EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS 2024 ESTORIL - PORTUGAL

Quality assurance of official statistics based on privately held data: the use of reference methodological pipelines

Peter Struijs[1], Fabio Ricciato[2]

[1]*Eurostat, Luxembourg, Peter.STRUIJS@ec.europa.eu*

[2]*Eurostat, Luxembourg, Fabio.RICCIATO@ec.europa.eu*

Abstract

The ESS is currently exploring a novel approach to quality assurance for the case of using Mobile Network Operator (MNO) data for official statistics. The core of the approach consists of data holders and NSIs agreeing on the use of a so-called reference methodological pipeline, that is, a detailed and modular description of the data processing flow, with a clear description of what each module does and how it operates, expressed in a formal, non-ambiguous language. This reference methodological pipeline is a prerequisite to keep NSIs in effective control of the quality of the statistics produced.

The paper discusses the applicability of this approach to other types of privately held data sources. It is argued that even if other data sources have distinctive characteristics and the data processing pipelines will differ from source to source, the adoption of reference methodological pipelines will be necessary in some cases to enable the ESS to exercise quality assurance when dealing with new data sources.

Keywords: Privately Held Data, Quality Assurance, Reference Methodological Pipelines

1. The quality assurance problem

The revision of the framework regulation of European statistics enables the ESS to use privately held data for official statistics[1]. In this way statistical authorities can produce better, richer, and timelier statistics to respond to the growing information needs of their statistical users, from public policy organisations and decision makers to citizens and business actors[2].

Basing statistics, partly or completely, on privately held data requires a novel approach to quality assurance. One reason is that such new data sources may have characteristics that differ from more traditional data sources (i.e., surveys and administrative data), in particular:

[1] At the time of writing, the text of the amending regulation had not yet been published in the Official Journal of the EU, but political agreement on the text had been reached: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0152_EN.html#title1

[2] https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13332-European-Statistical-System-making-it-fit-for-the-future_en

- The data may be difficult to interpret. This can be due to their possibly unstructured nature and to the process in which the data are generated (e.g., social media records or sensor data). Moreover, their relationship with the population and variables of interest may be overly complex or unclear (e.g., the relationship between the human population and the set of mobile phones).
- The volume of the data may be huge, and the data may be streamed ('big data'). The available data may be constantly changing (volatility), and new data sources may also be unstable in the sense that they appear and disappear at much higher rates and less predictably than more traditional data sources.

This may have several consequences for the production of official statistics. Considerably more complex methodologies may be required, including new modelling techniques. More complex methodologies amplify the sensitivity of the final statistics to methodological choices throughout the data processing workflow, including the initial stages of data selection, cleaning, and pre-processing, which requires tighter control of (and transparency of) the methods applied at all stages (Ricciato, 2022). Software implementations can become complex and intricate, requiring enhanced quality assurance. The sustainability of statistics can become an issue if volatility is high. Ensuring privacy and security may require new approaches if large volumes of citizens' data from private data sources are processed, or business interests come into play. Combining information from different sources may be necessary, for instance for calibration purposes, which may raise the level of complexity even more. Attaining harmonisation across countries may require new approaches, for instance in situations where businesses operate in multiple countries. Moreover, they may prefer to get single data requests for their combined international activities (e.g., economy platforms operating at EU level[3]).

Last but not least, in some cases the acquisition and processing of raw data by NSIs may simply not be feasible or may be very inefficient. Processing the data by the data holders at their premises may then be the best option. This could also take away the need for transferring personal data. In fact, the data holder and the NSI may share the production process of the statistics in several ways.

[3] https://ec.europa.eu/eurostat/web/experimental-statistics/collaborative-economy-platform

Nevertheless, the NSI remains fully responsible for the statistics produced, their quality, and the demonstrability thereof. All this leads to novel challenges in quality assurance.

The ESS is currently exploring a novel approach to address these challenges for the case of using MNO data for official statistics. This effort builds on the new legal mandate to (re)use data held by the private sector for European statistics, to which certain conditions and safeguards apply[4]. The core of the approach consists of data holders and NSIs agreeing to process the data and produce statistics based on a standardised and open reference methodological pipeline (RMP).

> *An RMP is a detailed and modular description of the data processing flow, with a clear description of what each module does and how it operates, expressed in a formal, non-ambiguous language.*

For the MNO case, the RMP is expected to be indispensable for comprehensive quality assurance in accordance with the framework regulation on European statistics[5]. The design principles of the RMP for the MNO case and the way in which the RMP can underpin quality assurance are described in section 2.

If the approach to quality assurance by means of an RMP works for the use of MNO data for statistics, it may also work for other types of private data sources, but this cannot be taken for granted. After all, private data sources have widely varying characteristics and can be used in diverse ways for various types of official statistics. The applicability of RMPs to various types of private data sources other than the MNO case will be discussed in in section 3, followed by a more general discussion in section 4.

2. The approach to using data from Mobile Network Operators for official statistics

The ESS Task Force on the use of MNO data for official statistics (TF-MNO) has produced a vision paper on the approach to be taken (TF-MNO, 2023). In this vision, the RMP will consist of a system of interconnected modules, each dealing with a functional part of the data processing

[4] See the provisional legal text, in particular the new articles 17b to 17e: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0152_EN.html#title1

[5] This implies, among other things, compliance with the European Statistics Code of Practice and the Quality Assurance Framework.

flow, with a specification of the inputs and outputs per module. The RMP is now being developed in the so-called Multi-MNO project, launched by Eurostat, which started in January 2023[6].

The Multi-MNO project is working on an RMP for several use cases, a use case being an envisaged statistic with specified quality requirements, based on MNO data sources, for which it is assumed that the balance of total benefits and costs at the level of society as a whole is positive. Examples are statistics on the temporal location of the European population and statistics on cross-border tourism. For such statistics, data from multiple MNOs are needed, within and across EU countries, but previous research has shown that passing the highly granular raw data to the NSIs is neither feasible nor desirable[7]. The MNO industry is subject to industry-specific legislation and a certain level of technological standardisation applies, but location data are not standardised and therefore may differ across MNOs along several dimension (syntax, frequency, information content). There are interpretability issues concerning the data points, related to their temporal and spatial uncertainty, and the relationship between the set of mobile subscriptions that is "observed" through MNO data and the general population that is of interest for statistics. Some MNOs have already developed their own data processing pipeline, as they use their data for offering commercial analytic services to the market. In respect of privacy, the data are highly sensitive.

In the Multi-MNO project, NSIs and MNOs collaborate, so that the jointly developed standard, the RMP, can count on support from the ESS as well as from the industry. Importantly, in addition to the RMP, an open-source reference software implementation will be made available. This makes it possible for the RMP to fulfil its role:

- The RMP is meant to serve as a reference for all NSIs and Eurostat. It represents the detailed specification of the entire data flow, from raw data to final indicators, covering the processing parts that are carried out at the premises of both the MNO and the NSI. The processing split must be agreed on with the data holders based on the end-to-end specifications represented by the RMP. In fact, keeping part of the data processing at the MNO premises brings multiple benefits. First, privacy risks and business confidentiality

[6] Multi-MNO project | Eurostat CROS (europa.eu)

[7] The ESSnet Big Data II, a European research project that investigated the potential of using new data sources for official statistics, included research related to MNO data. The final deliverables are available from https://cros.ec.europa.eu/ESSnetBDII-WPI.

exposure are reduced if the granular data are processed close to where they are generated. Second, it allows to reuse more effectively the computation resources that are already deployed at MNO premises.

- The adoption of an RMP does not necessarily imply the application of the software implementation thereof. In fact, different software implementations (possibly proprietary) can be developed that are compliant with the RMP. Compliance should be verified based on reference input/output test data, to be provided along with a suite of compliance test functions, complementing the reference open-source software implementation.
- The part of the production where data from different MNOs are aggregated is the domain of NSIs. This stage may involve much more than plain aggregation, for instance calibration and integration with other data sources, model-based adjustments, etc. Eurostat then has the task of producing the figures at EU level. If NSIs and Eurostat agree on an adapted division of labour for cases where MNOs that hold data on multiple countries prefer to deal with a single data collection authority rather than with multiple NSIs, the RMP can still be used.

Such intended use of the RMP has implications for its design. Principles and requirements include the following:

- As a standard, the RMP must be **open and transparent**. The working and intended functionality of each module must be clear, so that outputs are fully interpretable. The reference software implementation must be open source and must come with reference test data that allow to verify compliance of alternative implementations. This must be ensured for the RMP as a whole and at the level of individual modules. This approach makes it possible for private companies to develop and market proprietary software solutions and at the same time enable NSIs to remain in control of the methodology and quality of results.
- The RMP must be **fit for purpose**. Applicability to use cases for statistics at EU level entails strict harmonisation requirements. Plausible splits of the production pipeline between the data holder and the NSI should correspond to transitions between modules. The modules on the input side must accommodate the different situations that exist regarding the raw MNO data, since data holders cannot be required to adapt their raw data generation process to the needs of statistical production. As a consequence, there must be

some leeway in the specifications, especially for the modules upstream that are closer to (non-standardised) raw data. At the input side of the RMP it may even be necessary to limit the specification to general guidelines, to be applied by the MNO and adapted to the particular configuration of the network, about which the RMP must be transparent. In any case, the quality characteristics of the modules are the main considerations for their design. These include not only common quality aspects such as accuracy, coverage, and bias, but also conditions such as privacy protection and security [8]. Moreover, fitness for purpose implies the possibility to combine intermediate data with data from other sources where needed.

- The RMP must be designed in a way that maximises **evolvability and adaptability**. The RMP design should aim to ease future extensions and customisations to meet new needs and future use-cases. For instance, modules can have configurable parameters allowing for maximum flexibility of functionality, and the overall modular architecture should accommodate the possible later addition of modules introducing new functionalities. Similarly, the RMP should be robust to changes in the landscape of data holders. An RMP that is as much as possible agnostic to the choice of use cases and is adaptable to changes in the statistical environment, including future technical possibilities and restrictions, will result in evolvable production pipelines and more a more sustainable statistical production process.

3. The applicability of the approach to other private data sources

Would the RMP approach be useful for quality assurance purposes when accessing other private data sources? There are many types of data sources, with diverse characteristics[9], and many potential use cases[10]. The differences have to be considered. Various use cases for different types of data sources have already been investigated, for instance in the ESSnet Big Data mentioned earlier, but therein the RMP approach was not considered.

[8] The quality dimensions are defined and worked out in the European Statistics Code of Practice and the Quality Assurance Framework.

[9] See, for example, annex 3 to document ESSC/2020/43/8/EN, *Actions enabling the use of privately held data for official statistics,* 43rd meeting of the ESSC, Luxembourg, 13-14 May 2020.

[10] The document *ESS use cases for privately held data,* written for the 4th meeting of the PG Task Force on access to privately held data, 21 September 2020, provides an extensive overview of use cases for various data sources.

The most promising types of privately held data to consider for a possible RMP approach are those that have a wide range of potential use cases for official statistics, as this will not only address quality assurance issues, but will result in comparable statistics to boot. This applies to the MNO case. Another such case is the use of financial transactions data (FTD) for statistical purposes. Many phenomena of social interest have a financial dimension. The possible application of an RMP to the case of FTD will be considered first. Then the suitability of the RMP approach to some other types of data sources will be discussed, but in less detail.

Data sources that hold FTD that would be relevant for statistics are banks, credit and debit card companies, and other providers of financial services (including SWIFT). Such data might be used for various statistics, such as retail indicators or income statistics, and many more.

The financial sector is strictly supervised, with some level of standardisation due to regulatory and interoperability requirements. FTD are high-volume, but technically speaking it would still be feasible for NSIs to process the raw data themselves. However, this would have negative consequences for privacy and efficiency. Like the MNO case, for some use cases there may be a need for micro-level linkage with data from other sources.

If an RMP is designed for FTD-based statistics and this is done along the same lines as the RMP for the MNO case, it could also be used for specifying the split of the production pipeline, for each data holder, between the data holder and the NSI. Given the different and more structured nature of the data, it may be expected that the RMP for the FTD case would be easier to design than the one for the MNO case. The openness and transparency required would also allow for effective quality assurance, with the NSI in control of methodological design. Data holders could still adopt, develop and even market proprietary data processing software with verified compliance with the RMP. Again, NSIs would generate national aggregates, with Eurostat producing European figures. For the financial sector, it is not clear to what extent it would be desirable to centrally collect data from international data holders. This would depend on the use case. Sustainability is a less important consideration for developing an RMP for the FTD case, but the modular approach would still yield gains in flexibility in respect of adding use cases and adapting to circumstances.

All considered, there is strong case for developing an RMP for the processing of FTD. Does this also hold for other sources of privately held data? Table 1 mentions the main factors to consider

for some prominent types of data sources and gives an idea of the extent these factors favour an RMP approach. The table is indicative and is only meant to get a rough idea of the potential applicability of the RMP approach. It does not specify the use cases associated with the data sources listed. Nevertheless, it illustrates the reasoning that can be applied when considering a possible RMP approach for a specific type of source of privately held data.

Table 1: Suitability of data sources for the RMP approach

	number of use cases	need for process split	complexity of source data	need for combining sources	cross border complications	data processing for market	data source volatility	potential RMP efficiency gains	**strength of the case for an RMP approach**
MNO data	high	high	high	high	high	yes	high	high	high
FTD	high	medium	low	high	high	yes	low	high	high
Public social media messages	medium	low	high	medium	medium	no	high	high	medium
Accommodation platforms	low	low	low	low	low	no	medium	low	low
Smart meters	low	medium	low	medium	low	no	low	low	low
Retailers' scanner data	low	low	low	medium	low	no	low	low	low

A few comments to the contents of the table may be useful. **Public social media messages**, such as tweets, can be used for various indicators, such as for a public confidence index or for SDG indicators (i.e., indicators of the UN Social Development Goals), but the required language processing and developing a valid methodology can be quite involving. The social media landscape changes fast. The RMP approach may be worth considering. **Data from accommodation platforms** are already used by Eurostat for experimental statistics. The number of platforms is limited and an RMP approach may not be needed. The possible use of **smart meters data** for official statistics will likely have strong national characteristics, since countries typically have their own regulatory framework and infrastructure for smart meters, making this data source less apt for the RMP approach. Many NSIs already use **scanner data from retailers** for official statistics, in particular as input to certain modules of price statistics, and the need to apply an RMP approach seems to be low.

4. Discussion

There may be several **reasons for developing an RMP** for European statistics based on a specific source of privately held data. First of all, it allows NSIs to ensure quality assurance in cases where the data holder carries out part of the statistical production process. For the data holder it has the advantage that he, if he chooses to do so, can use proprietary software for his part of that process and keep using his software for marketing data services. In addition to situations where raw data processing by NSIs is not feasible, splitting the statistical production process can also be a matter of choice: if the first data aggregation is done at the source, this can alleviate privacy risks and concerns. When the protection of data confidentiality requires a very high level of protection also during the processing operation, the detailed specification of the RMP facilitates the application of advanced input-privacy enhancing techniques, such as those envisioned in (Ricciato, 2024).

The application of an RMP leads in all situations to more robust statistical production, characterised by consistency, transparency, harmonisation and standardisation, including the quality aspects of the statistics concerned. However, for cases where the use of an RMP is not indispensable, this can also be achieved, in principle, by coordinating the outputs of NSIs. After all, and in line with the subsidiarity principle, this is the default way of producing European statistics. Nevertheless, NSIs may decide to apply the RMP approach on a voluntary basis, for good reasons. If the data from the private source are hard to interpret and processing is complex, requiring substantial methodological development, it may be **efficient to jointly develop an RMP**. In cases of high data or data source volatility, and to make adding later use cases easier, a modular approach may anyway make sense, even at the national level. An RMP offers the possibility to develop and reuse methodological and software modules, some of those perhaps even between RMPs. This would go considerably beyond earlier initiatives to develop repositories of standard methods and software modules[11], as the RMP approach entails additional requirements related to the suitability of all components to be part of a single methodological and quality assurance architecture.

There are some **drawbacks to not applying an RMP** when making statistics based on privately held data, apart from efficiency. Attaining the level of harmonisation and standardisation required

[11] See, for instance, https://statswiki.unece.org/display/CSPA

for European statistics would be much harder, with the risk that in the end the quality of the statistics will be lower than had been the case with an RMP-based approach. If a data holder takes care of part of the statistical production process, approaches need to be developed to make it possible for the NSI to assume full responsibility of what is done at the side of the data holder. Some form of compliance statements combined with external auditing could then be considered, but even for this some agreed reference set of norms would be needed. In any case, it would be impossible to achieve the same level of transparency and quality allowed by the RMP approach. Still another drawback is that without an RMP it would be much more difficult to accommodate data holder requests for a single data collection point at European level.

Taking a **long-term perspective**, the production of statistics inexorably moves towards a model in which individual data sources are used for multiple statistics, and individual statistics are based on multiple data sources. A generally applied modular approach to the process of producing statistics with design standards for the inputs and outputs of the modules, including their quality, would greatly facilitate the reuse of intermediate and final data sets, harmonisation and standardisation, and integrated quality assurance (Struijs et al, 2013). The long-term perspective thus points in the direction of adopting the RMP approach *by default*. In fact, the choice of the modules of an RMP would ideally be based on an analysis aimed at optimizing the entire system of statistics. This reasoning would apply to the national as well as the European level.

Even though the RMP approach looks promising in many cases, and ultimately for the default approach, there are still **many questions and issues to be addressed**. The first RMP, the one for the MNO case, is currently being developed, and the results must be awaited and assessed. The actual adoption by the industry and the ESS of the results of the Multi-MNO project cannot be taken for granted. For developing other RMPs, much depends on the choice of data sources and use cases. For the time being, it is far from clear what new European statistics will have a positive business case. It is also not clear how many initial and additional use cases can be covered by a single RMP.

All in all, the RMP approach is very promising, in offering a solution to the quality assurance problem in situations where the data holder carries out part of the statistical production process, and it may also be the best approach in many other situations, but it will not be a cure-all. The road ahead will be challenging, but ultimately rewarding – for the users of statistics and the ESS itself.

Acknowledgment

The authors would like to thank Kostas Giannakouris for his comments on an earlier draft of this paper.

References

Ricciato, F., (2022), A reflection on methodological sensitivity, quality and transparency in the processing of new "big" data sources. Paper at Q2022 conference in Vilnius, Lithuania. Available at: https://q2022.stat.gov.lt/scientific-information/papers-presentations/session-20)

Ricciato, F., (2024), Steps Toward a Shared Infrastructure for Multi-Party Secure Private Computing in Official Statistics. Journal of Official Statistics 40(1), March 2024. Available at: https://doi.org/10.1177/0282423X241235259

Struijs, P., Camstra, A., Renssen, R. and Braaksma, B. (2013), Redesign of statistics production within an architectural framework: the Dutch experience, Journal of Official Statistics, Vol. 29, No. 1, 2013, pp. 49–71. Available at: https://sciendo.com/article/10.2478/jos-2013-0004

TF-MNO (2023), Reusing mobile network operator data for official statistics: the case for a common methodological framework for the European Statistical System – 2023 edition. Statistical Report, Eurostat. Available at: https://ec.europa.eu/eurostat/en/web/products-statistical-reports/w/ks-ft-23-001