Fabrice Gras (Eurostat)

New data sources: new modelling approaches?

Topic 3 – More rapid statistics and indicators on new phenomena

Keywords: administrative data, big-data, official statistics, modelling

Introduction

Statistical authorities need to produce data faster and in a cost effective way, to become more responsive to users´ demands, while at the same time continuing to provide high quality output, in particular concerning the existing constraint for official statistics to be ideally stable in time and/or having an enough long history in order to guarantee some continuity of the output.

One way to fulfil this is to make use of all new accessible data sources, as for example administrative data and big data. Those data could then be used either as a substitution and/or as a complement to information already collected by official statistics (including surveys' statistics), or as the basis for producing new statistics, or to improve timeliness in order to better answer users' needs.

Most of the time, it will require to follow in an explicit or implicit way a model based approach at some step of the statistical process. Then, using new sources will introduce some new features in the way statistical models will be used or adapted for producing and assessing the quality of future official statistics. This paper will thus investigate if and how modelling approaches in statistical offices would need to be updated.

Methods / Problem statement

After showing in the introduction why using new sources in statistical leads to an increase use of model based statistics, the first part will be based on the assumption that modelling approaches when using new sources essentially depend on three factors:

- The nature of the source (administrative data, big-data).

- The foreseen usage (replacement of existing data, improve timeliness, improve the granularity of the existing statistics, estimation of missing data, complement data, etc …).

- The nature of existing statistics (survey, accounting data, mixed data, robustness, sensitivity).

Taking into account these three dimensions, this paper will thus first review in a comprehensive way the possible usages of new sources and examine how it could have an impact on thechoise of a modelling approach, because of the characteristics either of the type of usage or of the type of new sources. Then, a tentative typology classifying and presenting the main possible cases when using new sources will be proposed. Based upon this tentative typology, the second part will focus on the identification of the main challenges that should be addressed when using new sources in term of modelling. Then, recommendations will be provided on how modelling approaches should be adapted in order to better achieve those challenges.

Results / Proposed solution

Using new sources most likely leads to an increase production of model based statistics. Statistical methods and tools already exist in most of the possible cases, but one of the most challenging feature concerns the way to measure uncertainty related to non-survey data.

Nevertheless quality constraint of official statistics, in particular the continuity of the official statistics, also needs to be taken into account before adopting any kind of model approaches. Possible solution could lay in the capacity of building bridges between existing statistical methods currently used by statistical offices and new methods in order to assess the comparability and robustness of official statistics over time.

Conclusions

Digital revolution will lead to an increasing number of possible sources that could be used for producing official statistics. Modelling techniques enabling the combination of several sources of information as well as extracting and structuring available information in a usable and valuable way will be more and more used by statistical offices. In most of the cases, whatever the methods and the types of sources used, continuity and robustness of official statistics will also have to be guaranteed.

Therefore, in addition to the choice of efficient technical solutions depending on the objectives and on the characteristics of the data, selection of modelling approaches should also paid attention to the building of bridges between existing statistics and statistics of the future. One possible promising approach could consist in using surveys for estimating parameters that would validate model specifications.