ESSnet Big Data II

Grant Agreement Number: 847375-2018-NL-BIGDATA

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

Work package K

Methodology and quality

Deliverable K6: Quality report template

Draft version, 28.2.2020

Prepared by:
Jacek Maślankowski (GUS, PL)
David Salgado (INE, ES)
Sónia Quaresma (INE, PT)
Gabriele Ascari, Giovanna Brancato, Loredana Di Consiglio, Paolo Righi and Tiziana Tuoto (ISTAT, IT)
Piet Daas (CBS, NL)
Magdalena Six, Alexander Kowarik (STAT, AT)

Work package leader:

Alexander Kowarik (STAT, AT)
alexander.kowarik@statistik.gv.at
telephone : +43 1 71128 7513

Contents

Introduction

The structure of this template is taken from the widely known SIMS (Single Integrated Metadata Structure). The definitions and guidelines are based on the recently updated version of the EHQMR (ESS handbook for quality and metadata reports)[1].

The members of the WPK went through each subconcept of the EHQMR taking into consideration feedback given by other work packages, and for each subconcept we asked

- if it is relevant for new data sources,
- if the definition of the (sub)concept has to be re-worded or
- if they can kept as they were.

If a subconcept was considered as not relevant, we deleted it. Since we kept the numbering of the subconcepts as in the EHQMR, the following subconcepts are not numbered consecutively.

When we had the impression that the existing subconcepts did not cover all relevant quality aspects for new data sources, we introduced new subconcepts. These new subconcepts are indicated by an "A" for "additional" in the subconcept number.

We came across the problem, that the SIMS is generally output-oriented, this means it measures the quality of Official Statistics. When it comes to new data sources, the output so far is almost never an Official Statistics, sometimes it is not even publishable.

In the case of the WPs of the ESSnet Big Data II, the “output” has often more the form of a throughput data set, which could further be used and processed. To avoid a problem with wording, we use the term "statistical output", which stands for the output of the WP, but does not have to be a publishable statistical product.

You can immediately see the changes made by us since they are written in color.

This quality template was tested by the WP members of the pilot track. It was used as basis for a questionnaire about quality issues for the pilot track intermediate meeting from 11th to 12th of December 2019 in Vienna. The feedback from the WPs was used to refine this quality report template. We appended two examples of completed quality templates in the Annex.

Helpful Documents

Here you can find the updated version of the EHQMR, which was officially published in February 2020: https://ec.europa.eu/eurostat/documents/3859598/10501168/KS-GQ-19-006-EN-N.pdf

[1] Both documents and further information can be found on Eurostat’s Quality reporting homepage: https://ec.europa.eu/eurostat/web/quality/quality-reporting

Here is the structure of the original SIMS (no adjustments were made from our side).

2. SIMS V2.0

Item No	Concept name
S.1	**Contact**
S.1.1	Contact organisation
S.1.2	Contact organisation unit
S.1.3	Contact name
S.1.4	Contact person function
S.1.5	Contact mail address
S.1.6	Contact email address
S.1.7	Contact phone number
S.1.8	Contact fax number
S.2	**Metadata update**
S.2.1	Metadata last certified
S.2.2	Metadata last posted
S.2.3	Metadata last update
S.3	**Statistical presentation**
S.3.1	Data description
S.3.2	Classification system
S.3.3	Sector coverage
S.3.4	Statistical concepts and definitions
S.3.5	Statistical unit
S.3.6	Statistical population
S.3.7	Reference area
S.3.8	Time coverage
S.3.9	Base period
S.4	**Unit of measure**
S.5	**Reference period**
S.6	**Institutional mandate**
S.6.1	Legal acts and other agreements
S.6.2	Data sharing
S.7	**Confidentiality**
S.7.1	Confidentiality - policy
S.7.2	Confidentiality - data treatment
S.8	**Release policy**
S.8.1	Release calendar
S.8.2	Release calendar access
S.8.3	User access
S.9	**Frequency of dissemination**
S.10	**Accessibility and clarity**
S.10.1	News release
S.10.2	Publications
S.10.3	On-line database

Item No	Concept name
S.10.3.1	AC1. Data tables - consultations
S.10.4	Micro-data access
S.10.5	Other
S.10.5.1	AC 2. Metadata - consultations
S.10.6	Documentation on methodology
S.10.6.1	AC 3. Metadata completeness - rate
S.10.7	Quality documentation
S.11	**Quality management**
S.11.1	Quality assurance
S.11.2	Quality assessment
S.12	**Relevance**
S.12.1	User needs
S.12.2	User satisfaction
S.12.3	Completeness and R1. Data completeness - rate for U
S.12.3.1	R1. Data completeness - rate for P
S.13	**Accuracy and reliability**
S.13.1	Overall accuracy
S.13.2	Sampling error and A1. Sampling errors - indicators for U
S.13.2.1	A1. Sampling errors - indicators for P
S.13.3	Non-sampling error and A4. Unit non-response - rate for U and A5. Item non-response - rate for U
S.13.3.1	Coverage error
S.13.3.1.1	A2. Over-coverage - rate
S.13.3.1.2	A3. Common units - proportion
S.13.3.2	Measurement error
S.13.3.3	Non response error
S.13.3.3.1	A4. Unit non-response - rate for P
S.13.3.3.2	A5. Item non-response - rate for P
S.13.3.4	Processing error
S.13.3.5	Model assumption error
S.14	**Timeliness and punctuality**
S.14.1	Timeliness and TP2. Time lag - final results for U
S.14.1.1	TP1. Time lag - first results for P
S.14.1.2	TP2. Time lag - final results for P
S.14.2	Punctuality and TP3. Punctuality - delivery and publication for U
S.14.2.1	TP3. Punctuality - delivery and publication for P
S.15	**Coherence and comparability**
S.15.1	Comparability - geographical
S.15.1.1	CC1. Asymmetry for mirror flows statistics - coefficient
S.15.2	Comparability - over time and CC2. Length of comparable time series for U
S.15.2.1	CC2. Length of comparable time series for P

Item No	Concept name
S.15.3	Coherence- cross domain
S.15.3.1	Coherence - sub annual and annual statistics
S.15.3.2	Coherence- National Accounts
S.15.4	Coherence - internal
S.16	**Cost and burden**
S.17	**Data revision**
S.17.1	Data revision - policy
S.17.2	Data revision - practice and A6. Data revision - average size for U
S.17.2.1	A6. Data revision - average size for P
S.18	**Statistical processing**
S.18.1	Source data
S.18.2	Frequency of data collection
S.18.3	Data collection
S.18.4	Data validation
S.18.5	Data compilation
S.18.5.1	A7. Imputation - rate
S.18.6	Adjustment
S.18.6.1	Seasonal adjustment
S.19	**Comment**

S01 Contact

SIMS	Concept Name	Defintion	Guidelines
S.01	Contact	Individual or organisational contact points for the data or metadata, including information on how to reach the contact points.	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.01.1	Contact organisation	The name of the organisation of the contact points for the data or metadata.	Provide the full name (not just code name). of the organisation responsible for the process and outputs (data and metadata) that are the subject of the report.
S.01.6	Contact email address	E-mail address of the contact points for the data or metadata.	Provide the email address(es) of the person(s) indicated as contacts. The address(es) can be (an) individual e-mail address(es) or a mailbox in the organisation to which the person(s) has (have) access.
S.01.7	Contact phone number	The telephone number of the contact points for the data or metadata.	Provide the telephone number(s) of the person(s) indicated as contacts.

S02 Metadata Update

SIMS	Concept Name	Definition	Guidelines
S.02	Metadata update	The date on which the metadata element was inserted or modified in the database.	(Information relating to this concept is provided by reporting on its sub-concepts.)

S.03 Statistical Presentation

SIMS	Concept Name	Defintion	Guidelines
S.03	Statistical presentation	Description of the statistical output.	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.03.1	Data description	Main characteristics of the data set, referring to the statistical output.	Describe briefly the main characteristics of the data in an easily and quickly understandable manner, referring to the main variables. More detailed descriptions of the variables and how they were derived are in S.03.4.
S.03.4	Statistical concepts and definitions	Statistical characteristics of statistical observations, variables.	Define and describe briefly the main statistical variables that have been observed or derived. Indicate their types~~. Indicate discrepancies, if any, from the ESS or international standards.~~ Note that any difference between these variables and the variables desired by users is a relevance issue and is discussed in S.12. Indicate discrepancies, if any, from variables which were previously collected in a different way (e.g. via surveys).
S.3.5	Statistical unit	Entity for which information is sought and for which statistics are ultimately compiled.	Define the type of statistical unit about which data are available, e.g. enterprise, local unit, private household, person, import transaction. If there is more than one type of unit, define each type. Summarize, if possible, the differences to units in traditional ways to collect data.
S.3.6	Statistical population	The total membership or population or "universe" of a defined class of people, objects or events.	Define the target population of the statistical units for which information is sought. ~~The *survey (frame) population* of statistical units (which is the approximation to the target population used in practice) is described in S.18.1.~~ The difference between target population and the actual (frame) population is a coverage issue and is discussed in S13.3 Describe if there are any differences to the populations in traditional official statistics.
S.3.7	Reference area	The country or geographic area to which the measured statistical phenomenon relates.	Describe the country, the regions, the districts, or the other geographical aggregates, to which the data refer. Identify any specific exclusions in the statistical data. ~~If coverage includes overseas territories this should be stated, and they should be specified.~~
S.3.8	Time coverage	The length of time for which data are available.	State the time period(s) covered by the data, e.g. first quarter 2018, or quarters 2015-2018, or year 2018, or years 1985-2018. Note that any issues concerning comparability over time are discussed in S.15.

S.04 Unit of Measure

SIMS	Concept Name	Definition	Guidelines
S.04	Unit of measure	The unit in which the variables of the statistical output are measured.	The statistical data usually involves several units of measure depending upon the variables. ~~Examples are: Euro, national currency, number of persons, and rate per 100,000 inhabitants.~~ ~~The magnitude (e.g., thousand, million) of numerical units should be included.~~ Examples: • Country in which a SIM card is located at a certain time, • position of a ship at a certain time, • consumption of electricity in Watt/Kilowatt in a certain time span, • classifying the negative or positive sentiments of a text input on a -1/1 scale

S.05 Reference Period

SIMS	Concept Name	Defintion	Guidelines
S.05	Reference period	The period of time or point in time to which the measured observation is intended to refer.	The value of a variable refers to a specific time period (for example, the last week of a month, a month, a fiscal year, a calendar year, or several calendar years), or to a point in time (for example, a specific day, or the last day of a month). The variables in a dataset may refer to more than one reference period. All reference periods should be stated. Note that the difference, if any, between the target reference period(s) and the actual reference period(s) is an accuracy issue and should be discussed in S.13.3. Note that if frame population does not include all the units in the target population for the specified reference period, this is a coverage issue and should be discussed in S.13.3.

S.06 Institutional Mandate

SIMS	Concept Name	Definition	Guidelines
S.06	Institutional mandate	~~Law,~~ Set of rules or other formal set of instructions assigning responsibility as well as the authority to an organisation for the collection, processing, and dissemination of statistics.	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.06.1	Legal acts and other agreements	Legal acts or other formal or informal agreements that assign responsibility as well as the authority to an agency for the collection, processing, and dissemination of statistics.	~~State the national legal acts and/or other reporting agreements, including EU legal acts, the implementation of EU directives.~~ Describe the (legal) agreement and other forms of cooperation with the data owner which allow the NSI access to the data source. Describe which forms of reciprocity (not necessarily financial) the NSI offers to the data source.
S.06.A	Data access and data transmission	Arrangements or procedures for data access and data transmission	Describe the arrangements, procedures or agreements for data access and data transmission. In particular, describe • Modes of data access (full access to raw data, access to pre-processed data, on-premise, off-premise) • In case of access to pre-processed data: transparency about the technological processes applied to the pre-processed data • Time and method of transmission • Time horizon of the cooperation - Is a long term access to the data guaranteed?

S.07 Confidentiality

SIMS	Concept Name	Definition	Guidelines
S.07	Confidentiality	A property of data indicating the extent to which their unauthorised disclosure could be prejudicial or harmful to the interest of the source or other relevant parties.	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.07.1	Confidentiality – policy	Legislative measures or other formal procedures which prevent unauthorised disclosure of data that identify a person or economic entity either directly or indirectly.	Describe all European or national legislation, or other formal requirements, that relate to confidentiality. Describe relevant policy (if any). Note that the existence of legislation and/or policy provides some assurance that methods necessary to assure confidentiality have been applied to the data.
S.07.2	Confidentiality - data treatment	Rules applied for treating the datasets to ensure statistical confidentiality and prevent unauthorised disclosure.	For aggregate outputs • Provide the rules that define a *confidential cell*. • Describe the procedures for detecting confidential cells (primary confidentiality) and checking for residual disclosure (derivation or secondary confidentiality); • Describe the procedures for reducing the risk of disclosure by treating confidential cells, for example by perturbation, controlled rounding, cell suppression, or cell aggregation. For micro-level outputs: • Describe the procedures that are used in protecting confidentiality.
S.07.A1	Privacy	How privacy sensitive is the information coming from external data holders?	State how privacy sensitive the information from external data holders is.
S.07.A2	Privacy-protecting treatments	Treatments applied to ensure privacy-sensitive information from external data holders	State any treatments prescribed to satisfy privacy concerns.

S.08 Release Policy

SIMS	Concept Name	Defintion	Guidelines
S.08	Release policy	Rules for disseminating statistical data to all interested parties.	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.08.A	Release policy for Experimental Statistics	Rules for dissemination of experimental data or experimental statistical products.	State if there exists a publicly available policy for the dissemination of experimental statistics and if there exists a designated area at your NSI's homepage.

S.09 Frequency of Dissemination

SIMS	Concept Name	Defintion	Guidelines
S.09	Frequency of dissemination	The time interval at which the statistics are disseminated over a given time period.	State the frequency with which the data are disseminated, e.g. monthly, quarterly, yearly. The frequency can also be expressed by using a code from the harmonised ESS code list so long as this is considered to be easily understandable by users.

S.10 Accessibility and Clarity

SIMS	Concept Name	Defintion	Guidelines
S.10	Accessibility and clarity	The conditions and modalities by which users can access, use and interpret data.	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.10.6	Documentation on methodology	Descriptive text and references to methodological documents available.	List national reference metadata files, methodological papers, summary documents and handbooks relevant to the statistical process. For each item provide the title, publisher, year and link to on-line version (if any). List reference metadata files, methodological papers, summary documents etc. relevant to the process of deriving statistical data from raw data and - if already available - for producing statistical output using the statistical data.
S.10.7	Quality documentation	Documentation on procedures applied for quality management and quality assessment.	List relevant quality related documents, for example, other quality reports, studies. Cross reference to descriptions of quality procedures in other chapters, especially S.13.

S.11 Quality Management

SIMS	Concept Name	Definition	Guidelines
S.11	Quality management	Systems and frameworks in place within an organisation to manage the quality of statistical products and processes.	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.11.1	Quality assurance	All systematic activities implemented that can be demonstrated to provide confidence that the processes will fulfil the requirements for the statistical output.	~~Describe the procedures (such as use of a general quality management system based on EFQM or ISO 9000 series) to promote general quality management principles in the organisation.~~ ~~Describe the quality assurance framework used to implement statistical quality principles.~~ Describe the quality assurance procedures specifically applied to the statistical process for which the report is being prepared, for example agreements with the big data providers, training courses, process monitoring, benchmarking, assessments, and use of best practices. Include descriptions of all forms of quality assessment procedures (such as user satisfaction survey, self-assessment, peer review, compliance monitoring, audit, labelling, certification) and when they most recently took place. Describe any ongoing or planned improvements in quality assurance procedures.
S.11.2	Quality assessment	Overall assessment of data quality, based on standard quality criteria.	Summarise the results of the most recent quality assessments and cross reference to the chapters in the report where the results are presented in more detail.

S.12 Relevance

SIMS	Concept Name	Defintion	Guidelines
S.12	Relevance	The degree to which statistical information meet current and potential needs of the users.	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.12.1	User needs	Description of users and their respective needs with respect to the statistical data.	Provide: • a classification of users, also indicating their relative importance; • an indication of the uses for which users want the statistical outputs; • an assessment of the key outputs desired by different categories of users and any shortcomings in outputs for important users; • information on unmet user needs and any plans to satisfy them in the future; and • details regarding those quality components which do not meet user requirements.
S.12.3	Completeness	The extent to which all statistics that are needed are available.	Provide qualitative information on the extent to which content requirements in relevant legislation, regulations and guidelines are met. Where such requirements are not fully met, reasons for this should be provided. Provide information on the extent to which user needs related to content are satisfied. ~~Provide values of indicator R1 Data completeness rate, for each required data item for each relevant regulation/ guideline at producer/user level of detail as appropriate.~~ ~~In the case where the indicator refers to data sent to Eurostat, this indicator can be compiled by Eurostat.~~
S.12.A	Added Value through new data source	The potential added value of a new data source to an existing statistical product.	Describe if and how the usage of a new data source provides an added value to an already existing statistical product. E.g., this could be more detailed data on particular subgroups, or information on grid level instead of district level or the potential replacement of questions of a survey through information of the new data source.

S.13 Accuracy and Reliability

SIMS	Concept Name	Definition	Guidelines
S.13	Accuracy and reliability	Accuracy of data is the closeness of computations or estimates to the exact or true values that the statistics were intended to measure. Reliability of the data, defined as the closeness of the initial estimated value to the subsequent estimated value.	(Information relating to accuracy is provided by reporting on S.13 sub-concepts. Information on Reliability is reported in S.17 Data Revision).
S.13.1	Overall accuracy	Assessment of accuracy, linked to a certain data set or domain, which is summarising the various components.	Describe the main sources of random and systematic errors in the statistical outputs and provide a summary assessment of all errors with special focus on the impact on key estimates. The bias assessment can be in quantitative or qualitative terms, or both, and may be expressed as bias risk. It should reflect the producer's best current understanding (sign and order of magnitude) and include actions taken to reduce bias. *European level* Provide a summary picture of accuracy across countries. The emphasis placed on various types of errors should depend upon the error profile of the respective process. For repetitive processes, describe how accuracy is developing over time and what efforts are underway to improve accuracy from an ESS perspective. For new data sources, there is a tendency to focus on the micro-level. Please include in this and subsequent sections that reporting at the group or aggregated level can/should be done when the units can not be identified. In general, one should be able to report any quality issues when working with event-based big data sources.

S.13.2	Sampling error	That part of the difference between a population value and an estimate thereof, derived from a random sample, which is due to the fact that only a subset of the population is enumerated.	State whether sampling error is relevant. If probability sampling is used: • for user reports, provide the range of variation of the A1 indicator among key variables at user report level of detail; • for producer reports, provide the range of variation of the A1 indicator among key variables at producer report level of detail; • indicate the impact of sampling error on the overall accuracy of the results; • state how the calculation of sampling error is affected by imputation for nonresponse, misclassifications and other sources of uncertainty, such as outlier treatment. If non-probability sampling is used, provide an assessment of representativeness, a motivation for the invoked model for estimation and risk of sampling bias. *European level* If probability sampling is used: • present sampling errors for key estimates across countries; • indicate which country to country differences are significant and which are not; • for a repetitive survey, describe at least broadly the trends in sampling error over time • provide sampling errors for ESS level estimates.

S.13.3	Non-sampling error	Error in estimates which cannot be attributed to sampling fluctuations	Summarise the most important aspects of coverage, measurement, non-response, processing and model assumption errors. Discuss the corresponding bias risks and actions undertaken to reduce them.
	A4. Unit non-response - rate	The ratio of the number of units with no information or not usable information to the total number of in-scope (eligible) units, at a level of detail appropriate for a user report.	Report A4 Unit non-response-rate
	A5. Item non-response - rate	The ratio of the in-scope (eligible) units that have not responded to a particular item and the in-scope units that are required to respond to that particular item, at a level of detail appropriate for a user report.	Report A5 Item non-response-rate

S.13.3.1	Coverage error	Divergence between the population of the Big Data source and the target population.	Provide information on the frame and its sources and actions performed to gather the population impacting on coverage (e.g. webscraping). Provide an assessment, whenever possible quantitative, of overcoverage and undercoverage, including an evaluation of the bias risks associated with the latter. Describe actions taken for reduction of undercoverage and associated bias risks
S.13.3.1.1	A2. Overcoverage – rate	The proportion of units accessible via the frame that do not belong to the target population.	Report A2, Overcoverage - rate
S.13.3.2	Measurement error	Measurement errors are errors that occur during data acquisition and recording and cause recorded values of variables to be different from the true ones	The main sources of measurement error should be reported and assessed. Their description should be accompanied by any available analysis, otherwise by the producer's best knowledge. Where available and relevant describe: • identification and general assessment of the main sources of measurement error, including errors arising from data acquisition and recording; • efforts made in questionnaire design and testing, information on interviewer training and other work on error prevention; • errors in measurement instruments (meters, satellites,...) • results of assessments based on comparisons with external data, re-interviews or experiments; • results of indirect analysis, for example, of the editing phase; and • actions taken to correct measurement errors.
S.13.3.3	Nonresponse error	Nonresponse errors occur when the big data source fails to collect one or all the variables for units belonging to the domain covered by the source	Provide qualitative/quantitative assessments of unit nonresponse and highlight the units that are most subject to nonresponse Highlight the variables that are most subject to item nonresponse (e.g. associated with sensitive questions) Provide a qualitative/quantitative assessments of the bias associated with nonresponse, comparing response rate for different sub-groups or distribution of auxiliary variables known for respondents and non-respondets (etc.) Provide a breakdown of nonrespondents according to cause for nonresponse, mainly focusing on unit dependent cause and data acquisition tools cause. Describe efforts to reduce nonresponse during data acquisition and follow-up. Define a strategy for reducing nonresponse during data acquisition and follow-up. Implement an estimator adjusted for nonresponse .

S.13.3.4	Processing error	The error in final data collection process results arising from the faulty implementation of correctly planned implementation methods, e.g., algorithms used to transform the data or extract information from raw data.	If processing errors are significant, identify the main issues regarding them. Present an analysis of processing errors, where available, otherwise a qualitative assessment. Report their extent, and impact on the outputs, of the most significant types of error. Include descriptions of linking and coding errors, if applicable. Where mistakes relating to programming or publishing have occurred, corrective measures taken as well as actions for avoiding them in the future should be reported. Example: For web data source, setting up a pipeline assures processing is comparable over time. Because texts were processed, the final results were highly affected by the various choices of text processing made.
S.13.3.5	Model assumption error	Error due to models used in the statistical production.	Describe process specific models, for example, as needed to define the target of estimation itself and models used for transformation of data into statistical data. Provide an assessment of the validity of each model. (Descriptions of models used in treatment of specific sources of error should be presented in the section dealing with those errors.) The assessment of the models used in treatment of specific sources of error should be presented in this section. Discuss the trade-off between the need to use proper model that can change over time (accuracy) and to use a constant model in order to ensure comparability over time

S.14 Timeliness and Punctuality

SIMS	Concept Name	Defintion	Guidelines
S.14	Timeliness and punctuality	(Defined by its sub-concepts)	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.14.1	Timeliness	Length of time between data availability and the event or phenomenon the data describe.	Outline the reasons for the time lag. Outline efforts to reduce time lag in future. Describe the envisioned time lag for producing statistical output from/with the help of a new data source. Describe if the use of the new data source has the potential to decrease the time lag as it exists at the moment for already existing statistical products.

S.15 Coherence and Comparability

SIMS	Concept Name	Defintion	Guidelines
S.15	Coherence and Comparability	Adequacy of statistics to be reliably combined in different ways and for various uses and the extent to which differences between statistics can be attributed to differences between the true values of the statistical characteristics.	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.15.1	Comparability – geographical	The extent to which statistics are comparable between geographical areas.	Describe any problems of comparability between regions of the country. The reasons for the problems should be described and as well an assessment (preferably quantitative) of the possible effect on the output values. Give information on discrepancies from the ESS/ international concepts, definitions, with reference to other chapters for more details.
S.15.2	Comparability – over time	The extent to which statistics are comparable or reconcilable over time.	Provide information on possible limitations in the use of data for comparisons over time. Distinguish three broad possibilities: 1. There have been no changes, in which case this should be reported. 2. There have been some changes but not enough to warrant the designation of a break in series. 3. There have been sufficient changes to warrant the designation of a break in series. Additionally, provide information about the comparability over time of the technological processes which produce the data, of the data access and changes in the covered population over time. Give also an assessment how the comparison over time will develop in the future.
S.15.3	Coherence-cross domain	The extent to which statistics are reconcilable with those obtained through other data sources or statistical domains.	An analysis of incoherence should be provided, where this is an issue of importance.
S.15.4	Coherence – internal	The extent to which statistics are consistent within a given data set.	Each set of outputs should be internally consistent. If statistical outputs within the data set in question are not consistent, any resulting lack of coherence in the output of the statistical process itself should be stated as well as a brief explanation of the reasons for publishing such results.

S.15.A.1	Coherence - with existing information/ Official Statistics	The extent to which information / statistical output from new data sources is consistent with information /statistical output from traditional data sources.	Provide information if it is meaningful to compare the information gained from new data sources with information from traditional data sources and if so, how consistent the information /statistical output gained from new data sources is with the one from traditional data sources.
S.15.A.2	Comparability - between information from several distinct new data sources	The extent to which information from several distinct new data sources is comparable among one another.	If you have raw data from several distinct new data sources, provide information how comparable the respective raw data sets and the information derived from them are among one other. Examples: MNO data from several mobile operators, smart meter data from several electricity providers

S.16 Cost and Burden

SIMS	Concept Name	Defintion	Guidelines
S.16	Cost and burden	Cost associated with the collection and production of a statistical product and burden on respondents.	***Cost*** Provide annual operational costs of the process, with breakdown by major cost component. Describe recent efforts to improve efficiency and comment on the extent to which information and communication technology is used. ***Burden*** Provide an estimate of the respondent burden imposed by the process. Describe all the means taken to minimise burden.
S.16.A	Potential savings in cost and burden	Description how the new data source might influence cost and burden in the future	Provide an overview how the new data source could be deployed in the future to save the NSIs cost and/or decrease the respondent burden. Provide a qualitative description of the additional efforts for the NSI and the data owners.

S.17 Data Revision

SIMS	Concept Name	Definition	Guidelines
S.17	Data revision	Any change in a value of a statistic released to the public.	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.17.1	Data revision – policy	Policy aimed at ensuring the transparency of disseminated data, whereby preliminary data are compiled that are later revised.	Describe the data revision policy applicable to data output from the statistical process being reported. In so far as they are relevant to the process being reported, summarise the general procedures for treatment of planned revisions, benchmark revisions, unplanned revisions, and revisions due to conceptual and/or methodological changes..

S.18 Statistical Processing

SIMS	Concept Name	Definition	Guidelines
S.18	Statistical processing	(Defined by its sub-concepts)	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.18.1	Source data	Characteristics and components of the raw statistical data used for compiling statistical aggregates.	Indicate if the data are based on a survey, an administrative data source, multiple data sources, big data source (machine generated, human sourced, process mediated), e.g., web data, and/or macro-aggregates. Refer to the accreditation document of the data source, if applies. In the event of multiple data sources or macro-aggregate processes describe each source. For each survey source, report the survey population, cross referencing the description of the target population in S.03.6, and summarise the sample design. For each data source ~~from an administrative source~~, summarise the source, its primary purpose, and the most important data items acquired. Indicate information in which form the metadata for the new data source is available, where it can be found, and if it is updated on a regular basis.
S.18.2	Frequency of data acquisition and recording	Frequency with which the source data are acquired	Indicate the frequency of data acquisition (e.g. monthly, quarterly, annually, or continuous).

S.18.3	Data acquisition and recording	Systematic process of gathering data for official statistics.	*For each survey data source:* • describe the method(s) used to gather data from respondents; • annex or hyperlink the questionnaires(s). *For each administrative data source* • describe the acquisition process and how it was tested. *For all sources* • describe the types of checks applied at the time of data entry. *For big data sources* • describe the methods used to for data acquisition and recording; • add hyperlink if it is web data or name of the API used to collect the data.
S.18.4	Data validation	Process of monitoring the results of data compilation and ensuring the quality of statistical results.	Describe the procedures for checking and validating the source data and how the results are monitored and used. Describe the procedures for validating the aggregate output data (statistics) after compilation, including checking coverage and response rates, and comparing with data for previous cycles and with expectations. List other output datasets to which the data relate and outline the procedures for identifying inconsistencies between the output data and these other datasets. Define the linkage method for big data sources and other data sources used for validation.

S.18.5	Data compilation	Operations performed on data to derive new information according to a given set of rules.	If there is missing data, give detailed description of the methods used for imputation. For big data sources, e.g., web data, indicate the reason why data were not collected (technical issues etc.). Describe the procedures for imputation, the most common reasons for imputation and imputation rates within each of the main strata. Describe the likely impact of imputation. Describe the procedures for adjustment for non-response and the corrections to the design weights to account for differences in response rates. Describe the calculation of design weights, including calibration (if used). Describe the procedures for combining input data from different sources.
S.18.5.1	A7. Imputation – rate	The ratio of the number of replaced values to the total number of values for a given variable.	Provide values of indicator A7 Imputation – rate

S.19 Comment

SIMS	Concept Name	Defintion	Guidelines
S.19	Comment	Supplementary descriptive text which can be attached to data or metadata.	Provide any information • that is pertinent to the report but does not fit under any of the other concepts, or • to repeat key issues, or • to make reference to annexes that might be attached to the report.

Annex I: Example of Completed Quality Questionnaire: MNO Data (WPI)

Specific Comment on MNO Data

We will fill in the quality questionnaire both for an intermediate dataset and a final statistical output. For the intermediate data set we can think of the location probabilities of each device per time period unit. Notice that it is intermediate inasmuch as it can be used as an input for the statistical layer in the Reference Methodological Framework to produce statistical outputs in different statistical domains (e.g. present population in demography, domestic tourists in tourism statistics, etc.). These location probabilities basically amount to the conditional probabilities and , where stands for the random variable location tile for mobile device *d* at time *t* in a chosen reference grid i=1, 2, 3,…, N.

For the statistical output we will consider the number of domestic tourists and their trips. These choices are not to be definitive in either cases. We can devise more intermediate data sets (e.g. those including the interactions between mobile devices such as calls and SMS/MMS). In this exercise, the statistical outputs in tourism statistics are preferred over simpler options like present population since they need more computation to detect domestic tourists among the mobile network data. Thus, we can assess this quality questionnaire from a wider perspective.

In any case, currently we do not have a definitive statistical product and the answers provided below are given in terms of the proposed methodological framework to build these statistical outputs (still ongoing work in WPI).

Answers for the intermediate dataset will be denoted with the initial I and for the statistical output with the initial O. Additionally, we skipped here some of the concpets (e.g. S01 Contact, S02 Metadata Update).

S.03 Statistical Presentation

SIMS	Concept Name	Defintion	Guidelines	Answer (I)	Answer (O)
S.03	Statistical presentation	Description of the statistical output.	(Information relating to this concept is provided by reporting on its sub-concepts.)	Location probabilities per working time period unit (x sec, x min,…) for each mobile device detected in the network.	Number of domestic tourists and their trips broken down per dissemination territorial cell and per dissemination time period unit. Depending on the statistical method used we can provide either just point estimations and confidence intervals or alternatively the (posterior) distribution.
S.03.1	Data description	Main characteristics of the data set, referring to the statistical output.	Describe briefly the main characteristics of the data in an easily and quickly understandable manner, referring to the main variables. More detailed descriptions of the variables and how they were derived in S.03.4.	For each device and each working time period (x sec, x min,…) we provide the probability of location in each cell of a reference grid together with the joint probability for successive time periods.	Number of domestic tourists and their trips broken down per dissemination territorial cell and per dissemination time period unit.

S.03.4	Statistical concepts and definitions	Statistical characteristics of statistical observations.	Define and describe briefly the main statistical variables that have been observed or derived. Indicate their types. Note that any difference between these variables and the variables desired by users is a relevance issue and is discussed in S.12. 32	1) Raw telco variables (cell/antenna ID, coverage area, antenna position coordinates, etc.) depend on the agreement with the MNOs and on the underlying technological infrastructure of the network. 2) Location probabilities are computed basically as conditional probabilities upon the observed telco variables. 3) Auxiliary variables such as those regarding the land use, geographical information, etc. may be used. 4) In the modelling exercise for the computation of probabilities, parameters such as the probability of staying at the same cell may be introduced but estimated using the variables indicated above.	1) Raw telco variables (cell/antenna ID, time advance, etc.) depend on the agreement with the MNOs and on the underlying technological infrastructure of the network. 2) Location probabilities are computed basically as conditional probabilities upon the observed telco variables. 3) Auxiliary variables such as those regarding the land use, geographical information, etc. may be used. 4) In the modelling exercise for the computation of probabilities, parameters such as the probability of staying at the same cell may be introduced but computed using the variables indicated above. 5) Local penetration rates of each MNO. 6) Official population data. 7) If available, survey estimates of the number of domestic tourists and of their trips. 8) In the modelling exercise for the number of domestic tourists and trips, parameters may be introduced but estimated using the variables and data introduced above.

S.3.5	Statistical unit	Entity for which information is sought and for which statistics are ultimately compiled.	Define the type of statistical unit about which data are available, e.g. enterprise, local unit, private household, person. If there is more than one type of unit, define each type.	The base unit is the network event, i.e. an event producing a digital trace in the network (calls, SMS/MMS, Internet connections, LA updates, etc.). There may be a difference in this set depending on the kind of telco data (CDRs, radio signaling data, core network signaling data, etc.)	The target statistical unit is the domestic tourist and the domestic trips. As far as possible, these population units will be identified as a result of the implementation of algorithms translating the reglementary official definitions.

S.3.6	Statistical population	The total membership or population or "universe" of a defined class of people, objects or events.	Define the target population of the statistical units for which information is sought. Note that a difference between the target population and the population desired by users is a relevance issue and is discussed in S.12; and the difference between target population and the actual (frame) population is a coverage issue and is discussed in S13.3 If there is more than one type of population, define each type.	The target population is the complete set of network events allowing us to geolocate each mobile device in the chosen working time period unit.	The target populations are the domestic tourists and the domestic trips according to the reglementary official definitions.

S.3.7	Reference area	The country or geographic area to which the measured statistical phenomenon relates.	Describe the country, the regions, the districts, or the other geographical aggregates, to which the data refer. Identify any specific exclusions in the statistical data.	Mobile network data corresponds to the whole national territory under analysis. There will be a minimal working territorial cell of no more than 1km x 1km and aggregation into city districts, municipalities, provinces, and regions will immediately be possible. These minimal territorial cells will not need to coincide with the dissemination cells (upon agreement with the MNOs).	Geographical breakdown of the whole national territory will scale down to commonly used dissemination cells. Finer cells will internally be explored, but the dissemination will be subjected to agreements with the MNOs (as part of the exploration of the partnerships).

S.3.8	Time coverage	The length of time for which data are available.	State the time period(s) covered by the data, e.g. first quarter 2018, or quarters 2015-2018, or year 2018, or years 1985-2018. Note that any issues concerning comparability over time are discussed in S.15.	The two involved time scales will depend on the agreements reached with MNOs: (i) duration of the whole period under analysis for research purposes is usually limited to weeks or perhaps few months, (ii) the time frequency of the network events will depend on the type of raw telco data used for the analysis (CDRs or signaling data). In the long term, the goal is to have sustainable access to MNO data. If either continuous or in given time periods (daily, weekly, monthly, etc.) will depend on details about the access. The data frequency will also depend on the technological infrastructure of MNOs.	Time breakdown of the disseminated aggregates will scale down to commonly used dissemination frequency (monthly). Shorter time periods will internally be explored, but the dissemination will be subjected to agreements with the MNOs (as part of the exploration of the partnerships).

Additional for big data				Location probabilities for each device must never be publicly disseminated. They must be an undisclosed intermediate dataset in a wider process.	Time and spatial breakdown is a critical issue regarding the collaboration with MNOs with a commercial business line around MNO data. If NSIs disseminate highly disseminated outputs, MNOs will possibly lose the incentive to monetise their statistical products. An agreement should be reached according to some preliminary experience (ongoing efforts).

S.04 Unit of Measure

SIMS	Concept Name	Definition	Guidelines	Answer (I)	Answer (O)
S.04	Unit of measure	The unit in which the variables of the statistical output are measured.	The statistical data usually involves several units of measure depending upon the variables. Examples: • Country in which a SIM card is located at a certain time, • position of a ship at a certain time, • consumption of electricity in Watt/Kilowatt in a certain time span, • Classifying the negative or positive sentiments of a text input on a -1/1 scale	Location probabilities of each device will be provided with respect to a grid of reference and time frequency dependent on the access agreement and available data. Cells not larger than 1km x 1 km will be explored and time frequencies no longer than 1 day will be analysed.	Number of domestic tourists and domestic trips will be provided in the time and spatial breakdown mentioned above.
Additional for big data				The current analysis is focused on population counts. The target variables are thus adimensional.	The current analysis is focused on population counts. The target variables are thus adimensional.

S.05 Reference Period

SIMS	Concept Name	Defintion	Guidelines	Answer (I)	Answer (O)
S.05	Reference period	The period of time or point in time to which the measured observation is intended to refer.	The value of a variable refers to a specific time period (for example, the last week of a month, a month, a fiscal year, a calendar year, or several calendar years), or to a point in time (for example, a specific day, or the last day of a month). The variables in a dataset may refer to more than one reference period. All reference periods should be stated Note that the difference, if any, between the target reference period(s) and the actual reference period(s) is an accuracy issue and should be discussed in S.13.3. Note that if frame population does not include all the units in the target population for the specified reference period, this is a coverage issue and should be discussed in S.13.3.	The set of reference periods will correspond to the whole period under analysis for which we have access to data. The duration of each reference period will be minimally divided into the usual regulatory timescale, but shorter time periods will be explored (days, weeks, etc.).	The set of reference periods will correspond to a shorter time period within the whole period under analysis (for research purposes usually limited to weeks or perhaps few months due to data access restrictions). The duration of each reference period will be minimally divided into the usual regulatory timescale, but shorter time periods will be explored (days, weeks, etc.).

Additional for big data				To the extent feasible, we shall try to provide output for a sequence of reference periods, and not just a single one-shot result.	To the extent feasible, we shall try to provide output for a sequence of reference periods, and not just a single one-shot result. Part of the data may be reserved to train algorithms (home/work detection, trip detection, etc.).

S.06 Institutional Mandate

SIMS	Concept Name	Definition	Guidelines	Answer (I)	Answer (O)
S.06	Institutional mandate	Set of rules or other formal set of instructions assigning responsibility as well as the authority to an organisation for the collection, processing, and dissemination of statistics.	(Information relating to this concept is provided by reporting on its sub-concepts.)	No legal regulation for location probabilities for each mobile device. Being an intermediate dataset for the production process in multiple statistical domains, here we detect an important difference with final statistical outputs.	Tourism Statistics for Resident Population is regulated by XXXXX.

S.06.1	Legal acts and other agreements	Legal acts or other formal or informal agreements that assign responsibility as well as the authority to an agency for the collection, processing, and dissemination of statistics.	Describe the (legal) agreement and other forms of cooperation with the data owner which allows the NSI access to the data source. Describe which forms of reciprocity (not necessarily financial) does the NSI offer to the data source?	In principle, data collection for official statistical purposes is legally supported by the Spanish National Statistical Act (RD12/89) when referred to a concrete statistical program included in the Spanish National Statistical Plan. Requesting access to raw data to compile this kind of intermediate dataset seems to require a new interpretation of the Spanish National Statistical Act (not discarded in the current wording but it would need some debate). If the intermediate dataset is compiled to produce diverse official statistics, the legal support is clearly stated in the Law. However, extraction costs and especially preprocessing costs (necessary for this data source) are not mentioned in the Law for data sources like this. A literal reading of the Law supports the Spanish NSI to request data, but this data do not exist unless extraction and preprocessing is conducted by MNOs, which imply non-negligible costs. Furthermore, if an MNO lacks the technological infrastructure to extract and preprocess raw telco data, there does not exist a solution to request data to this MNO. This would introduce a	When considering a final statistical output included in the Spanish National Statistical Plan there is no doubt about the legal support to request data. However, the issue about the costs mentioned for the intermediate dataset is still present.

S.06.A	Data access and data transmission	Arrangements or procedures for data access and data transmission	Describe the arrangements, procedures or agreements for data access and data transmission. In particular, describe • Modes of data access (full access to raw data, access to pre-processed data, on-premise, off-premise) • In case of access to pre-processed data: transparency about the technological processes applied to the pre-processed data • Time and method of transmission • Time horizon of the cooperation - Is a long term access to the data guaranteed?	The current solution considered is the on-premise processing. No data is transmitted to Statistics Spain (INE). No direct access to data is granted to official statisticians, but the MNO will implement the exact specifications provided by Statistics Spain (INE). This implementation is agreed to be sequentially approached in independent modules. Currently, this seems to be the optimal solution, up to the issue about the costs. No long-term agreement has been reached so far, but this aspect is part of the goal of the current research and is in the interest of both the MNO and Statistics Spain (INE).	The current solution considered is the on-premise processing. No data is transmitted to Statistics Spain (INE). No direct access to data is granted to official statisticians, but the MNO will implement the exact specifications provided by Statistics Spain (INE). This implementation is agreed to be sequentially approached in independent modules. Currently, this seems to be the optimal solution, up to the issue about the costs. No long-term agreement has been reached so far, but this aspect is part of the goal of the current research and is in the interest of both the MNO and Statistics Spain (INE).

S.07 Confidentiality

SIMS	Concept Name	Definition	Guidelines	Answer (I)	Answer (O)
S.07	Confidentiality	A property of data indicating the extent to which their unauthorised disclosure could be prejudicial or harmful to the interest of the source or other relevant parties.	(Information relating to this concept is provided by reporting on its sub-concepts.)	Location probabilities from mobile network data are extremely sensitive to disclosure and they should be protected at all costs.	Aggregates for tourism statistics from mobile network data have a priori the same degree of sensitivity than those from traditional sources, with the exception of the degree of breakdown (both in time and in space). Statistical disclosure control procedures will need to be investigated in this regard (not planned in the WPI).

S.07.2	Confidentiality - data treatment	Rules applied for treating the datasets to ensure statistical confidentiality and prevent unauthorised disclosure.	For aggregate outputs • Provide the rules that define a *confidential cell*. • Describe the procedures for detecting confidential cells, including checking for residual disclosure. • Describe the procedures for eliminating confidential cells, for example by controlled rounding, cell suppression, or cell aggregation. For micro-level outputs: • Describe the procedures that are used in protecting confidentiality.	No dissemination of location probabilities will be conducted.	Not investigated and not even planned in the current project.

S.07.A Additional for big data	Privacy	How privacy sensitive is the information coming from external data holders?	State which treatments are prescribed to satisfy privacy concerns	The privacy of mobile network data is a strong issue in many respects (access, processing, dissemination). No dissemination of intermediate datasets with location probabilities is intended.	Privacy of information for aggregate datasets of tourism statistics is subjected to the same regulations as with traditional data sources. There might be an issue regarding the potential degree of disaggregation.

S.08 Release Policy

SIMS	Concept Name	Defintion	Guidelines	Answer (I)	Answer (O)
S.08	Release policy	Rules for disseminating statistical data to all interested parties.	(Information relating to this concept is provided by reporting on its sub-concepts.)	No intermediate dataset with location probabilities is to be disseminated.	Aggregates for tourism statistics are to be shared as experimental statistics and for research purposes.
S.08.A Additional for big data	Release policy for Experimental Statistics	Rules for dissemination of experimental data or experimental statistical products.	State if there exists a publicly available policy for the dissemination of experimental statistics and if there exists a designated area at your NSI's homepage.	*This is an organisation-related topic and is kept for the deliverable, but not asked in the quality-related questionnaire for the track meeting in December.*	*This is an organisation-related topic and is kept for the deliverable, but not asked in the quality-related questionnaire for the track meeting in December.*

S.09 Frequency of Dissemination

SIMS	**Concept Name**	**Defintion**	**Guidelines**	**Answer (I)**	**Answer (O)**
S.09	Frequency of dissemination	The time interval at which the statistics are disseminated over a given time period.	State the frequency with which the data are disseminated, e.g. monthly, quarterly, yearly. The frequency can also be expressed by using a code from the harmonised ESS code list so long as this is considered to be easily understandable by users.	No dissemination of intermediate datasets with location probabilities is to be conducted.	Not yet known. An object of study is the degree of dissemination in time which we can achieve in producing reliable outputs. Depending on this, on the MNO-NSI agreements and on stakeholders' interests, this time frequency will be decided.
Additional for big data					

S.10 Accessibility and Clarity

SIMS	Concept Name	Defintion	Guidelines	Answer (I)	Answer (O)
S.10	Accessibility and clarity	The conditions and modalities by which users can access, use and interpret data.	(Information relating to this concept is provided by reporting on its sub-concepts.)	No access will be granted to location probabilities whatsoever.	Access to aggregate datasets for tourism statistics as experimental statistics will only be granted under agreement with the MNO (as part of the access conditions).
S.10.7	Quality documentation	Documentation on procedures applied for quality management and quality assessment.	List relevant quality related documents, for example, other quality reports, studies. Cross reference to descriptions of quality procedures in other chapters, especially S.13.	Quality assessment of location probabilities has not yet been designed. Certainly, comparison with population statistics, land use information, and traditional tourism statistics will be considered.	Quality assessment of location probabilities has not yet been designed. Certainly, comparison with population statistics, land use information, and traditional tourism statistics will be considered.

S.11 Quality Management

SIMS	Concept Name	Definition	Guidelines	Answer (I)	Answer (O)
S.11	Quality management	Systems and frameworks in place within an organisation to manage the quality of statistical products and processes.	(Information relating to this concept is provided by reporting on its sub-concepts.)	Processing will take place almost entirely within MNO's premises. Access is granted only indirectly via their own data scientists. Systems and frameworks are thus designed and deployed by the private organization.	Processing will take place almost entirely within MNO's premises. Access is granted only indirectly via their own data scientists. Systems and frameworks are thus designed and deployed by the private organization.

S.11.1	Quality assurance	All systematic activities implemented that can be demonstrated to provide confidence that the processes will fulfil the requirements for the statistical output.	Describe the quality assurance procedures specifically applied to the statistical process for which the report is being prepared, for example agreements with the big data providers, benchmarking, assessments, and use of best practices. Include descriptions of all forms of quality assessment procedures (self-assessment, peer review, compliance monitoring, audit) and when they most recently took place. Summarise the results of the most recent quality assessments and cross reference to the chapters in the report where the results are presented in more detail. Describe any ongoing or planned improvements in quality assurance procedures.	*Only quality assurance procedures which affect the new data sources are relevant for the track meeting.* Starting from preprocessed network event data producing a first set of geolocation variables for each mobile device, the construction of location probabilities will be undertaken in collaboration with MNOs' data scientists. They will implement the methodology proposed by our NSI to produce the location probabilities and assess them jointly in coordination. A protocol of quality assurance has not been designed yet since we are still in a preliminary stage. Audit and processing disclosure control may be part of the private-public agreement.	*Only quality assurance procedures which affect the new data sources are relevant for the track meeting.* Starting from preprocessed network event data producing a first set of geolocation variables for each mobile device and using the construction of location probabilities, we will be undertake the construction of aggregates in collaboration with MNOs' data scientists. They will implement the methodology proposed by our NSI to produce the location probabilities and assess them jointly in coordination. A protocol of quality assurance has not been designed yet since we are still in a preliminary stage. Audit and processing disclosure control may be part of the private-public agreement.

S.11.2	Quality assessment	Overall assessment of data quality, based on standard quality criteria.	Summarise the results of the most recent quality assessments and cross reference to the chapters in the report where the results are presented in more detail.	*Only quality assurance procedures which affect the new data sources are relevant for the track meeting.* Yet to be done (depending on the statistical methods to be used). Traditional bias and variance considerations will be taken into account, but probably more elements will be needed (e.g. model checking and model assessment). Probability theory is used from the onset so that uncertainty evaluation can be adequately formulated.	*Only quality assurance procedures which affect the new data sources are relevant for the track meeting.* Yet to be done (depending on the statistical methods to be used). Traditional bias and variance considerations will be taken into account, but probably more elements will be needed (e.g. model checking and model assessment). Bayesian techniques are under consideration so that the final posterior distributions and predictive posterior distributions can provide a way to assess quality.
Additional for big data					

S.12 Relevance

SIMS	Concept Name	Defintion	Guidelines	Answer (I)	Answer (O)
S.12	Relevance	The degree to which statistical information meet current and potential needs of the users.	(Information relating to this concept is provided by reporting on its sub-concepts.)	Location probabilities only meet internal users' needs. They are not intended for dissemination. They are to be used in the production of official statistics in a diversity of domains.	For aggregates, relevance is provided through the intended degree of time and spatial breakdown. Currently, focus is set on estimating totals, but in the future some new insights will be possibly analyses using mobile network data, thus reinforcing relevance.

S.12.1	User needs	Description of users and their respective needs with respect to the statistical data.	Provide: • a classification of users, also indicating their relative importance; • an indication of the uses for which users want the statistical outputs; • an assessment of the key outputs desired by different categories of users and any shortcomings in outputs for important users; • information on unmet user needs and any plans to satisfy them in the future; and • details regarding those quality components which do not meet user requirements.	Intermediate outputs with location probabilities are only intended for internal uses in NSIs to produce outputs in different statistical domains. A limitation with these location probabilities is that they can assist only in the production of displacements of statistical units (tourism, commuters, present population, etc.), This source cannot provide other information like expenditure, etc.	Tourism statistics is produced under both national and European regulations. Users are (i) internal users such as the National Accounts unit, (ii) other public administrations such as the Ministry of Tourism and municipalities, and (iii) private organizations in the industry of tourism such as hotel chains.

S.12.3	Completeness	The extent to which all statistics that are needed are available.	Provide qualitative information on the extent to which content requirements in relevant legislation, regulations and guidelines are met. Provide information on the extent to which user needs related to content are satisfied. Provide values of indicator R1 Data completeness rate, for each required data item for each relevant regulation/ guideline at producer/user level of detail as appropriate. In the case where the indicator refers to data sent to Eurostat, this indicator can be compiled by Eurostat. *European level* Summarise across countries the extent to which ESS requirements for data items are met	Location probabilities are not intended for public use, only for internal use to produce different statistics.	Not all variables requested in the legal regulations about tourism statistics can be obtained from mobile network data. Only those related to people displacements and stays (e.g. number of tourists, number of nights, etc.) can be possibly estimated. Other variables such as number of beds, number of bedrooms, etc. need another data source. Thus, to fully fulfill legal regulations a combination of sources is needed.

S.12.A Additional for big data	Added Value through new data source	The potential added value of a new data source to an existing statistical product.	Describe if and how the usage of a new data source provides an added value to an already existing statistical product. E.g., this could be more detailed data on particular subgroups, or information on grid level instead of district level or the potential replacement of questions of a survey through information of the new data source.	There does not exist an analog for location probabilities in traditional data sources. Thus the value is completely new in this line.	For aggregates, the level of time and spatial disaggregation potentially reached with mobile network data is unattainable with survey data.

S.13 Accuracy and Reliability

SIMS	**Concept Name**	**Definition**	**Guidelines**	**Answer (I)**	**Answer (O)**
S.13	Accuracy and reliability	Closeness of computations or estimates to the exact or true values that the statistics were intended to measure. Reliability of the data, defined as the closeness of the initial estimated value to the subsequent estimated value.	(Information relating to accuracy is provided by reporting on S.13 sub-concepts. Information on Reliability is reported in S.17 Data Revision).	Yet to be done. (Being probabilities an adequate accuracy measure should be chosen.)	For aggregates, we have chosen a Bayesian approach to produce posterior distributions so that bias, variance, credible intervals, and model checking can be assessed in a systematic way.

S.13.1	Overall accuracy	Assessment of accuracy, linked to a certain data set or domain, which is summarising the various components.	Describe the main sources of random and systematic errors in the statistical outputs and provide a summary assessment of all errors with special focus on the impact on key estimates. The bias assessment can be in quantitative or qualitative terms, or both, and may be expressed as bias risk. It should reflect the producer's best current understanding (sign and order of magnitude) and include actions taken to reduce bias. *European level* Provide a summary picture of accuracy across countries. The emphasis placed on various types of errors should depend upon the error profile of the respective process. For repetitive processes, describe how accuracy is developing over time and what efforts are underway to improve accuracy from an ESS perspective.	Since the intermediate datasets as outputs are comprised of probability distributions for the location of network events along the time period of study, the concepts of bias and variance should be properly understood (we are not producing point estimations). More work is still needed. In quantitative terms, we have not identified yet the target figure of merit to express the degree of "unbiasedness". Neither have we identified the equivalent to the variance. In qualitative terms, the "bias" will be the result of the modelling and understanding exercise for the production of the telco variables (network events) and for the dynamical behavior of the mobile devices (transition probabilities, etc.). The nature of data (CDR or signaling data) will certainly have an effect of the	For the aggregates, apart from the uncertainty accumulated in the location probabilities, we shall apply a Bayesian hierarchical model to compute the posterior distribution for the total number of individuals per time unit and per territorial cell of reference. The model will be central in the assessment of accuracy, both in terms of bias and variance, but also of model checking. We will use official data as benchmark in the estimation procedure. The assessment of these quantities will be empirically tested with the data simulator. The nature of data (CDR or signaling data) will certainly have an effect of the estimation of these location probabilities.

Additional for big data				The use of statistical models is central in the inference process. Accuracy assessment as well as robustness evaluation of both parameters and model needs to be evaluated. This depends very much on the choice of statistical methods. The use of auxiliary information (official data, land use, etc.) is important.	The use of statistical models is central in the inference process. Accuracy assessment as well as robustness evaluation of both parameters and model needs to be evaluated. This depends very much on the choice of statistical methods. The use of auxiliary information (official data, land use, etc.) seems to be important. It is important to notice that rigorously the exact meaning of bias is different in the case of sampling designs and in the case of statistical models. Design-bias is not the same as model-bias.

S.13.2	Sampling error	That part of the difference between a population value and an estimate thereof, derived from a random sample, which is due to the fact that only a subset of the population is enumerated.	State whether sampling error is relevant. If probability sampling is used: • for user reports, provide the range of variation of the A1 indicator among key variables at user report level of detail; • for producer reports, provide the range of variation of the A1 indicator among key variables at producer report level of detail; • indicate the impact of sampling error on the overall accuracy of the results; • state how the calculation of sampling error is affected by imputation for nonresponse, misclassifications and other sources of uncertainty, such as outlier treatment. If non-probability sampling is used, provide an assessment of representativity and risk of sampling bias. *ESS level*	For intermediate datasets with location probabilities for every mobile device, only those devices in a given MNO are considered at a time. The evaluation of these location probabilities is conducted individually per each mobile device. At this stage, we understand sampling error in the sense of the "sampling error" to estimate these location probabilities for each device. As before with bias, we need to better understand what we mean by "sampling error" for a probability distribution (we are not producing point estimations). Under this way of understanding "sampling error" for each location probability distribution, we may consider that CDRs are a form of sampling (in time) to geolocate the mobile device, whereas signaling data in some sense produces a "census" of all data generated by the device. Complementary, according to current technology, signaling data differ in spatial accuracy (lower) with respect to CDRs (higher).	For aggregates, apart from the uncertainty accumulated in the location probabilities, there is a sampling mechanism to generate our dataset. If we are working only with some of the MNOs in the country, the sampling is obvious. If we are working with all MNOs in the country (not the case right now), we still do have a sample of the target population. This could be assessed through the penetration rates. The sampling error is accounted for by the model connecting the dataset (mobile devices) with the target population (tourists). Since we have chosen a Bayesian approach, the posterior distribution provides a way to assess the variance in the estimation. It is important to notice that rigorously the exact meaning of sampling error is different in the case of sampling designs and in the case of statistical models. Design

S.13.3	Non-sampling error	Error in estimates which cannot be attributed to sampling fluctuations	Summarise the most important aspects of coverage, measurement, non-response, processing and model assumption errors. Discuss the corresponding bias risks and actions undertaken to reduce them.	Again, for the intermediate datasets of location probabilities, we assess the concept of non-sampling error with regard to the location probability distribution of each mobile device. Certainly, the model assumptions to compute these probabilities will impinge directly on the data. Details will depend on the input from the MNOs. For example, if antenna parameters are shared then we have the choice to use radio propagation models in the telecommunication literature (a variety of them depending on available data). If antenna parameters are not shared and only antenna ID or cell coverage are available, radio propagation models cannot be used and then location probabilities must be computed otherwise. Usual issues with traditional non-sampling errors are: • Coverage: for location probabilities, as explained above, this reduces to the nature of data (CDRs or signaling data). • Non-response:	Apart from the non-sampling errors accumulated in the location probabilities, for the case of aggregates the model connecting the dataset of mobile network data and the target population is based on the following basic assumption: the number of devices detected in the network can be considered as a binomial-distributed random variable over the total number of individuals in the total population with a given detection probability (that of being detected by the network at stake).

	A4. Unit non-response - rate (U)	The ratio of the number of units with no information or not usable information to the total number of in-scope (eligible) units, at a level of detail appropriate for a user report.		A priori, with signaling data the MNO can compute those devices with a failed connection to the network. For CDRs, non-response is more subtle since absence of communicating behavior cannot cleanly be considered non-response. Connection failure could possibly be reported by MNOs also in this case, but it depends on the access agreement.	Apart from the same consideration for the location probabilities as in the cell in the left, we see that unit-nonresponse is meaningless for the model connecting the dataset with the target population (except for their consideration in building this model).
	A5. Item non-response - rate (U)	The ratio of the in-scope (eligible) units that have not provided a particular item and the in-scope units that are expected to provide that particular item, at a level of detail appropriate for a user report.		This does not seem to be meaningful for mobile network data.	This does not seem to be meaningful for mobile network data.

				The penetration rates of mobile technology provide an insight about the proportion of target human population potentially detected in a telecommunication network. When working with a given MNO, the part of the target human population amounts to the penetration rate/market share of that MNO. According to these figures, only children, some elderly people, and singular cases (people in prison, extremely deprived people, homeless people, etc.) may be unreachable with this technology. The concept of over- and under-coverage is much related to the ultimate concept of target population under study (present population, inbound tourism, outbound tourism, domestic tourism, commuters, etc.).	
	Coverage	Divergence between the	Provide information on the frame and its sources and actions performed to gather the population impacting on coverage (e.g. webscraping). Provide an assessment,	Regarding intermediate datasets with location probabilities as output, the undercoverage is related both to the market share/penetration rate of the MNO(s) under analysis and those groups of society not	Apart from the over- and under-coverage present in the dataset with the location probabilities, when computing the posterior distribution for

S.13.3.1.1	A2. Overcoverage – rate (P)	The proportion of units accessible via the frame that do not belong to the target population.	Report A2, Overcoverage - rate	We do not have results yet. Our approach will be: (i) For overcoverage arising from multiplicity of devices we aim at analysing location probability distributions to estimate whether two given devices belong to the same individual or not (basically a classification algorithm). Density-based clustering algorithms techniques may be useful for this task (not yet investigated). (ii) For overcoverage arising from the detection of subscribers not in the target population we aim at using and/or developing detection algorithms implementing the official definition of statistical units.	Apart from the treatments at the location probabilities level, according the results of these treatments, the modelling exercise to compute the posterior distribution of the number of tourists will incorporate these results. We do not have results yet.

S.13.3.2	Measurement error	Measurement errors are errors that occur during data capture and cause recorded values of variables to be different from the true ones	The main sources of measurement error should be reported and assessed. Their description should be accompanied by any available analysis, otherwise by the producer's best knowledge. Where available and relevant describe: • identification and general assessment of the main sources of measurement error, including errors arising from data acquisition; • efforts made in questionnaire design and testing, information on interviewer training and other work on error prevention; • errors in measurement instruments (meters, satellites,...) • *results of assessments based on comparisons with external data, re-interviews or experiments;* • *results of indirect analysis, for example, of the editing phase; and* • actions taken to correct measurement errors	With mobile telecommunication technology, the line to define a measurement error is really thin. According to the technology there may be cases where the connection to the network could possibly provide misleading values for some of the raw telco variables (e.g. when an antenna is saturated and service is provided by another one, the cell ID may be misleading if no extra data about this situation is provided in the dataset –not usually the case). Whether to consider this a measurement error or not is subtle and subject to debate. Technology is working properly and no technological error is produced. However for statistical purposes this raw value is misleading. The computation of location probabilities we are proposing takes into account the dynamical behavior of mobile devices so that we hope this kind of situations may be accounted for. A potential editing and imputation strategy could be	Apart from the treatments at the location probability distributions level, the modelling exercise may be adapted to the presence of anomalous distributions in the datasets. We do not have a methodological proposal for this yet.

S.13.3.3	Nonresponse error	Nonresponse errors occur when the Big Data source fails to collect one or all the variables for units belonging to the domain covered by the source	Provide qualitative/quantitative assessments of unit nonresponse and highlight the units that are most subject to nonresponse Highlight the variables that are most subject to item nonresponse Provide a qualitative/quantitative assessments of the bias associated with nonresponse, comparing response rate for different sub-groups or distribution of auxiliary variables known for respondents and non-respondets (etc.) Provide a breakdown of nonrespondents according to cause for nonresponse mainly focusing on unit dependent cause and data collection tools cause. Define a stategy for reducing nonresponse during data collection and follow-up. Implement an estimator adjusted for nonresponse . *European level* Provide a qualitative/quantitative assessments of unit and item nonresponse across countries.	Nonresponse with mobile technology can only be understood as a system failure not providing service to a given mobile device, thus not generating network events. The reliability of this technology is in principle really high and this failure can be considered only as a special exception. The computation of location probability distributions we are using takes into consideration the dynamical behavior of mobile devices and thus, in principle, this lack of data could apparently be detected.	Once non-response is treated at the location probabilities level, it is meaningless to consider non-response for the computation of posterior distributions for the number of tourists. At most, the modelling exercise in the computation of this distribution must take into account the non-response detected at the device level.

S.13.3.3.1	A4. Unit nonresponse - rate (P)	The ratio of the number of units with no information or not usable information to the total number of in-scope (eligible) units, at a level of detail appropriate for a producer report.	Report A4: Unit nonresponse rate overall and at a level of detail appropriate for a producer report.	Not available (see above).	Not available (see above).

S.13.3.4	Processing error	The error in final data collection process results arising from the faulty implementation of correctly planned implementation methods, e.g., algorithms used to transform the data or extract information from raw data.	If processing errors are significant, identify the main issues regarding them. Present an analysis of processing errors, where available, otherwise a qualitative assessment. Report their extent, and impact on the outputs, of the most significant types of error. Include descriptions of linking and coding errors, if applicable. Where mistakes relating to programming or publishing have occurred, corrective measures taken as well as actions for avoiding them in the future should be reported. Example: For web data sources: Setting up a pipeline assures processing is comparable over time. Because texts were processed, the final results were highly affected by the various choices of text processing made.	No implementation has been executed so far, thus we cannot provide information at this moment.	No implementation has been executed so far, thus we cannot provide information at this moment.

S.13.3.5	Model assumption error	Error due to models used in the statistical production.	Describe process specific models, for example, as needed to define the target of estimation itself and models used for transformation of data into statistical data. Provide an assessment of the validity of each model. Descriptions of models used in treatment of specific sources of error should be presented in the section dealing with those errors. The assessment of the models used in treatment of specific sources of error should be presented in this section. Discuss the trade off between the need to use proper model that can change over time (accuracy) and the use a constant model in order to ensure comparability over time	The computation of location probabilities is extremely data-dependent. For example, if the access agreement considers sharing antenna parameters (such as power, azimuth, tilt, frequency, etc.), then radio-propagation models can be used to carry out the computation of these probabilities. If these parameters are not shared, then geospatial considerations (e.g. using Voronoi tessellations) must be used. The choice of model is expected to impact strongly on the accuracy of the outputs. Regarding the assessment of these choices, we do not have a methodological proposal yet since we are not producing point estimations, but probability distributions. WE need to find the most appropriate figure of merit for this purpose. As a key criterion in the choice of model, apart from data availability, we will consider robustness in the data in the sense that the results of the models do not violently change under an acceptable change of parameters (e.g. not using a radio-propagation model extremely sensitive to ever-changing conditions	Apart from the model assumptions for the location probabilities, the computation of the posterior distribution for the number of tourists also depends on a statistical model relating the total number of tourists, the number of devices detected in the network for these tourists, and the detection probabilities for tourists by the MNO. Auxiliary official data are to be used as benchmark in the model. Although not yet tested and implemented, usual model checking and model assessment techniques in Bayesian modelling (posterior predictive distributions, especially) are planned to be used to assess this sort of errors. Also, the detection algorithm to filter tourists from the rest of subscribers in the data sets is subjected to model assumptions in the translation of the official definition of statistical units under analysis (domestic tourist). Again, the robustness of all model parameters will be assessed

13.3.5.A Additional for big data				The assessment of model assumptions is extremely dependent on the selected model, which, among other things, also depend on data availability. Exact details cannot be provided (not even known yet) until a first test is conducted.	The assessment of model assumptions is extremely dependent on the selected model, which, among other things, also depend on data availability. Exact details cannot be provided (not even known yet) until a first test is conducted.

S.14 Timeliness and Punctuality

SIMS	Concept Name	Defintion	Guidelines	Answer (I)	Answer (O)
S.14	Timeliness and punctuality	(Defined by its sub-concepts)	(Information relating to this concept is provided by reporting on its sub-concepts.)	Both timeliness and punctuality depend heavily on the access agreement and the technological infrastructure in the MNO to retrieve and preprocess the raw telco data. Besides, the whole processing time has not been tested yet for a realistic dataset (say at a national level).	Both timeliness and punctuality depend heavily on the access agreement and the technological infrastructure in the MNO to retrieve and preprocess the raw telco data. Besides, the whole processing time has not been tested yet for a realistic dataset (say at a national level).
S.14.1	Timeliness	Length of time between data availability and the event or phenomenon the data describe.	Outline the reasons for the time lag. Outline efforts to reduce time lag in future.	For research purposes we have planned to access a long period of data, out of which we can produce location probabilities at the shortest time lag possible. For production conditions, no agreement is still envisaged.	For research purposes we have planned to access a long period of data, out of which we can produce location probabilities and the subsequent model construction at the shortest time lag possible. For production conditions, no agreement is still envisaged.

S.15 Coherence and Comparability

SIMS	Concept Name	Defintion	Guidelines	Answer (I)	Answer (O)
S.15	Coherence and Comparability	Adequacy of statistics to be reliably combined in different ways and for various uses and the extent to which differences between statistics can be attributed to differences between the true values of the statistical characteristics.	(Information relating to this concept is provided by reporting on its sub-concepts.)	A priori the underlying technological stratum in this industry introduces a layer of homogeneity thus potentially fostering the comparability of location probabilities. However, we expect some issues when datasets from different MNOs and even the same MNO in different countries are to be combined.	A priori the underlying technological stratum in this industry introduces a layer of homogeneity thus potentially fostering the comparability of location probabilities. However, we expect some issues when datasets from different MNOs and even the same MNO in different countries are to be combined. Furthermore, in the case of aggregates using official data as auxiliary benchmarking information, the differences in these statistical figures (accuracy, etc) may impinge on the comparability among countries.

S.15.1	Comparability – geographical	The extent to which statistics are comparable between geographical areas.	Describe any problems of comparability between regions of the country. The reasons for the problems should be described and as well an assessment (preferably quantitative) of the possible effect on the output values. Give information on discrepancies from the ESS/ international concepts, definitions, with reference to other chapters for more details.	Under the assumption that a given MNO is using the same technology across the national territory, we do not expect comparability issues among different regions inside a country. We do expect potential issues when using different technologies (thus with data from different MNOs and different countries).	Under the assumption that a given MNO is using the same technology across the national territory, we do not expect comparability issues among different regions inside a country. We do expect potential issues when using different technologies (thus with data from different MNOs and different countries). Furthermore, in the case of aggregates using official data as auxiliary benchmarking information, the differences in these statistical figures (accuracy, etc) may impinge on the comparability among countries.

S.15.2	Comparability – over time	The extent to which statistics are comparable or reconcilable over time.	Provide information on possible limitations in the use of data for comparisons over time. Distinguish three broad possibilities: 1. There have been no changes, in which case this should be reported. 2. There have been some changes but not enough to warrant the designation of a break in series. 3. There have been sufficient changes to warrant the designation of a break in series.	The underlying telecommunication technology is ever-changing and evolving. Accuracy in the location probabilities may be improved in the future, although this will depend on these technological changes and the corresponding access agreement. To preserve comparability over time the Reference Methodological Framework is to be put into place in which the implementation through the computation of the location probabilities will hopefully reduce the impact of the technological changes. Figures of merits for the location probabilities (yet to be identified) will report a change in the source data (e.g. by noting that accuracy has noticeably changed over time using the same computational procedure).	The underlying telecommunication technology is ever-changing and evolving. Accuracy in the location probabilities may be improved in the future, although this will depend on these technological changes and the corresponding access agreement. To preserve comparability over time the Reference Methodological Framework is to be put into place in which the implementation through the computation of the location probabilities will hopefully reduce the impact of the technological changes.

S.15.3	Coherence-cross domain	The extent to which statistics are reconcilable with those obtained through other data sources or statistical domains.	An analysis of incoherence should be provided, where this is an issue of importance.	No traditional counterpart for location probabilities is available.	It is expected to find some incoherence in the results coming from different data sources (especially when time and spatial breakdowns are severely different). To the extent feasible a comparative analysis will be conducted to understand these differences.
S.15.4	Coherence – internal	The extent to which statistics are consistent within a given data set.	Each set of outputs should be internally consistent. If statistical outputs within the data set in question are not consistent, any resulting lack of coherence in the output of the statistical process itself should be stated as well as a brief explanation of the reasons for publishing such results.	Since data are generated through the same technology and the same methodological proposals are to be appied throught the whole dataset, no internal inconsistency is expected.	Since data are generated through the same technology and the same methodological proposals are to be appied throught the whole dataset, no internal inconsistency is expected.

S.15.A.1	Coherence - with existing information/ Official Statistics	The extent to which information / statistical output from new data sources is consistent with information /statistical output from traditional data sources.	Provide information if it is meaningful to compare the information gained from new data sources with information from traditional data sources and if so, how consistent the information /statistical output gained from new data sources is with the one from traditional data sources.	No traditional counterpart for location probabilities is available in Official Statistics..	It is expected to find some incoherence in the results coming from traditional data sources (especially when time and spatial breakdowns are severely different). To the extent feasible a comparative analysis will be conducted to understand these differences.
S.15.A.2	Comparability - between information from several distinct new data sources	The extent to which information from several distinct new data sources is comparable among one another.	If you have raw data from several distinct new data sources, provide information how comparable the respective raw data sets and the information derived from them are among one other. Examples: MNO data from several mobile operators, smart meter data from several electricity providers	No access to several MNOs. Anyway, access to multiple MNOs should drive us to combine them for the production of a given statistics, not to a multiplicity of them. We woud find more interesting to compare these location probabilities from MNO data with location probabilties from other new digital data (e.g. from geolocated payment points in financial transaction data).	No access to several MNOs. Anyway, access to multiple MNOs should drive us to combine them for the production of a given statistics, not to a multiplicity of them

S.16 Cost and Burden

SIMS	Concept Name	Defintion	Guidelines	Answer (I)	Answer (O)
S.16	Cost and burden	Cost associated with the collection and production of a statistical product and burden on respondents.	***Cost*** Provide annual operational costs of the process, with breakdown by major cost component. Describe recent efforts to improve efficiency and comment on the extent to which information and communication technology is used. *European level* Describe recent initiatives and efforts to improve efficiency at the European level.	Cost is a strong issue in using MNO data for Official Statistics. Raw telco data are not intended for statistical purposes, thus they must be preprocessed prior to any statistical treatment. All these operational procedures are new, and we lack knowledge about their real cost. Moreover, is this cost to be considered as a cost for the NSIs (cost in this epigraph S.16) or as a cost for the MNOs (burden in this epigraph S.16).	Apart from the comments for location probabilities, in the case of aggregates, it may be the case that processing takes place in MNOs' premises (model construction and estimation). That is, MNOs will execute part of the statistical process usually carried out by NSIs.

			Burden Provide an estimate of the respondent burden imposed by the process. Describe all the means taken to minimise burden. *European level* Describe recent initiatives and efforts to minimise burden at the European level.	In survey data, responding questionnaires are considered just a burden for the respondents even despite the effort to collect the information. This is understandable and can be assumed. For mobile network data it is not clear if this view (for this data source) should prevail. It could be the case for those MNOs with an existing technological infrastructure for the statistical exploitation of their data (although complains could be formulated), but in the case of those MNOs with no such an infrastructure whatsoever who is going to pay for such a technological deployment? Part of the work on the access agreements tries to tackle this issue. No definitive solution is foreseen in the short term.	

S.16.A Additional for big data	Potential savings in cost and burden	Description how the new data source might influence cost and burden in the future	Provide an overview how the new data source could be deployed in the future to save the NSIs cost and/or decrease the respondent burden. Provide a qualitative description of the additional efforts for the NSI and the data owners.	We do not have reliable knowledge about the cost for NSIs in producing location probabilities. The burden for MNOs amounts to data retrieving, data preprocessing, implementation statistical methodologies and applying these to the data.	We do not have reliable knowledge about the cost for NSIs in producing both location probabilities and aggregates. The burden for MNOs amounts to data retrieving, data preprocessing, implementation of statistical methodologies, and application of these to the data.

S.17 Data Revision

SIMS	Concept Name	Definition	Guidelines	Answer
S.17.1	Data revision – policy	Policy aimed at ensuring the transparency of disseminated data, whereby preliminary data are compiled that are later revised.	Describe the data revision policy applicable to data output from the statistical process being reported. In so far as they are relevant to the process being reported, summarise the general procedures for treatment of planned revisions, benchmark revisions, unplanned revisions, and revisions due to conceptual and/or methodological changes. *European level* Describe the data revision policy and procedures at European level.	*Not relevant for the track meeting*

S.18 Statistical Processing

SIMS	Concept Name	Definition	Guidelines	Answer (I)	Answer (O)
S.18	Statistical processing	(Defined by its sub-concepts)	(Information relating to this concept is provided by reporting on its sub-concepts.)	A statistical process is run from the raw telco data to the location probabilities.	A statistical process is run from the raw telco data to the final aggregates..

S.18.1	Source data	Characteristics and components of the raw statistical data used for compiling statistical aggregates.	Indicate if the data are based on a survey, an administrative data source, multiple data sources, big data source (machine generated, human sourced, process mediated), e.g., web data, and/or macro-aggregates. Refer to the accreditation document of the data source, if applies. In the event of multiple data sources or macro-aggregates, reference each source and indicate how they are combined. For each survey source, summarise the sample design, cross referencing the descriptions of the target and survey populations, presented in S.03.6. For each ~~administrative~~ data source, summarise the source, its primary purpose, and the most important data items acquired. Information in which form the	Data are mobile network data generated by the interaction between a telecommunication network and a mobile device. No accreditation document is available. So far, only data from a single MNO is considered due to limited access issues. In the future, access to all MNOs in a country is to be pursued. The data generation process constitutes sensitive information for the MNO due to the high level of competitiveness in the telco market. More work is needed in the agreement with MNOs to incorporate this information into the statistical process. Mobile network data are	Data are mobile network data generated by the interaction between a telecommunication network and a mobile device. No accreditation document is available. So far, only data from a single MNO is considered due to limited access issues. In the future, access to all MNOs in a country is to be pursued. The data generation process constitutes sensitive information for the MNO due to the high level of competitiveness in the telco market. More work is needed in the agreement with MNOs to incorporate this information into the statistical process. Mobile network data are basically composed of pseudonymised ID variables, geolocation variables, time variables, and other

S.18.2	Frequency of data collection	Frequency with which the source data are collected.	Indicate the frequency of data collection (e.g. monthly, quarterly, annually, or continuous).	For research purposes in this project, access will be granted once. For production purposes in the long term, no agreement has been reached.	For research purposes in this project, access will be granted once. For production purposes in the long term, no agreement has been reached.

S.18.3	Data collection	Systematic process of gathering data for official statistics.	*For each survey data source:* • describe the method(s) used to gather data from respondents; • annex or hyperlink the questionnaires(s). *For each administrative data source* • describe the acquisition process and how it was tested. *For all sources* • describe the types of checks applied at the time of data entry. *For big data sources* describe the methods used to collect the data; add hyperlink if it is web data or name of the API used to collect the data. *European level* Provide a summary of the commonalities and differences in the collection methods, questionnaires and checks used in different countries.	Data will be retrieved, preprocessed and processed in the MNO's premises. Data entry is meaningless. No detailed description of how data are gathered from the highly distributed information systems across the national territory is available.	Data will be retrieved, preprocessed and processed in the MNO's premises. Data entry is meaningless. No detailed description of how data are gathered from the highly distributed information systems across the national territory is available.

S.18.4	Data validation	Process of monitoring the results of data compilation and ensuring the quality of statistical results.	Describe the procedures for checking and validating the source data and how the results are monitored and used. Describe the procedures for validating the aggregate output data (statistics) after compilation, including checking coverage and response rates, and comparing with data for previous cycles and with expectations. List other output datasets to which the data relate and outline the procedures for identifying inconsistencies between the output data and these other datasets. Define the linkage method for big data sources and other data sources used for validation. *European level* Provide a summary of the commonalities and differences in the validation methods used by countries.	No checking and validating of source data are foreseen. The location probabilities will be investigated comparing with official population figures. No individual identification is possible, thus no linkage will be conducted. Only analysis at different aggregation levels and time breakdown will be conducted.	Location probabilities will be validated according to the description in the left. Aggregate results will be compared with traditional and official estimates.

S.18.5	Data compilation	Operations performed on data to derive new information according to a given set of rules.	If there is missing data, give detailed description of the methods used for imputation. For big data sources, e.g., web data, indicate the reason why data were not collected (technical issues etc.). Describe the procedures for imputation, the most common reasons for imputation and imputation rates within each of the main strata. Describe the likely impact of imputation. Describe the procedures for adjustment for non-response and the corrections to the design weights to account for differences in response rates. Describe the calculation of design weights, including calibration (if used). Describe the procedures for combining input data from different sources.	The computation of location probabilities is conducted in two steps: a) Retrieval and preprocessing of raw telco data preparing them for statistical purposes. b) Application of the model to compute location probabilities (based on a hidden Markov model).	The computation of posterior distributions for the number of tourists is conducted in steps: a) Computation of location probabilities according to the description in the left. b) Application of detection algorithms for tourists (possibly including usual environment, home/work, trips and related concepts). c) Computation of the distribution of the number of devices of tourists per territorial cell and time period of analysis. d) Construction and application of the hierarchical model to obtain the posterior distribution of the number of tourists per territorial cell and time period of analysis.

S.18.5.1	A7. Imputation – rate	The ratio of the number of replaced values to the total number of values for a given variable.	Provide values of indicator A7 Imputation – rate	No imputation is included in the methodological proposal.	No imputation is included in the methodological proposal.

Annex II: Example of Completed Quality Questionnaire "Innovative Companies Webscraping"

S.03 Statistical Presentation

SIMS	Concept Name	Defintion	Guidelines	Answer
S.03	Statistical presentation	Description of the statistical output.	(Information relating to this concept is provided by reporting on its sub-concepts.)	Dutch companies with 2 or more working persons (WP) that are technological innovative.
S.03.1	Data description	Main characteristics of the data set, referring to the statistical output.	Describe briefly the main characteristics of the data in an easily and quickly understandable manner, referring to the main variables. More detailed descriptions of the variables and how they were derived in S.03.4.	Number of Dutch companies (WP >= 2) that are technological innovative at the zipcode-4 level. The text on the webpage is used to derive that characteristic. Both websites written in Dutch or English (read any other language) are included.

S.03.4	Statistical concepts and definitions	Statistical characteristics of statistical observations.	Define and describe briefly the main statistical variables that have been observed or derived. Indicate their types. Note that any difference between these variables and the variables desired by users is a relevance issue and is discussed in S.12.	1) Technological innovation is determined by a model based on the text on the webpage of the company; result is yes or no. This variable was previously collected via the CIS survey and defined there. It is a derived variable. 2) The webpage of a company is obtained from a (beta-version of a) frame linking each Dutch companies to the URL of the main page of the website; this is text. It is a new variable. 3) Number of WP is obtained from the business register; it is an integer 4) The zipcode-4 is provided by the business register; it is a integer
S.3.5	Statistical unit	Entity for which information is sought and for which statistics are ultimately compiled.	Define the type of statistical unit about which data are available, e.g. enterprise, local unit, private household, person. If there is more than one type of unit, define each type.	The statistical unit studied is a company identified by its Chamber of Commerce number (to which the website is linked). This unit can be linked to higher aggregates via the Business register (such as, Business unit etc).

S.3.6	Statistical population	The total membership or population or "universe" of a defined class of people, objects or events.	Define the target population of the statistical units for which information is sought. Note that a difference between the target population and the population desired by users is a relevance issue and is discussed in S.12; and the difference between target population and the actual (frame) population is a coverage issue and is discussed in S13.3 If there is more than one type of population, define each type.	The target population is all companies in the Netherlands with 2 or more WP included in the Business register. The Chamber of Commerce number is used to identify each company. Companies with no website cannot be scraped. We found that nearly all innovative companies have a website (99.9%). This is less the case for non-innovative companies have less websites (~95%). In other words: The number of technological innovative companies without a website is determined to be 0.1% maximum (based on the CIS survey and additional research). For comparison with the CIS survey data the population of companies WP >= 10 and WP >= 2 are discerned.
S.3.7	Reference area	The country or geographic area to which the measured statistical phenomenon relates.	Describe the country, the regions, the districts, or the other geographical aggregates, to which the data refer. Identify any specific exclusions in the statistical data.	The data refers to the Netherlands with zipcode-4 level as the most detailed geographical region. This data can be aggregated to municipality, COROP, Provinces, and country level.

S.3.8	Time coverage	The length of time for which data are available.	State the time period(s) covered by the data, e.g. first quarter 2018, or quarters 2015-2018, or year 2018, or years 1985-2018. Note that any issues concerning comparability over time are discussed in S.15.	The websites were scraped in July 2019. Webscraped data from 2017 onwards are available for a set of 6,000 companies (this is the training and testset).
Additional for big data				Stability of the model over time was found to be an issue. Dealing with this required data collected over a longer period of time for a lot of companies. The use of the text of a website to determine if a company is technological innovative is a derived (secondary) use and, hence, may be less accurate than directly contacting a company (primary data collection).

S.04 Unit of Measure

SIMS	Concept Name	Definition	Guidelines	Answer
S.04	Unit of measure	The unit in which the variables of the statistical output are measured.	The statistical data usually involves several units of measure depending upon the variables. Examples: • Country in which a SIM card is located at a certain time, • position of a ship at a certain time, • consumption of electricity in Watt/Kilowatt in a certain time span, • Classifying the negative or positive sentiments of a text input on a -1/1 scale	Companies at the Chamber of Commerce level with 2 or more WP. The text on the main page of the website of a company is classified. The website needs to contain 10 or more words to enable classification. The result of the classification is the finding if a company is technological innovative or not (0/1 scale).
Additional for big data				A census-based approach was followed. This means that all websites of all Dutch companies (with 2 or more WP) in the Business register were attempted to be scraped.

S.05 Reference Period

SIMS	Concept Name	Defintion	Guidelines	Answer

S.05	Reference period	The period of time or point in time to which the measured observation is intended to refer.	The value of a variable refers to a specific time period (for example, the last week of a month, a month, a fiscal year, a calendar year, or several calendar years), or to a point in time (for example, a specific day, or the last day of a month). The variables in a dataset may refer to more than one reference period. All reference periods should be stated Note that the difference, if any, between the target reference period(s) and the actual reference period(s) is an accuracy issue and should be discussed in S.13.3. Note that if frame population does not include all the units in the target population for the specified reference period, this is a coverage issue and should be discussed in S.13.3.	The period during which the websites were scraped was July 2019.
Additional for big data				Websites were attempted to be scraped multiple times to assure the website was active (really exists). The status code of the website was used to determine this.

S.06 Institutional Mandate

SIMS	**Concept Name**	**Definition**	**Guidelines**	**Answer**

S.06	Institutional mandate	Set of rules or other formal set of instructions assigning responsibility as well as the authority to an organisation for the collection, processing, and dissemination of statistics.	(Information relating to this concept is provided by reporting on its sub-concepts.)	CIS survey is held once every 2-years in the Netherlands. Compiling CIS data is voluntary to EU member states.
S.06.1	Legal acts and other agreements	Legal acts or other formal or informal agreements that assign responsibility as well as the authority to an agency for the collection, processing, and dissemination of statistics.	Describe the (legal) agreement and other forms of cooperation with the data owner which allows the NSI access to the data source. Describe which forms of reciprocity (not necessarily financial) does the NSI offer to the data source?	CBS-law (https://wetten.overheid.nl/BWBR0015926/2019-01-01) For webscraping, CBS has created an internal -legally verified- document (written in Dutch) with rules and instructions. Most important remarks are that i) any CBS webscraping tool needs to identify itself as a CBSbot, the ii) Robots Exclusion Protocol of a website has to be respected and that iii) websites need to be scraped from within the IT-infrastructure of CBS and any data may not be exported.

S.06.A	Data access and data transmission	Arrangements or procedures for data access and data transmission	Describe the arrangements, procedures or agreements for data access and data transmission. In particular, describe • Modes of data access (full access to raw data, access to pre-processed data, on-premise, off-premise) • In case of access to pre-processed data: transparency about the technological processes applied to the pre-processed data • Time and method of transmission • Time horizon of the cooperation - Is a long term access to the data guaranteed?	

S.07 Confidentiality

SIMS	Concept Name	Definition	Guidelines	Answer
S.07	Confidentiality	A property of data indicating the extent to which their unauthorised disclosure could be prejudicial or harmful to the interest of the source or other relevant parties.	(Information relating to this concept is provided by reporting on its sub-concepts.)	The classification of a company as technological innovative or not is –in principle- not considered a privacy concerned identifiable variable. In addition, aggregates at the zipcode-4 level are the most detailed results produced.
S.07.1	Confidentiality – policy	Legislative measures or other formal procedures which prevent unauthorised disclosure of data that identify a person or economic entity either directly or indirectly.	Describe all European or national legislation, or other formal requirements, that relate to confidentiality. Describe relevant policy (if any). Note that the existence of legislation and/or policy provides some assurance that methods necessary to assure confidentiality have been applied to the data. *European level* Summarise the commonalties and differences in national approaches to confidentiality policy..	Not needed in the pilot track meeting.

S.07.2	Confidentiality - data treatment	Rules applied for treating the datasets to ensure statistical confidentiality and prevent unauthorised disclosure.	For aggregate outputs • Provide the rules that define a *confidential cell*. • Describe the procedures for detecting confidential cells, including checking for residual disclosure. • Describe the procedures for eliminating confidential cells, for example by controlled rounding, cell suppression, or cell aggregation. For micro-level outputs: • Describe the procedures that are used in protecting confidentiality.	No issues because i) the output is limited to aggregates at the zipcode-4 level and ii) being classified as technological innovative or not is not considered privacy sensitive variable.
S.07.A Additional for big data	Privacy	How privacy sensitive is the information coming from external data holders?	State which treatments are prescribed to satisfy privacy concerns	The text is extracted from websites that can be viewed by anyone in the world.

S.08 Release Policy

SIMS	**Concept Name**	**Defintion**	**Guidelines**	**Answer**
S.08	Release policy	Rules for disseminating statistical data to all interested parties.	(Information relating to this concept is provided by reporting on its sub-concepts.)	Only aggregated results are released (Currently only as a beta-product).

S.08.A Additional for big data	Release policy for Experimental Statistics	Rules for dissemination of experimental data or experimental statistical products.	State if there exists a publicly available policy for the dissemination of experimental statistics and if there exists a designated area at your NSI's homepage.	*This is an organisation-related topic and is kept for the deliverable, but not asked in the quality-related questionnaire for the track meeting in December.*

S.09 Frequency of Dissemination

SIMS	Concept Name	Defintion	Guidelines	Answer
S.09	Frequency of dissemination	The time interval at which the statistics are disseminated over a given time period.	State the frequency with which the data are disseminated, e.g. monthly, quarterly, yearly. The frequency can also be expressed by using a code from the harmonised ESS code list so long as this is considered to be easily understandable by users.	*Not relevant for the track meeting*
Additional for big data				Findings for the whole target population have been produced once. The findings are available as a beta-product on the CBS website.

S.10 Accessibility and Clarity

SIMS	Concept Name	Defintion	Guidelines	Answer
S.10	Accessibility and clarity	The conditions and modalities by which users can access, use and interpret data.	(Information relating to this concept is provided by reporting on its sub-concepts.)	Experimental statistics are available as aggregates as a beta-product (and not at the individual level).

S.10.6	Documentation on methodology	Descriptive text and references to methodological documents available.	List national reference metadata files, methodological papers, summary documents and handbooks relevant to the statistical process. For each item provide the title, publisher, year and link to on-line version (if any).	The method has been described in a scientific paper that is currently under review. There is an internal CBS version of this paper available.
S.10.7	Quality documentation	Documentation on procedures applied for quality management and quality assessment.	List relevant quality related documents, for example, other quality reports, studies. Cross reference to descriptions of quality procedures in other chapters, especially S.13..	The webscraped findings for large companies (WP >= 10) are comparable with the most recent findings of the CIS survey for that part of the target population (also WP >= 10).

S.11 Quality Management

SIMS	**Concept Name**	**Definition**	**Guidelines**	**Answer**
S.11	Quality management	Systems and frameworks in place within an organisation to manage the quality of statistical products and processes.	(Information relating to this concept is provided by reporting on its sub-concepts.)	There is a centrally available (beta) version of the frame linking companies and their URLs. Hence, all webscrape-projects at CBS use the same URL for a particular company. Users of this service provide feedback on their experience with the newly created frame (so any errors can be corrected).

S.11.1	Quality assurance	All systematic activities implemented that can be demonstrated to provide confidence that the processes will fulfil the requirements for the statistical output.	Describe the quality assurance procedures specifically applied to the statistical process for which the report is being prepared, for example agreements with the big data providers, benchmarking, assessments, and use of best practices. Include descriptions of all forms of quality assessment procedures (self-assessment, peer review, compliance monitoring, audit) and when they most recently took place. Summarise the results of the most recent quality assessments and cross reference to the chapters in the report where the results are presented in more detail. Describe any ongoing or planned improvements in quality assurance procedures.	*Only quality assurance procedures which affect the new data sources are relevant for the track meeting.* Websites are visited regularly to maximize the chance that they are active and, hence, the text can be scraped. Websites that cannot be scraped after multiple attempts are considered inactive (and are excluded from the target population; this is more likely to occur for small companies).

S.11.2	Quality assessment	Overall assessment of data quality, based on standard quality criteria.	Summarise the results of the most recent quality assessments and cross reference to the chapters in the report where the results are presented in more detail.	*Only quality assurance procedures which affect the new data sources are relevant for the track meeting.* The findings of the number of innovative companies with WP >= 10 of the CIS survey could be replicated with the webscrape-based approach (both found around 19.500 companies and had overlapping confidence intervals).
Additional for big data				As a sanity check, we checked the words included in the (logistic regression) model with the largest positive and negative coefficients. These words were logical/explainable related in a positive and negative way; it made sense why they were included in the model.

S.12 Relevance

SIMS	Concept Name	Defintion	Guidelines	Answer
S.12	Relevance	The degree to which statistical information meet current and potential needs of the users.	(Information relating to this concept is provided by reporting on its sub-concepts.)	The approach developed produced similar findings for large companies but –in addition- also enables to: i) estimate the number of small (WP >= 2 & WP < 10) technological innovative companies in the Netherlands; such as Startups. ii) produce maps at a much more detailed level; e.g. zipcode-4 and municipalities. iii) at a much higher frequency (more than once per 2 years)

S.12.1	User needs	Description of users and their respective needs with respect to the statistical data.	Provide: • a classification of users, also indicating their relative importance; • an indication of the uses for which users want the statistical outputs; • an assessment of the key outputs desired by different categories of users and any shortcomings in outputs for important users; • information on unmet user needs and any plans to satisfy them in the future; and • details regarding those quality components which do not meet user requirements.	*Not relevant for the track meeting*

S.12.3	Completeness	The extent to which all statistics that are needed are available.	Provide qualitative information on the extent to which content requirements in relevant legislation, regulations and guidelines are met. Provide information on the extent to which user needs related to content are satisfied. Provide values of indicator R1 Data completeness rate, for each required data item for each relevant regulation/ guideline at producer/user level of detail as appropriate. In the case where the indicator refers to data sent to Eurostat, this indicator can be compiled by Eurostat. *European level* Summarise across countries the extent to which ESS requirements for data items are met	*Not relevant for the track meeting*

S.12.A Additional for big data	Added Value through new data source	The potential added value of a new data source to an existing statistical product.	Describe if and how the usage of a new data source provides an added value to an already existing statistical product. E.g., this could be more detailed data on particular subgroups, or information on grid level instead of district level or the potential replacement of questions of a survey through information of the new data source.	More detailed data, the CIS survey maximally enables an estimation of WP >= 10 companies at the provincial level based on a sample. The new approach enables estimation of ALL companies with WP >= 2 at the zipcode-4 level (and higher aggregates). The data can also be produced at a much higher frequency. Be aware that the use of text means that technological innovation is a derived use.

S.13 Accuracy and Reliability

SIMS	**Concept Name**	**Definition**	**Guidelines**	**Answer**
S.13	Accuracy and reliability	Closeness of computations or estimates to the exact or true values that the statistics were intended to measure. Reliability of the data, defined as the closeness of the initial estimated value to the subsequent estimated value.	(Information relating to accuracy is provided by reporting on S.13 sub-concepts. Information on Reliability is reported in S.17 Data Revision).	CIS survey WP >= 10 number, 19,916 ± 680 Web based method WP >= 10 number, 19,276 ± 23

S.13.1	Overall accuracy	Assessment of accuracy, linked to a certain data set or domain, which is summarising the various components.	Describe the main sources of random and systematic errors in the statistical outputs and provide a summary assessment of all errors with special focus on the impact on key estimates. The bias assessment can be in quantitative or qualitative terms, or both, and may be expressed as bias risk. It should reflect the producer's best current understanding (sign and order of magnitude) and include actions taken to reduce bias. *European level* Provide a summary picture of accuracy across countries. The emphasis placed on various types of errors should depend upon the error profile of the respective process. For repetitive processes, describe how accuracy is developing over time and what efforts are underway to improve accuracy from an ESS perspective.	The main sources of error affect bias or variance. Bias is the result of i) model classification results (FP and FN are not identical), ii) websites containing less than 10 words, and iii) innovative companies without a website. These all negatively affect the overall estimate. These effects are corrected for in 3 subsequent steps. Variance is affected by: i) specific model classification results, ii) specific model classification error (FP and FN ratio) and iii) exact number of websites included and classified. The variance is determined in two stages (variance resulting from the model and variance resulting from the number of websites classified) As stated above the estimate for the whole of the country were similar. At the provincial level the WP >= 10 companies numbers were also compared; this is the most detailed level of the CIS survey based results. These differed for only 2 provinces; the other 10 are ordered exactly the same. This could be due to time differences between both measurements.

Additional for big data				A model was created a 1000-times and in each case the results of the confusion matrix on the test set were used to obtain model specific bias and variance results.

S.13.2	Sampling error	That part of the difference between a population value and an estimate thereof, derived from a random sample, which is due to the fact that only a subset of the population is enumerated.	State whether sampling error is relevant. If probability sampling is used: • for user reports, provide the range of variation of the A1 indicator among key variables at user report level of detail; • for producer reports, provide the range of variation of the A1 indicator among key variables at producer report level of detail; • indicate the impact of sampling error on the overall accuracy of the results; • state how the calculation of sampling error is affected by imputation for nonresponse, misclassifications and other sources of uncertainty, such as outlier treatment. If non-probability sampling is used, provide an assessment of representativity and risk of sampling bias. <u>*ESS level*</u> If probability sampling is used: • present sampling errors for key estimates across countries; • indicate which country to country differences are significant and which are not; • for a repetitive survey, describe at least broadly the trends in sampling error over time • provide sampling errors for ESS level estimates<u>.</u>	*Not relevant since no sampling with new data sources*

S.13.3	Non-sampling error	Error in estimates which cannot be attributed to sampling fluctuations	Summarise the most important aspects of coverage, measurement, non-response, processing and model assumption errors. Discuss the corresponding bias risks and actions undertaken to reduce them.	Some websites only contain a few words. Websites with < 10 words cannot be classified by the model. We assume that these websites have the same ratio innovative/non-innovative as those with >= 10 words (there are no indications that this assumption is incorrect). Not all technological companies have a website. We have determined that this is at maximum 0.1%. We correct for this. The model developed has an accuracy of 0.88%. The FP and FN are not identical, resulting in a (slightly) biased estimate.
	A4. Unit non-response - rate (U)	The ratio of the number of units with no information or not usable information to the total number of in-scope (eligible) units, at a level of detail appropriate for a user report.		This concerns websites with too few words (< 10). The ratio is 0.027. Websites that could not be scraped were excluded from the frame.

	A5. Item non-response - rate (U)	The ratio of the in-scope (eligible) units that have not provided a particular item and the in-scope units that are expected to provide that particular item, at a level of detail appropriate for a user report.		Not relevant (see above)

S.13.3.1	Coverage error (P)	Divergence between the population of the Big Data source and the target population.	Provide information on the frame and its sources and actions performed to gather the population impacting on coverage (e.g. webscraping). Provide an assessment, whenever possible quantitative, of overcoverage and undercoverage, including an evaluation of the bias risks associated with the latter. Describe actions taken for reduction of undercoverage and associated bias risks	Only companies with a website that contain 10 or more words after processing can be classified. The estimate was bias corrected for that (See 13.3). From the CIS survey and from a detailed study on small innovative companies, we have determined the number of technological innovative companies without a website. This number was estimated to be maximally 1 in a 1000 (0.1%). All input on the relation between a company and its website available at CBS was used to create the (beta) version of the linking frame. This was done to (???)
S.13.3.1.1	A2. Overcoverage – rate (P)	The proportion of units accessible via the frame that do not belong to the target population.	Report A2, Overcoverage - rate	Only for companies included in the Business register, websites were scraped. Since the target population is companies at the Chamber of Commerce level, there is no overcoverage.

S.13.3.2	Measurement error	Measurement errors are errors that occur during data capture and cause recorded values of variables to be different from the true ones	The main sources of measurement error should be reported and assessed. Their description should be accompanied by any available analysis, otherwise by the producer's best knowledge. Where available and relevant describe: • identification and general assessment of the main sources of measurement error, including errors arising from data acquisition; • efforts made in questionnaire design and testing, information on interviewer training and other work on error prevention; • errors in measurement instruments (meters, satellites,...) • *results of assessments based on comparisons with external data, re-interviews or experiments;* • *results of indirect analysis, for example, of the editing phase; and* • actions taken to correct measurement errors.	Sometimes website cannot be scraped; for instance because of maintenance. Therefore multiple attempts on various days (max 4) were made to scrape as much websites as possible. Non existing websites were excluded from the target population. Various ways to scrape and extract the text from websites were compared. The approach that provided the most reproducible and reliable results (based on the accuracy of the model developed) was used in the end. Reproducibility of the number of technological innovative companies (WP >= 10) from the CIS survey confirmed the validity of the webbased approach developed. External validation of the webbased model was tested by applying the model on the websites of startups; these are small innovative companies. Of these websites 92% were classified as technological innovative. The method developed suffered from 'concept drift'; a change in the reliability of the model-based results over time. This was solved by developing a model on a large number of webscraped data as possible. This solution was inspired by streaming data studies and we found that it increased the number of relevant (???)

S.13.3.3	Nonresponse error	Nonresponse errors occur when the Big Data source fails to collect one or all the variables for units belonging to the domain covered by the source	Provide qualitative/quantitative assessments of unit nonresponse and highlight the units that are most subject to nonresponse Highlight the variables that are most subject to item nonresponse Provide a qualitative/quantitative assessments of the bias associated with nonresponse, comparing response rate for different sub-groups or distribution of auxiliary variables known for respondents and non-respondets (etc.) Provide a breakdown of nonrespondents according to cause for nonresponse mainly focusing on unit dependent cause and data collection tools cause. Define a stategy for reducing nonresponse during data collection and follow-up. Implement an estimator adjusted for nonresponse . *European level* Provide a qualitative/quantitative assessments of unit and item nonresponse across countries.	Websites that existed but provided less than 10 words after processing are considered nonresponse. Websites with < 10 words cannot be classified by the model. The words of websites for which this was the case were studied to check if any groups could be discerned that were different. This was not the case. It was therefore assumed that websites < 10 words have the same ratio of innovative/non-innovative as the websites with >=10 words. This was done to correct for 1 of the causes of bias. Websites that could not be scraped after multiple attempts were excluded from the target population as these websites were no longer active (did no longer exists).

S.13.3.3.1	A4. Unit nonresponse - rate (P)	The ratio of the number of units with no information or not usable information to the total number of in-scope (eligible) units, at a level of detail appropriate for a producer report.	Report A4: Unit nonresponse rate overall and at a level of detail appropriate for a producer report.	This number has been reported above as 0.027

S.13.3.4	Processing error	The error in final data collection process results arising from the faulty implementation of correctly planned implementation methods, e.g., algorithms used to transform the data or extract information from raw data.	If processing errors are significant, identify the main issues regarding them. Present an analysis of processing errors, where available, otherwise a qualitative assessment. Report their extent, and impact on the outputs, of the most significant types of error. Include descriptions of linking and coding errors, if applicable. Where mistakes relating to programming or publishing have occurred, corrective measures taken as well as actions for avoiding them in the future should be reported. Example: For web data sources: Setting up a pipeline assures processing is comparable over time. Because texts were processed, the final results were highly affected by the various choices of text processing made.	Since the detection of technological innovation is a derived use (it is not directly available on the website) processing is needed to extract the information from the words on the websites. This is a two-stage process; i) create a document term matrix and ii) build a logistics regression model. Since there is no intermediate result only the final accuracy can be reported (is 88%). The effect of various processing steps were compared on a training and testset. The process steps resulting in the highest classification accuracy were selected. Since the goal was to extract the maximum amount of information from the text available, the goal was to minimize the processing error (and maximize the accuracy of the model). No exact value can be assigned to the processing error.

S.13.3.5	Model assumption error	Error due to models used in the statistical production.	Describe process specific models, for example, as needed to define the target of estimation itself and models used for transformation of data into statistical data. Provide an assessment of the validity of each model. Descriptions of models used in treatment of specific sources of error should be presented in the section dealing with those errors. The assessment of the models used in treatment of specific sources of error should be presented in this section. Discuss the trade off between the need to use proper model that can change over time (accuracy) and the use a constant model in order to ensure comparability over time	The error resulting from applying the model. Overall accuracy is 88%. Because the FP and FN are not identical, the estimate is biased; in the innovative company case the FN is higher than the FP. This is corrected for by a method described in Meertens, Q.A., C.G.H. Diks, H.J. van den Herik, and F.W. Takes. 2019a. "A Bayesian Approach for Accurate Classification-Based Aggregates." Proceedings of the 2019 SIAM International Conference on Data Mining, May 2-4, 2019. 306 – 314. Calgary, Canada. Doi: https://doi.org/10.1137/1.9781611975673.35. Issues with model stability ('concept drift') forced us to focus on the development of a model that was stable over a long(er) period of time. This eventually resulted in the model that was 88% accurate and stable (by checking it on data scraped at various points in time). Please realize that the first model was 93% accurate, but started suffering from 'concept drift' after 6 months.

13.3.5.A Additional for big data				It's difficult to think of errors for this Big Data approach when describing them with words used in the context of survey data. Non-response is an example of that. I think in this example that non-response is a website that exists but contains too few words (< 10) to be classified.

S.14 Timeliness and Punctuality

SIMS	Concept Name	Defintion	Guidelines	Answer
S.14	Timeliness and punctuality	(Defined by its sub-concepts)	(Information relating to this concept is provided by reporting on its sub-concepts.)	Data can be scraped when needed. Because websites are checked multiple times and huge amounts of websites could be scraped (in this case 825,000) the entire scraping might take some time (here, around 2 weeks).
S.14.1	Timeliness	Length of time between data availability and the event or phenomenon the data describe.	Outline the reasons for the time lag. Outline efforts to reduce time lag in future.	Data can be scraped when needed. When huge numbers of websites have to be scraped this might take some time and may result in some delay. In our case the maximum time for scraping was 2 weeks. Indicating that, in principle, this statistics could be produced every 2 weeks.

S.15 Coherence and Comparability

SIMS	Concept Name	Defintion	Guidelines	Answer
S.15	Coherence and Comparability	Adequacy of statistics to be reliably combined in different ways and for various uses and the extent to which differences between statistics can be attributed to differences between the true values of the statistical characteristics.	(Information relating to this concept is provided by reporting on its sub-concepts.)	Since the websites are linked to the units in the business registers, this issue should pose no problems. Regarding the comparability of the findings with the webbased method for companies with different WP's: it was found that: WP >=10 and WP >=2 produced comparable findings. But companies with WP >= 0.1 and WP <= 1 (i.e. self-employed) behaved somewhat different regarding the words they used on their website. This and privacy concerns were the main reasons to exclude WP <= 1 from the study.
S.15.1	Comparability – geographical	The extent to which statistics are comparable between geographical areas.	Describe any problems of comparability between regions of the country. The reasons for the problems should be described and as well an assessment (preferably quantitative) of the possible effect on the output values. Give information on discrepancies from the ESS/ international concepts, definitions, with reference to other chapters for more details.	It is assumed that companies behave themselves identical regarding the words they use on their website for the various regions in the Netherlands. Both Dutch and English words were included in the model.

S.15.2	Comparability – over time	The extent to which statistics are comparable or reconcilable over time.	Provide information on possible limitations in the use of data for comparisons over time. Distinguish three broad possibilities: 1. There have been no changes, in which case this should be reported. 2. There have been some changes but not enough to warrant the designation of a break in series. 3. There have been sufficient changes to warrant the designation of a break in series.	It is expected that companies may use different words indicative for innovation on their website over time. For this reason, the model needs to be checked regularly (we propose every 6 months). This is done by rescraping the websites of the companies included in the training and testset. The same approach is used to check for 'concept drift'.
S.15.3	Coherence- cross domain	The extent to which statistics are reconcilable with those obtained through other data sources or statistical domains.	An analysis of incoherence should be provided, where this is an issue of importance.	The CIS survey data can be reproduced.
S.15.4	Coherence – internal	The extent to which statistics are consistent within a given data set.	Each set of outputs should be internally consistent. If statistical outputs within the data set in question are not consistent, any resulting lack of coherence in the output of the statistical process itself should be stated as well as a brief explanation of the reasons for publishing such results.	Only bias correction (and rounding of the estimates) effect the total number of innovative companies when comparing the results at various regions

S.15.A.1	Coherence - with existing information/ Official Statistics	The extent to which information / statistical output from new data sources is consistent with information /statistical output from traditional data sources.	Provide information if it is meaningful to compare the information gained from new data sources with information from traditional data sources and if so, how consistent the information /statistical output gained from new data sources is with the one from traditional data sources.	Findings for the total number of technological innovative companies (WP >= 10) are similar. At provincial level 2 of the 12 finings clearly differ.
S.15.A.2	Comparability - between information from several distinct new data sources	The extent to which information from several distinct new data sources is comparable among one another.	If you have raw data from several distinct new data sources, provide information how comparable the respective raw data sets and the information derived from them are among one other. Examples: MNO data from several mobile operators, smart meter data from several electricity providers	Not relevant for this study

S.16 Cost and Burden

SIMS	Concept Name	Defintion	Guidelines	Answer
S.16.A Additional for big data	Potential savings in cost and burden	Description how the new data source might influence cost and burden in the future	Provide an overview how the new data source could be deployed in the future to save the NSIs cost and/or decrease the respondent burden. Provide a qualitative description of the additional efforts for the NSI and the data owners.	Part of the data collected by the CIS survey can be replaced by the webbased approach. In addition, the findings for the entire population (WP >= 10) and of companies with WP >= 2 can be completely determined.

S.17 Data Revision

SIMS	Concept Name	Definition	Guidelines	Answer
S.17.1	Data revision – policy	Policy aimed at ensuring the transparency of disseminated data, whereby preliminary data are compiled that are later revised.	Describe the data revision policy applicable to data output from the statistical process being reported. In so far as they are relevant to the process being reported, summarise the general procedures for treatment of planned revisions, benchmark revisions, unplanned revisions, and revisions due to conceptual and/or methodological changes. *European level* Describe the data revision policy and procedures at European level.	N*ot relevant for the track meeting*

S.18 Statistical Processing

SIMS	Concept Name	Definition	Guidelines	Answer
S.18	Statistical processing	(Defined by its sub-concepts)	(Information relating to this concept is provided by reporting on its sub-concepts.)	Webpages are processed

S.18.1	Source data	Characteristics and components of the raw statistical data used for compiling statistical aggregates.	Indicate if the data are based on a survey, an administrative data source, multiple data sources, big data source (machine generated, human sourced, process mediated), e.g., web data, and/or macro-aggregates. Refer to the accreditation document of the data source, if applies. In the event of multiple data sources or macro-aggregates, reference each source and indicate how they are combined. For each survey source, summarise the sample design, cross referencing the descriptions of the target and survey populations, presented in S.03.6. For each ~~administrative~~ data source, summarise the source, its primary purpose, and the most important data items acquired. Information in which form the metadata for the new data source is available, where it can be found, and if it is updated on a regular basis.	Webdata is used. The text displayed on the main page of the website of a company (excluding text generated by scripts) is used in the classification. Websites written in English and Dutch websites are processed and classified.

S.18.2	Frequency of data collection	Frequency with which the source data are collected.	Indicate the frequency of data collection (e.g. monthly, quarterly, annually, or continuous).	This has been done once now for the entire population. It can be repeated every 2-weeks or at lower frequencies.

S.18.3	Data collection	Systematic process of gathering data for official statistics.	*For each survey data source:* • describe the method(s) used to gather data from respondents; • annex or hyperlink the questionnaires(s). *For each administrative data source* • describe the acquisition process and how it was tested. *For all sources* • describe the types of checks applied at the time of data entry. *For big data sources* describe the methods used to collect the data; add hyperlink if it is web data or name of the API used to collect the data. *European level* Provide a summary of the commonalities and differences in the collection methods, questionnaires and checks used in different countries.	A list of companies (Chamber of Commerce numbers) and URLs of their websites is requested from the maintainer of the frame. Websites are visited and the content of the main page of the website is scraped and stored locally (if the site is not active or does not respond, it is visited at a later date and time; max 4x). Various scraping methods were tested and the results were compared. The scraping approach that provided the most reproducible and accurate results on the training and testset was used.

S.18.4	Data validation	Process of monitoring the results of data compilation and ensuring the quality of statistical results.	Describe the procedures for checking and validating the source data and how the results are monitored and used. Describe the procedures for validating the aggregate output data (statistics) after compilation, including checking coverage and response rates, and comparing with data for previous cycles and with expectations. List other output datasets to which the data relate and outline the procedures for identifying inconsistencies between the output data and these other datasets. Define the linkage method for big data sources and other data sources used for validation. *European level* Provide a summary of the commonalities and differences in the validation methods used by countries.	The model was internally validated by using a random split training and testset (80%/20%). For stability over time, the internal validity on multiple datasets was determined and compared. The model was externally validated on a set of Dutch startup websites. The data was validated by comparing the findings for the number of technological innovative companies of WP >= 10 with those reported by the most recent CIS survey result. The findings of WP >= 10 were also compared at the provincial level (the most detailed level on which CIS survey results are available) The relation between a company and its URL was assumed to be correct. The centrally available linking frame was used for this.

S.18.5	Data compilation	Operations performed on data to derive new information according to a given set of rules.	If there is missing data, give detailed description of the methods used for imputation. For big data sources, e.g., web data, indicate the reason why data were not collected (technical issues etc.). Describe the procedures for imputation, the most common reasons for imputation and imputation rates within each of the main strata. Describe the likely impact of imputation. Describe the procedures for adjustment for non-response and the corrections to the design weights to account for differences in response rates. Describe the calculation of design weights, including calibration (if used). Describe the procedures for combining input data from different sources.	The scraped websites were processed as described below: i) removal of scripts ii) extracting the visible text iii) language detection iv) removal of stop words v) removal of words <= 2 characters v) stemming of the remaining words Websites with <10 words remaining were not classified. These were added later assuming the ration innovative//non-innovative is similar to those with >= 10 words. Websites that did not result in a scraped webpage, were assumed to be inactive. The status code was checked and stored for that purpose. These website were excluded from the population (this was mainly the case for small companies). The number of technological innovative companies without a website was determined to be 1 in a 1000. The estimates were adjusted accordingly.

S.18.5.1	A7. Imputation – rate	The ratio of the number of replaced values to the total number of values for a given variable.	Provide values of indicator A7 Imputation – rate	No imputation took place at the individual company level.

S.19 Comment

SIMS	Concept Name	Defintion	Guidelines	Answer
S.19	Comment	Supplementary descriptive text which can be attached to data or metadata.	Provide any information that is pertinent to the report but does not fit under any of the other concepts, or to repeat key issues, or to make reference to annexes that might be attached to the report.	Current studies are performed to check if other variables in the CIS survey (i.e. other forms of innovation) can be determined in a similar way.