

# Quality Concepts and Metadata for Statistical Frames

George Petrakos

Panteion University, Dept. of Public Administration

Agilis SA, Statistics and Informatics, Athens GR

e-mail: [george.petrakos@agilis-sa.gr](mailto:george.petrakos@agilis-sa.gr)

## Abstract

The selection, evaluation, maintenance and use of statistical frames are key procedures in the design and implementation of statistical surveys. The aim of this paper is twofold. First, to propose definitions of quality components, similar to those that Eurostat and other organizations use in order to define and evaluate quality in statistical surveys, which form a base for defining and assessing quality in statistical frames. Second, to present the design of a metadata repository for statistical frames, which contains, formally arranged information on descriptive and quality concepts. This metadata can be applied to the selection and quality assessment of a statistical frame but can also be used to the evaluation and improvement of regularly used statistical frames and registers hosted by Statistical Offices and other organizations.

**Keywords:** Metadata; Statistical frames; Quality

## 1. Introduction – Basic Definitions

Attempting a formal definition of the basic concepts related to populations and sampling frames, we start with the *population unit*, which is a complex concept that determines classifies and describes a physical or legal entity. It contains characteristics on identification, classification, access and communication as well as subject (quantitative and qualitative). A well defined, finite or at most countable set of units under study, is called *population*. This concept is determined (exclusively) by a set of logical statements connected with logical operations ( $\vee$  or/and  $\wedge$ ) in a way that uniquely defines whether a unit belongs to the population or not (logical filter). Every population is a dynamic entity which evolves with time regarding to its content, size and characteristics. As an example the population of informatics companies in Greece is changing rapidly with many new companies entering the population while fewer ones are leaving (closing, moving or changing subject) every year. At the same time many population units are growing bigger, incorporating with others, changing their names or legal status.

Entering the statistical world, we define two important mappings. The *statistical unit*, a finite vector of random variables that corresponds (exclusively and uniquely), to those characteristics of population unit which are used or examined by the study (identification, classification, communication, and subject characteristics) and also the

random variables, determined with the mathematical mapping from the real world to the line of real numbers. A well defined, through a logical filter, finite set of statistical units, 'alive' and accessible at the time of the survey, is called statistical population or target population. In order to design and contact a statistical survey we need certain pieces of information from the population under study, which are summarized in a file named statistical frame. Statistical frames are lists containing the statistical units with (at least) the following characteristics:

- (i) *Key variables (id)*: one to one correspondence of a population unit with a frame record (VAT number, Social Security Number, Personal ID Number, etc.).
- (ii) *Communication variables* (address, phone, e-mail, fax, etc.).
- (iii) *Classification variables* (sex, age group, region, legal status, size, etc.).

Statistical Frames can be constructed by using Registers, one or more, whole or parts. Registers, official records used by public administration, were noted in ancient Athens, kept in a building named Mitroo, dedicated to the mother of Gods. This was where they presented all the original resolutions written on papyrus, leather or wooden tables. The state would hire an employee especially for the arrangement and guard of these documents, where the registry office documents were also held. Mitroo is still the Greek name for the register. Now days statistical offices and other CNAs create, update and keep registers for several purposes, among them statistical survey sampling. The quality of a statistical frame relays upon the quality of the register(s) used and the quality of the process followed (merging, updating, etc.), in a similar way that the quality of a main dish in a restaurant relays upon the quality of raw products used and the quality of process(cooking).

The most important issue related to the quality of sampling frames is the closeness of the sampling frame (SF) to the target population (TP), which is known as *coverage*. Let P the set of all statistical units in the target population and R the set of statistical units (records) in the sampling frame. We also define, using the basic set operations, the following sets:

$P \cap R$  : Population units included in the sampling frame

$P \cap R^c$  : Population units not included in the sampling frame (under coverage)

$P^c \cap R$  : Sampling frame units not belonging to the Population (over coverage)

$D$  : statistical units with no one-to-one correspondence between SF and TP (duplication - clustering).

We can identify several sources of deviation between SF and TP with the most important being the time lag between the creation of the frame and the time (actual time and reference time) of the survey. Furthermore, incomplete or erroneous recording, lack of identity or relevance between SF and TP, duplication, clustering and misclassification are expanding the deviation of the two sets creating a number of coverage errors. Trying to identify the reasons why certain statistical units belong to  $P \cap R^c = \{\text{units} \in P \text{ and } \notin R\}$ , we presume incomplete recording, lack of relevance and/or identity, misclassification and last but not least units that enter the population after the creation (or the last update) of the frame (time lag). Working similarly for  $P^c \cap R = \{\text{units} \in R \text{ and } \notin P\}$ , we presume erroneous recording, lack of relevance and/or identity, misclassification and finally population units alive at the time of the frame creation but not at the time of the survey (time lag). Statistical units in  $\mathbf{D} = \{\text{units} \in P \text{ with no one-to-one mapping in } R\}$  are there because of duplication and/or clustering

two errors that commonly occur when a SF is produced by merging two or more registers. Specifically, Duplications occurs when several (more than one) units in the frame are mapped onto a single unit in the target population, while clustering occurs when several population units mapped on a single unit in the frame. Under coverage, over coverage and duplication are considered coverage errors.

## 2. Quality in Statistical Frames

Quality in Official Statistics is a complex (multidimensional) concept, which incorporates attributes and characteristics such as timeliness, relevance, accuracy, coherence, comparability, clarity and accessibility addressed to statistical data which is the main product of the various statistical systems and organizations (Statistics Canada, 1998, Statistics Finland, 2002, Eurostat, 2003). The quality of these products is defined and measured in accordance with the customer (data user) needs. In a similar way, we consider a statistical frame (or a register) as a product and we define quality with its major attributes and characteristics, focusing on the user, which is the statistical survey itself, needs. An excessive amount of work has been done in defining and improving the quality of registers (especially Business Registers) in Statistical Organizations (for example, Wallgren, 2007, Vale, 2001). In this chapter we attempt to define quality components addressed to any statistical frame, using the same attributes and characteristics redefined in context of statistical frames and registers, as follows:

***Timeliness** is the time difference between the creation of the statistical frame (or the last update) and the contact of the survey. The time difference between the creation of the statistical frame and the reference time of the survey will also be considered.*

Populations are dynamic structures changing over time. Only for a short period of time they can be considered stable (close population), while in general births, deaths and changes in characteristics occur (open population). This general statement applies, among others to target population under study. When the time lag between the creation of the statistical frame and contact of the survey is getting large a number of statistical units enter the population (without been recorded in the frame), while another number of units leaves the population (but still remains in the frame). At the same period of time, eligible units experience changes in characteristics adding to the deviation between SF and TP. Many types of coverage errors naturally appear in quantities analogous to elapsed time and since we cannot prevent them we can try to estimate them. The first population parameter under study is the birth rate, which is estimated by the average of new incomes in the population per unit time. In the same manner the death rate, the second population parameter, is estimated by the average of units that leave the population per unit time. Changes in characteristics such as enterprise name, size and legal status, or individual occupation, resident and marital status can be also treated with estimating changing rates for each characteristic. By knowing these rates, the statisticians can estimate the coverage errors resulted from the time lag between SF and TP and use these estimates in the design of the survey and the calibration of the survey estimates. Any auxiliary information on these estimates should accompany the sampling frame (or register) as metadata since it can also be useful in the quality assessment of the frame and of the results of the survey.

**Relevance** is the degree of closeness and completeness of the statistical frame to the target population at the time of its creation (no time lag).

Relevance consists of several quantitative and qualitative characteristics such as variable (vertical) completeness, record or strata (horizontal) completeness, item completeness, structural and definitional differences between SF and TP. Under the term of relevance we group all these imperfections that have to do with the sampling frame and not with the time lag. Variable completeness give to the statisticians the ability to identify the statistical units, to communicate with them (maybe in more than one ways using different sampling modes) and finally to perform the sampling scheme of their choice (stratified, multistage, etc.). Item incompleteness results in to the exclusion of some population units and more severe of some population groups from the survey, driving into some forms of bias in the final estimates. Also missing data in large volumes of certain variables results in poor performance in communicating or creating strata using these variables. Missing 40% of the telephone numbers in a business register make impossible to perform a post survey with telephone follow ups. Finally, when definitional and structural dissimilarities occurs between population and frame units (e.g. household and family units), we can experience practical problems in the implementation of a survey (Kish, 1965).

**Accuracy** is the precision of sampling frame data and depends on the erroneous values of the random variables included in the Sampling Frame.

Validity checks can reveal a number of measurement and processing errors from different stages of the frame construction such as data entry or transfer, register merging, variable updates, classification etc. These checks can be performed before and after the survey and their outcome (i.e. percentage of errors) can be coded as quality metadata. Therefore, issues such as misclassification, ineligible units, outliers, erroneous data, clustering, can be revealed, estimated and treated. Several approaches as edit specification using abstract data model and binary segmentation (Petrakos, 2004) or robust automatic methods (Chambers, 2004) can be used to detect outliers and erroneous data in large sampling frames.

**Accessibility** is consists of certain technical characteristics which make the frame easy to access and friendly to use.

These characteristics are, distribution channels, ordering procedure, time required for delivery, pricing policy, marketing condition (copyright), availability of micro or macro data (detail level of information), media/format (paper, CDrom, Internet, database in electronic format (Excel, Access, Oracle)), sampling capabilities, variable classifiability.

**Coherence** is the internal balance of those sampling frames, which consist of more than one registers and relays upon common id codes, definitions and classification as well as similar register time of creation and reference.

**Comparability** is relay upon the existence of standard (international) definitions and classifications along with reports on changes in definitions and classification, which allow the sampling frame to be comparable with others, used in different time, domain and place.

*Clarity is the degree of existence and access to certain metadata type of information such as variable list and definitions, classification, updates, registers involved and merging process, which clearly indicates to the potential user if this SF is suitable.*

Clarity depends on the existence of variables metadata (definitions, units and classifications), while coherence and comparability depends upon the nature of these characteristics (International standards and classifications, common definitions).

### **3. Metadata Systems**

Within a statistical frame we identify variables classified as identification, communication and stratification variables, basic concepts such as statistical unit, target population, coverage, etc. Information on names and titles, construction and size of the frame, variables by type, classifications used, reference time and area, etc., can be coded to form a subsystem of descriptive metadata. For example this system contains among others a complete variable list, recording the name, the type and the description of each variable. It also contains information on the register or registers used for the construction of the frame such as time of construction, updates, processing method, etc. This subsystem will provide a complete qualitative and quantitative, as much as possible, description of the relevant statistical frame.

The second subsystem will be based upon the quality concepts defined in the previous section and include statistical metadata concerning quality. Concepts such as timeliness, relevance, accuracy, accessibility, comparability, coherence and clarity are defined and measured through a set of relative metadata. Construction date, reference period, estimates on birth, death and change rates provide a quantitative picture of timeliness and give the input to the researcher in order to estimate coverage errors. Percentages on item, variable and unit completeness, along with a description of existing definitional differences between population and frame units will specify and access the relevance of the statistical frame. Accuracy can be assessed through validity checks and the corresponding metadata array will include description of the performed test and their results (i.e. number of occurrences and percentages of misclassification, ineligible units, outliers, erroneous data, clustering. Accessibility can be described and evaluated through metadata describing the distribution channels, the ordering procedure, the delivery time, the pricing policy, the detail level of information available, the format, the sampling capabilities, and the variable classifiability. Metadata on classifications, definitions, registers used, changes in definitions and classifications we enable us to evaluate coherence and comparability of the statistical frame. Finally the existence of most descriptive and quality metadata will ensure clarity.

The integrated, by the two subsystems, metadata repository will be able to answer four major questions. a) Is that frame suitable for my survey; b) Is that frame good enough for my survey; c) Was that frame good enough for my survey; d) What can be fixed or improved for the next survey. Clearly a) and b) are evaluated before the survey while c) and d) after. This metadata repository, organized in a SDMX standard, can be applied to the selection and quality assessment of a statistical frame

but can also be used to the evaluation and improvement of regularly used statistical frames and registers hosted by Statistical Offices and other organizations.

SDMX is a standard that facilitates the statistical data and metadata exchange. It offers artefacts for representing structures for reference metadata that can be used to produce metadata sets in XML (i.e. SDMX-ML the XML implementation of the SDMX standard). The artefact for creating such structures is called Metadata Structure Definitions (MSD) and can define all metadata and their definitions. This structure is generic and can be used for representing the metadata for the statistical frames presented in this work. Therefore these metadata values can be represented with SDMX-ML metadata sets and processed by any SDMX tool (e.g. related with storage, transmission, dissemination). This opens a broad area of application providing registers and frames with structural metadata repositories.

## Acknowledgements

The author would like to thank Mrs. Stavroula Chrysanthopoulou for her valuable inputs in descriptive and quality metadata and Mr. Spyros Liapis for his contribution regarding SDMX standard. The author is also grateful to the NTTS Scientific Committee for their constructive comments.

## References

- Chambers, R., Hentges, A., Zhao, X., (2004) Robust automatic methods for outlier and error detection, *Journal of the Royal Statistical Society(series A)*, 167, 323-339
- Biemer, P. P., Lyberg, L. E., (2003) *Introduction to Survey Quality*, Willey series in Survey Methodology, NJ, USA
- Eurostat. (2003) *Definitions of Quality in Statistics*, Doc. Eurostat/A4/Quality/03/General/Definition
- Groves R.M. et all. (2004) *Survey Methodology*, Willey series in survey methodology, NJ, USA
- Kish, L. (1965) *Survey Sampling*, Wiley, New York
- Petrakos, G., Conversano, C., Farmakis, G., Mola, F., Siciliano, R., Stavropoulos, P. (2004) New ways of specifying data edits, *Journal of the Royal Statistical Society(series A)*, 167, 249-274.
- SDMX: Statistical Data and Metadata eXchange <http://www.sdmx.org>
- Statistics Finland. (2002) *Quality Guidelines for Official Statistics*.
- Statistics Canada. (1998) *Quality Guidelines, 3<sup>rd</sup> edition*.
- Vale, St., Perry, J., Pont, M. (2001) Developing and accessing the quality strategy for business registers: a UK perspective, *NTTS Conference, Crete 2001*, pp.415-421
- Wallgren, A., Wallgren, B. (2007) *Register-based Statistics-Administrative Data for Statistical Purposes*, Wiley, New York