United Nations ECE/CES/2018/37

Economic and Social Council

Distr.: General
4 June 2018

English only

Economic Commission for Europe

Conference of European Statisticians
Sixty-sixth plenary session
Geneva, 18–20 June 2018
Item 3 of the provisional agenda
Measuring what matters - broadening official statistics:
Session 2: How to react swiftly

Experimental statistics new challenges for NSOs: Istat experience

Note by the Italian National Institute of Statistics

Summary

The note highlights the strategic vision of Istat that is aimed at ensuring its relevance by innovating processes and products to respond to the new needs with increased effectiveness. Experimental statistics are at the core of this innovation agenda, fully embedded in the research activities. Two examples are presented showing how to ensure a good balance between research *independence* on the one hand, and *relevance and pertinence* on the other hand.

The document is presented to the Conference of European Statisticians' seminar on "Measuring what matters – broadening official statistics", session 2 "How to react swiftly" for discussion.

GE.18-08850(E)
1808850

I. Introduction

1. The evolution of the demand for statistical information, also resulting from new data sources, technologies and methods, confronted the Italian National Institute of Statistics (Istat) with the urgent need to innovating processes and products, thus increasing its ability to respond to the new needs with growing effectiveness.

II. Istat as part of public research system

2. Since November 2016, Istat is officially part of the Italian public research system. The regulatory law of the Institute (passed at the end of 2017), defines research as "aimed at improving the quality of statistical information and processes for production, development and dissemination of official statistics, as well as at embedding the results of methodological and thematic research in the above processes".

3. Evidence of need and willingness to invest in research is also given by the Institute's strategic planning, wherein a specific development of methodological and thematic research is outlined. Methodological and thematic research are acknowledged as a factor of growth for the Institute and its human capital, to pursue structurally and organically.

4. Research contributes to improving statistical information, to experimenting and developing new techniques and methodologies, to implementing thematic analyses, and to the promoting initiatives aimed at integrating new sources, big data and open data in the production and dissemination of official statistics.

5. To outline the role of research in Istat, reference to Fellegi (2010) on research in the National Statistical Institutes may be useful. FeIlegi discusses in detail the antinomy between research *independence* (freedom in choosing fields to investigate), on the one hand, and *relevance and pertinence* (motivated by data and information production needs), on the other. Being able to identify the right balance between independence and relevance guarantees both a strong research capacity, and significant spillovers on production activities. Research is only relevant if motivated and guided by the activities carried out in the production processes.

6. On the other hand, independence gives researchers the opportunity to propose and manage research topics. Failing to secure a certain degree of autonomy to researchers, and engaging them entirely in immediate production support requests, will imply that those researchers will guarantee neither research nor development. If an Institute fails to give research activities the appropriate frames of reference and management, relevance and independence will inevitably clash. The point is therefore to find a balance to secure both. Results of research activities are a factor of structural and progressive improvement, able to feed a virtuous circle of continuous progress on both processes and final products.

7. Recently, Istat has adopted instruments and infrastructures to shape and structure research and innovation activities to conjugate independence and relevance, as highlighted by Fellegi.

8. The decision at Istat was to enhance two research lines, typical of a Statistical Institute: one on statistical methods and techniques and one on social and economic themes. To support the two areas, we have established two external expert committees: a Methodological Advisory Board will assist strategic research projects and relevant process innovations; a Scientific Committee will evaluate the thematic projects.

9. In the methodological field, the Institute is investing in projects that address research questions directly related to the modernization program: e.g., the integrated system of registers, census data obtained through data integration, use of Big Data for the production

of statistics and the unification of methods and tools for managing and running production processes of surveys on business.

10. Concerning thematic research, areas of interest have been identified by their relevance in the short and mid-term scientific debate, potential contribution to economic and social policies, impact on statistical production, also taking into account stakeholders' needs of at the Country's level and international projects.

11. In May 2017, a call for projects was launched, on the research areas thus identified, open to all Istat researchers. The response to the call was higher than expected, with 77 projects submitted.

12. The Laboratory is another instrument aimed at encouraging innovation and strengthening the role of research as a founding value and a tool for strategic growth of the Institute and its human resources. The Laboratory selects projects likely to produce outputs of interest for official statistics and experimental statistics.

13. The main impact of investing in research is on the relevance of the Institute, its capacity to produce new knowledge on emerging phenomena and to exploit information from new sources.

14. While information needs of users of statistical information grow wider and deeper, production of quality statistics needs time to test new methodologies, translate them into technological and organizational solutions and to assess their compliance with high standards.

III. "Experimental" statistics

15. In line with similar initiatives launched by Eurostat and other statistical institutes to meet the new demand, Istat recently decided to release the results of "unofficial" statistical analysis in form of "experimental" statistics.

16. Below are two examples of experimental statistics, stemming from investments in research. They offer new information, using the potential of big data.

A. First example. Experimental statistics on ICT use by enterprises from Internet data: e-commerce, job advertising and presence on the social media

17. Istat has experimented a new approach based on the combined use of survey, administrative and Big Data sources to obtain a subset of estimates currently being produced through the sample survey on ICT usage and e-commerce among enterprises. Target estimates of this survey include the characteristics of websites where enterprises display their business. To produce these estimates, data are collected by means of traditional questionnaires, following the usual design-based inference approach.

18. An alternative procedure has been implemented: data have been collected by accessing directly the websites, processing the texts thus collected to identify relevant terms, and modelling the relationships relating those terms and the characteristics we are interested to estimate. The sample of surveyed data has played the role of a training set, useful to fit models applicable to the generality of enterprises that own a website. Administrative data (recorded in the Business Register) have been used to solve the usual problem of representativeness, typical of Big Data sources. The sequential application of web scraping, text mining and machine learning techniques has produced estimates that have been compared to those obtained by the survey. Estimates (rate of enterprises offering

e-commerce and job advertisements on their websites, and rate of presence in the social media) have been produced by adopting a model-based estimator, together with a combined estimator (using both Internet and survey data). Estimates have been produced with reference to different domains: economic activities at different level of detail, dimension of the enterprises, territorial classification (regions), ICT and non ICT enterprises.

19. The new estimates (model-based and combined) are compatible with the current design-based survey ones: most of them fall inside the design-based confidence intervals.

20. In terms of quality (accuracy), the impact of the new estimators is both positive (reduction of the variability due to sampling variance and of the bias due to total non-response and to measurement errors in the survey) and negative (model bias and variance). Whenever the quality of estimates obtained by means of this new approach is not lower than the one of the traditional process, the former is to be preferred, as it allows, not only to produce aggregate estimates, but also to predict individual values, useful to enrich the information contained in registers. In addition, the survey does not produce estimates each year, while model based estimates can be obtained at any desired time interval.

21. In conclusion, this experience has proved the validity of the approach. As new information can be obtained by direct access to Internet data, new pilots are being carried out to verify the extent to which it will enable us to enrich the content of the Business Register.

B. Second example. Experimental statistics: the social mood on economy index

22. More and more people all over the world are using Social Media platforms to express their feelings and ideas, as well as to share or debate opinions about virtually every possible topic. Consequently, the interest towards the Social Media as a means for "measuring" public's mood is growing vigorously.

23. Istat has investigated whether social media messages might be successfully exploited to develop domain-specific sentiment indices, namely statistical instruments meant to assess the Italian mood about specific topics or aspects of life, like the economic situation, the European Union, the new migratory phenomena, etc.

24. To this end, Istat researchers have developed procedures to collect and process only those social media messages that contain at least one keyword belonging to a specific filter, namely a definite set of relevant Italian words. Domain-specific filters have been designed by subject-matter experts with the aim of filtering out messages that would very likely turn out to be off-topic for the intended statistical production goal.

25. The 'social mood on economy' index, that Istat is to publish soon as experimental statistic, provides a first example of domain-specific sentiment index derived from social media messages. Presently, the index only uses Twitter as a source, but other social media might be added in the future. This new statistical instrument has been devised to enable high-frequency (e.g. daily) measures of the Italian sentiment on the state of the economy.

26. Twitter's public interface is used to collect samples of public tweets matching a filter made up of 60 relevant keywords (individual words, as well as phrases). A subset of those keywords has been taken from questionnaire items of the Italian Consumer Confidence Survey, a sample survey that collects data monthly during the first fortnight and disseminates results by the end of the month. Note that the phenomenon tracked by the 'social mood on economy' index and consumer confidence only partially overlap. Still, the index can detect and promptly point out events that influence consumer confidence but

happen to occur after the interview period of the Consumer Confidence Survey: the Central Italy earthquake of 24th August 2016 is a striking example.

27. To compute daily index values, all the tweets collected in a single day (about 40'000, on average) are processed as a single block. Messages are first cleaned and normalized, then undergo a sentiment analysis procedure that clusters them into three mutually exclusive classes: Positive, Negative and Neutral. The daily index value is lastly derived as an appropriate central tendency measure of the score distribution of the tweets belonging to the Positive and Negative classes.

28. Special care is devoted to make the index robust against possible contaminations by off-topic tweets that might pass the filter. To this end, a surveillance system has been put in place, which continuously searches for anomalous values in the daily index time series by means of a dedicated outlier detection routine. Daily values detected as potential outliers are sent to human reviewers in charge of deciding whether they are actually proper data points, or instead truly anomalous. The latter case typically arises when an off-topic tweet that happened to pass the filter becomes "viral" on Twitter. Being re-twitted and quoted thousands of times in a day, viral tweets may have an unduly impact on the daily index and introduce bias. As a consequence, all the daily index values classified as truly anomalous are eventually imputed.

29. Further research is in process on possible uses of the 'social mood on economy' index. The hope is that this experimental index could either improve the performance of Istat's forecasting models, or enrich existing statistical products (e.g. the BES).

IV. Some conclusions

30. This document has highlighted the strategic vision of Istat, aimed at ensuring the relevance of the adoption of an agenda which deals with the urgent need to innovate processes and products, thus increasing the ability to respond to the new needs with growing effectiveness. In this conceptual setting, research plays a central role: experimental statistics are at the core of this innovation agenda, fully embedded in the research activities.

31. The examples presented on specific research infrastructures show how we are engaged in ensuring a good balance between research independence on the one hand, and relevance and pertinence (motivated by data and information production needs), on the other hand.