Quality Framework for combining survey, administrative and big data for official statistics

Yvonne Gootzen, Statistics Netherlands, yapm.gootzen@cbs.nl

Piet Daas, Statistics Netherlands, pjh.daas@cbs.nl

Arnout van Delden, Statistics Netherlands, a.vandelden@cbs.nl

Abstract

Creating statistics by combining data sources allows for the production of new, more timely and/or more detailed statistics. With an intended statistical output in mind, and various potentially useful data sources, a logical first step is to assess the ability of each source to contribute to the intended statistic. Quality frameworks provide tools for such tasks. This paper proposes a quality framework that includes dimensions applicable to survey, administrative and big data in such a way that each of them is informative for the intended statistic. The framework is applied to a case study of mobility data.

Keywords: combining data sources; multi-source statistics; data integration; quality framework

1. Introduction

Many multi-source statistics are based on a combination of surveys and administrative data. Meanwhile, the first examples of successful applications of big data in official statistics are appearing (ESSnet, 2018; Puts & Daas, 2021) Big data comes with its own challenges, which are partly different from the challenges regarding the use of surveys or administrative data. The experience at Statistics Netherlands reveals that quality frameworks created for surveys and administrative data alone (Groves & Lyberg, 2010; Bakker, 2018) cannot be fully applied to big data sources. In such a framework, the nature of a big data source is usually so different that the evaluation becomes very uninformative. Similarly, quality frameworks specifically developed for big data (Batini et al., 2015; Eurostat, 2020; UNECE, 2014) are not designed to include the most relevant information in surveys and administrative data. Therefore, the need arises for a multi-source quality framework that is relevant and applicable to surveys, administrative data, big data and combinations thereof.

Frameworks for two of the three types of sources have been developed for survey and administrative data (de Waal et al., 2020), and for survey and big data (Amaya et al., 2020).

Quality frameworks agnostic to the type of input data are often focussed on statistical output rather than on the input data (OECD, 2011; UN, 2018). Though the intended application of such frameworks is not specifically to assess a single data source for the purpose of creating a multi-source statistic, the underlying quality dimensions are relevant to consider for specific sources (Puts & Daas, 2021; Schober et al., 2016).

This paper reports on the similarities observed between existing quality frameworks and proposes a framework based on combinations of surveys, administrative data and big data. In the proposed framework, the sources that one wants to combine are assessed individually. After this, the assessments per source are combined to judge whether the combination of sources can be used to produce the intended statistic. In terms of hyperdimensions of Karr, the proposed framework mainly focusses on the data hyperdimension of quality (Karr et al, 2006). The framework takes into account the target variable and target population of the intended statistic. Additionally, it includes the intended aggregation level of the intended statistic and any available accompanying data.

2. Dimensions for categorisation of data sets

The quality framework presented in this paper is meant to categorise individual data sets. Each dimension consists of categories that were chosen such that the resulting categorisation is a summary of information relevant to the process of combining the data set with other data sets to create an intended statistic. The other data sets which are considered for combining with the current data set, are called accompanying data.

The *context* forms the basis for the perspective from which we will look at the data set, and should be determined before the data set is categorised. The context of the intended statistic consists of a target variable, target population and aggregation level of the intended statistic, as well as available accompanying data. The time dimension is considered part of the aggregation level.

More than one category may be applicable per dimension. We propose the following dimensions for categorising a data set for the purpose of combining with accompanying data sets to create an intended statistic, within a specified context:

Relevance. Does the data contain information relevant to the intended statistic?

- *Directly relevant.* The data contains the target variable, or a variable that so closely resembles the target variable which implies that no accompanying data, variable or model is needed to extract the target variable at the intended aggregation level.

- *Indirectly relevant.* The data contains information that can be relevant, but only in combination with an accompanying data set, variable or model. Or the aggregation of the target variable is only available for a unit type which is non-trivially linked to the intended aggregation level.
- *Irrelevant.* The data does not contain information that is relevant to the user. If this data was not available, it would not influence the final result. If new accompanying data become available in the future, the classification in this category should be reconsidered.

Population coverage. How complete is the population in the data compared to the target population?

- *Perfect coverage*. Every unit in the target population occurs exactly once in the data.
- *Duplication*. The data contains units of the target population. Some units of the target population occur more than once in the data.
- *Undercoverage*. The data contains units of the target population. Some units of the target population are not present in the data.
- *Undetermined.* No direct link is available between the units in the data and the target population. No claims can be made about the coverage of the data set.
- *No unit-type coverage.* The unit type of the data is different from the unit type of the target population. Some accompanying data or modelling is needed to convert the unit type to that of the target population, before the population coverage can be assessed.

Population representativity. To what extend can we derive whether the set of units in the data represent the target population?

- *Known inclusion probabilities.* The inclusion probabilities of units are known. This includes cases with a probability sample or deterministic selection.
- *Unknown inclusion probabilities*. The inclusion probabilities of units are not known. This includes cases with a non-probability sample.
- *Non-zero inclusion probabilities.* All inclusion probabilities of units are larger than 0.
- *Zero inclusion probabilities.* Some inclusion probabilities are 0.
- *Undetermined.* Population representativity cannot be determined if no unique identifiers of units are available, or the unit type of the target population is not covered in the data. No accompanying variables are available to measure the representativity.

Variable validity. How well does the data set measure the target variable?

- *Perfect.* The definition of the target variable is identical to the definition used in the data and no measurement errors occur.

- *Definition inconsistency*. The definition of the target variable or unit type is different from the definitions in the data. Definition inconsistency is also known as concept (in)validity.
- *Measurement error*. A measurement error causes the values in the data to be different than the intended definition in the data set.
- *Modelling error*. The variable in the data set was previously derived from a different variable by imperfect modelling.
- *Processing error*. The variable contains errors from a previous processing step.
- *Causal error*. Errors have been introduced by disregarding causal connections between variables in a previous version of the data (or multiple data sets if the current set is a combination of data sets).
- *Undetermined*. The definition, measurement process or modelling process of the variable in the data is (partially) unknown.

Concept stability. Does the assessment of the data set in the variable validity dimension remain stable over time?

- *Stable*. The level of definition consistency, measurement error and modelling error of the data set compared to the target variable are stable over time.
- *Concept drift*. The level of definition consistency of the data set compared to the target variable changes over time. Note that this can also be due to a change of definition in the target variable over time, when the change is not present in the data set.
- *Unstable*. Either the measurement error or modelling error of the data set changes over time.
- *Not applicable*. For the purpose of this study, the concept stability is irrelevant. This may be the case for sources where the target variable is not included in the data source.

Correctability. Can inaccuracies (such as bias) in the data be corrected by modelling or by combining with other data sets?

- *Unnecessary*. No correction is needed because the data accurately measures the target variable.
- *Self-correctable*. The inaccuracies in the data set can be corrected using accompanying variables in the data set itself, without usage of other data sets.
- *Supplement-correctable*. The bias in the data set can be corrected using other accompanying sets, possibly by linking them with variables from the current data set.
- *Uncorrectable*. The data cannot be corrected within the given context.
- *Undetermined*. It is unclear whether the data can be corrected within the given context.

Recentness. What is the nature of the time lag between the occurrence of a phenomenon and the moment it is first reported in the data?

- *Event-based.* The data related to an event becomes available relatively soon after the event occurred, without grouping multiple events into a single "delivery" of data, resulting in a stream of data.
- *Periodically.* A system is in place that guarantees a periodical release of data. The system includes some automatic processing of the data, which possibly includes aggregation.
- *Sporadically.* Availability of the data is dependent on individual actions that are hard to anticipate. Or there might not be any guarantee that a successor of a data set will become available in the future.

Processing timing. What is the nature of the time lag between obtaining access to the data and the intended statistic being ready for publication?

- *Instantly.* An automatic system is in place that ensures data can be processed virtually instantly, which is at least before the next instalment of data is available.
- *Automated.* Whenever a new instalment of data is available, it can be processed with few human interventions.
- *Individually.* The data is processed manually and the process is started on request each time a new instance of the data becomes available.

Accessibility. To what extent are there limitations to access the data?

- *Full access.* Legal access is guaranteed for the foreseeable future and does not limit the options based on the technical availability of the data. Usage of the data is allowed for the publication of the intended statistic.
- *Paid access.* The data is accessible for a financial compensation.
- *Limited access.* Legal issues either prevent the user from accessing the full data or limit the scope of the intended statistic.
- *No access.* There is currently no access to the data, or usage of the data for the intended statistic is not allowed.

Meta-data. To what extend are the definitions of the variables known?

- *Synergetic.* The metadata is complete and well defined and fits perfectly with metadata from accompanying data sets and the intended statistic.
- *Well-defined.* The metadata is complete and well defined, but does not fit well with metadata from accompanying data sets or the intended statistic.

- *Ill-defined.* The metadata is largely available, but vague and allows for multiple interpretations.
- *Incomplete*. The metadata is largely unavailable, and exploratory data analysis or assumptions are necessary to interpret the data.

Comparability. How well can the results of the research be compared to results of parallel research?

- *Fully comparable.* There is a general consensus of definitions used in the data and in parallel projects.
- *Semi comparable.* Some discrepancies between definitions in the data and in parallel projects can be expected, but a conversion allows for the outcome of the study to be compared to parallel studies.
- *Non-comparable.* The definitions used are different to such an extent that it is unlikely the results will be comparable to parallel studies.

The dimensions population coverage, population representativity, variable validity, concept stability and correctability all fall under the umbrella of accuracy. The dimensions recentness and processing timing both fall under the umbrella of timeliness and punctuality.

3. Case Study

We illustrate the proposed quality framework by applying it to a study -at Statistics Netherlands- where administrative data, survey data and big data were combined to Dutch road network data (Gootzen et al, 2022). A similar approach was applied to a metro network on city level (Gootzen et al, 2020). Traffic intensities on the road network in the Netherlands were studied by combining four different sources. The four data sources each have their own unit types and, at a first glance, do not seem suitable to be combined. The first source concerns a combination of different administrative data sets. The second source is the ODiN survey, where people are asked to report on their transportation movements during a particular day. The survey data contains a sample of persons that travel for work with a known transportation modality. It was used to train a model that determines the probability of a certain modality, given the background characteristics of a person. This model was applied to the combined set of administrative data, which was subsequently aggregated into an origin-destination (OD) matrix. The OD-matrix is composed of pairs of neighbourhoods and the expected number of people that travel to work by car. The third source is based on Open Street Map data, more specifically the road network of the Netherlands (Open Street Map, 2022). Using Open Trip Planner (2022), the OD-pairs were converted to routes consisting of road segments which resulted in an expected intensity for each road segment. The last source is traffic loop data which

contains observations of traffic intensities for road segments per minute (Puts et al, 2018). The minute-based data was aggregated to one value for each sensor by taking the sum of all observations during the morning rush hour. This way, most travel from home to work is taken into account while minimizing the inclusion of travel for leisure or travel from work to home, which tends to happen outside of the morning rush hour. In short, this resulted in two variables for a set of road segments: expected intensity and observed intensity. This allowed for validation and calibration of the model that led to the expected intensity. The complete approach applied for combining the four sources mentioned is described in detail in the references mentioned above.

Before applying the framework to the available data sets, let us first formalize the context of the study. The target variable is the number of cars and the target population is the road segments in the Netherlands. Since we are classifying one source at a time, the accompanying data sources are the three remaining sources not subject to classification. The aggregation level is defined as the morning rush hour peak per road segment. Table 1 shows the result of applying the quality framework on the data sets used in the case study, given this context.

Let us discuss some of the assigned classifications and how they were overcome by combining with accompanying sources. These notes are marked by corresponding numbers in the Table 1.

1) Since the target population consists of road segments and the unit type of administrative data is person, we can claim there is no unit-type coverage. The route planner (fed by the location variables in the administrative data) acts as a converter between the two unit types.
2) The dimension *concept stability* is not relevant for these sources since the target variable of the intended statistic is not present in these sources.
3) The available infrastructure for travel may change over time. This may affect the calculated routes from the infrastructure data. Though infrastructure data may seem robust over time, it is important to use a data set that corresponds to the time stamp of the administrative data and the sensor data.
4) The minute-based data was aggregated from 5 a.m. to 9 a.m. to resemble the morning rush hour. To correct for measurement errors, the aggregate was averaged for all regular working days during a full month. Since these corrections were applied without using accompanying data, the sensor data was categorised as self-correctable.
5) It is likely that other countries have similar data that could be used in the same role if this project were to be applied in another country.

Dimension	Admin. data	Survey data	Infra. data	Sensor data
Relevance	Indirectly relevant	Indirectly relevant	Indirectly relevant	Directly relevant
Population Coverage	No unit-type coverage (1)	No unit-type coverage	Perfect coverage	Undercoverage
Population representativity	Undetermined	Undetermined	Known inclusion probabilities	Known inclusion probabilities
Variable validity	Definition inconsistency, modelling error	Measurement error	Target variable not present	Measurement error, definition inconsistency
Concept stability	Not applicable (2)	Not applicable (2)	Concept drift (3)	Stable
Correctability	Supplement-correctable	Supplement-correctable	Unnecessary	Self-correctable (4)
Recentness	Periodically	Periodically	Periodically	Event-based
Processing timing	Automated	Automated	Automated	Automated
Accessibility	Full access	Full access	Full access	Full access
Meta-data	Synergetic	Synergetic	Well-defined	Well-defined
Comparability	Semi comparable	Semi comparable	Fully comparable	Fully comparable (5)

Table 1. Application of the proposed quality framework to the case study.

4. Conclusion & discussion

In this paper, we propose a quality framework for combining administrative data, survey data and big data for official statistics. The quality framework was applied to a case study where all three types of sources were combined. Applying the framework to each source separately and comparing the outcome as a whole allowed for an overview of the challenges that were encountered during the case study. This study illustrated a number of key points, particularly when non-trivial modelling steps were included to obtain the intended statistic. In the case study, one data source served as a link between the two populations. Here, it became clear that even though the populations differed in a number of sources, they could still be combined towards the intended statistic. The combination of categorisations for each source helps to identify the key points beforehand and helps to identify similar situations, which allows for re-use of the solution.

5. References

Amaya, A., Biemer, P. P. & Kinyon, D., 2020. Total error in a big data world: adapting the TSE framework to big data. *Journal of Survey Statistics and Methodology,* pp. 89-119.

Bakker, B. F., 2018. *Quality evaluation of register-based statistics.* Krakow, European Conference on Quality in Official Statistics.

Batini, C., Rula, A., Scannapieco, M. & Viscusi, G., 2015. From Data Quality to Big Data Quality. *Journal of Database Management,* pp. 60-82.

Citro, C., 2014. From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology,* pp. 137-161.

de Waal, T., van Delden, A. & Scholtus, S., 2020. Commonly used methods for measuring output quality of multisource statistics. *Spanish journal of statistics,* pp. 79-107.

De Waal, T., Van Delden, A. & Scholtus, S., 2020. Multi-source Statistics: Basic Situations and Methods. *International Statistical Review,* pp. 203-228.

ESSnet Big Data, 2018. *Report about possible new statistical output based on (European) AIS data,* sl: Eurostat.

eurostat, 2020. *ESSnet Big Data II, Workpackage K: Methodology and quality, Deliverable K5: First draft of the methodological report,* sl: eurostat.

Gootzen, Y., Roos, M. & Bostanci, I., 2022. *Data Collection for City and Subnational Statistics - Milestone: comparison between patterns from register- and big data sources in the mobility network,* sl: Statistics Netherlands.

Gootzen, Y., Roos, M. & Mussman, B., 2020. *Combining data sources to gain new insights in mobility.* [Online]

Available at: https://www.cbs.nl/en-gb/over-ons/innovation/project/combining-data-sources-to-gain-new-insights-in-mobility

Greenberg, J., 2017. Big Metadata, Smart Metadata and Metadata Capital: Toward Greater Synergy Between Data Science and Metadata. *Journal of Data and Information Science,* pp. 19-36.

Groves, R. M. & Lyberg, L., 2010. Total survey error: Past, present, and future. *Public opinion quarterly,* pp. 849--879.

Higginson, S. et al., 2018. Achieving Data Synery: The Socio-Technical Process of Handling Data. In: C. Palgrave Pivot, red. *Achieving Energy Policy: Lessons on the Integration of Social Sciences and Humanities.* : Springer Nature, pp. 63-81.

Karr, A. F., Sanil, A. P. & Banks, D. L., 2006. *Data quality: A statistical perspective,* : National Institute of Statistical Sciences.

Lovelace, R., Birkin, M., Cross, P. & Clarke , M., 2016. From Big Noise to Big Data: Toward the Verification of Large Data sets for Understanding Regional Retail Flows. *Geographical Analysis,* pp. 59-81.

OECD, 2011. *Quality Framework for OECD Statistical Activities,* : Organisation for Economic Co-operation and Development.

Open Street Map, 2022. [Online]
Available at: https://www.openstreetmap.org/

OpenTripPlanner, 2022. [Online]
Available at: https://www.opentripplanner.org/

Puts, M. J. & Daas, P. J., 2021. Machine Learning from the Perspective of Official Statistic. *The Survey Statistician,* pp. 12-17.

Puts, M. J. H., Daas, P. J. H., Tennekes, M. & de Blois, C., 2018. Using huge amounts of road sensor data for official statistics. *AIMS Mathematics,* pp. 12-25.

Schober, M. F. et al., 2016. Social media analyses for social measurement. *Public opinion quarterly,* pp. 180-211.

Scholtus, S., Bakker, B. & Van Delden, A., 2015. *Modelling Measurement Error to Estimate Bias in Administrative and Survey Variables,* The Hague: Statistics Netherlands.

UN, 2018. *UN Statistics Quality Assurance Framework,* : Committee of the Chief Statisticians of the United Nations System.

UNECE, 2014. *A Suggested Framework for the Quality,* : UNECE Big Data Quality Task Team.

van der Sloot, B., Broeders, D. & Schrijvers, E., 2016. *Exploring the Boundaries of Big Data,* The Hague: Amsterdam University Press.

Zheng, Y., 2015. Methodologies for cross-domain data fusion: an overview. *IEEE Transactions on big Data,* pp. 16-34.