Trusted Smart Statistics: how new data will change official statistics

Fabio Ricciato*, **Albrecht Wirthmann**

European Commission, EUROSTAT
fabio.ricciato@ec.europa.eu

Abstract

In this discussion paper we outline the motivations and the main principles of the Trusted Smart Statistics concept under development in the European Statistical System (ESS) to respond to the challenges posed by the prospective use of innovative digital data sources for the production of official statistics.

Keywords— Official Statistics; Trusted Smart Statistics; Big Data; Statistical Office.

1 Introduction

Historically, official statistics have provided the main basis for data-driven government and policy making. The production of official statistics is based on a *system* of scientific methods, regulations, codes, practices, ethical principles and institutional settings that was developed through the last two centuries (mostly at national level) in parallel to the development of modern states. At the centre of each statistical system is often a single Statistical Office (SO) or a federated network of multiple SO.

Statistical systems were formed in a world where “data" were a scarce commodity, consequently a large part of their capacity (and of SO resources) was devoted to the *collection* of (input) data. Surveys (including censuses) and administrative registers have been for decades the only sources of data for SO, and therefore the processing methodologies, data governance models and any other aspect of the statistical system were tailored to such particular types of data.

In the last two decades, following the digitalisation, *smartification* and *datafication* [1] of our societies, a wealth of new types of digital data sources have become available. In the official statistics community, the term “big data" has been often used to refer collectively to non-traditional data sources, as a wrapper term for any kind of data beyond survey data and administrative data. The statistical systems are now seeking to use also such “new data" (or “big data") sources, in addition to legacy ones, for the production of Official Statistics. This trend is motivated by multiple prospective benefits that (the statistical indicators based on) new data sources promise to deliver: improved timeliness, finer spatial and/or temporal resolution, increased level of detail, better accuracy, increased relevance and possibly (in the long-term) lower production cost of official statistics.

However, such weighty potential gains come with a set of not less weighty challenges to be addressed. In fact, it turns out that the “new" types of data bear fundamental differences from the traditional ones along multiple dimensions, and a systematic critical review of such differences reveals that adopting new types of data for official statistics would require (and induce) changes in almost every aspect of the statistical system, both at the technical level – processing methodologies, computation paradigms, data access models – as well as at the human level – regulatory, organisational, communications, etc.

By analogy, we may think of data sources as fuel, and the statistical system as an engine: the “new" fuel cannot be fed into the legacy engine, and statistical systems need to develop a new type of engine, with different operational principles than the legacy one, tailored to the peculiarities of the new data fuel. As new data sources will complement but not replace the legacy ones, the established system elements will not be dismissed, but augmented by the new ones. In other words, the prospective evolution path for the statistical system will take the form of a *systemic augmentation*. The future statistical system will eventually be a multi-fuel machine with multiple engines. The term *Trusted Smart Statistics* was put forward by Eurostat to signify this evolution [2] and officially adopted by the European Statistical System Committee (ESSC) in 2018 in the so-called Bucharest memorandum [3]. The design principles of Trusted Smart

Statistics (TSS for short) are outlined in the rest of this paper.

2 Micro-data and nano-data

For traditional data sources, personal data refer to features associated to an individual person, like for example her annual income, residence address, health status, number of trips abroad in the previous year, car ownership and alike. Such features are static, changing occasionally or anyway aggregated at coarse timescales (month, quarter, year). A single record refers to an individual data subject, and the term *micro-data* is used to distinguish them from aggregate indicators referred to groups and populations at the super-individual level, called *macro-data*. A set of enablers (rights to access) and safeguards (obligations to protect) are encoded in the legislation to regulate the use of personal (micro-)data for the production of official statistics.

Nowadays smart devices and sensors allow for the continuous collection of more detailed data at sub-individual level, for example the instantaneous location, every single transaction and social interaction, every single step, heartbeat, message, etc. With such new data sources, data points refer to single events, transactions, encounters or movements. Data are measured continuously and at fine timescales. We propose to use the terms *nano-data* to refer to data records at sub-individual level (event-based).

If micro-data are sensitive in terms of individual privacy, nano-data are much more so. Nano-data (also called “granular data" or “behavioural data" elsewhere) are potentially more invasive than micro-data: they qualify an individual subject well beyond a set of summary features, representing an ultra-detailed view of his/her behaviour. Considering the increased level of risk associated to (the misuse of) nano-data compared to micro-data, it should not be taken for granted that the set of enablers and safeguards developed for the latter is sufficient to handle also the former. Stronger safeguards might be required (by the law and/or by the public) to protect the confidentiality of personal nano-data compared to the established practices in place for personal micro-data. This makes the case for adopting “hard" technological solutions in addition to regulatory provisions (laws, codes of conduct, etc.) to strengthen confidentiality protection. While concentration of personal micro-data for the entire population at a single trusted administration might be considered acceptable, concentration of personal nano-data is another story.

3 Trusted Smart Statistics principles

3.1 From “sharing data in" to “sharing computation out"

The increased sensitivity of personal nano-data, and the need to avoid the risks associated to their concentration, drives towards the introduction in the statistical system of computation models based on *distributing the computation outwards* (towards the data sources) as opposite to *concentrating the data inwards* (from the sources) during the statistics production phase. This trend is further reinforced by the opportunity to use data produced in the private sector (privately held data) that might sensitive also from a business point of view. In other words, when external data gathered outside the SO are regarded as highly sensitive, whether for privacy and/or business reasons, the shift from *pulling data inside* (the SO) towards *pushing computation out* is functional to hardening the protection of data confidentiality. This shift represents a fundamental paradigm change for statistical systems, with important implications at multiple levels, not only technical.

First, the move from *sharing data* to *sharing computation* entails *sharing process control* with the data sources during the statistical production phase. It opens the way to adopt Secure Private Computation (e.g., Secure Multi-Party Computation [4]) as an enabling technology for processing confidential data that are produced (or anyway gathered) by multiple entities outside the statistical system. This combines well with other technical solutions practices aimed to strengthens transparency and public trust, like e.g. the adoption of open algorithms, fully auditable processing, public non-modifiable logging of processing instances (possibly but not necessarily based on distributed ledgers as proposed in the pioneering work [5]). Such combination of technological solutions bears the potential to further increase participation and engagement by the general public as well as individual data holders outside the statistical system. The latter remain technically in control of their data, and therefore can directly prevent any alternative use of their data, for purposes and in ways that were not previously communicated and agreed-upon.

Second, this trend pushes towards a full automatisation of the statistical production process, that must be necessarily encoded into binary code executable by machines (and not only into methodological handbooks targeted to human experts) as a prerequisite to be exported outside the SO. If the statistical methodology is encoded into a software program, we can clearly decouple the phase of methodological development (writing the source code) and production (executing

the binary code). Such decoupling allows to export the physical computation (code execution) outside the SO, partially or in full, without giving away control over the methodology (code writing). In other words, *where* the code runs remains independent from *what* the code does, and sharing control over the execution, in the production phase, does not imply any loss of control by the SO over the development of the statistical methodology. The latter might be developed internally by the SO, or co-developed by SO in cooperation with external experts (possibly but not necessarily from the same input data holder), or developed externally and then audited and approved by the SO.

Third, this change of paradigm fosters a shift of focus for statisticians and SO, from the input towards the output side of the computation process. With more and more (new, big) data available out there, statisticians can be partially relieved from the burden of collecting raw input data and can focus increasingly more on distilling high-quality output information (final statistics) from the available data.

3.2 Multi-purpose sources and multi-source statistics

By their nature, new data sources can provide multi-faceted information serving multiple statistical domains. Thus, rather than pursuing domain-specific approaches to data collection, processing and analysis, new data sources would represent the basis for multi-purpose extraction of different statistical indicators, as sketched in Fig. 1. One benefit of this approach is that the investment needed to exploit each single data source (e.g. for developing new methods and processes, building the new capabilities necessary to access and interpret new kinds of digital data) can be repaid across multiple application domains. The dual implication of this view, as sketched in Fig. 1, is that novel statistics and indicators can be developed integrating multiple data sources, including combinations of traditional (survey, administrative) data and new data sources For instance, surveys can be used to calibrate indicators computed from new data sources, e.g. against selectivity bias and/or under-coverage errors, similarly to what is done today in those countries where census is based on administrative data integrated by sample surveys (e.g. [6]).

3.3 Layered organisation of data processing workflow: the hourglass model

Considering that new data sources are often generated as a by-product of other technology-intensive processes, the development of new methodological approaches requires contribution by experts from disciplines that are outside the traditional knowledge field of official statistics (e.g., engineers, computer science experts). Generally speaking, in most cases the whole data processing flow (from raw input data to output statistics) can be split into distinct segments, or layers, as sketched in Fig. 2. In the lower layer, the raw data are transformed into intermediate data (and associated meta-data) with a clear structure, easy to be interpreted by statisticians. If the raw input data are unstructured (e.g, text or images) the lower processing layer shall include algorithms to transform them into structured data[1] (e.g. text classification and object recognition modules). Other sources of data originate by technological processes can be seen as semi-structured data with formats and semantics that are very complex and highly specific to the particular technology domain, e.g., mobile network operator (MNO) data, smart meter data, ship tracking data, airplane tracking data, etc. In most cases, only a small part of the information embedded in the raw data is relevant for official statistics, and the first layer of data processing should be devoted to extract that (and only that) component. This stage is logically homologous to the interpretation stage for unstructured data: it involves selection functions (for variables, events, etc.) but also some basic form of low-level transformations (e.g., geo-mapping of events in MNO data [7, 8]). The definition of such first layer of processing requires close involvement of specialists and technology experts from the specific source domain (e.g., engineers). Similar to the case of unstructured data, sources of errors and uncertainty must be understood by statisticians, represented and properly modelled in meta-data and data models.

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 rel-

[1] 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 in order 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 in, or adapted for, a given application context. With the help by ML experts, official statisticians must learn to *qualify* such tools, i.e., understand the relevant types of errors and uncertainty that affect the interpretation result, quantify the errors and develop models and meta-data to represent and properly account for such errors in the following processing stages.

evant for official statistics) into intermediate data (and associated meta-data) that can be more easily interpreted and further processed by statisticians – possibly in combination with other data sources following the multi-source paradigm discussed above. Close cooperation between statisticians and domain-specific technology specialists is required only at this first (lower) layer, in order to build functions with technology-specific logic for selecting and transforming the data components that are relevant for further statistical purposes.

In the uppermost layer, methods developed by statisticians transform the intermediate data and meta-data produced by the lower layer into statistical information and indicators as relevant to their respective application domains.

The intermediate data block between the lower and upper processing segments (exemplified for each data source by a coloured bar in Fig. 1) 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 a given *class of data sources*, independently 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 (see [9]). 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 generation technology (2G, 3G, 4G and forthcoming 5G) can be anticipated several years in advance, during the development and standardisation process. Similarly, in other technological domains the fundamental directions of future evolution can be anticipated to a certain extent 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.

In other words, 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. 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 instances of data sources within the same class, e.g. different MNOs) 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 structure for data and meta-data 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.

The resulting layered structure resembles the hourglass model sketched in Fig. 2. Notably, the principles of modularity, functional layering and the hourglass that we have sketched insofar are among the fundamental success factors of the Internet architecture [10, 11]. Such similarity is not coincidental at all, considering the similarities between the (actual) Internet and the (envisioned) Trusted Smart Statistics when seen as large socio-technological ecosystems. In both cases, they are large infrastructure (of hardware, software and *humanware* components) that ultimately move and transform data, involving multiple types of players with different roles (ecosystem) across multiple administrative domains, developed through a combination of (mostly centralised) *design* and (mostly distributed) *evolution*, flexible in enabling independent growth of the lower physical components (lower layer, network interfaces cards) and upper logical components (upper layer, application software).

3.4 Modular methodological frameworks

New data sources are often generated as a by-product of other technological processes that are not static, but instead are subject to change following the natural evolution of technologies and/or usage habits. This introduces temporal changes in the source data. Moreover, for some classes of data sources, certain detailed aspects of the technology are not completely standardised and may vary across countries. This introduces changes to the detailed formats and structures of data generated by different sources and/or across different countries (e.g. different mobile operators). Heterogeneity and non-stationarity of input data details poses addi-

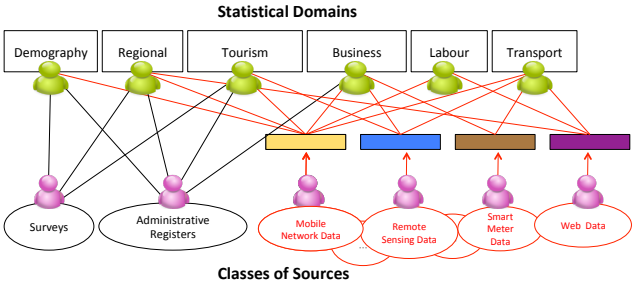

Figure 1: Each class of data serves multiple statistical domains (multi-purpose sources) and each statistical domain can benefit from different sources of data (multi-source statistics).

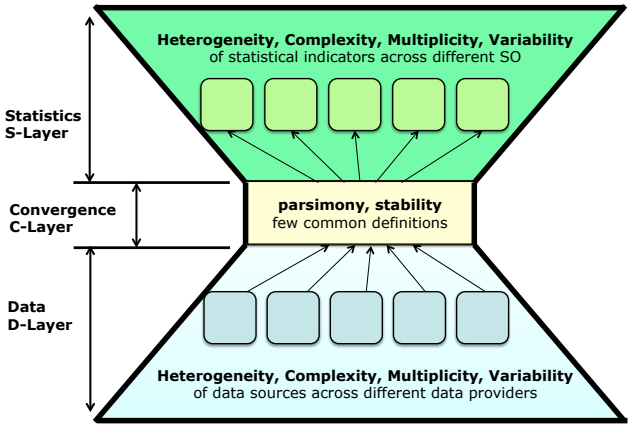

Figure 2: The layered hourglass model at the foundation of the Reference Methodological Framework being, worked out by Eurostat in cooperation with ESS members, for the specific class of Mobile Network Operator data.

tional challenges for the development of processing methodologies in terms of *evolvability* (to cope effectively with changes in time) and portability/interoperability (to address heterogeneity across countries and individual sources). The key to address such challenges is to develop highly modular methodological frameworks, within the layered approach outline above, where each module can be evolved or replaced without requiring changes to the rest of the processing workflow. The layering approach outlined above represents itself a form of modularity, but the pursue of modularity should inform also the development of statistical methodologies at more operational levels.

One advantage of the modular approach is that the *execution* of each processing modules can be assigned to the physical or logical environment that best fit for such module. Therefore, the allocation of processing modules to different physical/logical execution environments can be decided taking into account the relevant administrative and commercial scoping constraints and legal aspects. For instance, some of the processing modules at the lower processing layer can be executed at the source premises, following the principle of *pushing computation out* discussed above.

3.5 Trusted Smart Surveys

The term *smart surveys* has been used to refer to surveys based on smart personal devices, typically the smartphone. Smart surveys involve (continuous, low-intensity) interaction with the respondent and with his/her personal device(s) [12]. They combine (inter)active data provided explicitly by the respondent (such as responses to queries, or shared images) together with passive data collected in the background by the device sensors (e.g. accelerometer, GPS) on the same device or within other devices within the personal sphere of the respondent. The term *trusted smart surveys* refers to an augmentation of the smart survey concept by technological solutions aimed at increasing the degree of trustworthiness, hence promote public acceptance and participation. Constituent elements of a trusted smart survey are the strong protection of personal data based on privacy-preserving computation solutions, full transparency and auditability of processing algorithms. For instance, Secure Multi-Party Computation can be used to *use* the data of individual respondents for specific queries, without sharing the input data. In this case, every individual respondent will play the role of an input party, and citizen associations (e.g., non-governmental organisations for civil right protection) can be involved as intermediate computing parties. Together with other technological solutions, strong safeguards will guarantee also on the technical level (in addition to the legal one) that individual data can be used exclusively to compute statistics serving the collective good, ruing out the risk of causing collective or individual harm (e.g., to conduct criminal investigations or exerting public control). Furthermore, active and truthful participation should be promoted by means of a coherent strategy for public communication and individualised incentives (including personalised feedback, gamification, public rewarding, (pseudo)financial compensation, etc.) as done in other fields of Citizen Science [12].

The envisioned concept of Trusted Smart Survey blends together technological solutions and non-technical aspects into a coherent vision. Its design entails inter-disciplinary development requires cooperation with and input by with experts from multiple knowledge fields, beyond the traditional competence perimeter of statisticians, touching into disciplines – from cryptology to psychology, from behavioural economics to human-to-computer interaction design. It will

represent a novel way for SO to interact with citizen, and a pivot tool for participatory statistics — or *Citizen Statistics*.

4 Conclusions and outlook

In this discussion paper we have outlined the drivers and main principles of the Trusted Smart Statistics vision as currently seen by the Eurostat. The ESSC is in the process of elaborating further this vision and, from there, move towards the definition of an implementation action plan. Furthermore, several smaller activities undergoing within the ESS are moving forward on particular aspects and components of such complex socio-technical endeavour.

References

[1] K. Cukier and V. Mayer-Schoenberger. The rise of big data. *Foreign Affairs*, May/June 2013.

[2] Ricciato et al. Towards a reference architecture for trusted smart statistics. In *104th DGINS conference*, Oct. 2018. `https://tinyurl.com/y7pqbmze`.

[3] Bucharest memorandum on official statistics in a datafied society (trusted smart statistics), October 2018. `http://www.dgins2018.ro/bucharest-memorandum/`.

[4] D. Archer *et al.* From keys to databases ? real-world applications of secure multi-party computation. *The Computer Journal*, 6(12), December 2018. `https://eprint.iacr.org/2018/450.pdf`.

[5] Guy Zyskind, Oz Nathan, and Alex Pentland. Enigma: Decentralized computation platform withguaranteed privacy, 2015. `https://arxiv.org/pdf/1506.03471.pdf`.

[6] ISTAT. Linee strategiche del censimento permanente della popolazione e delle abitazioni: Metodi, tecniche e organizzazione, 2014. `https://www.istat.it/it/files/2014/11/Cens_Perm_pop.pdf`.

[7] F. Ricciato and G. Lanzieri. Towards a methodological framework for estimating present population density from mobile network operator data. In *IUSSP Research Workshop on Digital Demography in the Era of Big Data*, Seville, June 2019.

[8] M. Tennekes. R package for mobile location algorithms and tools, April 2017. `https://github.com/MobilePhoneESSnetBigData/mobloc`.

[9] F. Ricciato. Towards a reference methodological framework for processing mno data for official statistics. In *15th Global Forum on Tourism Statistics, Cusco, Peru*, November 2018. `https://tinyurl.com/ycgvx4m6`.

[10] M. Chiang, S. Low, A. Calderbank, and J. Doyle. Layering as optimization decomposition. *Proceedings of the IEEE*, 95, 2007.

[11] S. Akhshabi and C Dovrolis. The evolution of layered protocol stacks leads to an hourglass-shaped architecture. In *ACM SIGCOMM'11*, August 2011.

[12] E. Ruppert, F. Gromme, F. Ustek?Spilda, and B. Cakici. Citizen data and trust in official statistics. *Economie et Statistique / Economics and Statistics*, 505-506, 2018. `https://doi.org/10.24187/ecostat.2018.505d.1971`.