

Towards Integrated Architectures of Administrative Information Systems: New Opportunities for the Territorial Statistical Analysis

*Verso Architetture Integrate di Sistemi Informativi Amministrativi: Nuove
Opportunità per l'Analisi Statistica Territoriale*

Biancamaria Zavanella

Dipartimento di Statistica, Università degli Studi di Milano Bicocca, Via Bicocca degli
Arcimboldi, 8 - 20126 Milano, e-mail: biancamaria.zavanella@unimib.it

Riassunto: In seguito alla Legge n. 127 15/5/97 e alla grande disponibilità di nuove tecnologie informatiche, la Pubblica Amministrazione si è mossa verso obiettivi di semplificazione delle procedure amministrative. Su questa strada stanno nascendo molti progetti innovativi, volti alla costruzione di architetture integrate di sistemi informativi amministrativi, per i quali è essenziale l'apporto dello statistico in tutte le fasi costitutive, dalla progettazione all'utilizzo dei dati, sia a fini di controllo dei processi amministrativi e di servizio cui i dati si riferiscono, sia a fini di analisi statistiche su tematiche connesse ai dati medesimi, che sono del tutto innovativi e di buona qualità. In questo lavoro sono presentati e discussi alcuni esempi significativi in questo campo: il sistema integrato di archivi amministrativi (Unioncamere, INPS, INAIL, Ministeri dell'industria e delle finanze) l'integrazione tra Anagrafe comunale e dati del Ministero delle finanze (Comune di Milano e Università di Milano Bicocca)

Keywords: Sistemi informativi Amministrativi, Data Warehouse, Analisi territoriale

1. Introduction

1996 was a breakthrough year for the supply of statistical data. In such a year, indeed, the first statistical archive ASIA96 became operative. On its base it has been possible to realize the “*Censimento Intermedio dell'industria e dei Servizi*” (CIS96), the business census 1996, with quite innovative methodologies. In the same year the “*Registro delle Imprese*”, the business register, with the connected “*Repertorio delle notizie economiche e amministrative*” (REA), the inventory of economical and administrative information, in fulfilment of law 580/93, have been born. They had never been realized before, although provided for by the civil code.

The building of such statistical archives, carried out through the integration of various administrative archives and survey sources, was first imagined on account of the need of having, also at a local level, complete and detailed data, such as the Census' data, but updated more frequently. This need was generated by the speeding up of structural changes in the economy, which happen no more in the “long run”, but rather tend to overlap the short-term fluctuations.

A further goal in the building of statistical archives consists in making possible surveys, aimed at knowing the new dynamics of the economical systems, in order to carry out local dynamic analysis and forecasts: a good example is the analysis of local profession-markets, essential to realize employment and training policies, which may take into account the quick structural changes started in the 90's. At this purpose remember the

Excelsior project, which will not be discussed here, but whose good performances are well known.

In recent times two important innovation factors are becoming effective, bringing positive effects in the way administrative data can be used for statistical goals: the availability of new technologies and the renewed institutional framework.

Data Warehousing technologies were not yet available, when the first integrated statistical archives were built; their introduction and diffusion has brought a great value in term of quality, accessibility and usability into the procedures of integration of administrative archives. The organization by subject of the data into a Data Warehouse (DW) opens many doors to the statisticians, in connection with the building of the DW itself, the exploiting of an amount of data not available before in such a detail and the study of new methodologies for Data Mining.

The changed institutional framework (Bassanini Law 127 15/05/1997, constitutional reform Law 18/10/2001), together with the mentioned new DW technologies, has lead the Civil Service (CS) into a radically new path, whose keywords are: decentralization, simplification and customer-satisfaction orientation. This has brought to the building of extremely innovative architectures and of integrated information systems.

This new attitude of the CS is carrying out a new system, which is able to bring positive changes in the availability of data, and great statistical potentialities, in particular as far as the analysis of the territorial systems is concerned. Also the role of the statistician is taking a new form, particularly in the following six directions:

1. in the logical project of the architectures and of the metadata (definition of the units, of the states and of the events);
2. in the assessment of the quality of administrative data and of the relative processes of information, collection and storage into archives;
3. in the analysis of the quality of the internal processes and of the services of the various bodies (in a total-quality viewpoint);
4. in the joint analysis of different characteristics of the same unit, coming from different administrative archives and at different times;
5. in the realization of surveys to examine more specifically some themes;
6. in the statistical exploitation of partial statistical data, which can anticipate trends (for example declarations sent through Internet from a subset of businesses, thus, readily available).

In the present work it will be illustrated the potentiality of innovation in terms of new fields of research for the statistical methodology, quality and availability of data, possible contributions to classical statistical procedures, made possible by the two aforementioned innovative factors. The possibilities of statistical development induced by these two causes, apparently exogenous to the statistical world in its stricter meaning, will be presented by means of two examples, concerning two innovative projects: i) the REA (*Registro degli Agenti Economici*, register of economical agents) project, realized by Unioncamere, INPS and INAIL with the contribution of AIPA (*Autorità per l'informatica nella pubblica amministrazione*, Authority for informatics in Civil Service), ii) the project of integration of fiscal data (Ministry of Finance) with the population register (City of Milan).

2. New Data Warehousing technologies

The goal of the making of a DW is the construction of an environment, which is separate from the source archives, thought for the centralization of organized data, and such as to allow different levels of data aggregation. These aggregation capabilities answer to all the possible queries, which different analysts may question in the flow of their decision process.

The main feature of a DW, synthetically describable through its definition “subject oriented, integrated, not time-volatile, management-supporting collection of data” (W. H. Immon 1996), is allowing the statistical exploitation of data from different operative systems of administration.

The data inside a DW are therefore:

- Integrated, consistent and well established, as they are homogenous even when coming from different sources, using the same units of measure, uniform methods for codifying and pursuing a semantic homogeneity of all variables.
- Organized as “subject oriented”, that is for the final use in terms of analysis; hence it is built a structure able to favour the user in reading and eventual elaborations of data, for example imagining some homogenous subsets at various levels of aggregation, by using proper instrumental variables of hierarchical kind, that are consistent with the possible statistical analyses.
- Characterized by the information history, as they have a lifetime much wider than data base: the data of an archive usually represent only the current information, then cancelled and replaced by new data entry; on the contrary, in the DW every variation in the data is kept, in order to have marks of every correction or every variation in the process of data organization too.
- They are not volatile, that is they cannot be modified a posteriori in any way and the only recognized operations from final user are the reading and the answering, while the DW administrator can take part in the process of data production and modify it, but not cancel historical data.

From the data's quality point of view, the process of integration, normalization and imputation of missing data occurring in the making of a DW, is very similar to that of a classic data base; the thorough difference of the DW from a database is its characteristics of accessibility and exploitation of the data. The idea behind the DW is, in fact, to create a warehouse, physically independent from the base archives of the various systems integrated in it, “open to any query”, this meaning that there is no outline of questions a priori rigidly defined, but on the contrary every user can dynamically build his own interrogation to the system. Moreover, the answers of the system are consistent, because the answer does not depend on the way the Query is constructed and different users must be able to obtain the same answers to the same question, even constructing formally different questions.

One of the particularly innovative aspects of the DW is constituted by the multidimensionality of the data organization, allowing to carry out outlooks on data themselves from multiple views, with the possibility of navigating in all dimensions without any restrictions on the system's functionalities. Indeed, it is said that on the hypercube representing the data it is possible to apply the “slice and dice”.

But the most statistically relevant feature in the making of a DW is the data transposition phase, where the records of the source database (DB) are arranged in a matrix, whose rows represent the elementary events in the DB, while the columns

contain the variables, which the DW-designer hold as relevant for the statistical exploitation of the DW. This phase, just like the standardization phase of the nomenclatures and classifications, basically requires the intellectual direction of the statistician, who has the sensibility and the indispensable experience to organize data for their analysis.

3. Integrated architectures of administrative information systems

A transformation process of the organization system in the CS is currently ongoing similarly to what happened in the industrial market during the 1990s with the introduction of the BPR (Business Process Reengineering) theories; these introduced a vision on new I&CT (Information & Communication Technology) technologies as a fundamental support to services' process innovation. Important CS, both Central and Local, have started up information systems' restructuring projects aimed at innovating their own organizational areas through the introduction of market technological platforms; this, wanting to exploit both the technological and managerial new issues present in such platforms. In particular, starting from social-security non Economical Public Authorities to some important Ministries (as the Ministry of Finance) and Local Authorities (Regions and Municipalities), ERP (Enterprise Resource Planning) systems have been introduced in the past two years so as to restructure the internal management service processes pursuing integration and efficiency. Furthermore, significant experiences in the relationship management area with suppliers have been put forward with the start up of e-procurement and Market Place projects along with several initiatives aimed at improving the relationship with the users also through internet technologies, following principles of transparent information and administrative procedure simplification.

Particularly, the simplification process of the administrative procedures is able to affect in a relevant way the methodologies of collecting data for administrative archives, for example through the standardization and unification of the formats for the compulsory declarations in the constitution and transformation passages of the life of an enterprise. The opportunity of declaring, for example, the constitution of a new enterprise just one time, to a single body, even by internet and/or through a Publish&Subscribe system, can reduce in a sensible way the probability to make material errors.

Many of these innovative projects feature a DW technology system management component which allows on one side the controlling, the monitoring and the verifying of ongoing services relative both to economical and management figures; on the other side it allows data availability for statistical analyses and for the punctual issuing of information to the different user classes.

In this changing framework it is possible to identify a background in which the role of the statistician becomes more and more important in the rational construction of integrated architectures in the administrative field and to support the improvement of the organization of public utility services.

4. The RAE project

In year 2000, Unioncamere, INPS and INAIL, with the agreement of the Ministries of Industry and Finance, have promoted the RAE (Register of economical agents) project, aiming the simplification of the administrative relations between the economical agents (businesses, self-employed, non-profit) and the very promoting bodies. The building of the register has been carried out experimentally in year 2001, with the contribution of AIPA, and will be fully operative starting from June 2002.

The project is based on the following considerations:

- there exists a plurality of economical agents, subject to the action of different administrations (finance, social security, etc.), with partially overlapping fields of interest;
- the same agent is often required to declare changes of its state (for example the address, the ceasing or the variation of the activity, etc.) to every administration, with different modalities and forms and at different times.
- every administration activates accurate procedures of control, for those characters connected with specific obligations or rights (for examples payment of taxes, contributions, etc.), while tends to put a lesser attention to those features that have only a classificatory significance. Thus, the reliability of the different records of each archive is not homogeneous: for example, for INPS, the localization and the economic activity of a firm are not so relevant as the number of employees and their remunerations.

The RAE project is motivated by the following goals:

- virtual unification of the registers of the different administrations: the field of interest of the integrated virtual register consists in the union of the fields of interest of the different administrations;
- standardization of the characters, of the definitions and of the taxonomies, and identification of the administration of certification for each character;
- reconstruction of the stock of the economical agents (firms, self-employed workers, non-profit associations and relative local units) and control of the related constitutive and state characters;
- construction of a Publish&Subscribe system, which allow every administration that receive the declaration of a change to transmit it in real time to every other interested administration, and in particular to the certification body. For example, the communication of the ceasing of activity to any of the interested administration will be forwarded to all the interested bodies and particularly to the Business Register, which has the duty to definitively confirm the occurred cessation.

In the following tables some results about the treatment of the stocks are shown. Out of more that 20 millions records in the registers of the three administrations more than 4.5 millions result ceased, 14 millions concern firms and self-employed workers and about 1.2 millions relate to other bodies.

To every active position it has been attributed the constitutive characters and the state characters, which define their belonging to the fields of interest of the various administrations. These figures are the result of systematic operations of control of the quality of the coverage of every single character and archive. The next step has been the linking of the positions (table 2), in particular this has meant the punctual control carried out by each administration of about 580.000 problematic links, arisen in the

diverse phases of the procedure. These controls have been due to the administrative nature of the archive, which cannot rely only on probabilistic evaluations.

It is possible, particularly for the linked positions, to carry out cross-checks of the characters. As an example, table 3 gives an account of the check on the fiscal codes with the official one. Analogous indicators have been built for all the other characters.

Table 1: *Records extracted from the RAE reconstruction*

Archives	Total records treated	Active businesses or self-employed workers	Other active bodies	Ceased bodies
Business Register (RI)	6.611.555	5.914.032	0	697.523
INAIL register	7.065.365	3.051.973	322.851	3.690.541
INPS register	2.538.613	1.968.465	328.091	242.057
• <i>Businesses</i>	1.575.796	1.339.302	135.100	101.394
• <i>Employer co.co.co</i>	580.377	465.878	114.499	0
• <i>Farms</i>	382.440	163.285	78.492	140.663
INPS craftsmen	1.830.237	1.637.424	192.813	0
INPS retailers	1.741.197	1.489.172	252.025	0
INPS agriculturists	486.938	382.573	69.188	35.177
Totale record trattati	20.273.905	14.443.639	1.164.968	4.665.298

Table 2: *Active businesses and self-employed workers: linkage*

Archives	Total active records	Not linked	Linked	Linked RI-INAIL-INPS	Linked RI-INAIL	Linked RI-INPS	Linked INPS-INAIL
Businesses register (RI)	5.914.032	2.694.109	3.219.923	1.657.501	1.276.926	285.496	-
INAIL register	3.051.973	0	3.051.973	1.729.278	1.215.177	-	107.518
INPS register	1.968.465	11.191	1.957.274	1.586.123	-	282.478	88.673

Table 3: *Fiscal codes*

	Same code %	Different code %
Business Register (RI)	97,7	2,3
INAIL register	97,4	2,6
INPS businesses	97,8	2,2
INPS Employer co.co. co.	97,2	2,8
INPS Farms	97,7	2,3
INPS Craftsmen	66,7	33,3
INPS Retailers	65,8	34,2
INPS Agriculturists	93,2	6,8

The construction of the RAE brings many advantages both in the administrative and in the statistical frameworks.

On the administrative level, the RAE allows:

- reduction of the number of compulsory declarations;

- reduction of the time to collect, archive and control the data;
- immediate control of coherence of the characters needed by the different administrations;
- reduction of the costs of declaration and control both for the economical agents and for the administrations;
- simplification of the relations between the economical agents and the Civil Service.

On a strictly statistical level the RAE gives the up-to-date picture of the general field of observation and of the specific fields of observation of each administration, necessary condition in order to exploit partial administrative data and to design samples for surveys. In particular, the RAE can help to improve the ASIA, prompting transmitting to it the administrative data, already integrated and at a higher frequency than now. This way, the Statistical Institute can concentrate on completing administrative integrated data with sample or population surveys and investigations on specific issues.

While the integrated administrative data have nature of publicity-news and are, therefore, accessible to all the interested persons at any level of detail, the data of the statistical registry, arising from statistical elaborations, are covered by the statistical secrecy discipline.

5. The Milano Municipality Project

This project, currently in its implementation phase through the collaboration between the Milano Municipality and Università di Milano Bicocca, applies the DW technology to meet administrative requirements, but simultaneously is able to supply interesting statistical data, impossible to obtain in traditional ways.

In the spring of 1999 the Education Sector of Milano's Municipality asked the Statistics Sector for figures regarding the citizens' incomes. The object of the request was that of foreseeing differentiated tariff policies according to the different income brackets and, subsequently, formulating simulations relative to the different studied hypotheses to be submitted to political decision. The almost total absence of information regarding individual and family realistic income then became evident. The ISTAT survey on "Family income and consumption", both for its methodological assumptions and for the territorial detail, cannot supply the above mentioned information. The information derived from the National Accounting, besides featuring problems regarding the requested territorial detail, suggests income as being "produced" by enterprises, institutions, etc., on a given territory rather than "earned" by the individual; a possible subdivision of such figures among the subjects still wouldn't represent a realistic earnings figure; rather it would stand as a per capita, globally produced, wealth indicator.

Income figures can be directly obtained from the income-tax return instead. This is the source that fully highlights individually earned wealth following different economical or financial activities or coming from property. There are several strong points in using such typically administrative data with a statistical aim, for example:

- They have already been "standardized" and, in the case of income-tax returns, there is a certain degree of tutelage around the truthfulness of what has been declared due to administrative and penal sanctions;

- Their usage following a statistical objective doesn't have any direct "statistical" burden.

However, in order to obtain information regarding resident families in the municipality of Milano consulting the resident population's register is also necessary; this, too, is an administrative archive. It is precisely ruled and has a very important function in managing citizens' domicile and status aspects. The comparison of these two administrative archives brings in considerable information relative to resident citizens according to the family unit composition.

The normalization and standardization phases have been already performed without particular problems; the data checking phase for the data that are present in both the sources, but that result different from each other has been solved through the "certifier bureau principle": the personal data are certified by the resident population's register, while for the fiscal data (e.g. errors in fiscal codes) is assumed as correct what is present in the fiscal archive. The idea that supports this procedure is the following: each bureau check with particular care the characters that constitute its filed of interest, while is inclined to ignore other data.

Many variables have been considered, both of registry and income kind, to make possible a detailed analysis of different patterns of income for the different sort of family. Furthermore, these data are of official kind, so if a mistake is found in the integration procedure (for example a wrong fiscal code, that is not consistent with the population register data of the person) it is impossible to statistical correct it, because only the Ministry of Finance can change it, with a well defined administrative procedure.

The main objective of the project is the building of an integrated system of normalized archives containing information on the "whole of Milan citizens".

The Ministry of Finance placed at disposal three databases with the following fiscal information:

- Modello Unico;
- Modello 730;
- Modello 770.

The architecture of the registry office seems complex. It is composed of the following: register of the population;

- individual forms, in alphabetical order, containing for each individual: the personal details, the working position and educational qualification, the address;
- family forms, under the name of the householder, with indications for the family positions of the people in the family unit;
- cohabitation forms, relative to people cohabiting for religious, health, care, military reasons and so on, with regular domicile in the same municipality.

The studies on the resident population, its habits, income, behaviours (obtained in the full respect of the individual's privacy and of statistical secrecy) can be extended to the following fields:

- family unit data are actually available, such as the overall declared income and the different income categories (capital, self-employment, subordinate employment, pension, buildings, farmland, etc.) so it is possible to analyse the individual and family unit income;
- income data, family composition and professional condition of family components are available with a territorial detail which would be otherwise unimaginable; in fact, the information obtained by cross comparing the

databases within a DW system can be aggregated according to the address, the census sector, the neighbourhoods, the area of administrative decentralisation;

- another important aspect comes from the historical authenticity of the information: the yearly updating will allow to study behaviours, incomes, improvements within a context of equal opportunities, etc.;
- lastly, the opportunity to achieve an "improvement" in the sources must not be under evaluated: integration brings with itself the possibility of verifying positioning or mismatching information therein. It seems, at a glance, that it may be the population register to obtain the best out of the integration with the Ministry of Finance archive. Normalization and integration actually highlight errors in the databases (duplication of fiscal codes, errors in the codes which have been recalculated by the Municipality, date of birth, domicile, etc.).

At least, it is important to note a further statistical capacity of this kind of DW: the complete knowledge of the municipality population, split for family sort, income dimension and kind, professional condition, age, sex and so on, allows to easily perform very careful and well representative samples to improve traditional surveys (Labour Force and Family income and consumption surveys) or to implement new ones.

6. Conclusions

The new techniques proposed by local and central administrations in order to simplify communication with citizens and enterprises have practical effects in the rationalization and standardization of data collection procedures, which attain a territorial detail level and are available in an undreamt short time compared to traditional surveying techniques.

A further added value for the data is given by their storage in DW, organizing it by subjects and, this way, allowing specialized inquiries on behalf of the users wherever they are and in whatever moment in time, without any additional elaborations in the database. Moreover, the DW features historical authenticity, i.e. it is able to maintain the old data once these have been substituted or integrated so as not to lose the historical memory of phenomena of public interest.

Consequently it is necessary that the statistical methodology evolves, suiting the always moving reality, learning and using the new implementations available from the technique, and being able to choose the right integration level between traditional methods and modern ones, aiming at producing statistical information nearer and nearer to the various subjects' needs: enterprises, public and private bodies, communities, etc.

Besides the traditional survey methods it is now consolidated the use of integrated archives of administrative data; the availability of new technologies and the new institutional framework give the statisticians a multiple opportunity of research to exploit these new possibilities as tools for data production, quality assessment (in a total quality prospective) of data and of public administration internal process and services, integration to sample surveys and census and, last but not least implementations of new statistical analysis methodologies.

The need of fitting the reality and choosing the more suitable tool in every situation, determines the chance of studying and refining new Data Mining procedures, not only as implementation of statistical processes coming from the transformation of known methods but particularly as formalization of original methodology, thought out to be

applied to the massive quantities of data, made available by the new information systems; these procedure must be able to supply to the operator the controlling parameters necessary in order to take the decisions requested by the procedure and in order to answer to the questions that the demand of statistical information implies.

References

- Abbate C., Pezone A. (1999) Il metodo di controllo della qualità dell'archivio delle imprese ASIA per il censimento del 2001, *Atti del Convegno della Società Italiana di Statistica Verso i censimenti del 2000*, Udine, 7-9 giugno 1999.
- Biffignandi S. , Martini M. (1995) (A cura di) *Il registro statistico europeo delle imprese. Esperienze e metodi per la sua costruzione in Italia*, F. Angeli, Milano.
- Buzzigoli L., Martelli C. (1999), Le regioni come nodo di linkage di archivi amministrativi: limiti e potenzialità di un nuovo ruolo per la gestione di fonti statistiche, *Atti del Convegno della Società Italiana di Statistica Verso i censimenti del 2000*, Udine, 1999.
- Giudici P. (2001), *Data Mining: metodi statistici per le applicazioni aziendali*, McGraw-Hill, Milano, 371-386.
- Immon W. H.(1996) *Building the Data Warehouse*, Wiley, New York.
- Martini M., Aimetti P. (1989) *ASPO, Un archivio delle imprese per l'analisi economica*, Unione Regionale delle Camere di Commercio della Regione Lombardia, Milano.
- Martini M. (1993) Statistical aspects of the Business Register Integration, *International Conference on Establishment Survey*, Buffalo 27-30 giugno 1993.
- Martini M. (1995) *Documento conclusivo del Gruppo di lavoro avente incarico di progettare e realizzare il nuovo registro statistico delle imprese (ASIA)*, Centro Stampa ISTAT, Istituto Nazionale di Statistica.
- Martini M.(1999), Censimenti, dati amministrativi e sistema statistico delle imprese basato sui registri, *Atti del Convegno della Società Italiana di Statistica Verso i censimenti del 2000*, Udine, 1999.
- Martini. M. (2000), L'integrazione statistica di registri e indagini, in Filippucci C. (a cura di) *Tecnologie informatiche e fonti amministrative nella produzione dei dati*, Franco Angeli, Bologna, pp291-307.
- Schionato L. (1995) Tecniche di linkage statistico per il raccordo di una pluralità di fonti amministrative, in: Biffignandi S.,Martini M. (a cura di) "*Il registro statistico europeo delle imprese* "Franco Angeli editore, Milano, pp.333-343.
- Verrecchia F. (2001), Data Mining: tecniche di segmentazione: *Atti del convegno SMDM2001*, Pavia, 35-40.
- Zavanella B. (1999) Nuove possibilità di stima del lavoro sommerso attraverso l'uso congiunto di censimenti e registri, *Atti del Convegno della Società Italiana di Statistica Verso i censimenti del 2000*, Udine, 7-9 giugno 1999.
- Zavanella B. (2000) Uso integrato di fonti diverse per la stima del lavoro sommerso", in Filippucci C. (a cura di) *Tecnologie informatiche e fonti amministrative nella produzione dei dati*, Franco Angeli, Bologna, pp. 399-411.