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 2.1
Report on schema and criteria to collect information on non-MNO data

March 2024
Partner in charge: INSEE (France) - *Marie-Pierre Joubert*

Authors: DESTATIS (Germany) - *Gloria Deetjen, Maurice Brandt*, INSEE (France) - *Marie-Pierre Joubert*, *Chloé Breton, Julien Pramil,* INSE-PT (Portugal) - *Sonia Quaresma, Antonio Portugal*, INSSE (Romania) - *Marian Necula, Bogdan Oancea*, SCB (Sweden) - *Remy Kamali, Pär Hammarström, Wictoria Widén, Stefan Svanström*

MNO-MINDS | Eurostat CROS (europa.eu)

Co-funded by
the European Union

Deliverable 2.1
Report on schema and criteria to collect information on non-MNO data

Summary

This report proposes an assessment matrix, which allows to analyse the relevance of non-MNO data for being combined with MNO data. It explains the reasons that led to the choices of analysis criteria and presents these choices. It also provides a first list of non-MNO sources and potential target statistics, along with a systematic identification of their main pros and cons, costs and gains, obtained thanks to the assessment matrix. It concludes by describing WP2's future steps and perspectives.

The deliverable is composed of three documents : in addition to this file, two documents in spreadsheet format contain information on the data's assessment.

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 in order to fight against 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, allowing to capture 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[1].

Many new data sources have already proven their potential for informing public policies. In particular, Mobile Network Operator (MNO) data has been much used among Official Statistics. During the COVID 19 crisis, MNO data were used in France and Germany to document the population's movements during lockdown, allowing to better calibrate public services. These data have also provided short term conjunctural indicators in order to estimate the crisis impact on the economy in a shorter period of time 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's mobility or tourist attendance. The future evolution of the legislation regarding access to private data might allow NSIs to gain access to more detailed data, yet discussions on this matter are still on-going[2].

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 MNO data with other data is a promising way of obtaining better official statistics, which are more accurate and cover a larger spectrum of issues. The first part of this deliverable goes deeper 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 how accurate non-MNO data are to meet 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

[1] 'New perspectives and priorities for EU 2030 Indicators – Indicators and methodologies for describing society in the Information Age" ESAC Doc. 2018/30

[2] The ESS has published in 2011 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 position paper on the future Data Act proposal – June 2021

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". Among the key high-level requirements that this framework should fulfil figures supporting the fusion of MNO data with non-MNO data. Indeed, although many useful use cases 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[3]. This report won't detail these pitfalls, yet it will focus on three main ways in which combining MNO data and non-MNO data can improve the quality of official statistics. This analysis will help us identify the most promising non-MNO sources to combine with MNO sources.

2.1. Improving population's coverage

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 NSI's usual data bases 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. Yet on the contrary 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's coverage, whether in terms of representativeness or 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.

2.2. Improving spatial and temporal precision of analyses

Combining different data sources also allows to provide more precise analyses. For instance, MNO data alone often don't include socio-demographic information; combining them with fiscal data about people's earnings (even aggregated), will allow to improve the study of socio-spatial segregation. In the same spirit, MNO origin-destination matrix often lack information about the transportation mode, or it is of very bad quality. This angle of analysis will focus on themes that already exist, where the combination of sources makes it possible to overcome a problem of data quality or data completeness.

2.3. Broadening the scope of issues covered by official statistics

Lastly, combining MNO and non-MNO sources allows to cover new topics of interest for official statistics. As explained in the introduction, given the major changes taking place in our environment,

[3] For instance : *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*. Coudin, E., Poulhes, M., & Suarez Castillo, M. (2021) Data & Pol or *Estimating the Residential Population from Mobile Phone Data, an Initial Exploration* – Economics and Statistics – 2019 – Sakarovitch, Bellefon, Givord, Vanhoof

this is an essential point in order for Official Statistics to keep providing up-to-date and relevant information.

3. Which type of non-MNO data ?

In order to distinguish between different non-MNO sources, this report draws on the work by the Norwegian NSI and ESSNets Big Data 1 and 2[4]. 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. **Remote sensing, fixed** (smart meters readings, whether station readings, traffic loop signals, lorry tracking signals,…)
e. **Remote sensing, mobile** (satellite images, drone images, airborne laser scanning, maritime AIS, …)
f. **Internet** (web pages, social media posts,…)

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

Some non-MNO data are produced by official statistics, such as survey-based Census. In this case the whole data production's 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 in order to meet the NSI's quality requirements, yet they are in general structured in a way that allows classical statistical treatments and the data providers have also an interest in offering the most exhaustive view of their population of interest. The last non-MNO data that this report consider are initially produced for objectives far removed from those of the official statistical service. The

[4] *Non-survey big data for official statistics:Sources, usability and statistical design*

producer can be either from the private or the public sector and these data are often not structured on a classical way, some lack documentation about the interest variables for NSIs or about the quality issues encountered in the data collection process.

In order to analyse the potential of these sources, WP2's members have adapted the 'Big Data classification matrix' which was produced by ESSNet BigData 2[5], to the specific question of the combination of MNO and non-MNO sources. Some questions are unchanged and others needed some adaptations. The resulting criteria are the following:

Data Type
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 ? Rq : different systems, not all the data owned by one company.
Is it possible to get access ? Does it have to be paid ?
Are there potentially competing uses (operator publishing similar statistics, etc.)?
Are there legal or ethical problems to access the data ?
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 ?
Is there a possibility to access the data to study its relevance ?
Is this data available in all EU countries ?
Metadata
Is the definition of the population known ? If not do you already have a method to address this issue ?
Is the base unit of the dataset known ?
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 other source) ?
Are there known quality issues with any of the variables ?
Does the data contain sensitive variable ? (Meaning legal or ethical issues related to its use)

[5] ESSNet Big Data 2, Workpackage K Methodology and Quality – Deliverable K7 : Typification matrix for big data projects - 2020

One important aspect of ESSNet's MNO MINDS is that it aims at producing official statistics, and not only experimental statistics. The short list of the most promising data sources will therefore be established taking into account the availability of sources in all NSIs and their respects of the European statistics code of practice. Not all aspects of the code are relevant for analysing new data sources. The following aspects where selected :

Official statistics quality standards
Is the scope, detail and cost of this source commensurate with needs ?
Is it possible to regularly monitor output quality ?
Confidentiality and data protection : 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 ?
Impartiality and Objectivity : are information on data sources, methods and procedures publicly available ? Is there a process to inform the NSI in case an error is discovered ?
Appropriate statistical procedures : are the definitions and concepts used in this source a good approximation of the concepts required for statistical purposes ? Do the data holders collaborate with the NSI in improving data quality (take into account feedback,...)

5. First set of data evaluated

WP2's members began their focus on thirteen specific sources. The list will be broadened and enriched after the Sprint event in June. Results are presented in annexed excel files:

WP2_D2.1_matrix_census_registers entails the analysis of:

- the German census, which is register-based and integrates data from administrative registers with supplementary surveys

- the German microcensus, which is survey based and entails questions from SILC and LFS survey

- the French census, which is totally survey based contrary to the german one

- the Swedish total population register

- the Swedish national travel survey

- the Portuguese survey about non-resident tourist arrivals in Portugal

WP2_D2.1_matrix_nontraditional entails the analysis of:

- satellite data (Germany and Romania)

- land use and land cover register (Romania)

- traffic signalling (Sweden)

- electronic invoices (Portugal)

- social media (Germany)

- credit card transaction data (France)

- Google Popular time (Romania)

6. Conclusion and following steps

After these first months of working on work package 2, we have established a first proposition of assessment matrix. This proposition has been discussed among us and with the members of the Task Force MNO data. It will be adjusted at the margin in the course of future feedbacks.

Moreover, we have proposed a first assessment of data, through the matrix's lenses. In future months, we will focus on work in following directions:

- deepening our understanding of the data we have already identified in order to precise our assessment of their usefulness to be combined with MNO data

- exploring other sources, which we didn't have time to analyse for this first deliverable

- proposing some potential target statistics derived from the combination of MNO and non-MNO data and assessing their impact on existing statistics

- preparing the June's Sprint event, in order for it to allow us receiving as much new ideas and interesting feedback as possible.

- assessing actual or prospective availability of non-MNO data sources across different countries. If necessary, a short survey through the ESS will be carried out for this purpose.

7. Annexes

Annex 1: WP2_D2.1_matrix_census_registers

Sheet 1: DE_Census

Questions on the non-MNO source			
Data Type	**Description**	**Challenges**	**Treatments**
a.1 **Survey** (census, SILC, employment,…)			
a.2 **Combination of survey and register**	German Census is register-based and integrates data from administrative registers with supplementary surveys	High in terms of quality standards but published once every ten years.	
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			
d. **Remote sensing, fixed** (smart meters readings, whether station readings, traffic loop			
e. **Remote sensing, mobile** (satellite images, drone images, airborne laser scanning,			
f. **Internet** (web pages, social media posts,…)			
Do you know the size of the data set ? Will it be a problem to treat it at once ? Will you split it	Yes, the size of the data set is well known. Depending on the research question, not		
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 ?	Structure is known. Several options are available, depending on the research question.		
Relevance to the MNO data uses cases			
Improving population coverage : are these data complementary to MNO data ? Do they allow to improve MNO data population representativeness ? Are there difficulties to due to different irreconcialable concepts (definitions, time periods, …) ? Improve of timeliness of analyses and geographical precision.	Census results are complementary to MNO data because they provide a base for the total population size whenever the available MNO data does not comprise data from all MNOs. Census results are high in quality but only published once every ten years, which makes it challenging to obtain a recent population figure in the later following years after publication.	High quality data source but depending of the year, it might not provide the most recent information.	
Improving precision of analyses : do these data allow to provide preciser analyses (ex : socio-economic indicators, indicators of the mobility type, indicators of the commercial activity,…)	In general yes, but it depends on the research question.		
Broadening the scope of issues covered by official statistics : do these data, combined with MNO data, allow to cover new topics (exposure to air pollution, commuting CO2 emissions,…) ?	Census can serve as a base for further analysis together with MNO data in combination with other digital data. Whereas digital data can widen the scope for analysis, census provides a trustable population figure and potentials for validation and plausibility tests.		

Access			
Who owns the data ? Public administration, one company, several companies ? Rq : different systems, not all the data owned by one company.	NSIs hold census results.		
Is it possible to get access ? Does it have to be paid ?	Access: Yes. Paid: No.		
Are there potentially competing uses (operator publishing similar statistics, etc.)?	No.		
Are there legal or ethical problems to access the data ?	No.		
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 ?	Limitations are on the individual data level for confidentiality reasons. However, the official number of residents does not contain any data manipulation.		
Is there a possibility to access the data to study its relevance ?	Yes. Census results are published. In addition, access to micro data can be granted for research purposes (see below).		
Is this data available in all EU countries ?	Census results are available for all EU countries that conduct a census. However, the approaches might differ (e.g. register, survey, or both). All ESS members produce statistics on their number of residents.		https://ec.europa.eu/CensusHub2/selectHyperCube?countrycode=en&clearSession=true
Metadata			
Is the definition of the population known ? If not do you already have a method to address this issue ?	Yes.		
Is the base unit of the dataset known ?	Yes.		
Do the units have an identifier ?	Yes, depending on which level and the research question		
Do you have the necessary variables to reach the relevant granularity level for the statistical unit ?	Yes.		
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 other source) ?	Census already provides useful information together with MNO data but depending on the research purpose, it can be enhanced with e.g. further administrative or survey data, as well as with other digital data		
Are there known quality issues with any of the v	No.		
Does the data contain sensitive variable ? (Meaning legal or ethical issues related to its use)	No. Only when combining multiple data sources, confidentiality and privacy issues should always be considered (this holds for all non-MNO data sources).		

Official statistics quality standards			
Is the scope, detail and cost of this source commensurate with needs ?	Census usually represents one of the core statistics that NSIs publish and is therefore a trustable and detailed data source to be combined with MNO data. It is in scope, because NSIs are census experts themselves and possess the results anyway.		
Is it possible to regularly monitor output quality ?	Yes, for the NSI it is possible to monitor the output quality.		
Confidentiality and data protection : Is it possible to put in place the necessary regulatory, administrative, techincal and organisational measures to protect the security and integrity of statistical data and their transmission ?	Confidentiality and data protection are in place for census results.		
Impartiality and Objectivity : are information on data sources, methods and procedures publicly available ? Is there a process to inform the NSI in case an error is discovered ?	Data sources, methods and procedures are publicly available except for parts that are confidential. However, those are known to the NSI.		
Appropriate statistical procedures : are the definitions and concepts used in this source a good approximation of the concepts required for statistical purposes ? Do the data holders collaborate with the NSI in improving data quality (take into account feedback,...)	Yes.		
Accuracy and reliability : are source data, integrated data, intermediate results and statistical outputs regularly assessed and validated ?	Yes, they are assessed and validated during and after production. As the publication happens once in 10 years, there is no need for a very frequent assessment.	Results might be less accurate/relevant with time.	For more recent population figures, there is the intercensal population update (based on census) published annually. For analysis regarding years that are already quite far from the last available census, MNO data together with intercensal population updates (and census) could be used for more recent results.
Are sampling errors and non-sampling errors measured and systematically documented ?	Yes.		
Are revisions regularly analysed in order to improve source data, statistical processes and outputs ?	Published census results are the final results. Therefore, there are no preliminary results and neither revisions.		
Timeliness an punctuality : is it possible to release statistics based on this data source on a regular basis ?	Yes. Census is published every 10 years and represents the source e.g. for population figures with the highest quality available.	The current time interval of ten years is quite long and does not provide the most recent numbers after some time.	Census can be used in combination with MNO data and possibly further (more recent) non-MNO data. In this case, census can serve as a base and ensure quality.
Coherence and Comparability : are statistics based on this source coherent, consistent and comparable over a reasonable period of time ?	Yes, census is spatially comparable for definitions and classification (standardised EU-wide). Partial comparability to census prior 1987 (because of German reunification and changes in methods). In general, census results are coherent. Minor inconsistencies could result from confidentiality procedures.		As MNO data is relevant for the years later than 1987, this should not represent a challenge.
Is cross-national comparability of statistics based on this source possible ? -> has to be discussed between WP members for comparability	Because of EU standards, it is comparable across countries. Still, NSIs might use different strategies to obtain census results (e.g. survey based)		
Is access to microdata allowed for research purposes ?	Yes, access to microdata for research purposes is possible at the (federal and state) NSIs' Research Data Centres*		https://forschungsdatenzentrum.de/en

Sheet 2_DE_microcensus

Questions on the non-MNO source			
Data Type	**Description**	**Challenges**	**Treatments**
a.1 **Survey** (census, SILC, employment,…)	The "Mikrocensus" (microcensus) is a yearly survey of about 1% of the population (ca. 810 000 persons). It incorparates mandated EU-Surveys like SILC and LFS and national questions regarding: - 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	Legability of linkage with other datasources requires a legal check	
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 ?	Yes, the size of the data set is well known. Depending on the research question, not all information from the microcensus is needed and therefore a suitable data set can be selected.		
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 ?	Structure is known. Several options are available, depending on the research question.		
Relevance to the MNO data uses cases			
Improving population coverage : are these data complementary to MNO data ? Do they allow to improve MNO data population representativeness ? Are there difficulties to due to different irreconcialable concepts (definitions, time periods, …) ? Improve of timeliness of analyses and geographical precision.	Microcensus data is not usable for population coverage as the sample population is not large enough for total population estimates.		
Improving precision of analyses : do these data allow to provide preciser analyses (ex : socio-economic indicators, indicators of the mobility type, indicators of the commercial activity,…)	In general yes, but it depends on the research question.		
Broadening the scope of issues covered by official statistics : do these data, combined with MNO data, allow to cover new topics (exposure to air pollution, commuting CO2 emissions,…) ?	Microcensus can serve as a base for further analysis together with MNO data in combination with other digital data. Whereas digital data can widen the scope for analysis, microcensus provides potentials for validation and plausibility tests.		

Access			
Who owns the data ? Public administration, one company, several companies ? Rq : different systems, not all the data owned by one company.	NSIs hold microcensus results.		
Is it possible to get access ? Does it have to be paid ?	Access: Yes. Paid: No.		
Are there potentially competing uses (operator publishing similar statistics, etc.)?	No.		
Are there legal or ethical problems to access the data ?	Legability of linkage with other datasources requires a legal check		
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 ?	Limitations are on the individual data level for confidentiality reasons.		
Is there a possibility to access the data to study its relevance ?	Yes. Microcensus results are published. In addition, access to micro data can be granted for research purposes (see below).		
Is this data available in all EU countries ?	Microcensus results are available for all EU countries that conduct a microcensus. EU-wide surveys are (Silc, LFS, IKT) are availble for all EU-Countries		
Metadata			
Is the definition of the population known ? If not do you already have a method to address this issue ?	Yes.		
Is the base unit of the dataset known ?	Yes.		
Do the units have an identifier ?	Yes, depending on which level and the research question		
Do you have the necessary variables to reach the relevant granularity level for the statistical unit ?	Yes.		
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 other source) ?	Microcensus already provides useful information together with MNO data but depending on the research purpose, it can be enhanced with e.g. further administrative or survey data, as well as with other digital data		
Are there known quality issues with any of the variables ?	Yes. As a large yearly survey there are naturally issues with changes in methods for example. The yearly Qualitiy Report (https://www.destatis.de/DE/Methoden/Qualitaet/Qualitaetsberichte/Bevoelkerung/mikrozensus-2021.html - only available in German) contains detailed information on identified issues.		https://www.destatis.de/DE/Methoden/Qualitaet/Qualitaetsberichte/Bevoelkerung/mikrozensus-2021.html
Does the data contain sensitive variable ? (Meaning legal or ethical issues related to its use)	No. Only when combining multiple data sources, confidentiality and privacy issues should always be considered (this holds for all non-MNO data sources).		

Official statistics quality standards			
Is the scope, detail and cost of this source commensurate with needs ?	Microcensus usually represents one of the core statistics that NSIs publish and is therefore a trustable and detailed data source to be combined with MNO data. It is in scope, because NSIs are microcensus experts themselves and possess the results anyway.		
Is it possible to regularly monitor output quality ?	Yes, for the NSI it is possible to monitor the output quality.		
Confidentiality and data protection : Is it possible to put in place the necessary regulatory, administrative, techincal and organisational measures to protect the security and integrity of statistical data and their transmission ?	Confidentiality and data protection are in place for microcensus results.		
Impartiality and Objectivity : are information on data sources, methods and procedures publicly available ? Is there a process to inform the NSI in case an error is discovered ?	Data sources, methods and procedures are publicly available except for parts that are confidential. However, those are known to the NSI.		
Appropriate statistical procedures : are the definitions and concepts used in this source a good approximation of the concepts required for statistical purposes ? Do the data holders collaborate with the NSI in improving data quality (take into account feedback,...)	Yes.		
Accuracy and reliability : are source data, integrated data, intermediate results and statistical outputs regularly assessed and validated ?	Yes, they are assessed and validated during and after production.		
Are sampling errors and non-sampling errors measured and systematically documented ?	Yes.		
Are revisions regularly analysed in order to improve source data, statistical processes and outputs ?	Microcensus result is published in two waves. First results are based on a smaller sample size, whereas final results are based on the full sample. Data is only revised if made necessary by a significant error.		
Timeliness an punctuality : is it possible to release statistics based on this data source on a regular basis ?	Yes. Microcensus data is published every year		Microcensus can be used in combination with MNO data and possibly further (more recent) non-MNO data. In this case, census can serve as a base and ensure quality.
Coherence and Comparability : are statistics based on this source coherent, consistent and comparable over a reasonable period of time ?	Due to far-reaching methodological and structural changes made in 2020 the content and methodology of the results can only be related to results from 2019 and earlier to a limited extent.		
Is cross-national comparability of statistics based on this source possible ? -> has to be discussed between WP members for comparability	Because of EU standards, it is comparable across countries		
Is access to microdata allowed for research purposes ?	Yes, access to microdata for research purposes is possible at the (federal and state) NSIs' Research Data Centres*		https://forschungsdatenzentrum.de/en

Sheet 3_FR_census

Questions on the non-MNO source			
Data Type	**Description**	**Challenges**	**Treatments**
a.1 Survey (census, SILC, employment,…)	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. It also provides statistics on the characteristics of inhabitants (gender, age, occupation, nationality, mode of transport, etc.) and their homes (type of dwelling, type of construction, number of rooms, etc.).		**National results use 5 annual surveys**
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. Remote sensing, fixed (smart meters readings, whether station readings, traffic loop signals, lorry tracking signals,…)			
e. Remote sensing, mobile (satellite images, drone images, airborne laser scanning, maritime AIS, …)			
f. Internet (web pages, social media posts,…)			
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 ?	Yes, the size of the dataset is known. No difficulty in processing it in its entirety.		
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 ?	Data structure is known within INS. Geolocation variables are known.	Sub-municipal geolocation only available for the 2017 census results.	
Relevance to the MNO data uses cases			
Improving population coverage : are these data complementary to MNO data ? Do they allow to improve MNO data population representativeness ? Are there difficulties to due to different irreconcialable concepts (definitions, time periods, …) ? Improve of timeliness of analyses and geographical precision.	Census data allow to improve MNO representativeness, especially with night time comparisons (to ease concpt reconciliation with residential population).		
Improving precision of analyses : do these data allow to provide analyses (ex : socio-economic indicators, indicators of the mobility type, indicators of the commercial activity,…)	Yes, it brings multiple socio-demographic pieces of information		
Broadening the scope of issues covered by official statistics : do these data, combined with MNO data, allow to cover new topics (exposure to air pollution, commuting CO2 emissions,…) ?	Census data are a major part of official statistics.		
Access			
Who owns the data ? Public administration, one company, several companies ? Rq : different systems, not all the data owned by one company.	NSIs hold census data.		
Is it possible to get access ? Does it have to be paid ?	Access: Yes. Paid: No.		
Are there potentially competing uses (operator publishing similar statistics, etc.)?	No.		
Are there legal or ethical problems to access the data ?	Legability of linkage with other datasources requires a legal check		
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 ?	Limitations are on the individual data level for confidentiality reasons.		
Is there a possibility to access the data to study its relevance ?	Yes. Microcensus results are published. In addition, access to micro data can be granted for research purposes (see below).		
Is this data available in all EU countries ?	Yes		

Metadata			
Is the definition of the population known ? If not do you already have a method to address this issue ?	Yes.		
Is the base unit of the dataset known ?	Yes.		
Do the units have an identifier ?	Yes		
Do you have the necessary variables to reach the relevant granularity level for the statistical unit ?	Yes.		
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 ?	No.		
Is there auxiliary information to make the data set useful with auxiliary data (NSI or other source) ?			
Are there known quality issues with any of the variables ?			
Does the data contain sensitive variable ? (Meaning legal or ethical issues related to its use)	No.		
Official statistics quality standards			
Is the scope, detail and cost of this source commensurate with needs ?			
Is it possible to regularly monitor output quality ?	Yes		
Confidentiality and data protection : Is it possible to put in place the necessary regulatory, administrative, techincal and organisational measures to protect the security and integrity of statistical data and their transmission ?	Yes		
Impartiality and Objectivity : are information on data sources, methods and procedures publicly available ? Is there a process to inform the NSI in case an error is discovered ?	Yes		
Appropriate statistical procedures : are the definitions and concepts used in this source a good approximation of the concepts required for statistical purposes ? Do the data holders collaborate with the NSI in improving data quality (take into account feedback,…)	Yes.		
Accuracy and reliability : are source data, integrated data, intermediate results and statistical outputs regularly assessed and validated ?	Yes, as long as statistical confidentiality is respected		
Are sampling errors and non-sampling errors measured and systematically documented ?	Yes.		
Are revisions regularly analysed in order to improve source data, statistical processes and outputs ?	Yes		
Timeliness an punctuality : is it possible to release statistics based on this data source on a regular basis ?	Yes		
Coherence and Comparability : are statistics based on this source coherent, consistent and comparable over a reasonable period of time ?	Yes, at least since 2006		
Is cross-national comparability of statistics based on this source possible ? -> has to be discussed between WP members for comparability	Because of EU standards, it is comparable across countries		
Is access to microdata allowed for research purposes ?	Yes		

Sheet 4_SW_register

Total Population Register (TPR)	
Description	The Total Population Register contains data that is essential for information about Sweden's population. The system includes registers that are the bases for official population statistics and provide basic data for large parts of Statistics Sweden's operations. The Total Population Register contains information about the population and its changes, and to a large extent reflects the content of the population register of the Swedish Tax Agency. Contains these variables: * 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)
Challenges	Undercoverage of non-registered people in Sweden.
Treatments	The Total Population Register is 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 samples of individuals 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 restriction 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 - and- data for authentication in connection with the request of extracts of registers in accordance with the Personal Data Act.
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 ?	Not a dataset, register.
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 ?	See above.

Relevance to the MNO data uses cases	
Improving population coverage : a. are these data complementary to MNO data ? b. Do they allow to improve MNO data population representativeness? c. Are there difficulties to due to different irreconcialable concepts (definitions, time periods, …)? d. Improve of timeliness of analyses and geographical precision.	a. Yes b. Yes c. No d. No
Improving precision of analyses : do these data allow to provide preciser analyses (ex : socio-economic indicators, indicators of the mobility type, indicators of the commercial activity,…)	Combining with MNO to improve dynamic populations movements, day/night population patterns. See treatments information.
Broadening the scope of issues covered by official statistics : do these data, combined with MNO data, allow to cover new topics (exposure to air pollution, commuting CO2 emissions,…) ?	Yes definitely.
Access	
Who owns the data ? Public administration, one company, several companies ? Rq : different systems, not all the data owned by one company.	SCB.
Is it possible to get access ? Does it have to be paid ?	Yes.
Are there potentially competing uses (operator publishing similar statistics, etc.)?	No.
Are there legal or ethical problems to access the data ?	Depends.
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 ?	No.
Is there a possibility to access the data to study its relevance ?	Yes.
Is this data available in all EU countries ?	Subject to investigation
Metadata	See link https://metadata.scb.se/mikrodataregister.aspx?produkt=BE0102
Is the definition of the population known ? If not do you already have a method to address this issue ?	Yes
Is the base unit of the dataset known ?	Yes
Do the units have an identifier ?	Yes
Do you have the necessary variables to reach the relevant granularity level for the statistical unit ?	Yes
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 ?	No
Is there auxiliary information to make the data set useful with auxiliary data (NSI or other source) ?	Unsure
Are there known quality issues with any of the variables ?	See Quality declarations
Does the data contain sensitive variable ? (Meaning legal or ethical issues related to its use)	Probably yes
Official statistics quality standards	Official statistics
Is the scope, detail and cost of this source commensurate with needs ?	Yes
Is it possible to regularly monitor output quality ?	Yes
Confidentiality and data protection : Is it possible to put in place the necessary regulatory, administrative, techincal and organisational measures to protect the security and integrity of statistical data and their transmission ?	Yes
Impartiality and Objectivity : are information on data sources, methods and procedures publicly available ? Is there a process to inform the NSI in case an error is discovered ?	Yes
Appropriate statistical procedures : are the definitions and concepts used in this source a good approximation of the concepts required for statistical purposes ? Do the data holders collaborate with the NSI in improving data quality (take into account feedback,…)	Yes
Accuracy and reliability : are source data, integrated data, intermediate results and statistical outputs regularly assessed and validated ?	Yes
Are sampling errors and non-sampling errors measured and systematically documented ?	Yes
Are revisions regularly analysed in order to improve source data, statistical processes and outputs ?	Yes
Timeliness an punctuality : is it possible to release statistics based on this data source on a regular basis ?	Yes
Coherence and Comparability : are statistics based on this source coherent, consistent and comparable over a reasonable period of time ?	Yes
Is cross-national comparability of statistics based on this source possible ? -> has to be discussed between WP members for comparability	Probably yes
Is access to microdata allowed for research purposes ?	Yes

Sheet 5_ SW_National_Travel_Survey

The National Travel Survey (RVU Sweden)
Statistical Agency: Transport Analysis
Type: Survey
Source: *https://www.trafa.se/en/travel-survey/travel-survey/*

Description	Transport Analysis is responsible for the national travel survey, RVU Sweden. The survey is about people's daily travel, at what times the trips are made, what modes of transport are used, and what is the purpose of the trip. Since 2019, a travel survey has been underway and is presented in annual reports. The survey covers Sweden's population aged 6–84 years and is conducted as a combined paper and web survey. The sample is stratified by county, age group and gender of the dwelling. The sample for 2022 includes just under 12,600 people in Sweden.
Challenges	The survey has sources of uncertainty that are largely due to the fact that it is a questionnaire for a sample of people. The sources of uncertainty include problems with non-response, as well as coverage of target population and measurement errors. (Around 70 % non-response rate). Other challenges related to the cost of the survey and the burden for the respondents.
Treatments	Data on travel habits provide important background information for the formulation of national and regional transport policy, for the development of infrastructure and traffic supply, for road safety work and for research and development on people's travel and communication patterns.
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 ?	Normal data sets stemming from Surveys. See tables in the link above. No problem of treatment. No splits 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 ?	Simple data set structure.
Relevance to the MNO data uses cases	
Improving population coverage : a. are these data complementary to MNO data ? b. Do they allow to improve MNO data population representativeness? c. Are there difficulties to due to different irreconcialable concepts (definitions, time periods, …)? d. Improve of timeliness of analyses and geographical precision.	a. Yes. Travel Survey contributes with crucial background informations such as gender and age. The Survey also addresses people's daily travel, the dates and times when they travel, the modes of transport used, and the purposes of the trips. b. Yes, see above. c. No difficulties. d. The Survey will not improve timeliness, but probably improve/complement geo precisions.
Improving precision of analyses : do these data allow to provide preciser analyses (ex : socio-economic indicators, indicators of the mobility type, indicators of the commercial activity,…)	Yes, with comprehensive background information and different behaviour patterns.
Broadening the scope of issues covered by official statistics : do these data, combined with MNO data, allow to cover new topics (exposure to air pollution, commuting CO2 emissions,…) ?	Potentially Yes.
Access	
Who owns the data ? Public administration, one company, several companies ? Rq : different systems, not all the data owned by one company.	Public agency.
Is it possible to get access ? Does it have to be paid ?	Access free.
Are there potentially competing uses (operator publishing similar statistics, etc.)?	No.
Are there legal or ethical problems to access the data ?	No.
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 ?	No.
Is there a possibility to access the data to study its relevance ?	Yes.
Is this data available in all EU countries ?	Should be subject to investigation

Metadata	
Is the definition of the population known ? If not do you already have a method to address this issue ?	Yes, see the link.
Is the base unit of the dataset known ?	Yes, personal number.
Do the units have an identifier ?	Yes.
Do you have the necessary variables to reach the relevant granularity level for the statistical unit ?	Yes.
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 ?	No.
Is there auxiliary information to make the data set useful with auxiliary data (NSI or other source) ?	Unsure.
Are there known quality issues with any of the variables ?	See the link.
Does the data contain sensitive variable ? (Meaning legal or ethical issues related to its use)	No.
Official statistics quality standards	Official statistics standard
Is the scope, detail and cost of this source commensurate with needs ?	See official statistics
Is it possible to regularly monitor output quality ?	See official statistics
Confidentiality and data protection : Is it possible to put in place the necessary regulatory, administrative, techincal and organisational measures to protect the security and integrity of statistical data and their transmission ?	See official statistics
Impartiality and Objectivity : are information on data sources, methods and procedures publicly available ? Is there a process to inform the NSI in case an error is discovered ?	See official statistics
Appropriate statistical procedures : are the definitions and concepts used in this source a good approximation of the concepts required for statistical purposes ? Do the data holders collaborate with the NSI in improving data quality (take into account feedback,...)	See official statistics
Accuracy and reliability : are source data, integrated data, intermediate results and statistical outputs regularly assessed and validated ?	See official statistics
Are sampling errors and non-sampling errors measured and systematically documented ?	See official statistics
Are revisions regularly analysed in order to improve source data, statistical processes and outputs ?	See official statistics
Timeliness an punctuality : is it possible to release statistics based on this data source on a regular basis ?	See official statistics
Coherence and Comparability : are statistics based on this source coherent, consistent and comparable over a reasonable period of time ?	See official statistics
Is cross-national comparability of statistics based on this source possible ? -> has to be discussed between WP members for comparability	See official statistics
Is access to microdata allowed for research purposes ?	See official statistics

Sheet 6_ PT_tourist survey

Questions on the non-MNO source			
Data Type	**Description**	**Challenges**	**Treatments**
a.1 **Survey** (census, SILC, employment,…)	International tourism statistics: Non-resident tourist arrivals in Portugal		
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. **Remote sensing, fixed** (smart meters readings, whether station readings, traffic loop signals, lorry tracking signals,…)			
e. **Remote sensing, mobile** (satellite images, drone images, airborne laser scanning, maritime AIS, …)			
f. **Internet** (web pages, social media posts,…)			
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 ?			
Relevance to the MNO data uses cases			
Improving population coverage : are these data complementary to MNO data ? Do they allow to improve MNO data population representativeness ? Are there difficulties to due to different irreconcialable concepts (definitions, time periods, …) ? Improve of timeliness of analyses and geographical precision.	Together with other data sources, MNO data can be a valuable input to an alternative model for estimating the arrival of non-resident tourists in Portugal. The last survey was carried out almost 10 years ago and our aim is to develop an alternative model for compiling statistics on non-resident tourist arrivals from various complementary sources, including MNO.		
Improving precision of analyses : do these data allow to provide preciser analyses (ex : socio-economic indicators, indicators of the mobility type, indicators of the commercial activity,…)	Idem.		
Broadening the scope of issues covered by official statistics : do these data, combined with MNO data, allow to cover new topics (exposure to air pollution, commuting CO2 emissions,…) ?	The combination of different sources with MNO data should allow the continuous dissemination of more detailed international tourism statistics.		
Access			
Who owns the data ? Public administration, one company, several companies ? Rq : different systems, not all the data owned by one company.	Statistics Portugal		
Is it possible to get access ? Does it have to be paid ?	It is possible to have access to aggregated data.		
Are there potentially competing uses (operator publishing similar statistics, etc.)?	No		
Are there legal or ethical problems to access the data ?			
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 ?			
Is there a possibility to access the data to study its relevance ?			
Is this data available in all EU countries ?			
Metadata			
Is the definition of the population known ? If not do you already have a method to address this issue ?	Yes		
Is the base unit of the dataset known ?			
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 other source) ?	Yes. All other data sources used in the current estimation model.		
Are there known quality issues with any of the variables ?	No.		
Does the data contain sensitive variable ? (Meaning legal or ethical issues related to its use)	No.		

Official statistics quality standards			
Is the scope, detail and cost of this source commensurate with needs ?	Yes.		
Is it possible to regularly monitor output quality ?	Yes.		
Confidentiality and data protection : Is it possible to put in place the necessary regulatory, administrative, techincal and organisational measures to protect the security and integrity of statistical data and their transmission ?	Yes.		
Impartiality and Objectivity : are information on data sources, methods and procedures publicly available ? Is there a process to inform the NSI in case an error is discovered ?	Yes.		
Appropriate statistical procedures : are the definitions and concepts used in this source a good approximation of the concepts required for statistical purposes ? Do the data holders collaborate with the NSI in improving data quality (take into account feedback,...)	Yes.		
Accuracy and reliability : are source data, integrated data, intermediate results and statistical outputs regularly assessed and validated ?	Yes.		
Are sampling errors and non-sampling errors measured and systematically documented ?	Not applicable		
Are revisions regularly analysed in order to improve source data, statistical processes and outputs ?	Yes.		
Timeliness an punctuality : is it possible to release statistics based on this data source on a regular basis ?	Annually		
Coherence and Comparability : are statistics based on this source coherent, consistent and comparable over a reasonable period of time ?	Yes.		
Is cross-national comparability of statistics based on this source possible ? -> has to be discussed between WP members for comparability	Yes.		
Is access to microdata allowed for research purposes ?	Not applicable		

Annex 2: WP2_D2.1_matrix_nontraditional

Sheet 1: DE_satellite_image

Questions on the non-MNO source			
Data Type	**Description**	**Challenges**	**Treatments**
a.1 **Survey** (census, SILC, employment,…)			
e. **Remote sensing, mobile** (satellite images, drone images, airborne laser scanning, maritime AIS, …)	Satellite data are signals from one or several orbital or transport systems and provide detailed images of the earth surface. Images are created via passive optic and thermal and/or active radar sensors. Aerial pictures are created either photographed or via electronic recording systems and contain images of parts of the earth surface. Aerial images are usually created by air planes, helicopters, drones, or air ships.	Better data availability of high resolution images at lower costs, certain technical infrastructure is needed, images have to be processed/analysed before further usage or linkage with MNO data (e.g. not the images are linked/used but the information that the images provide, e.g. number/location of houses, green areas, …)	Use of appropriate infrastructure, expertise and techniques
f. **Internet** (web pages, social media posts,…)			
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 ?	The size of image data depends on certain aspects, an important factor is the resolution. Usually, image data is comparatively big. Processing requires specific methods.	Huge amount of data	Use of appropriate infrastructure and techniques
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 ?	Image data is no-structured data, depending on the observed area, normally multiple images need to be combined to cover the area of interest, there are established methods for merging individual images; Linkage with other data sources is possible with geospatial information	Non-structured data requires certain data handeling techniques, image analysis can be a quite complex field	Use of appropriate infrastructure and techniques
Relevance to the MNO data uses cases			
Improving population coverage : are these data complementary to MNO data ? Do they allow to improve MNO data population representativeness ? Are there difficulties to due to different irreconcialable concepts (definitions, time periods, …) ? Improve of timeliness of analyses and geographical precision.	Probably not directly, as satellite images identify objects and not people. However, it can be used for plausibility tests (e.g. when estimating resident population with the help of MNO data, satellite/aerial images or geoinformation derived from these images can provide information of the location of resident buildings).		
Improving precision of analyses : do these data allow to provide preciser analyses (ex : socio-economic indicators, indicators of the mobility type, indicators of the commercial activity,…)	Yes, generally in terms of timeliness and as a way of quality assurance. (But not regarding socio-economic indicators, mobility type, …)		
Broadening the scope of issues covered by official statistics : do these data, combined with MNO data, allow to cover new topics (exposure to air pollution, commuting CO2 emissions,…) ?	Yes, multiple (sometimes already tested) approaches especially regarding environmental issues possible, access to green areas	Lack of research / use cases that combine satellite images + MNO data to cover new topics.	Identification of suitable new topics and development of processes and methods
Access			
Who owns the data ? Public administration, one company, several companies ? Rq : different systems, not all the data owned by one company.	The ownership varies. Some imagery is freely accessible (e.g. EU funded missions), some belong to public administration, other are parts of public-private partnership and a significant amount is owned by multiple companies.	The situation in Germany has another layer of complexity as a considerable part of the data that is owned by public administration is owned by the different federal states	
Is it possible to get access ? Does it have to be paid ?	There are muliple ways of accessing the publicly available data, some of the private data can be accessed over master agreements or science funding	Imagery in a resolution and a frequency that is often necessary for statistical purposes is oftentimes owned by private companies and access has to be paid.	
Are there potentially competing uses (operator publishing similar statistics, etc.)?	No		
Are there legal or ethical problems to access the data ?	No		
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 ?	The better the resolution, the bigger the data, resulting in technical obstacles: Higher resolution is often more expensive, leading to financial limitations	High quality images are large and costly	Appropriate technical infrastructure and better access to high resolution images; Or trade-off between cost/quality/infrastructure
Is there a possibility to access the data to study its relevance ?	Yes, there is a significant amountof data publicly available that can be studied.		
Is this data available in all EU countries ?	Yes		
Metadata			
Is the definition of the population known ? If not do you already have a method to address this issue ?	There is no definition of population		
Is the base unit of the dataset known ?	Each image is composed of pixels		
Do the units have an identifier ?	No		
Do you have the necessary variables to reach the relevant granularity level for the statistical unit ?	Depending on the use case (the desired statistical unit) and on the used material, a deep granularity level can be reached		
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 ?	Yes		
Is there auxiliary information to make the data set useful with auxiliary data (NSI or other source) ?	Yes, e.g. additional geoinformation provided by the respective ONA.		
Are there known quality issues with any of the variables	Commonly known problems in image analysis in general (speckles,radiometric effects, cloud coverage etc.)		
Does the data contain sensitive variable ? (Meaning legal or ethical issues related to its use)	A personal reference cannot be ruled out	Specific IT security measures and data protection regulations need to be implemented	

Official statistics quality standards			
Is the scope, detail and cost of this source commensurate with needs ?	If you use the freely accessible pictures, yes.	Trade-off between image quality, costs (financial and time/expertise), etc. needs to be considered/defined for each use case or statistic	
Is it possible to regularly monitor output quality ?	Possible, but requires very specific knowledge	Potentially lack of expertise/ressources	
Confidentiality and data protection : Is it possible to put in place the necessary regulatory, administrative, techincal and organisational measures to protect the security and integrity of statistical data and their transmission ?	Partially	The processing of image data can consume a lot of computing ressources and environments that provide these might not provide the required measues to protect statistical data.	Workflow that processes the data seperately
Impartiality and Objectivity : are information on data sources, methods and procedures publicly available ? Is there a process to inform the NSI in case an error is discovered ?	Yes, a lot of the knowledge regarding remote sensing is based on open source		
Appropriate statistical procedures : are the definitions and concepts used in this source a good approximation of the concepts required for statistical purposes ? Do the data holders collaborate with the NSI in improving data quality (take into account feedback,...)	No	Data is not produced with statistical purposes in mind	
Accuracy and reliability : are source data, integrated data, intermediate results and statistical outputs regularly assessed and validated ?	Yes	Process very technical and might be difficult to assess from the side of the NSIs	
Are sampling errors and non-sampling errors measured and systematically documented ?	Does not apply		
Are revisions regularly analysed in order to improve source data, statistical processes and outputs ?	does not apply		
Timeliness an punctuality : is it possible to release statistics based on this data source on a regular basis ?	yes		
Coherence and Comparability : are statistics based on this source coherent, consistent and comparable over a reasonable period of time ?	As soon as the recording instrument flies and sends data stably		
Is cross-national comparability of statistics based on this source possible ? -> has to be discussed between WP members for comparability	Yes		
Is access to microdata allowed for research purposes ?	Yes		

Sheet 2_ RO_remote_sensing

Questions on the non-MNO source			
Data Type	**Description**	**Challenges**	**Treatments**
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. **Remote sensing, fixed** (smart meters readings, whether station readings, traffic loop signals, lorry tracking signals,…)			
e. **Remote sensing, mobile** (satellite images, drone images, airborne laser scanning, maritime AIS, …)	Satellite remote sensing data is considered by experts in official statistics to be a low-hanging fruit ready to be used in statistical production. Satellite remote sensing data for Earth observation could be defined as measurements of Earth's surface and atmosphere through active or passive sensors mounted on spacecrafts in mostly in low Earth orbit. Satellite remote sensed data comes in different sizes and shapes depending on the satellite mission objective and type of sensor as a continuous stream of data in need of pre-processing before any type of analysis could be carried.	The main challenges with satellite remote sensing data come, on the one hand from the pre-processing step (i.e. a set of corrections and calibrations needed to increase measurements accuracy), and on the other hand, from the size of data (one sensor can produce gigabytes of data each second). Another type of challenge is the internal capacity on the demand side (are NSI's ready to use in statistical production remote sensing data)	These two challenges are mainly solved through data products which hide most of the complexity (analysis ready data or ARD) and cloud computing infrastructure (e.g. Copernicus Data Space Ecosystem or CDSE, Google Earth Engine, Microsoft Planetary Computer. Etc.).
f. **Internet** (web pages, social media posts,…)			
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 ?	The size of the data is directly proportional with the resolution (spatio-temporal and spectral resolution), from several megabytes for low resolution data to exabytes for high or very-high resolution datasets.	The size of the data could present some challenges, such as internal expertise of using cloud computing infrastructure for big data processing.	Although analyses could be carried using on-premises computing infrastructure, it is more efficiently to use on-demand cloud computing infrastructure already available. Training courses are available for using cloud computing infrastructure on Earth observation data.
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 ?	The data comes in different file formats, usually as file archives containing multiple files(data structure files, data and metadata files, quality files, higher level data products already derived etc.). A. Yes, in order to obtain an useful data product it is needed to relate multiple files (eg. an area of interest could be covered by data coming from several data files). B. Yes, the variables and algorithms that enable data linking are known.	The different file formats can present some challenges. Also there are different types of algorithms and techniques used to link data coming from multiple sources (e.g. data fusion between Sentinel and Landsat missions).	Training courses are available for processing remote sensing data from satellite sensors. Moreover, analysis ready data products are available and, in practice, could greatly minimize costs of using remote sensing data in statistical production.
Relevance to the MNO data uses cases			
Improving population coverage : are these data complementary to MNO data ? Do they allow to improve MNO data population representativeness ? Are there difficulties to due to different irreconcialable concepts (definitions, time periods, …) ? Improve of timeliness of analyses and geographical precision.	To our knowledge is not applicable.	Not applicable.	Not applicable.
Improving precision of analyses : do these data allow to provide preciser analyses (ex : socio-economic indicators, indicators of the mobility type, indicators of the commercial activity,…)	Current spatio-temporal resolution of freely available remote sensing data from satellite sensors is to low in order to measure with great accuracy changes at at very high spatial resolution.	Not applicable.	Not applicable.
Broadening the scope of issues covered by official statistics : do these data, combined with MNO data, allow to cover new topics (exposure to air pollution, commuting CO2 emissions,…) ?	Satellite remote sensing data are complementary to MNO data and, theoretically, could be used to produce new statistical products which require auxilary geographical spatio-temporal information with respect to land change, land use and air quality statistics.		Solving spatio-temporal coverage requirements could be tackled by using in-situ informations (e.g. such as air pollution monitoring stations or data collected trough remote sensing by UAV) in order to provide continuous time series of observations.
Access			
Who owns the data ? Public administration, one company, several companies ? Rq : different systems, not all the data owned by one company.	With respect to ownership, remote sensing data comes in two flavors: state owned data and freely accessible, state owned and classified/restricted access data and privately owned data. Freely available data is provided through different platforms (data and information access services – DIAS) mainly by state-owned agencies (e.g. European Commission & ESA, NASA, USGS, Agenzia Spaziale Italiana). Commercial data is provided by private entities and currently is not targeted as feasible to be used in official statistics.	Access to state-owned and freely available data presents no challenges.	Not applicable to state owned and freely available data.
Is it possible to get access ? Does it have to be paid ?	In case of state-owned and freely available data this is not applicable.	Not applicable.	Not applicable.
Are there potentially competing uses (operator publishing similar statistics, etc.)?	Combining population statistics (including statistics derived from mobile phone network data) and remote sensing data is a field of active academic research, but to our current knowledge no institution or private entity provides statistics recurrent and timely statistics based on fusion between MNO and remote sensing data. Event if we run on the assumption that some statistics are produced by combining these two types of datasets, we can safely consider that the cost of matching the quality criteria used in official statistics are too restrictive or impose a set of constraints which makes it prohibitive (in terms of scope and trusted statistics) to compete with products delivered by official statistics.	Not applicable.	Not applicable.
Are there legal or ethical problems to access the data ?	Not applicable.	Not applicable.	Not applicable.
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 ?	For freely available, state-owned data, there are no limitations or legal restrictions.	Not applicable.	Not applicable.
Is there a possibility to access the data to study its relevance ?	Yes.	Not applicable.	Not applicable.
Is this data available in all EU countries ?	Yes.	Not applicable.	Not applicable.
Metadata			
Is the definition of the population known ? If not do you already have a method to address this issue ?	Yes. Satellite remote sensing data is highly standardized.	Not applicable.	Not applicable
Is the base unit of the dataset known ?	Yes.	Not applicable.	Not applicable.
Do the units have an identifier ?	Yes.	Not applicable.	Not applicable.
Do you have the necessary variables to reach the relevant granularity level for the statistical unit ?	Yes.	Not applicable.	Not applicable.
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 ?	Yes.	Not applicable.	Not applicable.
Is there auxiliary information to make the data set useful with auxiliary data (NSI or other source) ?	Yes, mainly from land use and cover registers and in-situ measurements.	Not applicable.	Not applicable.
Are there known quality issues with any of the variables ?	Mainly there are two types of issues which can arise: sensors dependent (calibration/correction) and sensor independent issues (e.g. cloud coverage). The first category is usually handled by the data producer. The second one is independent of the data producers and is handled by the users.	First category of issues is out of the scope and means of the users. The second category could limit the spatio-temporal coverage.	Auxiliary data and algorithms/methods can be used to improve spatio-temporal coverage due to issues from the second category.
Does the data contain sensitive variable ? (Meaning legal or ethical issues related to its use)	Not applicable for freely available data.	Not applicable.	Not applicable.

Official statistics quality standards			
Is the scope, detail and cost of this source commensurate with needs ?	There is a growing need, also from the part of Eurostat/NSI's, to use remote sensing data with other types of dataset in order to enhance or produce new official statistical products. Given that remote sensing data is freely available and entry-cost are at a historically minimum (analysis ready data, relative low cost of access to cloud computing infrastructure) official statistics could greatly benefit from using remote sensing data.	The main challenge is to develop NSI's internal capacities to use remote sensing data with MNO data, given that the cost of accessing and using MNO data is currently unknown.	This is beyond the scope of the current project and is to be tackled at the European Statistical System level.
Is it possible to regularly monitor output quality ?	Yes. Data producers provide also quality metrics.	Not applicable.	Not applicable.
Confidentiality and data protection : Is it possible to put in place the necessary regulatory, administrative, techincal and organisational measures to protect the security and integrity of statistical data and their transmission ?	Not applicable.	Not applicable.	Not applicable.
Impartiality and Objectivity : are information on data sources, methods and procedures publicly available ? Is there a process to inform the NSI in case an error is discovered ?	Yes.	Not applicable.	Not applicable.
Appropriate statistical procedures : are the definitions and concepts used in this source a good approximation of the concepts required for statistical purposes ? Do the data holders collaborate with the NSI in improving data quality (take into account feedback,...)	Yes, the definitions and concepts fulfill requirements for statistical production in most of the potential use cases.	Further enhance compatibility between satellite remote sensing data and statistical data needs of NSI's.	To our knowledge, currently Eurostat partners with ESA to enhance compatibility between remote sensing data provided through Copernicus Programme and statistical needs of the European Commission and Member States.
Accuracy and reliability : are source data, integrated data, intermediate results and statistical outputs regularly assessed and validated ?	Yes.	Not applicable. Handle by the data producers.	Not applicable.
Are sampling errors and non-sampling errors measured and systematically documented ?	Yes.	Not applicable. Handle by the data producers.	Not applicable.
Are revisions regularly analysed in order to improve source data, statistical processes and outputs ?	Yes.	Not applicable. Handle by the data producers.	Not applicable.
Timeliness an punctuality : is it possible to release statistics based on this data source on a regular basis ?	Yes.	Not applicable. Handle by the data producers.	Not applicable.
Coherence and Comparability : are statistics based on this source coherent, consistent and comparable over a reasonable period of time ?	Yes.	Not applicable. Handle by the data producers.	Not applicable.
Is cross-national comparability of statistics based on this source possible ? -> has to be discussed between WP members for comparability	Yes.	Not applicable. Handle by the data producers.	Not applicable.
Is access to microdata allowed for research purposes ?	No. Raw sensor data is not available and is not very useful for statistical purposes.	Not applicable. Handle by the data producers.	Not applicable.

Sheet 3_ RO_land use registers

Questions on the non-MNO source			
Data Type	**Description**	**Challenges**	**Treatments**
a.1 **Survey** (census, SILC, employment,…)			
a.2 **Combination of survey and register**			
b. **Register** (vital events, diagnoses, wage, income tax, welfare payments, …)	Land use and land cover registers hold systematic information regarding Earth surface coverage and purpose (e.g. CORINE LULC)		
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. **Remote sensing, fixed** (smart meters readings, whether station readings, traffic loop signals, lorry tracking signals,…)			
e. **Remote sensing, mobile** (satellite images, drone images, airborne laser scanning, maritime AIS, …)			
f. **Internet** (web pages, social media posts,…)			
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 ?	Yes. Data sets are proportional with spatio-temporal coverage.	Not applicable.	Not applicable.
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 ?	Data is stored, usually, in a GIS database. Database architecture could differ from country to country. A. There are several tables holding information. B. Yes linking is inherently possible, through the design of the database.	No challenges	Not applicable.
Relevance to the MNO data uses cases			
Improving population coverage : are these data complementary to MNO data ? Do they allow to improve MNO data population representativeness ? Are there difficulties to due to different irreconcialable concepts (definitions, time periods, …) ? Improve of timeliness of analyses and geographical precision.	Not applicable	Not applicable.	Not applicable.
Improving precision of analyses : do these data allow to provide preciser analyses (ex : socio-economic indicators, indicators of the mobility type, indicators of the commercial activity,…)	Land use and land cover data can be used to enhance geographical precision of MNO data (e.g. accurately pinpoint home location in order to produce mobility types indicators).	Out of date information.	Regular updates of the registers are required.
Broadening the scope of issues covered by official statistics : do these data, combined with MNO data, allow to cover new topics (exposure to air pollution, commuting CO2 emissions,…) ?	Land use and land cover registers can be used to produce point-of-interest statistics (e.g. coverage and usage of urban green spaces)	Out of date information.	Regular updates of the registers are required.
Access			
Who owns the data ? Public administration, one company, several companies ? Rq : different systems, not all the data owned by one company.	Data is owned by several national institutions, including NSI's.	In order to access all information regarding land use and land cover for a country partnerships and agreements are needed between all data holders. Different database systems and database designs can present data access challenges.	Agreements at national level are needed in order to insure database access and interoperability.
Is it possible to get access ? Does it have to be paid ?	State owned data and freely (no financial cost) available, but not publicly available.	In order to access all information regarding land use and land cover for a country partnerships and agreements are needed between all data holders.	Agreements at national level are needed in order to insure database access and interoperability.
Are there potentially competing uses (operator publishing similar statistics, etc.)?	Similar case with remote sensing data, combining population statistics (including statistics derived from mobile phone network data) and land use and land cover register data is a field of active academic research, but to our current knowledge no institution or private entity provides statistics recurrent and timely statistics based on fusion between MNO and land use and land cover data. Event if we run on the assumption that some statistics are produced by combining these two types of datasets, we can safely consider that the cost of matching the quality criteria used in official statistics are too restrictive or impose a set of constraints which makes it prohibitive (in terms of scope and trusted statistics) to compete with products delivered by official statistics.	Not applicable.	Not applicable.

Are there legal or ethical problems to access the data ?	Not applicable.	Not applicable.	Not applicable.
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 ?	For freely available, state-owned data, there are no limitations or legal restrictions.	Not applicable.	Not applicable.
Is there a possibility to access the data to study its relevance ?	Yes.	Not applicable.	Not applicable.
Is this data available in all EU countries ?	Yes. It is mandatory by national laws, although data holders could differ.	Not applicable.	Not applicable.
Metadata			
Is the definition of the population known ? If not do you already have a method to address this issue ?	Yes. Land cover and land use data is highly standardized.	Not applicable.	Not applicable
Is the base unit of the dataset known ?	Yes.	Not applicable.	Not applicable.
Do the units have an identifier ?	Yes.	Not applicable.	Not applicable.
Do you have the necessary variables to reach the relevant granularity level for the statistical unit ?	Yes.	Not applicable.	Not applicable.
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 ?	Yes.	Not applicable.	Not applicable.
Is there auxiliary information to make the data set useful with auxiliary data (NSI or other source) ?	Yes.	Not applicable.	Not applicable.
Are there known quality issues with any of the variables ?	The main issues with registers can arise from out of date information.	Not applicable.	Update the registers. Depends on the type of register and data producer/holder.
Does the data contain sensitive variable ? (Meaning legal or ethical issues related to its use)	Not applicable to all types of land use and land cover registers. Some registers can contain data about locations of household usual residence.	Disclosure risk.	Statistical disclosure control can be applied to ensure statistical confidentiality.
Official statistics quality standards			
Is the scope, detail and cost of this source commensurate with needs ?	Land use and land cover registers can be used to produce low cost and relevant statistics to a broad category of statistical products users taking into account the need to produce and disseminate geo-grid statistics.	Not applicable.	Not applicable.
Is it possible to regularly monitor output quality ?	Yes.	Not applicable.	Not applicable.
Confidentiality and data protection : Is it possible to put in place the necessary regulatory, administrative, techincal and organisational measures to protect the security and integrity of statistical data and their transmission ?	Yes. Statistical disclosure control can be applied.	Not applicable.	Not applicable.
Impartiality and Objectivity : are information on data sources, methods and procedures publicly available ? Is there a process to inform the NSI in case an error is discovered ?	Yes.	Not applicable.	Not applicable.
Appropriate statistical procedures : are the definitions and concepts used in this source a good approximation of the concepts required for statistical purposes ? Do the data holders collaborate with the NSI in improving data quality (take into account feedback,...)	Yes, the definitions and concepts fulfill requirements for statistical production in most of the potential use cases. Most of the national land use and land cover registers are produced using EU INSPIRE standard	Not applicable.	Not applicable.
Accuracy and reliability : are source data, integrated data, intermediate results and statistical outputs regularly assessed and validated ?	Yes.	Not applicable. Handle by the data producers.	Not applicable.
Are sampling errors and non-sampling errors measured and systematically documented ?	Yes.	Not applicable. Handle by the data producers.	Not applicable.
Are revisions regularly analysed in order to improve source data, statistical processes and outputs ?	Yes.	Not applicable. Handle by the data producers.	Not applicable.
Timeliness an punctuality : is it possible to release statistics based on this data source on a regular basis ?	Yes.	Not applicable. Handle by the data producers.	Not applicable.
Coherence and Comparability : are statistics based on this source coherent, consistent and comparable over a reasonable period of time ?	Yes.	Not applicable. Handle by the data producers.	Not applicable.
Is cross-national comparability of statistics based on this source possible ? -> has to be discussed between WP members for comparability	Yes.	Not applicable. Handle by the data producers.	Not applicable.
Is access to microdata allowed for research purposes ?	Yes.	Not applicable. Handle by the data producers.	Not applicable.

Sheet 4_ SW_traffic signals

Traffic signals Statistical agency: The Swedish Transport Agency Type: Sensor data Source: https://www.transportstyrelsen.se/en/road/road-tolls/congestion-taxes-in-stockholm-and-goteborg/	
Description	In Sweden there is a system of congestion taxes in Stockholm and Gothenburg. The tax applies to vehicles registered in and outside of Sweden. See Figure in the link below for Stockholm - city showing vehicle sensor locations for congestion tax https://www.transportstyrelsen.se/globalassets/global/bilder/vag/trangselskatt/karta_trangselskatt_sthlm.pdf
Challenges	* Limited to Stockholm and Gothenburg central parts of the cities. * Can be difficult to map with SCBs DeSo levels. * Data structure and formats need to be investigated further. * Covarage must be investigated.
Treatments	Combing relevant non-MNO such as traffic sensors and MNO sources can generate more useful data for public use as well as for calibration modelling.
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 ?	Needs to be investigated.
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 ?	Needs to be investigated.
Relevance to the MNO data uses cases	
Improving population coverage : a. are these data complementary to MNO data ? b. Do they allow to improve MNO data population representativeness? c. Are there difficulties to due to different irreconcialable concepts (definitions, time periods, …)? d. Improve of timeliness of analyses and geographical precision.	a. Yes b. Potentially c. Potentially d. Timeliness is unsure. But geographical precision potentially yes.
Improving precision of analyses : do these data allow to provide preciser analyses (ex : socio-economic indicators, indicators of the mobility type, indicators of the commercial activity,…)	Potentially yes.
Broadening the scope of issues covered by official statistics : do these data, combined with MNO data, allow to cover new topics (exposure to air pollution, commuting CO2 emissions,…) ?	
Access	
Who owns the data ? Public administration, one company, several companies ? Rq : different systems, not all the data owned by one company.	Public administration data.
Is it possible to get access ? Does it have to be paid ?	Access - Yes. Payment, No.
Are there potentially competing uses (operator publishing similar statistics, etc.)?	Potentially yes, no currently offficial statistics.
Are there legal or ethical problems to access the data ?	Potentially yes.
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 ?	To be investigated.
Is there a possibility to access the data to study its relevance ?	Possibly yes.
Is this data available in all EU countries ?	To be investigated.

Metadata	To be investigated.
Is the definition of the population known ? If not do you already have a method to address this issue ?	To be further investigated
Is the base unit of the dataset known ?	To be further investigated
Do the units have an identifier ?	To be further investigated
Do you have the necessary variables to reach the relevant granularity level for the statistical unit ?	To be further investigated
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 ?	To be further investigated
Is there auxiliary information to make the data set useful with auxiliary data (NSI or other source) ?	To be further investigated
Are there known quality issues with any of the variables ?	To be further investigated
Does the data contain sensitive variable ? (Meaning legal or ethical issues related to its use)	To be further investigated
Official statistics quality standards	No official statistics today. Following questions must be investigated.
Is the scope, detail and cost of this source commensurate with needs ?	Not known
Is it possible to regularly monitor output quality ?	Not known
Confidentiality and data protection : Is it possible to put in place the necessary regulatory, administrative, techincal and organisational measures to protect the security and integrity of statistical data and their transmission ?	Not known
Impartiality and Objectivity : are information on data sources, methods and procedures publicly available ? Is there a process to inform the NSI in case an error is discovered ?	Not known
Appropriate statistical procedures : are the definitions and concepts used in this source a good approximation of the concepts required for statistical purposes ? Do the data holders collaborate with the NSI in improving data quality (take into account feedback,...)	Not known
Accuracy and reliability : are source data, integrated data, intermediate results and statistical outputs regularly assessed and validated ?	Not known
Are sampling errors and non-sampling errors measured and systematically documented ?	Not known
Are revisions regularly analysed in order to improve source data, statistical processes and outputs ?	Not known
Timeliness an punctuality : is it possible to release statistics based on this data source on a regular basis ?	Not known
Coherence and Comparability : are statistics based on this source coherent, consistent and comparable over a reasonable period of time ?	Not known
Is cross-national comparability of statistics based on this source possible ? -> has to be discussed between WP members for comparability	Not known
Is access to microdata allowed for research purposes ?	Not known

Sheet 5_ DE_social media

Questions on the non-MNO source				
Data Type	**Description**	**Challenges**	**Treatments**	**Sources/Useful links:**
a.1 **Survey** (census, SILC, employment,…)				
f. **Internet** (web pages, social media posts,…)	Data from social media usually contain information that platform users have created themselves. Some of the social network platforms offer access to parts of the data via APIs (Application Programming Interface). To get data access, there is usually a registration and indication of research purpose/use of data required. Access has to be granted by the platform then.	Not all social media platforms offer access, different pricing models (basic-pro/enterprise), probably limited information on data selection, social media content does not represent the whole population (only useful for certain research questions, e.g. trending topics in digital 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/
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 ?	depends on the (purchased) access level, for Twitter/X: 1,500 Tweets per month are available for free and 1,000,000 Tweets per month available with the Pro level (>1 Mio Tweets/month available with the enterprise level). Free data contain a random sample of 1% of Tweets from the last 7-10 days. Purchased datasets contain 500 Tweets per Request and have geoinformation available. Only the most costly level offers filtering and real-time stream.	The challenge might not be handling a too large dataset but quite the opposite of potentially not having sufficient amount of data. But again, this depends on the access level and consequently on price/budget.		
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 ?	Twitter/X data contain: date, text, user name, location (if geo tagging available), further objects are possible, e.g. used hashtags, tagged other users, URLs, media objects (videos, GIFs, …)			

Relevance to the MNO data uses cases		
Improving population coverage : are these data complementary to MNO data ? Do they allow to improve MNO data population representativeness ? Are there difficulties to due to different irreconcialable concepts (definitions, time periods, …) ? Improve of timeliness of analyses and geographical precision.	social media data can somewhat be complementary to improving population coverage for certain events or to complement MNO data with socio-demographic information or to relate topics/events with a population number and location for cases with a very specific research question. However, it probably cannot contribute to improve total population coverage or geographical precision on a large scale because only a relatively small share of the total population actively and regularly posts on social media (with geo tagging).	Twitter/X (and further providers) not useful for broad analysis due to lack of user participation (to represent or draw conclusions to total population), Facebook/Meta might be more suitable (e.g. profile info, not posts). However, there are multiple challenges / quality issues. Usually, users are not verified and there are plently of accounts which do not represent a real person (fake-accounts, bots, multiple accounts, small business accounts not marked accordingly, ...)
Improving precision of analyses : do these data allow to provide preciser analyses (ex : socio-economic indicators, indicators of the mobility type, indicators of the commercial activity,...)		
Broadening the scope of issues covered by official statistics : do these data, combined with MNO data, allow to cover new topics (exposure to air pollution, commuting CO2 emissions,...) ?	potentially wider scope, especially for event or topic based analysis/statistics, e.g. are certain topics discussed more in certain geographical areas? Can there be short-term event based population figures estimated?	
Access		
Who owns the data ? Public administration, one company, several companies ? Rq : different systems, not all the data owned by one company.	Social media platforms	
Is it possible to get access ? Does it have to be paid ?	Free access available for very small random samples of data. Paid access (usually different access levels basic-pro/enterprise) available.	
Are there potentially competing uses (operator publishing similar statistics, etc.)?	Not to our current best knowledge	
Are there legal or ethical problems to access the data ?	common data privacy restrictions for private data	
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 ?	Amount is mainly restricted by price, aggregation less of a problem but more the general availability / lack of user activity	
Is there a possibility to access the data to study its relevance ?	Yes, small random samples are available after registration / application for research purposes from Twitter and Facebook/Instagram. Further social media providers to be investigated	
Is this data available in all EU countries ?	Yes (to our current best knowledge)	
Metadata		
Is the definition of the population known ? If not do you already have a method to address this issue ?		
Is the base unit of the dataset known ?		
Do the units have an identifier ?	User handle can serve as a unique identifier (in case of Twitter and Instagram)	
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 other source) ?		
Are there known quality issues with any of the variables ?		
Does the data contain sensitive variable ? (Meaning legal or ethical issues related to its use)		
Official statistics quality standards		
Is the scope, detail and cost of this source commensurate with needs ?		
Is it possible to regularly monitor output quality ?		
Confidentiality and data protection : Is it possible to put in place the necessary regulatory, administrative, techincal and organisational measures to protect the security and integrity of statistical data and their transmission ?		
Impartiality and Objectivity : are information on data sources, methods and procedures publicly available ? Is there a process to inform the NSI in case an error is discovered ?		
Appropriate statistical procedures : are the definitions and concepts used in this source a good approximation of the concepts required for statistical purposes ? Do the data holders collaborate with the NSI in improving data quality (take into account feedback,...)		
Accuracy and reliability : are source data, integrated data, intermediate results and statistical outputs regularly assessed and validated ?		
Are sampling errors and non-sampling errors measured and systematically documented ?		
Are revisions regularly analysed in order to improve source data, statistical processes and outputs ?		
Timeliness an punctuality : is it possible to release statistics based on this data source on a regular basis ?		
Coherence and Comparability : are statistics based on this source coherent, consistent and comparable over a reasonable period of time ?		
Is cross-national comparability of statistics based on this source possible ? -> has to be discussed between WP members for comparability		
Is access to microdata allowed for research purposes ?		

Sheet 6_ DE_smart_meters

Questions on the non-MNO source			
Data Type	**Description**	**Challenges**	**Treatments**
d. **Remote sensing, fixed** (smart meters readings, whether station readings, traffic loop signals, lorry tracking signals,…)	Smart Meters usually comprise Smart Electricity Meter and Gas/Energy Smart Meter. In Germany, legislation mandates the installation of Electricity Meter until 2032 for all users above a certain consumption level. In the past, analysis or usage of smart meter readings did not make much sense for NSIs due to lack of data. This might/will change in the next years. According to the European Commission, 13 EU-Countries have a relatively high coverage of Smart Meters.	Lack of data in certain EU-countries due to lack of installed smart meters. Ownership and accessability of data lies outside the NSIs (consumers own their data and electricity companies hold the data).	Reliable access to data through legislation or alternative mechanisms. Waiting for rollout of smart meters (at least in the DE case for which a higher coverage is expected in the upcoming years)
e. **Remote sensing, mobile** (satellite images, drone images, airborne laser scanning, maritime AIS, …)			
f. **Internet** (web pages, social media posts,…)			
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 ?	Data is mainly stored in large data hubs and can usually accessed via FTP-Server or API	Data can be large depending on the frequency and detail of the data sets (in case of a large household coverage with smart meters)	
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 ?	Data comprises energy usage and energy input of residential and commercial buildings, unit, location as well as aggregation levels by time and space. Data can be in different quality levels with respect to time and space. Good quality means that it can be associated with an exact address and poor quality means that data can be associated to an address of multiple buildings of different usage.		
Relevance to the MNO data uses cases			
Improving population coverage : are these data complementary to MNO data ? Do they allow to improve MNO data population representativeness ? Are there difficulties to due to different irreconcialable concepts (definitions, time periods, …) ? Improve of timeliness of analyses and geographical precision.	If coverage is high enough, it can be complementary to MNO data to represent present population because it can assign a home location to a household and the consumption provides assumptions on the household size. Electricity/Energy consumption can indicate household size.	No experience / use cases so far	
Improving precision of analyses : do these data allow to provide preciser analyses (ex : socio-economic indicators, indicators of the mobility type, indicators of the commercial activity,…)	potentially user ambiguity (can give a better idea about household size)	No results / use cases available so far	
Broadening the scope of issues covered by official statistics : do these data, combined with MNO data, allow to cover new topics (exposure to air pollution, commuting CO2 emissions,…) ?	Possibly yes, however no concrete use cases so far	No experience or identification of potential new topics so far	Identification, development and testing of potential new topics
Access			
Who owns the data ? Public administration, one company, several companies ? Rq : different systems, not all the data owned by one company.	Consumers and Electricity/Energy Companies		
Is it possible to get access ? Does it have to be paid ?	Possibly via collaboration/paid access with Energy Providers -> to be investigated		
Are there potentially competing uses (operator publishing similar statistics, etc.)?	No information / experience		
Are there legal or ethical problems to access the data ?	Data can probably be used if consumer agrees to it and/or if sufficient anonymization is ensured		
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 ?	No information / experience, probably legal and financial?		
Is there a possibility to access the data to study its relevance ?	No information / experience		
Is this data available in all EU countries ?	Yes, in most countries, but on a different scale. Some have a relatively high coverage of Smart Meters (80% of households or more) and others do not have a sufficiently high coverage (yet)		

Metadata			
Is the definition of the population known ? If not do you already have a method to address this issue ?			
Is the base unit of the dataset known ?			
Do the units have an identifier ?			
Do you have the necessary variables to reach the relevant granularity level for the statistical unit ?	To be investigated.		
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 other source) ?			
Are there known quality issues with any of the variables	To be investigated. Intuitively, there should not be large measurement errors nor false household information, etc. expected.		
Does the data contain sensitive variable ? (Meaning legal or ethical issues related to its use)	Yes, consumers probably would need to agree that their data can be used and/or suitable anonymization mechanisms need to be in place.		
Official statistics quality standards			
Is the scope, detail and cost of this source commensurate with needs ?	To be investigated. Intuitively, yes if it accessible and adds sufficient value to statistical production.	Only makes sense with sufficient amount of data available, data access and quality needs to be ensured, sufficient added value might not be there	Identification of potentials for statistical production, large rollout of smart/intelligent meters needed
Is it possible to regularly monitor output quality ?	To be investigated.		
Confidentiality and data protection : Is it possible to put in place the necessary regulatory, administrative, techincal and organisational measures to protect the security and integrity of statistical data and their transmission ?			
Impartiality and Objectivity : are information on data sources, methods and procedures publicly available ? Is there a process to inform the NSI in case an error is discovered ?			
Appropriate statistical procedures : are the definitions and concepts used in this source a good approximation of the concepts required for statistical purposes ? Do the data holders collaborate with the NSI in improving data quality (take into account feedback,...)			
Accuracy and reliability : are source data, integrated data, intermediate results and statistical outputs regularly assessed and validated ?			
Are sampling errors and non-sampling errors measured and systematically documented ?	Does not apply.		
Are revisions regularly analysed in order to improve source data, statistical processes and outputs ?			
Timeliness an punctuality : is it possible to release statistics based on this data source on a regular basis ?	Yes, if data access is ensured and if there is sufficient amount of data (houshold coverage).		
Coherence and Comparability : are statistics based on this source coherent, consistent and comparable over a reasonable period of time ?	Highly likely, as energy consumption is measured in a physical constant		
Is cross-national comparability of statistics based on this source possible ? -> has to be discussed between WP members for comparability	Probably yes, see above		
Is access to microdata allowed for research purposes ?			

Sheet 7_ RO_google popular time

Questions on the non-MNO source			
Data Type	**Description**	**Challenges**	**Treatments**
a.1 **Survey** (census, SILC, employment,...)			
a.2 **Combination of survey and register**			
f. **Internet** (web pages, social media posts,...)	Google Popular Times shows how busy a location typically is during different times of the day. Popular times are based on average popularity over the last few months. Popularity for any given hour is shown relative to the typical peak popularity for the business for the week. Is computed based on GPS location from Android-based smartphones.	Google does not provide an API to access data directly, Data is obtained through web scraping which presents legal challenges. Google's terms and conditions of service forbid web scraping.	An agreement between Eurostat and Google can be put in place in order to facilitate data access.
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 ?	It is proportional with spatial coverage.	Not applicable.	Not applicable.
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 ?	Date is obtained through web scraping. A. Depends on the design of the web scrapping procedure. B. Geographic location is available which can facilitate linking with this data.	Google does not provide an API to access data directly, Data is obtained through web scraping which presents legal challenges. Google's terms and conditions of service forbid web scraping.	An agreement between Eurostat and Google can be put in place in order to facilitate data access.
Relevance to the MNO data uses cases			
Improving population coverage : are these data complementary to MNO data ? Do they allow to improve MNO data population representativeness ? Are there difficulties to due to different irreconcialable concepts (definitions, time periods, ...) ? Improve of timeliness of analyses and geographical precision.	In theory, can be used to assess population density between certain time intervals in certain locations, which in turn can be used as a proxy variable for assessing mobile phone network data coverage. Also statistics regarding certain point-of-interest can be derived using a combination of MNO data and Google Popular Times data.	The data is a delivered as product of a black box. No control on measuring uncertainty.	An agreement between Eurostat and Google can be put in place in order to facilitate data access.
Improving precision of analyses : do these data allow to provide preciser analyses (ex : socio-economic indicators, indicators of the mobility type, indicators of the commercial activity,...)			
Broadening the scope of issues covered by official statistics : do these data, combined with MNO data, allow to cover new topics (exposure to air pollution, commuting CO2 emissions,...) ?			
Access			
Who owns the data ? Public administration, one company, several companies ? Rq : different systems, not all the data owned by one company.	Data is privately owned by Google (Alphabet Corporation).	Google does not provide an API to access data directly, Data is obtained through web scraping which presents legal challenges. Google's terms and conditions of service forbid web scraping.	An agreement between Eurostat and Google can be put in place in order to facilitate data access.
Is it possible to get access ? Does it have to be paid ?	Data is obtained through web scraping which presents legal challenges. Google's terms and conditions of service forbid web scraping.	Google does not provide an API to access data directly.	
Are there potentially competing uses (operator publishing similar statistics, etc.)?	Google produces statistics based on Popular times.	Google may view itself in direct competition with NSI's for producing relevant, timely and reliable statistics.	
Are there legal or ethical problems to access the data ?	Data is obtained through web scraping which raises legal challenges. Google's terms and conditions of service forbid web scraping.	Google terms and conditions of service.	
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 ?	Data is delivered a product of a black box, but can be aggregate or disaggregated at different spatio-temporal resolutions.	Data is deliverd a product of a black box	
Is there a possibility to access the data to study its relevance ?	Yes.		
Is this data available in all EU countries ?	Yes.		
Metadata			
Is the definition of the population known ? If not do you already have a method to address this issue ?	No and we don't have a method to address this issue.	Data is delivered a product of a black box.	
Is the base unit of the dataset known ?	No.		
Do the units have an identifier ?	No.		
Do you have the necessary variables to reach the relevant granularity level for the statistical unit ?	No.		
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 ?	No.		
Is there auxiliary information to make the data set useful with auxiliary data (NSI or other source) ?	Land use and land cover registers, along with open source spatial data (Google Maps, Open Street Maps) can be used to provide enhanced datasets with different statistical purposes.		
Are there known quality issues with any of the variables	Data is delivered as a product of a black box. No sufficient information to know in advance data quality issues (e.g. representativity)		
Does the data contain sensitive variable ? (Meaning legal or ethical issues related to its use)	Yes. Contains information regarding locations, location with private ownership.	Disclosure risk.	

Official statistics quality standards			
Is the scope, detail and cost of this source commensurate with needs ?	In certain use-cases Google Popular Times can be used to provide timely statistics with high spatio-temporal resolution (such as during the COVID 19 pandemics).	Data is delivered a product of a black box	
Is it possible to regularly monitor output quality ?	No.		
Confidentiality and data protection : Is it possible to put in place the necessary regulatory, administrative, techincal and organisational measures to protect the security and integrity of statistical data and their transmission ?	Yes. Statistical disclosure control can be applied.		
Impartiality and Objectivity : are information on data sources, methods and procedures publicly available ? Is there a process to inform the NSI in case an error is discovered ?	No.		
Appropriate statistical procedures : are the definitions and concepts used in this source a good approximation of the concepts required for statistical purposes ? Do the data holders collaborate with the NSI in improving data quality (take into account feedback,...)	No.		
Accuracy and reliability : are source data, integrated data, intermediate results and statistical outputs regularly assessed and validated ?	No.		
Are sampling errors and non-sampling errors measured and systematically documented ?	No.		
Are revisions regularly analysed in order to improve source data, statistical processes and outputs ?	No.		
Timeliness an punctuality : is it possible to release statistics based on this data source on a regular basis ?	No.		
Coherence and Comparability : are statistics based on this source coherent, consistent and comparable over a reasonable period of time ?	No.		
Is cross-national comparability of statistics based on this source possible ? -> has to be discussed between WP members for comparability	No.		
Is access to microdata allowed for research purposes ?	No.		

Sheet 8_ FR_creditcard_transaction

Questions on the non-MNO source			
Data Type	**Description**	**Challenges**	**Treatments**
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,…)	Bankcard payment data at the transaction level	The data's volume and intricacy weren't crafted for socio-economic analyses; instead, they were tailored for fraud detection. The bases were not documented enough for someone to use it without interacting with the data scientist team who work at the data provider's office. The table have sometimes more than 200 columns whithout a data dictionnary. Also to understand the collection of data one needs to be familiar with the mechanisms of the elctronic payment system	To effectively utilize the transaction-level data and work with it, there is a cost of entry in terms of time required to gain knowledge of the database. Learning by doing and creating a data dictionnary as we go with the help of the team on the floor
d. Remote sensing, fixed (smart meters readings, whether station readings, traffic loop signals, lorry tracking signals,…)			
e. Remote sensing, mobile (satellite images, drone images, airborne laser scanning, maritime AIS, …)			
f. Internet (web pages, social media posts,…)			
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 ?	The size of the data set is known	It is impossible to treat all of the data set at once. It has to be splitted monthly or at least yearly due to the excessive computational resources required	The dataset is splitted or aggregated according to the object of the study.
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 structure of the data set is known and is accessible through a Big Data architecture. (A) Depending on the specific need or study, different databases may be accurate. It might	The size and format of the data is a challenge as well as the infrastructure which is not always the most efficient for implementing new ways of studying the data. The infrastructure is closed and does	Creation of a data dictionnary. Data cleaning to be sure to identify the addresss of the companies. Try to infer a location to the cards through several hypothesis to treat mobility question
Relevance to the MNO data uses cases			
Improving population coverage : are these data complementary to MNO data ? Do they allow to improve MNO data population representativeness ? Are there difficulties to due to different irreconcialable concepts (definitions, time periods, …) ? Improve of timeliness of analyses and geographical precision.	The unit of observation are different from MNO. Regarding the timelines, the transactions are dated by the second	The MNO population are detected by IRIS (French statistical subdivision smaller than the municipality) whereas the transactions are located in a shop. We can then place a shop in an IRIS.	
Improving precision of analyses : do these data allow to provide more precise analyses (ex : socio-economic indicators, indicators of the mobility type, indicators of	The data should allow for more precise analysis of the mobility of the population, the commercial activity but it is more limited when it comes to improving socio-economic	The information about the individual owning the credit card are not known. We can rely on hypothesis elaborated with the help of the consortium's teams who know the data.	Mobility: By tracking the stores visited, it's possible to establish an individual's itinerary. Consumption : The spending and evolution of spending are a good proxi for the households consumptions
Broadening the scope of issues covered by official statistics : do these data, combined with MNO data, allow to cover new topics (exposure to air pollution, commuting CO2 emissions,…) ?	For the moment the combination of data is not used for studies about air quality or subject directly related to environmental studies		

Access			
Who owns the data ? Public administration, one company, several companies ? Rq : different systems, not all the data owned by one company.	A single company holds ownership of the data. The credit card consortium, possibly unique to France, was established by all French banks upon the introduction of credit cards to the market. Its purpose was to mitigate interbank fees. Operating its own transaction network, it competes domestically with Visa and Mastercard. The consortium oversees all transactions conducted through the CB network, comprising 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 conducted from a card to a payment terminal owned by the same bank do not appear on the database.	
Is it possible to get access ? Does it have to be paid ?	The access to aggregated data has to be paid	The credit card company has partnership with research	
Are there potentially competing uses (operator publishing similar statistics, etc.)?	Mastercard and Visa could also contribute, but their presence and coverage within the country are significantly smaller compared to CB's. As regards CB, their socio-economic analyses are all caried by research partnerships. The scientific program is elaborated in a way which allows diversified and non-competitive uses of data.		The existence of this unique data set of the national territory makes it hard for other operator to publish similar statistics
Are there legal or ethical problems to access the data ?	Accessing the data poses 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 indiviual'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)	
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 ?	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 can not identify their sales		
Is there a possibility to access the data to study its relevance ?	yes through the research chair		
Is this data available in all EU countries ?	This data may be accessible differently through different EU countries due to the uniqueness of the consortium		
Metadata			
Is the definition of the population known ? If not do you already have a method to address this issue ?	The population consists of credit card holders; however, access to specific data about individual cardholders (e.g., name, age, gender) is unavailable.	Some banks in France now issue cards that are only mastercards or visa when they used to be both CB and another transactional network for international purchases. The loss of the population is hard to estimate. Another point of attention to have is that most of the banks declare the information about the terminal of payment which is giving us the information. The accuracy of those declarations, the harmony and the precision can vary from one bank to another and can make it harder to use that data as a whole. Comprehensive knowledge can be needed	The best way to know more about the database is to spend time among the team of CB
Is the base unit of the dataset known ?	yes		
Do the units have an identifier ?	the credit cards are identified through an anonymised credit card number and the shops through their national id		
Do you have the necessary variables to reach the relevant granularity level for the statistical unit ?	The transaction-by-transaction data is accessible, providing highly detailed and granular information		
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 ?	The unit for the MNOs are the IRIS zones (small geographical zones for statistical purposes); we can regroup transactions by IRIS for comparision purposes		

Is there auxiliary information to make the data set useful with auxiliary data (NSI or other source) ?	We can link transactions to stores through the SIRENE database (the French National System for the Identification and Directory of Businesses and their Establishments).		
Are there known quality issues with any of the variables ?	The classification of the establishments can sometimes be outdated and does not correspond to the French Classification of Economic Activities (NAF). The terminal of payment are registered under the Merchant Category Codes (MCC) which answers an ISO norm but are not a parfect match to the NAF.		
Does the data contain sensitive variable ? (Meaning legal or ethical issues related to its use)			
Official statistics quality standards			
Is the scope, detail and cost of this source commensurate with needs ?	This source is very accurate for official statistics and is accessible through research partnerships which is an adequate way to access it.		
Is it possible to regularly monitor output quality ?	The company monitors the quality on its own and tries to signal when data are missing or when		
Confidentiality and data protection : Is it possible to put in place the necessary regulatory, administrative, techincal and organisational measures to protect the security and integrity of statistical data and their transmission ?	The server is accessed through a VPN and the data can only be analyzed in a cloud space dedicated to CB and highly securised		
Impartiality and Objectivity : are information on data sources, methods and procedures publicly available ? Is there a process to inform the NSI in case an error is discovered ?	The data are not meant to be publicly available		
Appropriate statistical procedures : are the definitions and concepts used in this source a good approximation of the concepts required for statistical purposes ? Do the data holders collaborate with the NSI in improving data quality (take into account feedback,...)	This has been a good source of approximation of consumption.		
Accuracy and reliability : are source data, integrated data, intermediate results and statistical outputs regularly assessed and validated ?	the source data are regulary assessed for the fight against fraud purposes		
Are sampling errors and non-sampling errors measured and systematically documented ?	rejoint point representativite		
Are revisions regularly analysed in order to improve source data, statistical processes and outputs ?	?		
Timeliness an punctuality : is it possible to release statistics based on this data source on a regular basis ?	yes		
Coherence and Comparability : are statistics based on this source coherent, consistent and comparable over a reasonable period of time ?	yes, the credit card data have been proven to help the consumption studies		
Is cross-national comparability of statistics based on this source possible ? -> has to be discussed between WP members for comparability			
Is access to microdata allowed for research purposes ?	yes to researcher affiliated to the company. The GIECB does not publish its data nor does it make it available		

Sheet 9_ PT_electronic_invoice

Questions on the non-MNO source			
Data Type	**Description**	**Challenges**	**Treatments**
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,…)	Electronic invoices (**e-Fatura**) that can be defined as a mandatory reporting invoices system implemented by the Tax Administration as part of the administrative simplification and anti-fraud measures. Electronic transmission of the invoices issued by individuals or legal entities that have their head office or permanent establishment in Portuguese territory to the Tax and Customs Authority is mandatory. This administrative data includes all the invoicing recorded electronically by the issuer, whether the acquirer / buyer has requested an invoice from.	- To ensure regularity in the monthly transmission by the data supplier; '- High volume of data (100 million records monthly)	- Identification and imputation of extreme outliers: for positive taxable values equal to or greater than 100 million euros and more than 3 standard deviations from their respective mean; for negative taxable values, the values of the highest magnitude are briefly analyzed. - Correction of negative taxable values (less than -100,000 euros), in cases where it is possible to identify a similar symmetrical value (between 95 and 100%), up to 4 months prior. These negative taxable values mainly result from corrections to values transmitted incorrectly in previous months. The total taxable value for the set of records involved remains unchanged, with only the temporal distribution being altered. - Identification and imputation of missing values in a small subset of more significant companies (in terms of number of employees and turnover). The identification of missing values and their respective imputation is carried out based on the company's behavior over time (historical series).
d. **Remote sensing, fixed** (smart meters readings, whether station readings, traffic loop signals, lorry tracking signals,…)			
e. **Remote sensing, mobile** (satellite images, drone images, airborne laser scanning, maritime AIS, …)			
f. **Internet** (web pages, social media posts,…)			
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 ?			
Relevance to the MNO data uses cases			
Improving population coverage : are these data complementary to MNO data ? Do they allow to improve MNO data population representativeness ? Are there difficulties to due to different irreconcialable concepts (definitions, time periods, …) ? Improve of timeliness of analyses and geographical precision.			
Improving precision of analyses : do these data allow to provide preciser analyses (ex : socio-economic indicators, indicators of the mobility type, indicators of the commercial activity,…)			
Broadening the scope of issues covered by official statistics : do these data, combined with MNO data, allow to cover new topics (exposure to air pollution, commuting CO2 emissions,…) ?			

Access			
Who owns the data ? Public administration, one company, several companies ? Rq : different systems, not all the data owned by one company.			
Is it possible to get access ? Does it have to be paid ?			
Are there potentially competing uses (operator publishing similar statistics, etc.)?			
Are there legal or ethical problems to access the data ?			
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 ?			
Is there a possibility to access the data to study its relevance ?			
Is this data available in all EU countries ?			
Metadata			
Is the definition of the population known ? If not do you already have a method to address this issue ?			
Is the base unit of the dataset known ?			
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 other source) ?			
Are there known quality issues with any of the variables ?			
Does the data contain sensitive variable ? (Meaning legal or ethical issues related to its use)			
Official statistics quality standards			
Is the scope, detail and cost of this source commensurate with needs ?			
Is it possible to regularly monitor output quality ?			
Confidentiality and data protection : Is it possible to put in place the necessary regulatory, administrative, techincal and organisational measures to protect the security and integrity of statistical data and their transmission ?			
Impartiality and Objectivity : are information on data sources, methods and procedures publicly available ? Is there a process to inform the NSI in case an error is discovered ?			
Appropriate statistical procedures : are the definitions and concepts used in this source a good approximation of the concepts required for statistical purposes ? Do the data holders collaborate with the NSI in improving data quality (take into account feedback,...)			
Accuracy and reliability : are source data, integrated data, intermediate results and statistical outputs regularly assessed and validated ?			
Are sampling errors and non-sampling errors measured and systematically documented ?			
Are revisions regularly analysed in order to improve source data, statistical processes and outputs ?			
Timeliness an punctuality : is it possible to release statistics based on this data source on a regular basis ?			
Coherence and Comparability : are statistics based on this source coherent, consistent and comparable over a reasonable period of time ?			
Is cross-national comparability of statistics based on this source possible ? -> has to be discussed between WP members for comparability			
Is access to microdata allowed for research purposes ?			