Use of Big Data in official statistics

Giulio Barcaroli
Italian National Institute of Statistics (Istat), Italy – barcarol@istat.it

Abstract

More abundant and higher-quality data are required to meet increasingly complex knowledge demands from users and stakeholders in modern societies. At the same time, budget pressures and costs associated with response burden limit the scope for new surveys. Harnessing new data sources in official statistics is therefore decisive to provide an adequate supply of information and knowledge while enhancing the effectiveness of statistical production processes. This paper presents the approach followed, and the results produced, in some experiments carried out by Istat in order to investigate the possibility to produce official statistical information by making use also of the new data sources represented by Big Data. Pilots have been planned by making reference to different possible scenarios: (i) leaving unaltered the classical sampling survey framework, data collection can be based on the availability of Big Data (use of *web scraping* and *text mining* techniques in "ICT in enterprises" survey); (ii) Big Data can be harnessed jointly with survey data, and considered as auxiliary information that can increase the accuracy and timeliness of estimates (*Google Trends* indicators for producing nowcasting and small area estimates); (iii) Big Data are used as the only source in alternative to the one represented by traditional statistical surveys (use of *mobile phone data* to estimate origin/destination matrix of daily mobility for work and study at the spatial granularity of municipalities). Though further research and development are needed, first results are encouraging, and new investigations are being launched: use of data from social networks (Twitter and Facebook) to estimate Consumer Confidence Index, estimation of road traffic flows by using sensors (webcams), use of scanner data for the calculation of Consumer Price Index.

Keywords: web scraping; text mining; Internet queries; mobile phone data; multiple data sources.

1. Introduction

Big Data represent a new source to be harnessed by official statistics organisations with the aim to produce additional information, or increase the quality of already available, while reducing related costs (Eurostat 2013, United Nations 2014).
As in the case of administrative data, also Big Data can be exploited in different ways in the context of the official statistics production process.
Considering a given target population (households and/or individuals, enterprises and/or institutions, etc.), the scheme in Fig.1 illustrates a general framework for the production of related statistical information, where different sources of data are originated by (or can be related to) the target population:

- *statistical data*, collected by means of traditional surveys: all identification and structural information related to the target population is organised in a frame, the basis for censuses or sample surveys; data are collected by directly contacting selected units in the population and processed in order to produce estimates that are disseminated to users;
- *administrative data*, resulting from administrative procedures (tax, social security, health, education, identity cards, internal accounting data, etc.): these data can be linked to other statistical sources either with certainty or with probabilistic record linkage procedures, and may coincide with variables of interest or can be used as auxiliary information;
- *Big Data*, originated by the use of digital devices (from transactions, sensors, tracking devices as GSM or GPS, online searches, comments on social media, etc.).

While the final target of producing reliable statistical information remains the same, different scenarios can be outlined on the basis of the degree of use of Big Data in the production process:

1. use of digital devices (in particular of those connected to the Internet) as media for data collection: the overall statistical framework remains unchanged (this scenario can be referred to as *Internet as Data source, IaD*);
2. use of Big Data in *conjunction* with statistical data, as additional information that can be used in order to improve the quality of statistical data (for instance in the edit and imputation phase), or to enhance the reliability of estimates using Big Data in the same way administrative data or census data can be used together with sampling data (for instance in composite estimators in Small Area Estimation, or in Statistical Matching);
3. use of Big Data as an *alternative to* the use of statistical data: when certain conditions occur, in order to reduce costs and response burden it is possible to completely substitute the classical process of production based on statistical survey and to adopt radically different processes based on the integral use of Big Data.

Istat planned to carry out pilot experiments with respect to each one of these different scenarios. In the following paragraphs they are illustrated and related results are reported.

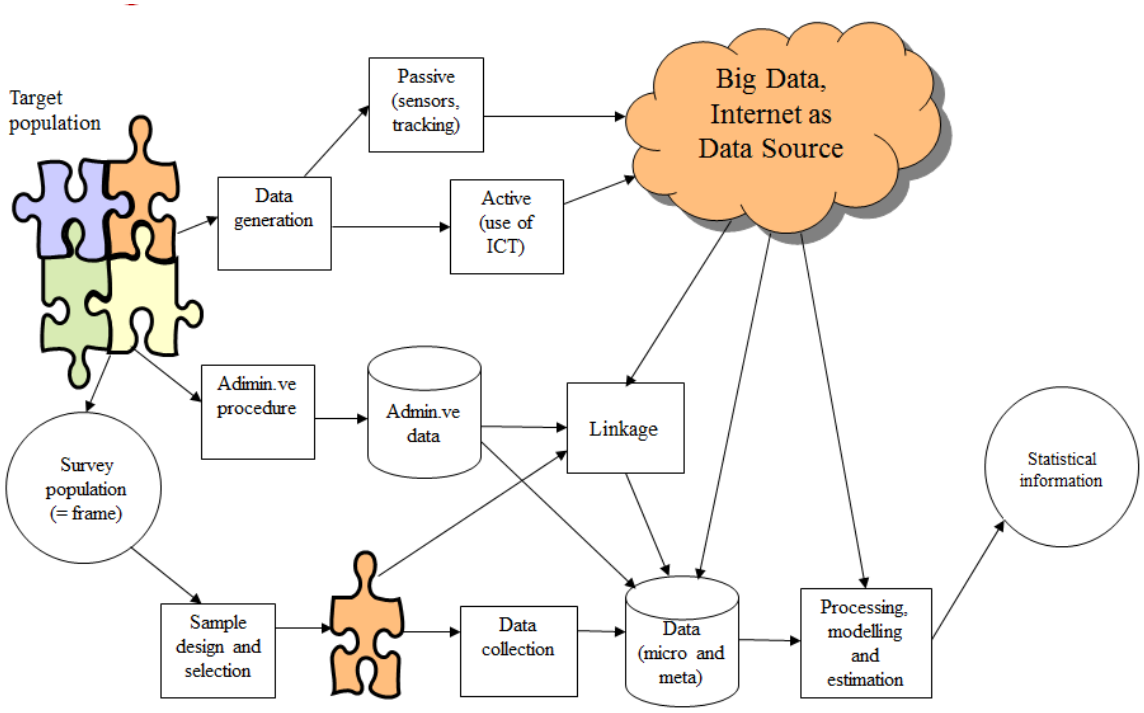

Figure 1 – A general framework for the production of statistical information in a multi-source environment

2. Internet as Data source: the "ICT in enterprises" survey

The Istat sampling survey on Information and Communication Technologies (ICT) in enterprises aims at producing information on the use of Internet and other networks by Italian enterprises for various purposes (e-commerce, e-skills, e-business, social media, e-government, etc.). For this purpose, data are collected by means of the traditional instrument of the questionnaire. Istat began to explore the possibility to use web scraping techniques, associated, in the estimation phase, to text and data mining algorithms, with the aim to replace traditional instruments of data collection and estimation, or to combine them in an integrated approach. The 8,687 websites, indicated by the 19,114 enterprises responding to the survey of year 2013, have been scraped and the acquired texts have been processed

in order to try to reproduce the same information collected via questionnaire. Preliminary results are encouraging, showing in some cases a satisfactory predictive capability of fitted models (mainly those obtained by using the Naive-Bayes algorithm, see Table 1). Also the method known as Content Analysis has been applied, and its results compared to those obtained with classical learners. (for a complete illustration see Barcaroli *et al.* 2015).

Table 1 – Results of the application of Naive Bayes to the complete set of website functionalities

QUESTION	Precision	Sensitivity	Specificity	Proportion Web sales = Yes (observed)	Proportion Web sales = Yes (predicted)
Web sales functionality	0.78	0.50	0.86	0.21	0.21
Orders tracking	0.82	0.49	0.85	0.18	0.11
Description and price list of goods	0.62	0.44	0.79	0.48	0.32
Personalised content for regular visitors	0.74	0.41	0.781	0.09	0.23
Possibility to customise online goods	0.86	0.53	0.87	0.05	0.14
Privacy policy statement	0.59	0.57	0.64	0.68	0.51
Online job application	0.69	0.521	0.78	0.35	0.33

The quality of predictions cannot be considered as satisfactory for all questions. Procedures for web scraping and text mining are under further development to be improved.
Next step will be the implementation of an integrated system harnessing both survey data and data collected from the Internet, based on systematic scraping of the near 100,000 websites related to the whole population of Italian enterprises with 10 persons employed or more, operating in industry and services. The whole set of survey data will be considered as the "training" dataset to fit a prediction model that will be applied on the whole information obtained by scraped websites of all enterprises in the target population, in order to get estimates whose reliability will be evaluated in comparison with the currently produced sampling estimates.

3. Estimation of *monthly unemployment rate* by using Internet queries (Google Trends)

The aim of this experiment was to evaluate the possibility of using, as auxiliary information, the time series of the query share related to terms such as "jobs", "job offers" and similar, obtained by Google Trends (GT), in order to produce (i) early estimates of *monthly unemployment rate* (*nowcasting*) (Choi and Varian 2012), and (ii) small area estimates for the same indicator (D'Alò *et al.* 2012).
In a preliminary step, the time series related to the monthly unemployment rate calculated from the Labour Forces (LF) survey has been compared with the one calculated with GT data. In Fig.2 the two series are jointly visualised.
Then, considering the LF data, models with different lag structures have been fitted, with the aim of identifying the best ones, which are found to be seasonal ARIMA models. Given these benchmarks, corresponding models using also GT data have been considered. After a model evaluation, carried out by means of *rolling regression* procedure on monthly LF (using data for the years 2013-2014, see Fig.4), the results obtained have been evaluated computing the Mean Absolute Relative Error (MARE) of both *one step* ahead and *three step ahead* estimates with respect to the true values. The prediction of

monthly unemployed rates computed on the basis of models with also GT information outperform those obtained on the basis of the benchmarking models. Moreover, the *three step ahead* estimates show greater improvements compared to the *one step ahead* estimates (Fasulo *et al.* 2015).

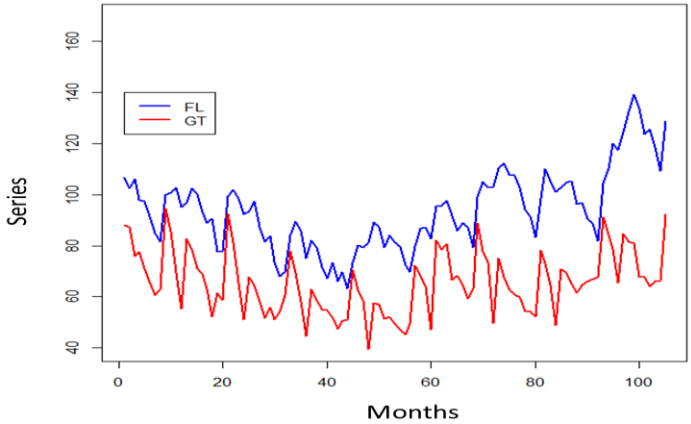

Figure 2 – Monthly unemployment rate from FL (Istat Labour Force Survey) and GT (Google Trends: results on the search term «job offers») - Time series 2004-2012

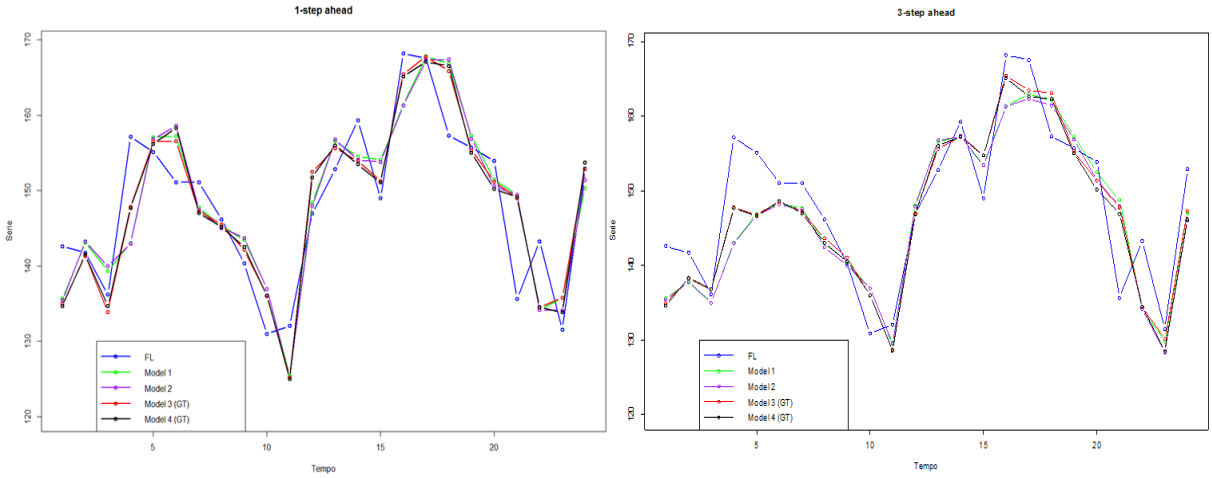

Figure 3 - Alternative model analysis and comparisons with benchmark models: Rolling Regression, 2013-2014 (one step ahead and three steps ahead)

4. Mobile phone (tracking) data to estimate mobility flows (Persons and Places project)

The Istat project "Persons and Places", based on the integration of administrative and census data, has the aim of producing an Origin-Destination matrix at municipality level from which it is possible, for each municipality A, to estimate the following sub-populations:

- *static residents* in A: persons who have formal residence and place of work (study) in the same municipality A, or who do not work (study);
- *dynamic residents* in A: people that spend long periods for working or studying in a municipality A (most days of the week), while being formally resident in another municipality B;
- *commuters* in A: people who commute for working or studying to municipality A, having formal residence in another municipality B;
- *visitors*: people occasionally visiting the municipality A.

Considering administrative data, for any given individual the coincidence between the city of residence and that of work (or study) is considered as a proxy of the absence of intercity mobility for a

person (therefore we define him a static resident). The opposite case is considered as a proxy of presence of mobility (the person is a dynamic resident or a commuter or an occasional visitor). Administrative data do not allow either to distinguish between dynamic residents and commuters or to estimate occasional visitors, as they do not contain information on the frequency of the mobility. The idea is to consider another source of data, namely the mobile phone data (GSM) that allow to estimate all the sub-populations of interest (Pedreschi *et al*. 2014).
The unit of information is the Call Detail Record (CDR), each one containing data on a phone call made by an individual in a given municipality at a certain time in a day. Each individual has been tracked for one month (7.8 million CDRs collected from Jan 9th to Feb 8th, 2012 in Pisa province). From CDRs have been defined individual patterns, that have been associated to the different sub-populations of interest (see Fig.4).

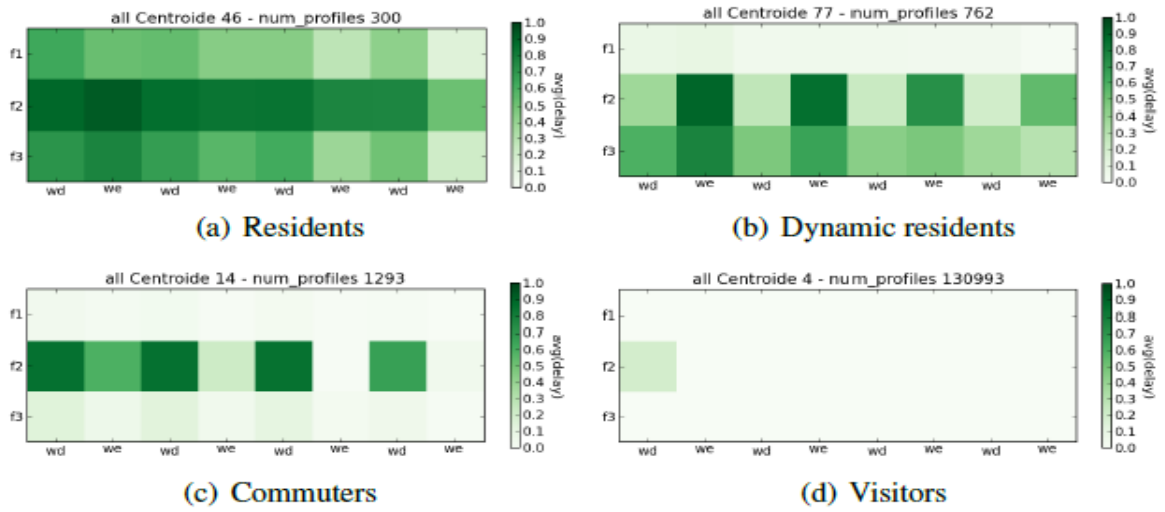

Figure 4 – Patterns defined by Call Detail Records

It is possible to compare estimates obtained by making use of CDRs with those obtained by administrative data and population census. For example, in Fig.5 the joint distributions for static residents and commuters are visualised together with fitted models.

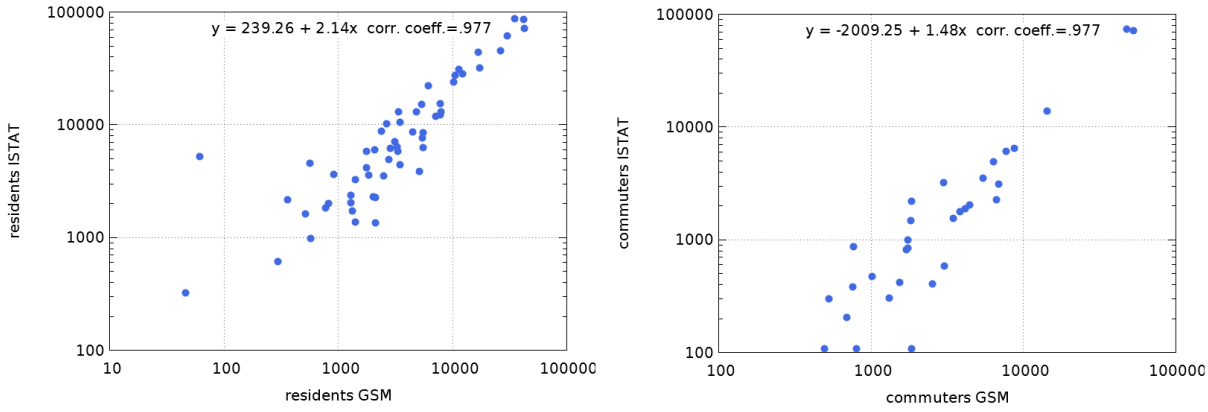

Figure 5 – Correlation between estimates by Istat administrative or census data and by GSM data

It can be seen that the alignment between the two sets of estimates is on average good, in particular very good for big municipalities and sometimes poor for small ones. In any case, these results can be

considered as acceptable, and in the next step it is planned to consider (i) the whole nation instead of one province and (ii) the plurality of mobile phone providers instead of only one.

5. Conclusions

In this paper, some experiments carried out by Istat in last two years are presented and related results illustrated. In general, these results are encouraging and promising, but further improvements are required before introducing related data sources and applications into production processes. The next 18 months will be dedicated to these refinements, also collaborating at international level in European and EU projects.
In the meanwhile, other areas of work will be considered, including (i) the application of text mining and sentiment analysis methods to social networks data (e.g. Twitter and Facebook) in order to estimate the Consumer Confidence Index; (ii) the estimation of road traffic flows using sensor data; (iii) the use of scanner data for the calculation of the Consumer Price Index; (iv) the use of mobile phone data to produce tourism statistics; (v) the production of smart city indicators (De Santis *et al.* 2014). Common issues on benefits and challenges of Big Data relate to quality concerns, accuracy and timeliness, data linkage, profiling methods and visualisation tools. Privacy issues are also a concern (legal frameworks, ethical guidelines and technological solutions to safeguard privacy), as well as IT and data science issues (cloud computing, data mining and machine learning).

References

Barcaroli, G., Nurra, A., Salamone, S., Scannapieco, M., Scarnò, M., Summa, D. (2015). Internet as Data Source in the Istat Survey on ICT in Enterprises. Austrian Journal of Statistics, Vol. 44, N. 2/2015

Choi, H., Varian, H. (2012). Predicting the Present with Google Trends. Economic Record, 88: 2–9

D'Alò, M., Di Consiglio, L., Falorsi, S., Ranalli, M.G, Solari, F. (2012). Use of spatial information in small area models for unemployment rate estimation at subprovincial areas in Italy, Journal of the Indian Society of Agricultural Statistic, Volume 66, December 2012, pp. 1-239

De Santis R., Fasano A., Mignolli N., Villa A (2014). Smart city: fact and fiction. Working Documents n. 100, LUISS-LAB of European Economics - LLEE.

Eurostat (2013). Scheveningen Memorandum on Big Data and Official Statistics.
http://ec.europa.eu/eurostat/documents/42577/43315/Scheveningen-memorandum-27-09-13

Fasulo, A., D'Alò, M., Falorsi S. (2015). Provisional estimates of monthly unemployment rate using Google Trend. Proceedings of ITACOSM. Rome, June 2015.

Pedreschi, D., Vivio, R., Giannotti, F., Nanni, M., Furletti, B., Garofalo, G., Gabrielli, L., Milli, L. (2014). Use of mobile phone data to estimate mobility flows. Measuring urban population and inter-city mobility using big data in an integrated approach. Proceedings of the 47th Scientific Meeting of the Italian Statistical Society. Cagliari, June 2014.

United Nations Statistical Commission (2014). Big Data and Modernisation of Statistical Systems, report prepared for the Forty-Fifth Session of the Statistical Commission, New York, 4–7 March 2014.
http://unstats.un.org/unsd/statcom/doc14/2014-11-BigData-E.pdf