

Stock-take of Statistical Developments – The use of administrative data

Statistics Canada

The demand being placed on NSOs for more granular, timely and integrated statistics is forcing NSOs to rethink their approach to collecting the basic data necessary to produce official statistics. It is quite likely that, in the future, most of the data that underpin our official statistics will be obtained from images and sensors, will be held by private firms or 'owned' by households and will represent a non-random portion of the population that we are attempting to measure.

It could be argued that this is not new. Statistical agencies have been using administrative data in the past, and in many ways have always had an 'admin data' first approach. In the case of Canada, the direction to use administrative data is embedded in the Statistic Act (the Act that governs Statistics Canada). The Act stipulates that the agency must work with other government departments in order to make maximum use of the administrative data collected by government agencies.

Today, many of Statistics Canada's programs are based, in whole or in part, on data available from administrative sources. Like most other statistical agencies, Statistics Canada uses administrative data in lieu of or to complement survey data, and to support statistical operations. The use of this data has several benefits including:

- improving data quality by integrating statistical information from several sources
- responding quickly to new and growing needs for quality statistical outputs
- reducing data collection costs
- reducing the response burden on individuals and companies
- obtaining data that may be difficult or costly to obtain by survey
- improving survey design
- providing a direct way to tabulate statistical outputs
- facilitating statistical modelling based on the integration of multiple data sources.

While most NSOs have a long history of using administrative and alternative data there are several significant changes occurring.

First, more and more of the administrative data that can be of use to statistical agencies is held ("owned") by private organizations. Scanner data, satellite imagery, financial transactions, travel information, housing data are often partly or entirely controlled by private firms. The authorities, powers and social license for statistical agencies to acquire and use these data is not always apparent.

Second, in the past administrative data files generally represented a census or near census of the population under observation. Taxation statistics, social benefit information, health information, justice information usually contain an exhaustive coverage of the target statistical universe. Even though, at times, these data had to be acquired from multiple players and partners, it was possible to reconcile the information such that it constituted a census or near census. Today, this is no longer the case. The data acquired from private firms are generally not exhaustive and embody a non-random sample of individuals, businesses or other institutional units. The 'statistical unit', the concepts and the classifications used by these firms to store and share their data are often not comparable with the NSO's statistical infrastructure nor are they comparable from one firm to the next.

The generalized statistical process business model (GSPBM) is a widely accepted standard that describes the statistical process carried out by most NSOs. At a high level, the GSPBM describes the statistical process as one of specify need, design, build, collect, process, analyze, disseminate and evaluate. The GSPBM is uniquely survey centric in its design. While it is true that, at a conceptual level, the idea of administrative data can fit neatly into this process, in practical terms, given the changing nature of administrative data, NSOs are facing challenges and need to rethink their business processes.

In many cases, the statistical process now starts with the collection of information. NGOs, government agencies and private firms are now showing up at the doorstep of NSOs with truckloads of data and asking NSOs to make sense of the information. Instead of a process where we first specify needs, describe the target population, design a questionnaire, draw a sample, collect and process we are now in a world where we receive data, where we need to determine the population to which it applies, identify the type of questions the data can address and disseminate the data in such a way that users have an accurate picture about what it represents and what it does not represent. This represents a fundamental shift in the NSOs business processes.

The future system of economic statistics therefore needs to consider a far different process, set of building blocks and methods than what is currently in use. It would be prudent moving forward that the system of economic statistics has an eye towards this new process. In some ways, at the extreme, you could argue that the system should, in part, be designed around the data that is available. While it may be irresponsible to fully swing the pendulum to the data driven design, at the same time it would be prudent to at least have an eye on the new sources of available data when we consider the future design of the system.