WHY MORE AND BETTER DATA ARE KEY TO AN EVIDENCE-DRIVEN CANADA

JAN KESTLE

Jan Kestle is a President and Founder of Environics Analytics, a leading provider of data and insights to Canadian businesses, governments and not-for-profits. She has worked in the statistics community in Canada for 50 years, is a member of the National Statistics Council and Director of the Board of the Canadian Marketing Association.

When the mandatory long-form census was cancelled and replaced with the voluntary survey in 2011, it was surprising to hear experienced researchers quoted as saying that we have so much big data we don't need a detailed census. The truth is that while many of the new sources can augment our production of data, official reliable statistics are required to build an evidence-driven economy and an effective social policy framework.

With the reinstatement of the long-form census, the dust has settled — at least for now. All agree that the results from the National Household Survey (NHS), which temporarily replaced the long-form census, were questionable at best. As a private company that builds small area data for use by businesses, governments and not-for-profits, we look back at the predictions from our models to see how they compared to what was subsequently measured. Most results from the NHS were not usable; for the data that were of sufficient quality, there were still obvious "breaks" in the series based upon historical review, despite applying best practices.

Hopefully our politicians have learned from this ill-conceived decision. But to ensure it doesn't happen again, the statistical community must convince voters that the census is the cornerstone of Canadian data. As StatCan moves into its second century, two things are essential. First, we must protect and expand the census, the economic accounts, the small area administrative data program and the key survey programs. Second, StatCan has to develop new methods of sourcing and producing data. The current modernization efforts that are underway, which include public-private cooperation, mining

big data and enhanced use of administrative data, are crucial to StatCan's future. But these initiatives will not succeed unless new approaches are combined with strong support for the fundamental programs.

A good example of the challenges and opportunities that StatCan can expect to encounter, as it tests and employs new approaches, emerged recently in another headline story. StatCan was going to conduct a pilot accessing detailed expenditures from personal bank records to test the use of these administrative bank records as a possible source to replace the Survey of Household Spending. Why? Because it's increasingly difficult to get Canadians to fill out extensive surveys, such as this one that is so crucial to the economy.

The plan was to obtain, under The Statistics Act, information about what people spend their money on. From a large roster, a sample would be developed and then the data would be aggregated and used to produce official statistics. A key test in the design of the pilot was to see if the random sample of bank records would be representative enough to replicate the survey results or if the data would be of better quality if they were weighted by the demographics that are typically used in survey design.

StatCan was planning to acquire the spending data along with personal identifiers in order to append select census attributes to the bank records. Those identifiers would be deleted after they were used to match the records. The anonymous records would be processed to produce the statistics. This approach, using appended census data, would be tested against using the same bank spending data without adding the census attributes for sample design and weighting. StatCan had discussed the approach with the Canadian Bankers Association and the Office of the Privacy Commissioner. The fact that the data would be transferred under the protection of the *Statistics Act* would have ensured that no personal information would be disclosed or stored.

This seemed like a reasonable plan to get good quality data without a significant burden on respondents while ensuring the full protection of confidentiality for the data being used. But it didn't go well. The press and the opposition took great exception to this "spying" on Canadians without their permission, and eventually the pilot was shelved.

It's worth adding that StatCan is well equipped to handle large volumes of data with all the privacy and security provisions that come with them. However, whether you are working with a single record or millions of files, what matters is how those data were collected. We believe that analysts can draw geographically pertinent conclusions without collecting personal information.

WHAT DID WE LEARN?

For citizens to share these kinds of sensitive data with their government, they have to trust their government, which apparently they do not. One critical report cited the possibility that allowing the government to look at the bank records would make it possible to identify individuals who cheat on their taxes. Of course, this would not be possible since no data would be permitted to leave StatCan (i.e. it could not legally be shared with any other government department), but in addition the concern seems a little ironic.

Clearly, we in the data community are not doing a good enough job at helping Canadians understand why good data are so important to their lives — for community safety, compensation and employment policies, as well as research, innovation, health, education and social services planning.

Also, the fact that StatCan has long been rated one of the top statistical agencies in the world and can be trusted with confidential information does not seem to be well known or accepted. When I pointed out to people that the personal information would only be used to append other data and then removed, many people indicated clear disbelief — once StatCan gets the data, they said, "the government (Big Brother) will be watching us."

Lost in the debate was the fact that many countries use these kinds of administrative data sources combined with census and other data as a substitute for surveys. If an independent and world-class organization like StatCan is prohibited from accessing administrative and alternate data sources, our policymakers and business leaders will have to rely on a number of different organizations that have those types of data to tell us their version of what is happening in Canada. Better that these important data be enhanced and integrated into the national statistical framework with oversight from StatCan.

SO WHERE DO WE GO FROM HERE?

A lot of data, from anonymized mobile phone location data to the Internet of Things reports on things like energy consumption, can be leveraged to design, develop and measure government programs. But they need to be normalized, cleaned, weighted and aggregated using more traditional data to ensure accuracy and reliability. Official data can make big data usable and reliable.

The experimenting has to continue to ensure that we not only preserve what we have, but that we fill in the many gaps. Silos across federal departments and between the federal and provincial governments need to be overcome. Producing the volume and variety of data that are needed in our changing world will require significant changes in our approaches.

Citizens certainly have a right to have their personal data protected under the laws that enshrine privacy. But if the data producers can show that not only will the data be protected but will actually be used to make people's lives better, we will be able to move closer to evidence-driven decision making in Canada.

Politics aside, these new approaches must be explored. We are striving to decipher fact from fiction. Statistics Canada is the right organization to lead these continued modernization efforts. Canadians and their government should give the agency a chance to do that and make these next 100 years as successful as the previous century.