"Science Stays True Here"
Journal of Mathematics and Statistical Science, 29-38 | Science Signpost Publishing

An Ontology-Based Approach To Administrative Data Sources' Documentation And Quality Evaluation

Giovanna D'Angiolini, Pierina De Salvo, Andrea Passacantilli

Italian National Institute of Statistics, Via Cesare Balbo 16, 00184, Rome, Italy.

Abstract

The paper discusses the documentation and standardization requirements related to the statistical utilization of administrative data sources and illustrates the main features of the Istat's strategy for satisfying such requirements, which is centred on an ontology based approach to the information content specification and data quality assessment.

Keywords: Administrative data source, Administrative data documentation, Administrative data quality, Data source ontology, Statistical data production

Introduction

A large number of NSIs (National Statistical Institutes) usually exploits administrative data for statistical purposes, in order to improve the quality of statistical outputs, to reduce the statistical burden on respondents and to minimize costs (Brackstone 1987, UNECE 2011). Moreover the official statistical data production is not the only context in which administrative data are used for statistical purposes: in recent years more and more non statistical organizations have been implementing their own decision support systems for monitoring the context and the effects of the organization's strategy. Such systems exploit the organization's administrative data sources together with other data sources for drawing inferences about those phenomena which are involved in the organization's activities. This is a kind of statistical

Corresponding author: Andrea Passacantilli, Italian National Institute of Statistics, Via Cesare Balbo 16, 00184, Rome, Italy. E-mail: passacantilli@istat.it.

exploitation of administrative data even if the purpose is to satisfy the organization's own knowledge requirements, while NSIs produce statistical data for the public.

Any statistical usage of an administrative data source implies a source evaluation activity, whose goal is to ascertain if the source can be used for studying the phenomenon of interest. Such an evaluation activity includes two main distinguished phases:

- evaluating if the collective and the variables of interest can be derived from the observed administrative collectives and characteristics
- evaluating if the administrative data concerning the collective and the variables of interest exhibit a good quality from a statistical viewpoint, that is, they can be used as dependable measures of the underlying phenomenon.

Often the statistical users of the available administrative data sources perform such an evaluation activity from their particular viewpoint without any reference to standard procedures and shared methodological and documentation tools.

As to the first evaluation phase, in many situations the administrative data users are compelled to analyze the whole source's information content for determining if the source satisfies their particular information requirements. Moreover they apply different approaches and models and therefore they cannot share the produced documentation. This is a serious drawback: in fact analyzing the information content of an administrative data source may require advanced conceptual modelling competencies, due to the complexity of many administrative data sources' observed part of the real world. The statisticians would take advantage of the availability of standard and shared documentation of the administrative data sources' information content.

Similar remarks apply to the administrative data sources' quality evaluation. The main Frameworks of quality indicators for administrative data sources propose sets of very general indicators which aim at documenting the overall quality of the analyzed data sources, and are not well-suited to leading the quality evaluators in assessing the data source's quality with reference to their particular collectives and variables of interest. In concrete situations the administrative data source's users perform more specific quality evaluation activities but they cannot share the results of such evaluation activities, because they generally apply empirical and not repeatable procedures.

In such a scenario the NSIs are required to play a new important role. They must devise and release guidelines, standard methods and tools for supporting any kind of user which need to evaluate the

administrative data sources' information content and quality in order to exploit administrative data for acquiring knowledge about real world phenomena.

Supporting the Statistical Utilization of Administrative Data Source: A New Task for Nsis

Today many NSIs are focusing their interest on exploiting new data sources in all the stages of the statistical data production process and dedicating. Researchers are working at defining methodologies for obtaining estimates from sets of integrated data sources, which may encompass surveys as well as other data sources, particularly administrative data sources.

However several new methodological and practical issues are related to the goal of consistently combining several data sources. In particular, in order to compare and effectively exploit the available administrative data sources we should be informed about their information content and we should be enabled to assess their data quality. In practice these basic requirements are not generally satisfied.

For any survey, the observed collectives and variables are often documented inside dedicated metadata management systems; if not, the survey's questionnaire provides the data users with some documentation about the survey's information content. On the contrary, the availability of administrative data sources as well as the available administrative data sources' information content are not documented in a systematic way, despite the efforts of statistical and not statistical institutions.

The surveys' data quality is largely controlled by the survey's designer. Several methods for reducing errors in the data collection stage are available, data editing is performed by means of proper standard tools. On the contrary, generally the statistical users cannot influence the administrative data collection stage and the administrative data editing stage, they only can assess the resulting administrative data quality and take it into account in choosing and combining data sources.

At present we have not yet proper approaches, methods and indicators for assessing the administrative data quality in a satisfactory way. We cannot directly apply methods and indicators which are used for surveys, mainly because: a) while the most surveys are designed as snapshots of the observed collectives at specified moments, the most administrative data sources collect new information at any moment, in a continuous way b) many administrative data sources observe a richer set of collectives when compared with surveys, linked by complex relationships, in particular they observe sets of events which

occur in the course of time, and the data users generally need a specific data quality evaluation for their particular collectives of interest c) the administrative sources' data are often affected by systematic errors, due to the purposes of the administrative data collection.

Some frameworks of quality indicators for administrative data sources exist, which may be useful for a first assessment of the administrative data source's overall quality. However often such a first assessment does not produce enough information for supporting the administrative data users in deciding if it is worth to use the administrative data source, especially if they are interested in specific observed collectives. In such a situation, often the administrative data users perform ad-hoc analyses of the acquired administrative data collections, and sometimes manipulate them, without fully documenting their operations.

The result is a scenario which points out the need for further investment in methods and standardization. In fact, a long history of methodological studies and standardization efforts has produced the survey techniques which are commonly used in official statistics. Nothing similar is today available for producing statistical data from administrative data sources.

Moreover, it is worth noting that today the usage in the NSI's data production processes is not the only possible statistical usage for administrative data anymore. Due to the spreading of the Data Warehouse approach in the latter years more and more non statistical organizations have been implementing their own decision support systems, which in fact employ statistical techniques even if their purpose is not the statistical data production. This decision support usage of administrative data requires that the exploited data exhibit a good quality when regarded as measures of real world phenomena, that is to say it requires a good data quality from a statistical viewpoint. As a consequence the interest in administrative data quality issues is spreading across several research communities, such as the database research community (Benedikt et al. 2006). More recently the open data vision is strengthening this trend.

In particular many organizations belonging to the national statistical systems, such as government institutions which need to monitor the effects of their adopted policies, are building large data warehouses which may encompass their own administrative databases together with survey data and external administrative databases.

In such a new scenario the NSIs are required to take responsibility for a new methodological coordination task. They must devise and release standardized guidelines, methods and tools for supporting any kind of user which need to exploit administrative data for acquiring knowledge about real world

phenomenaas they currently do for surveys (D'Angiolini et al. In press).

Supporting the Statistical Utilization of Administrative Data Sources: The Istat's Strategy

In order to carry out such a new task Istat (Italian National Institute of Statistics) has undertaken a general strategy aimed at making the available administrative data sources more understandable and usable.

Generally speaking the Istat's strategy is aimed at:

- collecting information about the available administrative data sources and producing standard documentation about their information content and quality
- modifying, when possible, the content of the available administrative data sources through adopting standard statistical definitions, classifications and data management conventions.

In order to provide the users with proper knowledge about the content and the quality of the administrative data sources Istat is launching several systematic documentation activities, which concern different kinds of administrative data sources.

The central government institutions manage large information systems made up of several administrative data sources which are fed and exploited by administrative procedures. In such a context, the Istat's experts together with the data source's experts jointly perform a dedicated investigation on each administrative data source and its related administrative forms, in a systematic way.

An administrative data source investigation is an analysis and documentation activity which follows a standard template in order to collect comparable information about the content and the quality of the data source. It is performed by means of analyzing the available documentation and interviewing the source's experts belonging to the owner institution as well as the source's users, and consist of three actions: 1) specifying a source's general description, 2) analyzing and documenting the source's information content, 3) collecting qualitative information about the source's data quality.

The collected information is managed by means of a dedicated web-based metadata management system, called DARCAP (Documenting Public Administration Archives) in order to disseminate it to any potential statistical user of the documented administrative data sources (D'Angiolini et al. 2014).

Such dedicated investigations enable us to thoroughly document the information content of the

available administrative data sources, but they collect only a limited amount of qualitative remarks about the quality of such data sources. For the most important and complex data sources, the statistical users may need additional information about the source's data quality. In order to support an in-depth quality analysis, we are studying a new Quality Assessment Framework for Administrative Data Sources.

Unlike central institutions, local government institutions often manage a large number of independent administrative databases aimed at supporting a large spectrum of heterogeneous administrative tasks which concern several subject matters, ranging from environment management to employment monitoring. In order to gain knowledge about such administrative data sources, Istat organizes dedicated administrative data sources' surveys together with those agencies which represent local government institutions. Such surveys enumerate the existing administrative data sources and classify them by subject matter. Moreover they collect some pieces of information for each administrative data source, such as the main observed collectives and variables. The collected information is stored into the DARCAP system too.

Istat is also launching another activity, which is aimed at making it easier to modify the content of the administrative data sources. This activity is the supervision on changes and innovation projects concerning the administrative data sources and their related forms.

For the most important administrative data sources, the owner institution is required to notify Istat each time it plans a change in the source's information content. Such a notification concerns all kinds of changes, periodic changes in the forms for income declaration as well as major innovation projects such as a new data warehouse. On the basis of the received notifications, Istat may give feedback and release proper recommendations. Examples of such recommendations are: using official instead of non-official classifications, improving the identification code system, improving the quality control procedures.

The DARCAP system provides the administrative data sources' owner institutions with a dedicated subsystem for supporting the change notification activities. All the received notifications together with their related recommendations are stored into the DARCAP system. Moreover Istat's experts may analyze the information content of the new designed administrative data sources and forms, as they do for already existing data sources and forms.

All the above described activities are coordinated by a Committee for Harmonizing Administrative Forms whose members are nominated by Istat and the most important administrative data sources' owner institutions, which is supported by a Network of experts.

Standardizing the Administrative Data Source Documentation

We have noticed that the availability of standardized documentation is an important prerequisite for properly exploiting the available administrative data sources.

Survey methodologies generally do not give directions for the preliminary phase of defining the survey's information content. Moreover it is theoretically assumed that there is one main observed population. On the contrary comparing the information content of available administrative sources is a complex task, that requires proper conceptual analysis capabilities. In fact the overall information content of any administrative data source, in terms of observed collectives, may be defined in a complex way, because the administrative sources collect data concerning events which occur in the course of time, and may observe several related populations due to the purpose of the administrative activity.

Similar remarks apply to the administrative data quality assessment. Due to the complexity of the administrative data sources' information content, we should equip any statistical user with proper methodological directions for enumerating possible errors and estimating their effects on any part of any administrative data source's information content. We need standardized procedures for a detailed enough practical assessment of any administrative data source's quality.

From a practical viewpoint we can satisfy such requirements for a detailed conceptual analysis of the administrative data sources' information content and quality by means of borrowing conceptual modelling notions and methods from computer science research and practice and embodying them in proper methodologies aimed at supporting the statistical utilization of administrative data sources.

From a theoretical viewpoint we should face a more important challenge. We have not yet a general approach for defining a probabilistic model of non-sample error for any given data source, a problem which acquire more importance when we have not an empirical control over the data collecting process.

An Ontology-Based Approach

In our approach, the administrative data source's information content is documented by means of defining the data source's own ontology. An ontology of an administrative data source is a structured description of its information content, based on a standard conceptual model.

Our Framework of quality indicators for administrative data sources organizes the quality indicators according to the well-established quality model which has been proposed by Statistics Netherlands (Daas

et al. 2009). However its distinguished feature is that it specifies a detailed set of indicators which are defined on the basis of the data source's ontology specification, particularly the quantitative indicators in the Data hyperdimension.

Our approach is innovative because the ontology-based description of the content of the available data sources is not yet a familiar documentation method among statisticians, despite the fact that today the ontology-based data documentation is a widespread practice in the database management field.

By means of anchoring the proposed indicators to the data sources' ontology we ensure a systematic specification of the indicators and we provide the quality evaluators with directions for choosing among indicators as well as for interpreting the calculated indicators, according to their particular requirements. Given that various distinguished factors influence the administrative data sources' quality, this is the only way for leading the quality evaluators in producing a standard assessment of the administrative data source's quality by applying standard and repeatable procedures.

Our Adopted Conceptual Model

For the purpose of a deeper analysis of the quality of administrative data sources, Istat is studying a more complete tool, namely the Quality Assessment Framework for Administrative Data Sources (D'Angiolini et al. 2013). It should be a handbook for driving anyone outside or inside a NSI, particularly the administrative data source's owners themselves, to assess the quality of any given administrative data source.

In order to meet such a requirement we are specifying proper quality indicators for each one of the different kinds of observed objects which form any data source's ontology. In such a way we ensure a systematic specification of the indicators and we provide the quality evaluators with directions for choosing among calculable indicators as well as for interpreting the calculated indicators.

The Framework is organized according to the well-known structure that has been proposed by Statistics Netherlands (Vis-Visschers 2009), which distinguishes three different views on quality, namely the Source view, the Metadata view, and the Data view. Each of these views called "hyperdimension" encompasses a number of dimensions and quality indicators, which may be qualitative indicators or quantitative indicators, the latter ones calculated from the source's data.

The first two hyperdimensions mainly concern the administrative data source as a whole. For such hyperdimensions the Framework proposes a set of qualitative indicators which are similar to those ones

which have been proposed in the BLUE-ETS project (Daas et al. 2011).

Our main effort is directed towards defining a richer and more structured set of quantitative indicators for the Data hyperdimension, which encompasses the traditional data quality measures such as the coverage of the observed collectives and the accuracy of the collected values for the observed characteristics.

We are defining such quality indicators on the basis of a careful analysis of the possible errors which may affect any observed collective (population or set of events), any observed characteristic, any observed relationship.

In order to single out such errors, we consider that for each object in an ontology, namely a collective, a characteristic, or a relationship, we can build belonging statements concerning observed elements. The administrative data sources continuously collect and store data which are in fact proper combinations of such belonging statements.

Therefore at any given time we may have in the administrative data source Inclusion errors, namely false belonging statements (definitely or temporarily) accepted in the data source, or Exclusion errors, namely true belonging statements (definitely or temporarily) excluded from the data source, even coincidental. Other errors may concern wrong identification of the involved elements, because of problems in the identification code system. We are analyzing and enumerating such possible errors and their combination for collectives, characteristics, relationships.

In order to define quality indicators, we put such errors into correspondence with the available quality check methods which are mainly: searching evident errors, such as duplicate identification codes, linking with other data sources, using logical constraints (mandatory or incompatible combinations between various belonging statements) and calculating delays between the moment events occur and the moment of their registration.

Until now, we've defined a quality indicators' frame concerning the collectives' coverage and the elements' identification by means of properly combining possible errors and quality checks methods. We are now analyzing the possible errors on characteristics and relationships in order to define two other quality indicators' frames concerning all kinds of nonresponses, measure errors, relationship errors.

Note that our proposed indicators are distinctly calculable for each collective, characteristic and relationship in the administrative data source's ontology, in order to effectively support any statistical usage of the collected information by any interested user.

Conclusions

Our experience is highlighting the various sources of complexity which make the specification of the administrative data sources' ontology and the analysis of the administrative data sources' quality two difficult activities. In the future, the standardization of such activities will provide the basis for automating particular evaluation tasks such as the reasoning on the derivability of new information from the existing administrative data sources and the comparison of quality indicators.

References

Benedikt, M., P. Bohannon, and G. Bruns. 2006. "Data Cleaning for Decision Support." In First Int'l VLDB Workshop on Clean Databases, September 11, 2006. Seoul. Available at: http://pike.psu.edu/cleandb06/papers/CameraReady_119.pdf.

Brackstone, G.J. 1987. "Issues in the use of administrative records for statistical purposes". In *Survey methodology*: Vol. 13 n.1 pp. 28-43. Available at: https://www.oecd.org/std/36237567.pdf.

Daas, P., S. Ossen, R. Vis-Visschers, and J. Arends-Tóth. 2009. *Checklist for the Quality evaluation of Administrative Data Sources*, The Hague: Statistics Netherlands.

Daas, P., S. Ossen, M. Tennekes, L.. Zhang, C. Hendriks, K. Foldal Haugen, F. Cerroni, G. Di Bella, T. Laitila, and A. Wallgren. 2011. "Report on methods preferred for the quality indicators of administrative data sources." In *BLUE-Enterprise and Trade Statistics*. European Commission. (Deliverable 4.2).

D'Angiolini, G., P., De Salvo, A. Passacantilli, and F. Pogelli. 2013. *Framework per la qualitàde gli archivi amministrativi* (Framework for the quality of administrative databases). Rome: Italian National Institute of Statistics.

D'Angiolini, G., P. De Salvo, A. Passacantilli, E. Patruno, T. Saccoccio, C. De Rosa, and E. Valente. 2014. "DARCAP: a tool for documenting the information content and the quality of the available administrative databases". In European conference on quality in official statistics, June 2-5, 2014. Vienna.

D'Angiolini, G., P., De Salvo, and A. Passacantilli. In press. "Istat's new strategy and tools for enhancing statistical utilization of the available administrative databases." *Statistika: Statistics and Economy Journal*.

UNECE. 2011. *Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices*. Geneva: United Nations Economic Commission for Europe Press.

Published: Volume 2017, Issue 1 / January 25, 2017