Bicycle and pedestrians' sensors Stockholm inner-city and city centre.

Description: Since 2015, Stockholm has deployed pedestrian and bicycle sensors to support its environmental goals and improve citizens' well-being. These sensors include:

- **Automatic Pedestrian Counters**: Seven fixed stations continuously monitor pedestrian traffic, supplemented by over 200 additional locations for periodic counts.
- **Bicycle Sensors**: Installed across the inner city and city center, these sensors track cycling trends.

Key insights include:

- **Walking Trends**:
  - Median walking trip length: ~1.6 km.
  - Proportion of trips made on foot: 37% (citywide) and 52% (inner city).
- **Cycling Trends**:
  - Significant increases in cycling attributed to health and environmental awareness, alongside infrastructure investments like bike lanes and parking spaces.

In 2017, the city also measured gender distribution among pedestrians, revealing an equal split between men (49.5%) and women (50.5%).

Challenges associated with these databases and ways of handling them

1. **Geographic Constraints**: Like vehicle sensors, bicycle and pedestrian sensors are clustered in select urban areas, limiting their representativeness.
2. **Coverage Gaps**: These sensors cannot capture comprehensive movement patterns, particularly in suburban or rural areas.

Sensor data enriches MNO datasets, particularly for calibration purposes. For example, pedestrian and bicycle sensors offer detailed counts for specific locations, complementing MNO estimates of broader movement patterns. This integration enhances the usability of both datasets for urban planning and environmental analyses.

Accuracy for Application Scenarios with MNO data

Both data sources are publicly accessible:

- Vehicle sensor data is managed by the Swedish Transport Agency.
- Bicycle and pedestrian sensor data is managed by the City of Stockholm.

While the target populations are partially known, limited coverage and other issues hinder the production of official statistics based solely on these sources. Instead, they hold substantial value as complementary datasets to MNO data, enabling more comprehensive analyses.

Data Scoring for combination with MNO data

The integration of vehicle, bicycle, and pedestrian sensor data with MNO data presents significant potential:

- **Enhanced Calibration**: Combined datasets improve accuracy in mobility statistics and provide actionable insights for policymakers.
- **Strategic Complementarity**: While standalone sensor data may be insufficient, their combination with MNO data enriches both datasets, offering mutual benefits.

Vehicle, bicycle and pedestrian sensor data			
Ease of handling the dataset			**high**
Ease of access to a temporally and geographically harmonized data source.	**low**		
Ease of access to detailed data			**high**
Range of possible application scenarios		**moderate**	
EU availability of these data	**low**		
Accuracy and robustness of the information available on these data			**high**
Non-sensitivity of these data			**high**

5.3.2 Pollution statistics

Data description

Currently, Sweden lacks a systematic, centralized data source for producing pollution statistics. Instead, various initiatives and research studies are conducted on an ad hoc basis. One notable study is **"High-resolution air quality modelling of NO_2, PM_{10}, and $PM_{2.5}$ for Sweden"**, which provides a comprehensive national assessment for 2019 using advanced dispersion modelling.

Key Highlights from the Study

- **Scope and Resolution**: The study calculated pollutant concentrations (NO_2, PM_{10}, and $PM_{2.5}$) across Sweden for 2019 using a seamless dispersion modelling methodology. This methodology integrates regional, urban, and street-scale models while avoiding double-counting emissions. The results were delivered at a high spatial resolution of **50x50 meters**, capturing critical concentration gradients for accurate exposure calculations.
- **Methodological Advances**:
  - New parameterizations and inputs were developed to reflect real-world dispersion conditions and physical environments.
  - The modelling linked directly to emission inventories and projections, enhancing its predictive capabilities.

- Storage optimization techniques ensured efficient handling of high-resolution data.

- **Findings**:
  - **NO_2**: High levels were observed in urban areas near roads with heavy traffic. Exceedances of air quality limits were identified in several locations.
  - **PM_{10}**: Exceedances of current air quality standards were relatively rare. However, future stricter standards could pose challenges.
  - **$PM_{2.5}$**: Levels were generally low, with no exceedances of current standards.
  - Validation results showed that quality objectives were met for $PM_{2.5}$ but not consistently for NO_2 and PM_{10}. Further investigation is needed at specific traffic-heavy sites where underperformance was noted.

Challenges associated with this database and ways of handling them

The study highlights several challenges associated with high-resolution air quality modelling:

1. **Model Performance**: Improving the accuracy of predictions for NO_2 and PM_{10} at traffic-heavy locations requires further investigation and prioritization.
2. **Memory and Storage**: Efficient handling of data at a national scale remains a critical obstacle, necessitating continued focus on storage optimization and computational capacity.
3. **National Application**: The results offer a comprehensive national overview of air quality, covering all Swedish municipalities. This dataset supports:
   - Identifying areas at risk of exceeding air quality thresholds.
   - Informing municipalities lacking their own air pollution measurement or modelling capabilities.
   - Supporting compliance with the updated EU Ambient Air Quality Directive, which introduces stricter air quality standards.

The data is freely available to municipalities, researchers, and the public via the **SMHI web portal "Luftwebb"** via this link: https://smhi.diva-portal.org/smash/record.jsf?pid=diva2%3A1939031&dswid=6707

Accuracy for Application Scenarios with MNO data

MNO (Mobile Network Operator) data presents a valuable opportunity to enhance the utility of pollution statistics by estimating **daytime population distributions**. This is critical for analyzing public health issues related to air pollution exposure:

- Traditional exposure calculations are based on the resident population, ignoring the significant daytime mobility of individuals.
- MNO data allows for dynamic modelling of exposure, capturing population shifts throughout the day and providing a more accurate assessment of health impacts from pollution.

Data Scoring for combination with MNO data

The integration of high-resolution pollution modelling with MNO data offers a powerful tool for addressing public health challenges. Combining pollutant concentration data with dynamic population estimates, policymakers and researchers can better understand and mitigate the risks associated with air pollution, ensuring a healthier environment for Sweden's population.

Air pollution data			
Ease of handling the dataset			**high**
Ease of access to a temporally and geographically harmonized data source.			**high**
Ease of access to detailed data			**high**
Range of possible application scenarios		**moderate**	
EU availability of this data			**high**
Accuracy and robustness of the available information (on this data)			**high**
Non-sensitivity of this data			**high**

5.3.3 Electronic invoices

Data description

Electronic invoices (e-Fatura) that can be defined as a mandatory reporting invoices system implemented by the Tax Administration as part of the administrative simplification and anti-fraud measures. Electronic transmission of the invoices issued by individuals or legal entities that have their head office or permanent establishment in Portuguese territory to the Tax and Customs Authority is mandatory. This administrative data includes all the invoicing recorded electronically by the issuer, whether the acquirer / buyer has requested an invoice from.

The detailed variables are described in annex 11.1.

Challenges associated with this database and ways of handling them

The two main challenges are to ensure regularity in the monthly transmission by the data supplier and to handle the high volume of data (100 million records monthly).

Some treatments are applied to these big data bases:

- Identification and imputation of extreme outliers: for positive taxable values equal to or greater than 100 million euros and more than 3 standard deviations from their respective mean; for negative taxable values, the values of the highest magnitude are briefly analysed.
- Correction of negative taxable values (less than -100,000 euros), in cases where it is possible to identify a similar symmetrical value (between 95 and 100%), up to 4 months prior. These negative taxable values mainly result from corrections to values transmitted incorrectly in previous months. The total taxable value for the set of records involved remains unchanged, with only the temporal distribution being altered.

- Identification and imputation of missing values in a small subset of more significant companies (in terms of number of employees and turnover). The identification of missing values and their respective imputation is carried out based on the company's behavior over time (historical series).

Accuracy for Application Scenarios with MNO data

For all of the following analyses, incorporating real-time and historical phone presence data will improve the geographic dimension of the studies. The integration with MNO data could also provide, particularly for hotels and restaurants, indications of an average transactional volume per visitor. The analyses of specific economic sectors like pharmacies and supermarkets together with the number of phones present per area can be invaluable for crisis management and government decision making.

- Descriptive Statistics: Summarizing the data to provide insights into average monthly transactions, standard deviations, and seasonal variation.
- Predictive Modeling: Building predictive models to forecast future transactions, taking into account identified trends and seasonal patterns.
- Tourism Analysis:
    - Visitor Trends: Identifying peak tourism months and correlating transactional data with tourism activities.
    - Economic Impact: Estimating the economic impact of seasonal tourism on local businesses by analysing transaction volumes.
- Comparative Analysis
    - Year-over-Year Comparison: Comparing transactional data from the same month across different years to detect growth or decline and to adjust for annual seasonality.
    - Month-over-Month Comparison: Evaluating changes in transactions from one month to the next to spot short-term trends and seasonal variations.

The most compelling statistical product under consideration involves indicators of business recovery in specific geographical areas following major disruptions, such as the COVID-19 pandemic or droughts. These events can significantly alter habits, devastate local establishments, or halt economic activity, leaving some areas unable to regain their former vibrancy.

Such indicators can be developed by combining mobile network operator (MNO) data with transactional data, such as e-fatura records. For instance, aggregated and anonymized MNO data can provide insights into population movements, foot traffic, and density patterns, helping to detect whether people return to an area after it has been rebuilt.

However, even if people return, their consumption habits may have shifted. Therefore, transactional data is essential for detecting changes in consumer spending and variations in average transaction values. It is important to note, however, that transactional data often lacks granularity at the establishment level and by itself would also not be able to provide economic recovery indicators in specific smaller areas. Transactional data alone can indicate whether an enterprise is recovering as a whole and reveal trends for specific business types (e.g., retail, dining,

entertainment) but provides limited insight into the geographic aspect of the recovery. This highlights the importance of integrating MNO data with transactional data to bridge this gap.

However, the statistical accuracy of this approach faces significant challenges:

- **Data Granularity Issues:** Transactional data is typically aggregated at the enterprise level rather than the establishment level. This limits its ability to provide geographically precise insights, particularly when enterprises operate multiple establishments in different locations. Consequently, transactional data may reflect overall enterprise recovery but fail to capture regional nuances or disparities.
- **Representation Bias:** MNO data, while useful for analyzing mobility patterns, is only as representative as the subset of the population using specific mobile operators. Differences in market share between operators may introduce bias, potentially skewing recovery indicators for certain areas.
- **Data Integration Complexity:** Integrating MNO data with transactional data is non-trivial. Both datasets operate on different scales and levels of aggregation, which may complicate accurate alignment and analysis. Errors in matching data points could lead to misleading conclusions.
- **External Confounding Factors:** Factors such as weather, public events, or policy changes can significantly influence both mobility and transactional patterns, potentially obscuring the direct relationship between the two. These external factors must be accounted for to ensure accurate recovery assessments.

In cases where enterprises have numerous establishments or where data is sparse, clustering techniques can be employed to identify regions or business types with similar recovery profiles. However, even this approach is subject to statistical limitations, particularly when dealing with incomplete or highly aggregated data.

Finally, the development of predictive models to forecast recovery patterns using MNO and transactional data must be grounded in a careful assessment of potential errors and assumptions. Even when combining these data sources with contextual predictors, weather, public events, or policy shifts, important limitations may remain. To align with the total error framework, it is essential to account for potential biases by incorporating statistical adjustments (e.g., correcting for market share differences among operators), validating against independent data sources (local business surveys), and conducting sensitivity analyses with respect to data aggregation levels and missing or biased inputs. These steps, reflected in the final stages of the roadmap (see WP3 deliverable 3.4), are crucial to enhance the robustness and interpretability of any resulting indicators.

Data analysis through the assessment matrix

Data is owned by the public administration (tax authority). It is possible to get access by the NSI and, to some extent, by the Scientific Community. It is free of charge, under a protocol. Outside the NSI it is only possible to access data aggregated by Issuers (sellers).

These data are currently only available in a limited number of countries, which significantly reduces the relevance of the use case at this stage. However, this could change in the future if wider availability is achieved.

As regards metadata, necessary variables to reach the relevant granularity level for the statistical unit are available at least for a large part of the population of acquirers. In other words, there is a significant number of individual issuers without an identifier. The dataset is directly usable and does not require additional data to be linked with. There are some quality issues, mainly related with outliers or missing values. Since personal identifiers are already encrypted, the data doesn't contain sensitive variables.

Information on the data source is publicly available in Portugal since it is administrative data held by the administration. Yet the fact that information is provided by companies rather than by individual establishments, leads to inaccuracies in the geographical variables.

Data Scoring for combination with MNO data

Regarding this analysis, these data seem sufficiently promising to be worth the cost of treatment for combining with MNO data. Yet the amount of remaining work for their integration make them a secondary priority.

Additionally, the country reporting this data source (PT) does not have access to MNO data, rendering the exploration of a practical application scenario unfeasible.

Electronic invoices			
Ease of handling the dataset		**moderate**	
Ease of access to a temporally and geographically harmonized data source.			**high**
Ease of access to detailed data			**high**
Range of possible application scenarios			**high**
EU availability of this data	**low**		
Accuracy and robustness of the available information on this data		**moderate**	
Non-sensitivity of this data			**high**

5.3.4 Tourism Surveys

Data description

NSIs widely capture statistics on tourism through surveys and therefore many NSIs hold results / data on tourism surveys. Some variables and the way and scope of the survey might differ by country but results should be comparable. In this analysis, the focus is on Tourism household survey in Germany and tourism border survey in Spain. In general, a linkage of these sources with MNO data (and further data sources) provides complementary information to gain a deeper understanding of tourism. Tourism Statistics usually differentiates three cases - inbound, outbound, and domestic tourism. Further, same-day trips and overnight trips are distinguished.

Tourism Household survey

Statistics on tourism demand is collected EU-wide by NSIs. It covers domestic and outbound tourism. In Germany, 10.000 households are surveyed annually via phone on their travel behaviour.

General information:

*base population: persons age 15+ living in Germany

*sample: 10.000 (non-)travellers

*statistical unit: persons in private households

*legal base: EU regulation on european tourism statistics and implementing regulation

Content and Variables:

Area	Description
Travel behaviour	…
travels with overnight-stay	point of time, nights spent, destination country, purpose, mode of transportation, mode of accommodation, expenditures (+ additional questions every three years)
domestic day-trips	number, purpose, expenditure (only every three years)
international day-trips	number, purpose, expenditures (delivered annually by the federal central bank since 2014)
socio-demographic information on the travellers	number, sex/gender, age

Table 4: Tourism household survey - variables

Tourism border survey

Inbound and outbound Tourism can be captured by border surveys. Some EU countries might conduct a border survey instead of/in addition to a household survey.

Advantages of a border survey in addition/comparison to accommodation statistics, platform data, and household surveys are:

- Information on non-residents (also: staying in non-rented accommodation).
- Accuracy on details regarding the trip(s) might be higher because of promptness.
- Can potentially be linked not only with MNO data but with further data sources from the same time and location (e.g. passenger counts).

For instance, the Spanish border Survey (Frontur) is conducted monthly at Road, Airport, Harbour and Train Station, and covers approximately 450.000 non-residents. Respondents are typically interviewed before exiting the country.

Challenges associated with these databases and ways of handling them

In household surveys, respondents may struggle to recall all travels or travel related details, particularly expenditures, during phone interviews. Additionally, certain population groups may be underrepresented. While survey results are generally comparable across countries, variations in data collection methods may affect consistency. Similar issues apply to border surveys, where respondents may forget or omit relevant information.

One way to address these limitations is through data linkage, especially with digital data sources, MNO data being a key example. In current applications for tourism statistics, e.g. at INE (Spain), the integration of MNO data and data from the border survey is applied and the linkage of further sources is foreseen. More information can be found on the website (annex 11.1) and in the report on methodologies (D3.2) from WP3.

Accuracy for Application Scenarios with MNO data

Combining Border Survey results with MNO data, or even using MNO data alone with appropriate definitions, can significantly improve population coverage. This is because MNO data enables the identification of tourists from a broader range of countries. In contrast, border surveys require a minimum number of respondents from each country for inclusion in the results, often due to privacy concerns. Moreover, integrating MNO data enhances geographical precision, allowing analyses to move from broader regions (e.g. NUTS2) to more detailed levels, like municipality.

Data analysis through the assessment matrix

Because of obligations towards Eurostat, all NSIs have to collect information on travel behaviour thanks to household surveys. Results are also published on Eurostat website. Official statistics quality guidelines are fulfilled.

Official statistics quality guidelines are fulfilled by those NSIs who conduct a border survey, yet not all NSIs conduct touristic travel surveys, which could limit the temporal comparability of these data.

Data Scoring for the combination with MNO data

Tourism Surveys			
Ease of handling the dataset			**high**
Ease of access to a temporally and geographically harmonized data source.	**low**		
Ease of access to detailed data			**high**
Range of possible application scenarios		**moderate**	
EU availability of these data		**moderate**	
Accuracy and robustness of the available information on these data		**moderate**	
Non-sensitivity of these data		**moderate**	

5.3.5 Tourism Platform data

Data description

Online platform data provides information on short-stay (touristic) accommodation which is often not covered in statistical registers because small accommodation establishments (e.g. those with fewer than ten beds), are not legally required to report to NSIs. Currently, there is an agreement between Eurostat and four major online platforms for short-stay accommodation (Airbnb, Booking, Expedia, Tripadvisor) which provide experimental data on short-stay accommodation classified under NACE 55.2. The data is delivered first to Eurostat and then shared with NSIs, in aggregated form. Consequently, booking figures by the single platforms are not available. The data includes the number of hosts, listings, bed places, stays, nights rented out, overnight stays.

Challenges associated with these databases and ways of handling them

The definitions used in official accommodation statistics differ from those used in experimental platform data on short-stay accommodation, thus limiting comparability. Nevertheless, experimental platform data serves as a complementary source offering valuable insights into touristic activity not covered by official statistics.

Furthermore, there are two additional challenges: there is no disaggregation by platform as results are delivered in aggregate form; there is incomplete market coverage as the data, even if covering a large share, does not capture the entire short-stay accommodation market.

Despite these limitations, experimental platform data is a useful source that enables to gain more insights into tourism trends. In addition, a new EU regulation requires all providers of short-stay accommodation - whether individuals or businesses - to officially register. If widely adopted and reliably implemented, this regulation could result in a promising additional administrative data source especially for short-stay accommodation establishments (with fewer than ten bed places or camping pitches) in the future.

Accuracy for Application Scenarios with MNO data

Currently, official statistics does not cover small, short-stay accommodation establishments. Therefore, platform data could help broaden the scope of issues covered by official statistics and improve population coverage for tourism application scenarios.

Detailed methodology and metadata can be accessed at:
https://ec.europa.eu/eurostat/web/experimental-statistics/collaborative-economy-platforms

Data analysis through the assessment matrix

There are different files but no relevant issues in processing are expected. Linkage can be realised at the geolocation.

Data Scoring for combination with MNO data

Tourism Platform data			
Ease of handling the dataset			high
Ease of access to a temporally and geographically harmonized data source.			high
Ease of access to detailed data	low		high
Range of possible application scenarios		moderate	
EU availability of these data			high
Accuracy and robustness of the information available on these data		moderate	
Non-sensitivity of these data			high

5.3.6 Satellite data

Data description

In the past decades, the advent of low-cost computational capabilities has transformed remote sensing data into an active field of research - both as a standalone discipline and as a supporting tool in other fields, including official statistics -, beyond applied geo-spatial domains. Continuous improvement in sensor capabilities and greater data availability are the main drivers behind the increasing use and testing of remote sensing data across a broader spectrum of statistical products, ranging from land use and land cover analysis to other applications.

Satellite data are usually referred to as remote sensing data and relate to measurements of the Earth's surface and atmosphere acquired without direct contact, using active (e.g., radar, LiDAR) or passive (e.g., optical or infrared) sensors that detect reflected, emitted, or scattered electromagnetic radiation. Mounted on satellites, spacecraft, airplanes or drones, these sensors generate continuous data streams characterized by varying spatial, temporal, and spectral resolutions depending on the mission's objectives. Satellite missions like Sentinel or Landsat can produce datasets ranging from several megabytes for low-resolution imagery to exabytes for very-high-resolution products. To build a cohesive dataset for analysis, it often required to combine multiple files that cover a specific area. This is done by using common geospatial identifiers and alignment techniques.

A major advantage of satellite data is its consistent metadata and universal base units, which make it easier to compare across countries and integrate with other geospatial datasets. The data is inherently non-sensitive, except at very high resolutions where it may enable the derivation of sensitive information. Apart from this specific constraint, the data can avoid ethical or legal constraints for the integration with MNO data. Freely accessible state-owned datasets, such as those from the EU's Copernicus Programme or NASA's Landsat, dominate official statistics applications due to their open licensing and analysis-ready data (ARD) formats, which minimize preprocessing burdens. ARD products, which are accessible via cloud platforms like the Copernicus Data Space Ecosystem or Google Earth Engine, heavily reduce the time and effort needed to preprocess the data, enabling NSIs to focus on analysis rather than raw data handling.

Accuracy for Application Scenarios with MNO data

The integration of mobile phone data with remote sensing data can provide a new category of enhanced data platform for the development of innovative statistics. In order to assess the current state of research on data fusion between mobile phones and remote sensing data, a bibliometric analysis was carried out using bibliometrix package (Aria and Cuccurullo, 2017). This analysis helped us to identify the most frequently studied research topic exploiting the combination of these two data sources. The results of the bibliometric analysis should not be interpreted as an estimation of the accuracy of this database to be combined with MNO data, but rather as a starting point for understanding the wealth of possibilities that such data fusion offers.

Data and methods

The methodology used for the bibliometric analysis is described in the appendix.

In the keyword map (Figure 6) several overarching themes of research can be identified, ranging from cities growth patterns in China (land use), pollution induced mortality/morbidity from air transmitted diseases and air pollution, ad-hoc statistical modelling mainly used for data fusion between mobile phone data and remote sensing data, to a plethora of applied statistics on different other themes raging from environment to agriculture (land cover and land use), resulting into relative heterogeneous coverage of different thematic fields of research by the two target data sources, including also some inherent overlaps. Themes are correlated with specific topics captured by figure 7, and are aggregated into 4 main trends split by relevance (number of papers) vs development (number of citations): motor themes (high relevance/high development), niche themes (low relevance/high development), basic themes (high relevance/low development) and lastly emerging or declining themes (low relevance/low development). Accordingly, we can observe that mainstream research is focused on using mobile phone data and remote sensing data for mobility inside and between cities relative to different topics (urban planning, health issues, etc.).

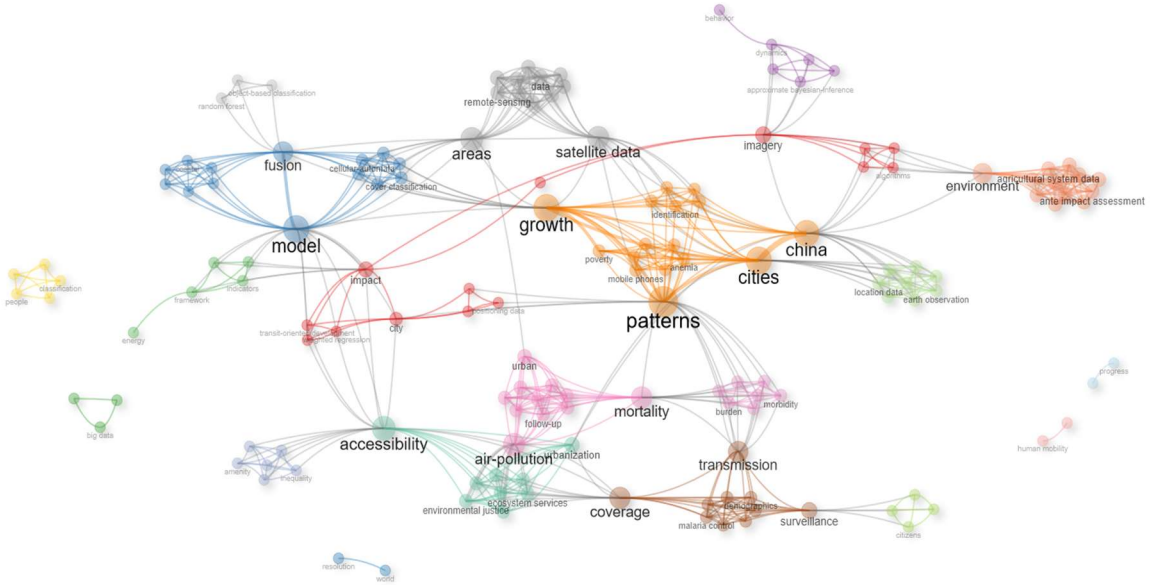


Figure 6: Network of research topics based on keyword frequency – Satellite data

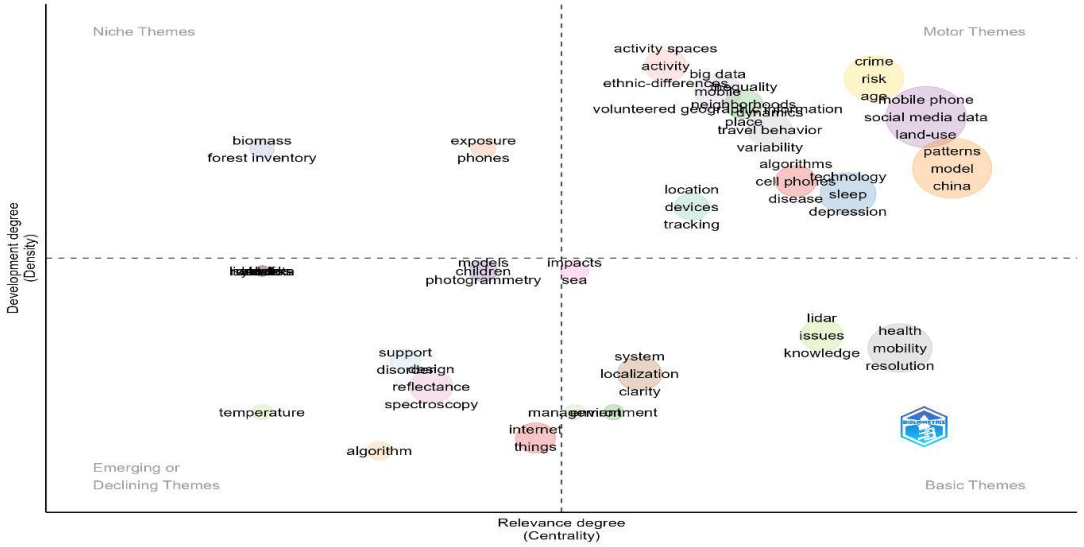


Figure 7: Thematic maps according to trends across time - Satellite data

Short summary of selected papers

Research on urban land use mapping and functional zone identification has advanced significantly through the integration of remote sensing imagery and mobile phone positioning data. (Jia, et al., 2018) proposed an efficient method for urban land use mapping by combining these data sources, showing improved classification accuracy in Beijing. Similarly, Tu et al. (2021) developed a framework integrating remote sensing and mobile data to analyze urban functional zones in Shenzhen, China, revealing the complex urban spatial structure and its deviation from classical urban theories. Tu et al. (2020) further explored the scale effect on fusing remote sensing and human sensing data for urban function portrayal, emphasizing the importance of scale in urban function inference.

The integration of mobile phone data and remote sensing has provided novel insights into socio-economic and poverty analysis. Steele et al. (2017) demonstrated how these data sources could map poverty in low- and middle-income countries, offering frequent and granular assessments. Pei et al. (2014) used mobile phone data to enhance urban land use classification, highlighting the importance of social functions in urban planning. Zulkarnain et al. (2019) improved geodemographic estimation using crowdsourced data, demonstrating significant accuracy enhancements in Jakarta.

The use of big data in disaster management and environmental monitoring has shown promising results. Pastor-Escuredo et al. (2020) proposed a multi-dimensional impact assessment framework for floods, integrating various data sources to support disaster management. Schnebele et al. (2015) discussed using non-authoritative data for disaster assessment, filling gaps left by traditional remote sensing methods. Goldstein and Faxon (2022) explored the role of new data infrastructures in environmental monitoring in Myanmar, emphasizing digital transparency in governance.

Urban activity and mobility patterns have been extensively studied using mobile phone data and remote sensing. Li et al. (2021) investigated the relationship between land use and population distribution around intercity railway stations in China, revealing significant spatio-temporal heterogeneities. Babkin (2014) reviewed the use of mobile phone data in economic geographical

research, highlighting its potential in demographic statistics, transport planning, and human behavior studies. Liu et al. (2015) incorporated spatial interaction patterns in urban land use classification, using taxi trip data to understand social functions of urban spaces.

Studies have also focused on health and environmental exposure using mobile data and remote sensing. Wang et al. (2021) combined high-resolution PM2.5 concentration data with population distribution to estimate human exposure in Beijing, uncovering significant exposure diversity among sub-groups. Chen et al. (2021) investigated ALAN exposure inequity in Tokyo, using mobile phone and satellite data to reveal disparities in light pollution exposure across different population groups.

The integration of various data sources has enhanced credit scoring and economic activity estimation. Simumba et al. (2021) improved credit evaluations for financially excluded individuals by combining mobile, satellite, and geospatial data, showing significant performance improvements relative to benchmark methods. Chang et al. (2016) used social network data to estimate economic activity distribution in Jiangsu Province, China, demonstrating the effectiveness of human mobility data in economic modeling.

Big data has also been utilized in agricultural monitoring and evaluating development outcomes. Fishman et al. (2020) presented a data-driven approach to precision agriculture in small farms, integrating IoT, satellite data, and social networks to support agricultural extension. Rathinam et al. (2021) mapped big data sources to development outcomes, emphasizing the potential of integrating big data with traditional methods to monitor and evaluate sustainable development goals.

Innovations in data collection methods have enhanced the accuracy and efficiency of various applications. Agustan et al. (2018) discussed the use of mobile phones for geolocation and pattern recognition in paddy growth stage reporting, highlighting significant improvements in reporting accuracy. Liddiard (2011) explored the application of microbolometer IR sensor technology for environmental monitoring, showcasing its potential in IoT and early warning systems when combined with remote sensing data.

Research has also focused on social and environmental equity, utilizing big data (including mobile phone and remote sensing data) to uncover disparities. Albanna and Heeks (2018) reviewed the potential of big data to address challenges in positive deviance, proposing its application in various development domains. Lam et al. (2021) discussed household wealth proxies for socio-economic inequality studies in China, integrating diverse data sources to enhance the accuracy of wealth estimation.

Challenges associated with these databases and ways of handling them

Regarding limitations and challenges, the authors highlight issues such as non-representative samples of mobile phone data and the coarse aggregation level of this data in certain instances, such as using call detail records for density estimation without supplementary datasets for calibration or validation. A prevalent issue across most of the selected studies is the lack of reproducibility and transparency, evidenced by the absence of published data and/or code.

Another significant challenge is the high variability of data formats and standards across different remote sensing platforms and MNOs. For instance, satellite data from various missions (e.g.,

Sentinel, Landsat) may have varying spatial, temporal, and spectral resolutions, complicating the integration process with mobile phone data. Similarly, MNO data can vary in terms of collection methods and privacy constraints, making it even more difficult to harmonize these datasets for unified analysis. Addressing these inconsistencies could require lengthy and computationally intensive pre- and post-processing.

As is often the case with dynamic data sources, remote sensing also poses challenges for real-time or near-real-time applications. To tackle this, researchers and practitioners are increasingly leveraging cloud-based platforms and HPC infrastructures to accelerate data processing and enable accurate, real-time analytics.

Data Scoring for combination with MNO data

The integration of remote sensing, mobile phone data and other big data sources has significantly advanced our understanding and management of urban planning, socio-economic analysis, disaster response, environmental monitoring, and agricultural practices. These studies provide some type of first-hand plausible evidence regarding the potential combining mobile phone data with remote sensing data in addressing complex real-world problems, providing a foundation for future research and practical applications across various statistical domains. Present study can be expanded by incorporating other scientific literature databases (e.g. SCOPUS), refine the query to specific targets from current areas of interest pertaining official statistics and to use more sophisticated methods of analysis (e.g. meta-analysis).

In the current state, handling and accessing the data is severely impacted by the complexity and extensive preprocessing required. The preprocessing step could be improved to better integrate the potentially inconsistent data sources. Nevertheless, there are compelling reasons to further explore this type of data and refine the processing techniques. Namely, the high EU-wide availability of satellite data ensures a consistent baseline for cross-national analyses, particularly when combined with MNO data. Moreover, the low sensitivity of satellite data greatly simplifies ethical and legal considerations, making it a valuable complement to MNO data, which often faces stricter privacy constraints. Perhaps the most appealing aspect of this data is the sheer range of possible application scenarios, which could greatly benefit the public.

Satellite data			
Ease of handling the dataset		**moderate**	
Ease of access to a temporally and geographically harmonized data source.			**high**
Ease of access to detailed data		**moderate**	
Range of possible application scenarios			**high**
EU availability of this data			**high**
Accuracy and robustness of the available information on this data		**moderate**	
Non-sensitivity of this data		**moderate**	

5.3.7 Credit Card Transaction data

Data description

This data consists of bank card payment transactions, recorded at the second-level granularity. Each transaction is available individually, providing highly detailed and granular information.

The dataset is hosted within a Big Data architecture. Depending on the specific needs or study, different databases may be accurate, and multiple databases might need to be combined to generate relevant outputs. The data is transmitted via payment terminals, which are registered under the SIRET number of the company it belongs to. The SIRET number, known to the NSI, serves as the key to access information about the company. This number allows transactions to be linked to stores via the SIRENE database (the French National System for the Identification and Directory of Businesses and their Establishments).

The population covered includes credit card holders; however, access to specific personal data about individual cardholders (e.g., name, age, gender) is unavailable.

Challenges associated with these data bases and ways of handling them

The data's volume and intricacy weren't designed for socio-economic analyses but rather for fraud detection. Therefore, to use it in a diverted use, close interaction is needed with the data science team at the private provider's premises. This is even more important because understanding the data collection process requires familiarity with the mechanisms of the electronic payment system.

Other challenges include: data size and format, infrastructure limitations, which may hinder implementation of new ways of studying the data. Moreover, linking companies via the national company identifiers to get extra information, comes with limitations, especially for the company classification and its location. The best way to know more about the database is to spend time with the data provider's team.

Accuracy for Application Scenarios with MNO data

Credit card transaction data should allow for more precise analysis related to population mobility and commercial activity. However, it is less effective for improving socio-economic indicators due to the lack of personal information about the individual owning the credit card. Inferences can rely on hypotheses elaborated with the consortium's teams familiar with the dataset.

Tracking store visits enables the reconstruction of user itinerary, and spending data serve as a good proxy for the household consumption behaviour.

Data analysis through the assessment matrix

The size of the dataset is known, yet it must be split monthly or at least yearly due to the excessive computational resources required.

The data is **owned by a single company:** a national credit card consortium was established by all French banks upon the introduction of credit cards to the market to mitigate interbank transaction fees. Operating its own transaction network, the consortium competes domestically with Visa and Mastercard and oversees all transactions conducted through the CB network, handling approximately 80% of transactions made in France using French cards. The network's design

involves data collection through the bank clearing system. As a result, transactions made from a card to a payment terminal owned by the same bank are not captured in this database.

Access to aggregated data is fee-based, but the credit card company has research partnership which can ease access, for instance through the intermediary of a research chair. Scientific programs are structured to allow non-competitive, diversified uses of the data.

Accessing the data raises legal challenges, as the detailed information is sensitive and can only be obtained under specific conditions of aggregation, the credit card number are anonymised and there is no information about the holder. There is no ethical problem to compute the anonymised data regarding the individual's privacy. However, the data is very sensitive as it contains the gross sales of all the shops (by adding all of the buys from the credit cards). Individual data must be retained for a maximum of two years in accordance with GDPR regulations for the personal data. CB has also rules concerning keeping strategic data from the companies so one cannot identify their sales.

This data may be accessible differently through different EU countries due to the uniqueness of the consortium.

Data Scoring for combination with MNO data

This source is very accurate for official statistics and is accessible through research partnerships which is an adequate way to access it. The company maintains its own quality monitoring system, alerting user when data is missing or quality issues detected. Credit card transaction data has proven valuable in consumption studies, and several ongoing research projects are exploring its integration with MNO data.

Credit Card Transaction data			
Ease of handling the dataset		**moderate**	
Ease of access to a temporally and geographically harmonized data source.			**high**
Ease of access to detailed data		**moderate**	
Range of possible application scenarios			**high**
EU availability of this data	**low**		
Accuracy and robustness of the available information on this data		**moderate**	
Non-sensitivity of this data	**low**		

5.4. Less promising sources

In the following, the analysis of data sources that are considered *less promising* for the integration with MNO data are presented. However, it should be noted that *less promising* refers to the present moment and to the combination of sources with MNO data for official statistics. Therefore, *less promising* does not mean that the reader should refrain from applying these data sources but instead to keep in mind that they currently lack substantial aspects to be applied in official statistics and that this can of course improve in the future or if applying to be aware of the limitations. For example, it might be that some of the following sources are promising but that integration with MNO data currently lacks additional value or it might be the case that some sources at the moment lack relevant applications for official statistics but may produce very insightful research analysis outcomes which lie outside the range of tasks that official statistics currently or even prospectively comprise. The different challenges and potentials as well as justification for scoring them as less promising are outlined for each data source accordingly.

5.4.1 Google Maps Popular Times

Data description

Google Maps Popular Times is a feature designed to help users understand relative crowd levels at various locations throughout the day. By providing both real-time and historical data on how busy places like restaurants, stores, parks, and other public venues are, it enables users to plan their visits more efficiently (Google, 2024).

Popular Times displays crowd levels in two primary forms: real-time data and historical data. Real-time data shows the current busyness of a location, which helps users avoid crowded places at the moment. Historical data, presented as bar graphs, indicates the average busyness for each hour of the day based on data collected over several weeks. This helps users anticipate peak times and choose quieter periods for their visits. When a user searches for a business or a public place in Google Maps, the Popular Times feature appears in the location's information card, typically including a "Live" indicator showing current crowd levels, hourly data for each day of the week, and an estimate of how long people typically stay at that location. The foundation of the Google Maps Popular Times feature lies in the data collected from users who have location services enabled on their mobile devices. This data is sourced from Google Location History, Google Maps usage, and third-party apps that utilize Google’s location services API. When users enable location history on their devices, they contribute anonymized data about their movements and the places they visit. As users navigate using Google Maps, the app collects location data, creating a real-time picture of crowd density. Additionally, third-party apps that use Google’s location services API share anonymized location data to supplement the dataset. Once the raw location data is collected, it undergoes several processing steps to ensure accuracy and privacy. The first step involves anonymizing the data by removing personal identifiers, ensuring that individual movements cannot be traced back to specific users. Following anonymization, the location data is aggregated over time, collecting data points from various users at the same location and time intervals to form a larger, more accurate picture. To further enhance accuracy, the data is smoothed and filtered to remove outliers and noise. For instance, a single user's brief visit to a location might be considered noise

and removed if it doesn't reflect typical user behavior at that location. The core algorithm for estimating popular times employs several statistical techniques. Historical averages are calculated for each hour of each day of the week, examining patterns over several weeks to determine typical crowd levels. These historical patterns help create the bar graphs that show expected crowd density for different times (see Figure 8). Bar graph values are between 0 and 100, where 100 stands for the highest density of visits. For real-time updates, the algorithm uses the latest location data to adjust the historical averages, ensuring that current crowd levels are more accurately reflected. When significant deviations from historical averages are detected, the algorithm adjusts the displayed crowd level accordingly. Machine learning models are also employed to predict crowd levels based on historical data and real-time inputs. These models consider various factors such as time of day, day of the week, holidays, and special events. Special cases, such as holidays, special events, or temporary changes in venue operation, are managed through dynamic adjustments. The algorithm detects unusual spikes or drops in location data that could indicate a special event or an atypical day by comparing real-time data with historical patterns and identifying significant deviations. Over time, the algorithm learns from these anomalies and adjusts future predictions to better accommodate similar events. For example, if a particular location experiences increased traffic every year during a specific festival, the algorithm will incorporate this pattern into its predictions. Google takes several measures to ensure the privacy of its users. All collected data is anonymized to prevent tracing back to individual users. Users must opt-in to location history tracking, and those who choose not to share their location data are excluded from the dataset. The data is also aggregated across many users, ensuring that individual contributions are not identifiable.

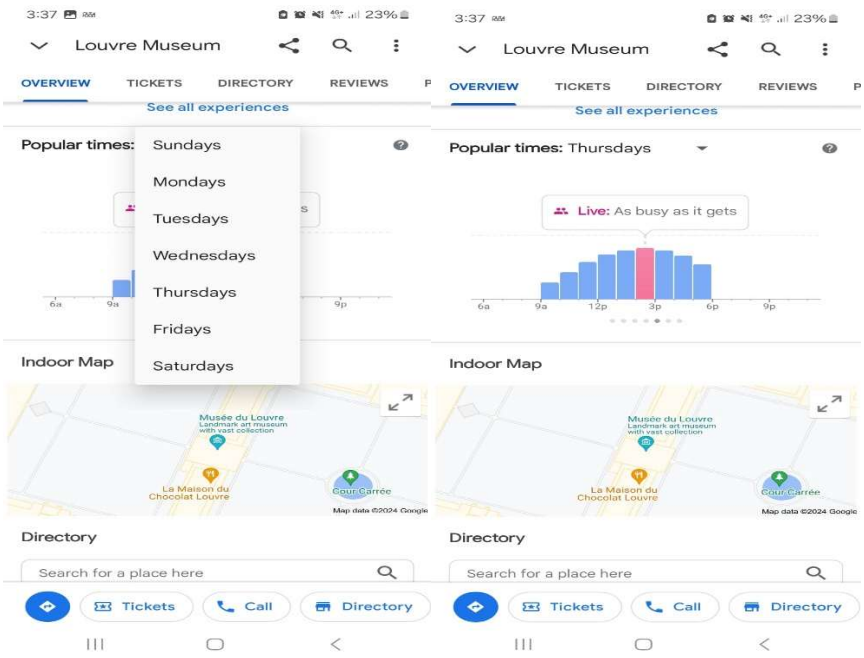


Figure 8: Google Popular Times (screen capture from Google Maps - Android OS)

Challenges associated with this database and ways of handling them

One important drawback is that Google does not provide an API (application programming interface) to download Google Maps Popular Times. The data is available only through the Google Maps graphic interface by manually inspecting the corresponding fields. In order to tackle this problem, web scrapers can be used to access and download data programmatically for particular areas of interest, see Github user m-wrzr (m-wrzr, 2017). But this in turn raises other serious problems regarding web scraping which breaks Google's Terms of Service and can lead to serious legal consequences. Moreover, Google does not provide documentation which contains a systematic and thorough description regarding the underlying algorithms used to compute Google Popular Times data.

Another significant drawback is the lack of details on how the data is processed and aggregated. The exact methodology and inner workings behind its algorithms are not publicly disclosed, making it quite tricky to assess the accuracy and reliability of the data. Without a clear understanding of these fundamental processes, integrating the data with other datasets—such as MNO data—for proper analysis would be a challenging task.

It is also important to mention that location data, by itself, is quite sensitive, and even more so when it's extracted by a third-party company like Google directly from its user base. Despite the data being anonymized, concerns remain regarding user privacy and data security. Furthermore, the reliance on opt-in location history raises questions about the representativeness of the dataset, as it significantly limits the potential sample size by excluding users who choose not to share their location data, making the service potentially unreliable—especially outside large Western urban centers. Properly addressing this issue will likely require collaboration between NSIs and Google to establish a framework for ethical data access and usage while ensuring compliance with privacy regulations.

Lastly, crowd patterns can be very volatile. Accurately modelling and tracking their shifts over time is a huge challenge for maintaining real-time data accuracy. For example, sudden changes in crowd density due to unforeseen events (e.g., emergencies, weather disruptions, or public health measures) may not be immediately reflected in the data, as computations often take into account historical data spanning several weeks rather than adjusting in real time.

Accuracy for Application Scenarios with MNO data

In order to test if there is some interest regarding combining mobile phone data and Google Maps Popular Times we employed a small survey on Web of Science (Clarivate, 2024) and using multiple combinations between the expressions "mobile phone data" and "Google Popular Times". After multiple iterations we failed to retrieve any paper who employed these two types of data sources simultaneously. In order to provide any kind of evidence regarding the potential usefulness of Google Maps Popular Times, we restricted the query only using the expression "Google Popular Times":

(Google NEAR/5 popular NEAR/5 times) (Title) and

(Google NEAR/5 popular NEAR/5 times) (Author Keywords) and

(Google NEAR/5 popular NEAR/5 times) (Abstract)

translatable into the retrieval of all papers which contain that approximate expression in the title, keywords and abstract fields. The query returned approximately 900 results from which after screening we kept only 5 papers which matched our goal, i.e. to identify some type of application which made use of Google Maps Popular Times data.

Some results

Recent studies have explored the use of Google Popular Times data and other digital sources to gain insights into human behavior across various domains, including urban planning, tourism, climate resilience, and electric vehicle infrastructure. These studies demonstrate the potential of using big data and crowd sourced information to understand and predict patterns in human activity and consumption.

Mahdi et al. (2023) investigated the use of Google Popular Times data to model the time spent at Points of Interest in Budapest. By integrating spatial and non-spatial parameters such as ratings, reviews, safety levels, and public transport access, the study developed robust regression models. These models highlighted the significant influence of Points of Interest categories on visitor behavior, providing valuable insights for optimizing activity chains in urban planning and transportation.

Santiago-Iglesias et al. (2023) conducted a case study on the impact of a heavy snowfall in Madrid using Google Popular Times data to assess urban resilience. The research found that essential services were less disrupted than leisure activities and that lower-income neighbourhoods were less affected than higher-income areas. These findings emphasize the importance of proactive urban management strategies to mitigate the effects of extreme weather events, contributing to climate resilience planning.

Moehring et al. (2021) examined tourist behaviour by leveraging Google Popular Times data to analyse customer patronage patterns at tourist destinations. The study found correlations between reviews, timing, and price segments, demonstrating the practicality of Google Popular Times data for understanding tourist behaviour. Similarly, Timokhin et al. (2020) used Google Maps and OpenStreetMap data, including Google Popular Times, to predict venue popularity. By applying models like gradient boosted regression, the research achieved improved accuracy in estimating venue occupancy, underscoring the potential of social media data in predicting consumer behaviour.

Dixon et al. (2020) utilized smartphone locational data from Google Popular Times to evaluate electric vehicle charging demand at popular amenities such as gyms and shopping centres. By adopting a Monte Carlo-based approach, the study provided insights into the potential electrical demand profiles for electric vehicle charging. This method offered a more realistic assessment of charging needs compared to traditional surveys, aiding in the planning of electric vehicle infrastructure to meet future demand.

Data Scoring for combination with MNO data

Google Maps Popular Times data provides high spatio-temporal resolution and has been used in various research applications, including urban mobility, tourism analysis, and infrastructure planning. Although the data is generated from aggregated and anonymized smartphone location history, it remains inaccessible through a public API, requiring either manual extraction or web scraping—both of which present legal and ethical challenges. Currently, there is limited structured access to Google Popular Times for statistical purposes.

Collectively, the aforementioned studies highlight some potential for Google Popular Times data, but judging by the small number of papers identified, the results are not encouraging enough to justify pursuing even a small research track on combining Google Popular Times with mobile phone data at this moment. If agreements between research institutions, NSIs, and Google were established, this data could provide valuable insights into population movement patterns. However, as previously discussed, the data provided by Google is effectively a black box. Without proper transparency into its methodology, using it in the context of official statistics is not entirely plausible due to significant concerns regarding its accuracy and reliability.

5.4.2 Vessel (boat) traffic data

Data description

The Swedish Maritime Administration (SMA) has played a significant role internationally in the development and implementation of the Automatic Identification System (AIS), which has been mandatory for all merchant vessels over 300 gross tons since 2007. This system has also been voluntarily adopted by many recreational vessels due to its utility. AIS transponder data is collected via the coastal radio system and has been systematically stored since September 28, 2006.

Extracts from this database serve various purposes, such as measuring maritime traffic flows, providing valuable insights for decision-making processes, and supporting a wide range of maritime activities. Specifically, the Swedish Maritime Administration utilizes AIS data to:

- Analyze traffic trends in Swedish waters.
- Evaluate and improve sea rescue operations.
- Support fairway projects and hydrographic surveys.
- Design offshore installations and wind farms, leveraging AIS data to ensure safety and optimize site planning.

Accuracy for Application Scenarios with MNO data

AIS data presents significant potential for statistical calibration when integrated with Mobile Network Operator (MNO) data.

Data analysis through the assessment matrix

The RAIS (Regional AIS) database allows for targeted data extraction based on geographical regions or specific passage lines. This flexibility enables diverse use cases, such as:

- Studying the environmental impact of shipping activities on coastal areas.

- Assessing liability in insurance cases involving damage to infrastructure (e.g., cables, piers, and diving equipment).

However, access to AIS data is not free of charge. The Swedish Maritime Administration retains the entire AIS message, which includes attributes such as:

- **Callsign**
- **Course over ground**
- **Draught**
- **Heading**
- **IMO number**
- **Latitude and Longitude**
- **MMSI (Maritime Mobile Service Identity)**
- **Type of cargo**

Data delivery is customizable according to the client's requirements, with formats available in:

- **Shape (vector data)**
- **PDF or PNG (raster images)**
- **Raw data (CSV table)**

Although AIS data is not currently utilized for official statistics, its integration with other data sources holds significant promise for enhancing its analytical value.

Data Scoring for combination with MNO Data

AIS data itself represents a rich and complementary source for official statistics. Researchers and policymakers can gain deeper insights into maritime traffic patterns, improve statistical accuracy, and support innovative applications in maritime management. However, there are currently no clear application scenarios for the integration with MNO data and no examples on what the additional benefits of integrating AIS with MNO data provide. Thus, AIS data should not be overlooked but is currently not prioritised for the integration with MNO data.

5.4.3 Smart Meters

Data description

Smart Meters typically include both Gas and Energy Smart Meter. In Germany, legislation mandates the installation of Electricity Meters by 2032 for all users exceeding a certain consumption threshold. In the past, the analysis or use of smart meter data by NSIs was limited due to a lack of data availability. This is expected to change in the coming years. According to the European Commission, 13 EU-Countries have relatively high Smart Meter coverage.

The data includes energy consumption and input for residential and commercial buildings, along with details on unit, location and aggregation levels by time and space. The quality of this data can vary (with respect to time and space): high-quality data can be associated with exact addresses,

whereas lower-quality data corresponds to an address covering multiple buildings with different uses.

Challenges associated with these databases and ways of handling them

In some EU countries, data is missing due to limited smart meter deployment. Additionally, data ownership and access lie outside the NSIs (consumers own their data and electricity companies hold it). Reliable access to data could be enabled through legislation or alternative mechanisms. Moreover, the rollout of smart meters, such as in Germany, will increase the coverage over time.

Accuracy for Application Scenarios with MNO data

When coverage is sufficiently high, smart meter data can complement MNO data for population and energy statistics. Energy consumption patterns may provide estimates of the household size and MNO data could assign a home location or show present population. Energy consumption can potentially indicate household size as well.

Data analysis through the assessment matrix

Depending on the frequency and detail of the datasets, data can be substantial in volume especially with extensive household coverage. Access possibilities have to be investigated and depends on specific partnership agreements with data providers. An anonymization mechanism must be put in place and user consent may be required.

Data Scoring for combination with MNO data

Data access and coverage varies among NSIs. Currently, most countries do not have access and there are no well-founded studies that would show relevant applications with MNO data. Moreover, handling these large datasets demands advanced data-science skills. As a result, smart meter data has not yet been prioritized.

5.4.4 Event Ticket data

Data description

This data source corresponds to ticket sales data for specific, planned, large-scale events. These include a wide variety of gatherings such as concerts (Altin, Ahas, Silm, S., & Saluveer, 2021); (Beaven & Laws, 2007), music festivals ((Mamei & Colonna, 2015), (Perez, 2016)), major sporting events ((Xavier, Silveira, Almeida, Malab, & Marques-Neto, 2012); (Mamei & Colonna, 2015); (Botta, Moat, & Preis, 2015); (Xu, Fader, & Veeraraghavan, 2015); (Pintér & Felde, 2021); (Solanellas, Muñoz, & Petchamé, 2022); (Nalin, et al., 2024)). These events typically draw significant crowds concentrated in specific venues (stadiums, arenas, festival grounds) over defined, often short, periods ((Altin, Ahas, Silm, S., & Saluveer, 2021); (Mamei & Colonna, 2015); (Nalin, et al., 2024)).

The data originates from and is managed by various entities within the ticketing ecosystem:

- **Primary ticketing agencies:** Large companies like Ticketmaster ((Xu, Fader, & Veeraraghavan, 2015); (Perez, 2016); (Thompson, 2025)) or other national/regional platforms authorized to sell tickets.
- **Secondary marketplaces:** Online platforms (e.g., StubHub, Viagogo, TicketCity, see (Perez, 2016)) facilitating the resale of tickets between individuals or brokers.

- **Event organizers:** Entities directly managing events may handle their own sales or provide sales data summaries ((Altin, Ahas, Silm, S., & Saluveer, 2021); (Solanellas, Muñoz, & Petchamé, 2022)).
- **Venue operators:** May manage ticketing for events held at their facilities.

The primary purpose of data collection is commercial (revenue management, dynamic pricing, sales tracking) and logistical (access control, capacity management) ((Xu, Fader, & Veeraraghavan, 2015); (Solanellas, Muñoz, & Petchamé, 2022)). Secondary markets facilitate liquidity for ticket holders and arbitrage opportunities ((Xu, Fader, & Veeraraghavan, 2015); (Perez, 2016)).

Key variables potentially available (though varying significantly between primary/secondary markets and data providers) include:

- **Event information:** Event name, type, artist/teams involved, date(s), time(s), venue name, venue location/address.
- **Ticket information:** Ticket ID, seat section/location/block (Xu et al. 2015), ticket type (General Admission, VIP - (Perez, 2016)), listed price, transaction price ((Perez, 2016); (Solanellas, Muñoz, & Petchamé, 2022); (Xu, Fader, & Veeraraghavan, 2015)), purchase/transaction date and time ((Perez, 2016); (Xu, Fader, & Veeraraghavan, 2015)), sales channel (primary agency, resale platform, box office - (Beaven & Laws, 2007); (Xu, Fader, & Veeraraghavan, 2015)).
- **Purchaser/seller information** (highly sensitive and often unavailable): Data is generally anonymized. Primary market data might contain purchaser details (name, address, email), but access is heavily restricted by privacy laws (GDPR) and commercial agreements.
- **Transaction information:** Number of tickets per transaction ((Xu, Fader, & Veeraraghavan, 2015); (Perez, 2016)).

Granularity is typically at the aggregated (event) level and to a lesser extent at the individual ticket transaction level. Most studies on the combination of MNO data and event ticket data had not access to actual ticket data but took advantage of the total number of attendees to a large event. This information is often shared by event organizers and published e.g. via newspaper articles shortly after the event took place.

Challenges associated with this database and ways of handling them

Utilizing ticket sales data for statistical purposes presents numerous significant challenges:

1. **Data access, fragmentation, and cost:** Accessing comprehensive ticket data is difficult. The primary market involves multiple agencies, promoters, and venues (Xu, Fader, & Veeraraghavan, 2015). The secondary market is also fragmented across various online platforms (Perez, 2016). Data is commercially valuable and sensitive, requiring specific, potentially costly, agreements with private entities (ibid.). There's no central repository.

2. **Data consistency and harmonization:** Data formats, variable definitions (e.g., seat categories, event types, pricing structures), and level of detail vary across platforms, event types, and time periods. Harmonizing data from different sellers, can be a complex task.

3. **Purchaser vs. attendee distinction:** Primary market data reflects the purchaser, not necessarily the attendee. Secondary market data reflects a resale transaction, further removed from the actual attendee's identity or characteristics ((Perez, 2016); (Nalin, et al., 2024)). This hinders using ticket data for accurate demographic profiling of event attendees.

4. **Incomplete and biased coverage:** Ticket data only represents paying attendees at ticketed events. It misses staff, complimentary tickets, VIPs not sold through standard channels, and personnel ((Nalin, et al., 2024); (Solanellas, Muñoz, & Petchamé, 2022)). It entirely misses non-ticketed events (protests, parades, free festivals) where MNO data can provide insights ((Mamei & Colonna, 2015); (Pintér & Felde, 2021)). Even for ticketed events, purchase doesn't guarantee attendance (Nalin, et al., 2024). Secondary market data represents only a fraction of total attendance and may be biased towards certain ticket types or events (Perez, 2016).

5. **Privacy and sensitivity:** Purchaser data in primary markets is personally identifiable information under GDPR. Transaction prices and sales volumes are commercially sensitive for both primary and secondary sellers. Anonymization/aggregation is typically required, limiting microdata analysis ((Perez, 2016) worked with anonymized transaction data).

6. **Lack of behavioural data:** Ticket data provides a transactional snapshot. It offers no insight into when attendees arrived or left, how long they stayed (dwell time), their mobility within or around the venue, or their pre/post-event journeys ((Nalin, et al., 2024); (Altin, Ahas, Silm, S., & Saluveer, 2021)).

7. **Time constraints**: If information of the total number of attendees is utilised and published in newspaper (online) articles, the required information is often only available for a limited period of time and requires manual search.

Accuracy for Application Scenarios with MNO data

The primary value of combining ticket sales data with MNO data lies in specific, event-focused applications:

1. **Calibration and validation:**

- Ticket sales figures (total sold, or category breakdowns) serve as the most common form of quantitative benchmark or "ground truth" for validating or calibrating MNO-based estimates of attendance at specific, ticketed events. Numerous of the abovementioned studies rely on this comparison. Mamei & Colonna (2015) explicitly use stadium ticket data for training their MNO model. Botta et al. (2015) use match attendance counts to validate correlations with mobile/Twitter activity.
- Comparing expected attendance (ticket sales) with actual presence (MNO data) can potentially shed light on no-show rates or the effectiveness of MNO data in capturing event populations.

2. **Broadening the scope of issues covered:**

- Event Context and Metadata: Ticket data provides accurate metadata about the event (type, official timing, venue, pricing structure) necessary for interpreting MNO activity patterns (Thompson 2025; Solanellas et al. 2022).

- Attendee Origins (Limited): Primary market purchaser addresses (if accessible and reliable) could offer some coarse geographical insight complementing MNO roaming data (Altin et al., 2021), but the purchaser-attendee issue severely limits its accuracy for profiling actual attendees. Secondary market data offers virtually no reliable origin information.
- Economic Insights (Limited): Provides direct data on ticket revenue (Perez 2016; Xu et al. 2015; Thompson 2025). Combining this with MNO-derived presence could potentially link spending (ticket purchase) to behaviour, but this link is weak due to timing differences and the purchaser-attendee gap. It does not capture on-site spending.
- Overall, the accuracy for improving general population coverage or fine-grained spatio-temporal analysis is low. Its main strength is providing a quantitative event size estimate for calibration. MNO data remains far superior for capturing actual dynamic presence, mobility, dwell times, and reach beyond ticketed attendees or venues (Nalin et al., 2024; Altin et al., 2021; Mamei & Colonna, 2015; Pintér & Felde, 2021).

Data analysis through the assessment matrix

- **Data type:** Size highly variable. Harmonization across disparate sources (primary/secondary, different vendors) is a major technical challenge. Technical cost: Moderate to High.
- **Access:** Fragmented across numerous private entities. Requires potentially costly, specific agreements. Stable, harmonized access is currently unrealistic for broad statistical use. Detailed purchaser microdata access is highly restricted (privacy/commercial). EU availability is non-existent in a standardized form. Scores: Easiness (harmonized) = Low; Easiness (detailed) = Low; Range of possible use cases = Moderate (focused on event calibration/context); EU availability = Low.
- **Metadata:** Event metadata generally good within a specific dataset. Purchaser/attendee link is weak. Harmonization of metadata across sources is poor. Accuracy high for the transaction record itself, low for inferring behaviour or representative attendance. Score: Accuracy/Robustness = High (for transaction facts) / Moderate (for attendance/behaviour inference).
- **Sensitivity:** High due to personally identifiable information and commercially sensitive (sales/pricing information). Score: Non-sensitivity = Low (for detailed data).

Data Scoring for the combination with MNO data

Ticket sales data offers a specific, valuable function as a calibration source for MNO-based attendance estimates at large, ticketed events, a use case demonstrated across multiple studies (Mamei & Colonna, 2015; Botta et al., 2015; Xavier et al., 2012; Solanellas et al. 2022). It provides a quantitative benchmark related to the potential or paying audience size.

However, the limitations regarding access, fragmentation, cost, privacy, the purchaser-vs-attendee problem and lack of behavioural information make it unsuitable for large-scale, regular integration with MNO data for producing general official statistics on population presence or mobility. Its utility

is largely confined to event-specific research where data partnerships are feasible and the focus is on validating MNO signals against known event parameters. Event data on the more aggregated granularity level, e.g. information in newspaper articles on the total number of attendees based on event ticket data is more easily accessible, however often only available for a certain period of time.

Therefore, the source "Tickets sold at mass events" is classified as *less promising* because there are only for limited use cases (event calibration/validation for ticketed events) and is generally unsuitable for broad official statistics production.

5.4.5 Social Media

Data description

Social media data usually consists of information that platform users have generated. Some of platforms offer partial data access via APIs (Application Programming Interface), typically requiring registration and the indication of a research purpose. Access is granted by the platform.

Data from the platform X typically includes the date, text, username, location (if geotagged), hashtags, user mentions, URLs and media objects such as videos or images

Challenges associated with these databases and ways of handling them

Not all social media platforms offer access and pricing models vary (e.g. basic, pro/enterprise): Probably limited information is available on data selection criteria and social media content does not represent the entire population (limiting its use or specific research questions, e.g. like trending digital topics).

The challenge might not be handling an overly large dataset but rather the opposite: potentially not having a sufficient amount of data. However, this ultimately depends on the access level and consequently on budget constraints.

Accuracy for Application Scenarios with MNO data

Social media data can complement MNO data by improving population coverage for specific events or by providing socio-demographic information/context or by relating topics/events with population number and location for very specific research questions. However, it is unlikely to improve overall population coverage or geographical precision on a large scale, due to the limited number of active, geotagged users.

X and similar platforms are generally not suitable for large-scale population analysis due to limited user participation (to represent or draw conclusions to total population), whereas Facebook/Meta may be more relevant given its profile data (e.g. profile info, not posts). However, challenges about data quality and verification persist. Many accounts are unverified or non-representative of a real person (e.g. fake-accounts, bots, multiple accounts, small business accounts not clearly identified).

These data can provide socio-demographic and geolocation insights for specific events but only to a limited extent and for certain population groups.

Data analysis through the assessment matrix

Data access can be purchased or free access may be granted via applying as a researcher, e.g. NSIs have a research department or utilise data in a research context. However, the volume and variables may vary.

Data Scoring for the combination with MNO data

Social media data could provide interesting complementary information when combined with MNO data. However, access complexity and uncertain and hard-to-estimate representativeness make this data less promising in the short term.

5.4.6 Connected vehicles

Data description

Connected vehicle data includes a wide range of information, primarily collected through digital sensors and geo-location trackers, and transmitted via mobile networks for external processing and use. Usually, car manufacturers (OEMs) hold ownership of this data. For statistical purposes, key variables of interest include:

- Vehicle location (via GPS and communicated through in-built sim cards): latitude, longitude, altitude, travelling direction, timestamp
- Vehicle information: build date, country code, drive type, number of seats, ...
- Navigation destination
- Parking ticket data: operator name, ticket ID, parking status, start and end times
- Sensor data
- Vehicle speed
- Battery/fuel level + consumption
- Mobile connection, GPS signal strength

Among these variables, vehicle location data is probably the most relevant for statistical use. Therefore, most of the following analysis focuses primarily on vehicle location data.

Challenges associated with these databases and ways of handling them

So far NSIs have little or no practical experience in using this data source. The main challenges lie in securing data access and obtaining detailed metadata on the variables.

Accuracy for Application Scenarios with MNO data

When linked, vehicle data can provide insights on road traffic patterns and help estimate people movement during specific periods. Variables such as vehicle speed, location, direction and stop duration can serve as inputs in origin-destination models. Data consistency is expected within manufacturers or at least car models and across countries. Some conceptual differences may arise between different car manufacturers; geolocation data is expected to be free from major inconsistencies.

Since MNO data generally exact location information, GPS data can help improve spatial accuracy. The range of application scenarios could also be extended, for instance concerning analysis of commuting patterns and estimation of CO2 emissions.

Data analysis through the assessment matrix

Frequent, raw or pre-processed data at fine spatial and temporal resolutions will result in large datasets. Potential issues depend on factors such as the total data volume (raw or aggregated data), number of vehicles tracked, number/share of OEMs involved. Data is owned by car owners while car manufacturers (OEMs) hold and can access the data anytime via mobile networks. Data and service providers generally purchase data from OEMs. NSIs may acquire data through purchases or establish agreements with OEMs or data providers. The question whether the involvement of multiple actors leads to heterogeneous data formats remains to be investigated.

Quality issues require thorough examination. Data gaps may occur in rural areas with poor mobile network coverage. Public documentation is not available on data construction methodology and quality issues It is partly available upon data acquisition. Given that this data is actively used by car manufacturers, insurance companies, car sharing services, etc. It is reasonable to assume that outputs undergo regular quality assessment.

Data Scoring for combination with MNO data

While connected vehicle data hold potential for future applications, at present too many uncertainties remain regarding both data access and quality.

6. Source Availability

To investigate the availability of data sources in the ESS, availability was first assessed as part of the primary scoring by a theoretical approach. This means that it was researched how well a respective data source is (in theory) available for NSIs. However, for some data sources the availability at NSIs cannot be identified or is hard to assess by a theoretical approach. Further, even though the source might be available, it may not be used or currently accessed by NSIs. Therefore, a short survey was conducted to enrich primary scoring with information directly from the individual NSIs in the ESS. Further, some additional information was asked to enhance the secondary scoring of data sources.

6.1. Short Survey on the availability of data sources among ESS-countries

To keep the short survey convenient to answer, it was crafted in Microsoft Excel. The survey consisted of three sheets: Introduction, Survey and Explanations. Responses were collected in the sheet “Survey” which consisted of three parts plus one box for contact information and one for comments. In Parts A-C, respondents could select pre-defined answers via drop-down. The full list of possible answers and explanation can be found in the annex.

Part A focused on access to data sources from official statistics (Figure 9).

Part A	Data source	In Production / Current Access?	Did you have or will you gain access in the last or next 12 months?	If yes, at which Granularity level?	If yes, at which Geospatial Resolution?
Official Statistics / Official Data	Land Cover and Land Use Register				
	Total Population Register				
	Census				
	National Travel Survey				
	Tourism Household Survey				
	Tourism Border Survey				
	Tourism Accomodation Statistics				
	Please add missing source here (if any)				
	Please add missing source here (if any)				

Figure 9: Survey Part A - Official Data

In Part B, information on new digital data sources was collected.

Part B	Data source	Do you currently have access?	Did you have or will you gain access in the last or next 12 months?	If no, what is the main reason?	If no, can you identify the provider?
New Data Sources	Tourism Platform Data				
	Google Maps Popular Time				
	Social Media				
	Vehicle, Bicycle and Pedestrian Sensors				
	Vessel (boat) Traffic Data				
	Pollution Data				
	Smart Meters				
	Connected Vehicles				
	Satellite Data*				
	Electronic Invoices				
	Credit Card Transaction Data				
	Ticket Data*				
	Please add missing source here (if any)				
	Please add missing source here (if any)				

**If applicable, please specify/describe briefly which kind of data (in the comment box)*

Figure 10: Survey Part B - New Digital Data

After collecting information whether NSIs have access to new data sources, and the reasons for any lack of access, respondents were asked to provide further details about the nature of their access to the sources currently available to them:

Part C	Data source	Mode of Access	Privately or publicly held data	Are there any costs?	Granularity	Coverage	Geospatial Resolution	Frequency of Data Delivery	Duration of Access	Period of Time until Access
New Data Sources										

Figure 11: Additional information on available new data sources

The variables were derived from the assessment matrix and the quality requirements from the code of practice that were identified as relevant from the secondary scoring.

To enable wide participation, MNO-MINDS partners, Task Force MNO members and Task Force TSS members were invited to participate in the short survey with the aim to receive one consolidated response per country. In total, NSIs from eleven countries listed in Figure 12 submitted their response to the survey.

AT	Austria
DE	Germany
ES	Spain
FR	France
IT	Italy
LT	Lithuania
NL	The Netherlands
NO	Norway
PT	Portugal
SE	Sweden
SI	Slovenia

Figure 12: List of Countries

The reference persons who submitted the responses for their NSI filled the information according to best knowledge and internal research. However, it cannot be guaranteed that the provided information is entirely complete.

6.2. Availability of data sources

To capture the availability of data sources analysed in this work package, respondents were asked to indicate for each data source whether their respective NSI currently has access to it by selecting “Yes” or “No”. For the most promising sources, summarised results are presented in the following. The full list of responses for each data source can be found in the annex.

Most countries reported having access to Census data and Total Population Register data, with all but one indicating access to at least one of the two (Figure 13). The country that answered “No” to both still relies on a combination of census and register data. The primary scoring in chapter 5 had labelled EU-wide availability of Census data as “high” and Total Population Register as “moderate”. According to the survey results, availability of Total Population Register seems to be higher at least among survey participants than previous research concluded.

For National Travel survey, four out eleven countries indicated having current access to it (Figure 13). For countries without access, background research could show in how many countries results from a national travel survey would be available. With the help of the survey results, primary scoring in chapter 5 is updated from previously unknown to “moderate”.

Regarding Land Cover and Land Use Register, nine out of eleven countries have access to it which confirms the high-ranked availability from previous scoring.

Tourism Accommodation Statistics is the only data source that all participating NSIs have access to which confirms previous scoring in chapter 5.

Overall, availability of the most promising data sources is either “high” or “moderate”. For cases scored as “moderate”, further investigation could look into reasons for indicating not having access to it.

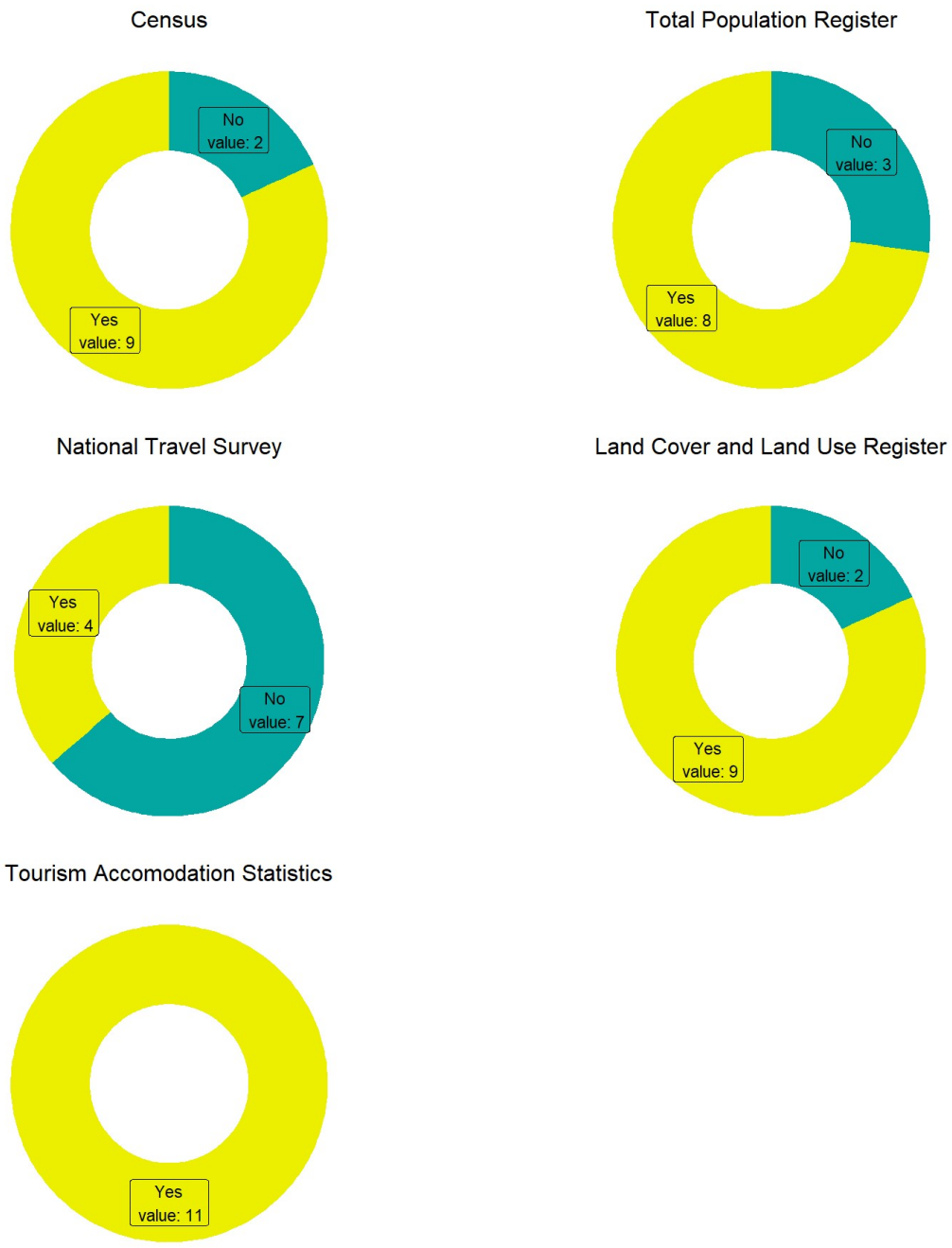


Figure 13: Survey Results - Access to the most promising data sources

Regarding the next category, *Promising sources but which would require substantial work, or which are not yet fully accessible,* availability of data sources looks more heterogeneous:

Best available is Satellite Data with 9 out of 11 countries having current access to it. Next, Credit Card Transaction Data is relatively well available with more than half of participating NSIs having access to it. Then, Tourism Platform Data as well as Vehicle, Bicycle, and Pedestrian Sensors are both available at five countries. For Pollution Data, and Electronic Invoices, there are four countries with current access.

The availability of most less relevant sources is consequently worse than for the other categories: For Vessel (boat) Traffic Data, four countries indicated having access to it. Two countries reported having current access to Smart Meter data. One country informed about having access to connected vehicle data and another one for social media data. There is only one NSI with access to ticket data. None of the participating countries is currently accessing Google Maps Popular Time.

The charts for all data sources can be found in the annex 11.4.

Consequently, the EU wide availability of data sources is updated as followed according to data analysis and results from the short survey:

Data source	Availability
Census	High
Population register	High
National Travel Survey	Moderate
Land Use and Land Cover register	High
Tourism Accommodation statistics	High
Vehicle, bicycle and pedestrian sensors	Moderate
Pollution data	Moderate
Satellite data	High
Electronic invoices	Moderate
Tourism Household and Border Surveys	High (household), moderate (border)
Tourism platform data	Moderate
Credit Card Transaction Data	Moderate
Event Ticket Data	Low
Google Maps Popular Time	Low
Vessel (boat) traffic data	Moderate
Smart Meters	Low
Connected Vehicles	Low
Social Media	Low

Table 5: Scoring for data availability

In Part B, respondents were asked to provide reasons for not having access to a certain data source. In most cases, there was no reason provided. For the cases that a reason was named, it was mainly financial or legal reasons, followed by bureaucratic and then technical reasons.

6.3. Additional Insights

To learn more about the availability of new data sources used by national statistical institutes, more information on relevant criteria from the quality requirements from the Code of Practice was provided by the respondents on new data sources that they have access to. Even though the collected information is limited to only a few data sources and to a small number of NSIs - therefore caution is advised to draw general conclusions - these additional insights support the secondary scoring.

First, regarding the mode of access, most countries acquired data access via partnerships. Another important access mode is open source data. This result aligns with the fact that most additional information was provided for new data sources and that NSIs carefully balance costs and benefits before utilising purchased data sources.

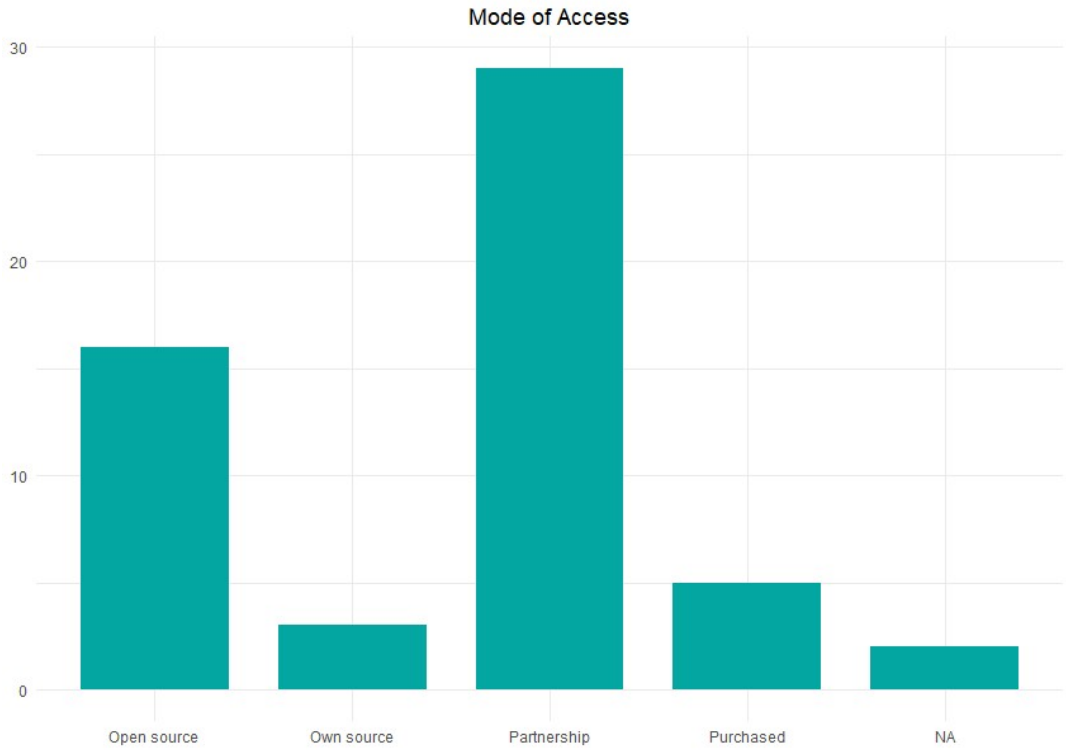


Figure 14: Mode of Access

Further, results show that most accessed data sources are publicly held and are obtained without any (direct) costs. This seems to underline the importance of low barriers of access for NSIs to use new data sources as there are typically constraints, e.g. financial.

Most new data sources that additional information was provided on are available as micro data but a notable amount is available as macro data or nano data. Regarding coverage, a lot of sources have either full or partly coverage of the total population. Geospatial resolution depends on the data source but in total counts most are available as precise point precision, as grid cells, and on municipality level.

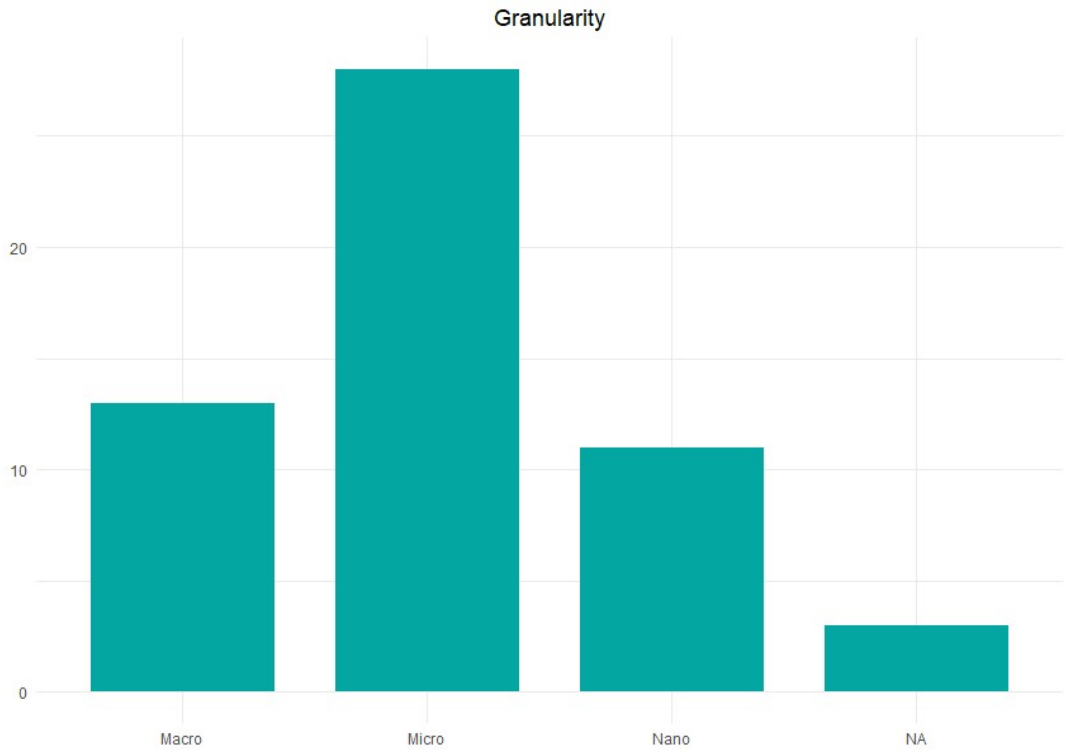


Figure 15: Granularity

Most new data sources that NSIs have access to come as a regular data delivery which is consistent with abovementioned characteristics that they are usually obtained through partnerships and mainly public data. Consequently, for most data sources, a permanent access is established (meaning there is no expiration of contract defined). For the majority of sources, access was established within less than three months after requesting or applying access. However, for a notable number of sources establishing access took longer than 12 months or period of time was unknown. It should be noted that preparations and negotiations for establishing data access can consume a lot of time even though data access can be granted within a very short time frame after officially requesting access.

Figure 16: Frequency of data delivery

Further charts for the abovementioned figures can be found in the appendix 11.4.

Again, the reader is advised to consider that this information refers only to new data sources that NSIs already gained access to. If a frequent data delivery is not possible and if there is a lack of sufficient granularity or lack of appropriate mode of access, an NSI possibly refrains from testing a specific data source.

In addition to the data sources listed, respondents added new data sources that they do have access to and that may be promising to be integrated with MNO data:

- E-turitas
- Webscraped data
- Scanner data
- Smart survey data (scanned invoices and measured geolocations)
- Lidar/3D data
- Weather and climate data
- Land regulation plan data
- Geological map data
- Data from electronic tolling system for vehicles >3.5 tonnes

7. Integrating MNO Data with external sources: From target statistics to surveys

As previously mentioned throughout this report, MNO data offers unique opportunities to enhance the quality and coverage of official statistics. However, its full potential can only be realised when it is effectively linked with other data sources such as administrative records, surveys, and sensor data.

This chapter explores how such integration can be achieved through three key approaches:

- Identifying potential target statistics and relevant data sources

- Applying metadata analysis to evaluate data integration possibilities without direct data access

- Designing dedicated surveys to fill information gaps and improve calibration

Together, these approaches provide a framework for National Statistical Institutes (NSIs) to build stronger, more representative statistical outputs.

In particular, this chapter outlines the connection with potential target statistics by referring to the commuter application scenario as an example, it provides support for combining diverse data sets with MNO data by a metadata analysis, and it finally suggests sets of questions that could enable a better linkage of data sources with MNO data. Further information on application scenarios can be found in the methodological report D3.2 from WP3 (Zhang, et al., 2025) and more information on a dedicated MNO survey in deliverable 4.2 from WP4 (Kowarik, et al., 2025).

7.1. Identifying potential target statistics

To begin the integration process, NSIs need to identify the most promising applications where MNO data can complement or enhance existing statistics. Drawing from WP2 and WP3, several potential target statistics have been identified.

Table 6 summarizes examples from WP3, where MNO data can be combined with other data sources for applications such as inbound tourism, commuter statistics, and accommodation statistics. More information on the application scenarios and potential target statistics can be found in D3.2 *Report on methodologies* from WP3.

Potential Target Statistic	Application Scenario (WP3)	Data Sources
Tourism Statistics (e.g. inbound tourism)	Inbound Tourism	MNO data Tourism surveys
Mobility and national travel statistics (Commuter statistics)	Trips	MNO data Origin + destination POI
Commuter statistics	Commuters	MNO data National travel surveys Administrative data
Tourism Statistics (e.g. flash estimates)	Nights-spent	MNO data Tourism Accommodation Statistics
Mobility, travel, commuter, tourism statistics, …	Sensor presence	MNO data Sensor data

Table 6: Potential target statistics - part 1

Table 7 below highlights additional opportunities identified in WP2's scoring analysis, including credit card transaction data and event ticket data.

Potential Target Statistic	Application Examples (WP2)	Data Sources
Population statistics	De facto population Night-time resident estimates	MNO data Census or population register (additional administrative data) (land use and land cover register)
Mobility, national accounts, retail trade (e.g. flash estimates)	Population mobility Commercial activity	MNO data Credit Card Transaction data Electronic invoices
Environmental topics	Access to green areas Exposure to pollution	MNO data Pollution data Satellite data
Tourism statistics and mobility	Transport and mobility	MNO data Event ticket data

Table 7: Potential target statistics - part 2

These lists are intended as starting points and should not be considered exhaustive. NSIs are encouraged to explore their own potential applications based on available data sources and user needs.

Tourism statistics: Consolidation and potential target statistics

Data scoring in chapter 5 considered several sources from tourism statistics that eventually were scored differently. Therefore, a summary and consolidation of all tourism sources is needed.

At first, some background information on tourism statistics is helpful: Tourism Statistics usually differentiates three cases - inbound, outbound, and domestic tourism. Further, same-day trips and overnight trips are distinguished.

Official Statistics (OS) currently uses three main sources to capture Tourism: Accommodation Statistics (domestic + inbound), Border Surveys (inbound + outbound), and Household Surveys (outbound + domestic). In general, a linkage of these sources with MNO data (and further data sources) provides complementary information, for example to get a deeper understanding of inbound tourism. Additionally, MNO data together with data sources currently not in use for OS in tourism, could potentially provide information on touristic activities that OS does not capture much yet: This is especially the case for same-day visitors and for (private) short-term accommodation. As a new data source, tourism platform data was analysed in chapter five. Further, some more data source may be relevant for specific applications:

Potential target statistics	Official sources	Non-traditional sources
Inbound tourism	• Tourism Accommodation Statistics • Border Surveys • Number of Passengers (e.g. from State Ports)	• MNO data • Sensor data • Transaction data (e.g. platform data, ticket info) • Points of Interest
Outbound tourism	• Border surveys	• MNO data

	• Household surveys	• Sensor data • Transaction data
Domestic tourism	• Tourism Accommodation statistics • Household surveys	• MNO data • Sensor data • Points of Interest
Same-day visitors	• Land Cover and Land Use registers (e.g. filter for areas marked as "sports and leisure" ad "historical building")	• MNO data • Points of Interest

Table 8: Data Sources for Tourism Statistics

In general, web data from various sources (Google Maps Popular Times, Tourism Platform data, Points of Interest from Open Street Maps) may be useful. Further, there might be private stakeholders that hold useful information on touristic activities, e.g. tourism associations, travel agencies and operators.

7.2. Application example: The commuter Scenario

The commuter scenario is a prime example of how MNO data can enhance official statistics. In many ESS countries, commuter statistics are derived from administrative sources and national labour force surveys. These sources, while valuable, have clear limitations:

- They are often based only on place of residence and place of work, limiting understanding of actual travel patterns.
- They cannot easily capture emerging trends such as remote work, part-time commuting, and cross-border movement.
- Flat-rate travel tickets (e.g., the Deutschlandticket in Germany) obscure trip-level data because single-ticket purchases are no longer a reliable proxy for travel behaviour.
- The spatial resolution of administrative data is frequently tied to municipal boundaries, creating uncertainties when grid-cell analysis is needed.

Example from Germany: Official statistics, such as the *Länder commuter statistics* and *Commuter Atlas Germany,* have long provided insights into commuter behaviour. These are based on social security registrations and administrative sources but cannot always capture nuanced changes, such as new workplace arrangements, georeferenced commuting paths, or changes in ticketing systems.

This data gap can be partially bridged by MNO data, which provides granular information on the presence and movement of devices. When combined with other sources, it enables:

- Better origin-destination matrices, beyond administrative boundaries
- Dynamic monitoring of commuting behaviour, reflecting real-world changes
- Greater insight into transport mode choices and trip purposes when linked with surveys or registers

However, the integration process is complex:

- The contribution of MNO data depends heavily on the variables and units available in both the MNO data and the complementary sources.

- As the number of potential integration scenarios grows exponentially with each additional data set or variable, manual evaluation becomes infeasible.

This is where *metadata analysis* plays a key role (see Section 7.3). It enables NSIs to evaluate, in a systematic way, which combinations of MNO and external data can produce the desired outputs. In this project, a user-focused Jupyter notebook was developed specifically to support such analysis for the commuter case study, allowing NSIs to test different integration scenarios even before accessing the actual data.

A deeper exploration of the commuter scenario allows National Statistical Institutes (NSIs) to clearly identify the most significant data gaps that persist in traditional statistical systems. It highlights how MNO data can complement and enhance existing sources, filling critical blind spots and offering richer insights.

7.3. Metadata analysis

This chapter presents the application of metadata analysis to support the combination of diverse data sets with MNO data, assisting in the creation of new statistical outputs. The presented application is designed to assist National Statistical Institutes (NSIs) in evaluating the feasibility of generating desired target outputs based on available data sets and models. Metadata analysis enables users to make a systematic overview of available data sets and models during the design phase of a new potential output, without the need for data access. Metadata descriptions can be shared and discussed between NSIs without privacy concerns, thus promoting collaboration and knowledge exchange across statistical organizations.

For given user input, metadata analysis can answer questions such as “Can an intended output be created from a given set of input data?” and “If so, what sequence of processing steps is then needed?”. The required input consists of several parts, which is discussed in the first paragraph. After that, the commuter example will be used to explore the potential of metadata analysis. A user-focused Jupyter notebook has been developed, specifically tailored for application to the commuter case study with MNO data.

7.3.1 What is metadata analysis?

To adequately analyze the process of combining data sets without the need for the values in those data sets, several key aspects must be known on a metadata level. The following input is required:

- A list of available data sets, including administrative records, survey data, and other data sets such as sensor data, along with the variables included in each data set.
- For each variable in a data set, their role as either a measurement variable or an identifying variable should be known.
- The relations between granularities of the variables across different data sets, in the form of conversion graphs and aggregation graphs for measurement variables and identifying variables respectively.
- Available models defined by input and output data sets only, without the need to specify algorithmic details.
- The intended target output is required in the form of a single data set, specified by a set of variables that the user wishes to generate.

A framework of logic concerning the above ingredients as well as pre-defined processing steps describes the process of combining data sources. This logic was implemented in a Python program that automatically searches to determine whether the available models and data sets can lead to the desired target output. If possible, a path will be provided consisting of all processing steps in chronological order required to create the target output. The path returned by the program may consist of the following processing steps:

- Conversion. This processing step converts a measurement variable from one granularity to another, as defined by the conversion graph for the variable. For example, the variable time can be converted from days to hours by dividing by 24 and vice versa by multiplication.
- Aggregation. This processing step and aggregates an identifying variable from one granularity to another, as defined by the aggregation graph for the variable. For example, the variable neighborhood can be aggregated to municipality. Note that this processing step cannot be reversed as it results in a loss of information.
- Combining. This processing step combines two or more data sets into a single one, through column-wise or row-wise combining.
- Modelling. The user defined models may be used to process one or more data sets (as specified by the model input) into the output data set. The effects of a modelling step cannot be achieved by chaining any of the above three processing steps.

The software is currently being prepared for open-source release as a Python package. It can be used in two ways. The first application requires manual input from the user, and can analyze any scenario defined in this way. The second application contains pre-loaded data sets and models the user can choose from to construct a scenario to analyze about the commuter case study (see paragraph 7.3.2). It is more user-friendly and was developed specifically for the current project and requires only minimal input from the user (see paragraph 7.3.3).

7.3.2 User-friendly implementation

Metadata analysis using the software mentioned earlier requires manual actions such as defining variables, data sets and models. Supplementary to the available software, a user-friendly notebook was developed for the application of metadata analysis on combining data sources with MNO data. It both illustrates the capabilities of metadata analysis, as well as enables those interested in combining MNO data with other data sets to analyze various scenarios.

The user is first presented with an introductory text and instructions on how to use the notebook, followed by a legend explaining the available variables and granularities used to define all other relevant objects for the analysis. The user may customise the metadata scenario by selecting which sources are available, which variables are available in the MNO data, which models are available and what is the intended output. The user may confirm and inspect the scenario to be analysed. If a path is found from the available input data sets to the target output, it is displayed.

Pre-defined metadata of several data sets are available for the user to choose from:

- Census data containing background characteristics and number of inhabitants per home and working location.

- MNO data containing a number of observed sim cards per current location. The user is asked to specify whether MNO data is available for a single provider or all providers in the country of interest, and whether home location may be available as a variable.
- National travel survey contains background characteristics, transport mode, trip purpose per person, route and time interval for a sample of the population.
- Population register contains background characteristics, home location, working location per person for the entire population.
- Route data contains information on which road segments are part of a route per route defined from an origin (home location) to destination (working location). This data is available for all possible routes and all road segments.
- Traffic loop data contains observed vehicle counts for the transport modality car per road segment on a minute level. This data is available for only those road segments where a traffic loop sensor is located.

The above list of data sets and their variables were chosen to illustrate the potential of metadata analysis for the application scenario of commuters in the user-friendly notebook. Similarly, the pre-defined metadata of several models are available for the user to choose from:

- Calibration models that either calibrate vehicle to person or sim cards to person. Several variants of such models are available to choose from. Calibrating the number of observed vehicles to number of persons is relevant for using traffic loop data.
- The model “create OD matrix” takes a dataset similar to the population register as input, and results in a count of number of persons per origin-destination combination, based on home location (origin) and working location (destination).
- Location estimation models are relevant for adjusting MNO data per cell tower area to municipality (“crude” model) or neighborhood (“detailed” model).
- The modality choice model uses background characteristics to estimate the probability of transport mode choice. It requires the national travel survey to be available as training data.
- The shortest path model is a route planning model that determines which road segments are present in routes from an origin to a destination.

Users may extend the framework by adding metadata of additional data sets and models, thereby broadening the range of scenarios that can be analyzed. The metadata of variables and data sets can be adjusted according to user preferences in a more customizable, though less user-friendly, notebook.

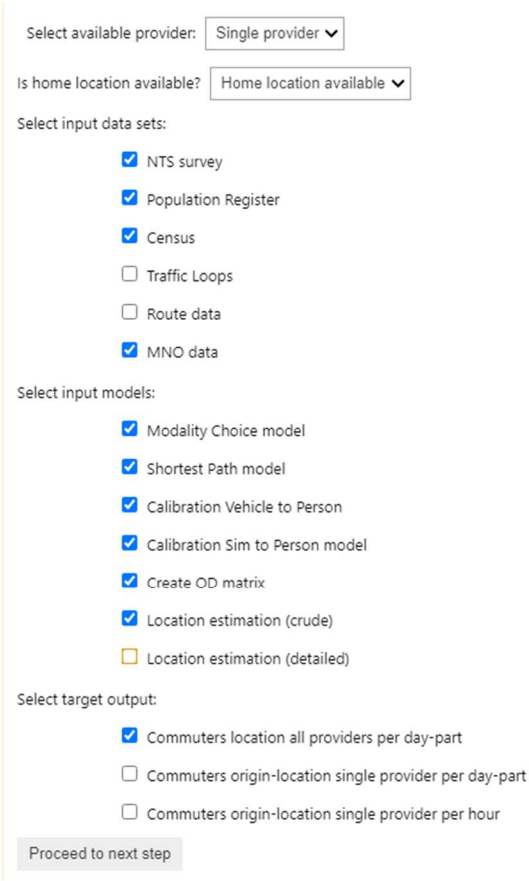


Figure 17: Screenshot a of the user-friendly notebook for metadata analysis: the user may customise the metadata scenario.

Given the scenario selected by the user, the framework is able to answer questions such as "Can an intended output be created from a given set of input data?" and "If so, what sequence of processing steps is then needed?". The underlying software automatically searches if the available models and data sets can lead to the desired target output. If possible, a path will be provided consisting of all processing steps in chronological order required to create the target output. Screenshots of the notebook are included in the annex. The Python code for the notebook and the underlying functionalities are available at https://github.com/YGootzen/MetadataFramework/tree/main/python .

'Target output can be created by the following path:'

step	method	method_detail	input	output
step 0	start set			
step 1	model	Location estimation (crude) l: 2→1	MNO data (n0 | l2, o2, t3)_III	MNO data* (n0 | l1, o2, t3)_III
step 2	aggregation	t: 3→2	MNO data* (n0 | l1, o2, t3)_III	MNO data** (n0 | l1, o2, t2)_III
step 3	model	Create OD matrix	admin data (b0, d0, o0 | p0, t2)_X	admin-based OD (p0 | d0, o0, t2)_I
step 4	aggregation	o: 0→1	admin-based OD (p0 | d0, o0, t2)_I	admin-based OD* (p0 | d0, o1, t2)_I
step 5	aggregation	d: 0→1	admin-based OD (p0 | d0, o0, t2)_I	admin-based OD* (p0 | d1, o0, t2)_I
step 6	aggregation	d: 0→1	admin-based OD* (p0 | d0, o1, t2)_I	admin-based OD** (p0 | d1, o1, t2)_I
step 7	aggregation	o: 0→1	Census (b0, p0 | d0, o0)_I	Census* (b0, p0 | d0, o1)_I
step 8	aggregation	d: 0→1	Census (b0, p0 | d0, o0)_I	Census* (b0, p0 | d1, o0)_I
step 9	model	Location estimation (crude) o: 2→1	MNO data** (n0 | l1, o2, t2)_III	MNO data*** (n0 | l1, o1, t2)_III
step 10	model	Calibration Sim to Person model	MNO data (sim) (n0 | l1, o1, t2)_Y	MNO data (persons) (p0 | l1, o1, t2)_I
			expected persons (p0 | o1, t2)_X	
step 11	subset	remove variables or units	MNO data (persons) (p0 | l1, o1, t2)_I	Commuters location all providers per day-part (p0 | l1, t2)_I

Figure 18: Screenshot a of the user-friendly notebook for metadata analysis: if a path is found from the available input data sets to the target output, it is displayed.

For users interested in analysing the pre-defined commuters application scenario, we recommend starting with the notebook case_essnet_user_friendly.ipynb. The application scenario on commuters is defined in the notebook case_essnet.ipynb. Before changing the case_essnet.ipynb notebook however, we recommend starting with the notebook examples.ipynb. It contains examples and explanations of the most important concepts of the implementation.

7.4. Enabling a better linkage of data sources with questions in a survey

While metadata analysis supports technical integration, surveys are essential for calibration and improving the representativity of MNO data.

Within MNO-MINDS (WP4), a dedicated survey has been designed to capture generic aspects of MNO data. Depending on the data sources combined with MNO data in a specific statistical domain, this survey can be adapted by adding tailored questions to fill information gaps and improve data integration.

Key complementary data sources include:

- Census
- Population Registers, combined Survey/Register data
- Transportation Surveys
- Land Use and Land Cover registers
- Tourism Accommodation statistics
- Satellite data
- Credit Card transaction data
- Tourism platform data

While some sources (e.g., Census, Transportation Surveys) already include the necessary questions, others may require additional questions to ensure full calibration with MNO data.

Below is a harmonised table with *data source – purpose – example additional questions*, which can be adapted for each specific integration scenario. The table provides a flexible framework for designing survey additions tailored to the data sources combined with MNO data. Each question can be selected or adapted according to the statistical domain and integration scenario to improve representativity, alignment, and temporal consistency.

Table 9: Example questions & purpose to enable better data integration

Data source	**Purpose of additional questions**	**Example questions (to adapt as needed)**
Population register / Combined survey & Register data	Address lags, discrepancies, and behavioural patterns that affect MNO data representativity	• Is your current address your officially registered residence? (Yes/No/Partially) • How long have you lived at your current address? • Do you have a secondary residence? How many nights per month are spent

		away from your main address? • Have all household members updated their registration? If not, why? • Do you use your mobile phone mostly at your registered address or elsewhere?
Transportation surveys	Improve calibration of trips, transportation modes, and temporal patterns inferred from MNO data	• How many distinct trips do you usually make per day? • At what times do you start your first trip and return from your last? • What is the main purpose of your most frequent trip (work, education, shopping, etc.)? • Do you regularly use more than one transport mode for a trip? • Do you usually carry your phone during all trips and is it connected to the network? • Do you work remotely or cross borders frequently?
Tourism accommodation statistics	Distinguish residents, tourists, and business travellers, and align stay duration and accommodation types with MNO data	• What is the main purpose of your trip (leisure, business, visiting family)? • Is this part of a multi-destination trip? • What type of accommodation are you staying in (hotel, rental, family, other)? Is it officially registered? • How many nights will you stay, and how many people are travelling with you? • Are you using your home SIM card or a local SIM during this trip?
Land use & Land cover registers	Align land use data with actual patterns captured in MNO data	• What is the main type of land use around your home (residential, commercial, agricultural, mixed, other)? • Has the land use around

		your home changed in the past five years?
Satellite data	Validate night-time light data as a proxy for human activity and development	• Do you spend time outdoors during the evening or night? • Is electricity consistently available where you live (Yes/No/Intermittent)? • How often do you travel outside your local area?
Credit card transaction data	Address under-representation of cash users and informal economies	• Do you own or regularly use a debit/credit card? • How often do you make purchases in cash (Always/Sometimes/Never)? • Where do you usually make your purchases (shops, online, informal vendors)?
Tourist platform data	Calibrate platform-based tourism estimates	• Do you use travel platforms such as Airbnb or Booking.com? • In the past 12 months, how many times have you stayed in paid accommodation away from home? • What type of accommodation do you usually choose when travelling?

8. Broad Overview

This section broadly summarises the main limitations, advantages, and possibilities of each scoring category. Even though some general conclusions can be drawn, the relevance of different aspects still may vary among data sources within the same category.

	Most promising	Promising	Less promising
Key Limitations	• sensitivity	• technical and financial cost • compliance with quality criteria • may require specific skills	• lack of potential target statistics • availability and accessibility • quality

Key Advantages	• (mostly) comply with quality criteria • are well-known to statisticians	• contain complementary and rich information	• additional information
Key Possibilities	• Improve precision and provide timely and granular statistics, broaden scope of statistics	• Improve precision and potentially broaden the scope of statistics	• Potential for research

Table 10: Broad overview on key limitations, advantages and possibilities

Considering the whole landscape analysis, the short list of data sources comprise those categorized as most promising and promising:

- Census
- Population registers
- National travel surveys
- Land use and land cover register
- Tourism accommodation statistics

- Vehicle, bicycle and pedestrian sensors
- Pollution data
- Satellite data
- Electronic invoices
- Tourism surveys
- Tourism platform data
- Credit card transaction data

The full list of data sources and all limitations and possibilities outlined in the detailed data scoring are found in chapter 5.

9. Conclusion

This Work Package has widely explored non-MNO data sources, potential target statistics, the importance of metadata-analysis and several elements relevant for data integration. It provides a theoretical background to identify an assess non-MNO data sources and delivers a detailed scoring for each source. Survey results regarding data source availability in eleven European countries offer insights on the current access situation and on important aspects for a potential implementation in official statistics, e.g. mode of access, granularity, and frequency of data delivery. It consolidates main results in different sections of this deliverable, e.g. as data scoring in chapter 5, as survey results in chapter 6, as connection to target statistics and a survey in chapter 7, and as a broad overview in chapter 8. Deeper insights on specific data linkage limitations and possibilities can be achieved once NSIs commonly apply MNO data together with non-MNO data sources in concrete applications working towards the implementation in official statistics.

10. Bibliography

[Anonymous]. (2004, February). ESA/inmarsat agreement to improve satellite mobile phone and data services. *ESA BULLETIN-EUROPEAN SPACE AGENCY*, 89.

Adams, M. W., Sutherland, E. G., Eckert, E. L., Saalim, K., & Reithinger, R. (2022, May). Leaving no one behind: targeting mobile and migrant populations with health interventions for disease elimination-a descriptive systematic review. *BMC MEDICINE, 20*. doi:10.1186/s12916-022-02365-6

Agustan, Yulianto, S., Sumargana, L., Sadmono, H., & Alhasanah, F. (2018). Innovation on Geolocation and Pattern Recognition for Paddy Growth Stages Reporting in Indonesia. *3RD INTERNATIONAL CONFERENCE OF INDONESIA SOCIETY FOR REMOTE SENSING (ICOIRS 2017). 165.* DIRAC HOUSE, TEMPLE BACK, BRISTOL BS1 6BE, ENGLAND: IOP PUBLISHING LTD. doi:10.1088/1755-1315/165/1/012001

Albanna, B., & Heeks, R. (2019, January). Positive deviance, big data, and development: A systematic literature review. *ELECTRONIC JOURNAL OF INFORMATION SYSTEMS IN DEVELOPING COUNTRIES, 85*. doi:10.1002/isd2.12063

Altin, L., Ahas, R., Silm, S., & Saluveer, E. (2021). Megastar concerts in tourism; a study using mobile phone data. *Scandinavian Journal of Hospitality and Tourism*, DOI: 10.1080/15022250.2021.1936625.

Aydogdu, B., Balcik, C., Gunes, S., Momeni, R., & Salah, A. A. (2023). Fine-grained mapping of migrants in Istanbul using satellite imaging and mobile phone data. *2023 31ST SIGNAL PROCESSING AND COMMUNICATIONS APPLICATIONS CONFERENCE, SIU.* 345 E 47TH ST, NEW YORK, NY 10017 USA: IEEE. doi:10.1109/SIU59756.2023.10223985

Babkin, R. A. (2021). The experience of using the mobile phone data in economic geographical researches in foreign. *VESTNIK OF SAINT PETERSBURG UNIVERSITY EARTH SCIENCES, 66*. doi:10.21638/spbu07.2021.301

Beaven, Z., & Laws, C. (2007). 'Never let me down again': Loyal customer attitudes towards ticket distribution channels for live music events: A netnographic exploration of the us leg of the depeche mode 2005–2006 world tour. *Managing Leisure, 12(2-3)*, 120-142. https://doi.org/10.1080/13606710701339322.

Botta, F., Moat, H. S., & Preis, T. (2015). Quantifying crowd size with mobile phone and Twitter data. *Royal Society open science, 2(5), 150162.* , http://dx.doi.org/10.1098/rsos.150162.

Cao, W., Dong, L., Wu, L., & Liu, Y. (2020, May). Quantifying urban areas with multi-source data based on percolation theory. *REMOTE SENSING OF ENVIRONMENT, 241*. doi:10.1016/j.rse.2020.111730

Chang, S., Wang, Z., Mao, D., Liu, F., Lai, L., & Yu, H. (2021, November). Identifying Urban Functional Areas in China's Changchun City from Sentinel-2 Images and Social Sensing Data. *REMOTE SENSING, 13*. doi:10.3390/rs13224512

Chen, Z., Li, P., Jin, Y., Jin, Y., Chen, J., Li, W., . . . Zhang, H. (2022, September). Using mobile phone big data to identify inequity of artificial light at night exposure: A case study in Tokyo. *CITIES, 128*. doi:10.1016/j.cities.2022.103803

Chi, G., Fang, H., Chatterjee, S., & Blumenstock, J. E. (2022, January). Microestimates of wealth for all low- and middle-income countries. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, 119*. doi:10.1073/pnas.2113658119

Clarke, K. M., & Kendall, S. (2019). `The beauty...is that it speaks for itself': geospatial materials as evidentiary matters. *LAW TEXT CULTURE, 23*, 91+.

CoP, E. (2017). *European Statistics Code of Practice.* Eurostat, European Satistical System.

Corches, C., Stan, O., Miclea, L., & Daraban, M. (2019). Embedded RTOS for a Smart RFID Reader. *2019 IEEE 25TH INTERNATIONAL SYMPOSIUM FOR DESIGN AND TECHNOLOGY IN ELECTRONIC PACKAGING (SIITME 2019)* (pp. 220-223). 345 E 47TH ST, NEW YORK, NY 10017 USA: IEEE. doi:10.1109/siitme47687.2019.8990817

Coudin, E., Poulhes, M., & Suarez-Castillo, M. (2021). The French official statistics strategy : Combining signaling data from various mobile network operators for documenting COVID-19 crisis effects on population movements and economic outlook. *Data and Policy*.

Dixon, J., Elders, I., & Bell, K. (2020, June). Evaluating the likely temporal variation in electric vehicle charging demand at popular amenities using smartphone locational data. *IET INTELLIGENT TRANSPORT SYSTEMS, 14*, 504-510. doi:10.1049/iet-its.2019.0351

Enescu, F. M., & Bizon, N. (2017). SCADA Applications for Electric Power System. In N. M. Tabatabaei, A. J. Aghbolaghi, N. Bizon, & F. Blaabjerg (Eds.), *REACTIVE POWER CONTROL IN AC POWER SYSTEMS: FUNDAMENTALS AND CURRENT ISSUES* (pp. 561-609). HEIDELBERGER PLATZ 3, D-14197 BERLIN, GERMANY: SPRINGER-VERLAG BERLIN. doi:10.1007/978-3-319-51118-4_{1}{5}

ESAC. (2018). New perspectives and priorities for EU 2030 Indicators – Indicators and methodologies for describing society in the Information Age. *ESAC - Sapienza Univeristy of Rome.*

ESS. (2021). Access to privately held data is urgently needed for producing new, faster, more detailed official statistics. *European Statistical System (ESS) position paper on the future Data Act proposal*.

Fishman, R., Ghosh, M., Mishra, A., Shomrat, S., Laks, M., Mayer, R., . . . Shacham-Diamand, Y. (2020). Digital Villages: A Data-Driven Approach to Precision Agriculture in Small Farms. In N. Ansari, A. Ahrens, & C. BenaventePeces (Ed.), *PROCEEDINGS OF THE 9TH INTERNATIONAL CONFERENCE ON SENSOR NETWORKS (SENSORNETS)* (pp. 161-166). AV D MANUELL, 27A 2 ESQ, SETUBAL, 2910-595, PORTUGAL: SCITEPRESS. doi:10.5220/0009373101610166

Goldstein, J. E., & Faxon, H. O. (2022, March). New data infrastructures for environmental monitoring in Myanmar: Is digital transparency good for governance? *ENVIRONMENT AND PLANNING E-NATURE AND SPACE, 5*, 39-59. doi:10.1177/2514848620943892

Google. (2024). Retrieved from Popular times, wait times and, visit duration: https://support.google.com/business/answer/6263531?hl=en

Hu, R. M., Liu, W. M., Wang, S., Zhang, X. Z., & Li, Y. (2017). The application of mobile GIS in mine land reclamation monitoring. In Z. Hu (Ed.), *LAND RECLAMATION IN ECOLOGICAL FRAGILE AREAS* (pp. 121-125). PO BOX 11320, LEIDEN, 2301 EH, NETHERLANDS: CRC PRESS-BALKEMA.

Jia, Y., Ge, Y., Ling, F., Guo, X., Wang, J., Wang, L., . . . Li, X. (2018, March). Urban Land Use Mapping by Combining Remote Sensing Imagery and Mobile Phone Positioning Data. *REMOTE SENSING, 10*. doi:10.3390/rs10030446

Jiang, W., Meng, Y., Zhang, Y., Wu, J., & Li, X. (2022). Response of Urban Park Visitor Behavior to Water Quality in Beijing. In H. Wu, Y. Liu, J. Li, J. Meng, Q. Guan, X. Song, . . . G. Li (Ed.), *SPATIAL DATA AND INTELLIGENCE, SPATIALDI 2022. 13614*, pp. 231-249. GEWERBESTRASSE 11, CHAM, CH-6330, SWITZERLAND: SPRINGER INTERNATIONAL PUBLISHING AG. doi:10.1007/978-3-031-24521-3_{1}{7}

Kowarik, A., & members, E. (2020). Typification matrix for big data projects. *ESNet Big Data 2 - Grant Agreement Number: 847375-2018-NL-BIGDATA*.

Kowarik, A., Tuoto, T., Di Consiglio, L., Deetjen, G., Brandt, M., Kamali, R., . . . Pichiorri, T. (2025). *Proof of concept of an ad-hoc survey to imrpove MNO data.* cros.ec.europa.eu/mno-minds: Eurostat CROS.

Kumagai, M., Ura, T., Kuroda, Y., & Walker, R. (1998). New AUV designed for lake environment monitoring. *PROCEEDINGS OF THE 2000 INTERNATIONAL SYMPOSIUM ON UNDERWATER TECHNOLOGY* (pp. 78-83). 345 E 47TH ST, NEW YORK, NY 10017 USA: IEEE.

Lam, J. C., Han, Y., Bai, R., Li, V. O., Leong, J., & Maji, K. J. (2020). Household wealth proxies for socio-economic inequality policy studies in China. *DATA & POLICY, 2*. doi:10.1017/dap.2020.4

Li, X., Zhang, M., & Wang, J. (2022, January). The spatio-temporal relationship between land use and population distribution around new intercity railway stations: A case study on the Pearl River Delta region, China. *JOURNAL OF TRANSPORT GEOGRAPHY, 98*. doi:10.1016/j.jtrangeo.2021.103274

Liang, L., Shrestha, R., Ghosh, S., & Webb, P. (2020, November). Using mobile phone data helps estimate community-level food insecurity: Findings from a multi-year panel study in Nepal. *PLOS ONE, 15*. doi:10.1371/journal.pone.0241791

Liddiard, K. C. (2011). Further applications for mosaic pixel FPA technology. In B. F. Andresen, G. F. Fulop, & P. R. Norton (Ed.), *INFRARED TECHNOLOGY AND APPLICATIONS XXXVII. 8012.* 1000 20TH ST, PO BOX 10, BELLINGHAM, WA 98227-0010 USA: SPIE-INT SOC OPTICAL ENGINEERING. doi:10.1117/12.886676

Liu, X., Kang, C., Gong, L., & Liu, Y. (2016, February). Incorporating spatial interaction patterns in classifying and understanding urban land use. *INTERNATIONAL JOURNAL OF*

GEOGRAPHICAL INFORMATION SCIENCE, 30, 334-350. doi:10.1080/13658816.2015.1086923

Loven, L., Peltonen, E., Pandya, A., Leppanen, T., Gilman, E., Pirttikangas, S., & Riekki, J. (2019). Towards EDISON: An edge-native approach to distributed interpolation of environmental data. *2019 28TH INTERNATIONAL CONFERENCE ON COMPUTER COMMUNICATION AND NETWORKS (ICCCN).* 345 E 47TH ST, NEW YORK, NY 10017 USA: IEEE.

Ma, S., Li, S., & Zhang, J. (2023, February). Spatial and deep learning analyses of urban recovery from the impacts of COVID-19. *SCIENTIFIC REPORTS, 13*. doi:10.1038/s41598-023-29189-5

Mahdi, A. J., Tettamanti, T., & Esztergar-Kiss, D. (2023). Modeling the Time Spent at Points of Interest Based on Google Popular Times. *IEEE ACCESS, 11*, 88946-88959. doi:10.1109/ACCESS.2023.3305957

Mamei, M., & Colonna, M. (2015). Estimating Attendance From Cellular Network Data. arXiv preprint arXiv:1504.07385.

MNO, T. F. (2023). Reusing Mobile Network Operator data for Official Statistics: the case for a common methodological framework for the European Statistical System. *Eurostat Position Paper*.

Moehring, M., Keller, B., Schmidt, R., & Dacko, S. (2021, May). Google Popular Times: towards a better understanding of tourist customer patronage behavior. *TOURISM REVIEW, 76*, 553-569. doi:10.1108/TR-10-2018-0152

m-wrzr, G. u. (2017). *Populartimes*. Retrieved from https://github.com/m-wrzr/populartimes

Nalin, A., Simone, A., Lantieri, C., Cappellari, D., Mantegari, G., & Vignali, V. (2024). Application of cell phone data to monitor attendance during motor racing major event. The case of Formula One Gran Prix in Imola. *Case Studies on Transport Policy, 18, 101287. https://doi.org/10.1016/j.cstp.2024.101287*.

No223, E. (2024). European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council amending Regulation (EC) No 223/2009 on European statistics.

Pastor-Escuredo, D., Torres, Y., Martinez-Torres, M., & Zufiria, P. J. (2020, May). Rapid Multi-Dimensional Impact Assessment of Floods. *SUSTAINABILITY, 12*. doi:10.3390/su12104246

Pei, T., Sobolevsky, S., Ratti, C., Shaw, S.-L., Li, T., & Zhou, C. (2014). A new insight into land use classification based on aggregated mobile phone data. *INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE, 28*, 1988-2007. doi:10.1080/13658816.2014.913794

Perez, J. (2016). Music Festivals: A Secondary Market Analysis. *CMC Senior Theses. Paper 1338.* , http://scholarship.claremont.edu/cmc_theses/1338.

Pesonen, H., & Piche, R. (2008). Numerical integration in bayesian positioning. In L. L. Bonilla, M. Moscoso, G. Platero, & J. M. Vega (Ed.), *PROGRESS IN INDUSTRIAL MATHEMATICS AT ECMI*

2006. 12, pp. 908-912. HEIDELBERGER PLATZ 3, D-14197 BERLIN, GERMANY: SPRINGER-VERLAG BERLIN.

Pintér, G., & Felde, I. (2021). ANALYZING THE BEHAVIOR OF SOCCER FANS FROM A MOBILE PHONE NETWORK PERSPECTIVE: EURO 2016, A CASE STUDY. arXiv preprint arXiv:2108.09291.

Rathinam, F., Khatua, S., Siddiqui, Z., Malik, M., Duggal, P., Watson, S., & Vollenweider, X. (2021, September). Using big data for evaluating development outcomes: A systematic map. *CAMPBELL SYSTEMATIC REVIEWS, 17*. doi:10.1002/cl2.1149

Reynolds, M., Kropff, M., Crossa, J., Koo, J., Kruseman, G., Milan, A. M., . . . Vadez, V. (2018, December). Role of Modelling in International Crop Research: Overview and Some Case Studies. *AGRONOMY-BASEL, 8*. doi:10.3390/agronomy8120291

Sakarovitch, B., Bellefon, M.-P., Givord, P., & Vanhoof, M. (2019). Estimating the Residential Population from Mobile Phone Data, an Initial Exploration. *Economics and Statistics*.

Santiago-Iglesias, E., Carpio-Pinedo, J., Sun, W., & Garcia-Palomares, J. C. (2023, September). Frozen city: Analysing the disruption and resilience of urban activities during a heavy snowfall event using Google Popular Times. *URBAN CLIMATE, 51*. doi:10.1016/j.uclim.2023.101644

Schnebele, E., Oxendine, C., Cervone, G., Ferreira, C. M., & Waters, N. (2015). Using Non-authoritative Sources During Emergencies in Urban Areas. In M. Helbich, J. J. Arsanjani, & M. Leitner (Eds.), *COMPUTATIONAL APPROACHES FOR URBAN ENVIRONMENTS* (Vol. 13, pp. 337-361). 233 SPRING STREET, NEW YORK, NY 10013, UNITED STATES: SPRINGER. doi:10.1007/978-3-319-11469-9_{1}{4}

Shi, Y., Qi, Z., Liu, X., Niu, N., & Zhang, H. (2019, November). Urban Land Use and Land Cover Classification Using Multisource Remote Sensing Images and Social Media Data. *REMOTE SENSING, 11*. doi:10.3390/rs11222719

Shi, Y., Yang, J., & Shen, P. (2020, January). Revealing the Correlation between Population Density and the Spatial Distribution of Urban Public Service Facilities with Mobile Phone Data. *ISPRS INTERNATIONAL JOURNAL OF GEO-INFORMATION, 9*. doi:10.3390/ijgi9010038

Simumba, N., Okami, S., Kodaka, A., & Kohtake, N. (2021, April). Spatiotemporal Integration of Mobile, Satellite, and Public Geospatial Data for Enhanced Credit Scoring. *SYMMETRY-BASEL, 13*. doi:10.3390/sym13040575

Solanellas, F., Muñoz, J., & Petchamé, J. (2022). An Examination of Ticket Pricing in a Multidisciplinary Sports Mega-Event. . *Economies, 10(12), 322.* , https://doi.org/10.3390/economies10120322.

Somantri, L. (2021). The Role of GIS and Remote Sensing for Population Mobility Mapping. In S. B. Wibowo, & P. Wicaksono (Ed.), *SEVENTH GEOINFORMATION SCIENCE SYMPOSIUM 2021. 12082.* 1000 20TH ST, PO BOX 10, BELLINGHAM, WA 98227-0010 USA: SPIE-INT SOC OPTICAL ENGINEERING. doi:10.1117/12.2617180

Song, Y., Huang, B., Cai, J., & Chen, B. (2018, September). Dynamic assessments of population exposure to urban greenspace using multi-source big data. *SCIENCE OF THE TOTAL ENVIRONMENT, 634*, 1315-1325. doi:10.1016/j.scitotenv.2018.04.061

Steele, J. E., Sundsoy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., . . . Bengtsson, L. (2017, February). Mapping poverty using mobile phone and satellite data. *JOURNAL OF THE ROYAL SOCIETY INTERFACE, 14*. doi:10.1098/rsif.2016.0690

Tatem, A. J., Huang, Z., Narib, C., Kumar, U., Kandula, D., Pindolia, D. K., . . . Lourenco, C. (2014, February). Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning. *MALARIA JOURNAL, 13*. doi:10.1186/1475-2875-13-52

Thompson, D. (2025). The evolution of ticket pricing strategies in the North American concert industry: evidence from two decades of data. *Applied Economics*, DOI: 10.1080/00036846.2025.2464817.

Timokhin, S., Sadrani, M., & Antoniou, C. (2020, September). Predicting Venue Popularity Using Crowd-Sourced and Passive Sensor Data. *SMART CITIES, 3*, 818-841. doi:10.3390/smartcities3030042

Tu, W., Hu, Z., Li, L., Cao, J., Jiang, J., Li, Q., & Li, Q. (2018, January). Portraying Urban Functional Zones by Coupling Remote Sensing Imagery and Human Sensing Data. *REMOTE SENSING, 10*. doi:10.3390/rs10010141

Tu, W., Zhang, Y., Li, Q., Mai, K., & Cao, J. (2021, January). Scale Effect on Fusing Remote Sensing and Human Sensing to Portray Urban Functions. *IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, 18*, 38-42. doi:10.1109/LGRS.2020.2965247

Wang, Y., Li, Q., Luo, Z., Zhao, J., Lv, Z., Deng, Q., . . . He, K. (2023, December). Ultra-high-resolution mapping of ambient fine particulate matter to estimate human exposure in Beijing. *COMMUNICATIONS EARTH & ENVIRONMENT, 4*. doi:10.1038/s43247-023-01119-3

Xavier, F. H., Silveira, L. M., Almeida, J. M., Malab, C. H., & Marques-Neto, H. (2012). Analyzing the Workload Dynamics of a Mobile Phone Network in Large Scale Events. *Proceedings of the first workshop on Urban networking (UrbaNE '12). Association for Computing Machinery, New York, NY, USA*, 37–42. https://doi.org/10.1145/2413236.2413245.

Xiaomeng, C., Guozhen, L., Yang, Y., & Qingquan, L. (2014). Estimating the distribution of economy activity: a case study in Jiangsu Province (China) using large scale social network data. In Z. H. Zhou, W. Wang, R. Kumar, H. Toivonen, J. Pei, J. Z. Huang, & X. Wu (Ed.), *2014 IEEE INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOP (ICDMW)* (pp. 1126-1134). 345 E 47TH ST, NEW YORK, NY 10017 USA: IEEE. doi:10.1109/ICDMW.2014.145

Xu, J., Fader, P., & Veeraraghavan, S. (2015). Evaluating the Effectiveness of Dynamic Pricing Strategies on MLB Single-Game Ticket Revenue. *2015 MIT Sloan Sports Analytics Conference*.

Yabe, T., Rao, P. S., Ukkusuri, S., & Cutter, S. L. (2022, February). Toward data-driven, dynamical complex systems approaches to disaster resilience. *PROCEEDINGS OF THE NATIONAL*

ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, 119. doi:10.1073/pnas.2111997119

Yang, Y., Wang, H., Qin, S., Li, X., Zhu, Y., & Wang, Y. (2022, December). Analysis of Urban Vitality in Nanjing Based on a Plot Boundary-Based Neural Network Weighted Regression Model. *ISPRS INTERNATIONAL JOURNAL OF GEO-INFORMATION, 11*. doi:10.3390/ijgi11120624

Yoneki, E., & Crowcroft, J. (2014, February). EpiMap: Towards quantifying contact networks for understanding epidemiology in developing countries. *AD HOC NETWORKS, 13*, 83-93. doi:10.1016/j.adhoc.2012.06.003

Zhang, G., Rui, X., Poslad, S., Song, X., Fan, Y., & Wu, B. (2020, August). A Method for the Estimation of Finely-Grained Temporal Spatial Human Population Density Distributions Based on Cell Phone Call Detail Records. *REMOTE SENSING, 12*. doi:10.3390/rs12162572

Zhang, Haug, Fosen, Consiglio, D., D'Orazio, Faricelli, . . . Oancea. (2025). *Report on methodologies.* cros.ec.europa.eu/mno-minds: Eurostat CROS.

Zhang, L.-C., Haraldsen, G., Pekarskaya, T., & Hole, B. (2018). Non-survey big data for official statistics:Sources, usability and statistical design .

Zulkarnain, F., Manessa, M. D., Suseno, W., Ardiansyah, Bakhtiar, R., Safaryanto, A. N., . . . Rokhmatuloh. (2019). People in pixels: developing remote sensing-based geodemographic estimation through volunteered geographic information and crowdsourcing. In T. Erbertseder, N. Chrysoulakis, Y. Zhang, & F. Baier (Ed.), *REMOTE SENSING TECHNOLOGIES AND APPLICATIONS IN URBAN ENVIRONMENTS IV. 11157.* 1000 20TH ST, PO BOX 10, BELLINGHAM, WA 98227-0010 USA: SPIE-INT SOC OPTICAL ENGINEERING. doi:10.1117/12.2533230

11. Annex

11.1. Main variables and useful links

Census and Micro census

Main Variable of German Census
Main variables are:
- Age
- Sex
- Marital Status
- Education (class level, highest school-leaving/vocational/professional qualification, ...)
- Employment & Occupation (activity status, duration, gainful activity by occupation, status in employment, ...)
- Citizenship
- Commuters
- Country of birth
- Migration (migrant background, migration experience, immigration history)
- Religion
- Buildings (type, year of construction, type of heating, energy source used for heating, ...)
- Dwellings (number, floor area, rent, number of rooms, equipment, ...)
- Dwellings rental information (rent, ownership, reason for and duration of dwelling vacancy)

Main Variable of French Census
- gender, age
- occupation
- nationality
- mode of transport
- inhabitants' homes: type of dwelling, type of construction, number of rooms, etc.

Main Variable of German Microcensus
- Information on the household (e.g. household size) and the person (e.g. gender, year of birth, nationality)
- Living expenses, income
- Childcare, school, universities
- Training and further education
- Employment, occupation, job search
- Retirement provision
- Internet usage
- Living situation

SE_TPR: https://metadata.scb.se/mikrodataregister.aspx?produkt=BE0102
FR_Census: https://www.insee.fr/en/metadonnees/source/serie/s1321
DE_Census database: https://ergebnisse.zensus2022.de/datenbank/online
DE_Census info: https://www.zensus2022.de/EN/How-does-the-census-work/_node.html#_nuo3ntz9p
Micro data access: https://forschungsdatenzentrum.de/en
ESS Census info: https://ec.europa.eu/CensusHub2/selectHyperCube?countrycode=en&clearSession=true

Microcensus quality report:
https://www.destatis.de/DE/Methoden/Qualitaet/Qualitaetsberichte/Bevoelkerung/mikrozensus-2021.html

Main Variables of the Swedish Population Register

* Social security number, sex, age
* Name, address
* Civil registration conditions
* Marital status, change of marital status
* Citizenship
* Country of birth
* Foreign background/Swedish background
* Information about births and deaths
* Domestic relocation
* Immigration/emigration
* Relationships (husband/wife, registered partner, biological parents, adoptive parents, caregiver)
* Basis for residence (for persons who have been granted a residence permit or have received the right of residence in Sweden)

Tourism Household Survey

Eurostat, Trips of EU residents, metadata:
https://ec.europa.eu/eurostat/cache/metadata/en/tour_dem_esms.htm

Quality report on travel behaviour survey (Destatis, DE):
https://www.destatis.de/DE/Methoden/Qualitaet/Qualitaetsberichte/Gastgewerbe-Tourismus/tourismus-reiseverhalten.pdf?__blob=publicationFile

Tourism Border Survey

Eurostat (2015). Methodological manual for tourism statistics - 2014, v.3.1. [Available at: https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-gq-14-013]

Border Survey, Spain:
https://www.ine.es/dyngs/INEbase/es/categoria.htm?c=Estadistica_P&cid=1254735576863
(general)

https://www.ine.es/dynt3/metadatos/es/RespuestaDatos.html?oe=16028
(methodological note on the border survey)

Tourism Platform Data

Destatis, Exp. Statistics: https://www.destatis.de/DE/Service/EXSTAT/Datensaetze/buchung-online-unterkuenfte.html

Eurostat Method. Note: https://ec.europa.eu/eurostat/documents/7894008/12961561/CETOUR-Methodological-note.pdf/1dee049f-5612-1b47-c7ce-75eacaf49790?t=1624886311053

Eurostat Database:
https://ec.europa.eu/eurostat/databrowser/view/tour_ce_omr/default/table?lang=en&category=tour.tour_ce.tour_ce_om

EU regulation 2024/2018 on data relating short-term accomodation: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401028

Social Media

https://developer.twitter.com/en/docs/twitter-api ;
https://research.facebook.com/blog/2021/03/new-analytics-api-for-researchers-studying-facebook-page-data/

Variables of Portuguese electronic invoices

E-Invoices (V_TF_EFAT_ENCRIPTADA_AAAA)

Atributo	Tipo	Descritivo	EN
ANO	TEXTO (4)	Ano de emissão de fatura	Year of invoice issuance
MES	TEXTO (2)	Mês de emissão de fatura	Month of invoice issuance
VERSAO	NUMERO	Versão dos dados (relativo ao ano, mês)	Data version (relative to year, month)
ID_SEQ	NUMERO	Número sequencial de registo (relativo ao ano, mês)	Sequential registration number (relative to year, month)
NIF_EMITENTE	TEXTO (9)	Número de Identificação Fiscal da entidade (singular ou coletiva) que emitiu fatura	Tax identification number of the entity (individual or collective) that issued the invoice
NIF_ADQUIRENTE_NAC_COL	TEXTO (9)	Número de Identificação Fiscal da entidade coletiva (e nacional) adquirente	Tax Identification Number of the acquiring collective (and national) entity
NIF_ADQUIRENTE_NAC_SING_ENCR	TEXTO (64)	Número de Identificação Fiscal encriptado da entidade singular adquirente; Inclui também os NIF 999999990 que representam faturação nacional com ausência de NIF, por forma a manter num mesmo atributo os NIF adquirentes singulares	Encrypted Tax Identification Number of the acquiring individual entity; It also includes NIF 999999990 that represent national invoicing without a NIF, in order to maintain the NIF of individual acquirers in the same attribute.
NIF_ADQUIRENTE_ESTR	TEXTO (200)	Número de Identificação Fiscal de entidades adquirentes estrangeiras	Tax Identification Number of foreign acquiring entities
VALOR_TRIBUTAVEL	NUMERO	Valor tributável (correspondente ao valor do adquirente agregado no mês, para um determinado emitente)	Taxable value (corresponding to the aggregate acquirer value in the month, for a given issuer)
TIPO_VALOR_TRIBUTAVEL	TEXTO (1)	Identifica o tipo de valor tributável (por defeito='O', de Original); Descodifica com TD_TIPO_VALOR_TRIBUTAVEL	Identifies the type of taxable value (default='O', for Original); Decode with TD_TIPO_VALOR_TRIBUTAVEL
TIPO_EMITENTE	NUMERO	Identifica se a entidade emitente é do tipo Singular ou Coletivo; Descodifica com tabela TD_TIPO_EMITENTE	Identifies whether the issuing entity is of the Individual or Collective type; Decode with table TD_TIPO_EMITENTE

TIPO_MERCADO	NUMERO	Identifica o tipo de mercado que adquiriu; Descodifica com TD_TIPO_MERCADO	Identify the type of market you acquired; Decode with TD_TIPO_MERCADO
PAIS_DSG_SMI	TEXTO (200)	Designação do país adquirente	Designation of the acquiring country
PAIS_COD_SMI	TEXTO (5)	Código ISO Alpha 2 do país adquirente	ISO Alpha 2 code of the acquiring country
NUTIII_EMITENTE	TEXTO (3)	NUTSIII do Emitente	NUTSIII of the Issuer
STA	TEXTO (2)	Situação perante a atividade do Emitente	Situation regarding the Issuer's activity
CAE3	TEXTO (5)	Código de atividade económica (CAE Rev3) do Emitente	Economic activity code (CAE Rev3) of the Issuer
FJR	TEXTO (3)	Código da forma jurídica do Emitente	Code of the Issuer's legal form
SIN	TEXTO (10)	Código do Setor Institucional (SIN) do Emitente	Institutional Sector Code (SIN) of the Issuer
ZONA_FRANCA	TEXTO (1)	Código de Zona Franca do Emitente	Issuer Free Zone Code
DDCCFF_EMITENTE	TEXTO (6)	Código DDCCFF do Emitente	Issuer Code DDCCFF (concatenation of district, municipality and parish)
FONTE_CARACTERIZACAO_EMITENTE	TEXTO (2)	Identifica a fonte de dados usada para caracterização do Emitente; Descodifica com TD_FONTE_CARACTERIZACAO	Identifies the data source used to characterize the Issuer; Decode with TD_FONTE_CARACTERIZACAO
CLASSE_ADQUIRENTE	TEXTO (2)	Tipifica o Adquirente, com base no seu NIF e origem; Descodifica com TD_CLASSE_ADQUIRENTE	Types the Purchaser, based on their NIF and origin; Decode with TD_CLASSE_ACQUIRENTE
DTCCFF_ADQUIRENTE	TEXTO (6)	Código DDCCFF do Adquirente	Purchaser Code DDCCFF (concatenation of district, municipality and parish)
FONTE_CARACTERIZACAO_ADQUIRENTE	TEXTO (2)	Identifica a fonte de dados usada para caracterização do Adquirente; Descodifica com TD_FONTE_CARACTERIZACAO	Identifies the data source used to characterize the Acquirer; Decode with TD_FONTE_CARACTERIZACAO
NUTIII_2024_EMITENTE	TEXTO (3)	NUTSIII (versão 2024) do Emitente	NUTSIII (version 2024) of the Issuer

Table 11: Variables of Portuguese electronic invoices

11.2. Parameters of the bibliometric analysis – Satellite data

Data consists from bibliometric entries recorded by Web of Science (Clarivate, 2024) in the following database field values: title, authors, authors keywords, abstract, affiliations, year, journal, keywords. We used the query, which was refined over multiple iterations:

(mobile NEAR/5 phone NEAR/5 data) and ((remote NEAR/2 sensing) or (satellite NEAR/4 data)) (Title) and

(mobile NEAR/5 phone NEAR/5 data) and ((remote NEAR/2 sensing) or (satellite NEAR/4 data)) (Author Keywords) and

(mobile NEAR/5 phone NEAR/5 data) and ((remote NEAR/2 sensing) or (satellite NEAR/4 data)) (Abstract)

which translates into retrieval for papers which contain mobile phone data and remote sensing in the paper's title, author keywords and abstract. The keywords in an expression of the type 'mobile phone data' may contain, from 0 to 5, other words between them, e.g. 'mobile phone network data', respectively between 0 to 4 words in 'remote sensing' expression. The query was refined by inspecting each paper retrieved on each iteration. In the last iteration (the above query) all papers retrieved address, in some specific form or another, the combined use of mobile phone data (either from network events, or crowdsourced through installed application, i.e. collected through INTERNET) and remote sensing data. Last update of the data extracted was carried out on 10th of July 2024.

From bibliometrix package we employed some basic summary statistics and some form of topic related aggregation across research domain and time, from simple types of aggregation such as number of papers published each year, to more complex types such as thematic maps of trends (Cobo et al. 2012).

Results of bibliometric analysis

Figure 19: Annual scientific production (number of published papers) - Satellite data

In figure 1 we provide the number of papers retrieved and some summary statistics regarding scientific productivity, i.e. 48 papers from 241 authors, published between 1998 and 2023 with an annual growth of 4.5%, registering a drop in past year, from a peak of 8 papers published in 2022 to just 3 in 2023 (figure 19). The drop may be caused by lag between the moment when a paper is accepted in a journal and when it is actually indexed by the database. The number of average citations per paper, which in this case amounts to almost 30 citations, signals some moderate presence of activity and interest, mostly due to novel approaches in integrating different data sources.

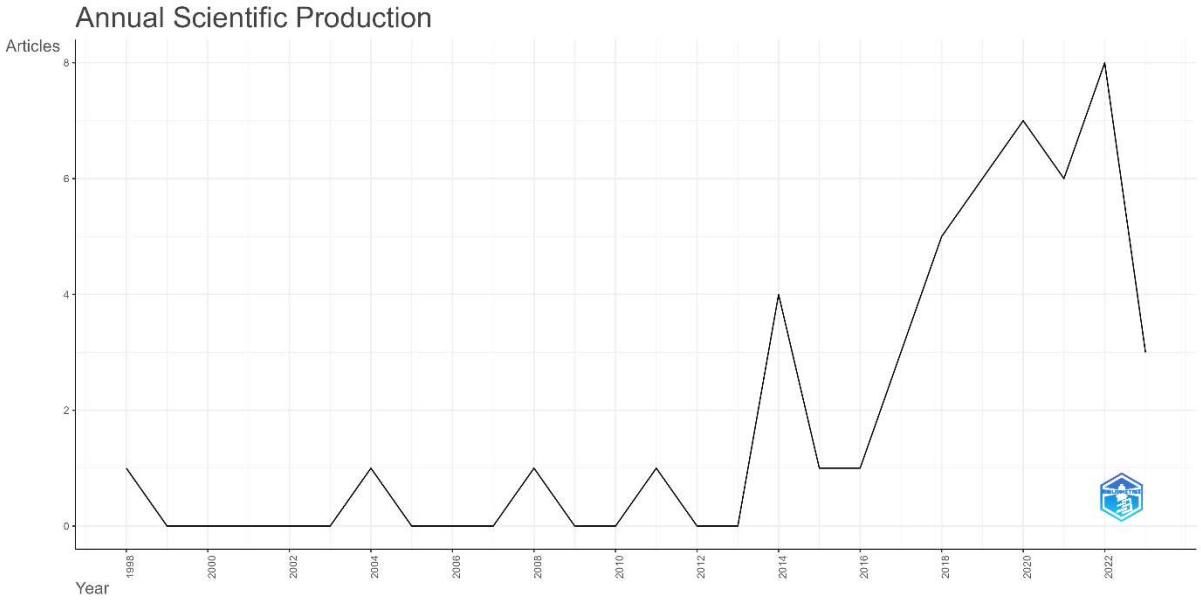


Figure 20: Annual scientific production (number of published papers) - Satellite data

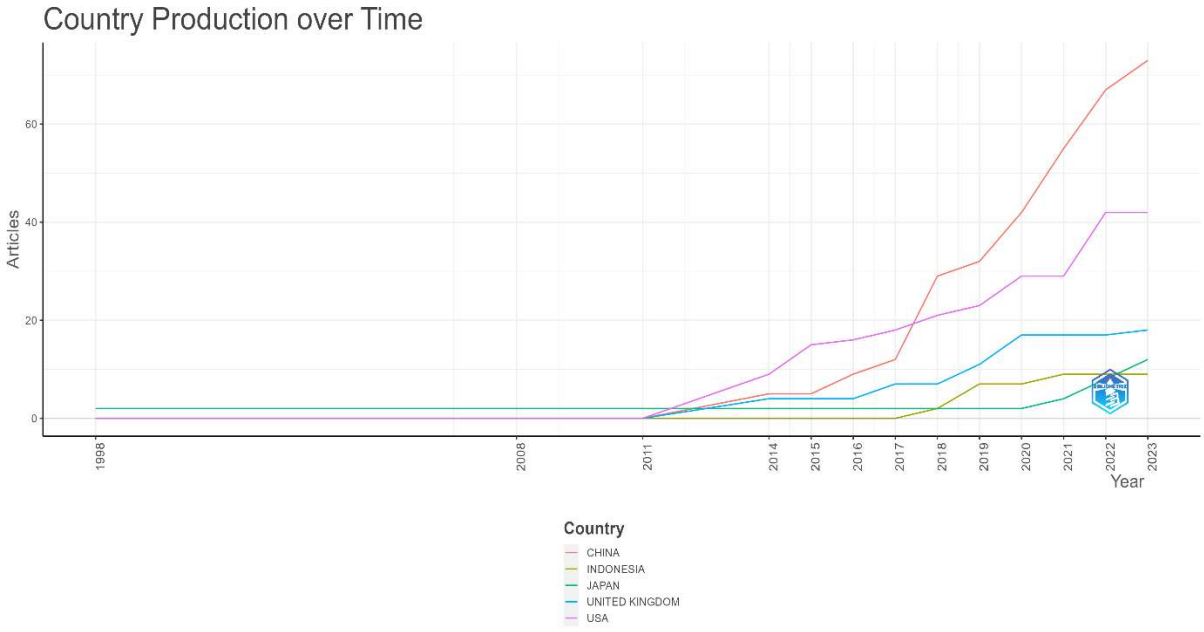


Figure 21: Articles per country over time - Satellite data

(a country may be counted multiple times; depends on the number of authors)

In terms of contributing countries (figure 20) China and United States of America (USA) lead the pack, accounting for more than 80% of published papers.

11.3. Short Short Survey on the availability of data sources

Please find the full survey with Introduction, Survey, Explanations and Definitions in the following.

Dear Participant, by answering this short survey you help us to collect information on the availability of data sources across the ESS.

In the ESSnet Project MNO-MINDS, we are researching the integration of different data sources with Mobile Network Operator (MNO) data and aim to identify the most promising ones to be integrated with MNO data. For this, the EU-wide availability of a certain data source is one evaluation criteria.

In **Part A**, please select for each official data source if you currently have access to it. Next, please indicate if you have had or will gain access.
In **Part B**, please select for each new data source if you currently have access to it. Next, please indicate if you have had or will gain access and if no, please provide more information.
In **Part C**, please provide more information for all new data sources that you currently have access to. For all parts, if there is any source missing that you believe is relevant to be combined with MNO data, please add it at the end of the respective table.

If there is any additional information you wish to share or if you have any remarks, please leave them in the comment box at the end of the survey.

Start survey

Survey on the EU-wide availability of Data Sources as part of the landscaping analysis of data sources to be integrated with Mobile Network Operator (MNO) data in the ESSnet Project MNO-MINDS

More Information

NSI	Please name a reference person from your National Statistical Institute.	
Contact Information	National Statistical Institute:	
	First and Last Name:	
	E-Mail:	

Please note: In the following, you can click on any question or data source to find additional helpful information (links to sheet "Explanations"). To provide an answer, please click on the corresponding cell and select your option from the drop-down menu.

Part A	Data source	In Production / Current Access?	Did you have or will you gain access in the last or next 12 months?	If yes, at which Granularity level?	If yes, at which Geospatial Resolution?
Official Statistics / Official Data	Land Cover and Land Use Register				
	Total Population Register				
	Census				
	National Travel Survey				
	Tourism Household Survey				
	Tourism Border Survey				
	Tourism Accomodation Statistics				
	Please add missing source here (if any)				
	Please add missing source here (if any)				

Part B	Data source	Do you currently have access?	Did you have or will you gain access in the last or next 12 months?	If no, what is the main reason?	If no, can you identify the provider?
New Data Sources	Tourism Platform Data				
	Google Maps Popular Time				
	Social Media				
	Vehicle, Bicycle and Pedestrian Sensors				
	Vessel (boat) Traffic Data				
	Pollution Data				
	Smart Meters				
	Connected Vehicles				
	Satellite Data*				
	Electronic Invoices				
	Credit Card Transaction Data				
	Ticket Data*				
	Please add missing source here (if any)				
	Please add missing source here (if any)				

**If applicable, please specify/describe briefly which kind of data (in the comment box)*

Please note: In Part C, please provide more information for all data sources that you have access to. In case of several valid options, please select the most applicable one.

Part C	Data source	Mode of Access	Privately or publicly held data	Are there any costs?	Granularity	Coverage	Geospatial Resolution	Frequency of Data Delivery	Duration of Access	Period of Time until Access
New Data Sources										

Do you have any further comments? (E.g. specifications, clarifications, further information, …)	
Further comments	

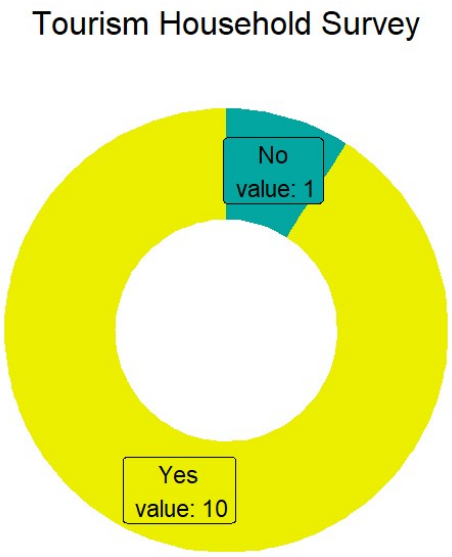
1. Data Source	Explanation	Key Words	Further Links
Land Cover and Land Use Register	Land use and land cover (LULC) registers provide systematic information on Earth's surface coverage and purpose, such as urban areas, agricultural land, and forests.	GIS, surface coverage and purpose	
Total Population Register	The register serves as the foundation for official population statistics and includes information on the population and its changes.	Official Population Statistics	
Census	Census determines the official population number. Whether it is conducted purely by a survey or register-assisted with supplementary surveys, results are derived eventually from extrapolation (unless it is a full survey).	Official Population Statistics	https://ec.europa.eu/CensusHub2/selectHyperCube?countrycode=en&clearSession=true
National Travel Survey	National travel survey collects information on modes of transport, trip purpose, and travel distance.	Mode of Transport, Trip purpose	
Tourism Household Survey	Conducted to capture tourism demand (domestic and outbound tourism) and includes questions on travel behaviour, travels with overnight-stay, domestic day-trips, and socio-demographic information on travellers.	Travel behaviour for domestic and outbound trips	
Tourism Border Survey	Inbound and outbound Tourism can be captured by border surveys, interviewing tourists about their stay at country entry/exist points.	Travel behaviour for inbound and outbound trips at entry/exit points	
Tourism Accomodation Statistics	Monthly Tourism Statistics comprise statistics on capacity and occupancy of tourist accommodation.	nights spent	https://ec.europa.eu/eurostat/documents/3859598/6454997/KS-GQ-14-013-EN-N.pdf/166605aa-c990-40c4-b9f7-59c297154277?t=1420557603000
Tourism Platform Data	Online platform data provides information on (touristic) short-stay accommodation. Examples for such platforms are: Airbnb, Booking, Expedia, and Tripadvisor.	bookings, nights spent	
Google Maps Popular Time	help users understand relative crowd levels at various locations throughout the day. By providing both real-time and historical data on how busy places like restaurants, stores, parks, and other public venues are, it enables users to plan their visits more efficiently.	relative crowd levels at locations, POIs	https://support.google.com/business/answer/6263531?hl=en
Social Media	Data from large social media platforms, e.g. X, TikTok, Facebook and Instagram. It mostly refers to public posts created by users which can come with a variety of variables, e.g. text, location, hashtags, ...	X, Meta, TikTok, Social Media Posts	
Vehicle, Bicycle and Pedestrian Sensors	Static sensors are usually placed in cities and count the number of vehicles, bicycles, and pedestrians for a certain location. They also capture walking and cycling trends.		
Vessel (boat) Traffic Data	Refers to Automatic Identification System (AIS) transponder data which is mandatory for all merchant vessels over 300 gross tons since 2007.	AIS transponder data	
Pollution Data	Static detection usually placed in cities to measure air pollution, e.g. NO_2, PM_{10}, and $PM_{2.5}$	Air pollution, air quality, emissions	
Smart Meters	Smart Meters usually comprise Smart Electricity Meter and Gas/Energy Smart Meter. Data comprises energy usage and energy input of residential and commercial buildings.	Energy Usage, Energy Input	
Connected Vehicles	Connected vehicle data comprise a variety of data, mostly coming from digital sensors and geo-location tracker, and transmitted via mobile network (for external processing and use of data).	Smart Cars, GPS	
Satellite Data	Satellite remote sensing data for Earth observation is defined as measurements of Earth's surface and atmosphere collected via active (e.g., radar) or passive (e.g., optical) sensors mounted on spacecraft, primarily in low Earth orbit.	Copernicus	
Electronic Invoices	Electronic invoices can be defined as a mandatory reporting invoices system implemented by e.g. the Tax Administration as part of the administrative simplification and anti-fraud measures. This administrative data includes all the invoicing recorded electronically by the issuer, whether the acquirer / buyer has requested an invoice from.	Invoices, Tax Administration	
Credit Card Transaction Data	These data consist in Bankcard payment data at the transaction level, dated at the second.	Bankcard payments	
Ticket Data	information on ticket sales, e.g. derived from the news. Other kinds of tickets, e.g. flight tickets or number of passengers / people based on ticket information can be included as well. Please specify which kind of ticket (information) you have access to in the information box.	Big Events, Concerts, Number of Visitors, Traffic Analysis	

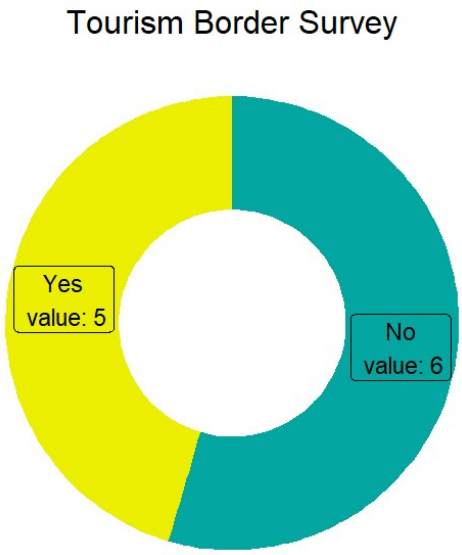
2. Question-specific information	Selection Option	Definition
Production / Current Access	Yes / No	Please indicate if you have access to this data source in any way at your NSI
Did you have or will you gain access in the last or next 12 months?	will gain access, have had access, both, none	Please indicate if you have had or aim to gain access to this data source in the past or next 12 months at your NSI.
If no, what is the main reason?	financial, legal, technical, bureaucratic	If you do not have current access, please indicate the main reason
If no, can you identify the provider?	Yes / No	If you do not have current access, please if you know who is the data holder and where you potentially could request data access.
Mode of Access	partnership	Please select if data sources is access via a partnership or collaboration agreement (with or without financial transactions).
	purchased	Please select this option if data is (simply) purchased (without a collaboration).
	open source	Please select this option if data is open source.
	own source	This option refers to data sources or bases that NSIs hold/produce themselves (e.g. Census)
Privately or publicly held data?	public	Is the data holder public? E.g. Other National Authority, NSI, …
	private	Please select this option if data is held privately.
Are there any costs?	Yes / No	Are there any costs / financial transactions directly related to data access and/or usage?
Granularity	macro	aggregated counts, e.g. per cell or day
	micro	individual counts, e.g. without further information
	nano	This refers to a very high level of granularity, e.g. (nearly) raw data.
Coverage	full	Please select if data source covers the full (or nearly full) population or area
	partly	Please select if data source covers a reasonable part of the full population
	limited	Please select if data source covers only a fraction of the full population
Geospatial Resolution (please select the finest available resolution. Please provide more information in the comment box if you wish to specify)	NUTS-0	Country / Member State level
	NUTS-1	major socio-economic regions
	NUTS-2	basic regions (for regional policies)
	NUTS-3	small regions (for specific diagnoses)
	LAU / municipality level	local administrative units
	grid cells	INSPIRE grid, e.g. 1x1km, 100x100m, …
Frequency of Data Delivery	regular	e.g. hourly, daily, weekly, quarterly, yearly, …
	one-time	one-time data delivery (or unregular data delivery with e.g. years in between)
	project-based	data delivery happens only during a predefined period, e.g. during a specific project and ends when the project ends
Duration of Access	temporary	Is data access restricted to a certain time period?
	permanent	Is data access established permanently?
Period of Time until Access	less than 3 months, between 3-6 months, between 6-12 months, more than 12 months	How much time does / did it take to eventually gain access to this source? We consider the time from requesting / applying data until actually gaining access. If you wish please specify directly in this field or in the comment box on the top.

Thanks to the feedback provided by respondents, we suggest to include “Polygons” as an additional selection option for geospatial resolution.

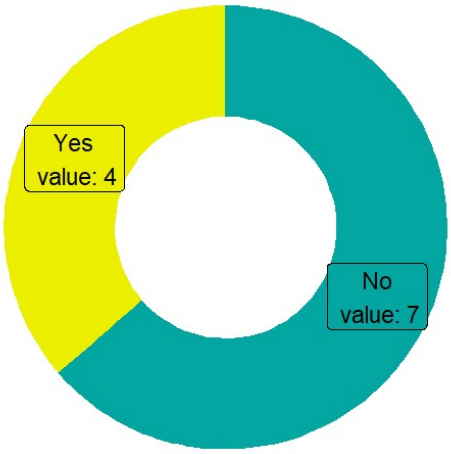
11.4. Survey responses regarding availability of data sources

This annex section includes results from the survey on data source availability for all “promising” and “less promising” sources as well as some general results.

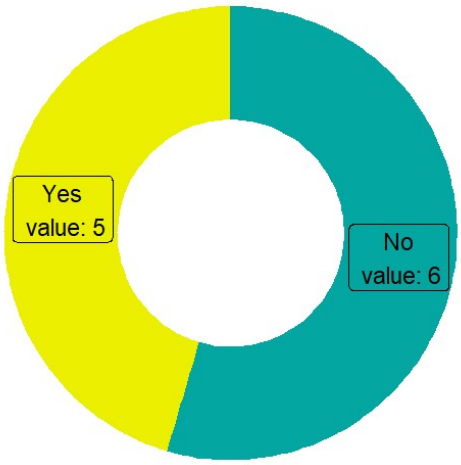




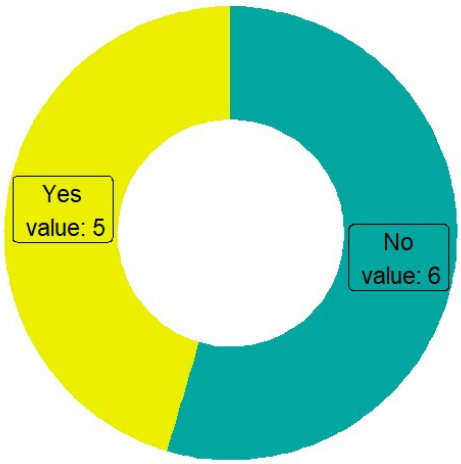
Vessel (boat) Traffic Data



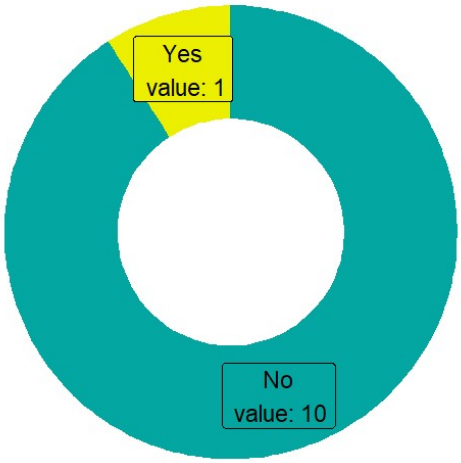
Vehicle, Bicylce and Pedestrian Sensors



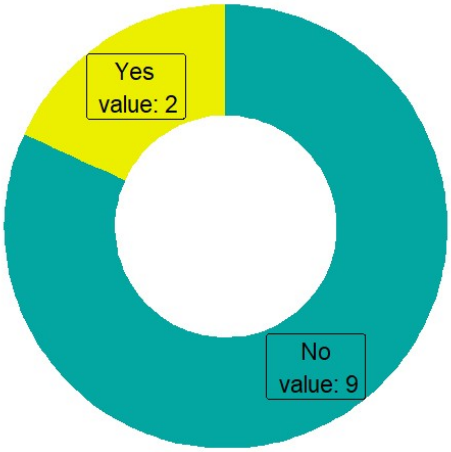
Tourism Platform Data



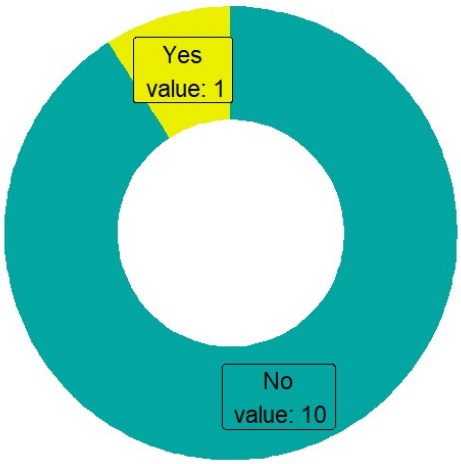
Ticket Data



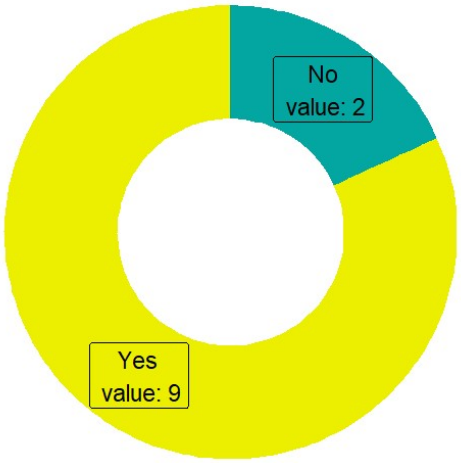
Smart Meters



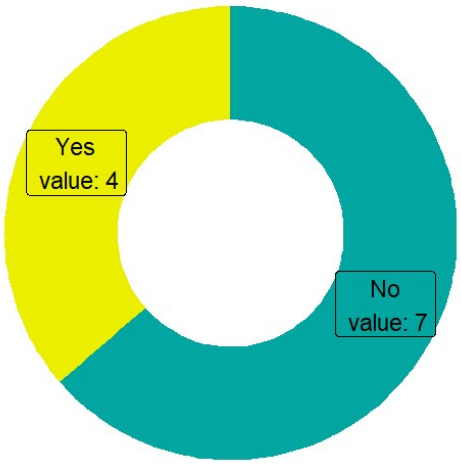
Social Media



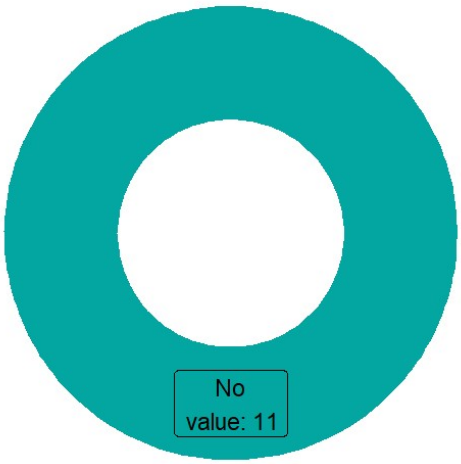
Satellite Data

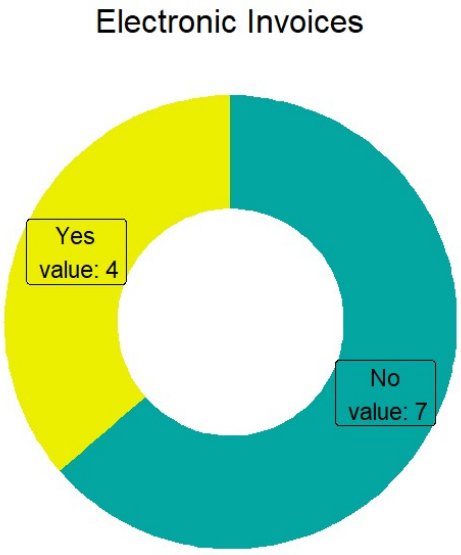


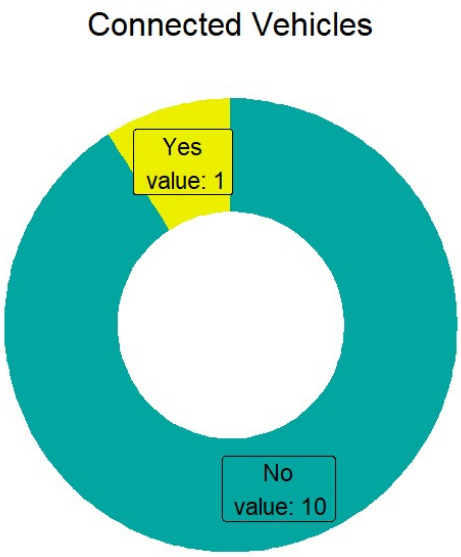
Pollution Data

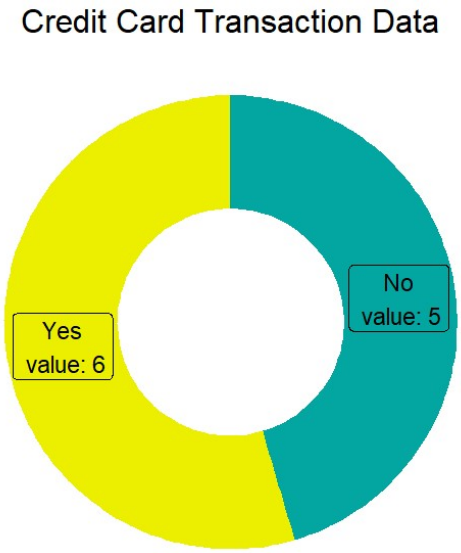


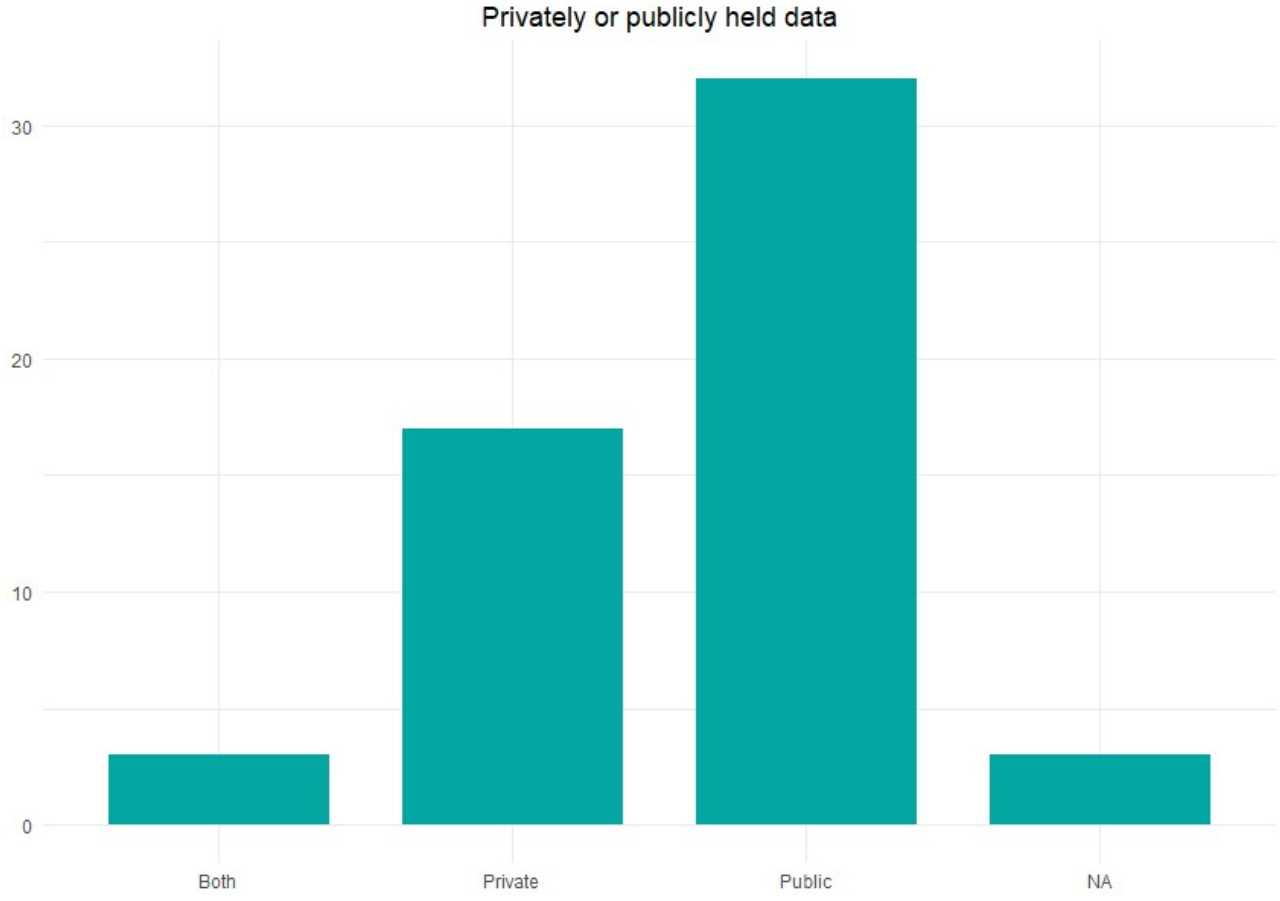
Google Maps Popular Time

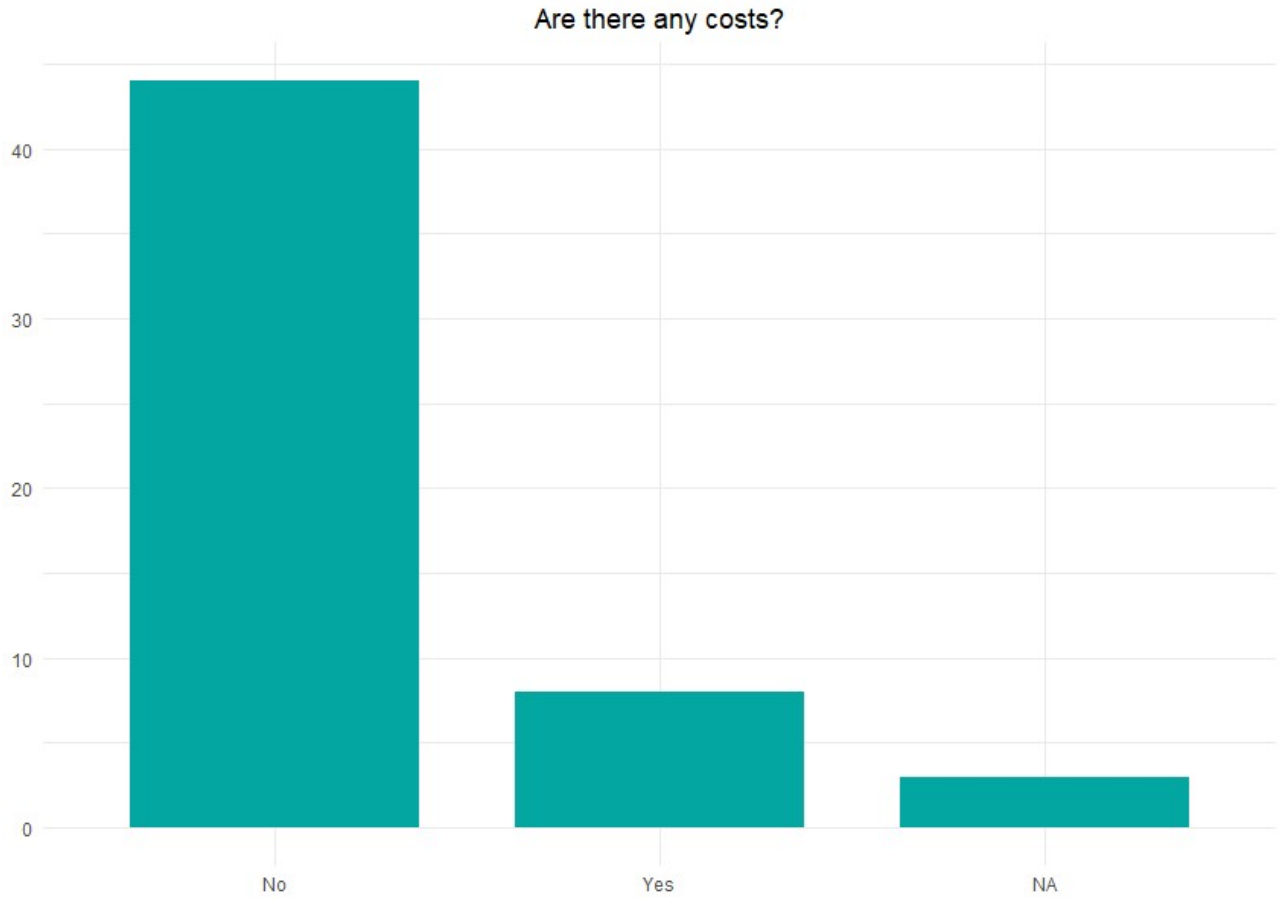


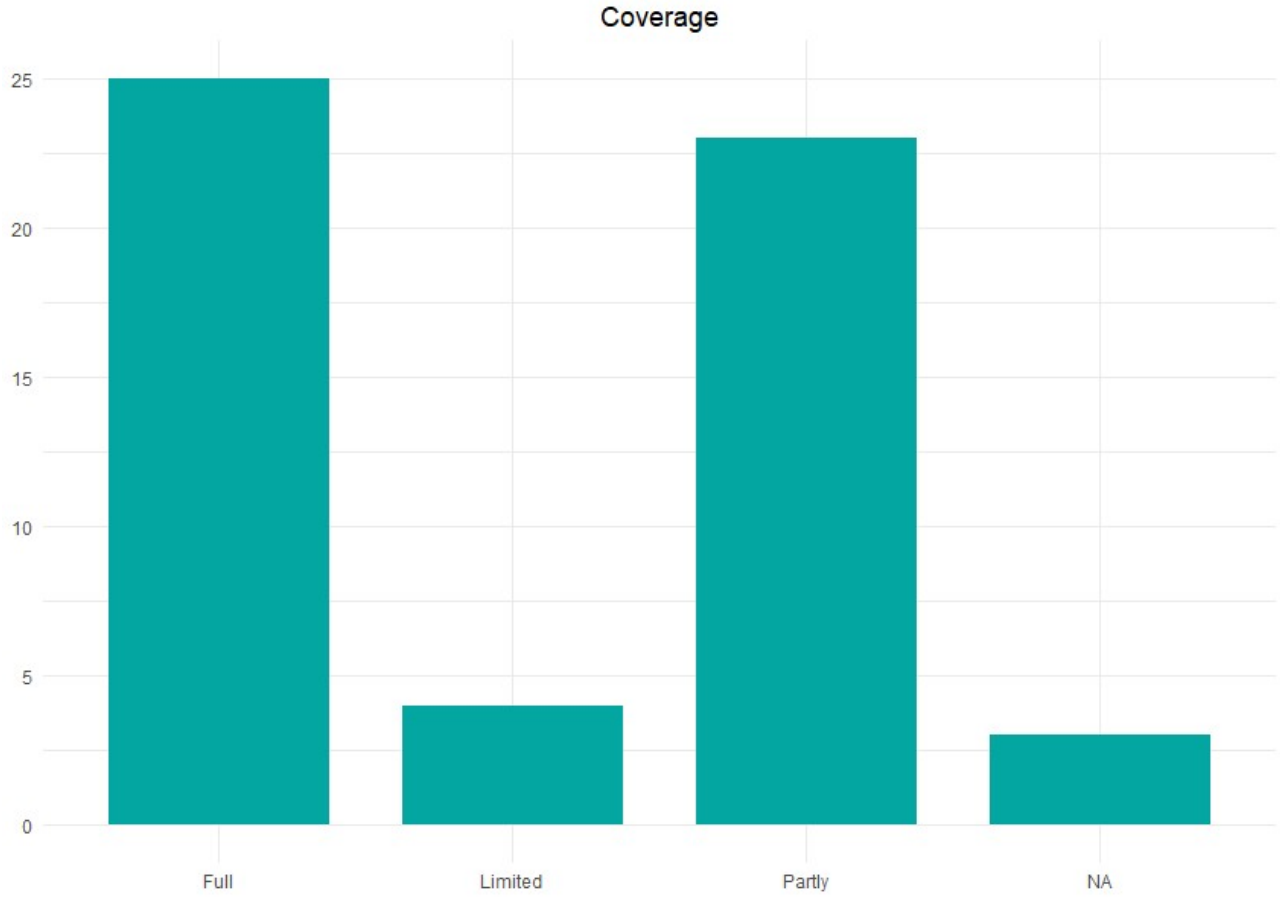

Electronic Invoices
Yes
value: 4
No
value: 7

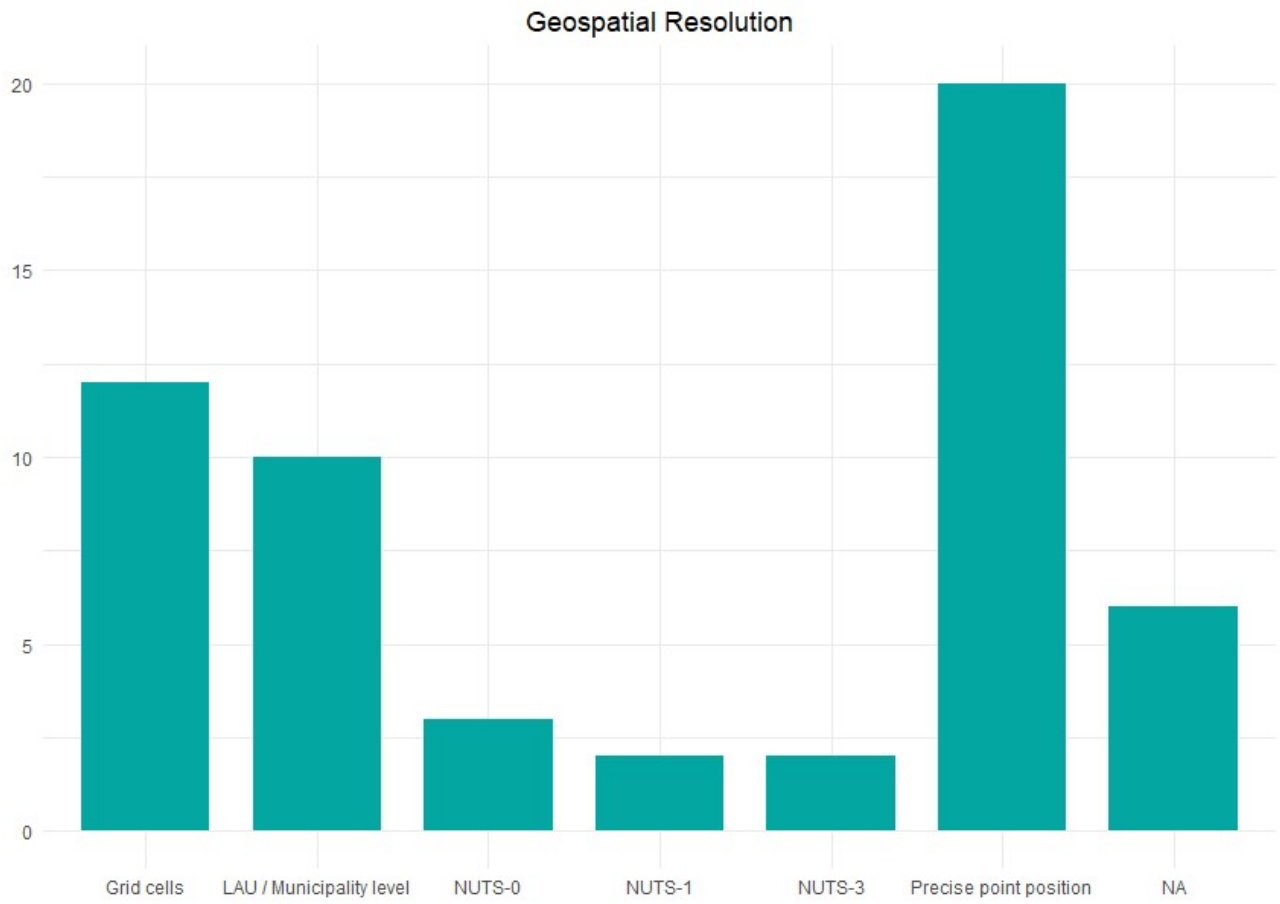

Connected Vehicles
Yes
value: 1
No
value: 10


Credit Card Transaction Data
No
value: 5
Yes
value: 6

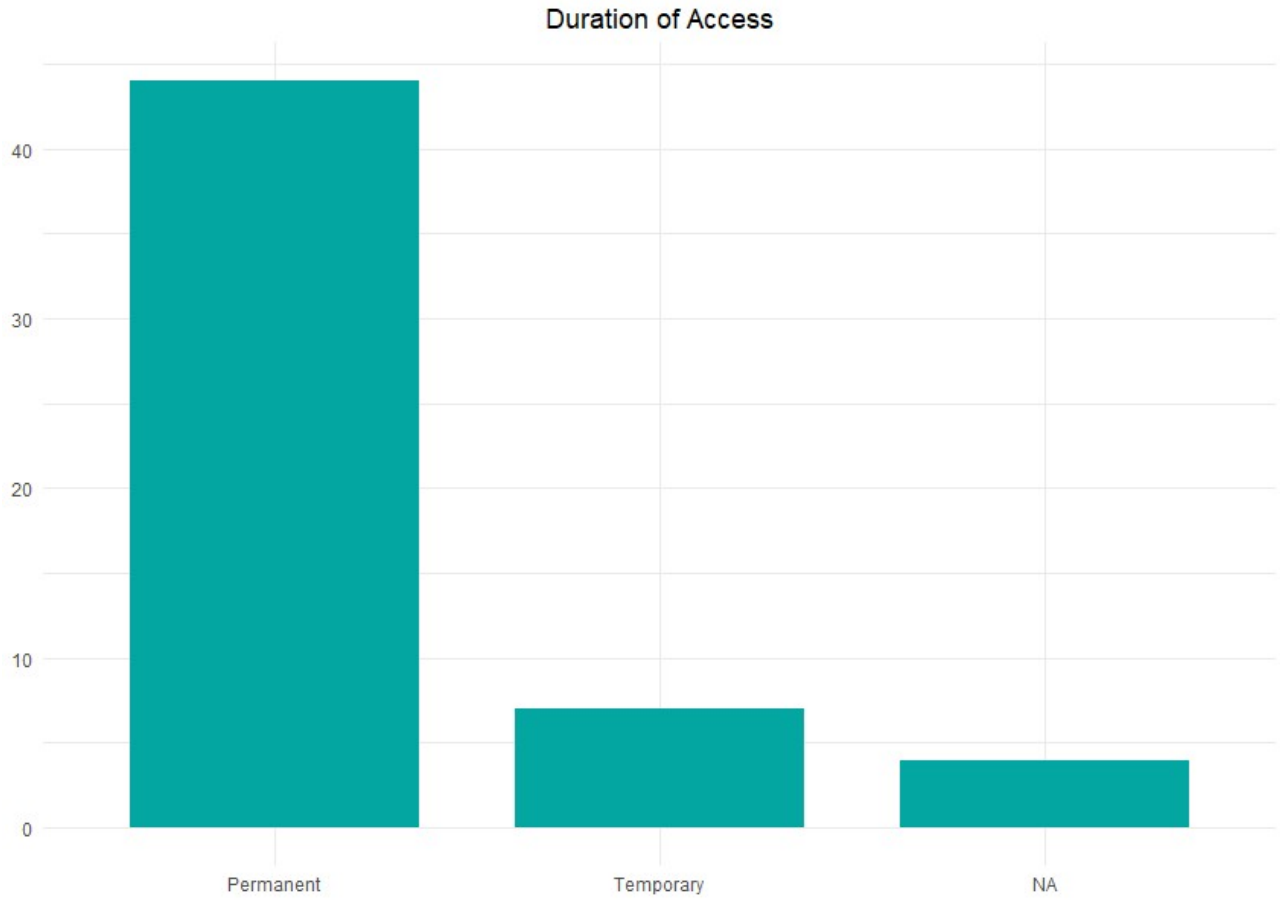

Privately or publicly held data
30
20
10
0
Both
Private
Public
NA

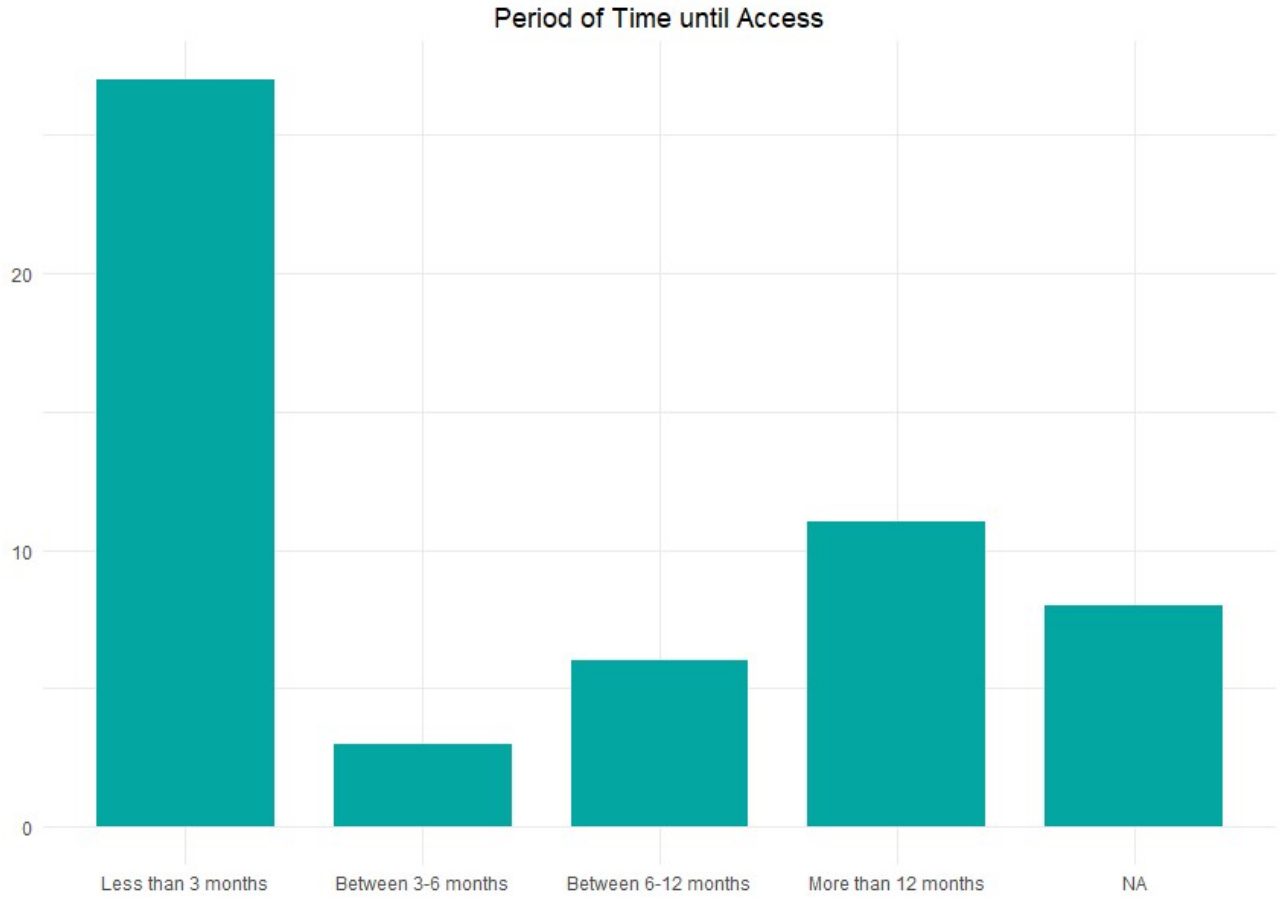

Are there any costs?
40
30
20
10
0
No
Yes
NA


Coverage
25
20
15
10
5
0
Full
Limited
Partly
NA



For Geospatial resolution, one relevant answering option was missing: Polygons. It was noted in the comments and it should be considered that at least two countries reported “Polygons” as the most suitable resolution for one data source.


Duration of Access
40
30
20
10
0
Permanent
Temporary
NA


Period of Time until Access
20
10
0
Less than 3 months
Between 3-6 months
Between 6-12 months
More than 12 months
NA

11.5. Group work on Tourism Applications at the ESTP-Course

As part of this project, the ESTP[1] course "Mobile Network Operator (MNO) data: Methods for Integrating New Data Sources in Official Statistics" was conducted[2]. To allow participants to explore possibilities and challenges regarding the integration of MNO and non-MNO data, a group work was designed and the following guiding questions hint towards relevant aspects with a focus on tourism statistics.

Generic Guiding Questions:

- **Identifying Data Sources:**
    - What traditional data sources (census, surveys, administrative records) are already used in this area?
    - What key limitations exist in these data sources?
    - What indicators from MNO data could fill these gaps?
- **Defining a Data Integration Approach:**
    - How will the data be combined? (e.g., direct aggregation, weighting, calibration with survey data)
    - What methodologies should be used to align MNO data with official definitions?
    - How can biases (e.g., phone ownership, MNO market share) be corrected?
- **Addressing Data Quality and Representativeness:**
    - How can we ensure MNO data correctly reflects population patterns?
    - Should we apply reweighting or correction factors based on official statistics?
- **Privacy, Ethics, and Accessibility:**
    - How will we handle privacy concerns (e.g., aggregation, anonymization)?
    - What legal or regulatory frameworks must be considered?
    - How can collaboration with MNOs be structured (e.g., data-sharing agreements)?

Estimating the Number of International Visitors in Real-Time

- Traditional border surveys and administrative records are slow and may undercount short-term visitors.
- Can roaming MNO data provide **real-time estimates** of the number of foreign tourists in a country?

[1] European Statistical Training Programme
[2] Link to the programme: https://cros.ec.europa.eu/book-page/mobile-network-operator-mno-data-methods-integrating-new-data-sources-official-statistics

- How can we **correct for biases** (e.g., tourists without SIM cards, multiple SIM usage)?

Understanding Domestic Tourism & Regional Travel Patterns

- National tourism surveys often **underrepresent** short domestic trips.
- Can MNO data help track **weekend travel, seasonal tourism flows**, and **urban-to-rural movement**?
- How can we **align MNO data with household survey results** for better accuracy?

Measuring the Economic Impact of Tourism in Key Destinations

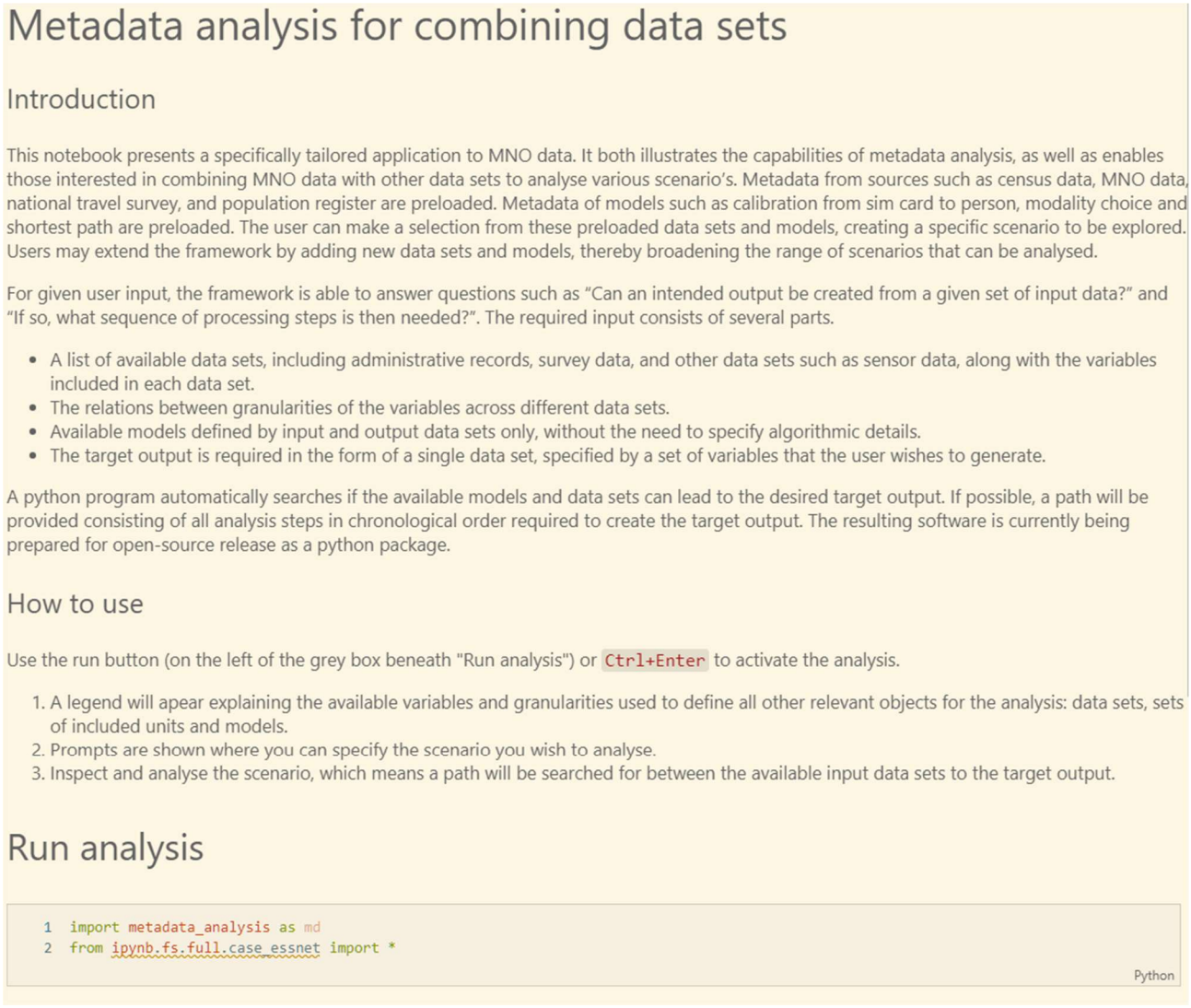
- Economic impact assessments rely on **business surveys and hotel occupancy rates**, which are **lagged**.
- Can mobile phone activity in tourism hotspots be a **proxy for spending levels** and **economic activity**?
- How can **official transaction data (e.g., card payments, tax records)** be integrated with MNO indicators?

Monitoring the Effects of a Crisis on Tourism (e.g., COVID-19, Natural Disasters)

- Official tourism statistics often lack **timely crisis impact assessment**.
- Can MNO data provide **near-instant estimates** of declines in visitor numbers and travel disruptions?
- What **baseline tourism levels** from previous periods could help measure recovery trends?

11.6. Screenshots of metadata analysis notebook

This annex presents screenshots of the user-friendly notebook for metadata analysis. The figures are ordered similarly to what the user encounters in the notebook.



Metadata analysis for combining data sets

Introduction

This notebook presents a specifically tailored application to MNO data. It both illustrates the capabilities of metadata analysis, as well as enables those interested in combining MNO data with other data sets to analyse various scenario's. Metadata from sources such as census data, MNO data, national travel survey, and population register are preloaded. Metadata of models such as calibration from sim card to person, modality choice and shortest path are preloaded. The user can make a selection from these preloaded data sets and models, creating a specific scenario to be explored. Users may extend the framework by adding new data sets and models, thereby broadening the range of scenarios that can be analysed.

For given user input, the framework is able to answer questions such as "Can an intended output be created from a given set of input data?" and "If so, what sequence of processing steps is then needed?". The required input consists of several parts.

- A list of available data sets, including administrative records, survey data, and other data sets such as sensor data, along with the variables included in each data set.
- The relations between granularities of the variables across different data sets.
- Available models defined by input and output data sets only, without the need to specify algorithmic details.
- The target output is required in the form of a single data set, specified by a set of variables that the user wishes to generate.

A python program automatically searches if the available models and data sets can lead to the desired target output. If possible, a path will be provided consisting of all analysis steps in chronological order required to create the target output. The resulting software is currently being prepared for open-source release as a python package.

How to use

Use the run button (on the left of the grey box beneath "Run analysis") or `Ctrl+Enter` to activate the analysis.

1. A legend will apear explaining the available variables and granularities used to define all other relevant objects for the analysis: data sets, sets of included units and models.
2. Prompts are shown where you can specify the scenario you wish to analyse.
3. Inspect and analyse the scenario, which means a path will be searched for between the available input data sets to the target output.

Run analysis

```
1  import metadata_analysis as md
2  from ipynb.fs.full.case_essnet import *
```

Python

Figure 22: Screenshot a of the user-friendly notebook for metadata analysis: introductory text.

```
----------Variables and granularities (legend)----------
a: MNOOperator
b: BackgroundCharacteristics
c: VehicleCount
d: Destination
   d0: Neighbourhood
   d1: Municipality
   d2: Cell tower
e: SampleInclusion
   e0: NTS sampling design
f: HasSensor
   f0: Has traffic loop sensor
l: Location
   l0: Neighbourhood
   l1: Municipality
   l2: Cell tower
m: Modality
n: SimCount
o: Origin
   o0: Neighbourhood
   o1: Municipality
   o2: Cell tower
p: Persons
q: TripPurpose
r: Route
s: RoadSegment
t: Time
   t0: Minute
   t1: 5 minute interval
   t2: Day part
```

Figure 23: Screenshot a of the user-friendly notebook for metadata analysis: legend of available variables.

```
----------Sets of included units (legend)----------
I: {p0 -- }
II: {p0 -- e_0: {1}}
III: {p0 -- a_0: {0}}
XI: {s0 -- }
XII: {s0 -- f_0: {1}, m_0: {car, motorbike}}
----------Data sets (legend)----------
NTS survey (b0, m0, q0 | p0, r0, t1)_II:
   BackgroundCharacteristics[], Modality[], TripPurpose[]
   per
   Persons[], Route[], Time[5 minute interval]
   for set of included units II.
Population Register (b0, d0, o0 | p0, t2)_I:
   BackgroundCharacteristics[], Destination[Neighbourhood], Origin[Neighbourhood]
   per
   Persons[], Time[Day part]
   for set of included units I.
Census (b0, p0 | d0, o0)_I:
   BackgroundCharacteristics[], Persons[]
   per
   Destination[Neighbourhood], Origin[Neighbourhood]
   for set of included units I.
Traffic Loops (c0 | m0, s0, t0)_XII:
   VehicleCount[]
   per
   Modality[], RoadSegment[], Time[Minute]
   for set of included units XII.
Route data (s0 | d0, o0, r0)_I:
   RoadSegment[]
   per
```

Figure 24: Screenshot a of the user-friendly notebook for metadata analysis: legend of sets of included units and pre-loaded data sets.

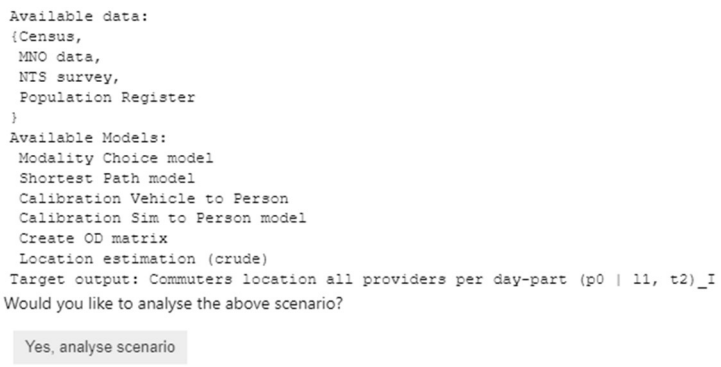


Figure 25: Screenshot a of the user-friendly notebook for metadata analysis: the user may confirm and inspect the scenario to be analyzed.