Big data–based national statistical production

By Alessandra Righi and Monica Scannapieco

DISCUSSION by Giuseppe Arbia

Workshop on “Big Data & Machine Learning Applications for Central Banks”
Bank of Italy, Rome, 21-22 October 2019

ISTAT Big Data activities

- Pilot
- Experimental
- Official

General comment: When data are not drawn according to a probabilistic sample design it is not possible to assign the probability of inclusion of each unit and as a consequence and draw any probabilistic inference.

Estimates are maybe still unbiased, but most probably not efficient.

In some cases there is the need to process the data before treating them if we wish to make inductive inference (e. g. post-sampling).

Piloting surveys

1. UCAS survey every 3 years since 2006 in order to estimate the Land Cover (LC) and Land Use (LU) within the EU up to NUTS-2

   1st phase: Master Sample of ~1.1 million points in a square grid of (2 km x 2 km) cells

   2nd phase: ~330,000 random points from the Master Sample

   Direct data collection, mainly on the ground (~70% of 2nd phase points), the rest by clerical photo-interpretation

   → **(details on sample design ?)**

2. Computer Vision methods (e.g. Deep Learning) **+ Satellite Imagery** data (e.g. Sentinel-2) are used for LC estimation:

Can a fully automated approach provide LC estimates of satisfactory accuracy?

→ **(any test to show ? E. g. simulation ?_**

Experimental surveys

1. The annual Survey on ICT Usage in Enterprises ('ICT survey') collects data on the usage of Information and Communication Technologies, the Internet, e-business and e-commerce in enterprises.

Its is a webscraping exercise with a sample design (→ **details**) sampled proportion 11 %, but with a non response ratio of 35 % (→ **any guess why**?) however only 70 % of the population owns a website.

Unbiased estimates, (→ **what about efficiency ?)**

2. *Istat domain-specific* sentiment index to assess the mood about the economic situation of the Italian-speaking Twitter users

(→ **Sample selection bias-correction ?)**

Official statistics

Scanner data to replace price collection for price indices of 79 grocery products

Scanner data for 2,146 outlets, including 534 hypermarkets and 1,612 supermarkets of the main 16 RTCs covering the entire national territory are monthly collected by Istat on a weekly basis at item code level.

(covering 40% of turnover but selecting no more than the first 30 GTINs in terms of turnover). (**clarify**)