The use of Administrative Data at Statistics Canada

Wesley Yung[1,2], Pierre Lavallée[1], and Julie Trépanier[1]
[1] Statistics Canada, Ottawa, CANADA
[2] Corresponding author: Wesley Yung, e-mail: Wesley.Yung@statcan.gc.ca

Abstract

It is well known that administrative data have many advantages including, but not limited to, supporting surveys in producing better quality results, reducing respondent burden and collection costs and possibly filling data gaps. Administrative data can be integrated into survey programs in many different places in the survey process. For example, administrative data can be used to improve sampling frames, validate responses at the editing stage or be used as auxiliary data at estimation. Statistics Canada, like many other national statistical agencies, has recognized the advantages of administrative data and has integrated it into many of its statistical programs. In this paper, an overview of the use of administrative data at Statistics Canada will be given covering both economic and social statistical programs. The integration of administrative data into several of these programs will be described in some depth. In addition, some of the challenges and potential risks of incorporating administrative data into survey programs will be discussed.

Keywords: Census of population, data integration, response burden, tax data, Unified Enterprise Survey

1. Introduction

Statistics Canada has a long history of using administrative data in its statistical programs. For instance, provincial and territorial vital statistics have been directly tabulated to produce national level statistics since the early 1920s and customs import and export data since the 1930s. For more examples of such early use, we refer the reader to Rowebottom (1978). As mentioned in that paper, the development of computers in the 1970s allowed easy access to records stored in databases and to share data amongst different agencies. As readily accessible administrative data became more available, official statisticians explored ways to incorporate them into survey programs as opposed to using them for direct tabulations. Reasons for this include cost savings, reduction of burden on respondents (from both surveys and administrative requirements) and possible quality improvements in the data.

Since the early 1970s, Statistics Canada has made many advances in the use of administrative data, to the point where they are now highly integrated into many of its statistical programs. Unfortunately, the integration of administrative data into statistical programs has not happened in a coordinated and uniform manner so that their use is somewhat program dependent. In response to this, Statistics Canada has developed a corporate strategic vision for the use of administrative data:

> *Statistics Canada will use administrative data in its statistical programs when they lead to a better outcome — that is, a better balance between relevance, quality, costs and respondent burden. In doing so, Statistics Canada remains committed to respecting the privacy of individuals and keeping the information provided to us confidential.*

In order to help attain this strategic vision, the Administrative Data Secretariat was created in the fall of 2012 with the mandate to develop and implement a corporate approach to increasing the use of administrative data at Statistics Canada.

In this paper, we present the current state of utilization of administrative data at Statistics Canada in the context of both business and social surveys. The various sources of administrative data used at Statistics Canada will be discussed in section 2 and the current uses will be presented in section 3. Section 4 will illustrate these uses for a business survey and a social survey. The cost and risks of using administrative data will be discussed in section 5, with a few closing remarks in section 6.

2. Sources of Administrative Data

As laid out in Section 13 of the Federal Statistics Act 1970-71, C. 15, S.1, Statistics Canada has the legal right to access administrative data held by other organizations. Even with such far reaching legal access, Statistics Canada has adopted an approach of forming partnerships with potential data providers in order to ensure a continuous flow of high quality administrative data (Yung, Leblanc and St-Louis, 2011). With this approach, Statistics Canada has formed partnerships with various organizations to obtain the following administrative data (Lavallée, 2007):

- The Canada Revenue Agency (CRA) provides various annual income tax files and schedules covering incorporated and non-incorporated businesses, monthly Goods and Services Tax (GST) data, statement of remunerations paid for employees, Business Number (BN) administrative files, child tax benefit files and administrative data from the program for the reduction of the GST for new construction.
- Various provincial and territorial agencies provide information on births, deaths, health insurance, drivers licenses and vehicle registration.
- Citizenship and Immigration Canada provide administrative information on non-permanent residents and landed immigrants.
- Data are obtained from the National Coroners and Medical Examiners database.
- Federal, provincial and municipal files on court appearances, correctional involvement and police files of incidence reports are obtained.
- Administrative files from universities and colleges are received.
- Human Resources and Skills Development Canada provides employment insurance data.
- Documents on travelers and vehicles entering Canada are received from the Canada Border Services Agency.
- Telephone files are obtained from multiple sources.
- Aviation administrative data are obtained from Transport Canada, NAV Canada and other sources.
- Canada Housing and Mortgage Corporation provides data on housing starts and completions.
- Investment data from Morningstar Corporation, Investment Funds Institute of Canada and published financial statements are obtained.
- The Indian Register from Aboriginal Affairs and Northern Development Canada is received.

Clearly, Statistics Canada has embraced the use of administrative data but unfortunately there has not been a coordinated approach. This has resulted in some individual divisions managing the acquisition and use of administrative data themselves. Statistics Canada's Tax Data Division was created in the late 1990s to coordinate the utilization of tax data received from the CRA, but there has been little done to govern the other many sources of administrative data. The previously mentioned Administrative Data Secretariat will take a lead role in developing and implementing a corporate approach to the acquisition, use and management of administrative data at Statistics Canada.

3. Current Uses of Administrative Data at Statistics Canada

Brackstone (1987) categorized the use of administrative data into direct tabulation, indirect estimation, survey frames and survey evaluation. However, in this paper we discuss the use of administrative data at Statistics Canada under the following headings: improving survey results, reduction of respondent burden, reduction of statistical program costs and filling data gaps.

3.1 Improving Survey Results

Administrative data are used in many stages of the survey process. In terms of the frame and sample design steps, administrative data are used extensively in Statistics Canada's Business Register (BR) and the Address Register (AR). Data from the CRA are used by the BR to identify births to the business population through the BN administration file since businesses must register with the CRA to obtain a BN before operating in Canada. Once a business appears on the BR, income tax and remuneration data are used to provide measures of size (annualized revenue, number of employees and salaries), which are used as design variables in business surveys.

Statistics Canada's AR was originally developed as a coverage improvement tool for the 1991 Census of Population and was created by combining personal income tax data from the CRA, telephone billing files, electricity billing information and municipal property tax records. As the quality of information on the AR increased, its use has grown to being used as a list frame for about 80% of addresses in the country for recent censuses and as a tool for cluster listing activities in Statistics Canada's Labour Force Survey. The sources of administrative data have expanded to include commercial telephone directory files, new housing tax rebate records and child tax benefit records.

In addition to being used in the frame and sample design stage, administrative data are also heavily used during the processing stage. For instance, they are used to validate survey data through the form of edits and as auxiliary data to detect outliers. During the imputation process, they are used to directly replace survey data for non-respondents, as auxiliary data in imputation models or as matching variables used to find donor records. During the estimation stage, they are used as auxiliary data for calibration estimators, which are typically more efficient, in terms of variance, than estimators that do not use auxiliary information. Finally, they are also used during the data confrontation stage to validate aggregated estimates coming from surveys.

3.2 Reduction of Respondent Burden

It has long been recognized that the heavy burden placed on respondents by statistical agencies can be reduced through the use of administrative data (Rowebottom, 1978, and Brackstone, 1987). In fact, some Statistics Canada programs are based solely on regulatory administrative data such as trade data, vital statistics and justice data. Other programs supplement survey data with administrative data to reduce respondent burden. For example, many business surveys use monthly or annual tax data to replace financial survey data for a portion of their sample. Social surveys commonly use income tax data from the CRA instead of collecting them from respondents. These surveys include the National Household Survey, the Survey of Labour Income and Dynamics, the Survey of Household Spending and the Longitudinal and International Survey of Adults. A less direct reduction of respondent burden is the use of administrative data as auxiliary data to improve the efficiencies of estimators through calibration, as mentioned in 3.1. By having more efficient estimators, sample sizes

required to attain targeted levels of precision can be lowered, thus reducing the overall burden placed on the survey population. Another example would be the use of administrative data to improve the information at the design stage, which allows for more efficient sample designs and thus smaller sample sizes.

3.3 Reduction of Statistical Program Costs

The use of administrative data can decrease survey costs by reducing collection effort and helping in frame creation and maintenance. Collection costs can be reduced by using administrative data directly, either in a survey or direct tabulation context, or by using them as auxiliary data to improve the efficiency of survey estimates. Frame creation and maintenance costs can be reduced by using administrative data, as discussed in section 3.1.

3.4 Filling Data Gaps

A common use of administrative data today is to combine multiple sources into a single database to fill data gaps and to provide analytical opportunities. These databases commonly combine administrative data with survey data to allow researchers to perform analyses on varied topics. Statistics Canada has produced the Linked File Environment (LFE) that has combined 13 different data sources, of which six are administrative data. The LFE provides researchers cross-sectional and longitudinal information to inform analysis of business strategies and their outcomes. See Jiang and Kozak (2011) for more details.

On the social side, Statistics Canada has developed several longitudinal databases using administrative and survey data. For example, the Longitudinal Immigration Database combines immigration and tax records. It allows the analysis of relative labour market behavior of different categories of immigrants to assess the impact of characteristics such as education and knowledge of French or English. Another example is the Longitudinal Administrative Database (LAD) that combines a 20% sample of persons from the income tax with the Longitudinal Immigrant Database. The LAD is a research tool for the analysis of income and demographics and is used by many government departments to evaluate programs and support policy recommendations. For more information on these products, visit www.statcan.gc.ca.

4. Illustrations of Administrative Data Use

As one can see, administrative data can be found in many aspects of the work done at Statistics Canada. In this section, we illustrate their use in two particular projects, the Unified Enterprise Survey, which encompasses approximately 60 annual business surveys, and the 2011 Census of Population and the related National Household Survey.

4.1 The Unified Enterprise Survey (UES)

Administrative data are used by the UES in the following steps of the survey process:

- Statistics Canada's BR is used as the source of the UES frame. The size measure used as the stratification variable is either the annual revenue based on sales from annualized monthly tax data, the annual sales from annual tax data or the profiled revenue (available for only a small percentage of enterprises).
- Tax data are used to impute for non-respondents if other methods are not applicable. For instance, if historical data are not available, then tax data will be used. In addition, for a predefined subset of units, identified at sampling, tax data are used to

replace survey data. These units, although selected in the sample, are not sent a questionnaire and tax data are used directly or in imputation models.
- Since the 1990s, Statistics Canada's business surveys have had a take-none stratum consisting of the smallest businesses who collectively contribute less than 10% of the total revenue within the industry/geography domain. These units are not eligible to be selected in the sample and their contribution is estimated through the use of tax data. Although these units are small, they are large in number.
- During the analysis stage, tax data are available at the individual business level (or micro level) and at aggregate levels (or macro levels). Analysts frequently use tax data to validate both micro-records and the macro estimates.

4.2 The 2011 Census of Population and the National Household Survey

Statistics Canada is mandated to carry out the Census of Population every five years, with the last one occurring in 2011. Historically, the Census Program consisted of two questionnaires, a short form covering basic demographic questions that was delivered to 80% of dwellings in Canada and a long form, including those same questions as well as a number of more detailed questions, that was delivered to 20% of dwellings. In the 2011 Census Program the short form or Census questionnaire was delivered to all dwellings and the long form was replaced by the voluntary National Household Survey (NHS). We now present the use of administrative data in the 2011 Census Program, which covers both the 2011 Census of Population and the 2011 NHS.

- Starting in 2006, the AR has been used as a source of addresses for mailing out the Census Program questionnaires. In 2006 close to 70% of the questionnaires were mailed out and in 2011, this percentage increased to 80%.
- The NHS questionnaire contains content related to income. Respondents are asked if they would prefer to have Statistics Canada link to their income tax records instead of filling out the section on the questionnaire. In the 2006 Census, over 80% of long form respondents replied positively to this question. Unfortunately, at the time of writing this paper, the figure for the 2011 NHS is not yet available.
- With the voluntary nature of the NHS, there were concerns that the non-response could be higher than that experienced with the long form in the past. While administrative data were not used directly to treat non-response, they were used as auxiliary data in the total non-response adjustment procedure improve the matching of donors with recipients.
- Before Census Program estimates are published, they are thoroughly reviewed by analysts. During this certification process, several administrative data files are used to validate the estimates. For instance, estimates of intercensal immigration to Canada are compared to counts from Citizenship and Immigration Canada.
- Census coverage studies in Canada make extensive use of administrative files. Without going into details, administrative files are used to build frames for intercensal births (vital statistics and Canada Child Tax Benefits records), intercensal immigration (immigration files) and the three northern Territories in Canada (Territorial health records). For more details on the use of administrative data in the Census Program, see Dolson (2011).

5. Costs of Using Administrative Data

It is clear that using administrative data for statistical purposes has many advantages, however there is a cost and some challenges to the use of them. Obviously, there is the cost of acquiring the administrative data themselves. Statistics Canada currently pays a significant amount for the data from CRA and if the trend of using data from for-profit sources continues, we can expect to pay even more in the future. Over and above the acquisition cost, Statistics Canada has invested a large amount of money, time and

effort to ensure that the administrative data are fit for statistical purposes. Statistics Canada's Tax Data Division's (TDD) primary role is to liaise with CRA, obtain administrative data and then process them for use by Statistics Canada's statistical programs. TDD currently has approximately 50 employees within the division, an extensive IT infrastructure in place and supporting players, such as methodologists, dedicated to the purpose.

In addition to the tangible costs, there are other aspects that should be considered. For instance, as more and more administrative data are integrated into statistical programs, national statistical agencies become more reliant on external sources for their data. This introduces the risk that a statistical program may not receive the required data due to changes outside of their control. Yung, Leblanc and St.-Louis (2011) discuss the importance of developing partnerships with data providers to mitigate this risk.

Finally, as national statistical agencies move towards increasing the use of administrative data, care must be taken so as not to allow their survey taking infrastructure to be reduced or eliminated. As powerful and useful as administrative data are, they are not always available in a timely fashion. Without an adequate survey taking infrastructure, a national statistical agency may not be able to react to demands for short term statistics as a result of particular events such as the economic downturn experienced in 2009/10. Without a readily available survey infrastructure, statistics required by policy makers could not have been made available from administrative sources in a timely enough fashion for them to have reacted.

6. Summary

Administrative data have been integrated into many different statistical programs at Statistics Canada and will be more so in the future. Due to the availability of taxation data for quite some time, it is natural that business surveys may be slightly ahead of social surveys in the use of administrative data, but as more and more sources are found, it is expected that social surveys will quickly catch up to business surveys. Despite the numerous advantages of administrative data, national statistical agencies must recognize the importance of maintaining solid partnerships with data providers and also realize that administrative data also have some costs and limitations.

References

Brackstone, G. (1987). Issues in the Use of Administrative Records for Statistical Purposes. *Survey Methodology*, **13**, pp. 29-43.

Dolson, D. (2011). Administrative Data Use in a Traditional Census. *Proceedings of the 58th World Statistics Congress in Dublin*, Ireland.

Jiang, M. and Kozak, R. (2011). *Record Linkage for the GBPS Linked File Environment Version 7 (v07)*. Internal report, Statistics Canada.

Lavallée, P. (2007). *Administrative Data usage in the Framework of Social Statistics: Current and Future Picture.* Internal report, Statistics Canada

Rowebottom, L.E. (1978). The Utilization of Administrative Records for Statistical Purposes. *Survey Methodology*, 4, pp. 1-15.

Yung, W., Leblanc, D. and St-Louis, G. (2011). The Use of Tax Data in Official Statistics – The Canadian Experience. *Proceedings of the 58th World Statistics Congress in Dublin*, Ireland.