

Statistical Information Systems and Data Warehouses for Job Marketplaces

Sistemi informativi statistici e Data Warehouse per il mercato del lavoro

Paolo Mariani, Mario Mezzanzanica, Biancamaria Zavanella

Università degli Studi di Milano-Bicocca, Dipartimento di Statistica,

Via Bicocca degli Arcimboldi 8, I-20126 MILANO, Italy

e-mail: mario.mezzanzanica@unimib.it

Riassunto: Questo lavoro presenta un progetto di integrazione di tre archivi amministrativi (Anagrafe comunale, Anagrafe tributaria e Centri per l'impiego di Milano), che fornisce dati di estremo interesse ai fini dello studio della situazione socio-economica dei cittadini milanesi. I dati ottenuti dopo un processo di cleaning, descrivono la popolazione interessata con un elevato livello di dettaglio e possono essere utilizzati per svolgere analisi molto specifiche, significative anche se condotte su sottoinsiemi ristretti della popolazione in oggetto. L'analisi riportata ha valore esclusivamente esemplificativo delle potenzialità del progetto e riguarda la distribuzione per età, per reddito e per tipo di contratto, dei cittadini milanesi con età compresa tra 20 e 49 anni, che nell'intervallo di tempo considerato (2000-2003) hanno perso, ritrovato e perso definitivamente il lavoro.

Keywords: statistical information systems, taxation archives, data quality, integrating heterogeneous data sources, data warehouses.

1. Introduction

Traditionally the job search market has been mostly prerogative of specialized entities (private/public employment agencies, head-hunting and job search companies). However its social implications and enormous impact onto the political/strategic decisions of governments have more and more seen an increasing implication of National and Local Public Administrations (PA).

In this context, information related to individual and collective needs plays a large and relevant role, therefore a deep knowledge of the job marketplace phenomenon is required.

Data deriving from administrative sources (e.g. government registries, tax registries, etc.) assume a basic value to gather information concerning the community. Moreover, the introduction of the information and communication technologies has enlarged and empowered the availability and fruition of administrative databases, making information accessible to Organizations and Institutions for further surveys (Sundgren, 1996).

The Administrative archives managed by PA nowadays contain data describing the whole population with fine grained details about each individual. These administrative archives are a rich source of information for statistical analysis and allow to perform meaningful and detailed analysis both on the whole population and on small subsets. Furthermore, by using heterogeneous archives coming from different administrations and describing different domains, it is possible to analyze a reality from many points of view, achieving in this way a deep knowledge of the observed phenomena.

In these scenarios, a key role is assumed by the development of *Statistical Information Systems* (SIS). An official definition of a SIS is given in (UNECE, 2000) where a SIS is described as an informative system aimed at collecting, storing, managing and distributing statistical information. A SIS is usually built starting from heterogeneous data sources, and in particular, using the data stored in *administrative archives*.

The main problem in using administrative databases for statistical and decision making purposes is the presence of errors, duplications, inconsistencies, and unsuitability. Such errors do not affect the regular use of the archives for administrative purposes, therefore they are hardly noticed, and, even when discovered, they are usually tolerated. However these errors and low quality of data (Aimetti and Zavanella, 2004) do negatively affect statistical analysis. Therefore, before being an input for any kind of contents analysis, the data sources need to undergo a *quality improvement pre-processing* which is performed by applying a set of data cleaning actions, each one fixing a specific problem. The output archives (namely the archives being the outcome of the cleaning actions) are subsequently integrated into a global one.

The result of the cleaning and integration operations is an integrated and huge archive of data where the statistical analysis can be performed. Such a huge amount of data (embracing different aspects of the same reality) is usually stored in a Data Warehouse which facilitates further analysis and the execution of queries aimed at computing synthetic information. Data Warehouses are software tools useful to manage data framed according to a multidimensional model composed of *facts* and *dimensions*. A *fact* is (one of) the main topic of interest in a set of data (e.g. the individual in a set of data concerning job information); *topics* or *analysis perspectives* that are connected via associations to facts are called *dimensions*, and they are usually the focus of the data warehouse analysis (in the previous example the dimensions are: the working history of an individual, academic and school experiences, employment status, ...).

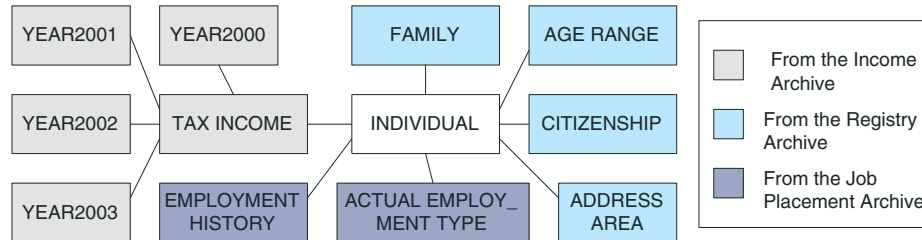
2. Statistical Information Systems supporting the Job Marketplace

In this section we are going to present an example of analysis that has been performed by integrating data coming from the Registry Office, from the Income Office, and from the Job Placement Office of the Milan Municipality.

Some international experiences show that the integrated use of tax-related databases together with Registry databases enable to obtain rich information regarding both income trends and citizens (Statistics Denmark, 2000). Still, the only available data in the context of national statistics about the investigated field come from the national statistic institutes and are constructed on the basis of *sample surveys*: the significance of these data decreases relevantly when small territorial environments are investigated, like the municipality of Milan which is the focus of our analysis.

The Income Office archives collect data regarding the tax system; the Registry Office collects data about citizens' residence, marital status, and family relationships; The Job Placement Office collects information about curriculum, job beginnings and cessations, and type of employment contract regarding employed people (excluding Public Administration ones). The Data Warehouse schema used to manage the data coming from the archives is reported in Figure 1. In the remaining of this section, with the aim of showing the potentiality provided by the SIS here described, we are exposing an

Figure 1: *The Data Warehouse Schema. “Individual” is the Fact, the surrounding Boxes are the “Dimensions”.*

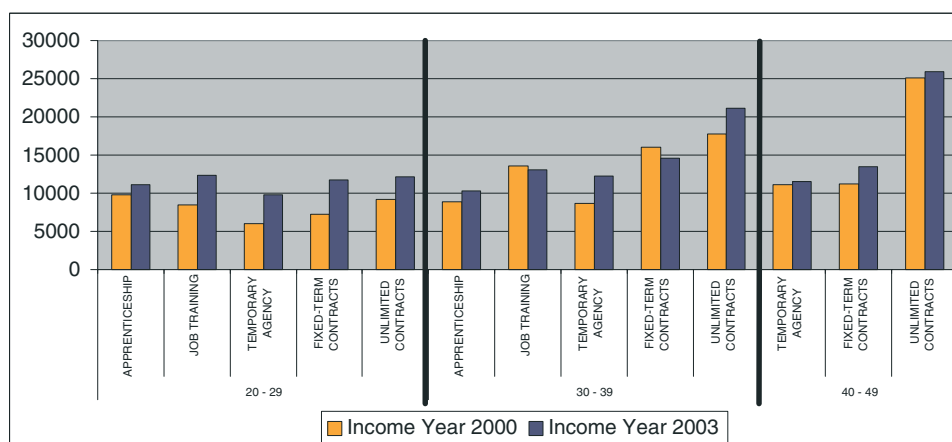


analysis performed by using the Data Warehouse built by the Department of Statistics of the University of Milano-Bicocca integrating the aforementioned administrative archives.

As a first investigation of the SIS we have built, we have done a sample of people in working age (20-49 years old) who have lost, found, and then definitively lost a job in the time interval ranging from 2000 to 2003. In fact, among the 60,000 persons who have changed their job in that time interval, we have identified a subset of approximately 28,500 people, which represents nearly the 5% of the Milanese population having the same age range. The identified people have been classified in 3 age classes: 20-29 years old (8,909 people, ~ 31%), 30-39 years old (11,361 people, ~ 40%), 40-49 years old (8,319 people, ~ 29%). Concerning the age distribution, none of the identified age classes shows particular suffering conditions, since the distribution of the classes is almost identical to the population's one. The subset has been also classified according to the type of employee contract used. According to the Italian law, the following type of contracts are considered: “Apprenticeship” (Apprendistato), “Job Training Contracts” (Contratto Formazione Lavoro), “Temporary Agency Contracts” (Interinale), “Fixed-term Contracts” (Tempo determinato), “Unlimited Contracts” (Tempo Indeterminato).

The graphic in Figure 2 shows the distribution of average incomes, grouped by age class and by contract type. Basing on the diagram reported, some considerations can be drawn: the graphic shows that the income gets higher as the working age increases. Moreover, the average income for young people (namely people between 20 and 29 years old) is almost independent from the contract type; on the contrary, in the 40-49 years class we can observe that people with unlimited contracts have a quite higher income than ones having atypical-flexible contracts (namely: Temporary Agency and Fixed-term Contracts). The approach just used provides some hints for further analysis: an atypical-flexible contract in the 40-49 years class, seems to introduce instability in the working life of an individual (although in the considered subset of people, the unlimited contracts have an average duration of approximately one year, therefore the unlimited one doesn't seem to provide more security with respect to the atypical-flexible ones), and it is characterized by an average income quite lower with respect to the average income of the unlimited contracts in the same age class. This phenomenon requires further screening and analysis, with the aim to provide information to politicians and to people in charge of enacting active policies and to support the entry in the job market place of people experiencing difficulties.

Figure 2: Average Income by Age Classes and by Working Contract Type.



3. Concluding remarks and Future Work

As previously remarked, the aim of this paper is only to highlight the potentialities of the Statistical Information System introduced in the previous section; in the future, the research activity may prosecute toward many directions. The first one will be a further investigation of the subset this paper has focused on; by acquiring data of more and more years it will be possible to track and to study the working career and the income modifications of the individuals composing the subset. A further research step will be the extraction and analysis of different subsets of persons. Moreover, the Data Warehouse that has been built to store the data allows to perform specific and thematic investigations, according to the information needs expressed by the people responsible of enacting active policies supporting the job market place.

References

- Aimetti P., Zavanella, B. (2004) *Qualità degli archivi amministrativi e qualità dell'informazione statistica*, Franco Angeli, Milano.
- 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), 23-38.
- UNECE (2000) Statistical metadata, in *Conference on European Statisticians Statistical Standards and Studies - N. 53*, Geneva (CH).