ANALYSIS-SENSITIVE CONVERSION OF ADMINISTRATIVE DATA INTO STATISTICAL INFORMATION SYSTEMS

Mirko Cesarini, Mariagrazia Fugini,

Politecnico di Milano, Dipartimento di Elettronica e Informazione
Via Ponzio, 34/5 I-20133 MILANO, Italy

Mario Mezzanzanica
Università degli Studi di Milano-Bicocca, Dipartimento di Statistica
Via Bicocca degli Arcimboldi 8, I-20126 MILANO, Italy

Keywords: Statistical Information Systems, Taxation Archives, Decision Support Systems, Data Quality, Integrating Heterogeneous Data Sources, Data Warehouse.

Abstract: In this paper we present a methodological approach to develop a Statistical Information System (SIS), out of administrative archives of the Public Administrations. Such archives are a rich source of information, but an attempt to use them as sources for statistical analysis reveals errors and incompatibilities that do not permit their usage as a statistical and decision support basis. The proposed methodological approach encompasses building a SIS out of administrative data, such as design of an integration model for different and heterogeneous data sources, improvement of the overall data quality, removal of errors that might impact on the correctness of statistical analysis, design of a data warehouse for statistical analysis, and design of a multidimensional database to develop indicators for decision support. We present a case study, the AMeRIcA Project.

1 INTRODUCTION

Public Administrations (PA) are facing institutional and organizational changes requiring managers, stakeholders, and politicians to increase quick decision making processes. A key role is assumed by the development of *Statistical Information Systems* (SIS) aimed at providing support for decisions, analysis, monitoring, and control activities. In particular, data deriving from administrative sources (e.g., government registries, tax registries) assume a basic value to gather information concerning the community and to feed the SIS. However, administrative data are often incorrect and unsuitable to be used for statistics and decision making. Hence, they need to be cleaned up from errors, and pre-processed before being reversed into statistical databases. This paper illustrates the AMeRIcA project (Anagrafe Milanese e Redditi Individuali con Archivi - Milan Registry Office and Individual Income with Archives), where the administrative archives available from the Registry Office of the Milan Municipality and of the Italian Income Office are used to derive statistical information about actual income of subjects and families in Milan. Some experiences show that the integrated use of tax-related databases together with Registry databases enables to obtain rich information (Statistics Denmark, 2000). In such streamline, AMeRIcA, applies statistical analysis to data gathered from PA *administrative sources* (representative of the whole population) rather than to *sample surveys*. An innovative aspect of AMeRIcA from the statistical and the ICT viewpoints is the use of a Data Warehouse designed to integrate different administrative sources. This enables to apply statistical analysis models encompassing different facts of the whole population, deriving in this way significant and accurate results in terms of the observed universe.

2 BUILDING A STATISTICAL INFORMATION SYSTEM

Within an organization, a SIS is loaded and continuously fed using data sources derived from the administrative and management systems. A SIS has two main purposes (UNECE, 2000): to support decision-making processes through the construction of *directional indicators* which are the final result of data collection, analysis, and processing activities; to return information to the management systems useful for update, evolution and quality management along time.

The first operation to be performed to build a

Cesarini M., Fugini M. and Mezzanzanica M. (2006).
ANALYSIS-SENSITIVE CONVERSION OF ADMINISTRATIVE DATA INTO STATISTICAL INFORMATION SYSTEMS.
In *Proceedings of the Eighth International Conference on Enterprise Information Systems - DISI*, pages 293-296
DOI: 10.5220/0002494402930296

SIS is a detailed study of the source archives. The data sources quality should be checked and some data cleaning operations should be performed in order to remove all possible errors that might negatively impact the statistical analysis. Then archives are checked for cross inconsistencies, and finally data are integrated in a global archive.

2.1 Data Integration and Cleaning

The first steps required to build a SIS are a detailed analysis of the archives and the development of a global integration schema which will drive the subsequent steps. Further activities are the establishment of a mapping schema between the global integrated schema and the single archive schemas (local schemas). Finally the steps of a process of data migration towards the integrated archive should be detailed. During data migration some low quality data issues might occur and should be resolved, as we will show in Sec. 2.2. Moreover, data loaded into the global integration schema instance might reveal unsuitable for the analysis leading to misinterpretations. For this reason the SIS development process should be an iterative one, with the aim of progressively tuning the global integration schema and the migration procedure. Moreover, schemas may not completely capture the semantics of the data that they describe, and there may be several plausible mappings between two schemas. This subjectivity makes it valuable to have user input to guide the match and essential to have user validation of the result.

2.2 Data Quality Improvement

The main problem in using administrative databases for statistical and decision making purposes is the presence of errors that do not affect the regular use of the archive for administrative purposes. Such errors are hardly noticed, and, even when discovered, they are usually tolerated. However, this errors and low quality of data can negatively affect statistical analysis. Therefore, data sources need to undergo a *quality improvement pre-processing* before being an input for any kind of analysis. Administrative databases are employed to access information describing a *single item* at a time (e.g., the address of a person), while statistical analysis deals with *collection of items* (e.g. how many people live within an area). This different usage of archives may unveil simple errors like duplicate records, or more complex ones, e.g. some inhabitants that are registered in the Registry Office of a neighbour town and not in the town where they live. Some of the problems may be fixed by performing *data cleaning actions* whose results have a certain degree of reliability, therefore requiring manual evaluation employing various data quality metrics such as accuracy, consistency, completeness, timeliness, and so on (*integration quality* criteria). Many cleaning techniques can be used, we won't investigate this topic anymore, we would like to highlight that these techniques have different *costs* in term of execution time required (both to humans and computers) and "optimal mix selection" issues arise when resources are scarce. The optimal mix selection is performed by evaluating an *execution cost* and a *quality improvement rate* for each candidate operation. The estimation of both values is a heuristic operation, based on experience as well.

3 THE AMeRIcA PROJECT

The concepts illustrated are presented for the AMeRIcA Project. The approach comprises various and independent phases: from data integration and quality analysis, to the definition of statistical indicators, via the analysis of information sources, database design, transformation and data management process, and definition of a multidimensional model for data analysis as a decisional support. The reference population is provided by the Registry of the Milan Municipality. Data on such population are fundamental, since it is impossible to obtain a data provisioning from the Income Office bounded to a geographic area. A cross reference between the Registry Archives and the Income Archives allows one to obtain the desired information. The process of data interpretation, cleaning, and normalization, applied both to single source and to integrated data, has required a great effort and a deep data domain knowledge.

The Income Archive holds also some registry information about people, however preference has been given to data derived from the Registry Archive, since it is usually more up to date. In fact, an individual notifies address changes to the Registry Office quickly, while the Income Office is notified once per year with the tax declaration form. Records describing the same person in different archives are identified by the Fiscal Code (FC, similar to the US Social Security Number). Once different records on the same individual have been identified, further information (e.g., profession, qualification, education, and so on) significant for analysis and not present in the Income Archive, may be used. However, the scarce freshness of some archives would violate the information quality criteria; thus, such additional information has not been included in the analysis. The portion of data in the AMeRIcA SIS coming from the Income Office refers to the income returns of both companies and people. Individuals declare income data by filling in different forms, according to the received type of income and properties. Three common basic *macro-information*

types can be identified: the total incomes grouped by income source; the deductions and detractions; the physical person taxation necessary to determine the tax drag. Around this information core, an integration model has been constructed able to drive the migration process and to highlight information relevant for statistical analysis. Once the integration model has been selected, the delivered archive undergoes a preprocessing aimed at improving the *quality and reliability* of information, and aimed at framing the classifications to the adopted standards. Two types of preprocessing procedures are used: *semantic* and *syntactic cleaning*. Hence, *two different integration levels* can be identified: integration at a single archive level, regarding provisioning over different years, and integration at a global level where different archives are involved. 1) *Integration at a single archive level*: Provisions over different years of the same archive can comprise heterogeneous information and hence must be reconciled to a unique data model taking into account information common to the different deliveries. The *selection* of the common information is *driven by the analysis* to be performed later, privileging relevant information or data present over different years, and hence comparable. A meaningful example in this case is the delivery of an archive from the Income Office: in the considered years, the tax laws have undergone many changes which caused the information record of tax income to change every year. 2) *Integration in the system*: this includes the link among different information, coming from distinct sources. The goal is to enrich the information content of the subjects to be analyzed (and consequently the range of possible queries) by collecting different information about the same subject that are scattered among different sources. The process described in the previous steps can be summarized in terms of the flow reported in Fig. 1.

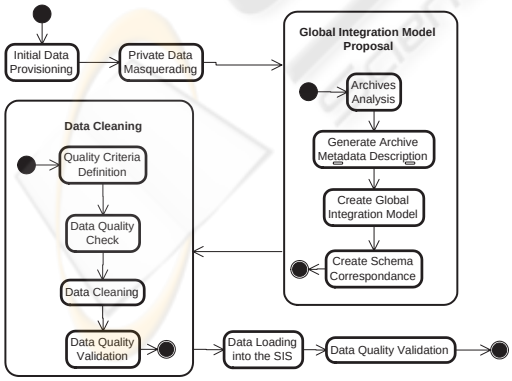

Figure 1: Data loading and cleaning workflow.

Due to the increasing number of treated information and of examined subjects, it is desirable to perform a selection/aggregation of information to be used for further analysis. For example, it is possible to identify family groups (using Registry data), and to aggregate the income revenue for the whole family group. The level of aggregation/selection of the information requires a trade off between the computation required by and the desired granularity of the analysis. Anyway by acquiring knowledge and some more data about the field of analysis, it is possible to reprocess data in order to build more suitable aggregations/selections.

3.1 Multidimensional Data Model

The adoption of a multidimensional data model at information source integration time introduces and outlines the statistical information needs of the project. Both the data model designed for each single source and the global model put into evidence a set of possible subject of analysis, and a set of dimensions along which the analysis can be performed, aggregating or detailing the information, according to the different analysis needs. To favor subsequent analysis, the final phase of data model design includes the definition of *facts of interest*, of their respective *dimensions*, and of *aggregation levels* along which combining the data. An example is illustrated in the Fig. 2.

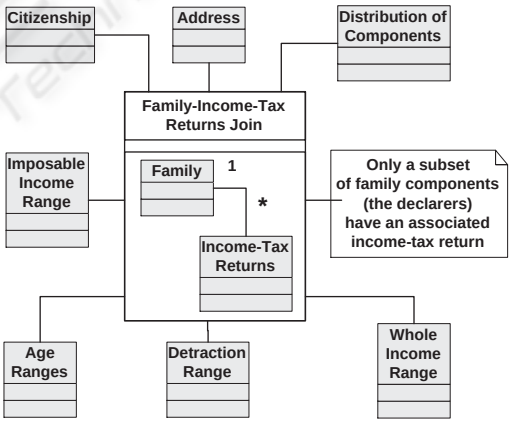

Figure 2: Facts: Family-IncomeTax.

The AMeRIcA Data Warehouse hosts the results of the integration phases. Data are organized along facts of interest for the analysis as reported in Fig. 3.

4 CONCLUDING REMARKS AND RELATED WORK

A strict link exists between an administrative and management system and a SIS. Such consideration al-

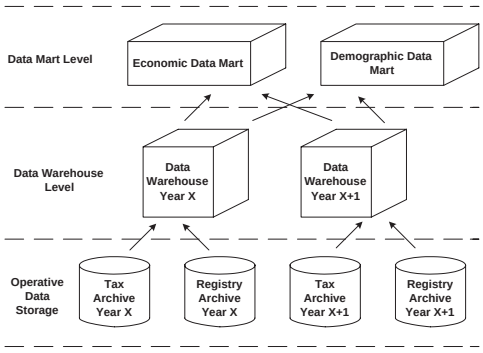

Figure 3: The AMeRIcA Data Warehouse.

lows to outline the *guidelines to define decision making policies*. In recent years, the reuse of statistical data (Hoffmann, 1995), (Thomson and Holmy, 1998) has increased the demand for easy access to a variety of pre-existing data sources (Sundgren, 1996). Some works address the integration of existing data sources of national or regional statistical offices, or providers of comparable nature (Denk and Froeschl, 2000), (Hatzopoulos et al., 1998). Other works leverage metadata classification to drive data integration and elaboration (Papageorgiou et al., 2001); another category of works refer to quality of data (IQ1, 2005), and specific quality assurance for census data (Census, 2005). An attempt to feed a SIS using PA's or large enterprises' archives is reported in (Buzzigoli, 2002) for efficient information system integration in a PA structure (e.g., the census of archives within an administration). However a discussion concerning quality of data, consistency, and archive integration issues is still missing. The link between an administrative and management system and the SIS is bidirectional: the administrative, management system feeds the SIS, while the SIS provides indications to the administrative and management one to support ameliorations along time. Such link is strong, although poorly implemented in practice. Administrative systems are designed using an auto-referential logic that privileges the definition of services functional to the organizational model rather than to the stakeholders or to the statisticians. This reflects in expensive activities to normalize, ensure data quality and standardization as required. An enabling factor for SIS construction is the ability of a PA to take into account the transversally and reciprocal acknowledgement of concepts, even if used in different administrative processes, and to obtain that such concepts are in relation with standard codifications. Another factor is related to the quality of documentation provided by the sources which is often scarce, or not present, making the SIS conceptual design harder. A current development of AMeRIcA regards the use of social security data. Using social security data owned by employment centres, it will be possible to correctly identify the available wealth of a larger set of citizens.

REFERENCES

Buzzigoli, L. (2002). The new role of statistics in local public administration. In *Proceedings of the Conference Quantitative Methods in Economics (multiple Criteria Decision Making XI)*, pages 28–34, Faculty of Economics and Management, Slovak Agricultural University, Nitra (SK).

Census (2005). Census bureau section 515 information quality guidelines, OFFICE OF MANAGEMENT AND BUDGET, guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of information disseminated by federal agencies. Available at `http://www.census.gov/quality/`.

Denk, M. and Froeschl, K. (2000). The *IDARESA* data mediation architecture for statistical aggregates. *Research in Official Statistics*, 3(1):pp.7–38.

Hatzopoulos, M., Karali, I., and Viglas, E. (1998). Attacking diversity in *NSIs' Storage Infrastructure*: The *ADDSIA* approach. In *Proceeding of International Seminar on New Techniques and Technologies in Statistics*, pages 229–234, Sorrento (IT).

Hoffmann, E. (1995). We must use administrative data for official statistics - but how should we use them? *Statistical Journal of the United Nations/ECE*, 12:pp. 41–48.

IQ1 (2005). Information quality I, 2005. Principles and foundation, the MIT total data quality management program. Available at `http://web.mit.edu/tdqm/www/index.shtml`.

Papageorgiou, H., Pentaris, F., Theodorou, E., Vardaki, M., and Petrakos, M. (2001). A statistical metadata model for simultaneous manipulation of both data and metadata. *J. Intell. Inf. Syst.*, 17(2-3):pp. 169–192.

Statistics Denmark (2000). The use of administrative sources for statistics and international comparability (invited paper). In *Conference of European Statisticians, 48th plenary session*, Paris (FR). Statistical Commission and Economic Commission for Europe.

Sundgren, B. (1996). Making statistical data more available. *International Statistical Review*, 64(1):pp. 23–38.

Thomson, I. and Holmy, A. (1998). Combining data from surveys and administrative record systems - the norwegian experience. *International Statistical Review*, 66(2):pp. 201–221.

UNECE (2000). Statistical metadata. In *Conference on European Statisticians Statistical Standards and Studies - No. 53*, Geneva (CH).