INTEGRATING GEO-REFERENCED DATA FROM DIFFERENT SOURCES: LIVESTOCK SURVEYS AND ADMINISTRATIVE DATA

Colomba Sermoneta[1], Roberto Gismondi[2], Daniela Fusco[3] Valerio Moretti[4]

[1] *Istat, Rome, Italy; sermonet@istat.it*
[2] *Istat, Rome, Italy; gismondi@istat.it*
[3] *Istat, Rome, Italy; dafusco@istat.it*
[4] *Istat, Rome, Italy; vmoretti@istat.it*

Abstract

The statistical data collection is often based on combined information from different sources, both to increase the knowledge and to ensure higher quality of statistical data. In particular, the geo-referenced database provides the possibility of their integration through the use of geographic information systems. In small area estimation models which take into account geo-referenced data, the estimates are more reliable respect to the use of traditional methods. The creation of a national farm register, to use as a reference list as support to surveys, implies the need to have a unique code to identify the farm. For example, in the estimate of the livestock, the administrative data from registry can be associated with survey data to obtain more detailed information about the kind of farms involved in the study, a particular animal disease or about the environment (eg. localization of

polluting emissions, information on the use of the territory, the presence of harmful substances related to the disposal of organic waste, etc.). The availability of the geographical coordinates of the farms provides an accurate and objective linkage between the database collected by administrative sources. In this way it's possible to go beyond some limits of matching procedures done using the identification number or the VAT number like matching variables. In fact, using multiply sources we could have different frames, so that a farm can be identified as technical economic unit or a specific geographical location, with significant differences of estimates. This paper presents the preliminary results of a geo-referred area analysis using map charts. They will illustrate how the territorial reference improves the integration of data from different sources and provides auxiliary information. Preliminary tasks concern the implementation of suitable techniques of record linkage, extensively used for improving the use of administrative data for statistical purposes in official statistics.

Keywords: geo-referenced, geo-spatial, integration, administrative data.

1. Geographic Information Systems as tool of data integration

Since the late '80s there has been a growing increase of interest in the data related to the territory and on the integration between surveys and administrative data. The development of information tools, such as Geographic Information Systems (Geographical Information Systems - GIS), can handle this type of information (Boffi, 2004). The data integration functions provided by GIS, which allow the connection of information from different subject areas, have led to a much wider use of statistical information. The need to develop spatial data has given a new point of view to the branch of statistics, called spatial statistics, which contains the methods that can analyze statistical observations taking into account the position in which they manifest themselves in a particular space. This procedure can have a variable level of detail that goes from the given macro-administrative areas, such as regions or

provinces, to come down to levels more detailed as provinces, municipalities or even more generically locations. One of the crucial points concerns the properties required of a statistical variable that must be defined with the maximum precision, so that each one describes exactly the same thing. Geo-referencing requires to acquire the accurate coordinates of the statistical variable object of study. Using GIS leads to combination among geo-referenced datasets on the same region in a single database via linkage. For example an agricultural holding can be identified as technical economic unit or as a geographical location, with significant differences of the estimates. Combined datasets can contain more information that the starting dataset holds, as in the case of public health studies, where data relating to a particular disease (eg. a register of congenital malformations of a certain region) can be associated, in the administrative data source (eg. livestock registry) to obtain demographic information about individuals involved in the study (consumers of milk or meat) and environmentally relevant data (eg. punctual measurements of pollutants, information on the use of the territory, the presence of harmful substances related to the decay of the substances used in water treatment, etc.). Let’s imagine what has been happening as regards the so called "Land of Fires" in the Italian Campania Region.

Finally, high spatial detailed information facilitates the realignment of information about statistical units detected from several sources, which typically do not refer to the same survey sites. The geo-reference approach also supports the integration of data from different sources by providing auxiliary information appropriate to the implementation of the widely used record linkage techniques, especially in the economic field, for the use of administrative records for statistical purposes (Liseo, Montanari, Torelli, 2006). Many territorial statistics are for the purpose of studying phenomena in terms of geographical area, highlighting and

possibly explaining the spatial variability that characterizes them. The spatial interaction is formalized through concepts such as distance (not necessarily meant as a metric in strict sense) or as adjacent sites on which events occur. The distance is a useful tool to explain the spatial variability of a phenomenon when it is sensitive to the deviation of the occurrences of sites with particular characteristics study. In this case, geo-referenced data with high level of detail facilitate the analyst who can prove their results by adopting different spatial references.

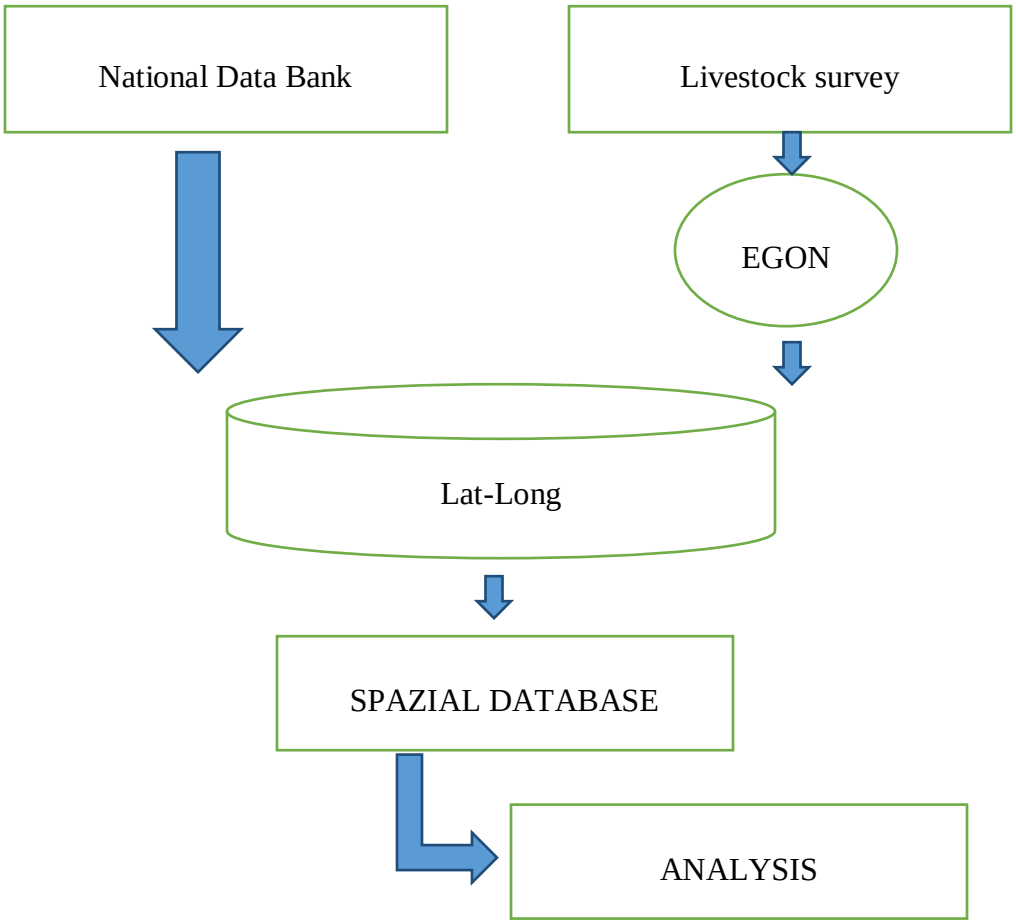
2. Goals and analysis

The proposal will illustrate how the territorial reference may improve the integration of data from different sources eg. the administrative data for statistical purposes and may provide auxiliary information on techniques of record linkage. The data collection in relation to the territory allows, the construction of appropriately themed cards that represent the territorial dynamics of the events of interest. The geo-referencing procedures are, in general, more complex and expensive than geocoding a variable by assigning a specific code (eg. area code). Geo-referencing in associations with addresses (address matching files) requires a phase of preparation of electronic records before the record linkage phase. The goal is to use as linking key geographic coordinates instead of VAT number, which is normally used in record linkage practices. The administrative source used comes from the Ministry of Health (National Data Bank) and the Italian statistics on livestock (Reg. CE 1165/2008). In the analysis we have compared, as example of small area estimation, farms in the provinces of Campania who own or have even temporarily buffalo heads.

In order to validate the addresses and to obtain geographical coordinates consistent and comparable each other, a specific software which corrects errors, validates the addresses and

standardizes them into the official format has been applied (Kuzma, 2013). The software used for this validation step is available at http://www.egon.com/en/solutions/address-validation.html

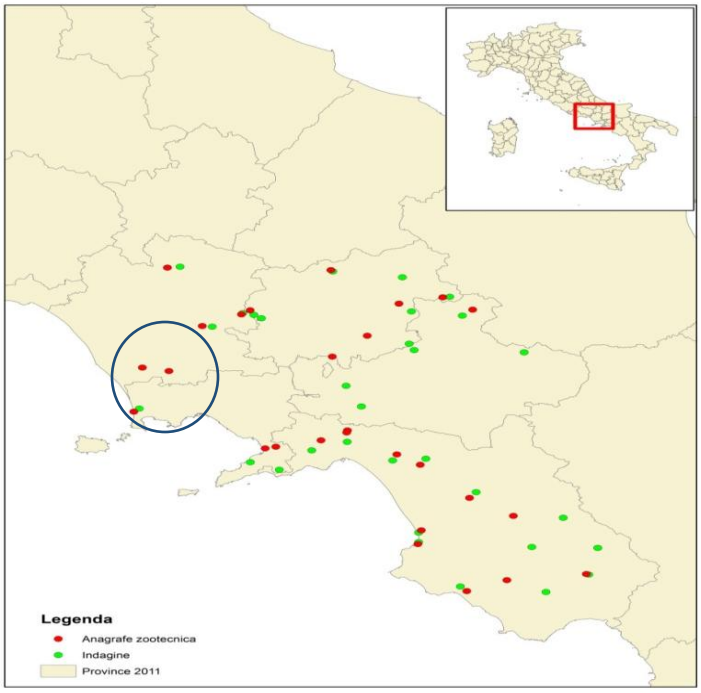
Figure 1: Linkage system design



3. Maps

The two statistical sources have been compared through mapping. The case study concerns the Italian Campania Region, which has been affected by a dangerous land contamination process due to not legal use of large parts of the territory as garbage dump. The regional territory concerned has been called "Terra dei Fuochi" ("Fire Land", since garbage is fired and drawn into large holes).

Figure 2: Agricultural holdings localization - Livestock survey (green) and BDN (red)



In the figure 2 Livestock data are the green points, while BDN data are the red ones. Red points have been obtained on the basis of green coordinates. Any red point may have been coupled with more than one green point (for this reason there are more red points than green ones). The best situation consists in having only red points, which overlap exactly the correspondent green ones. In practice, green points on the map represent geo-localization of agricultural holdings as derived from the Livestock survey, while red points show position of holdings as derived from the BDN which are nearest to a given survey holding.
Even though only a few number of holdings with livestock falls into the circle which identifies the "Terra dei Fuochi" (compare figure 2), other holdings not included in the livestock survey may fall inside that; in any case, errors occurred in the geo-referencing phase may lead to misleading conclusions as regards the dangerousness of this phenomenon.

4. Future perspective and conclusion

First results are promising and will imply further work regarding different issues. First of all, it would be useful to map additional reference points, as environmental areas, health, wholesale markets, manufacturing industries, in order to have a broader view of the role played by territory on the agri-food chain.

More in details, since the "Terra dei fuochi" is a part of the Italian territory particularly at risk as regards human health, it will be useful to evaluate the potential increase of health diseases due to pathologies connected to cow milk consumption. This evaluation may take into account the physical distance between each holding having livestock and specific dangerous location, which may have been contaminated by toxic garbage and other kinds of land pollution.

The analysis of geo-referenced data brings with it several problems. The survey carried out shows the need to put attention to the geo-referenced data in the collection phase. The coordination and integration of data from different sources is needed, as well as sharing of data gathered at different levels, making them more easily available, accessible, analyzed and interpreted by the researchers through issues that might not otherwise be assessed. This task is particularly relevant in the frame of "official statistics", where the territory should be seen as a place of interaction of a number of activities.

5. References

Australian Bureau of Statistics (2012), Report of the Australian Bureau of Statistics on developing a statistical-geospatial framework, http://www1.unece.org/stat/platform/download/attachments/81297858/2013-2-ProgReview-E.pdf.

Boffi M. (2004), Scienza dell'informazione geografica, introduzione ai GIS, Zanichelli. Bologna.

CZSO (2012), ESS Coordination of Geospatial Information and Statistics (2012), in http://www.czso.cz/dgins2012/dgins.nsf/i/session_iii_eurostat/$File/Contribution_Eurostat%20rev.pdf.

UN (2013), Committee of Experts on Global Geospatial Information Management, Linking of geospatial information to statistics and other data, in http://ggim.un.org/docs/meetings/3rd%20UNCE/E-C20-2013 9%20Linking%20Geospatial%20Information%20to%20Statistics%20Report.pdf

Kuzma I. (2013), An interactive web mapping application for presenting and dissemination of small area statistics, in NTTS New Techniques and Technologies for Statistics.

Liseo B., Montanari G.E., Torelli N., (*eds*) 2006, Metodi statistici per l'integrazione di dati da fonti diverse, Franco Angeli, Milano.