Trusted Smart Statistics: methodological developments based on new data sources

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

Work Package 2
Landscape of non-MNO data and potential target statistics

Deliverable 2.2
Report on a broad landscaping analysis of non-MNO data

August 2024

Partner in charge: INSEE (France) - *Marie-Pierre Joubert*

Authors[1]: DESTATIS (Germany) - *Gloria Deetjen, Maurice Brandt*
INSEE (France) - *Marie-Pierre Joubert, Chloé Breton, Julien Pramil*
INE-PT (Portugal) - *Sonia Quaresma, Antonio Portugal, Pedro Cunha*
INS (Romania) - *Marian Necula, Bogdan Oancea*
SCB (Sweden) - *Remy Kamali, Pär Hammarström, Wictoria Widén, Stefan Svanström*
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 not the grating authority can be held responsible for them.

Acknowledgement

Special thanks to M. Francesca D'Ambrogio and Immacolata Fera (Istat) - MNO-MINDS Coordination team - for the English language assistence provided.

Deliverable 2.2
Report on a broad landscaping analysis of non-MNO data

Summary

Work Package 2 of ESSnet MNO MINDS is dedicated to the landscaping of non-MNO data that could be usefully integrated with MNO data. This landscaping aims at giving insights into which of these data are the most promising. Criteria of analysis therefore integrate both the cost of accessing and handling these data and their relevance for being combined with MNO data.

This second deliverable of Work Package 2 goes in detail through a wide list of data sources. Each of them is systematically analysed through an assessment matrix, which is associated with a scoring along several dimensions. Data bases are then classified into three broad categories: most promising data sources to be combined with MNO data (census, Population Register, fiscal data, , combination of survey and register and transportation Surveys); Promising sources (but) which would require substantial work or which are not yet fully accessible (Vehicle, bicycle and pedestrian sensors, Vessel (boat) traffic data, Pollution data, Satellite data, Electronic invoices, Tourism Household and Border Surveys, Tourism platform data, Credit Card Transaction Data); and lastly less relevant sources (Google Maps Popular Time, Smart Meters, Connected Vehicles, Social Media).

In Work Package 2's last deliverable, due in August 2025, some complementary sources will be analysed (Tickets sold at mass events, Tourism accommodation statistics, Land use and land cover registers). Moreover, the analysis of the most promising sources will be deepened as well as their interest for some specific application scenarios.

Index

1. Introduction

Society's increasing digitalization has led both to the emergence of new data sources (smart sensors, mobile phone data, transaction data, satellite imagery…) 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 used in Official Statistics. During the COVID-19 crisis, MNO data were used in France and Germany to document (the) 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 EU countries have access to aggregated and anonymised MNO data. These data are bought by NSIs or obtained thanks to specific research funding. They are used by NSIs 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 it still has to be adopted by the new parliament and discussions have to be held regarding its concrete application in every EU country.

MNO data used by NSIs are mostly signalling data. These data are initially recorded by the operator for technical reasons (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 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. It aims to provide a broad view/an overview of all possible data to be combined with MNO data. It will then classify them among/into three groups:

- the most promising ones, for which a detailed analysis will be provided including some examples of usages/use.

- the ones which are foreseen as promising but that would need to overcome some methodological, technical, legal obstacles, (difficult to access to but with a perspective at middle term). For those a basic analysis will be provided.

- the ones which can already be considered as not worth investing in, at least in a middle term perspective (5 years). For instance: data whose usage would raise considerable problems of social acceptability. For this last group of data, the analysis will be less detailed.

This scoring of the data will serve as input for WP3, which will develop the methodology for combing MNO data with the most promising non-MNO data. Reciprocally, reflections led by WP3 about the theoretical frame of data combination will serve as guide for the data analysis.

The first part of this deliverable delves into the reasons for combining MNO and non-MNO data. This is indeed an essential step for identifying accurate non-MNO data. Then, the second part of this deliverable describes the main types of non-MNO data which are considered in our analysis. The third part provides some criteria to evaluate accuracy of non-MNO data in meeting the goals identified in the first part. Lastly, some non-MNO sources are analysed according to these criteria.

2. Why combining MNO and non-MNO data?

The interest of MNO data for official statistics has been well documented. ESSnet BigData 1 and ESSnet BigData 2 produced a considerable amount of methodological work and open-source software dedicated to their analysis. Based on these considerations, among others, members of 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). Among the key high-level requirements that this framework should fulfil includes support for the fusion of MNO data with non-MNO data. Indeed, although many useful application scenarios of MNO data have been identified and documented (such as present population or mobility behaviours), the limitations of these data – when used alone – have been highlighted in several NSIs publications (Sakarovitch, Bellefon, Givord, & Vanhoof, 2019). This report will not detail these pitfalls, yet it will focus on two main ways in which combining MNO data and non-MNO data can improve the quality of official statistics. This analysis will help identify the most promising non-MNO sources to combine with MNO sources.

2.1. Improving population statistics

It is hard to envisage the existence of a 100% exhaustive database covering the entire population. Even administrative directories have their limits and most of the NSIs’ usual databases cover only residential population and usual mobilities (such as commuting). The official statistical system lacks information about day-time population and mobility behaviours at a fine geographical and temporal scale. Conversely new data sources such as mobile phone data, traffic loop sensors or transaction data are not initially collected for statistical purposes, and their representativeness of the whole population is often biased due to the data provider’s market shares or some technical specificities (such as people switching off their phones).

The first goal of combining MNO and non-MNO data is (therefore) to improve population statistics, whether in terms of representativeness of the whole population or in terms of temporal and spatial precision. One important point to consider in the comparative analysis of different sources is the

potential difficulty due to some irreconcilable concepts: for instance, different definitions of fundamental units or different time periods. Data access also has to be guaranteed on a regular basis to produce official statistics in this fundamental topic. Following the concepts introduced in WP3, here MNO data would be a proxy of the *target measure*. Indeed, they entail some unique information about day-time population, with a spatial and temporal precision that cannot be reached by other sources. Non-MNO sources would here be used as *auxiliary data* to improve the robustness of the analysis. Indeed, as explained in the position paper (MNO, 2023), the observed population of mobile devices does not correspond exactly to the target population of humans (and the gap between the two populations is dynamic rather static). Therefore the combination of MNO and non-MNO data aims to stabilise, correct or calibrate the final statistics.

Several areas of application can be foreseen: for instance, daytime population also entails tourism statistics, which are often difficult to capture, especially for countries which do not conduct border surveys; mobility statistics could also be one application case of this combination of MNO and non-MNO data. Indeed, once areas have been characterised according to their purposes, MNO data allow for detailed statistics about the frequentation of these various locations, based on people's place of residence. For instance, combining MNO data with register data from the tax administration containing individuals' socio-demographic and income information would allow to better analyse people's mobility patterns according to their characteristics.

2.2. Broadening the scope of issues covered by official statistics

Combining MNO and non-MNO sources also enables the coverage of new topics of interest for official statistics. As explained in the introduction, this is essential for ensuring that official statistics continue to provide timely and relevant information considering the significant changes occurring in our environment. In this context, MNO data would serve as *auxiliary information*, offering insights into fine-grained presence and mobility patterns while non-MNO data sources would remain the primary basis for information. Potential application scenarios include the analysis of car-sharing patterns, social media usage, and causes of local increases in air pollution, among others. Both types of approaches will be considered when assessing the accuracy of sources to be combined with MNO data. However, priority will be given to application scenarios focused on day-time population, tourism and mobility, —areas collectively identified by ESSnet MNO-MINDS as the most important. More detailed information on the potential application scenarios of MNO data for official statistics, along with the related challenges, is currently being studied in Eurostat's Multi-MNO project.

3. Which criteria to analyse the relevance of non-MNO data?

3.1. Data sources to consider

In order to distinguish between different non-MNO sources, this report draws on the work by colleagues from this ESSnet WP3, the Norwegian NSI (Zhang, Haraldsen, Pekarskaya, & Hole, 2018) and ESSnets Big Data 1 and 2 (Kowarik & members, 2020). A category has been added: the combination of survey and register. This type of data has indeed different specificities from survey alone or administrative alone and they are more and more used by NSIs. The resulting categories are the following:

Data Type
a.1 **Survey** (census, SILC, employment, …)
a.2 **Combination of survey and register**
b. **Register** (vital events, diagnoses, wage, income tax, welfare payments, …)
c. **Transaction** (scanner data price, point-of-sales receipt, bankcard or giro payment, P2B or B2B invoice, P2P (ex: paypal...), property sales contracts, ownership registration, …)
d. **Static detection of connected objects and environmental phenomena** (smart meters readings, whether station readings, traffic loop signals, lorry tracking signals, maritime AIS…)
e. **Mobile airborne sensing** (satellite images, drone images, airborne laser scanning, …)
f. **Internet** (web pages, social media posts, …)

Some non-MNO data are produced by official statistics, such as survey-based Census. In this case the whole data production process is designed from the outset to meet the needs of NSIs, whether in terms of quality or variables of interest. Others are traditionally used by NSIs although they are originally gathered for administrative purposes. These data need some specific treatments to meet the NSI's quality requirements, yet they are generally structured in a way that allows classical statistical treatments, and the data providers also have an interest in offering the most exhaustive view of their population of interest. The last non-MNO data that this report considers are initially produced for objectives far removed from those of the official statistical service. The producer can be either from the private or the public sector and these data are often not structured in a classical way; some lack documentation about the variables of interest for NSIs or about the quality issues encountered in the data collection process.

3.2. Primary scoring dimensions

In order to analyse the potential of these sources, WP2's members have used as a starting point the 'Big Data classification matrix' which was produced by ESSnet BigData 2 (Kowarik & members, 2020). This matrix has then been adapted to the specific question of the combination of MNO and non-MNO sources (see figure 1). The reasons for these choices and adaptations are detailed below, with a focus on the analysis of new data sources, since for sources already used by national statistical offices, these criteria are already met.

Data type: some non-MNO sources are of a size that can be handled on NSI's premises, with a regular data analysis software. For instance, air pollution data at the municipality level. Others, on the contrary, need specific big data infrastructure, software dedicated to their analysis and specific data scientist skills (e.g., card transaction data, social media post, etc.). This information is important for the cost-benefit analysis of using these sources.

Access: the data owner is a crucial point for accessing and handling new data sources. First, its private or public status will impact the ease of access to the data and the partnership terms. Then, the number of different data owners will influence the ease of harmonisation and the coherence inside a same country and between countries. For instance, if each city produces public transport ticketing data in a different data format and with no harmonized meta-data, this will lead to almost insurmountable difficulties in handling these data for producing official statistics.

The stability of data access is also an important guarantee: is it a one-shot access to specific data or can a regular statistical production be envisaged? As regards the financial cost of accessing the data, a distinction has to be made between a minimal compensation cost of the data extraction and a cost that would lead to financial benefits for the operator. In this matter, the project of regulation 223 mentions that "As official statistics are a public good, the access to data should be free of charge. Where data are requested by an NSI, Member States may provide compensation to the private data holder that is limited to the processing service according to the specifications requested, except where national legislation does not allow NSIs or other national authorities to compensate data holders". The cost of the "processing service" can be non-negligible and therefore has to be considered in the data scoring.

Some of the new data sources can also be used by the data owner for commercial use. This could lead to some restrictions in the topics to cover, or the timeliness of the analysis to avoid direct competition with the data producer. In the same spirit, some data producers make the use of their data by official statisticians, conditional on avoiding some topics, which they consider as too sensitive. All these aspects must be considered in scoring the data.

The data aggregation level will be an important input for elaborating the methodology of combination with MNO data. WP3 distinguishes between M-enabler methods, which allow to improve data configuration, and M-executor methods, which apply to available data in the given configuration. M-enabler methods generally require as a pre-requisite to have access to micro-data, that is pseudonymized individual data with the geolocation and the time stamp. Yet accessing micro-data is often complicated by regulatory and confidentiality issues. Macro-data, already aggregated at some broader geographical and temporal scale, is more easily accessed at the moment. In this sense, WP3 methods mostly deal with M-executors' methods. Yet attention is still given to M-enabler methods, thus this point has to be dealt with in the assessment matrix (ESSnet MNO-MINDS (2024).

Metadata: issues addressed in this section are classical prerequisite for combining multiple data. The reference period of non-MNO data to combine with MNO data is important information, all the more as one of MNO data's interest is their high temporal frequency. If it is planned to combine MNO data with, for instance, census data, the observation years have to be as near as possible.

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 the information available on these data
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 1: assessment matrix - first part

3.3. Secondary scoring dimension

One important aspect of ESSnet's MNO-MINDS is that it aims at producing official statistics, and rather than only experimental statistics. The short list of the most promising data sources will therefore be established considering the availability of sources in all NSIs and the extent to which they respect the European Statistics Code of Practice (CoP, 2017). Naturally, non-MNO sources which already are in the official statistics domain fulfil the Cop requirements, which is not necessary the case of new data sources. This should therefore not be an essential prerequisite for considering a source as promising for being combined with MNO data. Yet going through the adequacy of a source to the CoP principles gives a rough estimation of how costly and complicated it would be to produce official statistics using this source. This adequacy will therefore be considered as a secondary level criterion in the sources scoring process. The detailed secondary level scoring will be conducted in deliverable D2.3. In D2.2 some preliminary elements regarding compliance with the CoP principles are mentioned.

As well known, the ES CoP includes 16 Principles organised in 3 areas, Institutional Environment, Statistical Process and Statistical Output. For each principle there is a set of indicators that represent best practices and guidance for the implementation of the principle. Not all the principles and indicators are relevant or specific enough to analyse and evaluate non-MNO data to be integrated with MNO data to produce Official statistics; a selection is proposed in figure 2.

Concerning Principles related to Institutional environment and Statistical Processes area, the selection was based on the following motivation: for the Institutional environment area, only the principles that could be connected to the actual usability of the data have been considered, and only if the issue was not already addressed in the previous part of the assessment matrix: for example, Professional Independence (principle 1) is a general principle not connected with the use of a specific data source, while the topic of data access (principle 2 Mandate for Data Collection and Access to Data) is already included in the assessment matrix. As regards adequacy of resources, indicator 3.2 about the cost-benefit analysis is an interesting follow-up to the question in the first part about whether the data producer has to be financially compensated for the data preparation. A regular monitoring of the output quality (indicator 4.3) is even more important as many new data sources aren't collected for official statistics purposes. Therefore, the data producer can have other objectives which don't necessarily impose the same quality monitoring requirements. This goes

hand in hand with indicator 6.3 about the existence of a process to inform the NSI in case an error is discovered. The security of data transmission has of course to be respected (indicator 5.5). Reaching these requirements at the level expected from the CoP will induce higher technical and organizational costs of access, especially for data sources which are not already included in these NSI processes.

For the Statistical process area: indicator 8.1 highlights the importance of the coherence between the concepts or at least the fact that there is a good approximation between them. A good cooperation between the NSI and the data producer (indicator 8.7) allows to build a trusting relationship, which greatly facilitating communication and therefore eases the exchange of information about specific methodological issues (concerning for instance the way data are produced). The fluency of this exchange of information is essential to understand the specificities of these data. Principle 9 Non-excessive Burden on Respondents could be a significant advantage of some new data sources, which are collected passively without imposing respondents to answer actively to a questionnaire. An important issue in this matter will be the possibility of linking different data sources (indicator 9.6). Principle 10 in its global formulation of "Cost effectiveness" covers broadly the challenges associated with the use of new data sources. These challenges and associated costs are already a core element of many of the previously described items of the analysis matrix. Yet among indicators described in principle 10, indicator 10.4 highlights the importance of implementing standardized solution.

As regards the statistical outputs: principle 11 Relevance is already included in a specific part of the assessment matrix, dedicated to the relevance of this non-MNO source to contribute to the identified application scenarios of the combination of sources. Principle 12 goes in more details about the monitoring and documenting of the potential errors. Some of the indicators won't apply to all data, yet it seems important (for the data which will be selected as the most promising ones), to assess these details. Principle 13 will be important to consider in the scoring of the data sources aimed at distinguishing between the most or least promising ones. Indeed, it brings the needs of official statistics into line with the temporal frequency of the data. Among indicators from principle 14, two in particular were selected. Indicators 14.1 and 14.2 are linked to the "data access" considerations of the first part of the assessment matrix. Indeed, guaranteeing a regular access to data which are coherent and comparable, both temporally and geographically (for instance if there are different data providers according to different geographical areas) is vital. Cross national comparability is particularly not obvious for data provided by private operators, who may have different data pipelines or might follow different data regulations according to their countries. Yet ensuring availability in as many European countries as possible is an important criterion to score the data. Lastly, as stated in indicator 15.4, access to micro data for research purposes is an important criterion for ensuring transparency in the production leading to official statistics production. Art. 23 of 223 revision reads that "Access to confidential data, including data made available by private data holders, […] may be granted to researchers carrying out statistical analyses for scientific purposes […]" This quite soft formulationsuggests this criterion as less crucial than previous ones. Yet one important direction for NSIs is that their work is a public good, ideally available on an open-source basis. Therefore, it seems a good practice that some researchers are granted a specific access to the data and can reproduce the work done by the NSI.

Quality requirements (derived from ES 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 2 : Assessment matrix, part 2 - official statistics

4. Source scoring

4.1. Source scoring in a nutshell

Most promising data sources to be combined with MNO data
Census
Population Register (fiscal data, …)
Combination of survey and register
Transportation Surveys
Promising sources but which would require substantial work, or which are not yet fully accessible
Vehicle, bicycle and pedestrian sensors
Vessel (boat) traffic data
Pollution data
Satellite data
Electronic invoices
Tourism Household and Border Surveys
Tourism platform data
Credit Card Transaction Data
Less relevant sources
Google Maps Popular Time
Smart Meters
Connected Vehicles
Social Media
Sources still to be analysed
Tickets sold at mass events
Tourism accommodation statistics
Land use and land cover registers

The following more detailed description of the data builds on the analysis led by country members. It is therefore based on data to which they have access or which they took the time to examine theoretically even without having access to it. The generality of these analyses for other EU countries will be analysed in WP2's last deliverable, such as the scoring according to the secondary scoring dimension presented in 3.3.

4.2. Most promising sources

Unsurprisingly, the most promising sources at the moment, that is the ones with the best cost-benefit analysis for being combined with MNO data, are the ones which are traditionally used by National Statistical Offices. Indeed, they fulfil almost all the requirements of the assessment matrix, and they are also fit with the application scenarios which aim at improving population's coverage. Here is a summary of the analysis regarding these sources:

4.2.1 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 eventually derived from extrapolation (unless it is a full survey). In this analysis, there are three sources 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 Micro Census (SILC, LFS)

The "Mikrozensus" (micro census) 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 detailed in annex 7.1.

Challenges associated with these databases and ways of handling them

German census is conducted once every ten years and provides the official population number. For the time between censuses, an intercensal population figure is estimated annually based on the latest census and administrative information. Conducting a census is complex and costly (in terms of time, resources, etc.).

For the German register-assisted census, copies of administrative registers serve as base data. Then, less than 10% of the population is surveyed to correct inaccuracies from the registers. The surveys collect further information as well, e.g. information on education and employment. The information on buildings and housing comes from the owners who were surveyed by post. In addition, a further survey is conducted in residential establishments and collective living quarters.

The German and the French censuses are determining the official population figures. Both provide additional information on certain variables listed above and both include information on housing and buildings. The difference from the German micro census is that the micro census does not aim to produce the official population number but to provide more details on a variety of topics. The micro census 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. Micro census 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 provide a database for science, politics and society.

Accuracy for the Application scenarios of the combination with MNO data

Experimental application scenarios have been conducted using total population figures and/or additional data from the Total Population Register and Census, combined with MNO data. Many MNO application scenarios require some form of resident, de facto population, or socio-demographic information, even when the research focus differs.

The following (potential) application scenarios may be relevant:
*De-jure/resident/night-time population
*De-facto population
*capturing non-registered persons
*improving temporal and spatial granularity of population figures
*investigating differences between total population by official statistics and by MNO data
*socio-demographic information by census and TPR as auxiliary variables

For most application scenarios and target statistics, census data is used as an auxiliary variable when integrated with MNO data. In fewer cases, there might be scenarios for which census data may serve as a proxy (target variable). Census results complement MNO data by providing a base for the total population size, especially when available MNO data does not cover data from all MNOs (extrapolation). Census data improves the representativeness of MNO data, particularly for night-time comparisons, facilitating concept reconciliation with the residential population. Census data also provide multiple socio-demographic variables.

Data analysis through the assessment matrix

Here, we focus on specific points; otherwise, these data sources comply with the requirement of the assessment matrix.
The 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).

For German census, data is available in the database and can be accessed online (https://ergebnisse.zensus2022.de/datenbank/online).
Access is also possible via API and software packages, such as {restatis}. Linkage with MNO data can occur at the geographical level, typically using the 1x1km INSPIRE grid.
For the French census, data structure geolocation variables are known. Sub-municipal geolocation is only available for the 2017 census results.

There are no legal or ethical issues regarding data access. However, for the micro census, linkage with other data sources requires a legal check. Data can be accessed at a detailed level, but there are limitations at the individual and very fine geographical levels for confidentiality reasons. Not every EU country conducts a census. The remaining countries use alternative data source with comparable figures (e.g. the Total Population Register). Micro census results are available for all EU countries that conduct a micro census. EU-wide surveys (Silc, LFS, IKT) are available for all EU Countries.

Data Scoring for the combination with MNO data

The costs associated with processing total population numbers and information from Censuses are relatively moderate. These sources are produced by NSIs, and statistical officers are accustomed to

working with this data. The infrastructure, quality standards, and knowledge about these data sources are already in place. Depending on the desired target statistics, legal checks may be required to clarify which microdata linkages are permitted. For published results, there should not be issues, although caution is advised when combining multiple data sources.

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**	

4.2.2 Total Population register

Data description

Swedish Total Population Register

The Swedish Total Population Register contains data that is essential for understanding Sweden's population. The system includes registers that are the basis for official population statistics and provide basic data for many of Statistics Sweden's operations. The Total Population Register contains information about the population and its changes and largely reflects the content of the population register of the Swedish Tax Agency.

Challenges associated with this database and ways to handle them

The main challenge is the under-coverage of non-registered people in Sweden. The Total Population Register (TPR) is indeed the foundation of official population and household statistics. Examples of population statistics include population by sex, age, marital status etc. in counties and municipalities. Statistics on population changes may relate to internal migration, births, deaths, marriages, divorces, immigration and emigration, etc. The statistics are produced several times per year.

TPR serves as a sample frame for many individual samples at Statistics Sweden. Samples are made for both appropriation and commissioned operations and are mainly used for different types of questionnaire and interview surveys. Some examples include Labour Force Surveys, the Political Party Preference Survey, and the Surveys on Living Conditions.

TPR is also a database which can supply supplementary data to other registers and surveys. This usage enables a reduction in the number of questions in a survey, which reduces the burden on respondents.

For example, the Total Population Register provides:
- background data for the registers of labour market statistics, economic welfare statistics and education statistics (answers the socio economic information);
- the basis for the coordination of register populations and surveys within statistics on individuals;
- data for statistical packages and for the statistical basis of population projections;
- procedures for updating data in samples for questionnaires and interview surveys;
- data for supplementation of personal identity numbers in different material;
- data for authentication in connection with the request of extracts of registers in accordance with the Personal Data Act.

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

The application scenarios and the quality issues are similar to those identified for the census. The availability in all EU countries has to be further examined.

Data Scoring for the combination with MNO data

The costs to process total population numbers and information from Population Register are similar to those incurred for Censuses.

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		**moderate**	
Accuracy and robustness of the information available on these data			**high**
Non-sensitivity of these data	**low**		

4.2.3 Transportation surveys

Data description

The National Travel Survey (RVU Sweden)

Transport Analysis, the Swedish government-appointed authority, is responsible for Sweden's official statistics in the area of Transport and Communications.

Since 2019, a travel habits survey has been underway, with results presented in annual reports. The survey covers Sweden's population aged 6–84 years and is conducted as a combined paper and online survey. The sample is stratified by county, age group and gender. In 2023, the sample included approximately 12,200 people in Sweden.

For organizations and regions that want an expanded sample for a specific interest population, it is possible to buy extra samples in the survey.

Statistics on travel habits provide important background information for formulating national and regional transport policy, developing infrastructure and transport services, improving road safety, and supporting research on people's travel and communication patterns.

Variables

The tables published by the Agency represent only a small portion of the data collected. Stakeholders can request customized tables or access the database

Table 1. Overview of variables in the Travel Survey database

Area	Description
The individual and Household	Gender, age, driving licence, access to a bicycle, disability: vision, hearing, mobility, transport services. domicile, municipality of residence, municipal group, household composition.
Cars and parking	Household car ownership: type of ownership, in traffic, parking at dwelling.
Travel Cards	Season tickets for public transport
Method	Answers via paper or web.
Measurement day movements	Mode of travel, route, errand, addresses for start, end, and target point, start and end time, travel time

The statistics are presented by sex, age, municipality group, mode of transport, errands and distance travelled.
The breakdown by mode of transport used is as follows:

- on foot
- bicycle
- car
- modes of public transport (including public transport by bus, train, tram, or metro) and,
- Other modes of transport (including transport services, taxis, air, and sea transport)

Errands are divided according to:

- work, business, and school trips.
- service and purchasing,
- leisure and
- Other matter

Challenges associated with this database and ways of handling them

The survey has sources of uncertainty that are largely since it is a sample survey. The sources of uncertainty include problems with non-response, as well as coverage of the target population and measurement errors.

The main challenge is related to the Survey per se, and its many sources of uncertainty, among them, the relatively low response rate, around 17 %. The response rate is slightly lower among men than among women. Examined by age group, the response rate is lowest in the age groups 15–24 years and 25–44 years.

Another challenge concerns the sample size that imposes limitations on the ability to break down results into small accounting groups. This applies to small geographical areas, modes of transport that are used to a limited extent, errands that are not carried out frequently or narrow intervals for age groups.

Users of the statistics include The Swedish government, Transport Analysis, the Swedish Transport Administration, and other authorities. Users are also found at universities and colleges, as well as in the media and with the general public.

External users contact Transport Analysis with questions about the statistics and sometimes also with requests for changes or special extracts from the database for the national travel survey.

Transport Analysis is aware that the statistics do not cover all current and potential users' needs, and thus the relevance of the statistics can be improved.

Accuracy for the Application scenarios of the combination with MNO data

These data are complementary to MNO data. The travel survey provides crucial background information, such as gender and age. It also captures individuals' daily travel behavior, the dates and times of travel, the modes of transport used, and the purposes of the trips.

They allow improving MNO data population representativeness. Indeed, the Travel Habits Survey provides essential context for understanding travel behaviour —such as trip purposes—which cannot be derived from mobile network data alone.

Modes of transport are also difficult to identify accurately through mobile network data, and due to confidentiality concerns, it is challenging to obtain detailed individual-level information.

According to WP2 members, there are no major conflicts regarding definitions or time periods. It is possible that a close integration of these sources could lead to changes in the survey questionnaire, which would be beneficial if it results in improvements to both data sources.

In terms of improving temporal or geographical precision, mobile data makes it possible to more rapidly track transport developments in areas such as new business zones or residential developments. More up-to-date and geographically broken-down statistics can also serve as a basis for changed travel patterns, such as during a pandemic. Research conducted by the Transport Analysis Agency has shown that mobile network data is a possible complement to travel habit survey to describe travel flows in total in Sweden, within counties and per year. In addition, with mobile network data, it is possible to describe the number of journeys at finer levels and in periods shorter than one year.

Regarding the expansion of topics covered by official statistics, the focus for the Transport Analysis Agency is on improving the relevance of existing statistics. In this sense, the MNO-data can potentially significantly improve its relevance, in terms of more accurate total counts, finer granularity, richer mobility patterns, and improved timeliness.

Illustration of a concrete use-case: collaboration between Statistics Sweden and Southeast region of Sweden

Sweden is a very decentralised country with Regional Authorities having an important role to play and clear mandate. In this context, 3 big regions together with 13 municipalities – the so called "Southeast triangle" have started a joint project with the aim of improving the cross-border exchange to boost the potentials of the economic growth of the Regions.

In this endeavour, the Regions realised the information about travel across county borders of the interest was very limited. Of course, there is the Swedish travel habit survey along with some register data (housing, study, workplace) and that was all. It became clear to the regions that there was a information about for instance leisure trips, visitors trips, work trips, how many people do not travel even though they have a place of work in another place, and so on. They quickly realised that they needed a complementary source to enhance insights, and the mobile network data was identified as one of the most suitable sources.

The Regions contacted Statistics Sweden to investigate the possibility of a partnership in order to get access to MNO-data. This joint project is on its initial phase. The aim of the project is to be able to answer the following questions:

- How do the flows down to the level of urbans areas look like in these regions?
- What does the exchange look like between regional locations
- What does it look like in time/days of the week – does it differ in the evenings and weekends/holidays (leisure trips)
- How many people are in a town during the day/evening/night/weekend (basis for municipality services)
- How does the night/day population vary over the year

Furthermore, if possible, the Regions would like to put the traffic on the road network through a route selection engine (OpenTripPlanner) to build some scenarios:

1. Everyone travels roughly as predicted by the travel habits surveys

2. Everyone travels as the regions expect and wish 1. Walk, 2. Bicycle, 3. Public transport, 4. Car.

Data quality analysed through the assessment matrix

The data is stored in a database which is easy to access. As such, each question of interest can be asked, processed, and summarized in a data set. There is no problem in handling the data set which can be obtained from the agency website in a simple Excel format.

The data sets extracted from the database are in a form of Excel files, in a very simple structure, accessible for everyone. All the variables that enable linking of the data are known.

The data is owned by the Transport Analysis, and it is a "public good". Access is therefore free and without competing use or specific legal issues.

It is plausible that all EU countries have some kind of Travel habit survey, yet the availability of this data in all EU countries has not been surveyed yet. This question will be examined in deliverable D2.3.

The target population is well defined: it consists of all persons registered in Sweden aged 6–84 years. The survey is a sample survey based on a stratified sample from the population register. The sample is drawn from the Total Population Register (TPR from Statistics Sweden, previously described). The selection is made stratified by **county**, **age group** and **gender**.
The sample for 2023 includes approximately **12,200 people** in Sweden. In the survey, the respondents are only asked to respond about a specific measurement day, but together they cover the entire reference period of the survey. The anonymized unique personal number is the base unit of the dataset.

Since 2019, the survey has been conducted as a combined paper and online survey, and the results of 2023 can be compared with the 2019–2022 survey.
Surveys from 2018 or earlier were instead conducted with telephone interviews, and the change in method means that this survey is not fully comparable to these surveys. Comparisons with these surveys should therefore be made with caution, as the differences may to some extent be due to methodological differences rather than actual changes in travel.

The survey is considered to have comparability between groups. In published tables, for example, municipal group division is used, which is a common division for reporting by functional regions. In addition, the travel habits database contains information that allows many categorizations by age, mode of transport and errand.

Transportation 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		**?**	
Accuracy and robustness of the information available on these data			**high**
Non-sensitivity of these data			**high**

4.3. Promising sources but which would require substantial work, or which are not yet fully accessible

4.3.1 Vehicle, bicycle and pedestrian sensor data

Data description

Sensor-based data is a valuable source of information, especially in the field of transport and the environment. In Sweden, a lot of analysis and political decision-making is increasingly reliant on these sources. Unfortunately, their main weakness is linked to the fact that they are limited in their coverage of the territory, and therefore are not located everywhere, in addition to the fact that their distribution is highly uneven. In some places, they are very concentrated, and in other places, they are almost non-existent.

The two predominant sources in Sweden (and probably in some other countries of the European Union) are Vehicles sensor locations for congestion taxes in Stockholm Gothenburg and the environmental barometer in Stockholm that measures bicycle and pedestrian activity. The focus in this document will be on these two sources.

There is also an informal collaboration between SCB and Statistics Norway (WP 3) where it is empirically explored how these sources can be used together with MNO. See accompanying document in annex 2.

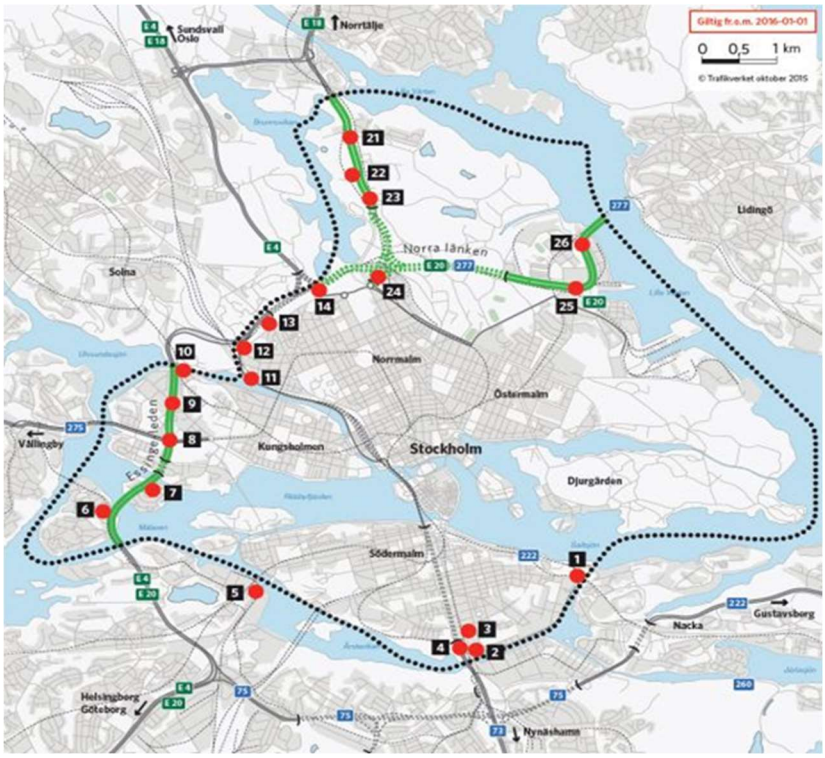
Vehicle, bicycle and pedestrian sensors in the city of Stockholm

Vehicle sensor locations for congestion tax – Stockholm and Gothenburg

Description: In Sweden congestion tax systems are in place in Stockholm and Gothenburg. The tax applies to vehicles registered in and outside of Sweden. The primary goal of the congestion tax is to reduce traffic congestion and covers both Swedish and foreign-registered vehicles.

The Swedish Transport Agency compiles data from the system and publishes some statistics. For example, a time series over the number of vehicle passages per month can be obtained. One can also see information that shows the number of passages at the “toll stations” and the proportion that were not taxed. Passages are recorded on weekdays between 06:00 and 18:29.

Below is a map showing the toll stations for congestion tax in Stockholm municipality. At the toll station, passages are automatically registered during the times when congestion tax is levied.



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

Challenges associated with these data bases and ways of handling them

The main limitation as previously mentioned is the limited coverage. The sensors are clustered in Stockholm inner-city and Gothenburg inner-city. In the rest of the country, there is no coverage available.

Another limitation relates to the target measure or statistics of interest. If the aim of the statistics is to measure the number of persons-crossings into (or out of) Stockholm inner-city during some given hours and days, only the number of vehicle passages per hour or per day is covered.

Accuracy for the Application scenarios of the combination with MNO data

The number of person-crossings refers to crossings made by persons. It is not the same as the number of (distinct) persons making the crossings, nor the number of persons present given the time and place. While mobile network operators (MNOs) can count the number of SIMs crossing into (or out of) Stockholm inner-city, it covers neither all the persons nor all the crossings. While there are non-MNO sources that can either provide the target measure of person-crossings by a given travel mode or a proxy of it, they lack the related data of movement (e.g. inside the inner-city). Therefore, combing relevant non-MNO sources of this sort with MNO data can generate more useful data for public users, and this combined data can also benefit the MNOs.

The main value of combining this data with MNO data is for statistical calibration. See the accompanying note from WP 3 for more details.

Data analysis through the assessment matrix

The data is "owned" by the Swedish Transport Agency and is publicly accessible on their website. The definition of the population is known. In this case, the target measure is not the number of persons crossing but the number of vehicle passages at specific places and times.

The vehicle sensors for congestion tax provide the number of vehicle-crossings by the type of vehicle and the point-of-crossing, as a proxy to the corresponding target measure of person-crossings.

No official statistics are based on this data source. Due to the limited coverage and other issues, it is difficult to envision official statistics based on this source. Rather, it is an interesting source to combine with the MNO-data.

Data description

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

Since 2015, the city of Stockholm has carried out annual manual counts of pedestrian passages in connection with counting bicycle passages. These take place during the spring and autumn. The main objective of the sensors is to support the city's environmental goals and enhance the well-being of its citizens.

In line with the Accessibility Strategy of Stockholm, the city is investing in increasing knowledge about the number of pedestrians. The city has therefore commissioned seven fixed measuring stations for automatic pedestrian counting, which deliver continuous data. There are four in the outer city and three in the inner city. The city has also invested in increased measurements in 2017 and 2018 so that pedestrian data from more than 200 locations are now available.

In 2017, the city counted the gender distribution of pedestrians in a selection of places. In total, 49.5 percent were measured as men and 50.5 percent as women.

Three quick facts about walking in Stockholm:

• About 1.6 km is the median length of a walking trip (where the entire journey is done on foot) in Stockholm.

• 37 percent of all trips in the city are made on foot.

• In the inner city, 52 per cent of all journeys are made on foot. In the outer city, the figure is 35 percent.

As for the bicycle traffic, the city of Stockholm has invested heavily in a plan that includes several measures to build cycle lanes, facilitate bicycle commuting, build more bicycle parking spaces, change the prioritisation of traffic signals, and improve operation and maintenance.

Bicycle measurements show a sharp increase in the number of cyclists. The city's cycling promotion measures together with an increased health trend and environmental awareness may be behind the increase.

The city has placed bicycle sensors in inner-city and city centre to measure the number of cyclists in Stockholm.

Challenges associated with these databases and ways of handling them

The main challenges are basically the same as for vehicle sensors. The sensors are not placed everywhere. They are rather clustered in few places of the cities. Hence, the coverage is very limited.

Accuracy for the Application scenarios of the combination with MNO data

The main value is for statistical calibration. See the accompanying note from WP 3 for more details.

Data analysis through the assessment matrix

The data is "owned" by the City of Stockholm city and is publicly accessible on their website.

Although it can be said that the target population is known, it can still be stated that the coverage is very limited. The information available can be found on the website.

No official statistics are based on this data source. Due to the limited coverage and other issues, it is difficult to envision official statistics based on this source. Rather, it is an interesting source to combine with the MNO-data.

Data Scoring for the combination with MNO data

We believe that this type of source contains huge potential to integrate with MNO data, not least to enrich the MNO data source whenever possible. Therefore, this source should not be dismissed but considered an important complement to MNO data.

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**

4.3.2 Vessel (boat) traffic data

Data description

The Swedish Maritime Administration has been involved internationally in developing the Automatic Identification System (AIS) and in establishing requirements for its use. As a result, since 2007 all merchant vessels over 300 gross tons worldwide are required to have this equipment on board. Additionally, many recreational vessels have recognised its benefits and have started using the system. Transponder data is collected via the coastal radio system and has been stored since 28 September 2006. Extracts from this database are used to measure traffic flows and can also serve as an important input for decisions making.

The Swedish Maritime Administration uses this information to gain an overview of traffic trends in Swedish waters. Historical AIS data is also used to evaluate sea rescue operations and to create a better basis for fairway projects or hydrographic surveys.

This data is also an invaluable resource in the design of offshore installations and wind farms.

Accuracy for the Application scenarios of the combination with MNO data

This data could be used for statistical calibration. A literature review and more detailed hypotheses will be presented in Deliverable D2.3.

Data analysis through the assessment matrix

Data from the AIS database can be extracted either according to geographical distribution or desired passage lines. For example, the information can be used to study the impact of ships on shore or in insurance investigations involving damage to cables, piers, diving cables or similar incidents.

Access to this data is not free of charge.

The Swedish Maritime Administration stores the full AIS message including all its attributes, such as Callsign, Course over ground, Draught Heading, IMO, Latitude, Longitude, MMSI and Type of cargo.

Data can be delivered in various formats according to the customer's wishes: Shape (vector data), PDF and/or raw data in comma-separated table (CSV) and in PNG format (raster). No official statistics are based on this data.

Data Scoring for the combination with MNO data

We believe that this type of data source holds huge potential for integration with MNO data, not least to enrich the MNO data source whenever possible. Hence, this source should not be completely disregarded, but considered as an important complement to MNO data.

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

4.3.3 Pollution statistics

Data description

There is no systematic data source in Sweden for producing pollution statistics. Instead, initiatives and various research studies are carried out on an ad hoc basis. One such study is titled 'High resolution air quality modelling of NO2, PM10 and PM2.5 for Sweden. A national study for 2019 based on dispersion modelling from regional down to street canyon level'. Below is the summary:

In this development project, concentrations of NO2, PM10 and PM2.5 were calculated for the entire country of Sweden for the year 2019. Simulations were conducted using a new methodology that enables a completely seamless combination of dispersion modelling on three scales - regional, urban and street scale - without double-counting emissions. Pollution levels were calculated at 50x50 m2 resolution, providing a comprehensive and detailed national dataset. This fine spatial resolution captures concentration gradients crucial for high-resolution exposure calculations. A key strength of using dispersion models to estimate pollutant levels is their direct connection to emission inventories and projections. New functionality, parameterizations and inputs were developed to increase model calculation performance while preserving storage capacity. This was crucial to carry out a comprehensive national modelling with high geographical resolution. Parameterizations and detailed input data were further developed to better represent real dispersion conditions, physical environment and the level of emissions from, for example, traffic. High levels of NO2 were observed in urban areas near heavily trafficked roads, with several locations exceeding the air quality limit values. The number of exceedances of current air quality standards were relatively low for PM10. PM2.5 levels were often low, with no exceedances of current standards. In a future perspective with stricter requirements for clean air, the situation will likely be different. With potentially tighter limit values, there is a risk for exceedances in several Swedish municipalities, especially for PM10. Validation of the modelling results compared to measurement data has shown that modelling quality objectives were achieved for PM2.5 at both urban and local traffic stations. For NO2 and PM10, the modelling quality objectives were not met. The model underperformed at a number of stations, failing to meet 90 % requirement. However, the Relative Directive Error (RDE) indicator was met at several stations, except for NO2 at traffic stations where the margin to fulfilment was very close. Further investigation of these sites is required and should be prioritized to understand the discrepancies and improve the accuracy of modelled concentrations.

Challenges associated with these data bases and ways of handling them

Model performance, memory and storage capacity remain major challenges for performing high-resolution calculations efficiently. Work in this area needs to be prioritised in future projects. The national modelling results provide a comprehensive description of the current state of air quality across all of Sweden's municipalities. The dataset enables the identification of locations where air pollution levels risk exceeding threshold values for air quality standards and environmental quality objectives. This information can be particularly valuable for municipalities that lack their own pollution measurements and modelling capabilities and support Swedish air pollution assessment and mitigation efforts. Having a comprehensive national assessment is especially important as the updated EU Ambient Air Quality Directive, with stricter requirements for clean air, is to be implemented in the coming years. The dataset can also support the design of measurement networks, the selection of appropriate measurement site locations and provide valuable

information to experts, researchers and interested members of the public. The results will be made freely available on the SMHI web portal "Luftwebb" by the turn of the year 2023/2024.

Accuracy for the Application scenarios involving the combination with MNO data

MNO data enables the estimation of day-time population. This information is often missing when analysing public health issues related to air pollution exposure. Indeed, exposure assessments are based on the resident population, whereas, during the day, people can be exposed to very different levels of air pollution depending on their locations and activities which can have substantial impact on their health.

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**

4.3.4 Electronic invoices

Data description

Electronic invoices (e-Fatura) can be defined as a mandatory invoice reporting system implemented by the Tax Administration as part of administrative simplification and anti-fraud measures. The electronic transmission of invoices issued by individuals or legal entities with their head office or permanent establishment in Portuguese territory to the Tax and Customs Authority is mandatory. This administrative data includes all invoices recorded electronically by the issuer, regardless of whether the buyer has requested an invoice. A detailed description of the variables is provided in the annex.

Challenges associated with these databases and ways of handling them

Two main challenges have been identified: ensuring regularity in the monthly transmission of data by the supplier and handling the high volume of records (100 million per month).

To address these challenges several treatments are applied to these big data bases:

- Identification and imputation of extreme outliers: positive taxable values equal to or exceeding 100 million euros and those exceeding 3 standard deviations from their respective means are flagged; for negative taxable values, the most extreme cases are briefly analysed.

- Correction of negative taxable values: values lower than -100,000 euros are corrected when a similar symmetrical value (between 95 and 100%) can be identified within the previous 4 months. These negative taxable values mainly result from corrections to incorrect values in prior

transmissions. The total taxable value for the set of related records remains unchanged, with only the temporal distribution being adjusted.

- Identification and imputation of missing values: in a small subset of more significant companies (in terms of number of employees and turnover) missing values are identified and imputed based on the company's behavior over time (historical series).

Accuracy for the Application scenarios involving the combination with MNO data

In all the following analyses, incorporating real-time and historical mobile (phone) presence data enhances the geographic resolution of the results. Particularly for hotels and restaurants, the integration with MNO data could also provide estimates of average transactional volume per visitor. For sectors such as pharmacies and supermarkets, the combination of transactional data with mobile presence data by area is highly valuable for crisis response and policymaking.

- Descriptive Statistics: Summarizing the data to provide insights into average monthly transactions, standard deviations, and seasonal trends.

- Predictive Modeling: Developing models to forecast future transactional volumes, 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 trends or declines and adjusting for annual seasonality.

Month-over-Month Comparison: Evaluating changes in transactions from one month to the next to detect short-term trends and seasonal variations.

Data analysis using the assessment matrix

The data is owned by the public administration (tax authority). The NSI and, to some extent, the scientific community can access the data. Access is free of charge and governed by a protocol. Outside the NSI it is only possible to access data aggregated by Issuers (sellers).

The detailed availability of such data across all EU countries still needs to be assessed, yet it is likely that these data are available only in a few countries.

As regards metadata, the variables required 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, a significant number of individual issuers lack 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 dataset does no't contain sensitive variables.

Information on the data source is publicly available since it is administrative data held by a public body. However, since the information is provided by companies rather than by individual establishments, this may result in inaccuracies in geographical variables.

Data Scoring for the combination with MNO data

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**

Regarding this analysis, the data seems sufficiently promising to be worth the cost of processing for integration with MNO data. Yet, the remaining work required for their integration makes them a secondary priority.

4.3.5 Tourism Household and Border Survey

Data description

Tourism Household survey

Statistics on tourism demand are collected EU-wide by NSIs, covering both domestic and outbound tourism. In Germany, 10.000 households are surveyed by phone about their travel behaviour.

General information:

- ★ base population: persons aged 15 and over living in Germany
- ★ sample: 10.000 (non-)travellers
- ★ statistical unit: Individuals in private households
- ★ legal basis: EU Regulation on European tourism statistics and implementing Regulation

Content:

*travel behaviour

*information on trips taken:

a) trips with overnight-stay: timing, nights spent, destination country, purpose, mode of transportation, type of accommodation, expenditures (+ additional questions every three years)
b) domestic day-trips: number, purpose, expenditure (only every three years)
c) international day-trips: number, purpose, expenditures (delivered annually by the Federal Central Bank since 2014)

★ socio-demographic information on travellers: number of persons, gender, age

Tourism border survey

Inbound and outbound Tourism can be captured through border surveys. Some EU countries may conduct a border survey instead of, or in addition to, a household survey. Advantages of border surveys compared to accommodation statistics, platform data, and household surveys include:

*Information on non-residents (including those staying in non-rented accommodation).
*Higher accuracy of trip(s) details due to promptness of response.
*Potential to be linked with MNO data and other data sources from the same time and location (e.g. passenger counts).

For instance, the Spanish border Survey (Frontur) is conducted monthly at Roads, Airports, Harbours and Train Stations, and covers approximately 450.000 non-residents.

Challenges associated with these data bases and ways of handling them

Challenges for the household survey

Respondents may not accurately remember all trips and/or travel details (especially expenditures) during a phone interview. Furthermore, some population groups might be underrepresented.
Although results are comparable across countries, data collection methods may differ between countries.

- Respondents may not accurately recall all travel details, particularly **expenditures**, during a phone interview
- Certain population groups may be **underrepresented**
- While results are **comparable across countries**, data collection methods may vary.

Accuracy for the Application scenarios involving the combination with MNO data

These data can contribute to improving population coverage. A detailed analysis of this data base will be presented in Deliverable D2.3.

Data analysis through the assessment matrix

Because of reporting obligations to Eurostat, all NSIs are required to collect information on travel behaviour via household surveys. Results are also published on the Eurostat website. Official statistics quality guidelines are met.

NSIs conducting border surveys compliy with official statistics quality guidelines. However, not all NSIs conduct travel surveys, which may limit the temporal comparability of these data across countries.

Data Scoring for the combination with MNO data (to be consolidated in D2.3)

Tourism Household and Border Survey			
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		**?**	
EU availability of these data		**moderate**	
Accuracy and robustness of the available information on these data		**moderate**	
Non-sensitivity of these data		**moderate**	

4.3.6 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 data bases 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 the Application scenarios involving 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 Scoring for the 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		**?**	
EU availability of these data			**high**
Accuracy and robustness of the information available on these data		**moderate**	
Non-sensitivity of these data			**high**

4.3.7 Satellite data

Data description

In the past three 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 of remote sensing data across a broader spectrum of statistical products, ranging from land use and land cover analysis to other applications.

Accuracy for the Application scenarios of the combination 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 4) several overarching themes of research emerge. These range from patterns of urban growth in China (land use), pollution, induced mortality/morbidity related to air-transmitted diseases and air pollution and ad-hoc statistical modelling mainly used for data fusion between mobile phone data and remote sensing data to a plethora of applied statistics on additional themes such as environment and agriculture (land cover and land use), resulting in a relatively heterogeneous coverage of research areas using the two target data sources, including also some inherent overlaps. These themes correspond to specific topics captured by figure 5, and are aggregated into four main trends based on their 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 emerging or declining themes (low relevance/low development). Accordingly, the analysis shows that mainstream research predominantly focuses on the use of mobile phones and remote sensing data for studying mobility within and between cities, relative to different topics (urban planning, health issues, etc).

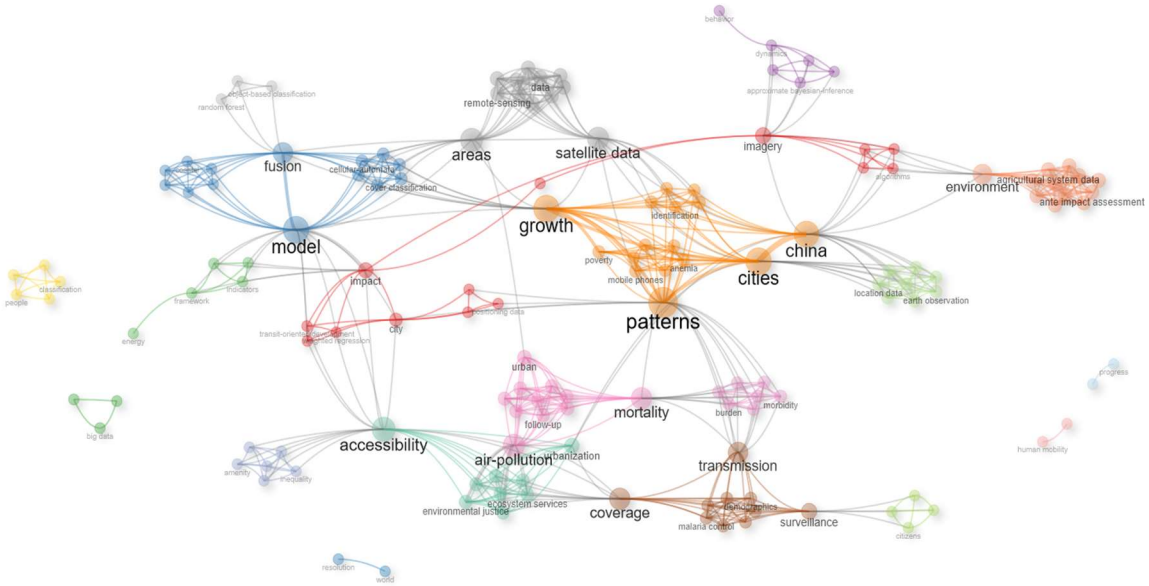


Figure 4. Network of research topics based on keyword frequency

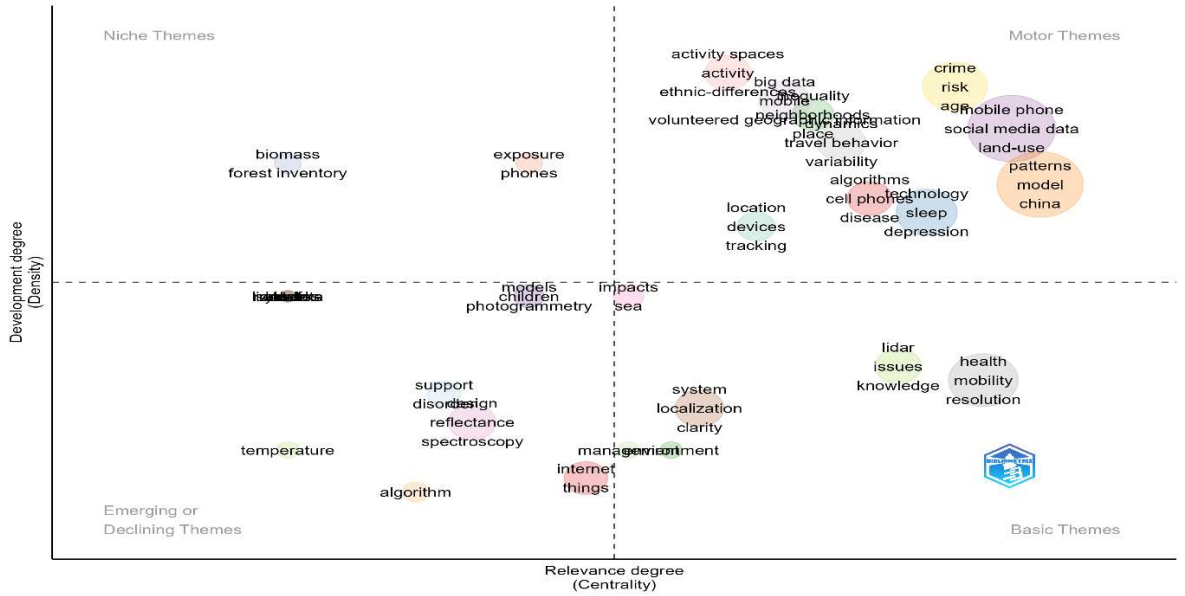


Figure 5. Thematic maps according to trends across time

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) Jia et al. 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 that integrates remote sensing and mobile data to analyze urban functional zones in Shenzhen, China, revealing a 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 important role that scale plays 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 be leveraged to 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, achieving notable improvement in accuracy for the Jakarta region.

The use of big data in disaster management and environmental monitoring has also 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) explored the use of non-authoritative data for disaster assessment, helping to fill gaps left by traditional remote sensing methods. Goldstein and Faxon (2022) examined the role of new data infrastructures in environmental monitoring in Myanmar, emphasizing the potential of digital transparency to enhance 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 for demographic statistics, transport planning, and studies on human behavior. Liu et al. (2015) incorporated spatial interaction patterns into urban land use classification, using taxi trip data to better understand the social functions of urban spaces. Health and environmental exposure have also been key areas of research. 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 inequities in artificial light at night (ALAN) exposure 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 value of human mobility data in economic modeling. Big data has also proven valuable in agricultural monitoring and evaluating development outcomes. Fishman et al. (2020) presented a data-driven approach to precision agriculture in small farms by 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 effectively monitor and evaluate progress towards sustainable development goals.

Innovations in data collection methods have enhanced both the accuracy and efficiency of a wide range of 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 on social and environmental equity has also benefited from 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 the use of household wealth proxies for studying socio-economic inequality in China, integrating diverse data sources to enhance the accuracy of wealth estimation.

Challenges associated with these databases and ways of handling them

Key limitations in the literature include the non-representativeness of mobile phone data (samples) and the coarse level of aggregation in some cases, such as using call detail records to estimate population density without calibration or validation from supplementary datasets. A common challenge across most of the selected studies is the lack of reproducibility and transparency, evidenced by the absence of published dataset and/or code.

Data Scoring for the 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. The reviewed studies provide plausible first-hand evidence of the potential of 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. The present study could be further enhanced by **expanding the literature base** to incorporate additional scientific databases (e.g. SCOPUS), refining search queries to better target themes pertaining official statistics and using more sophisticated methods of analysis (e.g. meta-analysis).

Satellite data			
Ease of handling the dataset	**low**		
Ease of access to a temporally and geographically harmonized data source.		**moderate**	
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			**high**

4.3.8 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 the Application scenarios of the combination 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 the 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**		

4.4. Less relevant sources

4.4.1 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.

Popular Times displays crowd levels in two primary forms: real-time data and historical data. Real-time data shows how busy a location is in real-time, which helps users avoid crowded places in the moment. Historical data, presented as bar graphs, indicate 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 on 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 with location services enabled on their mobile devices. This data is sourced from Google Location History, Google Maps activity, 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, helping to create a real-time picture of crowd density. Additionally, third-party apps that use Google's location services API share anonymized location data, supplementing the dataset. Once collected, the raw location data 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 multiple users at the same location and time intervals to form a more accurate picture. To further enhance accuracy, the data is smoothed and filtered to remove outliers

and noise. For instance, a brief visit by a single user to a location might be considered noise and removed if it does not reflect typical user behavior at that location. The core algorithm for estimating popular times relies on 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 at different times (see Figure 1.). Bar graph values range from 0 to 100, with 100 indicating the highest observed 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 refines future predictions to better accommodate similar events. For example, if a particular venue sees increased foot traffic every year during a specific festival, the algorithm incorporates this pattern into its predictions. Google takes several measures to protect 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. Furthermore, the data is aggregated across large groups of users, making it impossible to identify individual contributions.

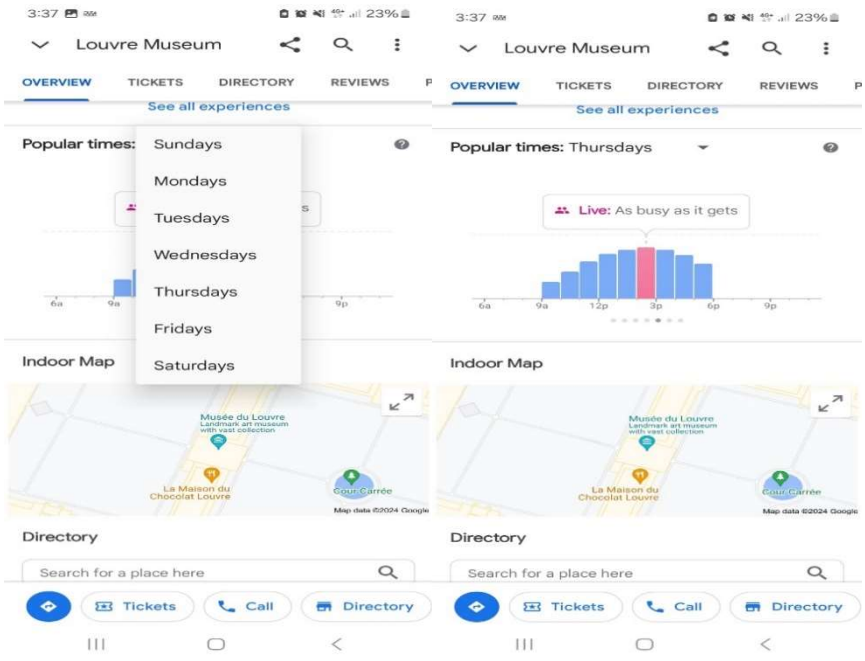


Figure 1. Google Popular Times (screen capture from Google Maps - Android OS)

Challenges associated with these data bases and ways of handling them

One important limitation is that Google does not provide an API (application programming interface) to download Google Maps Popular Times. The data is only available through the Google Maps graphic interface by manually inspecting the corresponding fields. In order to tackle this problem, web scrapers can be used to programmatically access and download data for specific locations, e.g. the Github repository m-wrzr/populartimes (2017). But this, in turn, raises serious legal and ethical concerns as web scraping breaks Google's Terms of Service and can result in legal consequences.

Moreover, Google does not provide documentation which contains a systematic and thorough description of the underlying algorithms used to compute Google Popular Times data.

Accuracy for the Application scenarios of the combination with MNO data

In order to explore whether previous studies had combined mobile phone data with Google Maps Popular Times, we conducted a preliminary search on the Web of Science database (Clarivate, 2024). We used multiple combinations of the terms "mobile phone data" and "Google Popular Times". Despite multiple iterations of the query, we failed to retrieve any published paper that used both data sources simultaneously. To provide any kind of evidence regarding the potential usefulness of Google Maps Popular Times, we refined our search by querying only the expression "Google Popular Times" to retrieve of all papers containing that expression in the title, keywords and abstract fields. The query yielded approximately 900 results. After a screening process we retained only 5 papers which matched our goal, i.e. to identify (some type of) applications which made use of Google Maps Popular Times data. The results of the bibliometric analysis should not be seen as an assessment of the accuracy of this data base to be combined with MNO data, but rather as a tool to start understanding the wealth of opportunities that could be offered combining these data with MNO data.

Some results

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

Mahdi et al. (2023) investigated the application 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 access to public transport, the study developed robust regression models. These models highlighted that Points of Interest categories significantly influenced visitor behaviour, 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 study found that essential services experienced less disruption than leisure-related activities, and that lower-income neighborhoods 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 better 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) combined Google Maps and OpenStreetMap data, including Google Popular Times, to predict venue popularity. The application of models such as gradient boosted regression 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 demand for electric vehicle charging at frequently visited locations such as gyms and shopping centres. By adopting a Monte Carlo-based approach, the study provided insights into potential electrical demand profiles for electric vehicle charging. This method offered a more realistic assessment of charging needs compared to traditional surveys, thus supporting more effective planning of electric vehicle infrastructure to meet future energy needs.

Data Scoring for integration with MNO data

Collectively, these studies highlight the potential of Google Popular Times data. However, judging by the limited number of effective publications identified, at present, the results are insufficient to justify even a small research track combining Google Popular Times with mobile phone data.

4.4.2 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 data bases 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 the Application scenarios of the combination with MNO data

When coverage is sufficiently high, smart meter data can complement MNO data to represent present population by assigning a home location to households. Energy consumption patterns can provide estimates of the household size. Energy consumption can indicate household size.

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 the combination with MNO data

This data appears to be a promising complement to MNO data for estimating day-time population. Yet access to data is complex, requiring partnerships with multiple private operators. Moreover, handling these large datasets demands advanced data-science skills. As a result, smart meter data has not yet been prioritized.

4.4.3 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.

Twitter/X data 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 data bases 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, like. trending digital topics s

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 the Application scenarios of the combination 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.

Twitter/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

Dataset size depends on purchased access level. For example, Twitter/X offers 1,500 free tweets per month and 1,000,000 tweets per month available at the Pro level and more at the enterprise level. Free data represents a 1% random sample of Tweets from the last 7-10 days. Purchased datasets include 500 tweets per request and provide geolocation information. Only the most premium subscription level offers filtering capabilities and real-time streaming access.

Free access is limited to very small random data samples. Paid access is available through tiered subscription levels (typically ranging from basic to pro/enterprise). Quality issues include representativeness, lack of user verification, and machine-generated content.

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.

4.4.4 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 data bases 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 the Application scenarios of the combination 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 the 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.

5. Conclusion and following steps

This second deliverable of Work Package 2 marks an important milestone in the landscaping analysis of non-MNO sources for integration with MNO data. It indeed sets the foundations of the scoring methodology that will enable the production of a consolidated short list of the most promising non-MNO sources. A preliminary, well-justified proposal of distinction in three major categories of sources, ranging from most promising to less relevant sources is presented. This proposal will guide the work planned for the remaining year of the ESSnet project. which consists in a deeper analysis of the most promising sources and broader discussions, facilitated by presentations at official statistics conferences. The final deliverable of WP2, due in August 2025, will present the consolidated version of this list along with the associated supporting arguments.

6. 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

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

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*.

ESSNet MNO-MINDS (2024) Deliverable 3.1 Preliminary Report on Methodology, Work Package 3 Methodologies and open source tools for integrating MNO and non-MNO data sources https://cros.ec.europa.eu/system/files/2025-01/WP3_D3.1_submitted.pdf

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

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*.

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

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

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

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.

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

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

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

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

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, 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

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

Github user m-wrzr (2017) “Populartimes”. URL: https://github.com/m-wrzr/populartimes

7. Annex

7.1. Useful links and information

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 Micro-Census
- 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

Micro census 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/

7.2. Parameters of the bibliometric analysis

7.2.1 For the 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

Timespan 1998:2023	Sources 41	Documents 48	Annual Growth Rate 4.49 %
Authors 241	Authors of single-authored docs 4	International Co-Authorship 35.42 %	Co-Authors per Doc 5.4
Author's Keywords (DE) 190	References 0	Document Average Age 5.79	Average citations per doc 28.75

Figure 1. Annual scientific production (number of published papers)

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 2.). 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.

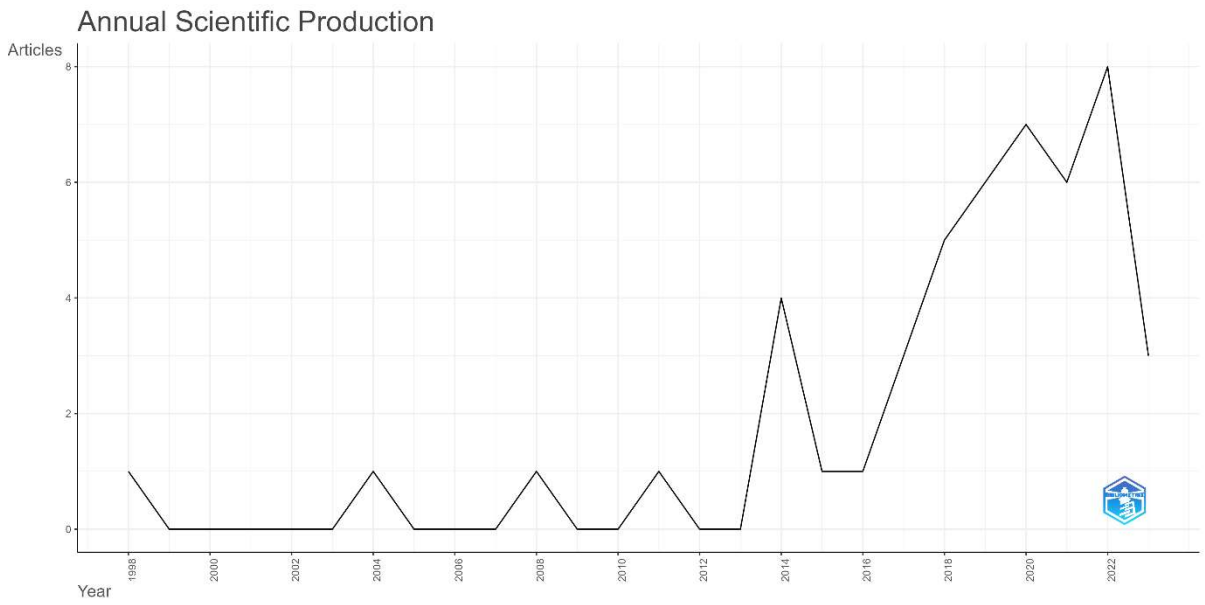


Figure 2. Annual scientific production (number of published papers)

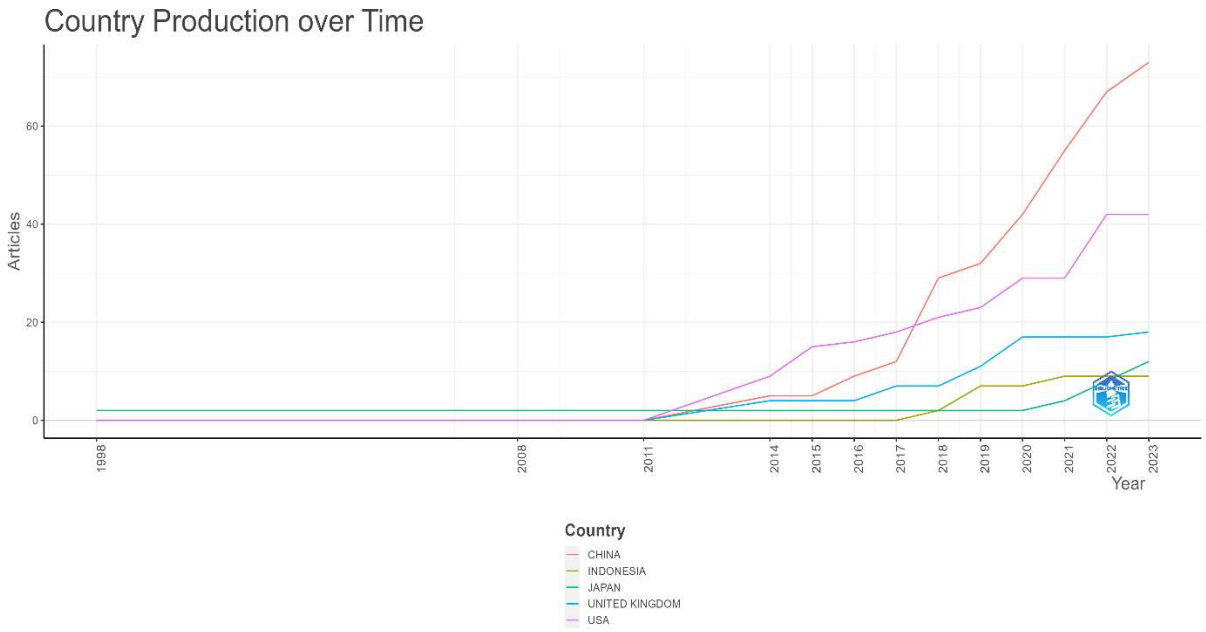


Figure3. Articles per country over time

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

In terms of contributing countries (figure 3.) China and United States of America (USA) lead the pack, accounting for more than 80% of published papers.

7.3 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