Proceedings of Statistics Canada Symposium 2014
Beyond traditional survey taking: adapting to a changing world

Making Use of Administrative, Big and Survey Data: An Assessment of the Quality of Canadian Wetland Databases

Herbert Nkwimi Tchahou, Claude Girard and Martin Hamel[1]

Abstract

While wetlands represent only 6.4% of the world's surface area, they are essential to the survival of terrestrial species. These ecosystems require special attention in Canada, since that is where nearly 25% of the world's wetlands are found. Environment Canada (EC) has massive databases that contain all kinds of wetland information from various sources. Before the information in these databases could be used for any environmental initiative, it had to be classified and its quality had to be assessed. In this paper, we will give an overview of the joint pilot project carried out by EC and Statistics Canada to assess the quality of the information contained in these databases, which has characteristics specific to big data, administrative data and survey data.

Keywords: Data analysis; quality assessment; environment.

1. Introduction

Traditionally in surveys, the data used to produce inferences came from a planned process of collection from a sample (selected probabilistically) of a population of interest. Over time, data from outside the collection process gradually became more available. As a result, these data are now being used to make inferences even though they were originally collected for other purposes. The need to reduce the response burden, the quality of the external data available, and budget constraints are reasons commonly cited in support of this change. Consequently, national statistical agencies are making increasing use of administrative records, which are required for administering various non-statistical programs, to complement, or even replace, survey data.

Another option generating interest among pollsters involves the use of big data. This is not surprising, since the information technology revolution and the increased use of social networks in recent years have made available a large volume of varied, constantly renewed data. A profound transformation in survey-taking is under way: survey statisticians are increasingly expected to produce statistics from survey data, administrative data and big data at the same time. An example of this is a consultation that Statistics Canada carried out on behalf of Environment Canada (EC) to assess the quality of the information contained in the EC mega-databases on wetlands. To make sense of the mass of information in these databases, we used the particular attributes of big data, administrative data and survey data as references.

In this paper, we describe the preparatory work for the statistical analysis that was the primary purpose of this consultation. This work shows how we dealt with these three types of data in the context of a statistical consultation. In Section 2, we briefly discuss the statistical consultation, which falls outside the traditional survey activities that we conduct at Statistics Canada. In Section 3, we describe the challenges that may be encountered with very rich but unstructured data, and how to make sense of such data using three main references. Having established this framework, we then describe how it was applied in the wetlands consultation that we conducted on EC's behalf. Since wetlands are not a common topic, we also provide a brief overview of the subject, before moving on to a detailed analysis and summary of the available information.

[1] Business Survey Methods Division, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa ON, Canada, K1A 0T6.

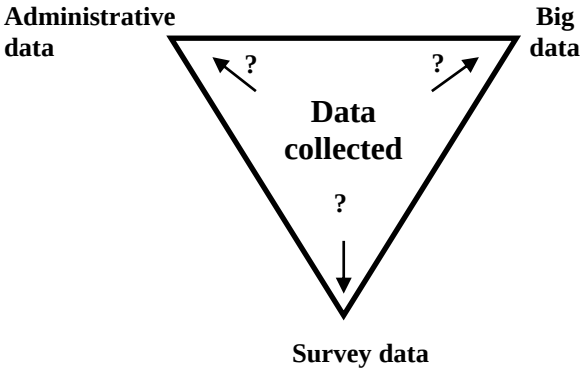
2. The fundamentals of consultations

Statistical consultations are generally structured around three main points, each of which involves a particular challenge. The first step in a consultation is to understand the scientific question being asked. In many cases, the question may not be entirely clear even to the client requesting the consultation. One or more meetings may be necessary to pinpoint the problem. Once the problem has been defined, the second step is to identify an appropriate statistical model. It is imperative to have both statistical tools (explore the various methods) and logistical tools (availability of software applications) in addressing the question. This depends largely on what data are available, how reliable the available data are (data coherence, missing values, outliers), and what resources have been allocated to the project. At this stage, the consultant must find a compromise between efficiency and operationality, and above all, set up a contract with the client that provides a detailed explanation of what the consultant plans to do, and verify that this solution addresses the client's objectives. Following the statistical analyses, the last, but not least important, step is to communicate the results to the client. Though it may seem straightforward at first glance, this can be the most difficult step in the entire process. Work groups are often composed of people with different levels of scientific awareness. If that is the case, the report must be stripped of highly technical content to make it accessible to all concerned.

As mentioned in the introduction, the advent of big data presents an additional challenge for statistical analysts in general. Data that were designed and collected for administrative use cannot be employed for statistical purposes without researchers having to first overcome a number of pre-existing issues. In the past, researchers collected data related to a clear, precise research question. On the whole, the subsequent statistical analyses were logical, albeit sometimes compromised by the way in which data collection was planned. In the current context, the key challenge is no longer a lack of data at the outset, but the avalanche of data now available. How can one make sense of it? To make the transition from this raw information to summarized, usable information, one first has to be able to make sense of it. We want to emphasize that this abundance of data can lead just as easily to very high-quality inferences as to completely erroneous conclusions if one fails to take certain precautions in advance. In the next section, we consider the question of how to make sense of this mass of information by introducing three key references.

3. Making sense of a mass of very rich but unstructured data

In this section, we address a fundamental issue that is central to any statistical process: the nature of the data to be analyzed. It is a key factor in deciding which methods to use. What type of data are we dealing with? This question is especially timely when data collection planning is completely beyond the statistician's control. To make sense of a large mass of information, we think it is crucial for statisticians to fully understand the attributes of the available data so that they can assess their quality and the role they might play in a statistical analysis. This will also help to clarify the researcher's purpose and the limitations of using such data. One way of doing so is to take one's bearings in relation to three key references as shown below.



Each of the three data types will play a different role in a statistical analysis. For example, administrative data will generally be available for all units in the database (with the exception of a few missing entries), while survey data are limited to the sample. In the former case, it is quite straightforward to produce an estimate of a parameter for the population of interest based on the administrative file, but in the latter case, much less so. We make inferences about the quantities of interest using only the sample units and the set of associated sampling weights. To dispel any doubts, it is important to clarify how we visualize each of these data types before we move on to the case study.

3.1 Administrative data

Administrative data are usually collected for operational reasons and not for statistical purposes. For example, the Canada Revenue Agency gathers and regularly updates information about citizens and residents to facilitate tax collection. Using administrative data for statistical purposes is increasingly attractive because it lowers collection costs significantly, and reduces the response burden. However, the concepts used in the administrative data are specific to the service being provided and therefore often vary from source to source. This raises a potential issue with their quality, such as the coverage of the population of interest. As noted by Michael Brick (2011), the use of administrative data over many years has not produced the desired effect. Referring to Jabine and Scheuren (1985), who described six objectives set by the U.S. government for improving the use of administrative data over a 10-year period, he observes that, 25 years later, some of these objectives still have not been met.

3.2 Big data

The use of big data in surveys is growing. Big data are often described with the three Vs: volume, variety and velocity. Douglas Laney was one of the first authors to use the three Vs to describe this type of data, in 2001. The volume part of this description reflects the fact that these data present us with a deluge of information (of the order of 1 zettabyte, or 10^{21} bytes). Since the data being analyzed are no longer necessarily structured as they are in traditional statistical analyses, but may consist of text, images, multimedia content, digital trails, connected objects and so on, they are referred to as varied data. In some respects, big data are similar to administrative data in that, ultimately, the concepts measured by these data are not necessarily the same as those measured by surveys. Despite the promise of reduced collection costs, big data can generate costs associated with information storage and the difficulty of processing and extracting information. Velocity refers to big data's capacity for very rapid turnover. In an inference context, big data also raise quality issues such as those relating to coverage and coherence.

3.3 Survey data

These data are intended for statistical analysis. They are collected on the basis of clear concepts for survey purposes, and, as a general rule, their collection is planned and carefully controlled. This ensures high-quality data, the data type most commonly found at Statistics Canada. Following collection, the data are formatted, structured and verified in an automated process. The final product is generally of good quality and ready for use in producing inferences. The main drawback lies in the fact that the units are collected for only part of the population: the sample. Then inferences must be made for the entire population. This generally results in sampling error in the estimates.

4. From principle to practice: Case study

In this section, we describe, as a case study, the consultation that we conducted for EC on wetlands. EC has a large mass of disparate data on every aspect of Canada's territory, stored in huge databases. As part of one of its environmental protection projects, EC would like to use this information for wetland monitoring. Before using the data, however, EC would like to assess their quality. How can we get the most out of the data? How much confidence can we have in this mass of information? These are some of the issues of concern to EC. First, though, we examine some of the reasons for the interest in wetlands.

4.1 Wetlands “101”

While wetlands represent only about 6.4% of the world’s surface area, they play a key role in the survival of the species around us. Wetlands are of particular interest to Canada, since 25% of them are found in Canadian territory. The following is a partial list of the key roles that wetlands play in the environment:

- A giant natural sponge: wetlands store and release water, regulating and supplying water courses.
- A natural filter: their composition promotes the biochemical transformation of organic materials and minerals.
- A natural temperature regulator: storage of carbon helps regulate the climate.
- A natural reservoir of life: many species live in wetlands because they provide an ideal habitat.

According to the geophysical literature, experts generally agree that there are five main types of wetlands: shallow water, marshes, swamps, ombrotrophic bogs and minerotrophic bogs (fens).

Following this brief introduction to wetlands, we now look at where EC’s large collection of information stands in relation to the three data types. Are we dealing entirely or partly with administrative data? Big data? Or just survey data? We want to emphasize the fact that classifying data in this way will help us to understand their strengths and weaknesses and to assess their future usefulness in a statistical analysis.

4.2 Summary of the available information

4.2.1 Administrative data

As previously noted, a large portion of EC’s information has the attributes of administrative data. These data were collected through various provincial government departments and agencies that are partners of EC. They were used in this project to divide Canadian territory into geographic parcels referred to as wetland polygons. These polygons serve as record entities or units in our database. It is important to note that this information was not originally collected for the purpose of wetland monitoring. In most cases, the data were collected to facilitate management of various government services. Consequently, the challenge here is enormous. The polygons must be defined consistently and in such a way as to maximize the available information. For example, consider the slope. Without the slope, one might think that Canada’s geography can be depicted in two dimensions. The slope adds a third dimension to the picture. If this information is not collected for all units or is inconsistent from unit to unit, it cannot be taken into account in dividing the territory into polygons.

Another of EC’s objectives was, following the division into polygons, to assign each of the resulting polygons to one of the predefined categories described in subsection 4.1. A definitive way of doing so would have been, for example, to hire experts to visit each polygon one by one and determine which category to assign it to. However appealing this solution might seem, it is obviously unrealistic given the amount of time and the resources required. For example, the division process gave rise to hundreds of polygons for the smallest database, the one used as a pilot project. To remedy this problem, an approximate classification based on big data modelling was adopted. That is the subject of the next paragraph.

4.2.2 Big data

The EC databases had some of the attributes of big data. To perform the above-mentioned classification, EC used the huge mass of information provided by satellites. Satellites do a continual periodic sweep of the territory and produce information that is a rich as it is varied. This information was used in this project to perform an automated classification, in which each parcel was assigned to one of the five wetland categories. The advantage of this

approach is that the polygons can be classified in a very short time, despite their large number, since it is essentially a computerized process. However, this processing yields a classification that is imperfect because it results from a model and not from an expert study of the wetlands.

4.2.3 Survey data

We further noted that, in EC's mass of information, there was a third type of data, with characteristics different from those of the first two types. A very small portion of the database had been classified by an expert (an independent consultant) into one of the five categories; this classification agreed with the model in some cases, and not in others. For the database used in the pilot project, we had an expert opinion for a sample consisting of about 2% of the polygons. EC sensed the potential of this new information source, as the consultant's classification was based on a more detailed study or even a visit to the site. At first glance, the arbitrary nature of the sample of polygons examined by the consultant made it difficult to use. Is there some way of quantifying the discrepancy between this classification and the one provided by the model for this sample? If so, can one extrapolate these results to the entire database? A detailed study of the reports produced by the consultant revealed that we were dealing with a probabilistic sample. More precisely, the consultant had taken the trouble to divide Canada into various strata and select from those strata a sample of polygons. Thus, statistical inference for the entire database was possible. One of the ways in which these data were used is summarized in the table below.

Table 4.2.3-1
Summarized data

id	X	Y
1	BOG	∅
2	FEN	
		Bog
		Marsh
N	Swamp	Fen

The administrative data are used to construct the ***id*** sampling units; the automated classification based on the collection of big data from satellites is used as auxiliary variable ***X*** (collected for all units in the database); and the consultant's classification is used as variable of interest ***Y*** (collected for the subset of selected polygons). In particular, we produced an estimate of the proportion of each type of polygon that the consultant would have found if he had had the opportunity to go through the entire database.

5. Conclusion

To sum up, in this paper, we showed how various types of data can be dealt with in the context of a statistical consultation. We started with a mass of data that was very rich but unstructured, making it difficult to use. Before conducting the statistical analyses that were the primary purpose of the consultation, we organized, analyzed and summarized the available information. The consultation focused on three main types of data: administrative data, big data and survey data.

6. Acknowledgements

We would like to thank Arthur Goussanou, Christian Olivier Nambeu, Nathalie Hamel and Wesley Yung for all their efforts in reviewing this paper.

References

Brick, J. Michael. (2011), "The Future of Survey Sampling", *Public Opinion Quarterly* 75(5), p. 872-888.

Jabine, Thomas, and Fritz Scheuren. (1985), "Goals for Statistical Uses of Administrative Records: The Next 10 Years", *Journal of Business and Economic Statistics* 3(4), p. 380-391.

Laney, D. (2001), "3D Data Management: Controlling Data Volume, Velocity, and Variety", Technical report, META Group.