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 K3: Revised Version of the Quality Guidelines for the Acquisition and Usage of Big Data

Final version, 18.5.2020

Prepared by:
Sónia Quaresma (INE, PT)
Jacek Maślankowski (GUS, PL)
David Salgado (INE, ES)
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 4

Objective of these guidelines 4

Why do we not take into account some of the traditional quality dimensions? 4

About the structure of this document 5

Why we look separately at the input phase, why we split the throughput phase into two parts and why we do not address the output phase 5

Why the input phase is structured along different ways of acquisition and recording of the data 6

Why the first part of the throughput phase is structured along different data classes 6

Why the second part of the throughput phase is structured along the applications of new data sources 7

Comparison between these guidelines and the Quality Guidelines for Multisource Statistics (QGMSS) from the Essnet KOMUSO 7

Input phase - data source 8

Acquisition and accreditation of privately owned data 10

Data scout / new data sources appointee 10

Foundational principles 11

Initial examination of source, data, metadata and different variants of data access 12

Clarification of data access possibilities and acquisition of (test) data 13

Forensic investigation of the test data 13

Statistical office decision 14

Formal agreement with source 14

Literature 15

No cooperation with the data owner needed to gain access to raw data 15

Webscraping 15

All technological processes happen on-premise (at the NSI) 16

AIS Data 16

Smart meter 17

Some technological processes happen off-premise (at the source) 18

MNO data 18

Getting access to MNO data 19

Literature 20

Throughput phase I: Deriving statistical data from raw data of a big data source 20

Definitions and explanations 20

Mobile network data 30

Smart meter 41
Earth observation (satellite images) 44
AIS (tracking ships) 48
Web scraping (online job vacancies, enterprise characteristics) 51
Social Media 58
Throughput phase II: Usage of the derived statistical data for the production of statistical output 64
Big data sources as input for the production of official statistics 64
Replacement of questions from surveys 68
Validation / comparison of results with results from traditional data source 70
Survey based estimation with auxiliary information / calibration 71
Flash estimates based on leading or correlated indicators 73
Appendix I - List of abbreviations 76

Introduction

Objective of these guidelines

As the European statistical system (ESS) gains experience in the acquisition, processing and use of new data sources, also the demands on quality become clearer. These guidelines use the quality based experiences in the ESSnet Big Data II to formulate guidelines for those who already use as well as those who are planning to use new data sources for the production of official statistics in the future.

It has become clear that the access, as well as the processing and the usage of new data sources include very source and data-specific processes. Due to the diverse nature of the new data as well as the new data sources, it is a huge challenge to formulate generally applicable quality guidelines with practical relevance, which are more than a general reassertion of the very abstract principles.

To overcome this challenge, we follow the principle: "as general as possible, as specific as necessary".

These guidelines guide users and potential future users of new data sources in the following relevant questions:

- What are the key quality issues with respect to the data access?
- What quality dimensions are relevant while processing the new data?
- What are the key quality issues with respect to the usage of new data in the statistical production process?

For each of these questions we list general guidelines in a sub-chapter worded in general terms, and we try to overcome the difficulties with the diverse nature of the data and the data sources by drafting specific sub-chapters based on a case differentiation. This means we list guidelines in sub-chapters for different forms of data access, for different data classes (coinciding with the work packages (WPs) in the ESSnet Big Data II) and for different applications of new data sources. As a consequence, users can focus on the guidelines relevant for their intended form of data access and their intended data usage as well as on the data class in question.

This modular approach does clearly not raise the claim of giving an exhaustive list of guidelines. On the contrary: if a new way to cooperate with new data owners or a new data class emerges, it is clear that this document should be extended by new guidelines.

This document incorporates already existing guidelines like the guidelines for an accreditation procedure of raw data from private data owners, but it also includes newer forms of data access to pre-processed data by the data owner.

Most of the quality aspects described in the quality report of WP8 of the ESSnet Big Data I are again covered in the Section "Deriving statistical data from raw data"

Why do we not take into account some of the traditional quality dimensions?

In this document, we focus on quality aspects which are affected by the involvement of new data sources. Some well-known and traditional quality dimensions - mainly output related - like relevance,

reliability, timeliness and punctuality are barely affected by the usage of new data sources: Statistical output has to be relevant and reliable, independent from the sources, it also has to be published on time, independent from the sources. For other traditional quality dimensions like comparability and coherence additional aspects become relevant: Is a data source comparable (in the sense of stable) over time? Is a statistical output for which new data sources are used coherent with a statistical output produced on the basis of traditional data sources?

From a production process point of view, the usage of new data sources mostly affects input and throughput related quality aspects. Thus also this document concentrates on the input and the throughput phase.

About the structure of this document

Why we look separately at the input phase, why we split the throughput phase into two parts and why we do not address the output phase

The structure of these guidelines follows a production process logic, with focus on those phases and processes in the production process, which are affected by new data sources. The usage of new data sources introduces or changes many processes in the production of official statistics. The most obvious change happens in the input phase, where the acquisition and the recording of the data can look completely different than in the case of survey data or administrative data. Especially in case of privately owned data, completely new ways of cooperation have to be developed.

We further split the throughput phase into two parts. The idea is to look separately at a lower processing level, where potentially unstructured raw data is processed into well structured intermediate ("statistical") data, and an upper layer in which the statistical data is used to produce statistical output. This idea is often found in modern literature about the usage of new data sources for the production of official statistics, but a commonly used language has not fully emerged yet. We will address the subdivision of the throughput phase in more detail in Section Troughput phase, where we also present definitions and explanations for newly introduced vocabulary.

We do not address the output phase with a chapter of its own, since the usage of new data sources does not alter the typical processes of the output phase like dissemination and evaluation.

For the moment, the following figure gives a rough overview of the phases, and what topics we cover in the respective chapters.

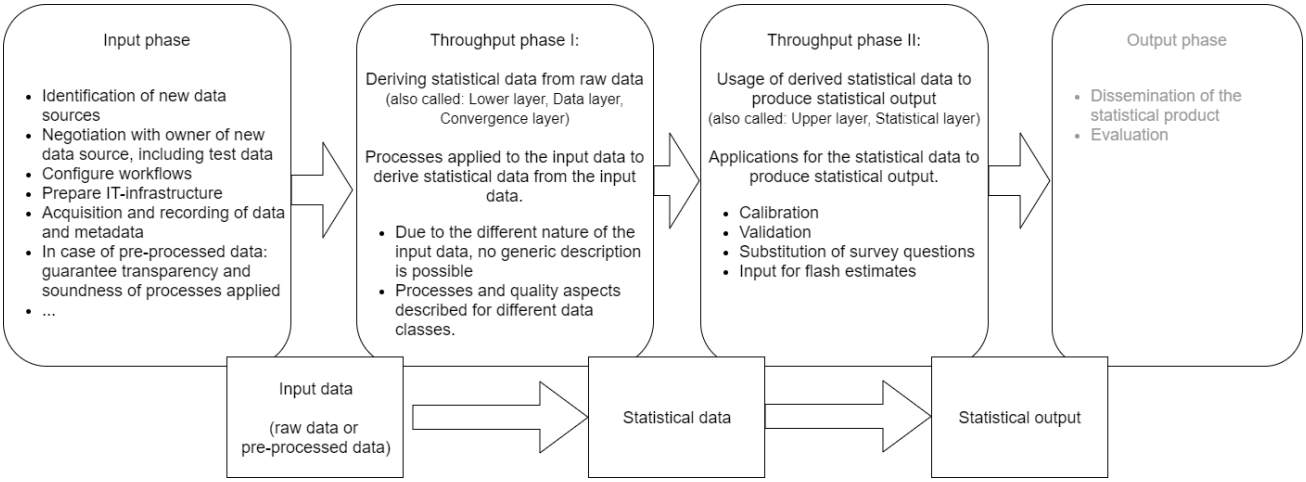

Figure 1 Illustration of the different phases of the production process with new data sources, including an illustration of the data resulting from each phase and serving as input for the next phase

Why the input phase is structured along different ways of acquisition and recording of the data

It is almost impossible to state generally applicable quality guidelines for the input phase. The reason behind is the very diverse nature of new data sources. In the input phase, the salient point is the divergence in the ways the National Statistical Institutes (NSIs) access the data from a new data source. Just think of two examples of data acquisition and recording: In the case of web-scraped data by the NSI, the data owner is the statistical institute itself. In the case of data from mobile network operators (MNOs), the data is privately owned by the mobile network operator, with whom the NSI has to negotiate an access to (part) of the potentially pre-processed data. These two examples illustrate that guidelines are only then meaningful and applicable when they take into account different forms of data acquisition and recording. This is the reason why we group this chapter about the input phase by the form of the data access. General guidelines for the acquisition and recording of privately owned data are presented in an extra sub-chapter prepended to the other subchapters.

It should be noted that with new data sources, the input phase and the throughput phase may sometimes be difficult to distinguish. The reason is, that sometimes the NSI only gains access to data pre-processed by the data source, and processes normally happening at the premise of the NSI ("on-premise") take place at the data source. Typical processes are selection of variables, some form of aggregation, but also some form of validation can happen at the premise of the data source ("off-premise"). Depending on the differentiation between on-premise and off-premise, the NSI has different insights in the processes applied and thus, different guidelines become relevant.

Why the first part of the throughput phase is structured along different data classes

In the field of official statistics, survey data and more recently, administrative data are seen as traditional data sources, all other data sources are subsumed as "new data sources" or "non-traditional data sources", or even less accurately "big data". Ricciato et at (2019) state: "This view is overly simplistic: the difference e.g. between satellite data and MNO data are so fundamental, in virtually every relevant dimension, that any attempt to blend them into a single joint methodological framework is likely to result into nothing more than a general reassertion of the very abstract principles, with little practical relevance". This is why Ricciato et al (2019) advocate the introduction of data classes, which are defined by methodological, technical and governance aspects. We are

convinced that the same observation also holds for a quality framework. We believe that it is barely possible to write down meaningful quality guidelines, which can be applied to all kinds of new data sources. Instead, we group the chapter covering the lower and the convergence layer into sub-chapters according to so called data classes, for which we write down data-class specific quality guidelines. General quality guidelines, non-specific for a data class, are collected in an extra sub-chapter prepended to the other subchapters.

Why the second part of the throughput phase is structured along the applications of new data sources

At the end of the first part of the throughput phase, statistical data has been derived from the input data. This statistical data can then be used for different purposes. Just imagine, an NSI gets access to MNO data of all mobile service providers of the respective country, and one of the derived statistical data sets would contain the information about all stays abroad of local SIM cards. Very distinct statistical applications could be applied to this statistical data: The NSI could produce statistical output like the number of stays abroad per week and per country based solely on this data. Another application would be to validate the number of stays abroad, produced with traditional survey data, by the new information. And further, the newly derived information could be used as auxiliary information for traditional survey data, for example for calibration.

Whereas the ways to derive statistical data vary vastly for different data classes, the applications seem to be more uniform. This is the reason why it seems reasonable to structure this chapter along the applications.
Still, for the time being, we have to add that the wealth of experience for producing statistical output (or even official statistics) with big data sources involved is limited. The output for some of the WPs of the pilots track will rather be of the form of statistical data than in the form of a real statistical product. Therefore, the chapter on the second part of the throughput phase is much shorter than the chapter on the first part. And also the variety of given examples is at the moment rather limited.

Generally speaking, one of the biggest challenges of this document is to find an optimal trade-off between project-specificity and generality. The advantage of the ESSnet Big Data is that we can draw guidelines from the actual experiences made by the members when working with new data sources. On the other hand, one would like to have guidelines which are valid and useful also for "new" new data sources. Our way out of this dilemma is to choose a modular approach. We formulate general guidelines when possible but project specific ones when necessary. If completely new data classes are used in the future, these guidelines should easily be extendable by adding new sub-chapters.

Comparison between these guidelines and the Quality Guidelines for Multisource Statistics (QGMSS) from the Essnet KOMUSO

The current quality guidelines have a different structure to other quality guidelines, in particular the "Quality Guidelines for Multisource Statistics" (QGMSS), developed in the ESSnet KOMUSO project. Those quality guidelines deal with the case where official statistics are produced on the basis of two combined data sources, namely survey data on the one hand and one or more administrative data sets on the other hand. Of course, with both sets of guidelines covering the topic of enhancing the possibilities of traditional statistics, the question arises about the reasons for having different structures in the two documents.

On closer inspection one can motivate the differences between these two sets of quality guidelines:

First, work with administrative data and work with new data sources fundamentally differ in the maturity of the processes involved. Whereas official statistics has gained in the meantime a lot of experience in including administrative data, the work with big data is mostly in the exploration phase or in the piloting phase. Even for those projects in the implementation phase producing official or at least experimental statistics, the processes involved have barely been standardized, especially not across countries, the processes are still revised very often and also a common language is only evolving. This situation justifies the different focus of these guidelines in comparison to the QGMSS: the latter focuses on the official output, whereas the former focuses on more general statistical products, which very often are not yet considered for publication as official statistics.

Second, with new data sources the focus of quality considerations shifts away from the traditional output quality towards the quality of all the new processes involved. This is also reflected in the structure of the two documents. Whereas the QGMSS are structured along the traditional ESS output quality dimensions, this document - as motivated in the previous sections of this chapter - is structured along the production process. The input phase includes completely new considerations about acquisition and recording of new data sources, which barely play a role for administrative data sources. As an example, the whole topic of pushing computation out, with technical processes happening on-premise of private data-owners, is completely new and is not covered by any of the traditional quality categories focusing on output quality.

Third, with new data sources, it becomes more important to do a data-class-wise quality assessment than to go through one general error type after the other. The reason is that processes diverge hugely across data classes, and so do potential errors.

In addition, the QGMSS focus on the direct use of new data source (more specifically, of administrative data sources). Indirect uses for new data sources such as validation and calibration are not covered by this document. The QGMSS focus on some basic data configurations, i.e. situations in which micro and macro data from the two types of sources are combined under different hypothesis on the coverage errors. Contrary, the whole Chapter "Throughput phase II" is structured along possible applications of new data sources, including direct and indirect uses of the new data source, as well as single-source applications as well as multisource applications.

In summary it can be stated, that of course there are many quality aspects which are covered in both documents, but due to the very different nature of the processes involved, the focus and as a result the structure of both documents differ fundamentally. At this stage of maturity of production of official statistics based on big data sources, the two manuals complement each other, and provide a set of quality standards for guiding the statistical producers in shifting from traditional to more complex approaches, involving multiple or new sources. Potentially, an attempt to merge the two manuals would make more sense when the production of official statistics based on new data sources has reached a higher level of maturity.

Input phase - data source

This chapter is grouped by the nature of the data access.

New data sources do not only differ with respect to structure and content from traditional data sources, but also with respect to the way data is acquired or recorded. The figure below shows one way to classify different scenarios of data access for the NSI.

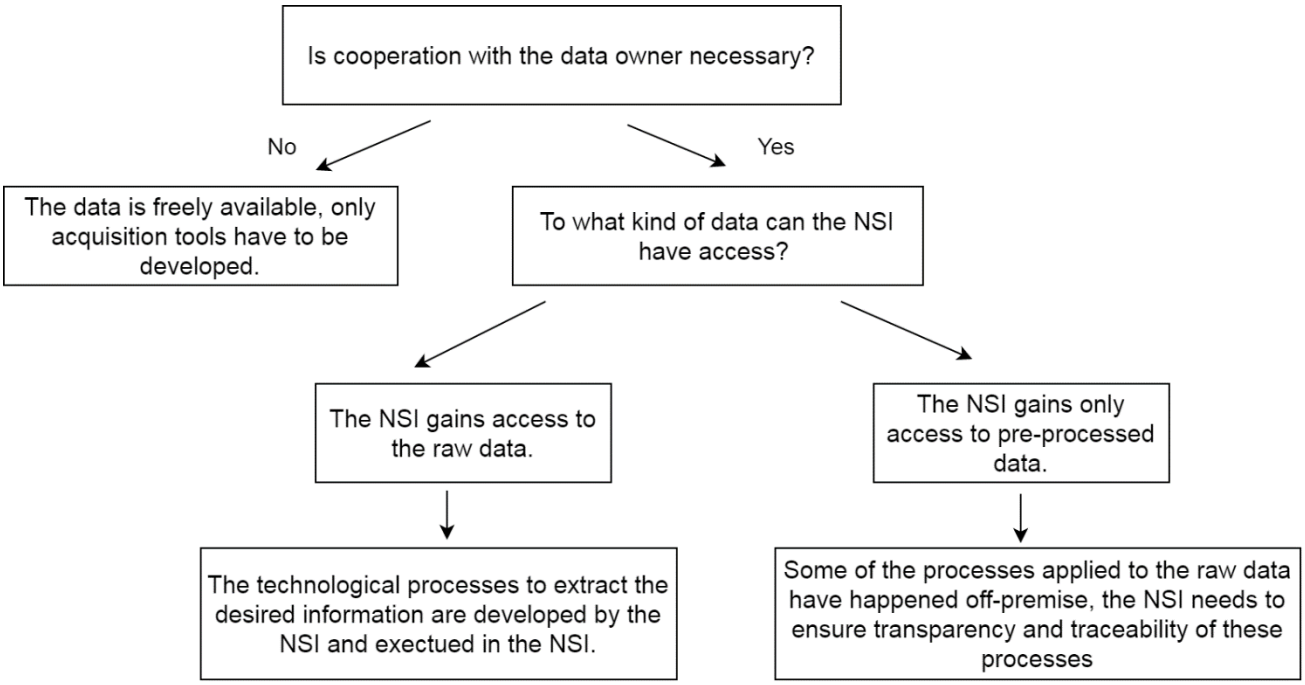

Figure 2 Illustration of the decision tree to classify different scenarios of data access to a new data source

The fundamental question is, if the access to a new data source depends on some form of cooperation with the data owner. In most of the cases, such a cooperation is necessary, there are only some exceptional cases such as freely available information on the web. In these cases, the NSIs only need to develop acquisition channels for harvesting this information, such as web scraping and the use of APIs.

Already in the past, NSIs gained experiences in cooperating with data owners outside the statistical system, namely administrative data owners. But the decisive difference to the access to new data sources is the current lack of legal entitlement for a data access to privately held data.

Further, cooperation with a data owner does not always include the full access to the raw data. In fact, the focus of the statistical system should be on the extraction of the desired output information. The acquisition of the (raw) input data should be seen merely as an ancillary (but not always necessary) task. Instead of access to the raw data, the cooperation with the data owner might include the access to pre-processed data. In these settings, some of the technological processes applied to the raw data are executed at the source premises (from an NSI's perspective, this is called "off-premise"). Further, one could differentiate between the case, where the NSI has developed the technological processes applied to the raw data, and the case where these technological processes were developed by the data owner or a third party.

Depending on the necessity of a cooperation with a private data owner and on the nature of the data access, differing quality guidelines apply. This is why we subdivide this chapter into sub-chapters

according to the decision tree in the diagram above. We prepended a general sub-chapter about the acquisition and accreditation of privately owned data to the other sub-chapters.

Acquisition and accreditation of privately owned data

Data scout / new data sources appointee

By their nature, the potential usefulness of a new data sources is unknown for NSIs, sometimes not even the existence of a potentially useful new data source is known. At the very beginning, an NSI has to become aware of a new data source and its potential usefulness for official statistics.

It is clear that it can not be fully determined in advance whose responsibility it is to become aware of a useful new data source. It can be the department producing an official statistic which becomes aware of a potential new data source that would help to improve the production of the statistic. It is just as possible that there exists a unit within the NSI specifically working with new data sources who finds out about a new data source and makes an approach to the specific department. In some NSIs there even exists a "big data scout" - a person whose responsibilities include the lookout for new data sources. Similar to the suggested organisational structure for managing quality in the Indicator 4.1 of the Quality Assurance Framework (QAF), a clear organisational structure for managing the access to new data sources is needed.

Guidelines for a new data sources appointee

New data sources raise new questions about governance and allocation of responsibilities. Only a clear governance structure and a clear distribution of responsibilities with respect to new data sources guarantees an efficient handling of new data sources.

Therefore, a person or an organisational unit should be appointed within the NSI who is responsible for the lookout for and the acquisition of new sources.

The new data sources appointee(s) should have the mandate from the top management of the agency to speak for the NSI when approaching the data owner and starting the negotiation process to access the data.

The compliance with these guidelines has the following advantages:

- The appointee(s) can gather experience with the acquisition procedure from different data sources and can profit from experiences made in previous acquisition procedures.
- A new data source can often provide multi-faceted information serving multi-domains. Thus, it falls into the appointee's remit to inform and coordinate all domains/departments who might potentially be interested in the new data.
- The appointee can also coordinate and involve the technical and legal units involved in the acquisition process.

Example Data Scout in the Netherlands

Data Scouting at Statistics Netherlands is described as follows (Presentation "Datascouting at Statistics Netherlands, What is it and how do we approach it", Florian Henning, 2019)

- Coordinating function: bridge-building between intern and external stakeholders
- Define data needs together with internal and external users of CBS statistics
- Exploring possible relevant new data sources
- Building relations with other organisations
- Liaising with legal, technical and domain experts on what is possible (and what not)
- Negotiate conditions, define business models, formalize agreements

The following sub-chapters follow in many aspects the paper "Proposal for an accreditation procedure for big data source" by Wirthmann et al (2015). Whereas the sub-chapter "Foundational Principles" is a direct citation of Wirthmann et al (2015), the subsequent sub-chapters were modified by us. The paper implicitly assumes the negotiation of a full access to the raw data, it does not investigate the various possibilities for partial data access / data usage as described above. We extend the scope of this paper to these cases, where the cooperation of the NSI and the data owner includes only the access to pre-processed data.

Foundational principles

The following set of principles is proposed as a basis for any procedure of accreditation of non-official data sources. They are 'designed' to remain stable over time, even if changing conditions necessitate modifications in the accreditation procedure itself. The principles ensure that the procedure is fit for use in a statistical system, efficient, cost effective and reliable and that it takes into account all possible impacts of the adoption of a new data source on the producer of official statistics.

Guidelines

Principle 1: Accreditation must be fully compliant with the well-established principles and quality frameworks that guide the world of official statistics and consistent with quality assurance practices embedded deeply in the work of statistical offices.
Principle 2: Any accreditation procedure must be flexible in a way that does not unduly prejudice or rule out new opportunities without serious examination.
Principle 3: An accreditation procedure should include sequential decision-making based on a pragmatic step-wise approach, so that new data sources that will not work are spotted early on, while investment in those that will work is not jeopardized.
Principle 4: The accreditation procedure must contain an empirical assessment with real data and it must be carried out by statistical offices directly. It cannot be delegated to filling out questionnaires by the source owners.

Principle 5: A systematic accreditation procedure must assess the quality of the statistical inputs (including the source and metadata), of the statistical outputs, as well as of the statistical processes involved.
Principle 6: The final decision for the accreditation of a new data source must also incorporate a combination of corporate criteria, broader than strict data quality. The accreditation procedure must compile adequate supporting documentation, including measurements.

Initial examination of source, data, metadata and different variants of data access

At the very beginning, the overarching question about a new data source is the potential usefulness.

Guidelines

When the NSI gains knowledge about a new, potentially useful data source, all units/departments within the NSI, who might have an interest in/use of the new data, should be informed.

The new data sources appointee should glean the exact information, the different units within the NSI hope to get from the new data source. This is important since the different intended purposes entail different requests for variables, different depths of detail as well as different ways of data access to the same new data source.

Example Mobile Network Operator data

Mobile network operators (MNOs) have very rich data about their clients, which includes also very different types of data (e.g. CDR data, signalling data, RAN data,...). It is unrealistic, that an NSI gains access to all of the available information. Thus, already in advance, the NSI has to figure out, which information is needed for the intended purpose(s). MNO data includes information about the presence as well as the mobility of clients. Whereas statistics related to the presence (e.g. de fact population) need information about the stay of persons, statistics related to mobility (e.g. tourism flows but also statistics on commuters) need information about the movement from one position to the other, but on very different levels of detail (countries versus much more granular geographical positions).

Already before any official contact to the data owner, an early assessment of the data, the metadata and the source is needed.

Guidelines

Anything that can be gauged from the outside or through limited and rather inofficial interaction with the working level at the source organisation should be collected.

Detailed questions can examine the following areas:

- the population coverage
- the units of measurement
- variables
- timeliness and frequency
- information on the organisation

Clarification of data access possibilities and acquisition of (test) data

This stage entails negotiations with the source with a view to acquire a set of data adequate for rigorous testing. Formalizing a legal agreement is not part of this stage.

Guidelines

At this stage, potential modes of data access should be clarified with the data owner.

- Is the data owner willing to share raw data with the NSI?
- If the raw data cannot be acquired, is there a possibility to gain access to pre-processed data? (For more details see the Subchapter "Technological processes off-premises")

It should be clarified if a future access to data is on the same process-level (raw data, pre-processed data) as the test data.

It should be clarified whether the source is willing and able to deliver test data.

Specifications of the test data files must be discussed in a professional manner, including:

- Modes of data access
- In case of access to pre-processed data: transparency about the technological processes applied to the pre-processed data
- Time and method of transmission
- Metadata

At the end of this stage, a test data set - consisting either of raw or of pre-processed data - should be acquired, which is of the same form as the "real one", so that it can be thoroughly tested.

Forensic investigation of the test data

The acquired data set from the previous stage has to be tested thoroughly.

Guidelines

During this stage it should be clarified,

- which (main) processes - technical and statistical - are necessary to use the new data source,
- if the skills necessary to process the data are available in the statistical office,
- if the available tools of the statistical office can adequately deal with the new data, particular attention should be given to the IT-issues of storage and processing.

The forensic investigation of the data involves

- all the known steps involved in data cleaning,
- the production of aggregate statistics and the production of outputs,
- the linking of the new data with existing data.

Statistical office decision

At this stage, all the gathered information has to be used to make a corporate decision about the usage of the new data source.

Guidelines

The following questions about the statistical production should be considered:

- What are the exact uses of the new data and what are their impacts?
- Which existing statistical outputs could benefit from the new data source, and what are the implications and trade-offs? These trade-offs could for example include a more granular data source but with an unknown coverage bias.

Further, a top-level cost-benefit analysis should be carried out, which focuses on the financial picture.

The following questions about risks beside the statistical production should be considered

- How vulnerable will the outputs involved will become?
- Could there be any consequences to the reputation and the trustworthiness of the statistical office?
- Which legal aspects have to be considered?
- Are there socio-political aspects to be considered?
- What risk mitigation strategies can the statistical office develop?

Formal agreement with source

All the information needed for high level negotiations are now available.

Guidelines

Long-term access has to be guaranteed.

Issues of reciprocity have to be explicitly clarified - what kind of benefits (not necessarily financial) can the NSI offer to the data source?

Issues of governance need to be articulated, including change management and a dispute resolution mechanism.

Literature

Wirthmann, Albrecht & Stavropoulos, Photis & Petrakos, Michalis & Petrakos, George. (2015). Proposal for an accreditation procedure for big data source.

No cooperation with the data owner needed to gain access to raw data

Webscraping

There are several methods to acquire or record statistical data. The methods include questionnaires, administrative data sources, machine generated data etc. Generally, web scraping is the fastest and least expensive way of acquiring statistical data for official statistics purposes. However, to acquire high quality data, there is a need to ensure that the data in raw format is readable and linkable with other formats.

Guidelines

Ensure that each data set will have a corresponding metadata set. Use the unified format for data and metadata store.

When collecting the data, ensure that there are reliable attributes that can be used to link to other data (e.g., geolocation, NACE, etc.).

If possible, allow to access the raw data with the unified interface, i.e. the same name of fields for the specific dimension, e.g. company_id, NACE.

If there are any methodological differences in the interpretation of the same dimension, e.g. job vacancy vs. job offer, please use the metadata.

Ensure that all data is stored in a secure way and try to create different groups of users, e.g. external users vs. internal users to allow limited access to the data.

Try to estimate the target population size, if possible, and use metadata to store this information.

For webscraping, follow the document „ESS web-scraping policy“ prepared by ESSnet Big Data WPC.

Use similar classifications, if possible, or at least create the transition key to encode/decode the list of possible values from one data source to another,

i.e. level of education, recode lower secondary and upper secondary to secondary.

Store the data in machine readable format, which can be processed by the computer. It means that the data must be collected in the column or raw two dimensional tables, e.g. ID; dim1; dim2; dim3; value1; value2.

If possible, allow to access raw data in standard formats like JSON or CSV, to be easily loaded into most common data science environments, like Apache Hadoop, Python or R.

Replication and possibility of reproducing the data set for other purposes is one of the key issues with the framework presented in this document. Therefore, please use the most common unified formats to store and access this information.

All technological processes happen on-premise (at the NSI)

AIS Data

Automatic Identification System (AIS) data was originally developed to ensure safety at sea. Based on Global Positioning System (GPS) technology, the AIS system of a vessel broadcasts its location and status information over a radio channel, making it possible to be detected by other ships wherever they are, as long as a radio signal can be sent and received. AIS data is generally considered to be privacy sensitive as it can be used to track vessel owned by individuals.

Getting access to AIS data

AIS data is collected at several places around the world. In Europe, for example, the organizations EMSA, Kystverket, Hellenic Coastguard, Dirkzwager and Marine Traffic all collect AIS data. These organizations can be contacted to get access to large amounts of AIS data.

AIS data is, however, available at various levels, e.g. at national, European and world wide level. The higher the level and – hence - the size of the data, the higher the costs. Another important consideration is the legal ground on which the data can and may be accessed. At the national (country) level this is usually already arranged for the NSI. However, getting access to AIS data at the European level, for a particular country, requires taking an additional hurdle. In the workpackage that studied AIS data in the first ESSnet Big Data they looked at the various options and it was concluded that buying AIS data at the European level - from Dirkzwager - sufficed for the planned studies. The costs for worldwide data were estimated to be above the available budget and were therefore not further investigated. In the end, a 400 GB European dataset was obtained. Options to get the data free of costs were also investigated but were not successful in the relative short time available.

Transforming AIS data to be used for statistics

Raw AIS data needs to be converted prior to analysis (see Figure AIS for an overview). In this pre-processing step the data is decoded, after which individual AIS messages can be extracted. For WP4 of the ESSnet Big Data I, the pre-processing step was performed on the premises of Statistics

Netherlands for all data at once. The decoded data was subsequently transferred – in a secure way - to the secure UNECE Sandbox in Ireland, as all European partners needed to access the data. Analysis of pre-processed AIS data greatly benefits from the availability of a distributed and scalable cluster, such as Spark. The latter enables analyzing data by multiple machines in parallel which speeds up this process considerably. The researcher needs to assure that the interdependencies between records are reduced to a minimum as this will seriously affects the speed of analysis. This can, for instance, be done by placing all messages from the same vessel on the same partition of the cluster.

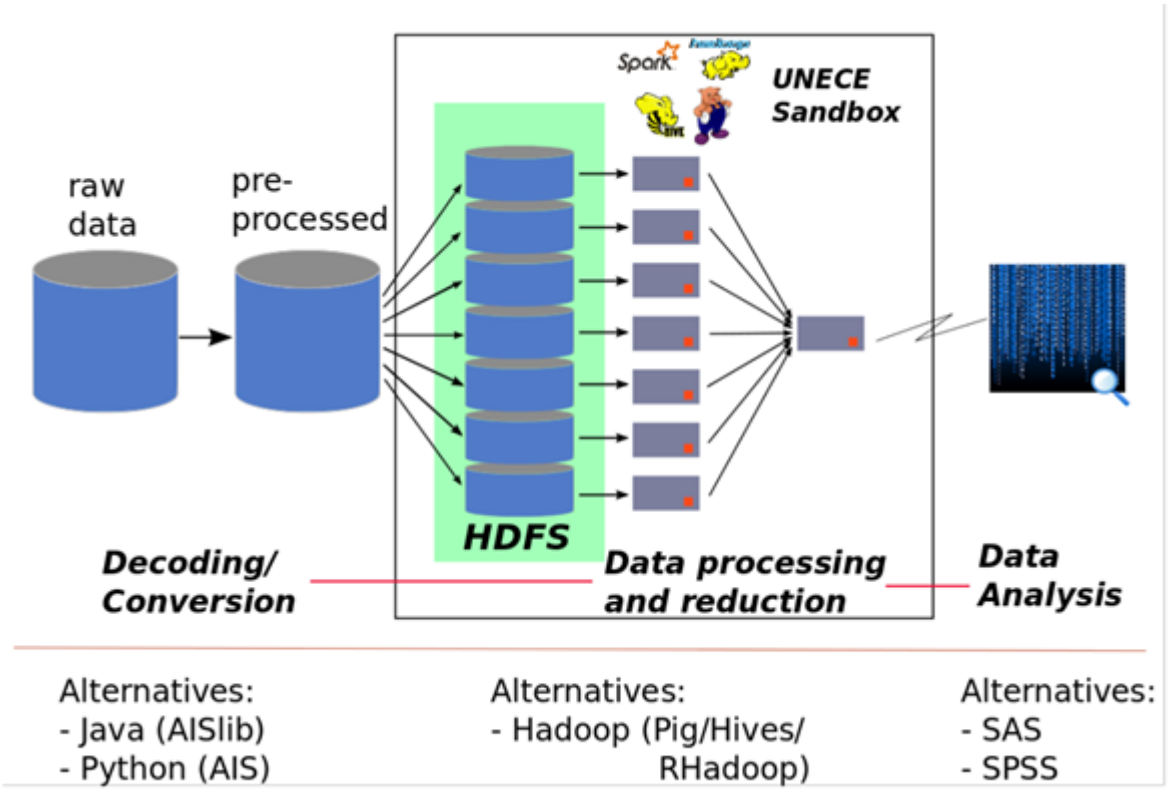

Figure 3 Overview how AIS data is converted before it can be analysed

Smart meter

A smart meter is an electronic device that records consumption of electric energy at regular intervals, usually at 15, 30 or 60 minute intervals. The device also communicates the information - at least daily- back to the system operator for monitoring and billing. Apart from this two-way communication, there is usually a central system were the data is stored which also determines the interval of recordings. Smart meter data raises privacy concerns as, depending on the frequency of data collection, significant personal details about the lives and private activities of customers can be revealed.

Getting access to smart meter data

Despite the large interest in smart meter data there were only two countries involved in the first ESSnet Big data project that had actual access to the data; these were Denmark and Estonia. The other countries involved faced either technical issues or legal restrictions. Legal issues affected 6 out of the 18 countries surveyed. To get access, there must be a legislative basis which allows collecting

personalized smart meter data by the NSI. It helps if the data is located in a centralized data hub; which was both the case for Estonia and Denmark. The Estonian situation is the most advanced and is discussed below as a typical example.

In Estonia the smart meter hub contains all agreements related to electricity transfer and consumption and all measurements. Via the hub, electricity consumers can: look at their electricity consumption points and their agreements; view historical electricity consumption data; and authorize one or more electricity sellers to access your data, so they can make personalized offers. Statics Estonia gets a copy of all data in the hub, which amounts to around 1.5TB. This data is transferred to the statistical office and analyzed in a cluster. To enable its use for statistics, a number of data enrichment and cleaning steps were needed. These focused on adding geolocation data to the addresses and timestamp checking and transformation. The geolocation enabled linking the data to other administrative sources. Timestamp conversion was needed to enable its use by the systems available at the statistical office.

Guidelines

When the pre-processing of the raw data is taking place on the premise of the NSI, an IT infrastructure to handle large amounts of data is necessary. The NSI should plan the needed resources (e.g., estimate the size of smart meter measurements per year) beforehand and invest in such an infrastructure already before the actual raw data is delivered.

If there exist several options of data sources for the same or at least very similar data, make a thorough cost-benefit analysis, including the needed level of detail, and take into account the transparency and soundness of methods and processes for the metadata and the data of each data source.

Keep in mind that even if the NSI gets access to so-called raw data, this data can depend on decisions at the data source, as even for raw data some kind of minimal processing at the source has to take place to be able to store them as the difference in coverage of several AIS data sources shows.

When the data is encoded in a specific format (e.g., special AIS binary format) , the NSI should already develop or acquire in advance tools to decode/convert the data with the help of a test data set.

Some technological processes happen off-premise (at the source)

MNO data

Mobile network data is generated from the interaction between mobile devices and mobile telecommunication networks. It is not data about the contents of the communication activities of subscribers (calls, SMS, data connection) but technological metadata needed for the network to provide these different services to each mobile device. Even despite this data is pseudonymised before any statistical processing, it is extremely sensitive due to confidentiality and privacy issues.

MNO data does not only offer information about the geolocation of individuals, but also about internet traffic (e.g. types of downloaded mobile applications) and about interactions between individuals. In consequence, privacy concerns about this data source are an outstanding element when getting access to it.

Getting access to MNO data

In this line, MNO data arises as a natural candidate for off-premise processing. Furthermore, this off-premise processing will need to be executed by MNOs themselves - not only for the aforementioned confidentiality and privacy issues, but also due to the high complexity of the technological environment and the data ecosystem at MNOs. This amounts to integrating these companies into the statistical process at the first stages of the production.

As a consequence, the access to MNO data has become an intricate issue for NSIs, especially for a production sustainable over time for several statistical domains. At the research stage, collaborations do already exist between NSIs and MNOs and even between some research centres and universities, on the one hand, and MNOs, on the other hand. But they are mainly one-off projects providing highly valuable insights into the potential of this data source.

For this reason, among others, in the context of the European Commission, an expert group was constituted which focused on business-to-government (B2G) data sharing recommendations (B2G, 2020). This group gathered independent experts from both the public and private sectors to agree on several recommendations. Complementary, within the ESS Task Force on Big Data/Trusted Smart Statistics, a group of experts from several NSIs has also been constituted to work on principles for the access to privately held data such as MNO data.

The following guidelines are included always under the assumption of seeking public-private partnerships between MNOs and NSIs under a collaborative framework.

Guidelines

Agree on roles

Agree with MNOs on the different roles played by MNOs and the NSI. The NSI should be part of the design of the whole end-to-end statistical process. MNOs should also participate in this design (at least for the initial stages) and will need to assume an execution role of some production tasks.

Audit raw data extraction

Agree with MNOs on the raw telco data to be used in the statistical process. This should be the result of a trade-off between the adaptation of the ESS Reference Methodological Framework for the Production of Official Statistics with Mobile Network Data and the technological and business feasibility to use this data.

Audit raw data pre-processing

Agree with MNOs on the statistical processing of raw telco data to generate intermediate data for the further statistical analyses. This intermediate data is identified by the Reference Methodological Framework for the Production of Official Statistics with Mobile Network Data and should be able to decouple the technological environment of telecommunication networks from the statistical processing for official statistics purposes.

Document data provenance and pre-processing

Document both the data provenance (which raw telco data exactly to use) and the data pre-processing (generation of intermediate data: method, parameters, etc.). This documentation must find an optimal trade-off between public transparency, citizenship privacy and confidentiality, and industrial secrecy and intellectual property rights. Any modification in either data provenance or data pre-processing must be duly communicated.

Literature

B2G (2020). High-level Expert Group on Business-to-Government Data Sharing. Towards a European strategy on business-to-government data sharing for the public interest. https://ec.europa.eu/digital-single-market/en/news/meetings-expert-group-business-government-data-sharing.

Throughput phase I: Deriving statistical data from raw data of a big data source

Definitions and explanations

This chapter covers the lower level and the convergence level as depicted in the diagram below.

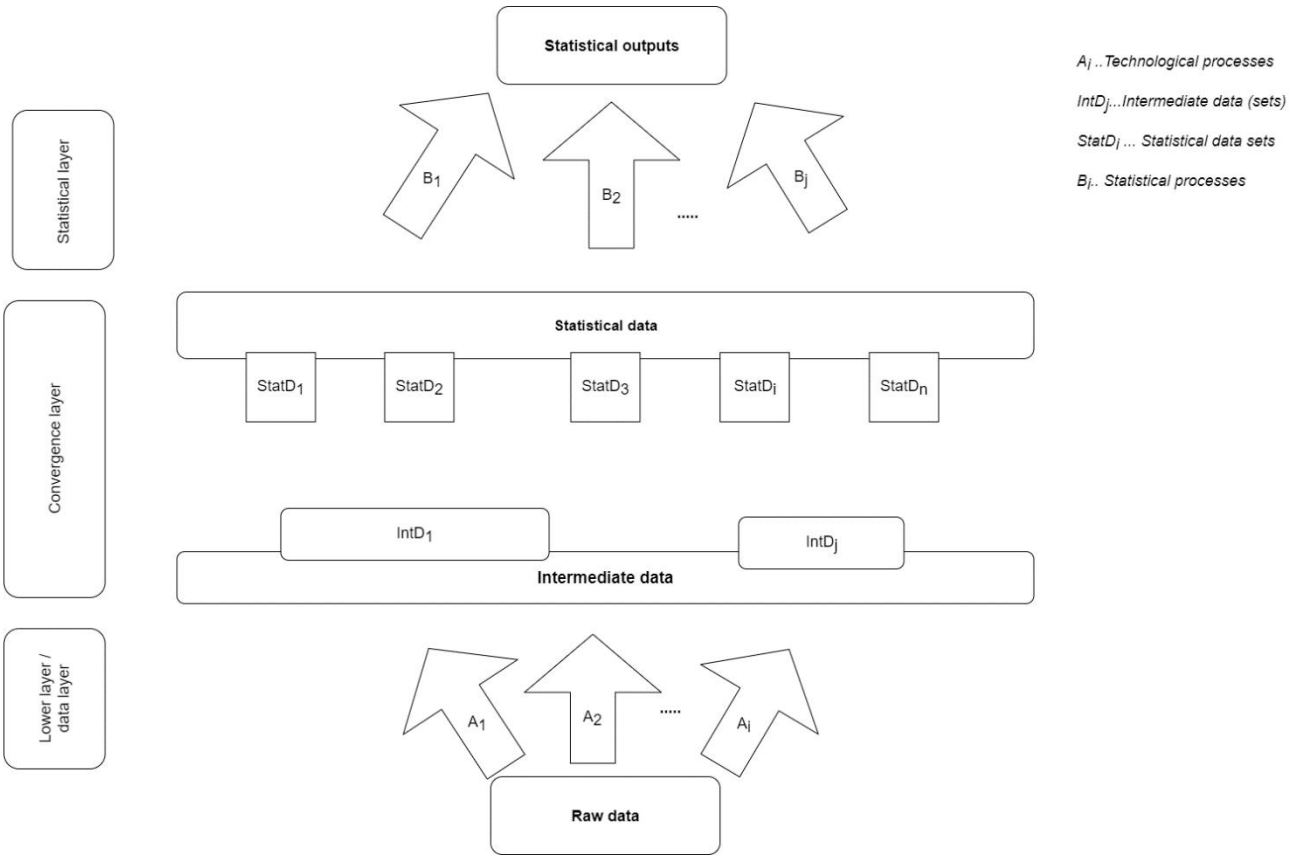

Figure 4 Illustration of the different layers, the different forms of data, as well as the different processes

A commonly used language has not emerged yet, thus we explain first what the terminology used by us means:

A **(big) data class** is defined by a similar data source and a similar data structure. As a consequence, we expect that the processes needed to derive statistical data from raw data to be similar among one class. As a consequence, it makes sense to group the quality guidelines by the big data classes. In the ESSnet Big Data the data classes coincide (mostly) with the WPs: MNO data, AIS data, earth observation data, web-scraped data, etc.

Raw data often comes in formats and semantics that are very complex and highly specific to the particular technology domain (e.g mobile network operator data, smart meter data, ship tracking data). Raw data might further be unstructured and/or rich of technology-specific information that is not relevant for official statistics. It can be unstructured like images, videos, audios and require a layer of interpretation (image and object recognition, speech interpretation etc) to be turned into categorical and/or quantitative data. It can be semi-structured like homepages or already structured in tabular form. If only a small component of information embedded in the raw data is of interest, selection functions (of variables, events, etc) may be required in the lower level. (c.f. Ricciato 2019)

Intermediate data is the intermediate stage between the raw data and the statistical data. Depending on the complexity of the data source, it can also coincide with the statistical data. Ideally, the semantics, format and structure of such intermediate data meet the following requirements (c.f. Ricciato 2019):

- It should follow a common structure and format for the whole data class, independent from the technological details that may vary across different instances within the same data class. For example, a single (intermediate) data format and semantics should be defined for the data class of MNO data, not specific to the configuration details of any particular MNO infrastructure. It other words, it should be "operator agnostic".
- It should be "technology agnostic", meaning that the intermediate data are in a form which does not depend on the technological details caused by the physiological evolution of the technological processes that produce the raw data. For example, in case of MNO data, the intermediate data should be independent from the technology generation 2G, 3G, 4G or the forthcoming 5G.

Depending on the characteristics of the new data source, all kind of (technological) processes have to be applied until we get so called **statistical data**. We define statistical data as the data which serves as input for the traditional statistical processes. The wording can be a bit tricky - it does **not** mean that statistical processes such as imputation or editing have already been applied to the data set. It only means that the data is in a format - namely in tabular form or a bit more general in a form manageable in relational databases - in which it can be further processed by statisticians.

It should be mentioned here that the relationship between the raw data and the intermediate data is not (necessarily) a one-to-one relationship. Content and even format of the intermediate data can depend on the intended use; different forms of intended use can lead to different intermediate data from the same raw data (simple example: earth observation and information derived from it - intermediate data from the same raw data differs depending on the variable of interest.)

It is clear that the difference between the statistical data and the raw data varies depending on the data class. Whereas e.g. in the case of smart meter data, the statistical data and the raw data are very similar, statistical data derived from earth observation data, from social media messages, from scraped homepages or AIS data is very different from the raw data.

Illustrative Examples for the Different Data Layers

The first example starts from raw data from MNOs, so this is already a quite diverse data source by itself. In this example we define three intermediate data sets:

- Domestic time-location data (sometimes referred to as C-Location Layer) in its most basic form consists of
  - an ID,
  - a timestamp,
  - and a probabilistic location in a given reference grid.
- Interaction data might consist of
  - two IDs
  - an interaction type (e.g. call or message)
  - a timestamp
  - a probabilistic location in a given reference grid for both units or other additional information

- Foreign roaming data, which might include at least the following information
  - an ID,
  - a timestamp,
  - and a country.

It is neither trivial nor unique to layout a specific convergence layer, especially how to define the intermediate data set(s); it is influenced by various parameter. In our example a quite complex data set including all the information from the three intermediate data sets is possible, however this would increase privacy risks.

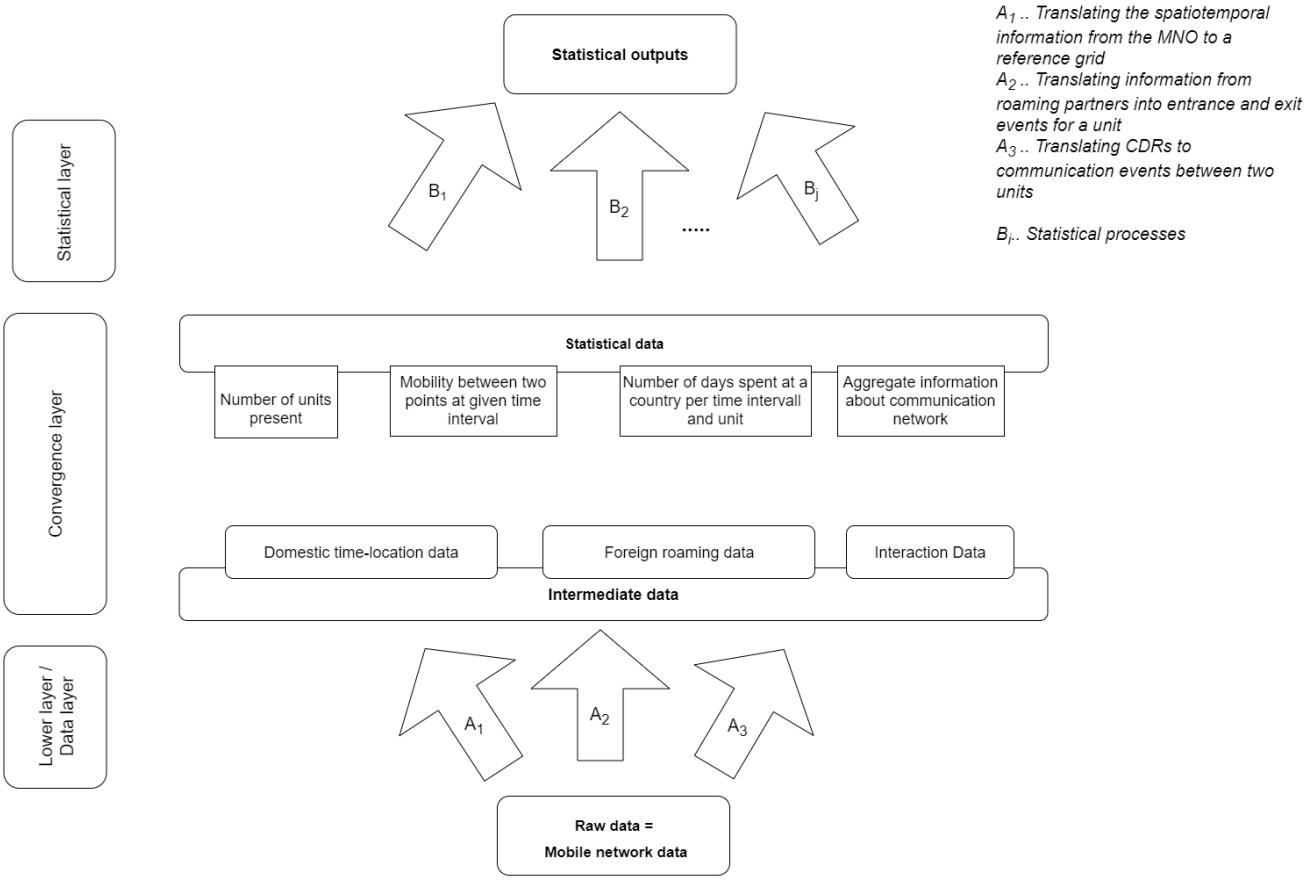

Figure 5 Illustration of the different layers and data forms with the help of MNO data

Illustrative Example of the Different Data Layers

The second example starts from the raw data from earth observation systems and the intermediate data sets could be defined quite close to the original data, where only harmonization steps in terms of the used coordinate system and image formats are applied. Afterwards processes like image classification might be used to create statistical data, e.g. the number/area of solar panels in a certain area.

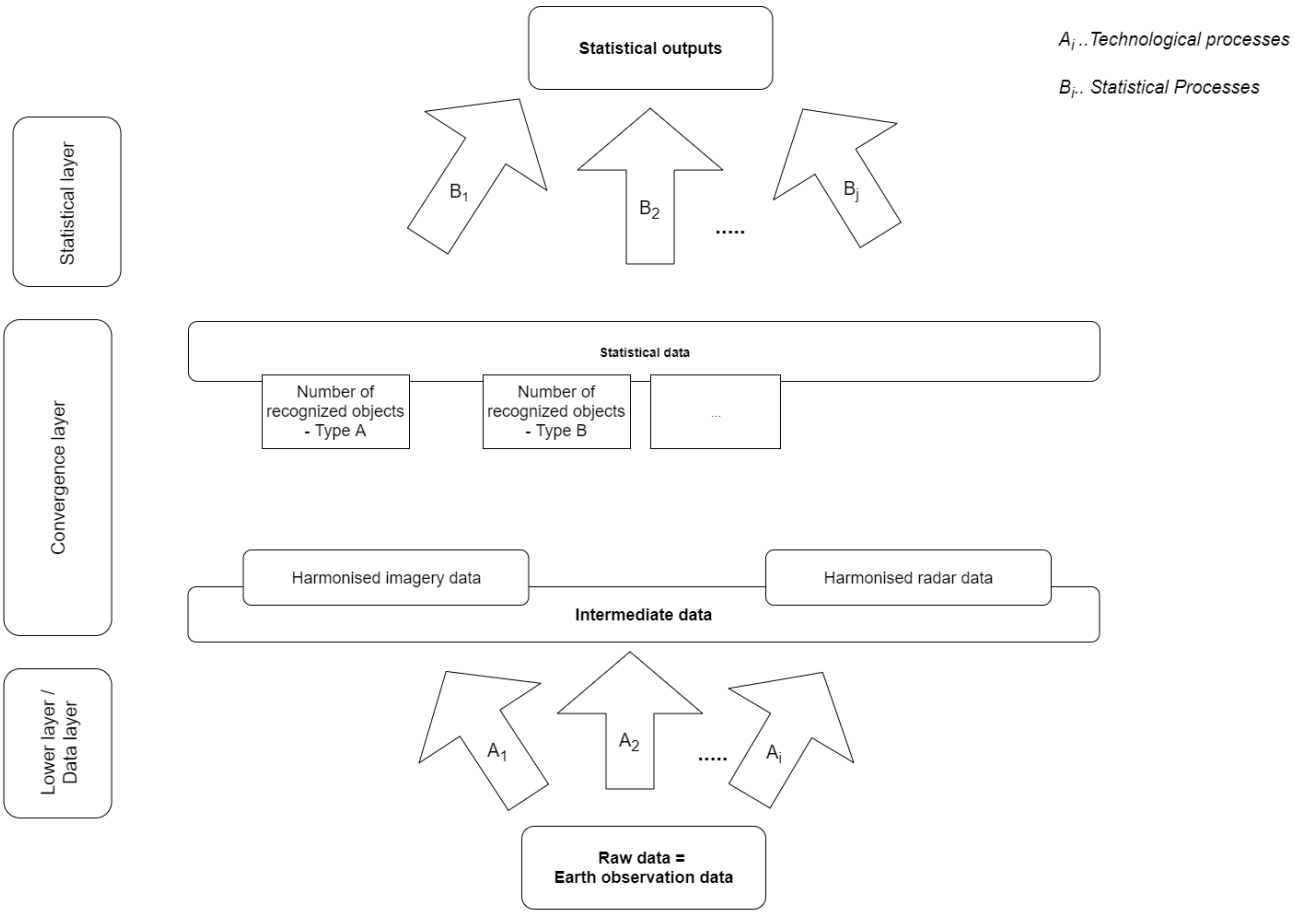

Figure 6 Illustration of the different layers and the data for the example of earth observation data

The nature of the following guidelines

The quality guidelines for this chapter can never be exhaustive, since a new data source can and most probably will entail new processes to arrive at statistical data and thus also new quality guidelines. In this document, we concentrate on data classes covered within the ESSnet.

In general terms, the guidelines will be presented following a meaningful order: from the fundamental and applicable to the more advanced and difficult to apply. They will reflect three levels of maturity: i) the awareness of the quality problems and consequences on estimates; ii) the existence of alternatives to prevent the errors or limiting them and choice among them; iii) the possibility to apply statistical techniques and models to handle the errors. Depending on the sources and the specific production situations, different levels of implementation of the guidelines are possible.

For some data classes it does not seem meaningful to list prescriptive quality guidelines for each quality aspect. The reason for this might be that the quality aspect is not relevant for the respective data class, or that the project is not mature enough to derive prescriptive quality guidelines from it. In these cases, it was a deliberate decision to leave the boxes with the quality guidelines away.

Quality aspects and error types

For survey based statistics, there exists a typification of possible errors with broadly accepted definitions of errors types. The statistical framework can be seen as (almost) exhaustive with respect to error types, meaning that the framework covers (almost) all potential errors which can occur in the statistical production process. Further, error types can be ascribed to non-overlapping error categories.

For multisource statistics, providing an exhaustive list of possible errors and categorizing these errors is much more challenging. New error sources emerge, definitions of traditional error sources might not fit perfectly any longer and definitions for new error sources have to be established. Also the classification of error sources in non-overlapping categories becomes more advanced.

For the case of statistics based on survey data combined with administrative data sources, a good overview of error types, for each data source alone as well as for the integration of the two data sources, is given in (KOMUSO, 2019). The authors also succeed in categorizing the error types in non-overlapping error categories. But even in this elaborated framework, it becomes sometimes difficult to determine to which error category an observed error belongs. An example are "holes" in the data set from survey data collection or administrative data acquisition, where it is unclear if the missing data is due to error sources errors attaining to the measurement line or the representation line.

For these quality guidelines for statistics based (also) on big data sources, we do not claim to provide an exhaustive list of error types. The diversity of the new data-class specific processes and the thereby emerging error types render this approach impossible. Neither is it always possible to ascribe potential errors to traditional error categories (see e.g. the description of the two error categories "processing errors" and "model errors" below). Instead, we focus in error types and quality aspects which were identified as particularly relevant in the context of big data.

The following quality aspects will also be covered within each data class, but the general comments on the respective quality aspects valid for all data classes are collected already here.

Coverage

Among the differences between big data and traditional sources is the degree of control NSIs have on the data acquisition and recording. In traditional surveys, NSIs can plan, design and carry out the acquisition procedures; for the use of administrative data sources, they may have agreements with the data providers and some knowledge on the reliability and quality of the data. Big data, on the other hand, share many characteristics with “found data”.

Indeed, in a survey life cycle, coverage errors attain to the dimension of the representation line, i.e. to the target population or the set of units to be studied (“who”) (Groves et al, 2004). Since the generation of the data depends on external factors and not on the NSIs' decisions, it is often the case that such data will not be representative of the whole population but just a fraction of it, which will probably have specific characteristics that differ from the broader population. In other words, the problem is similar to the one faced by studies dealing with non-probabilistic samples. In such cases, more than the sampling variance, the sampling bias is the most dangerous drawback that the estimates face.

For example, when considering Twitter data, it is evident that the collected tweets refer only to specific subsets of the more general population: the subset of people with a Twitter account and the subset of Twitter users that have chosen to share some of their messages publicly. Thus, inferences should not be made from a collection of tweets and any result from the analysis should be limited to the population underlying those tweets.

Guidelines

Establish the population of interest.
The definition and study of coverage errors require the definition of the target population, that should be explicitly identified in terms of type, time and place.
Surveys on potential bias.
Short surveys may be launched in order to identify the characteristics of an observed population, this might be done with traditional means. If target population characteristics are available from other sources, this analysis allows understanding the representativeness of the observed population with respect to those characteristics.
Surveys to obtain coverage estimates.
Capture-Recapture modeling is a well known class of methods that can be applied to estimate coverage, under given assumptions and informative scenarios.

Examples

A sample of persons or households could for example be questioned about their usage behavior of mobile phones, the number of phones per person and per household or the usage of specific social network sites. The results should offer an idea of the demographic characteristics of the users and the differences between them and the target population. It is also possible to include questions into existing surveys with the same aim.

Comparability over time

One of the main concerns about introducing big data in official statistics is the process generating data. Usually, the NSIs control the overall production process starting from the data collection phase, namely the survey. On the other hand big data is generated by a non-statistical purpose and provided by independent actors. In a fixed time point this issue could introduce errors such as coverage errors. In a longer time period an uncontrolled process can introduce troubles on the data comparability between two or more reference times. These troubles mainly depend on the stability of the data structure over time. When NSIs have limited chance to manage the data structure over time, it is important that they are aware of the risk of basing the statistical output on these type of data sources.

Guidelines

To deal with the concerns on the comparability over time of the statistical products, NSIs should rely on a suitable statistical framework. Here some relevant precautions to take into account are listed:

Closely monitor the structure of the data. Check each data generation on structural changes in comparison to the previous one.

Integrate use of different data sources. Rely the statistical output on more than one source of data . The sources can be of different typology: big data, administrative data, survey data.

Continuous updating of the data acquisition and recording tools: Web scraping, text processing and machine learning tools have to be agile to follow the necessary changes of the data source. For example, if the website (e.g. a job vacancy portal) changes its structure, a person at the NSI responsible for web scraping has to change the web scraper to record the appropriate data. In other words, to scrape the data in a long time series, we need to monitor changes on the website and modify web scrapers.

Fit an appropriate statistical methodology for producing the output. According to the Analyse Stage of data generating process by AAPOR (2015), apply a statistical method not sensitive to extreme data and define statistical tools for smoothing the break in the time series related to structural changes of the data source or coverage changes over time .

Measurement errors

Measurement error is the difference between the true value of the measurement and the value obtained during the measurement process. Measurement errors for survey based statistics were mostly attributed to human flaws (e.g. the inability of the respondent to give the true answer or the influence of the interviewer on the answer). Contrary, for new data sources, measurement errors are more related to defects in the instruments for data recording.

Guidelines

Establish the target information.
The definition and study of measurement errors require the definition of the target variable of interest.
Research on measurement errors.
If possible measurement errors should be evaluated (on a small sub-sample) with an appropriate method, e.g. manual reviewing or comparison with other data.
Track changes need to be observed.
If values are changed or imputed because of detected errors or implausibilities , these changes should be tracked.

Model errors (raw data to statistical data)

Big data based estimates are likely produced by models. The specifications of these models may be incorrect, which negatively affects the reliability of the estimates.

Working with big data sources instead of survey data, NSIs do mostly not find the information about the target variable in the data source directly; instead the information of interest has to be inferred from other variables in the data. This is not a completely new situation; the deduction of information about the target variable from other variables is also a common process when working with administrative data. Also text mining algorithms were already known before new data sources were considered, e.g. when working with open questions in a survey. Still, modelling the information about the target variable plays a prominent role when working with big data, and often, new models have to be developed. The complexity of the algorithms and models needed to arrive at the desired information about the target variable depends on how directly the information of interest and the available information from the big data source are connected.

Models can also be needed in the creation of statistical units, when they are not directly available in the sources. An example would be the deduction of the statistical unit "person" when the observation units in the raw data are SIM cards, with zero, one or more SIM cards per person. Errors in the creation of statistical units are defined as unit errors and are often categorized as processing errors (c.f. KOMUSO, 2019). When the creation of statistical units involves models, this error type serves as good example that in the context of big data, the separation between error categories becomes blurry. This observation is also mentioned in the quality guidelines for multisource statistics (KOMUSO 2019): "When the processing steps mentioned are done via a model they may result in model errors".

Statistical models have the special property of a foundation in probability theory, whereas machine learning algorithms are often of an ad-hoc/heuristic nature. A model in general is a simplification or an idealized form of the data-generating process (the truth), so model misspecifications can occur for classical statistical models, e.g. linear regression, but also for advanced machine learning algorithms, like random forest or deep learning, e.g. simply by not including an important variable.

Guidelines

Estimating the quality of models is of great importance:
Apply appropriate model selection and evaluation criteria. Techniques like cross validation, out-of-sample tests, etc. should be applied wherever possible to assess the model quality and possible errors.
Compare multiple machine/statistical learning methods.Since it is not always straightforward to choose the right tool for the job, different methods should be tested and evaluated.
Evaluate the bias of the training data set. In supervised learning, an unbiased training data is very important to not estimate based on a biased model.

Examples

To reduce under-fitting (high bias, low variance) and over-fitting (low bias, high variance) due to extreme matching of the algorithm models to the training set, different re-sampling techniques are used, such as bootstrapping or cross validation, in order to obtain a model that tends to give more general results.

Processing errors (raw data to statistical data)

During the processing of big data, errors may be introduced that negatively affect the quality of the data. Examples of this are the way outliers and missing values are treated.

Processes involved for producing statistical output with the help of big data sources are as varied as the big data sources itself, and thus also the potential processing errors are very diverse. The expression “Processing errors” itself is not completely clearly defined either. Some error types are counted among processing errors by some authors, other authors do not list them among this category (e.g. imputation errors). Further, the expression “processing errors” was defined – as most elements of the usual quality framework for official statistics – for statistical processes based mainly on surveys.
The quality framework was gradually extended when multisource statistics, which are based on more than one data source (e.g. survey data and administrative data sources), became increasingly important. For multisource statistics, additional processing errors can occur, which do not play a role for unique data sources (e.g. linking errors, errors when integrating data). The quality guidelines for multisource statistics (KOMUSO, 2019) describe processing errors as follows. "Processing errors are errors with manual activities." They list data entry error, coding or mapping error or misclassification, editing and imputation error, unification error, unit error and linkage errors in this error category. In the context of big data, it becomes clear that manual activities play a minor role. Instead, machine processing - including the use of models - gains significance. Therefore, also the error category of processing errors has to be extended to errors occurring during machine processing. As mentioned already above, the distinction between processing errors and model errors can become blurry and thus we treat them in the following sometimes within one chapter.

In some quality frameworks linkage errors are listed among the error category of processing errors, since linkage is one of the process steps when dealing with several data sources. Still, If linkage errors play a prominent role in the following chapters, we cover them separately in a sub-chapter of its own.

One can infer from the previous description that we treat the category of processing errors as a remainder category for potential errors in processes, which are often also very specific for each data class.

Literature

AAPOR (2015). Big Data in Survey Research. AAPOR Task Force Report. *Public Opinion Quarterly, 79*, pp. 839–880.

Groves R. M., Fowler F.J.Jr, Couper M, Lepkowsky J.M, Singer E., Tourangeau R. (2004). Survey Methodology. Wiley, New York.

Ricciato, F., Wirthmann, A., Hahn, M. (2019), Integrating alternative data sources into official statistics: a system-design approach.

ESSnet KOMUSO (2019), Quality Guidelines for Multisource Statistics (QGMSS), Version 1.1

Mobile network data

Description

The complex and rich data ecosystem in a mobile telecommunication network shows high volume and high velocity in data generation but it is highly structured with little variety. Data structures and semantics for statistical purposes depend strongly on the technology underlying the telecommunication services and different types of telco data can be reused to produce statistical outputs. Networks have a nested cellular (tree) structure with antennas in the leaves and the Network Management System (including the Billing System) at the root. The closer the origin of the data to the antennas (the leaves), the richer the data is for statistical purposes. Basically, we identify two large groups of data.

On the one hand, if data comes from the billing system we have so-called Call Detail Records (CDRs) or, when other services are also provided, Data Detail Records (DDRs). There is no closed consensus on the variables contained in a CDR/DDR, since it may depend on the technological solution chosen by the company. Furthermore, since this may vary with changes in the technology, we describe the main semantic content of CDRs/DDRs. From a statistical standpoint we identified the following variables:

- A pseudonymised ID.- This basically identifies each mobile device. The relation between phone number, SIM card, mobile device, and individual is indeed subtle, but in a given short time scope we can safely assume that the MNOs can assign such an ID to each mobile device/SIM card/individual. In some countries, the Law compels the MNOs to change this ID every 24 hours for privacy reasons. This will restrict somehow the statistical analyses thereof.
- Time attributes.- The billing system registers with a timestamp with a precision of one second when each network event (call, SMS, data connection, etc.) takes place. For more technical data, the precision can even be higher (unnecessary for statistical purposes).
- Spatial attributes.- The system provides also spatial attributes whose details and precision depend on both the technological solutions and the agreements about data access/usage for statistical purposes. For CDRs/DDRs these spatial attributes basically boil down to the radio cell of the antenna in which the network event has been detected. Sometimes, due to restrictions in the agreements to access the data, spatial attributes are coarse-grained to broader territorial units.
- Event information.- The system may also provide information about the network events detected, i.e. the type of event (call, SMS, data connection), duration of event, etc.
- Complementary variables.- Possibly, socio-demographic information from the subscription contract or a subscriber profiling in terms of the volume of data calls, SMS, and data connection, etc. may also be added to the CDRs/DDRs.

On the other hand, if more technical data regarding the interaction between devices and antennas can be accessed (which usually requires some investment), the complexity of data increases a great deal. These are called signalling data. Here we cannot even provide a superficial description of the

high number of these technical variables. Indeed, the identification of variables useful for statistical purposes is still a matter of investigation. The goal is to identify the optimal set of variables needed for the further processing for statistical purposes. Nonetheless, we still aim at providing both spatial and time information of each network event. Now these network events do not arise only from the communication activities by the subscribers, but also from passive events generated due to network technical operations. As examples of these technical variables, we may cite the signal strength, the timing advance (basically, the time needed for a signal to travel from a device to an antenna) or the angle of arrival to an antenna. Accessing these variables requires a higher technological investment and the subsequent data processing for statistical purposes.

In any case, raw telco data are not ready for statistical purposes and clearly need a non-negligible amount of pre-processing and transformation. As an essential element of this research we need to identify exactly what data we need for our statistical production in such a way that the production process is fully modular: Once these datasets are identified, the further processing will be independent of the underlying technology in the mobile telecommunication network ecosystem. The main lines of this transformation are described below and constitute the central idea behind the so-called Reference Methodological Framework for Mobile Network Data in the ESS.

The role of the big data class in the ESSnet

The analysis of mobile network data was already present in the first ESSnet on Big Data project during the period from January 2016 to May 2018 in the form of a dedicated work package to build a pilot study (WP5). In the current ESSnet Big Data II it is again present as another work package now aiming at building an end-to-end production process going from the information arising in the network events to the final statistical output to be disseminated (WPI). It is still a pilot study but with a clear aim at a production framework, not just showing the potentiality of these data for official statistics, which is beyond doubt.

The research in both projects has been and is still strongly affected by the access issue. In the first ESSnet, we were able to propose a sketch of the whole production process and make some initial partially disconnected proposals for different stages of this process (geolocation of network events, inference framework, quality issues,…). The lack of agreements in some cases and the restrictive conditions of these agreements in other cases prevented us to share data and investigate the process workflow from beginning to end.

The original approach was linear: to access the data, then to develop the methodology, to identify the necessary computer and technological tools to implement these methodological proposals, to produce statistical outputs thereof, and to investigate quality issues in these products. The restrictions upon the access the data put a severe limitation on the final outputs.

In the current ESSnet a modular approach has been chosen so that different aspects of the production framework are worked out in parallel. One of these modules is the generation of semi-synthetic mobile phone data as close as possible to real data in order to avoid the difficulties arising from the access issue. These synthetic datasets make it possible to investigate the different methodological proposals and even to assess the results with respect to synthetic complete target populations (something which we cannot do with real data). In this simulation exercise, the knowledge accumulated during the first ESSnet is vital to produce realistic datasets.

In the first ESSnet on Big Data, the reader may consult the results in the following deliverables:

- Deliverable WP5.1.- This deliverable collects a first analysis regarding the access and its status in the ESS as of the time of the project (2016).
- Deliverable WP5.2.- This deliverable collects different guidelines regarding the access to mobile phone data, divided into three big categories, namely (i) technical guidelines, where the reader may consult a description of the structure of the networks and the variables to be used for statistical purposes, (ii) business guidelines, where different considerations for the collaboration between MNOs and NSIs are offered, and (iii) guidelines based on ESSnet partners' own experience in dealing with access issues in contact with MNOs.
- Deliverable WP5.3.- This deliverable offers a first generic sketch of the production process with mobile networks data going from raw telco data to statistical microdata to aggregates and finally inferred estimates for the target population under study. Many of these ideas are used for the simulation module of the current project and to provide concrete methodological solutions in the ESS Reference Methodological Framework.
- Deliverable WP5.4.- This deliverable contains a description of the computer tools in development to implement the Reference Methodological Framework. This software is the basis for the current ESSnet.
- Deliverable WP5.5.- This deliverable focuses both on quality issues and on future lines of actions, many of which are under implementation in the current ESSnet.

Regarding this second ESSnet, as of time of this writing, the reader may consult the different modules and the ongoing work in the different research tracks in the working area of the project wiki page. The modules focus on (i) access, (ii) simulation, (iii) methodology, (iv) information technologies and computer tools, (v) standards and metadata, (vi) quality, (vii) application on real data, and (viii) visualization. Each of these modules will produce a deliverable by the end of the project.

Raw data to statistical data

The whole production framework under construction going from the raw telco data to the final statistical products is represented in Figure Figure_telco1. In this figure you can see the three layers of the production framework, namely the data layer, the convergence layer, and the statistical layer. Basically, the data layer comprises all process steps dealing with raw telco data, thus with a very high technological content (GSM, UMTS; LTE,…) and specific methods to preprocess this data, especially for positioning methods (point methods, tracking, etc.). The statistical layer, on the contrary, deals only with data with statistical content. Among other things, in this layer we need to connect our dataset with the target population at stake (i.e. the inference problem). In between these two layers, the convergence layer needs to apply different algorithms and combine data from diverse sources to detach the technological substratum to produce the statistical datasets.

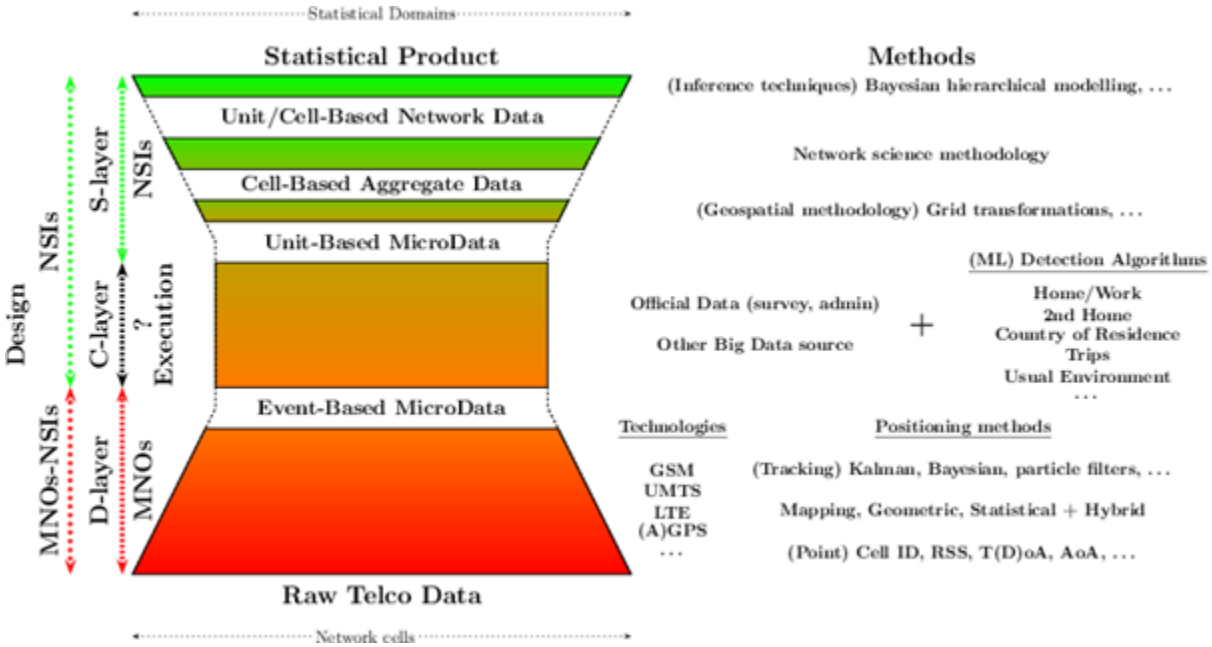

Figure 7 Representation of the production framework from raw telco data to the final statistical product

The whole proposed process is depicted in Figure Figure_telco2, where a collection of concrete production steps is assigned to the different layers. Each production step is characterised by input data (in), a process, and output data (out). The collection of production steps is explained and detailed elsewhere. Here we summarise the initial stage transforming the raw telco data described above into intermediate microdata ready for the upper-level statistical processing.

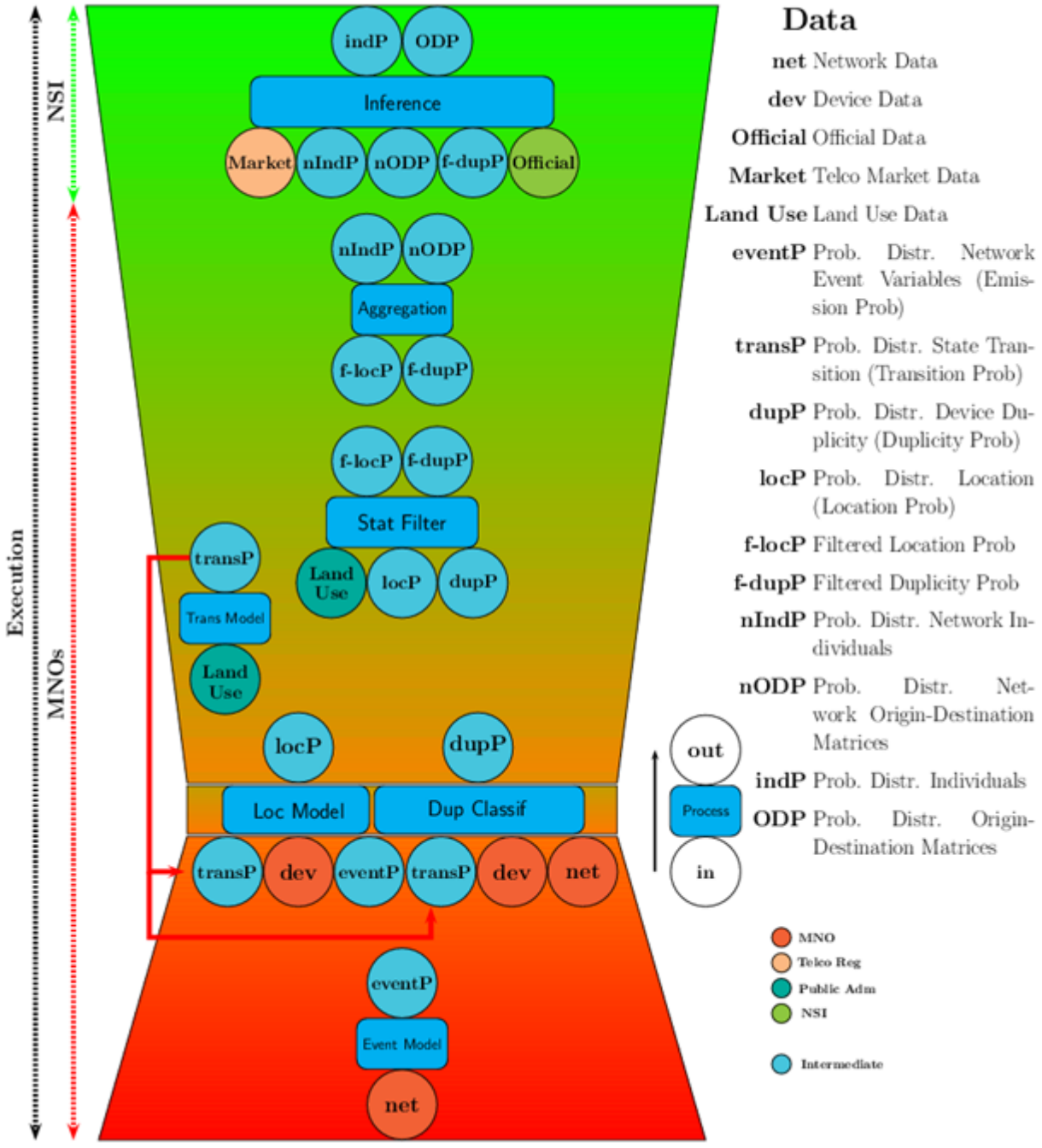

Figure 8 Illustration of the production process including production steps for MNO data

We are focusing on the geolocation information provided by telecommunication networks. The key idea is to translate this spatiotemporal information from the telecommunication network into a reference grid (e.g. the INSPIRE grid) constructing thus a random variable for each mobile device providing the probability distribution for it to be located at each tile of the grid at different time instants together with a probability for each device d to correspond to the same individual (2:1 device-individual correspondence). The computational procedure is highly modular to adapt to the diverse implementation conditions and agreements between MNOs and NSIs. We need to consider several sets of input variables:

- Let $\mathbf{E}_d(t)$ denote the set of variables in the network regarding a network event (call, SMS, data connection, location area update, …) of a mobile device *d* at time *t*. These variables can be the cell identity in which the device *d* has been detected at time *t* because of a phone call or an SMS or an Internet connection or any other technical circumstances registered by the network. They can also be the signal strength or the timing advance or even a combination of all this information. This generality is introduced on purpose to embrace all potential situations arising both from the technology and the concrete agreements between MNOs and NSIs. This is identified as **dev** in Figure Figure_telco2.
- Let $\mathbf{I}_{\mathbf{Net}}(t)$ denote the set of auxiliary variables regarding the parameterization of the network (antenna position, antenna parameterization, etc.). This is identified as **net** in Figure Figure_telco2. The concurrence of MNOs is compulsory to use this data.
- Let $\mathbf{I}_{\mathbf{Land}}(t)$ denote the set of auxiliary variables regarding the land use (transport networks, residential space, etc.). This is identified as **LandUse** in Figure Figure_telco2. Notice that this information is available to NSIs without the concurrence of MNOs.
- Let $\mathbf{I}_{\mathbf{Official}}(t)$ denote the set of auxiliary variables regarding the sociodemographic information from official figures. This is identified as **Official** in Figure Figure_telco2. Notice that this information is available to NSIs without the concurrence of MNOs (indeed produced by them probably with higher degree of detail than disseminated figures).
- Let $\mathbf{I}_{\mathbf{Market}}(t)$ denote the set of auxiliary variables regarding the telco market (penetration rates, market shares, etc.). This is identified as **Market** in Figure Figure_telco2. This information can be provided, in principle, by national telco regulators, although the collaboration with MNOs is advisable.

The core of the target intermediate data comprises two elements:

- Location probabilities: Let $T_d(t)$ be the random variable for the location of mobile device d being in the tile j of the grid at time instant *t*. This probability distribution is identified as **locP** in Figure Figure_telco2. By and large, this step also embraces the computation of joint location ($T_d(t)$, $T_d(t-1)$), which will be necessary for the inference stage.
- Duplicity probabilities: Let p_d be the probability for device d to have a 2:1 device-individual correspondence. This probability distribution is identified as **dupP** in Figure Figure_telco2.

Notice that some of these datasets (**LandUse**) only appear in the statistical layer, since there is no strict need to get MNOs involved in their processing. However, the computation of location and duplicity probabilities are recommended to be undertaken in a collaborative way. Figure Figure_telco2 clearly distinguishes between location and duplicity probabilities computation. In practice, these two business functions may be integrated. The computation of location probabilities is further broken down into the construction of a so-called emission model (producing emission probabilities $P(\mathbf{E_d}|T_d(t), \mathbf{I_{Net}})$) and of a so-called transition model (producing transition probabilities $P(T_d(t)|T_d(t-1), \mathbf{I_{land}})$). These two together with the information from each device produce the location probabilities. Details are given elsewhere (see Work Package I). It is important to underline the modularity and evolvability of this approach.

With these intermediate data, we can begin to produce more intermediate data for different statistical domains (demography, tourism,…) and then conduct purely statistical processing detached from the underlying technology to produce final estimates. Basically different modular processes are to be identified such as the identification of the devices related to the target population within this

data set (e.g. domestic tourists, commuters, …), the aggregation (whose output is the number of individuals detected by the network), the inference (producing estimates for the whole target population beyond the observed dataset), the quality indicators (e.g. assessing the accuracy of the estimates), etc. We underline that the current methodological proposals by work package I do not contain the number of devices as an intermediate dataset, but the number of individuals detected by the network. This is why the duplicity probabilities must be computed at the data-convergence interface. This will avoid several identifiability problems in the statistical layer to unravel the duplicities of mobile devices per individual.

These proposals are focusing on the geolocation information provided by telecommunication networks. This does exhaust the potentiality of this data source: information from internet traffic, and more far-reaching, from the interaction between individuals must be considered as soon as possible also to produce official statistics. The ESS Reference Methodological Framework is extendible to include this information, but more work than conducted in this ESSnet is needed.

Quality guidelines relevant for this big data class

Linking

We may consider the linkage between a mobile network dataset and an auxiliary dataset from both ethical/legal or statistical points of view. In the former case, linkage is directly related to the identifiability of individuals, which is intimately connected to the issue about the access to this data source. Mobile network data for statistical purposes are pseudo-anonymised, as described above, which avoids a direct identification of individuals in a population of interest, thus no direct record linkage with other auxiliary data sets is possible. In principle, linkage at the microdata (device) level is not possible and currently not considered as a resource to build the final statistical products.

We shall focus on the statistical point of view in order to assess the consequences on the production of official statistics. After pre-processing and aggregation (e.g. number of domestic tourists detected by the network at a given territorial area in a given time interval) the resulting aggregate dataset may be potentially combined with other aggregate auxiliary dataset (e.g. land use dataset per these territorial areas). This data integration is domain-specific and will depend on the objectives of the analysis (tourism, demography, labour market,…). In principle, this integration is statistically possible and even recommendable, but it is also related to the issue about the access to the data (e.g. integrating datasets from two different MNOs). As of this writing, work is in progress, but the natural linking variables are the spatiotemporal variables (grid tile and time interval – see above).

Guidelines

> To link MNO data at an aggregate level, check both territorial and time identifiers in the dataset(s). These identifiers must be linking variables with any auxiliary dataset providing more variables for the analysis.

Coverage

Beyond the direct concept of coverage for the input (raw telco) data, this notion becomes a subtle issue in relation to the potential target populations under study and the target estimates. We can always think of the natural notion of coverage as basically the fraction of human population using

any kind of mobile telecommunication technology detected and registered through a network. In this way, the official figures about penetration rates and market shares can straightforwardly illustrate a very first concept of coverage for this data. The trend is obvious and in the future more and more citizens will carry one or two mobile devices providing data through a telecommunication network. You can easily find more devices than individuals in many territorial cells in a country. This is a direct expression of the utility of this data source for official statistics, but also a methodological challenge for the reliability of statistical outputs thereof.

However, a more detailed concept of coverage arises when considering both target populations and target variables. On the one hand, regarding target populations the assessment of coverage is not a closed issue. Let us consider inbound tourism statistics. In principle, roaming information from the network can make us think that we can easily estimate the number of foreign visitors in a country. However, details about roaming agreements among MNOs show that a given visitor can easily change from one hosting network during his/her stay in a foreign country, thus making the estimation process more complex. When considering several MNOs at the same time, this change of hosting network makes us lose trace of each visiting individual, thus introducing severe methodological complexities in the process. These difficulties in the relationship between mobile network datasets and different target populations (present population, commuters, domestic tourists, etc.) are always present and will be present as we identify more and more target populations upon which we can apply these analyses.

On the other hand, mobile network data are rich enough to enable us to conduct network science analyses. We can produce both population totals (at different scales of geographical and time breakdown) and more interestingly interactions between population units (mobile devices, thus individuals). At this point, the concept of coverage is also important, although we do not have an answer yet: are the interactions among mobile subscribers representative of interactions among individuals in general in the target population under study? The issue is important and complex. Think e.g. of a present population and the interactions between children and adults or between children and elderly people. Are these interactions representatively detected in a mobile telecommunication network? This needs further study, but the potentiality is huge if the interactions in a given dataset can reliably represent the phenomena in the whole target population.

Guidelines

Compare with proxy rates and aggregates

Coverage cannot be directly assessed, the following highly relevant proxies must be collected either from the MNOs themselves or the corresponding National Telco Regulator:

- Penetration rates with the highest possible territorial and time breakdown.
- Market shares with the highest possible territorial and time breakdown.
- Roaming volume allocation among MNOs (number of subscribers per nationality with breakdown per territorial cells and time).

Comparability over time

Undeniably one of the quality issues regarding the production of official statistics is the sustainability over time of a data source feeding any concrete set of official statistics. Mobile telecommunication technology seems to be extended in the future and in this sense, there appears to be no further risks of losing this data source. However, the key issue is indeed in the technology dependence. In other words, will technology change so abruptly as to make official statistics produced thereof not comparable over time? We have already witnessed a strong evolution in this line from 2G over 3G to 4G and imminently to 5G. Beyond doubt, this affects the quality of data: the more recent the technology, the higher amount of information we have.

It is at this point where the central idea of the ESS Reference Methodological Framework needs to play its role by making the statistical analysis layer as much independent as possible from the raw data layer. The generic approach formulated above aims at ameliorating this dependence as much as possible so that methodologically the statistical microdata produced entering the statistical layer will be comparable over time. However, the semantic content will be somewhat affected: the potential for the computation of the event locations will be higher for modern technologies than for old technologies (e.g. more precision in the event geolocation computation), but the concept itself of event location will be the same.

Guidelines

Although the ESS RMF aims at decoupling technological (bottom) and statistical (top) production layers, the following information must be collected to assess comparability over time:

Audit technology updates
Be aware of changes in technology.

Audit spatiotemporal disaggregation
Be aware of changes in time frequency or spatial geolocation of network events. Spatiotemporal profile of events must be monitored.

Audit data provenance
Be aware of the origin of data generation, i.e. data generated by conscious behaviour of subscribers (making calls, sending SMS, connecting to Internet, etc.) or unconscious behaviour (people wandering while network detects the displacements for optimal service).

Model errors, measurement errors and processing errors / data source specific errors

Raw telco data is generated automatically by a complex information system and a complex transformation process needs to be applied. In this sense, the distinction between measurement error, model error and process error becomes much subtler, hence our joint treatment of all errors simultaneously.

We revise potential errors following the description of variables conducted above:

1. Pseudonymised ID.- A priori, beyond technological contingencies affecting the network (the system goes down, technical failures, etc.), there seems to be no errors in assigning a pseudonymised ID to each mobile device. Thus, from the technological point of view, no further considerations need to be made. From the statistical point of view, the limitations arise from the sustainability over time of each ID for each mobile device, firstly due to legal restrictions (in some countries this ID needs to be changed e.g. every 24h, which puts severe limitations on the analyses to be performed, at least, in longitudinal studies) and secondly due to reuse of mobile phone numbers by different subscribers. This must be clarified with the MNOs when reaching an agreement for the access/use of data. Phone number portability however is ameliorating this effect.
2. Time attributes.- From both a technological and a statistical point of view there seems to be no issues regarding the time attribute assignment to each network event. The system registers this variable which is ready for statistical purposes with no further transformation. However, it may be the case that depending on the agreement between the MNO and the NSI some coarse-graining procedure is applied so that a stricter disclosure control is applied (possibly depending on the legislation and on request by the national DPA). These potential disclosure control measures need to be clearly stated so that the upper-level statistical analyses can take them into account (in the computation of the location probabilities) to reach reliable final outputs.
3. Spatial attributes.- The situation with spatial attributes is much more complex because the network is not geolocating with the current technologies every single device for its operation. Indeed, this is the ultimate reason why we focus on a probabilistic concept of event location. At this point, the distinction between measurement error, model error, and process error becomes much fainter. We assess the capacity of the network to provide the statistical variable of physical geolocation of a network event. Firstly, the physical geolocation of a network event in the operation conditions of the network we are considering only enables us to reach a given amount of precision, i.e. we never determine the exact position of a mobile device in a map. Secondly, the raw telco variables used to construct our statistical notion of geolocation depends very sensitively on the raw telco data which we can access/use. This depends on the agreements with the MNOs and the investment on technology for statistical analyses carried out by the companies. For example, with CDRs/DDRs, only the radio cell ID detecting the event is used. With signalling data, timing advance and/or signal strength can be used to provide a finer estimation of the geolocation. Thus, an overall assessment of errors cannot be undertaken at this moment. However, we must be conscious that raw telco data may provide misleading information. Let us consider handovers, i.e. those situations in which an antenna close to the device cannot provide the connection service, and this connection is moved to another antenna located further way. The Cell ID as raw geolocation variable would be clearly misleading. This is equivalent to a classical measurement error. Ultimately, all these considerations must be taken into account in the computation of event locations. Thirdly, even when having precise raw telco data, since we are computing a probabilistic notion of geolocation, we need to make some modelling hypotheses. Thus, in the ESS Reference Methodological Framework which we are proposing, modelling errors are present from the very beginning, as in any statistical modelling exercise (this clearly departs from the traditional survey methodology). As part of this framework, we also provide model assessment indicators to evaluate this misspecification. The hypotheses underlying any computation of both the priors and the likelihoods must be clearly stated and subsequent

assessment indicators need also to be computed. The exact figures of merit are still under study, although some proposals can already be found in the ESSnet on Big Data I.

4. Event information.- The situation regarding the complementary variables about the events (duration, type, etc.) is similar to that about spatial attributes but with much lower degree of complexity. Firstly, this information will depend on the concrete raw telco data to be used and the agreements between MNOs and NSIs. From a technological point of view, unless some technical failure or other contingency is present, no further consideration is needed: the actual conditions must be taken into account in the computation of location probabilities. From a statistical standpoint, however, everything again depends on the modelling assumptions. Let us consider for instance the case of domestic tourists. This event information may be used to find patterns in the dataset enabling us to discern when an individual can be considered a domestic tourist or a commuter or a common individual not possibly being considered a tourist (even e.g. the same person in different seasons of the year). To make use of this information, we need to make some modelling assumptions, as with the spatial attributes. The same considerations apply here.
5. Complementary variables.- Again the situation is similar for any kind of complementary variable collected by the MNO (socio-demographic profile of subscribers, etc.). No further considerations need to be made regarding the technological aspects of the generation of these variables. However, their usage in the statistical analyses (hence its semantic content) will depend on underlying modelling assumptions. There are situations in which this information provide a more limited insight into the statistical analysis (e.g. personal contracts vs. enterprise contracts). Again, transparency regarding the modelling hypotheses and model assessment indicators thereof are mandatory.

Guidelines

Compare with auxiliary information
A comparison with penetration rates and market shares collected above must be carried out.

Monitor spatiotemporal disaggregation
Changes in time frequency of events must be monitored. Changes in the geolocation level of events must be monitored.

Monitor event variables
Changes in the number of variables for events must be controlled (event duration, event type, …). Empirical distribution of these variables must be monitored to detect uncontrolled changes.

Monitor complementary variables
Changes in the number of complementary variables must be controlled (event duration, event type, …). Empirical distribution of these variables must be monitored to detect uncontrolled changes.

Smart meter

Description

The class “Smart meter” consists of data collections related to metering information in general, mostly related to consumption or production of energy or the usage of water. This metering is done in short time intervals and the measurements can be read from a distance automatically or is transmitted automatically after a given time interval. In general the data can include inflow and outflow of a specific unit (metering point) in a certain interval. The most common occurrence of smart meters is smart electricity meters measuring the consumption and (most of the time) also production of electrical power.

The structure of the raw data is very simplistic:

- Metering ID
- Timestamp
- Inflow
- Outflow

Often additional background variables are available per metering ID which will be used to transform metering IDs into useful statistical units. These background variables may consist of:

- a unique identifier to link the metering point to administrative units or to a building or dwelling, e.g. business ID,
- geographical information which may be as precise as an address or a coordinate or just a geographical unit, e.g. district
- information on the kind of metering point, e.g. household, business, producer.

The role of this big data class in the project

This big data class was the main input for WP 3 of ESSnet Big Data I and is now handled in WPD Smart energy. The output of the previous ESSnet delivered insights in the area of data access, data handling, the production of statistics and methodology, future perspective and recommendations.

An obvious statistical output is the consumption of electricity and identifying specific patterns of consumption might be of additional interest, e.g. to find inherent socio-demographic factors to explain "energy-saving" and "energy-wasting" households. The energy consumption of businesses could be related to business cycle effects and it could therefore be used as an auxiliary variable in estimating economic indicators. Construction sites of new buildings, discrimination into vacant/non-vacant homes as well as prices/spending statistics are additional possible output.

Raw data to statistical data

The process of transforming raw metering data into meaningful statistical data can be split into the following sub processes:

- Linking metering point to statistical units, e.g. to register data about businesses or households. The linkage might be done based on a unique identifier, the address string or the coordinate of a metering point, depending on the available information.
- Non linkable metering points need to be classified according to the necessary classification not present in the background information, e.g., household / business.

Quality guidelines relevant for this big data class

Linking

Linking with other data sources gives additional knowledge and enables to produce more varied outputs. From the smart meter data itself it might not be possible to distinguish between different sectors of industry or building/household characteristics.

Guidelines

Make use of unique keys whenever possible.
Depending on the situation in the different countries, unique identifiers might be available and they should be used preferential compared to other methods, e.g. text based record linkage methods.

Develop a sequence of linking operations.
First use direct identifiers whenever possible.
Second apply unique linkage based on the textual information, e.g. addresses.
Third apply probabilistic record linkage methods.
As a last resort statistical matching could be applied if enough background information is available.

Check the quality of probabilistic record linkage on a sample.
When non deterministic record linkage methods are applied, a sample of units should be audited.

Coverage and Accuracy

The accuracy of the measurement itself is probably of limited relevance to the overall quality of statistical output as it is used to invoice customers, but checks should be in place to check for extreme and implausible values.

Guidelines

Establish a series of basic checks for the measurements.
All measurements should be checked for basic plausibility, such as non negative values, values below a suitable upper boundary depending on the type of metering point.

Currently, coverage of smart electricity meters is increasing Europe-wide, but the roll-out is still on-going and therefore undercoverage is present and has to be handled accordingly.

There are no known examples of overcoverage in the raw data, but there can be artificial overcoverage due to linking or classification errors.

Guidelines

Research smart meter adoption rates.
The current rate of deployed smart meters should be observed and also checked on a regional level.

Compare consumption data on macro level.
In some cases data on a macro level, e.g. energy consumption of all household or all households per region, is available from surveys. This data could be compared with the corresponding aggregates from the smart meter data.

Establish final electricity consumption.
The data might include smart meters of grid points, where no actually consumption takes place. These devices should be identified in the data and for most use cases they should be deleted.

Comparability over time

Once the setup is complete, and the NSIs have negotiated long-term contracts with the data owner(s), this should be a fairly stable data source.

Guidelines

Continuously monitor smart meter adoption rates.
The development of the rate of deployed smart meters should be observed and also checked on a regional level.

Model errors

Estimated classification such as the division into household or business might suffer from poor accuracy and therefore, all used models should be checked thoroughly and confusion matrices should be estimated.

Guidelines

Make an audit sample of classified units for quality assessment.
If possible a sample of the classified metering points could be followed up manually to check if the selected classification seems suitable.

Process errors / data source specific errors

A very specific source of error for consumption and production of electricity estimated based on smart electricity meters is the own consumption of produced energy, since this is not recorded by most smart meters.

Because of the use of models to estimate some of the important characteristics, it is important to have good training data which is not easy to get for some indicators, e.g. vacant/non-vacant dwellings.

Guidelines

Make an assessment about the consumption of energy produced directly on site.
Either the definition has to be stated clearly that this kind of energy consumption is excluded or attempts to estimate it have to be made and later added to obtained results.

Assess quality of estimation of vacant/non-vacant.
Especially, if no training data is available for supervised learning, a quality assessment of the estimation should be done.

Earth observation (satellite images)

Description

The class "Earth observation" is related to processing satellite images. It has been classified by UNECE in the following group

Internet of Things (machine-generated data): derived from the phenomenal growth in the number of sensors and machines used to measure and record the events and situations in the physical world. The output of these sensors is machine-generated data, and from simple sensor records to complex computer logs, it is well structured. As sensors proliferate and data volumes grow, it is becoming an increasingly important component of the information stored and processed by many businesses. Its well-structured nature is suitable for computer processing, but its size and speed is beyond traditional approaches.

3.1 Data from sensors /category/

3.12 Mobile sensors (tracking) /subcategory/

3.123 Satellite images /group/

Therefore, *satellite images* is a group within *mobile sensors (tracking)* subcategory of the wider category *data from sensors*. It is worth to note that other two groups in this subcategory are *mobile phone location* and *cars*. Because of the fact, that data is generated automatically by dedicated systems, all aforementioned types were classified by UNECE as *machine generated-data*, named also as *Internet of Things*.

Data sources includes both radar (Sentinel-1) and optical (Sentinel-2) data. Data can be acquired from the European Copernicus program (https://sentinel.esa.int).

Two examples of raw data are presented in Figure Figure_satellite1 (radar data) and Figure Figure_satellite2 (optical data),
(source https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/0/04/WP7_Deliverable_7_7_2018_05_31.pdf, https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/0/04/WP7_Deliverable_7_7_2018_05_31.pdf).

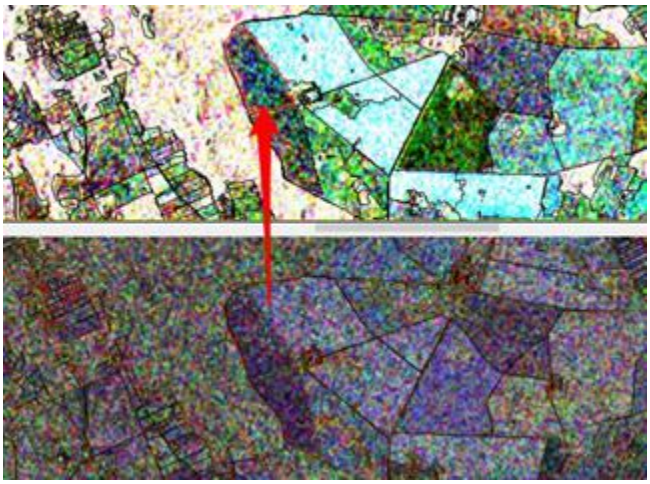

Figure 9 Example of radar data, including results of segmentation

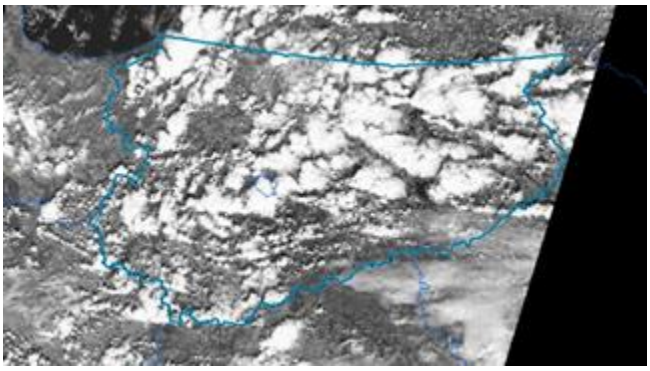

Figure 10 Example of optical data including cloud cover

Earth observation class is an example of the use of machine learning algorithms to provide reliable data on different objects identified from the images.

Raw data to statistical data

The role of the big data class "Earth observation" in the ESSNet is related to various use cases conducted by NSIs to acquire the data directly from the Internet. It includes especially crop types identification, which supplement the current agricultural surveys. Therefore, according to the classification, the statistical domain is Agriculture and Fisheries and the surveys supplemented are Annual Crop Statistics.

The process of crop identification relies on several processes, starting with integration of administrative data, field surveys and satellite images. It includes the following steps:

- insitu sample collection: ground truth as training input data for machine learning and accuracy assessment,
- administrative data collection: support data used for insitu,
- plots selection and raster data segmentation,
- remote sensing data processing: Sentinel 1 and Sentinel 2 data processing, creating time series SAR and optical raster orthomosaics,
- data fusion: fusion of raster and vector datasets,
- image segmentation: extracting segments sharing similar spectral characteristics, input for object based image classification,
- object based image classification: learning classifier based on insitu learn sample, SAR and optical image classification based on machine learning algorithms,
- accuracy assessment: computing confusion matrix based on insitu control sample.

Quality guidelines relevant for this big data class

Coverage and accuracy

Accuracy is the major measure because of the specification of algorithms used that are using machine learning algorithms. For example, accuracy was measured by the number of fields with crop types identified correctly and varies from 75% to 85% depending on the crop type and machine learning algorithm used (KNN and SVM are the most accurate).

The following factors can be considered when assessing accuracy:

1. Total Survey Error approach for analysing accuracy; including in particular (but not restricted to), over-coverage, under-coverage, selectivity, missing data (non-observation and non-response), adjustments made to the data and the presence of anomalies.
2. Reference datasets: Many analyses require the use of reference data sets due to respondent related error or instrument generated error.
3. Selectivity: Imperfections in coverage.

The potential indicators that can be constructed just to estimate the possibility of creating a good training dataset:

1. If a reference data set is available, assess coverage error. For example, measures of distance between the big data population and the target population (e.g. Kolmogorov-Smirnov Index, Index of dissimilarity)
2. Does the file contain duplicates?
3. Are the data values within the acceptable range?
4. Assessment (also qualitative) of sub-populations that are known to be under/over-represented or totally excluded by big data source.
5. Assessment of spatial distribution of measurement instrument and of periodicity of observations
6. Selectivity: Derive R-index for unit composition

Guidelines

In order to have a high accuracy it is necessary to prepare a training dataset with a large number of observations. The first step is to prepare and classify manually different satellite images with the label relevant for the picture, e.g. “wheat crop type”. The second step is to deliver images from different fields (different angles, different seasons, no clouds) to fit labels needed. The next step is to test the dataset. If the accuracy is too low (i.e. below 80%), then the training dataset must be extended. There should be the same number of observations for a specific label/class.

Guidelines

According to the Sentinel dataset, the coverage is the territory of European countries. However, because of cloudy weather, there may be some missing data in most cloudy months (e.g., February). Land can also be covered by snow which makes it impossible to make any analysis during snowy days in wintertime.

Comparability over time

For Earth observation comparability over time is not an issue, since the data source is stable and will be available also in the long term. Usually this quality dimension is referred to sustainability of the data sources. Therefore, sustainability through time factors (internal and external) which could affect the sustainability of the data provider’s data in relation to the NSI requirements. If the data provider will not be available, will similar data providers or comparable data sources be available in the future? The following indicators can be used to measure the sustainability of the data source.

1. What is your estimate of the overall risk that the data provider will not meet the quality requirements of the NSI?
2. What is the risk that the big data source will not be available from the data provider in the future? If it will not, will there be comparable data sources in the future?
3. How relevant is the data, what would happen if it is available for only a short period of time?
4. How long does this data need to be available to be relevant?
5. Is it likely that we can replace this data with similar (or next generation) data, once the data source or technology becomes obsolete?

Guidelines

Be sure to use the same data sources for a long time. Collect the data and process it to acquire time series. Because of different views on pictures in different seasons (e.g. crop types), it is necessary to prepare training datasets for the specific month/week of the year. That is why it is possible to change the training dataset over time. Otherwise, the dataset may be not comparable.

Process errors / data source specific errors

Because of the use of machine learning algorithms, it is very important to have a stable and reliable fundamental sources to compare. Therefore, there is a need to compare the results with a reliable data source, such as in-situ survey.

Guidelines

Four different indicators can be used to evaluate process errors:

1. the training fields classification error matrix,
2. the calculations accuracy,
3. comparisons with administrative data,
4. comparisons with survey data.

AIS (tracking ships)

Description

AIS provides data about ship position, speed, ship ID, headings etc. It was originally developed to ensure safety at sea. Based on Global Positioning System (GPS) technology, the AIS system of a vessel broadcasts its location and status information over a radio channel, making it possible to be detected by other ships wherever they are, as long as a radio signal can be sent and received. This data is very useful for maritime and transport statistics but also provides information about economic processes.

The role of this big data class in the project

In the first ESSnet Big Data, the structure and information content of AIS data has been studied including potential applications (WP4 Del 1). For the latter deriving port visits, linking the data to other data sources and deriving sea traffic routes were found essential (WP4 Del 2, Del 3) to improve existing statistics and derive new ones. The advantage of using one AIS-dataset for the entire European territory is that it provides: a) a better comparison of international traffic between the countries, b) more synergy as all participating countries work on the same dataset c) reduced pricing. A disadvantage is that this data is stored by private companies and handling fees have to be paid compared to national or EMSA data. A number of new applications were tested in SGA2 of the ESSnet Big Data (WP4 Del 6, Del 7), examples are emissions, distances traveled within port and potential new economic indicators.

Raw data to statistical data

The AIS messages used in ESSnet Big Data I have a text encoded binary format that needs to be decoded prior to use (see the 6th field in Figure Figure_AIS), which has to be decoded. It is important to note that some messages are distributed over several lines. Line 3 and 4 in Figure Figure_AIS encode one message. You can see this because the first element after the AIVDM tag is a 2, denoting that the message is split into 2 parts. The element after that denotes which part of the message it is, so we can see that the third line is the first, and the 4th line is the second part. The two strings in the 6th column have to be concatenated before sending to the decoder.

Figure 2. Example of raw NMEA encoded AIS data

```
!AIVDM,1,1,,A,13aEOK?P00PD2wVMdLDRhgvL289?,0*26
!AIVDM,1,1,,B,16S`2cPP00a3UF6EKT@2:?vOr0S2,0*00
!AIVDM,2,1,9,B,53nFBv01SJ<thHp6220H4heHTf2222222222221?50:454o<`9QSlUDp,0*0
9
!AIVDM,2,2,9,B,888888888888880,2*2E

description of the NMEA format, in order:

!AIVDM - The NMEA message type: !AIVDM (received data from other vessels)
and !AIVDO (your own vessels information)
2 - Total numer of sentences needed to transfer the message (1 to 9)
2 - Sentence number (1 to 9)
9 - Sequential message ID (0 to 9)
B - AIS Channel (A or B)
888888888888880 - The Encoded AIS Data
2*- End of Data
2E - NMEA Checksum
```

Figure 11 Example of raw NMEA encoded AIS data

The decoding of the raw data is covered in more detail in the subchapter "Technological processes on premise - AIS".

Quality guidelines relevant for this big data class

Linking

Linking of AIS data with survey data was investigated for AIS data from Dirkzwager and survey data on maritime statistics from Poland (Puts et al 2016). Readers need to realize that the survey data does not provide the MMSI numbers used as unique identifiers in AIS data; they provide other unique numbers, so-called IMO number. To tackle this issue, a reference frame of ships for Poland was linked with data containing coordinates of ports. After that, the combined dataset was split into two groups: “ships in ports” and “ships not in ports”. This work verified that “ships in ports” based on the coordinates were indeed in Poland’s dataset from survey data.

Coverage

At the unit level, errors in the unique identifier of an AIS message may result in the creation of a new (non-existing) number for non-existing ships. Messages from these so-called ghost ships cannot be assigned to a unit in the population and hence need to be removed. Another coverage issue is the fact that non-maritime ships, which did not belong to the target population for maritime statistics such as fishing ships and yachts, also transmit AIS data. These ships need to be identified and excluded. Both unit-level issues were resolved by creating a population frame (based on all the AIS messages available), from which non-maritime ships and the non-existing vessels were removed. This frame, the so-called reference frame, formed the target population and was used in all subsequent studies. Another coverage issue was identified but could not be completely solved. This was the differentiation between maritime ships loading and/or unloading goods at the port and those that don’t. Only vessels that load or unload goods are of interest. However, this information is not available in the AIS messages. There are also small numbers of ships, such as tugs, that may or may

not carry goods from a vessel to a port. There is no way of telling this from AIS data itself, so there is a chance that ships that do not carry goods are included in the reference frame.

When a number of different AIS data sources were compared, it was found that they differ in coverage of ships in the European waters. National AIS data usually contains data on more ships and more data points per ship compared to the tested European and satellite AIS data. Satellite data was found to be a very good additional source of data when looking at a ship's journey.

For new data sources it is not always clear under which category potential errors should be classified. The listed problems would also fit into the category "measurement errors".

Guidelines

To enable use, raw AIS-messages first need to be decoded. Next it is important to keep in mind that:

- AIS is a radio signal, which means that parts of the messages can get lost or scrambled due to factors such as meteorology or magnetics.
- Messages are transmitted encoded. As a result, an error in one transmitted 'byte' can result in an error in one or multiple fields in the decrypted message. Most of the times, these errors are detectable as the result yields an invalid variable, but sometimes they yield valid variables. For instance, a pre-processed MMSI can be coincidentally technically valid, yet incorrect. These errors can arise for every variable, so this can for example result in erroneous latitude and longitude, yielding faulty locations that are quite far away from the actual location of the ship. If not filtered out, this can result in a very high journey distance of ship.
- Receivers have time-slots in which data is received. In busy areas with many ships, not all data from all ships may fit into this time slot. This may result in the loss of data on some ships in that time slot.
- Ships can turn off their AIS transponder, resulting in the disappearance of a ship.
- AIS was originally intended for safety at sea, to warn nearby ships. As it was not meant for producing statistics, the variables entered manually by the shippers are not always reliable.

Comparability over time

AIS data follows an international standard which ensures that the data structure is comparable over time (AIS, 2019).

Measurement errors

Errors in the geolocation coordinates of an AIS message results in ships appearing at impossible or illogical locations. For instance, in the European data set some of the ships were located in the Sahara desert. Obviously, this must be a data error. Another example of such an error is a ship suddenly bouncing a few kilometers back and forth during its journey. By carefully looking at the subsequent

locations of a ship during its journey or by noting that a ship had reached a port, the team was able to develop a cleaning procedure (a median filter) to remove these erroneous locations for vessels.

Model errors

Errors may be caused by the model used to derive a journey from a ship. However, the WP4 members found that individual journeys can vary greatly because of all kind of (unforeseen) intermediate stops some vessel take (Consten et al 2018b). Because of this, the median filter applied to derive a journey from 'jumpy' AIS location data is not expected to be a major issue (Consten 2018a).

Process errors

Since the AIS data used in the studies of WP4 was obtained from an external organization, i.e. DirkZwager, transparency and soundness of methods and processes for the metadata and the data are essential (Consten et al. 2017). This was insufficient for this data which indicates that any decisions made by DirkZwager employees during the processing of AIS data may affect the final results. Because raw AIS data was used in WP4 the major effect of processing errors most likely affects coverage; as is demonstrated by the difference in coverage of different sources of AIS data.

Literature

AIS (2019) AIS standard webpage. Available at: http://www.allaboutais.com/index.php/en/ais-standards (Accessed June 2019).

Consten, A., Chavdarov, V., Daas, P., Horvat, V., Maslankowski, J., Quaresma, S., Six, M., Tuoto, T. (2018a) Report describing the methodology of using Big Data for official statistics and the most important questions for future studies. ESSnet Big Data I, Workpackage 8, Deliverable 8.4.

Consten, A., Puts, M., de Wit, T., Bisioti, E., Pierrakou, C., Bilska, A., Bis., M., Grøndal, O., Langsrud, Ø. (2017) Report about sea traffic analyses using AIS-data. ESSnet Big Data I, Workpackage 8, Deliverable 4.3.

Consten, A., Puts, M., de Wit, T., Bisioti, E., Pierrakou, C., Bilska, A., Bis., M., Grøndal, O., Langsrud, Ø. (2018b) Report about possible new statistical output based on (European) AIS data. ESSnet Big Data I, Work package 4, Deliverable 4.7.

Kowarik, A., Stolze, P., Grøndal, O., Ilves, M., Kirt, T., Jannson, I., Wu, D. (2016) Report on Data access and Data handling. ESSnet Big Data I, Workpackage 3, Deliverable 3.1.

Puts, M., Grøndal, O., Pouwels, M., Consten, A., Perrakou, C., Langsrud, O. (2016) Creating a database with AIS data for official statistics: possibilities and pitfalls. ESSnet Big Data I, Work Package 4, Deliverable 4.1.

Web scraping (online job vacancies, enterprise characteristics)

Description

Web scraping – the automatic collection of data on the internet – is being used increasingly by NSIs to reduce the response burden, to speed up statistics, to derive new indicators, to explore

background variables or to characterise (sub-)populations. It has proven to be a valuable way to study the dynamics of a phenomenon before designing a new costly statistical production chain or to supplement administrative sources and metadata systems. In principle, web scraping comes down to running a program that automatically collects data from previously indicated web pages and, if needed, defined locations on these web pages. Because web pages are collected the data consists of part or complete html-coded web pages which provide structure in the form of html-tags. This may aid subsequent processing. It is based on the variety of methods to filter, process and analyze the data.

Sources of the data include all websites publicly available on the internet, accessed via web browser engines or via API (Application Programming Interface). Therefore, there are three basic types of web scrapers to access information from the Internet:

- crawlers which access directly web data and extract information based on HTML/XHTML meta tags,
- robots which access structured or semi-structured data in various formats (e.g., JSON, CSV),
- applications which are using API to transfer the data directly from the hosting server to the local machine.

Data is extracted from the raw websites by the use of CSS (Cascade Style Sheets) tags and classes. For instance, an example of a website raw data is presented in Figure Figure webscraping.

Figure 12 Example of a website raw data

According to Figure Figure webscraping, to extract information in the green boxes, we have to prepare a web scraping robot to extract all readable information from the tag “div” with class “news priority”.

The structure of output data can be delivered in various data formats: CSV, JSON, SQL-like tables etc.

The role of this big data class in the ESSnet

The role of the big data class “Web scraping” in the ESSnet is related to various use cases conducted by NSIs to deliver the data directly from the Internet.

It includes web scraping:

- job vacancies

- enterprise characteristics
- comments/news
- Twitter data
- prices of products and services
- tourism accommodation establishments
- border movements

The list of use cases is not limited to the ones described above. However, this list contains the most reliable examples for the use of web scraping in the ESSNet.

WP1 (job vacancies) and WP2 (enterprise characteristics) from the ESSnet I both used data and texts available on the internet to extract information to be used for official statistics. To enable the use of web pages, legal aspects and netiquette are important prerequisites (Stateva et al. 2017a). It is important to realize that in some European countries it is unclear if data collection via web scraping is legally allowed for NSIs. From the study of online vacancies it became clear that the online data is not representative of the overall labour market and that there are various definitional issues which make it difficult to compare statistical output from web scraping directly with existing official statistics (Swier 2018). For the study of enterprise characteristics, the results were much more promising, the overall impression was between good and excellent (Stateva et al. 2018). A number of promising experimental statistics were produced.

Raw data to statistical data

When scraping web pages the raw html-code needs to be converted to a format that can be handled by employees of the NSI. However, the internet changes rapidly which requires a vast amount of technical knowledge and experience. It all depends on the web source of interest. One has to continuously monitor the validity and usefulness of the tools chosen and the scraping mechanism applied. However, the main concept of automatically reading data from web sources has proven to be stable. Ten Bosch et al (2018) concludes that web scraping can be managed successfully by people with the right knowledge and experience in choosing the right tools based on the latest technologies. The logical reference architecture developed in WP2 can be used as a general guide for this task. In addition, one can distinguish between specific and generic scraping. The first focuses on the collection of specific variable values on web pages, such as prices of products, which requires prior knowledge of the structure of the web page, whereas the latter requires no prior knowledge as it collects the entire web page.

Since the majority of the data extracted from web pages is text based, this needs to be processed to enable its use for official statistics. The column "Data preparation" of Figure Figure_webscraping2 provides an overview of the often used steps.

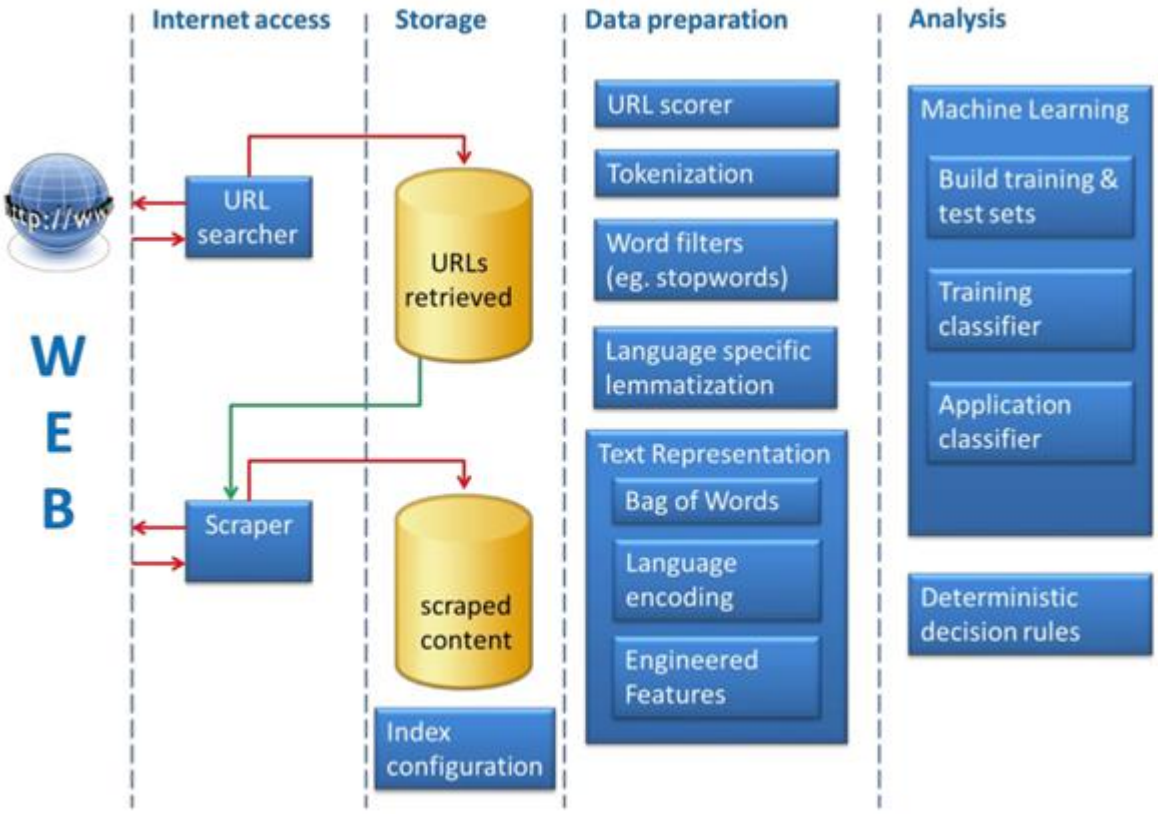

Figure 13 Production process for web scraped information

There are five general steps that should be used when transforming the raw data into statistical data:

1. Data acquisition
2. Pre-processing the raw dataset (including tag identification)
3. Processing data into machine readable format (including data cleansing and text mining methods)
4. Data evaluation and improving (including imputation of missing data/data linkability)
5. Data in usable format (e.g., CSV or JSON file)

Quality guidelines relevant for this big data class

Linking

For many use cases such as job vacancies and enterprise characteristics, it is essential to be able to link web-scraped information with a company in the business register. This is often not possible for all businesses. With the use of URL addresses, a list of websites can be retrieved and related to a given enterprise. Usually this list is obtained by querying a search engine on the web using the name of the enterprise as a search term. The underlying assumption is that - if an enterprise has an official website - this should be found within the results provided by the search engine. By comparing any company specific data displayed on the web page, such as a Chamber of Commerce number, it is be possible to check this relation. Using official documents of a company, such as an annual report, is another way of obtaining the appropriate URL of the website. For online job vacancies, several countries have explored the matching of online job ads with their own job vacancy survey micro data or business register data. The results have been rather mixed and vary from good linkage, when a

common identifier was present, to bad linkage results, when probabilistic linkage had to be used (Swier et al., 2018).

Guidelines

Find possible classifications that can be extracted from the website, e.g., date, territorial unit, sector as well as enterprise: business id, name, address etc.
When you are referring to official statistics data sources, it is important to start with official classifications (e.g. NTS, business ID) and prepare algorithms to extract this information from websites.

Coverage

Due to the fact that the sample of web data is in many cases unknown, coverage is a relevant issue we have to tackle during web scraping. According to the specification of the dataset, it includes various aspects of coverage. For example, web scraping Twitter data leads to the conclusion that the data is not representative for the whole population, because it is not used by everybody and there is usually under-coverage of different groups of individuals. When web scraping websites, we should be aware of the fact that not all enterprises are present in the web. Especially, small enterprises and self-employed persons may not be present in web or social media, which leads to under-coverage.

One of the fundamental quality issues in using online job advertisement is that not all job vacancies are advertised online. Hence, a considerable number of vacancies might be missed. Also some vacancies are listed more than once (e.g. on several different job portals) and some are listed but are not real vacancies (more in Fig. 3 in Swier 2016), so both over- as well as under-coverage is present. Combing data from several sources is the strategy recommended, but this does not solve all issues and increases deduplication work.

For enterprise web sites it is essential to obtain the corresponding URL of all companies in the Business register. However, not all businesses have a web site and some type of businesses have a higher chance of having a web site than others. This affects coverage considerably. By specifically searching for the website of a company via a search engine, this issue is attempted to be solved on a case by case basis. In job vacancies data coverage was measured at three different levels: micro-level, aggregation by industrial sector and advertising channels. All three were found to be selective in some aspects (Swier et al. 2018).

Guidelines

Representativeness: Try to estimate the population size and compare with traditional data
For example, when you are scraping enterprise characteristics, try to count the number of websites that are accessible and can be used for web scraping. Compare this number with the data from your business register.

Coverage: Relevant data available on website: Make a pilot web scraping to assess what information is included on the websites.
Check if specific information, e.g., territorial unit or industrial sector can be extracted from the website. When information on the website is limited, it is

also not very likely to monitor enterprise activity (e.g., innovations in enterprises) on the website. It is also important to monitor if the information is up to date and if changes over time can be identified (in longer time series).

Comparability over time

The internet changes continuously. Both the content as well as the structure of a web page may change. So on the one hand, web scraping may become increasingly challenging as the internet evolves, as more and more data may be encoded in forms that are harder to extract – audio or video files, for example - or because of the increasing use of interactive or user-specific content. It is therefore easy to imagine that websites of interest, such as those of some enterprises for instance in the creative industries, become much more challenging to extract information from. Such web-scraping may therefore need to be increasingly carried out by specialists inside or outside of NSIs, rather than only data scientists. On the other hand, when the content of web pages is continuously updated it is a challenge to assure that all data is scraped from a web page. For some web sites this may require a daily visits and puts a burden on data management. Time series analysis of online job vacancies and the JVS survey gave interesting results when focused on the overall development over time (Swier et al., 2018).

Guidelines

Check if the modification/update date can be extracted from the website.

When webscraping specific information from the website (e.g. job vacancies), try to extract the data of publishing this information.

When the website is not up to date it is unlikely to detect enterprise activity in longer time series.

Measurement Errors

The concepts measured (or derived) by scraping web pages may not be identical to the ones required by the NSI. This may cause measurement errors. An example of this is found in online vacancies were classification of jobs was extremely difficult because different web sites used different standards producing errors. Similar errors occur for important concepts such as occupation and economic activity. For these reasons, NSIs need to consider what the best use of their time and resources is. NSIs should consider options that will minimize the amount of data handling required to reduce these errors.

Guidelines

Verify if the data in the web fits the definition from official statistics.
It should be noted that sometimes the same data may have different definition. For example, online job vacancies data cannot be used as the official statistics data on demand on labour market.

Model Errors

The non-representativity of web scraped data introduces bias in the estimates. Some of the pilots in the ESSnet have highlighted the issue of bias – for example, they were more likely to identify websites for businesses conducting e-commerce. Thus, e-commerce statistics purely based on data scraped from the websites found would be biased without adjustment. They also found that some websites – for example, those that make heavy use of Javascript – are harder to scrape than others, and this may also introduce bias. A simpler source of bias, for some statistics, will simply be that some businesses are less likely to have a website than others.

A key challenge will be to understand these biases for different use-cases and work out how to adjust for them so that scraped data can be used for estimates qualifying as official statistics. This is likely to involve methodological work around combining web-scraped data and survey or administrative data, potentially picking up on some of the methods that are being developed as a part of WP1 (Stateva et al., 2017b).

In a number of projects, machine learning methods were applied and compared with the more commonly used deterministic approaches in official statistics. An important lesson was that with different methods similar results could be achieved and that some findings even converged (Stateva et al., 2018b). One of the factors that influenced the choice between deterministic and machine learning methods is the complexity of the relationship between the input variables or the features derived from the input data and the statistical target variables. If this relationship is fairly straightforward a deterministic approach seems obvious. But if this relationship is complex, unknown or difficult to model in a deterministic algorithm, which might well be the case when working with web data, a machine learning approach might perform better. For the latter, generalization and overfitting were found to be import quality issues.

Process Errors / data source specific errors

As mentioned above, the biggest issue is a continuous change of the internet. It means that we should monitor changes on the internet and modify the web scraping software if necessary.

For both online vacancies as for enterprise characteristics a process pipe-line was strived for. Not every partner succeeded but for online job vacancies some came close (Swier et al., 2018) whereas for enterprise websites nearly all succeeded (Stateva et al., 2018b). 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.

Literature

Consten, A., Chavdarov, V., Daas, P., Horvat, V., Maslankowski, J., Quaresma, S., Six, M., Tuoto, T. (2018a) Report describing the methodology of using Big Data for official statistics and the most important questions for future studies. ESSnet Big Data I, Workpackage 8, Deliverable 8.4.

Stateva, G., ten Bosch, O., Maślankowski, J., Righi, A., Scannapieco, M., Greenaway, M., Swier, N., Jansson, I., Wu, D. (2017a) Legal aspects related to Web scraping of Enterprise Web Sites. ESSnet Big Data I, Workpackage 2, Deliverable 2.1.

Stateva, G., ten Bosch, O., Maślankowski, J., Barcaroli, G., Scannapieco, M., Greenaway, M., Jansson, I., Wu, D. (2017b) Methodological and IT Issues and Solutions. ESSnet Big Data I, Workpackage 2, Deliverable 2.4.

Swier, N., Jansson, I., Wu, D., Nikic, B., Pierrakou, C., Körner, T., Rengers, M. (2016) Interim Technical Report. ESSnet Big Data I, Workpackage 1, Deliverable 1.2.

Swier, N., Hajnovic, F., Declite, T., Rengers, M., Islan, C-G., Jansson, I., Wu, D., Elezovic, S., Crahonja, C. Pierrakou, C., Biotti, E., Bergat., M., Eidelman, A., Alves, R., Fernandes, M-J. (2018) Final Technical Report. ESSnet Big Data I, Workpackage 1, Deliverable 2.2.

Ten Bosch, O., Windmeijer, D., van Delden, A., van den Heuvel, G. (2018) Web scraping meets survey design: combining forces. Paper for the BigSurv18 Conference, Barcelona, Spain.

Social Media

Description

The UNECE Task Team on Big Data, in June 2013, classified, among others, **Social Networks (human-sourced information),** under this umbrella they intend all the collection of human experiences digitized and stored everywhere from personal computers to social networks. This data is loosely structured and often ungoverned, and include

1. Social Networks: Facebook, Twitter, Tumblr etc.
2. Blogs and comments
3. Personal documents
4. Pictures: Instagram, Flickr, Picasa etc.
5. Videos: Youtube etc.
6. Internet searches
7. Mobile data content: text messages
8. User-generated maps
9. E-Mail

The use of social networks is strongly related to characteristics independent from the NSI's government: some of them like Twitter and Youtube can be accessed somehow without subscribing an account, Facebook and Instagram can only be accessed by subscribers, employees' company's e-mails can be accessed by the company in some countries (e.g. India) whereas this is definitively not

acceptable in many European countries. User-generated and internet search (e.g. Google trends) are somehow publicly accessible.

In this section, we mainly refer to publicly available social media, mainly reporting experiences considering Twitter as case study, due to the largest availability.

The role of this big data class in the project

In the previous ESSnet on big data, SGA-2, the use case "Population" in the work package 7 "Multi-Domains" was dedicated to show the structure of population in different regions according to the specific facts – e.g., public opinion on a topic (in the pilot Brexit) and life satisfaction in different regions, by means of social networks as data source. Therefore, three examples have been conducted. One of them was to identify the scale of depression in different countries based on Google Trends. The second was to find social mood according to public events or facts (e.g., Brexit). The goal of the third example was to identify life satisfaction in the population according to their comments/posts/tweets.

Raw data to statistical data

Let us consider Twitter as social network source. Twitter's Streaming API is used to collect samples of public tweets. The tweets sample can be filtered according to relevant keywords. The sampling algorithm is entirely controlled by Twitter's Streaming API and very little is known about it. The API is allowed to return *at most* a 1% sample of all the tweets produced on Twitter at a given time. When a filter is specified, the API returns *all* the tweets matching the request up to the "1% of all tweets" limit.

There would exist, however, the possibility to buy from the Twitter company different samples of tweets.

The sampled tweets need to be processed, somehow cleaned and normalized, and then the target information needs to be extracted. Actually, even when a filter is applied, the observed tweets represent only a fuzzy representation of the phenomenon of interest. To extract the relevant information, usually natural language processing algorithms are applied, often in an unsupervised, lexicon-based approach; often there are not the conditions to apply supervised machine learning methods due to the lack of a proper training set of labeled tweets. The presence of potential out-of-topic tweets should be checked, hopefully introducing a diagnostic step into the extraction and processing models.

Quality guidelines relevant for this big data class

Coverage errors

As stated in the general part, NSIs have usually little control over the characteristics of big data and their potential drawbacks.

This is true especially for data from social networks, which is shared willingly by individuals. First of all, the activities of preserving and maintaining the data, allowing their re-use, are based on the decision of the data curator, who might not have any interest in maintaining the data in a way, which

is useful for statistical purposes. Second, the databases can be organized in a way which does not allow to trace the provenience and the origins of the data (Baker R., 2017).

In the throughput phase, big data are characterized by a set of complex treatment operations. For Twitter data, they can be grouped into three steps (Hsieh Y.P and Murphy J., 2017):

- the coverage delimitation, i.e. the establishment of time, territory and language of the tweets;
- the identification of the topical keywords and the definition of data extraction queries;
- the automated text analysis (or machine learning algorithm) to assign the sentiment and the analyses (and imputation) of demographic data on the profiles behind the Twitter.

A big data source may represent a specific segment of a population and ignore other subsets which, for some reason, may not have been included in the process of data generation. Social network data are especially affected by this limitation: as social media data (text, images, videos) are shared willingly by participants in social platforms, the subset of the population who doesn't participate in social media activities obviously cannot be captured when using these sources. This, of course, holds true also for Twitter data: It is evident that the collected tweets refer only to specific subsets of the more general population: the subset of people with a Twitter account and the subset of Twitter users that have chosen to share some of their messages publicly. The representativity of the social network data with respect to the target population is still an open issue that needs further investigation and proper methodological treatments, as for instance in the case of the mobile phone data. In the absence of a sound adjustment for potential selection bias with social network data, the inferences made from a collection of tweets and related results should be limited to the population underlying those tweets.

This concerns another point in the coverage aspect of big data derived from social networks: we often have no means to trace back the data on the events we collected to the units behind such events. Indeed, while we may be interested in studying some characteristics of a population, social networks often do not offer direct information about the units of a population, just what those units have willingly shared with the world. In other words, we collect events derived from the target population, not data on the population itself. Considering the Twitter example, this means that we do not have data regarding the person behind a Twitter account; we actually cannot know if it is a person at all, since the account could be linked to an organization or to multiple individuals. As such, the data collected could be affected both by undercoverage with respect to the target population and overcoverage with respect to specific subpopulations.

Guidelines

Establish the population of interest.
In particular, with Twitter data, it should be possible to identify the target population throughout the metadata (some of it is optional) of the Twitter account, e.g. whether the account is connected to an individual or to an enterprise.

Research the background of the units.
Social network data comprises events that are generated by units. By analyzing the content of the messages or related metadata through profiling techniques, it may be possible to identify some characteristics of the units at individual or aggregated level. This is often necessary as some social networks, including Twitter, do not require users to submit real personal information as age or occupation, leaving to them the choice to do so. Once unit characteristics have been derived, an analysis of the characteristics of the "observed" units with respect to the target population should be carried out to assess the presence and entity of the coverage error.

Surveys to obtain coverage awareness.
Since user-generated content in social network data is not usually accompanied by metadata and information about the users themselves, short surveys can be a way to better assess the observed population. Such surveys may uncover the demographic characteristics of social media users and their habits, for example the topics they are more interested in or what they most write about. However, one should consider that participation to such surveys may be related to social media participation itself (e.g. the most active social media users may be the ones responding to the survey), so caution should be exercised due to potential biases in the results. Furthermore, focus should be put not only on population coverage but on topic coverage as well.

Comparability over time

We focus on comparability over time considering the social media data for producing social mood analysis. In the following, we show a not exhaustive list of the conditions for carrying out a correct comparability analysis:

- Stability of the data provider. It has to provide data over an extended period of time: the question is whether comparable data will be available in the future, from similar providers or sources. Note that several cases of social media platforms commonly used years ago now have completely disappeared.
- Stability of the social media data characteristics: functionalities of the social media platform can change over time. A prominent example for such a change happened 2017 when Twitter allowed its users to send tweets with up to 280 characters instead of a limit of 140 characters per tweet before. The changes of functionalities affect the way of using Twitter as well as the data processing (i.e. text mining analysis), which could not understand these changes and produced misleading output.
- Stability in the data access policy. Even if social media platforms allow a free data access for a certain time period (partially or in some sense even completely), data access policies can change and affect the time series. For instance, the Facebook platform allowed partial access to its data some time ago whereas they virtually have shut down all access to data from their platform recently. Even if the policy to data access is not modified (i.e. remain free to download), it is important that this policy is stable for the automatic procedures of downloading as well.

- Stability of the algorithms for the intermediate social media data. When we consider Google Trend, which represents some sort of throughput data of the big data generating process (intermediate level of data processing), the Google algorithms change over time in a black-box context performing different results for the same query. NSIs need to take into account this issue.

Comparability over time can depend not only on the stability of the data provider and on the characteristic of the data source (stability of the data structure) but also on exogenous conditions:

- Different social media platforms can compete over time on the market share of the social media and a given platform can have different levels of the target population coverage over time affecting the statistical output.
- Technological innovations (software and hardware) undermine the use of a given social media platform, the appealing of using it, and finally they affect the coverage of the target population over time which in return affects the statistical output.

Guidelines

To deal with the concerns about the comparability over time of the statistical products NSIs should rely on suitable statistical framework. In the following, we list some relevant precautions to take into account:

Integrate use of different data sources.
Rely the statistical output on more than one source of data. The sources can be of different typology: big data, administrative data, survey data.

Continuous updating of the Data Science techniques.
Web scraping, text processing and machine learning tools have to be ready to catch the changes of the data structure.

Fit an appropriate statistical methodology for producing the output.
According to the Analyse Stage of data generating process by AAPOR (2015), apply a statistical method not sensitive to extreme data and define statistical tools for smoothing the break in the time series related to structural changes of the data source or coverage changes over time .

Measurement, process and model errors

In the case of Twitter data, the existence of the true value for reported opinions could be debated. A measurement error could be due to the tweets that report purposely fake opinions. It is worthwhile noting that the measurement errors are hardly disentangled from the model errors, for instance in the case of ironic tweets the model that interprets the opinion/attitude cannot understand the underlying sentiment.

Processing errors introduced by manual activities can mostly be neglected for Twitter data. In fact, with this data there is a large use of modeling for extracting and interpreting, so potential errors mostly fall under the category of model errors, again the distinction betweeen model errors and

process errors becomes blurry: whenever a model is adopted, additional variance on the estimates could be the result as well as bias, if the model assumptions are not valid.

Measurement/model errors in Twitter data may origin in the choice of the topical keywords and the search queries for data extraction. It has been studied that also small variations in the choice of the topic keywords can lead to wide differences in the extracted data (Hsieh Y.P and Murphy J., 2017). Once the tweets of interest have been extracted, an automated text analysis or machine learning algorithm predicts the sentiment, e.g. positive or negative. These operations are associated with a degree of sensitivity and specificity, *precision* (i.e. proportion of retrieved tweets that are relevant for the target of the search query) and *recall* (proportion of relevant records obtained by the search query). Such concepts are usually translated into the sensitivity and specificity of the method to extract and interpret the tweet.

Errors propagate: an error in the selection of the tweets may lead to a coverage error. Different sources of errors may also interact. For example, errors in the interpretation of the tweets, leading to measurement errors can depend on model errors in the specification of the algorithms applied for the query and interpretation of the tweets.

Guidelines

Establish the target information.
The definition and study of measurement errors requires the definition of the target variable of interest. Twitter data is used to infer on politics, spare time activities, sentiments, and so on. Therefore there might not be a direct relationship between the statistical target variable of interest and the measurement. Such a relationship should be clearly identified and stated.

Research on measurement/model errors.
Since the Query and Interpretation operations are those more risky for measurement error, the sensitivity and specificity of the query and interpretation algorithm could be tested on simulated data.

Literature

AAPOR (2015). Big Data in Survey Research. AAPOR Task Force Report. *Public Opinion Quarterly, 79*, pp. 839–880.

Baker R. (2017), *Big data. A survey research perspective.* Chapter 3 in *Total survey error in practice*. Wiley and Sons.

Essnet Big Data SGA2, WP7, Multidomains: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/0/04/WP7_Deliverable_7_7_2018_05_31.pdf

Hsieh Y.P and Murphy J. (2017) *Total Twitter error. Decomposing public opinion measurement on Twitter from a Total Survey Error perspective.* Chapter 2 in *Total survey error in practice.* Wiley and Sons.

Groves R. M., Fowler F.J.Jr, Couper M, Lepkowsky J.M, Singer E., Tourangeau R. (2004). Survey Methodology. Wiley, New York.

UNECE task team on big data:
https://statswiki.unece.org/display/bigdata/Classification+of+Types+of+Big+Data

Throughput phase II: Usage of the derived statistical data for the production of statistical output

This chapter covers the statistical layer: After statistical data was derived from raw data - as described in previous chapters - it can now be used to produce statistical output. The sub-chapters are defined as different usage scenarios of the statistical data. We concentrate on applications described in other work packages of the ESSnet and we do not claim to be really comprehensive with respect to all possible types of usage of new data sources. For some data sources (WPs) in this project, especially the ones in the pilot phase, no typical statistical output is produced, or to state it differently, the statistical data is the statistical output for these WPs. In these cases, this chapter is not really applicable. Depending on the application, traditional statistical process steps such as editing and imputation can happen in this process phase.

In all except one sub-chapter, we discuss multi-source statistics with at least one data source being a "big data source". The sub-chapters contain a general description, examples (use-cases) and the quality guidelines relevant for this application.

Big data sources as input for the production of official statistics

Description

This chapter covers the usage of big data sources as input for the production of official statistics. It differs from the applications described in the following chapters due to the direct use of the new data source as input, contrary to indirect uses of new data sources for validation or calibration purposes. It has to be noted, that a clear distinction between direct and indirect uses is not always possible, as the next chapter on replacement of questions from surveys shows.

An overview of official statistics currently produced that make use of big data sources reveals that two cases can be positively identified. These are: the Consumer Price Index of some countries and Traffic Intensity Statistics of Statistics Netherlands. All other examples found are, at the moment of writing, still of an experimental nature (World Bank & UNSD 2019, Eurostat 2019b).

Disclaimer: this overview is limited to all big data based official statistics produced by NSIs or similar institutes in countries that have been inventoried by a number of International and European initiatives (see Literature) and the ones that the members of WPK of the ESSnet Big Data II are aware off. As a consequence, some big data based official statistics or some very recent developments may be missed.

Examples

Traffic intensity statistics (not in the ESSnet Big Data)

The national statistical office of the Netherlands was the first to produce an official statistics completely based on big data: Traffic intensity statistic (Statistics Netherlands 2015). For this statistic, vehicle counts data produced on a minute-by-minute basis by the 20,000 road sensors on the Dutch highway network were used. Main challenges were dealing with the large amounts of data and the fact that a considerable number of measurements was missing. Quality indicators were developed to enable the quick identification of bad and good performing sensors (Puts et al. 2019). At the moment this statistic is produced once a year.

Every night, a few hours after midnight, the data of the previous day is processed – according to the specifications of Statistics Netherlands – at the location of the maintainer of the road sensor data. The processed data compressed in a single zip-file is sent over a secure line to the office. The zip-file contains three files: i) the vehicle counts per minute for each sensor, ii) the vehicle speeds per minute for each sensor and iii) a metadata file of all the sensors active during that day. In the morning, the count data and metadata are checked at Statistics Netherlands and for each sensor 5 quality indicators are determined (see Puts et al. 2019). The vehicle counts for sensors with missing data are imputed with a Bayesian filter and the total number of vehicles detected by each sensor are estimated and stored in a database.

Consumer Price Index (CPI)

In some countries, NSIs include the prices of products scraped from the web (Griffioen and ten Bosch 2016) or those derived from so-called scanner data (Eurostat 2017) in the production of the CPI. The latter is a data source that captures information about individual transactions, quantities, values and descriptions of products sold in (large) files obtained directly from businesses; such as supermarkets.

Bar codes of the products sold are the basis of the scanner data files. Currently the NSIs in Austria, Belgium, Denmark, France, Italy, Luxembourg, the Netherlands, New Zealand, Norway, the United States, Sweden and Switzerland are using scanner data in statistics production. In a considerable number of other countries, their potential is being investigated or access to scanner data is being negotiated. Examples of such countries are Denmark, the United Kingdom and Portugal. In all cases where scanner data is being used, this source only provides a part of the prices for the products needed. Hence, in all cases the CPI is based on a combination of sources, of which scanner data is one. As an example, Figure Figure_scanner provides an overview of the data sources used for the CPI of the Netherlands. Usually, monthly statistics are produced.

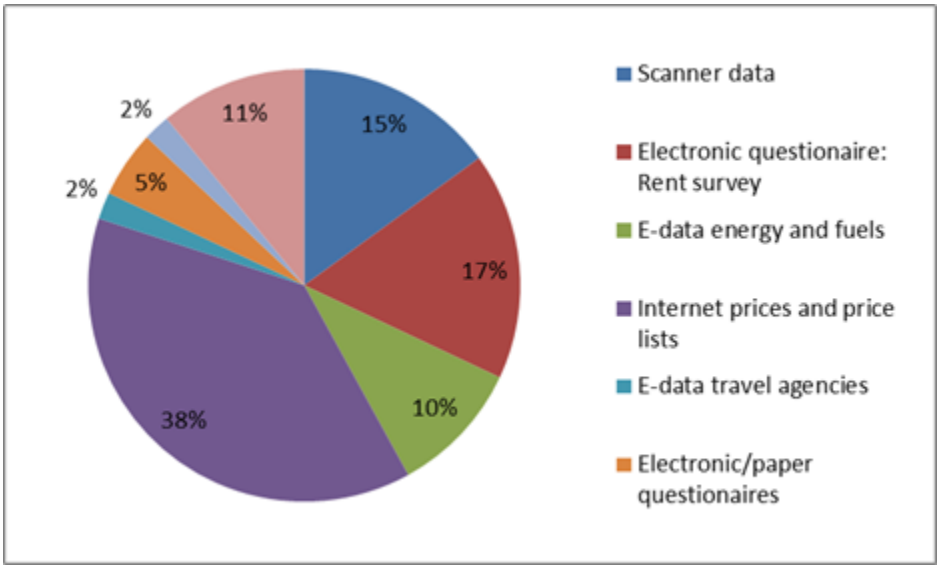

Figure 14 Overview of the data sources used for the Dutch CPI (based on Hoogteijling 2016, slide 25)

Prices scraped from the web are part of the CPI for a considerable number of countries, for example in Italy, the Netherlands and the United Kingdom (Powell et al. 2017). Many other countries are considering using it, such as Brazil and Belgium. The Billion Prices project of MIT (now called PriceStats) demonstrates that a CPI can be produced on a daily basis by using prices of products scraped from the web. However, to include the effect of the volumes of products sold, i.e. to derive consumer expenditure weights, additional official information is being used (Cavallo and Rigobon 2016).

The consumer price index (CPI) is computed by measuring the prices of a sample of representative items for which prices are collected periodically. The representative items are grouped into different categories with weights that reflect their shares in total consumer expenditures. In Europe, the European regulations and national regulations (CPI) determine the categories of products that are used for computation. Example of consumer goods and services included are food and beverages, products for personal hygiene, newspapers and periodicals, expenditure on housing, water, electricity, gas and other fuels, health, transport, communications, education, restaurants and hotels (Eurostat 2019a). In all cases where big data is used, a combination of a number of sources is additionally required to produce the CPI (see Figure Figure_scanner for an example). Scanner data are commonly obtained every week (Eurostat 2017) and websites are scraped at regular intervals, which may be as often as every day of the week (Griffioen and ten Bosch 2016).

Both sources - scanner data and scaped data - improve the quality and the efficiency of price collection and reduce the burden on the data suppliers.

Literature:

Cavallo, R., Rigobon, R. (2016). The Billion Prices Project: Using Online Prices for Measurement and Research. Journal of Economic Perspectives 30(2), pp. 151-178.

ESSnet Big Data II (2019). Second European Statistical System project on Big Data. Website located at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data

Eurostat (2017). Practical Guide for Processing Supermarket Scanner Data. Working paper, September. Located at: https://circabc.europa.eu/sd/a/8e1333df-ca16-40fc-bc6a-1ce1be37247c/Practical-Guide-Supermarket-Scanner-Data-September-2017.pdf.

Eurostat (2019a). Glossary: Consumer price index (CPI). Located at: https://ec.europa.eu/eurostat/ statistics-explained/ index.php/Glossary:Consumer_price_index_(CPI).

Eurostat (2019b). Overview of websites with experimental statistics within the European Statistical System. Located at: https://ec.europa.eu/eurostat/web/ess/experimental-statistics.

Griffioen, R., ten Bosch, O. (2016). On the Use of Internet Data for the Dutch CPI. Paper for the

Meeting of the Group of Experts on Consumer Price Indices, 2-4 May, Geneva, Switzerland. Located at: https://www.researchgate.net/publication/324896024_On_the_use_of_Internet_data_for_the_Dutch_CPI.

Hoogteijling, E. (2016). Modernisation of Price Collection at Statistics Netherlands. Presentation for the ESS Modernization workshop, Bucharest, Romania. Located at: https://ec.europa.eu/eurostat/cros/system/files/els_hoogteijiling_modernisation_of_price_collection_at_statistics_netherlands.pdf.

MacFeely, S. (2018). Big Data and official statistics. In: Kruger Strydom, S. and Strydom, M. (eds.) *Big Data Governance and Perspectives in Knowledge Management*, IGI Global, Chap. 2, pp.25 – 54. Doi: 10.4018/978-1-5225-7077-6.ch002.

Powell, B., Nason, G., Elliott, D., Mayhew, M., Davies, J., Winton, J. (2017). Tracking and modelling prices using web-scraped price microdata: Towards automated daily consumer price index forecasting. Journal of the Royal Statistical Society Series A 181(3), pp. 737-756. Doi: 10.1111/rssa.12314.

Puts, M.J.H., Daas, P.J.H., Tennekes, M., de Blois, C. (2019). Using huge amounts of road sensor data for official statistics. AIMS Mathematics 4(1), pp. 12-25.

Statistics Netherlands (2015). A13 busiest national motorway in the Netherlands. Statistics Netherlands report. Located at: https://www.cbs.nl/en-gb/background/2015/31/a13-busiest-national-motorway-in-the-netherlands.

Worldbank & UNSD (2019). Big Data project inventory web page. Located at: https://unstats.un.org/bigdata/inventory/.

Replacement of questions from surveys

Description

One way to benefit from the use of big data sources is to replace questions in traditional surveys or at least supplement them. There are three different scenarios to use big data in traditional surveys:

- Replace questions in surveys – if the big data source is of equal or better quality than the traditional data source (e.g., when the traditional data source is representative and the target population in big data source is full);
- Supplement surveys – if the data in the big data source can be used to verify data or to give a different perspective for data analysis;
- New data added to the survey – if the big data source contains more data than the traditional surveys.

Examples

The number of questions in surveys to replace is usually very limited. One example is the survey "ICT usage in Enterprises", which is conducted every year in EU countries.

Figure Figure replacement lists examples of questions that can be replaced or supplemented. Depending on the year of the ICT survey, these questions might slightly differ.

	Use of a website *Optional*		
C8.	**Does your enterprise have a website?** (Filter question)	Yes ☐	No ☐ ->go to **C11**
C9.	**Does the website have any of the following?**	Yes	No
*5	a) Description of goods or services, price lists	☐	☐
	b) Online ordering or reservation or booking (e.g. shopping cart)	☐	☐
	c) Possibility for visitors to customise or design online goods or services	☐	☐
	d) Tracking or status of orders placed	☐	☐
	e) Personalised content on the website for regular/recurrent visitors	☐	☐
	f) Links or references to the enterprise's social media profiles	☐	☐
C10.	**Does your enterprise use information about visitors' behaviour on its website (e.g. clicks, items viewed), for example for advertising or improving customer satisfaction?**	Yes ☐	No ☐

	Use of Social Media		
	Enterprises **using** social media are considered those that have a user profile, an account or a user licence depending on the requirements and the type of the social media.		
C11. *6	**Does your enterprise use any of the following social media? (not solely used for paid adverts)** *(add national examples; replace existing examples if necessary)*	Yes	No
	a) Social networks (e.g. Facebook, LinkedIn, Xing, Viadeo, Yammer, etc.)	☐	☐
	b) Enterprise's blog or microblogs (e.g. Twitter, Present.ly, etc.)	☐	☐
	c) Multimedia content sharing websites (e.g. Instagram, YouTube, Flickr, SlideShare, etc.)	☐	☐
	d) Wiki based knowledge sharing tools	☐	☐

The following question (C12) should only be answered if any of the above social media is used (i.e. C11 has at least one "Yes"), otherwise go to D1.

C12. *6	Does your enterprise use any of the above mentioned social media to:	Yes	No
	a) Develop the enterprise's image or market products (e.g. advertising or launching products, etc.)	☐	☐
	b) Obtain or respond to customer opinions, reviews, questions	☐	☐
	c) Involve customers in development or innovation of goods or services	☐	☐
	d) Collaborate with business partners (e.g. suppliers, etc.) or other organisations (e.g. public authorities, non-governmental organisations, etc.)	☐	☐
	e) Recruit employees	☐	☐
	f) Exchange views, opinions or knowledge within the enterprise	☐	☐

Figure 15 Examples of questions from "ICT in Enterprises (2019)" to be replaced by big data

It is rather easy to use web scraping to replace questions that are indicated in Figure Figure replacement. All questions are answered with the software developed by WP-C Enterprise Characteristics. The question C8 (if an enterprise has a website) is replaced with URL Finder software. Questions C11 (whether an enterprise is present in social media) and C12 (what is the type of social media activity) are replaced with the Social Media Presence software. The software mentioned in this paragraph is using search engine API's, text mining and supervised machine learning as methods to deliver answers to the questions from Figure Figure replacement. The use of this methods also helps to answer the questions from C9 (what components are present on the website, e.g. list of products offered).

Since the traditional survey is representative for a certain defined population, limited on specific economic activities and on having 10 or more employees, it is worth noting that a big data source based on the business register, could have a target population even bigger than the traditional survey.

Quality guidelines relevant for this application

Guidelines

Compare coverage.
First, it is important to compare the coverage of the traditional survey with the possibilities of the big data source. Coverage is one of the most important aspects. Sometimes, for example in Online Job Vacancies data, the definition of job vacancy in the traditional survey may be different than the one used in the big data source (online job vacancies).

Compare definitions.
The second issue is to have a unified metadata set – it is necessary to compare all definitions of data gathered in traditional data sources vs. metadata in big data sources.

Measure and report accuracy of applied models.
Due to the complexity of new data sources, e.g. the data of websites may lead to the use of machine learning algorithms, it is also important to measure accuracy of the data set and the information provided.

Validation / comparison of results with results from traditional data source

Description

When using new data sources for statistical production, both as primary source of data and in combination with other sources, i.e. in a multi-source approach, a usual step is to compare/ validate results from traditional data sources with results from the new data sources. As largely discussed above, the results in the traditional setting are usually based on design-based survey sampling theory and model-assisted inference, while in the usage of big data we move forward towards a model/algorithmic-based inference, and the interpretability of the models is often replaced by their ability in correctly predicting values at unit level and in estimating the parameters of interest. Hence, the comparison of the results from the two approaches is often useful to validate the algorithmic procedures used to extract the information of interest from the big data sources.

Examples

A subset of the estimates currently produced by the sampling survey on "Survey on ICT usage and e-Commerce in Enterprises", yearly carried out by EU member states, includes as target estimates the characteristics of websites used by enterprises to present their business (for instance, if the website offers web ordering facilities; job vacancies; presence in social networks). To produce these estimates, data is collected by means of traditional questionnaires.

These results can be compared with results based on new data sources, e.g. data collected by accessing the websites directly (i.e. via web scraping). The collected internet texts have then been processed to individuate relevant terms, finally the relationships between these terms and the characteristics of interest for the estimates are modeled.

Hence, the sequential application of web scraping, text mining and machine learning techniques represent the prediction approach to produce estimates that can be compared to the ones based on surveys.

Quality guidelines relevant for this application

Guidelines

In this kind of applications, the comparison allows a large number of quality evaluations, e.g.:

- Comparison between the variability and the bias due to sampling variance, total non-response and measurement errors in the traditional survey vs the model bias and variance in the prediction approach.
- Ability to produce aggregate estimates as well as to predict individual values.

Quality recommendations:

- Assess the coverage of the population considered by the new data sources compared to the target population (mainly risk of undercoverage);
- Assess the prediction errors of the model based approach.

Literature

Barcaroli, Golini, Righi (2018) Quality evaluation of experimental statistics produced by making use of Big Data, Q2018 European Conference on Quality in Official Statistics

Barcaroli, Golini, Nurra, Righi, Piersimoni, Salamone, Scarnò (2017) Joint use of sampling data and Big Data: the experience with the Istat survey on the use of ICT by enterprises, ITACOSM 2017

Righi, Barcaroli, Rinaldi, Bianchi (2019) Producing contingency table estimates integrating survey data and Big Data, ITACOSM 2019-Survey and Data Science

Survey based estimation with auxiliary information / calibration

Description

All applications with the main input being survey data, but with auxiliary information from big data sources are relevant in this sub-chapter. The main methods are known as model assisted survey sampling, well described in Särndal et al. (2003). These methods range from quite simple ratio estimations to calibrations (generalized regression estimator) with complex models. The motivation can be two-fold: the goal of being coherent between different statistical outputs, e.g. that survey results resemble the same population count as a register-based statistics or the goal of improving the quality (mostly in terms of precision and bias reduction) of the results. Having well-known auxiliary information at hand with strong correlation (or also non-linear dependency) to the target variable opens up possibilities to improve accuracy by incorporating this information in the survey estimates.

The most common applications of these methods are those which apply certain adaptions to the survey weights to make use of the available information. This step in the weighting process of a survey can be seen as a final adjustment of the weights that incorporate the following (general) steps in many cases:

- Design weighting (The design weight in a probability sample corresponds to the reciprocal value of the probability of being included into the sample. It depends on the sample design.)
- Non-response weighting (either based on a non-response model or just by adapting to the net sample size within strata or in total)
- Calibration (Adjusting the weights so they fulfill certain known properties, e.g. known population totals)

Traditionally, auxiliary information is derived from administrative sources, either only available as totals or as information for each individual in the sampling frame. Population totals are very common examples, e.g. the number of persons by sex, age group and state might be known from population registers and can be used. In survey theory this information is mostly considered to be "true", so to be error-free. But, for example with replication methods, it is also possible to incorporate the error of auxiliary information in the survey error estimation process.

As with administrative sources, big data sources can be envisioned to be used to provide just totals or information on each unit in the sampling frame and the examples in the next section present one example for each category.

Examples

Business survey with web-scraped information

In a business survey the question if the company has a web page is asked. Additionally, for all enterprises in the frame it is tested with web-scraping methods if a certain enterprise has an online presence or not. As this information might not be totally equivalent to the survey question definition, it can not be used directly to estimate the total number / or ratio of enterprises with a web page. However, the web-scraped information will probably be strongly correlated to the response to the survey question and is known for the whole population. A straight forward way to improve the precision of the survey estimates might be to calibrate the survey weights in such a way that the survey estimates for the number of enterprises with an online presence (according to the web-scraping definition) match with the same number for the whole population.

Survey of individuals with aggregated information from MNO data

With roaming data from mobile network operators, the number of trips by sim cards issued in one country to other countries can be estimated. In surveys about trips the number of trips to a specific country might be an important question. Mainly due to privacy reason the actual information inferred from the MNO data is not available for each individual (frame or survey) or at least not in a way so that it could be linked to a specific unit. Therefore, the totals from MNO data could be adapted in a way to resemble the data from the survey question more closely. This could be done by observing both variables for a longer time period and then hopefully find a suitable model to adjust the totals. The adapted totals can then be used in a calibration procedure to fix the number of outbound trips to a specific country.

Quality guidelines relevant for this application

Guidelines

Check definitions.
The variables from the big data source are checked regarding contents and definitions before used in a non-response analysis, weight adjustment or in general in a model assisted survey estimate.

Information must be trustworthy.
The quality of the information needs to be checked before it is used in such methods, since the survey theory regards the information to be known true population values in most scenarios.

Prefer auxiliary information on unit level.
If the auxiliary variable is available at the unit level, it is preferable to a situation with only information on the macro level, e.g. totals.

Estimators based on base weights are compared with adjusted estimators.
The base weight is a factor; usually the product of the design weight and a non-response factor assigned to each sampling unit before calibration. Estimators of the relevant key figures of the concerned statistics are analysed (e.g. the number of unemployed in LFS). Marginal totals of persons, households or businesses for important breakdowns are analysed.

Describe methodology and short-comings.
It should be described and publicly available how the method is applied and what effect can be seen compared to the base weights (see previous guideline). Possible short-comings should be clearly stated. One example of a short-coming might be to only have data from one MNO and this MNO is not fully representative for the whole population.

Literature

Särndal, C.-E., Swensson, B. and Wretman J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

Flash estimates based on leading or correlated indicators

Description

What is a flash estimate? According to the "Eurostat Handbook on Quarterly National Accounts": An early estimate produced and published for an economic variable of interest over the most recent reference period. The flash estimate is normally calculated on the basis of incomplete data, however produced using the same statistical or econometric model as for regular estimates. Flash estimate models should incorporate hard data as much as possible."

In the scope of this chapter, we extend the previous definition a bit, so the statistical indicator should be an estimate produced by a similar methodology and according to the same definitions. The main idea is to add new data sources to the estimation process that are either:

- leading indicators, so indicators that change before the statistical indicator itself is changing or

- correlated (or co-integrated) indicators that are available shortly after the reference period and before the statistical indicator could be estimated otherwise.

Depending on the data source, this line of work could also lead to a (near) real-time estimation and the opportunity of providing estimates with a higher frequency than the classical statistical indicators.

Various modelling techniques could potentially be used to compute a flash estimate, simple linear regression, regularized regression, time series methods, tree-based methods, neural networks and many more from the field of statistical or machine learning.

Examples

For the GDP estimation based on traffic data and firm-level sales data in Finland, different models are trained and compared. Forecasts are then combined, which could be described as an ensemble prediction. The traffic data is produced by traffic loops counting vehicles passing by. Since the vehicle types can be distinguished, only truck data is used. The second data source is a traditional survey of the 2000 most important Finnish companies reporting on turnover and sales (see deliverable 6.7 ESSnet Big Data I).

Statistics Poland tested to estimate the ILO unemployment rate based on the Labour Force Survey with the following two additional new data sources: the registered unemployment rate - using data from the Ministry of Family, Labour and Social Policy - and online job offers - which are then used to compute a so-called job vacancy barometer. Structural time series models are then used to produce flash estimates. (see deliverable 6.6 ESSnet Big Data I)

Quality guidelines relevant for this application

Guidelines

Evaluation and comparison of models.
Applied models should be evaluated and compared to different methods.

Comparison over time.
The estimation should be compared to the original values over multiple reference periods. Ideally, at least a full year should be observed, so seasonal effect on the estimation can be observed.

Reduce dimensions.
For some new data sources a wide variety of possible data series might be available. The option to reduce the dimensionality should be assessed systematically.

Comparability over time - Access.
The sustainability of the new data sources should be checked, especially if it can be expected that it will exist in an unchanged manner for a long time and remains accessible for the NSI.

Assess dependency (over time).
The dependency structure between the new data sources and the statistical indicator must be assessed. Furthermore, correlation between time series should be stable over time.

Examples

Evaluation and comparison of models.

For the comparison of the original time series with the series of flash estimates, several error measurements could be used, commonly used methods are:

- ME - the mean error,
- MSE - the mean squared error,
- RMSE - the root mean squared error,
- MAE - the mean absolute error,
- MAPE - the mean absolute percentage error and
- MAXE - the maximum error.

ME should be close to zero to indicate that the estimator is unbiased in comparison to the target indicator, which is important. MAXE can be important because very large errors can cause negative publicity (see deliverable 6.7 ESSnet Big Data I).

A very useful point of assessing the accuracy dimension in the quality of nowcasts is to compare the nowcasting accuracy to the revisions of the statistical office, because these revisions are deemed acceptable by the statistical offices.

Reduce dimensions.
The truck traffic data is available for different vehicle types and many locations, to reduce this high dimensionality a"Principal Component Analysis" could be applied, where new variables are derived that are linear combinations of the input variables, which are orthogonal to each other and explain as much variability as possible. A threshold on explained variability can be used to limit the number of principal components, e.g. 80 or 90 percent.

Assess dependency (over time).
The cross-autocorrelation function can be used to assess the correlation structure between time series data. When the time series are quite long, it can also be applied to subsets and then compared to see if the correlations are changing over time.

Appendix I - List of abbreviations

ESS, European statistical system

WP, Work Package in the Essnet Big Data I and II

NSI, (national statistical institute)

MNO, mobile network operator

JASON, data format, stands for JavaScript Object Notation

CSV, file format, stands for coma-seperated values

AIS, Automatic Identification System

CDR, Call detail records

DDR, data detail records

DPA, Data Protection Authority

API, Application Programming Interface

CPI, Consumer Price Index