United Nations ECE/CES/2019/32

Economic and Social Council

Distr.: General
3 May 2019

English only

Economic Commission for Europe

Conference of European Statisticians

67th plenary session
Paris, 26-28 June 2019
Item 2 (a) of the provisional agenda
New data sources – accessibility and use
Session 1: Accessing new data sources

Integrating alternative data sources into official statistics: a system-design approach

Note by Eurostat

Summary

New types of digital data sources (or 'big data') are now available as by-product of other technological processes. New data sources differ from the traditional data sources in use for official statistics, namely survey data and administrative records, along multiple dimensions. Therefore, the adoption of new data sources for the regular production of official statistics requires innovations at multiple levels, including new processing paradigms, computation methods, data access and governance models, staff skills, etc. The term "Trusted Smart Statistics" was put forward by Eurostat to indicate a comprehensive framework to evolve official statistics towards adoption of new data sources along with traditional ones.

In this document we focus on the need to take a systemic view, and in general a system-design approach, towards the development of novel processing methodologies for new types of data. We argue for the need to identify selected 'classes' of new data types (e.g., mobile network operator data, smart energy meters, satellite images, etc.) and, for each class, to build a general Reference Methodological Framework as basis for developing specific methodologies for particular use-cases and statistical products. We discuss the principles that should inform the construction of such framework, and briefly report on the ongoing work being conducted at Eurostat for one particular class of data, namely mobile network operator data.

This document is presented to the 2019 Conference of European Statisticians seminar on "New data sources – accessibility and use", session 1 "Accessing new data sources" for discussion.

GE.19-07258(E)
1907258

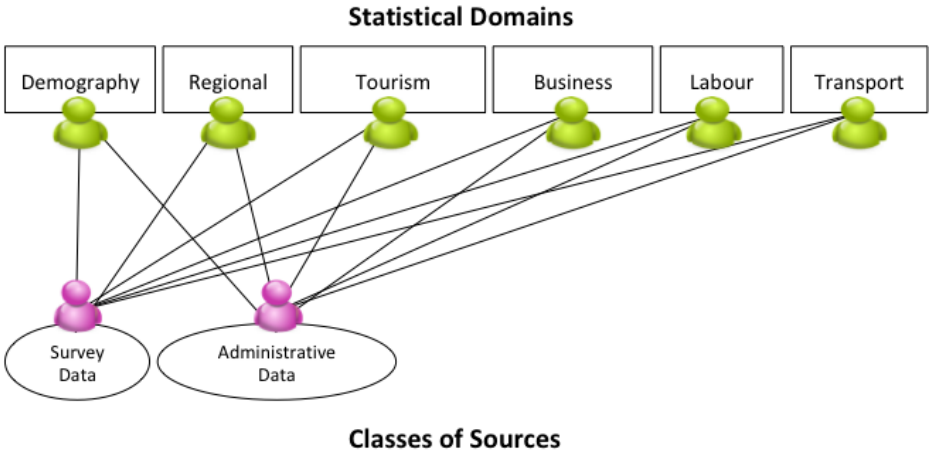
I. Introduction

1. Modern official statistics has been developed around two classes of data sources, namely survey data (SD) and administrative data (AD). All statistical domains nowadays rely on some combination of SD and AD in input, as depicted in Figure 1, while the specific instances of SD of AD might differ across domains. Note the difference between the notion of class (of data sources) and the particular elements thereof, i.e., instances. A specific survey represents a particular instance of the general SD class. Similarly a specific set of administrative records from a single register represents a particular instance of the general AD class. In Figure 1 we are representing classes, not individual instances.

2. The identification of suitable classes of source data is required to abstract away from details that are specific to particular instances, and focus on the common aspects. Abstraction, generalization and formal modeling are the necessary pre-requisites towards actionable methodological frameworks that guide and norm all stages of the data workflow. They also foster a holistic view, allowing for synergies, economies of scale and a harmonised approach.

3. The classification exercise enables the development of theory, hence the advancement of practice, only to the extent that the instances within each class are sufficiently similar to each other (intra-class homogeneity) with respect to the dimensions (features) that matter for methodological development. For traditional data sources, the classification into SD and AD is quite natural and anyway well established. For new types of digital sources, the classification task is still open, and this paper advocates the importance of conducting a careful source classification task as a pre-requisite for building meaningful methodological frameworks for each class.

Figure 1
Classes of sources – the current picture



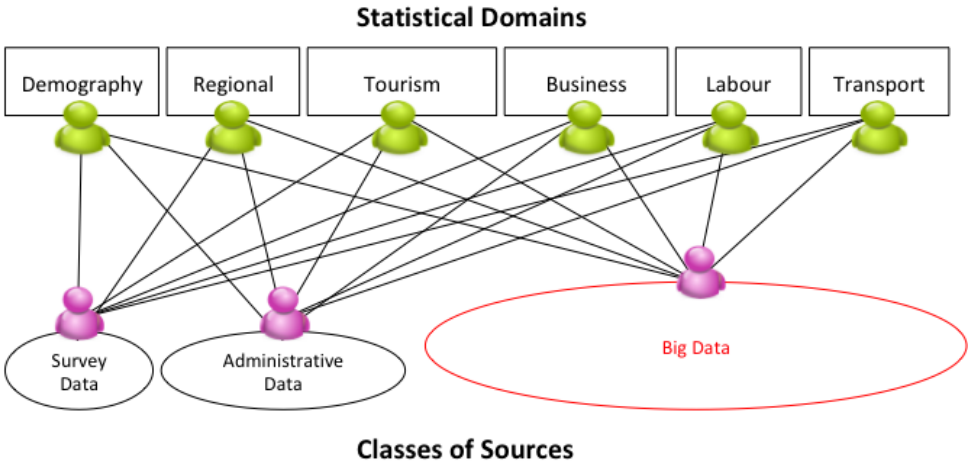
4. Figure 1 exemplifies the role of specialization within official statistics. The (green) personas on the top and the (magenta) personas at the bottom represent, respectively, domain-specific experts and source class-specific experts. Within each statistical office, there are functions and experts that are logically associated to the AD and SD classes, while others are more focused on the particular application domains. However, both types of experts typically follow the same training tracks, and any domain-specific expert has at least a basic understanding of the fundamental aspects of SD and AD. This is not necessarily the case when extending this picture towards new (non-traditional) data sources.

II. Towards classification of new data sources

5. In the last two decades, several new types of data source become potentially available for official statistics in addition to SD and AD. Such development is the direct consequence of technological progress along multiple directions, and of the increasing datafication of our lives. In the past, the umbrella term ‘big data’ has been used to refer collectively to essentially all kinds of new data sources. In the field of official statistics, ‘big data’ is often used as a

synonymous for ‘all non-traditional data sources’ other than SD and AD. To some extent, such usage has contributed to spread the misleading idea that all new data sources could be treated in a single, unified class, as depicted in Figure 2. This view is overly simplistic: the differences e.g. between satellite data and mobile network operator data are so fundamental, in virtually any 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 very abstract principles, with little practical relevance. We advocate instead the need to partition the range of candidate (new) data sources into separate classes, as a pre-requisite to conduct a meaningful dialogue about (and possibly a selection of) new data sources in the production of official statistics.

Figure 2
A wrong picture



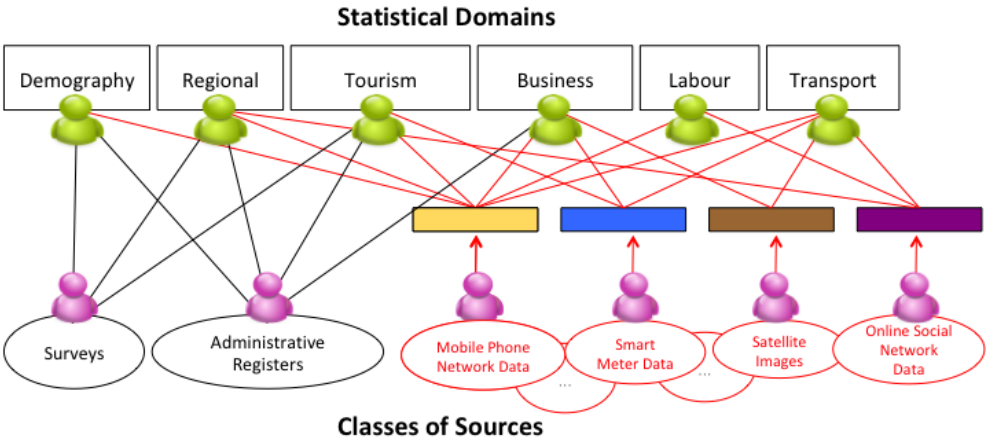
6. Among the dimensions to be taken into account in such classification exercise, the canonical “three Vs” that are associated to the ‘Big Data’ buzzword, might not be the most important in our context[1]. Other dimensions should be given higher priority and attention in the discussion, starting from the following (non exhaustive) list of items:

- Data type, structure and semantic; unstructured vs. structured or semi-structured, along with technology-specific aspects of data generation, structure and semantic
- Data confidentiality: personal vs. non-personal data, business sensitive vs. non-sensitive
- Who collects, holds and owns the data: public institutions vs. private companies; national vs. trans-national entities
- Whether sector-specific regulatory frameworks apply to the data, as is the case e.g. in telecom, energy, transport and other highly regulated market sectors.

7. Among the dimensions to be taken into account for the classification exercise, and later for the development of per-class methodological frameworks, both technical (methodology, IT infrastructure) and non-technical aspects (legal, contractual) must be addressed. In fact, recognizing the close inter-dependencies between technical and non-technical aspects, that jointly shape *how* and *where* new data can be accessed, transformed and processed for official statistics production, calls for a holistic approach where methodological, technical and governance aspects related to data and processes are addressed together.

[1] Unpinning the “Big Data” term would probably help to refocus the discussion about new data sources on more compelling and practically relevant dimensions for official statistics than the “three Vs”.

Figure 3
From *collecting* to *connecting* data



8. After the identification of data sources classes, the new scenario of official statistics augmented by new data sources would be represented graphically as in Fig. 3. Based on this picture, we highlight the following considerations:

(a) Each class of (new) data can serve multiple application domains, as highlighted in Figure 4. For example, MNO data can be used to derive information about presence and mobility of individuals serving different statistical domains: tourism, immigration, demography, labour market, etc. The multi-purpose nature of new data sources means that the related investments (including accessing, capacity building, methodological development etc.) could be repaid across a range of different statistical domains and indicators.

(b) Each application domain has the opportunity to define new, enriched indicators obtained by the fusion of data from multiple sources (new and traditional ones), as sketched in Figure 5. A prominent example is given by the use of survey data to calibrate and correct statistics derived from new data sources, that are often affected by selectivity biases.

(c) The data processing workflow, from source input data towards final output statistics, can be thought as the combination of two distinct processing layers, connected by intermediate data. Segmentation into ‘lower’ and ‘upper’ processing layers is one of the common design principles for the Reference Methodological Frameworks to be developed for each class of new data, as elaborated further in the next section.

Figure 4
Multi-purpose data sources …

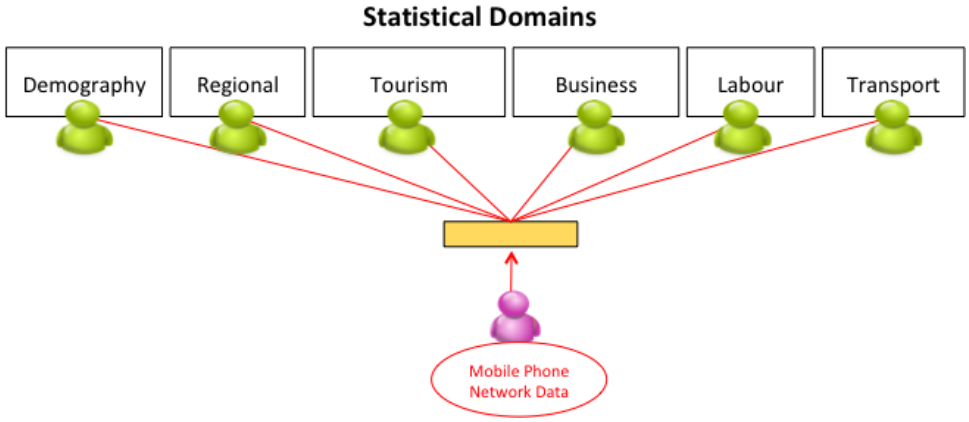
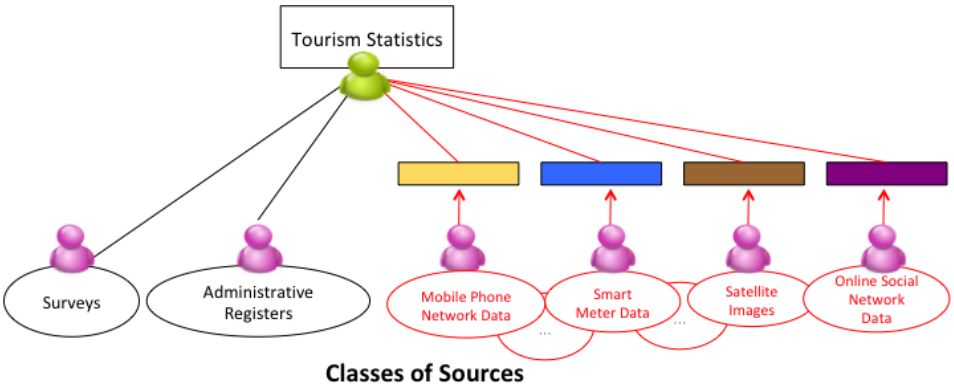


Figure 5

… for multi-sources statistics



III. Structuring the data processing flow within each class

9. One of the most important dimensions to be taken into account relates to the data type and structure (or lack thereof). Unstructured data sources like images, videos, audios, written text and spoken speech all require a layer of interpretation (image and object recognition, speech interpretation, etc.) to be turned into categorical and/or quantitative data. Nowadays, this processing stage can be performed automatically by specialised algorithms, e.g., deep learning networks and other algorithms from the field of Machine Learning (ML), that are quickly becoming commodity computing tools. Official statisticians do not need to acquire in-depth ML knowledge to use such tools (pretty much like the regular use of file compression tools, e.g. to "zip a large file", does not require in-depth understanding of the information theoretic principles of data compression). They can consult and seek guidance by computer science experts to select the most appropriate kind of ML tools to be adopted or adapted for a given application context. With the help by ML experts, official statisticians must learn to qualify such tools: understand the relevant types of errors and uncertainty that affect the interpretation result, quantify the errors and develop models and metadata to represent and properly account for such errors in the following processing stages. In other words, involvement of external experts is strongly required, but only at the first layers of the data processing workflow.

10. Other sources of data can be seen as structured or semi-structured data, with 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, airplane tracking data, etc.). In most such cases, only a tiny component of the information embedded in the raw data is relevant for official statistics, and the first layer of data processing should be dedicated to extract that (and only that) component. This stage is logically homologous to the interpretation stage for unstructured data. It requires selection functions (of variables, events, etc.) but also some basic form of low-level transformations (e.g., geo-mapping of events in mobile network data). The definition of such first layer of processing requires close involvement of specialists and technology experts from the specific source domain (e.g., electrical engineers, computer scientists). Similar to the case of unstructured data, sources of errors and uncertainty must be understood by statisticians, represented and modelled in meta-data and data models.

11. In all examples above, a first 'lower' layer of data processing is required to transform raw data (possibly unstructured and/or rich of technology-specific information that is not relevant for official statistics) into intermediate data (and associated meta-data) that can be more easily interpreted and further processed by statisticians. Close cooperation between statisticians and external technology experts (ML specialists, engineers, etc.) is required only at this first (lower) layer, characterized by functions with technology-specific logic to select and transform the data components that are relevant for further statistical purposes. In the second 'upper' layer, methods developed by statisticians transform the intermediate data

produced by the lower layer, possibly in combination with other data sources, into statistical information and indicators are relevant to their respective application domains.

12. The intermediate data block, between the lower and upper processing segments (exemplified by coloured bars in Figure 3-5) has a critical role. Ideally, the semantics, format and structure of such intermediate data should meet the following requirements:

- It should follow a common structure and format for the whole class, independent from technological details that may vary across different instances within the same class. For example, a single (intermediate) data format and semantics should be defined for the class of MNO data, not specific to the configuration details of any particular MNO infrastructure. In other words, it should be "operator agnostic".
- It should be designed in order to accommodate for the future changes in the technological details caused by the physiological evolution of the technological processes that produce such data. For example, in case of MNO data, the evolution of architecture and the principles of the next "new generation" technology (2G, 3G, 4G and forthcoming 5G) can be anticipated several years in advance, during the development and standardization process. Similarly, in other technological domains the fundamental directions of future evolution can be largely anticipated by technology experts.
- It should encode all and only the data component that are relevant for different statistical purposes in a way that is agnostic to the particular application domain and/or statistical use case.

13. The first two items above require the intermediate data structure to be 'input agnostic', i.e., independent from the detailed characteristics of the input data that may vary across instances (e.g., particular mobile operators, specific types of satellite images) and/or in time, while the third item requires it to be 'output agnostic', i.e., independent from the particular application domain and use case.

14. If the intermediate data block fulfils the above requirements, changes in time (due to technological evolution) and differences in space (across countries and/or across specific instances of data sources within the same class, e.g. different MNOs) of the underlying technology, hence of the raw input data at the bottom, can be resolved by adapting the processing functions at the lower processing layer only, with no need to modify the upper processing functions. Conversely, the modification or extension of particular use cases will be resolved by changes in the upper layer, with no need to modify the lower processing function. In other words, the presence of an intermediate data structure that is both input-agnostic and output-agnostic allows decoupling the complexity, heterogeneity and temporal variability on the two sides, easing the development and enabling independent evolution of the processing functions at both layers.

IV. Towards per-class methodological frameworks

15. A Reference Methodological Framework should be developed *for each source data class* to guide the development of specific methodologies for the various instances within the class.

16. The division into lower and upper processing layers, as described above, is one of the key design principles that should inspire the methodological frameworks. More in general, the framework should include a highly modular structure of the required processing steps. It should address methodologically relevant aspects of quality, meta-data etc. Another important general principle is the distinction between *development* and *execution* of processing methods (algorithms, methodologies, etc.). Statistical offices might be involved in developing, co-developing or at least auditing processing modules that may be executed at the premises of other entities (e.g., source data holder). This approach enables a fundamental paradigm change in official statistics, from 'pulling data in' (towards the statistical office) towards 'pushing computation out' (towards the sources), shifting the focus from the (raw) input data towards the (desired) output statistics. This approach is particularly compelling in case of confidential data (due to privacy and/or business sensitivity), for

instance to avoid risk concentration at a single institution, but it can serve to optimize the allocation of computation resources and IT infrastructure across multiple institutions. Also in this sense, having a clear modular view of the processing workflow as provided by reference methodological frameworks, is propaedeutic to clarify *what* (algorithm, function) is run *where* by *whom*, considering the peculiarities of each source class (e.g., stakeholder scenario, regulatory and business aspects, confidentiality requirements, etc.).

V. Outlook and recommendations

17. The principles outlined above underlie the development of a Reference Methodological Framework for the processing of MNO data for official statistics, an activity that is being conducted by Eurostat in cooperation with other members of the European Statistical System in the context of the ESSnet on Big Data 2018-2020. The layered structure and the 'hourglass' model adopted in such framework, sketched graphically in Fig. 6, have inspired the reflection proposed in this paper. Among the critical success elements we highlight the ability to team up professional statisticians with senior technology experts (telecom engineers in this specific case) to jointly co-develop a general framework spanning knowledge and expertise from both domains. Also, we stress the importance of adopting a unified top-down approach to modular system design oriented to production (rather than independent bottom-up developments based on individual case studies, as typically done during the previous research, exploration and capacity building stages).

18. We invite the statistical community to carry out a systematic analysis of new data sources oriented to identify 'classes' with sufficient similarity to warrant a unified treatment of the technical and non-technical aspects connected to data access and processing. For each class, the next step is to establish inter-disciplinary teams (statisticians and technology experts) to work out Reference Methodological Frameworks.

Figure 6

The layered "hourglass" model at the foundation of the Reference Methodological Framework for processing MNO data for official statistics, being worked out by Eurostat in cooperation with ESS members