Applying the extended Total Survey Error approach to statistics based on new data sources: the case of Mobile Network Operators data

Gabriele Ascari, Giorgia Simeoni[1]

[1]Istat, Italy

Abstract

Among the new data sources that National Statistical Institutes (NSIs) have begun to explore, data from Mobile Network Operators (MNOs) occupy a relevant position. The informative potential of these data covers many domains, from population statistics to mobility and tourism. However, as with other data sources that originate from non-statistical purposes, MNO data should be thoroughly manipulated before statistical offices can extract meaningful information from them. Similarly to administrative data, MNO data are generated out of the control of NSIs. Furthermore, in order to fully exploit the potential of these data, integration with other sources – even traditional ones – becomes a crucial passage. Additional difficulties derive from the fact that the initial preprocessing operations have to be carried out by the MNOs themselves to preserve confidentiality and also because a relevant infrastructure is needed to elaborate such high volume “big” data. Therefore, the issues that may affect the overall process, along with the ones affecting the source itself, can be hard to identify and assess. In this work we propose a structured approach to explore the quality of statistics based on MNO data, focusing in particular on the quality issues arising during data processing. The errors arising during the process are studied under the lens of the well-known Total Survey Error approach: although developed in the context of sample surveys, this approach has already been extended for processes based on administrative data and combined data and, in our opinion, can offer valuable insights for processes based on non-traditional data.

Keywords: mobile network operators data, Total Survey Error, big data quality

1. Introduction[1]

In the last few decades, official statistics underwent profound changes in the way data are produced. The increasing adoption of administrative sources has allowed statistical offices to respond to some institutional and logistical challenges that the evolution of society has introduced, such as the continuous falling of response rates, the shrinking of available resources and the need to mitigate the burden on respondents.

Due to such developments, it was only natural that National Statistical Institutes (NSIs) and other statistical authorities have started looking at the rise of big data with cautious interest. However, as projects like the *Big Data* and *Big Data II* ESSnets have shown[2], big data cannot be investigated as a monolithic entity, especially from a quality perspective: the errors that

[1] This paper resumes the outcomes of a work jointly carried out by the authors, however Section 4.2 is attributable to Giorgia Simeoni while the others to Gabriele Ascari.

[2] See for example Eurostat (2020).

affect a particular category of big data may have limited influence for other data types, and vice versa. This is also the direction of the present work, which will study the quality implications of a single, specific category of big data source: the data collected and made available by mobile network operators (MNOs).

Quality can be studied under different perspectives (quality of the input data; quality of the process that incorporates the data; quality of the output that arises from the input data transformation); in this paper, we chose to study the quality of MNO data adopting the approach of the Total Survey Error (TSE), widely used in official statistics. This model, originally developed for survey statistics (Groves et al., 2004), has been applied to many other categories of data (Biemer *et al*., 2017).

An extension of the TSE is needed also for the assessment of MNO data. As an aid to this task, we will not start from the original model, but from the extended version developed for the assessment of errors in administrative data sources: the two-phase life cycle model by Zhang (2012). Studying the errors and the quality aspects of MNO data under the life cycle perspective will help to frame the errors of MNO data under the TSE approach.

The work will proceed as follows. Section 2 will illustrate the main characteristics of MNO data, why they can be useful for official statistics and how they differ from other data sources. Section 3 will introduce the TSE approach and the two-phase lifecycle model as the starting point for the application of the TSE approach to MNO data. The core of the work will be presented in Section 4, where the life cycle model will be adapted to explore the errors that may occur when including MNO datasets into a statistical process. In order to do so, it will be necessary to introduce some changes to the original model, the main one of which is the split of the initial phase into two phases, one concerning the mobile phones event data and the devices data. It is important to note here that the integration of MNO data with other data sources, which should be assigned to the second phase of the original life cycle model, is out of the scope of this work. Section 5 will conclude the paper outlining some indications for future work and potential applications.

2. Characteristics of MNO data

Just as big data cannot be reduced to a single category, MNO data themselves can be considered as made up of different classes, so it is necessary to clarify what is included within such data and what not. From the point of view of the granularity of the data, we can differentiate between event data and device data.

- Event data: an *event* is the result of any potential interaction between a mobile device and the network infrastructure; event data are thus the most granular data resolutionwise in the MNO ecosystem. They are collected by the operators to monitor network functioning and can be aggregated by device, thus building the devices datasets. An event can be identified by its timestamp and can be thought of as a triplet of information concerning the time reference, the geographical location and the device involved. We define an event as the product of a potential interaction because the actual connection between the device and the network may not succeed, due to technical failures on either counterpart. This will be illustrated in more detail in section 4.

- Device data: a device dataset contains all the information related to the activities of one or multiple devices; as mentioned, they are derived from the granular event data which are aggregated to provide, for example, all the geo-located activities carried out by a specific device.

In the description of event and device data, reference was made to geographical location of the device. This information is usually not punctual and should be estimated on the basis of topology data. These data include information about the counterpart of the devices' activities, that is the network infrastructure. In particular, they contain information about the network cells to which the devices connect and the areas they cover.

Some observations are helpful here. First, the informative potential of MNO data is fully available when the classes of data described above are merged; for simplicity's sake, for the rest of the work we will assume that information on the coverage area of the cells is already integrated with the event data. Then, another consideration is that the aforementioned categories do not exhaust the possibilities of the data involved in a hypothetical process to produce statistical information based on MNO data: other categories of data can be useful, as for example the administrative geographical subdivisions of a country. We can refer to these additional data as contextual or auxiliary data, but they will not be discussed in the framework that will be presented. Lastly, although in official statistics there is no established workflow yet for incorporating MNO data into production, it is assumed here that the sequence of data transformations is such that the event data are the starting point and the aggregated device data the end point, with single device data as an intermediate output.

3. Total Survey Error and its extensions

As mentioned, TSE was developed as an approach to assess the accuracy of the results of surveys in a holistic manner, shifting from the concerns derived from sampling errors alone and introducing the different sources of non-sampling errors that can have an impact on output quality. The identification and classification of the sources of errors can facilitate the adoption of the best methods to correct and evaluate them. Such work has been done, among the others, by Groves et al. (2004), developing a visual representation of the lifecycle of survey data and the non-sampling errors they may be subjected to. An extension of this model for processes using administrative data sources has been done by Zhang (2012). As shown in Appendix 1 the extended model depicts two distinct phases: one for the source microdata and the second for the integrated microdata, as administrative data are often used in conjunction with survey datasets and other kinds of sources. Each phase can be differentiated into a measurement and a representation line (as in the original model by Groves), which follow the journey of the data under focusing on the variables and the units, respectively.[3] In turn, the extension proposed here aims to represent the basis for the adaptation of the TSE approach to statistical processes using MNO data.

4. Applying the Total Survey Error to MNO data

In this section the mapping of the main elements of the TSE to statistical processes that make use of MNO data will be shown, extending the two-phase lifecycle model and considering both the measurement and the representation lines. However, an important difference emerges in such an attempt: event data are objects with a different level of granularity with respect to the statistical unit of interest, that are individuals. In this scenario, event data are considered nanodata, while device level data are the microdata level objects with the same granularity level as the statistical unit of interest.

Due to this distinction, the two-phase model can be slightly modified to describe the initial stages that involve an MNO data source. The solution proposed here is to split phase 1 in two distinct phases – phases 1a and 1b – referring to the transformations made on event nanodata and device microdata, respectively. The distinction in two phases for the representation line of the framework is important because it tacitly implies that the target set conceived in phase 1a may be different from the target set of phase 1b, although the data originate from the same set of phones activity. The target set is an ideal list which includes all the objects belonging to a population, regardless of their accessibility. The target set in phase 1a is the complete set of events involving the devices linked to a specific MNO; in phase 1b, we will shift to the statistical population of interest. In subsection 4.1 we will start our analysis from phase 1a, following a chronological order within the overall process. Phase 1b will be described in subsection 4.2.

[3] More specifically, the representation line in the first phase deals with data objects or records, which are yet to be transformed in their statistical counterpart, as they are microdata based on the original administrative concepts.

4.1. Event data

Concerning the events nanodata, it would be impossible to observe each one of them within a defined timeframe: there are events that are simply not registered by the network antennas for circumstances depending on the device's behalf (and, in turn, on the user's behaviour): users turn off their phones and therefore any movement or action taken during the time the devices are not active are lost. From a user's perspective, events do happen (e.g. a user moves from A to B), but the connection events do not. We thus lose all the events that fall into a non-observation category, and the dataset of events that can be obtained is not the target set anymore but the accessible set (all events that can actually be accessed by the operator's instruments). This error corresponds to what is called frame error in the two-phase lifecycle model. The instruments, however, may fail due to technical circumstances; some of the observable events may never be registered by the antennas. It is the case, for example, of widespread network failures that involve specific areas, in which no event will be registered during the interval of the network blackout. The accessible set is reduced to the accessed set, after a selection error that, generally, is of technical nature, as most errors described in this phase. Not all the events that have been observed by the antennas and correctly acquired by the operator may be eligible for further processing: the events – that are mainly constituted by a timestamp and geolocation information – may contain different kinds of error but in particular missing values, due to technical reasons. These errors may be summarised in the missing/redundancy category of errors, which separate the accessed dataset from the observed dataset, that is, the dataset of events that the operator has acquired and purged out of the main inconsistencies. It can be expected, in fact, that at this stage no correction of the errors in the event data is made. Data are simply pruned incomplete or duplicate records, and the observed set can be expected to be smaller in size than the accessed set in the defined time interval.

A single event record does not contain much information, per se. It can be considered as a string containing a temporal reference in an established format, a reference to the device and some information to geo-localize the event which, in the simplest of circumstances, assume the format of GPS coordinates. No identifier is needed (as the combination of device id and timestamp identifies the record) and, of course, these data cannot be considered as statistical units. They are similar to the events registered in an administrative data source, such as the event of birth, the event of starting a new job and so on, although there is a significant difference from such events in terms of frequency and granularity. From the measurement line's perspective, this means that the target concept is the correct localization of where the event happened, since that is the main information that the event carries. The position of the device is usually estimated through the position and characteristics of the antenna to which the event has been registered: our target measure, therefore, can be considered as the cell where the event happened and the related information, such as the cell's coverage area. Uncertainty plays a role here, because generally events are assigned to different areas with specific probabilities, depending on the signal strength of the antenna. As in the original model, the difference between the target concept and the target measure is the validity error. Of course, measurement errors, like malformed data or out of range values, can happen, usually due to malfunctions or wrong configurations of the antennas. Such measurement errors lead to an obtained measure that is imperfect, since it can include malformed data presenting wrong values. This is actually a situation analogous to the one already described regarding the end of the representation line: erroneous data are initially included in the accessed dataset, and then purged to obtain the observed set. As mentioned, we assume that no editing is made on such data. This is the reason why in the figure in Appendix 2, unlike the original model, there are no processing error and edited measure in this phase, and the arrow flows from the obtained measure of the measurement line to the observed set of the representation line, depicting the elimination of data carried out by the MNO.

4.2. Device data

As already mentioned, the first phase of the two-phase life cycle introduced by Zhang is related to a single microdata source while the second one is aimed at representing integration of different sources. It also includes, on the one hand, the shift of the perspective from the primary purpose of the single data source to the statistical one, on the other hand, the representation of transformations applied to obtain data referred to statistical units starting from objects that are in original sources. In the application of the two-phase model to device level data, the doubt arises whether to consider the first or the second phase. MNO data are still not integrated with other data but they are produced applying a transformation to event data to obtain information on an intermediate object, the device, that is closer to the statistical unit of interest, usually the individual. However, attempting to apply the second phase to device data, we realised that most of the data concepts and the errors considered were not applicable since generally connected to integration issues. Consequently, we came back to re-apply phase 1, but proposing some adjustments and extensions. As we will see, most of the steps in phase 1b are addressed to further transform the data to make them closer to the statistical concepts we are interested in.

First of all, our target population and target concepts will be the statistical ones. Our target population could be the resident population or the inbound tourists, in any case a population of individuals. The "transformed set" of data that we can obtain aggregating the event data in the final observed set from the first phase, will instead be related to devices for which events have been registered by the MNO, and this implies several substantial discrepancies when compared to the target population: people that do not hold a device will represent a relevant source of undercoverage, while IoT devices or people holding more than one device are examples of overcoverage source. In general terms, these errors could be well identified as coverage errors. Starting from the transformed set, the next steps are oriented to removing devices that are not relevant or erroneous. Indeed, filtering should usually be applied to discard devices that seem uninteresting for the statistical purpose (for example, in the case we are interested in inbound tourists from abroad we can limit our data to devices related to roaming). Obviously, a selection error can accompany this process. The resulting data could be named "filtered set". Finally, the data obtained can present situations that can be identified only considering the different events from the same device and can lead to excluding the device from the analysis or applying some correction measures. For example, a device with too few observations or a device with incoherent events (e.g. too far with respect to the time gap). The final data obtained could also be referred to as the "observed set" as in the original model and the errors as missingness/inconsistency errors.

With regard to the measurement line, we should firstly take into account that the information we can derive from device data is almost always a proxy measure of the statistical concept we intend to measure. Thus, comparing the "obtained measure" coming from the combination of event data by device with the target statistical concept we have always a discrepancy that can be associated to validity error. As mentioned in section 2, MNO data are basically given by the triplet Device/Time/Location. In this step, time, that was an identifier of the object in the event data set, becomes a measurement variable and in particular a variable subject to derivation through data processing (e.g. applying algorithms) on the basis of specific assumptions. Indeed, while with event data we have punctual observations, we often need to know the device position for time periods, so we have to decide how to assign to specific locations time periods without events. For example, it can be assumed that a device remains in the same cell of an event till the next event registered, or that the time gap between two events is split equally between the two locations, or other configurations. Obviously, different assumptions can lead to different results that impact on the output statistics. Errors generated in this step can be considered model or processing errors and lead to what can be called "derived measure". To be noted that also the assignment of the geographical location of the device over time could be further adjusted and improved using probability models or algorithms that can exploit the information derived from the combination of different events for the same device: if close-in-time events are registered by different, but partially overlapping, cells, the probability that the device is in the overlapping area can be augmented.

At this point MNO microdata at device level, mainly in order to preserve confidentiality, are subject to further aggregations over time and territorial areas in preparation of the integration step. Such aggregation should not be the source of additional errors even if errors committed in previous steps can propagate here. The next steps in which data from MNO are combined with other data can be easily described by the second phase of the two-phase life-cycle model.

5. Conclusions

It is quite possible that the next few years will witness a significant increase of the use of data from mobile network operators in official statistics. Quality actions and evaluations will be needed for such data and to this purpose it is natural to frame quality considerations within the scope of existing approaches. This paper has introduced an adaptation of the two-phase life cycle model for the inclusion of MNO data into a statistical production workflow, highlighting the errors that could emerge at different steps and which could be included in the concept of a total survey error extension specific to this kind of data. The model illustrated here could serve as a canvas on which additional information could be added, for example quality indicators for the evaluation of specific errors or metadata concerning specific data objects.

This theoretical work, however, presents some limitations, which could be overcome by further studies. First of all, in this paper only a single operator has been considered, while in practice data will come from multiple operators; as a consequence, integration between MNO datasets has not been considered, along with the integration between MNO data and other sources as well. Furthermore, specific data provided by the operators, such as datasets on the characteristics of the network, have not been examined. Lastly, as statistical offices acquire familiarity with MNO data, new types of error could be identified or new processing procedures could be adopted, enriching the structure of the proposed model.

References

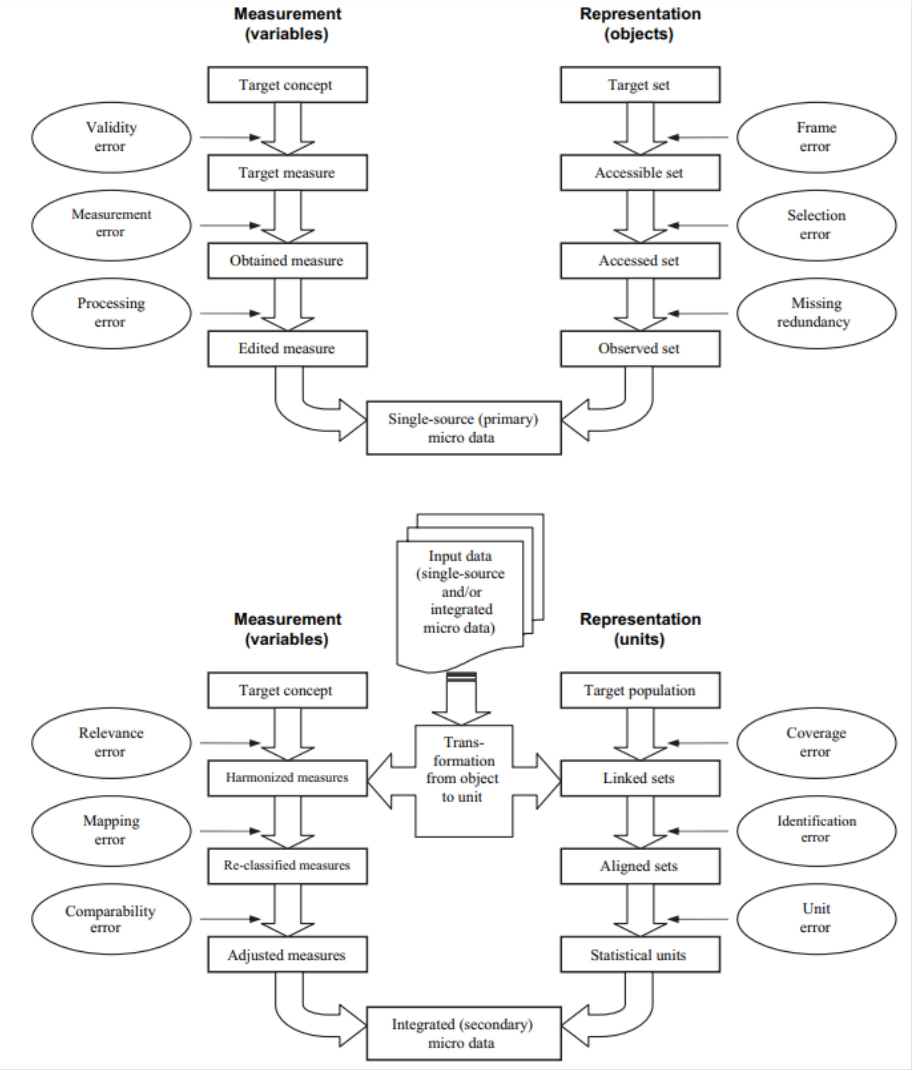
Biemer P., de Leeuw E., Eckman S., Edwards B., Kreuter F., Lyberg L., Clyde Tucker N., West B (2017). Total survey error in practice. Wiley and Sons.

Eurostat (2020). ESSnet Big Data II, work package K: methodology and quality. Deliverable K3: Revised version of the quality guidelines for the acquisition and usage of big data.

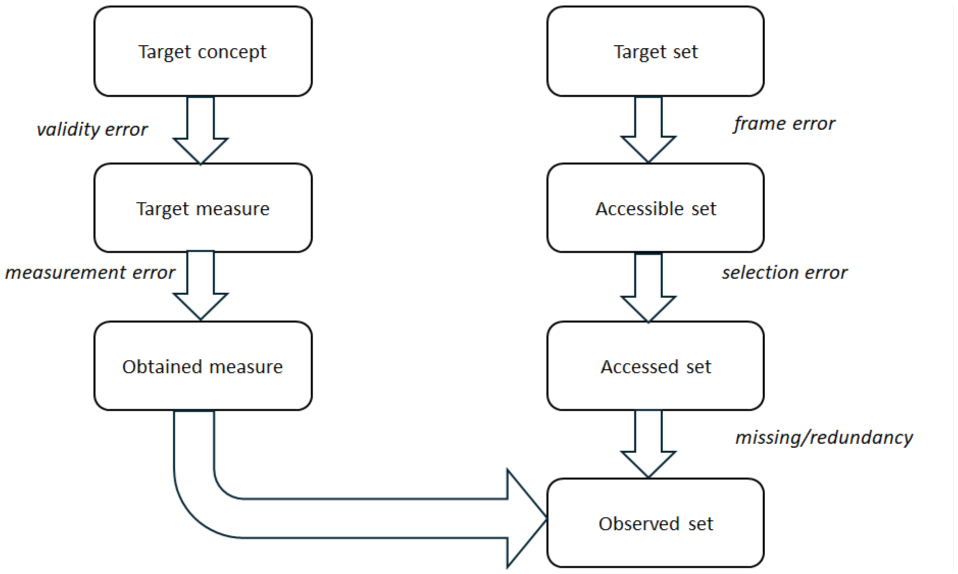
Groves R., Fowler F., Couper M., Lepkowski J., Singer E., Tourangeau R (2004). Survey Methodology. Wiley.

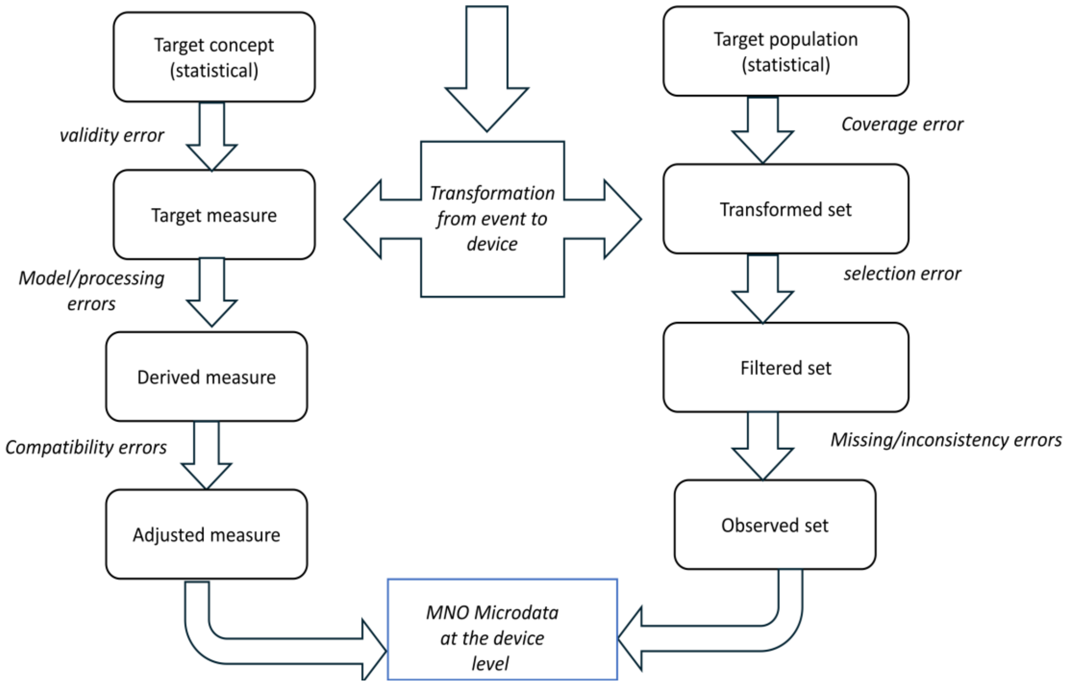
Zhang L.C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, vol. 66, n. 1.

Appendix 1. Original two-phase life cycle model by Zhang (2012)



Appendix 2. Adapted phases 1a and 1b applied to MNO data life cycle


Target concept
validity error
Target measure
measurement error
Obtained measure
Target set
frame error
Accessible set
selection error
Accessed set
missing/redundancy
Observed set


Target concept (statistical)
validity error
Target measure
Model/processing errors
Derived measure
Compatibility errors
Adjusted measure
Transformation from event to device
Target population (statistical)
Coverage error
Transformed set
selection error
Filtered set
Missing/inconsistency errors
Observed set
MNO Microdata at the device level