Proceedings of Statistics Canada Symposium 2016
Growth in Statistical Information: Challenges and Benefits

The Alternative Data Solution – Experience of the Producer Prices Division

Gaétan Garneau and Mary Beth Garneau[1]

Abstract

Over the last decade, Statistics Canada's Producer Prices Division has expanded its service producer price indexes program and continued to improve its goods and construction producer price indexes program. While the majority of price indexes are based on traditional survey methods, efforts were made to increase the use of administrative data and alternative data sources in order to reduce burden on our respondents.

This paper focuses mainly on producer price programs, but also provides information on the growing importance of alternative data sources at Statistics Canada. In addition, it presents the operational challenges and risks that statistical offices could face when relying more and more on third-party outputs. Finally, it presents the tools being developed to integrate alternative data while collecting metadata.

Key words: data; alternative; administrative; challenges; risks; prices; production; indexes.

1. Introduction

1.1 Description

Over the last decade, Statistics Canada's Producer Prices Division has expanded its service producer price indexes (SPPI) program and continued to improve its goods and construction price indexes program. While the majority of price indexes are based on traditional survey methods, efforts were made to increase the use of alternative data, which include administrative data and other data sources. This search for data sources other than traditional surveys aligns with the agency's efforts to reduce respondent burden and to stabilize or reduce its collection costs. These new data sources come in many forms, such as list prices on the Internet, third-party data files, and microdata from turnover surveys.

Section 2 summarizes the various alternative data sources available to national statistical offices (NSOs) and Statistics Canada, along with the challenges associated with using these data and potential solutions. It also presents a few examples of how Statistics Canada uses alternative data to produce its price indexes in the transportation, financial, construction, goods and professional services industries. The third section describes how Statistics Canada and the Producer Prices Division are working to develop tools to integrate a wide variety of data. Finally, Section 4 concludes the paper by providing ideas for consideration. This paper is an adaptation and update of a paper written for the Voorburg Group on Services Statistics (Garneau 2015).

2. Alternative data sources

Many NSOs, including Statistics Canada, use a wide range of data sources that are not directly collected from surveys managed by the agency. Furthermore, seeking alternative data sources is increasingly encouraged to reduce respondent burden and costs while maintaining a high level quality for Statistics Canada. Indeed, program quality is very often enhanced by using these alternative data directly in the production or the validation of data from so-called traditional surveys. In his annual address in March 2016, Statistics Canada's Chief Statistician reaffirmed the importance of alternative data (administrative and other) to the agency's programs.

[1] Gaétan Garneau, Statistics Canada, Tunney's Pasture, Ottawa, Canada, K1A 0T6 (gaetan.garneau@canada.ca);
Mary Beth Garneau, Statistics Canada, Tunney's Pasture, Ottawa, Canada, K1A0T6 (marybeth.garneau@canada.ca)

Generally, a national statistical office has strong control over the regular survey data it compiles and processes, but in the case of data from sources other than regular surveys, it is normal to find that each source has its own unique challenges and risks. In this section, we will look at some of these issues, describe how to mitigate the risks and, finally, offer a few examples.

2.1 Main alternative data sources

The five main data sources used at Statistics Canada and, more specifically, in the Canadian producer price indexes program are:

- prices on websites
- prices in catalogues or newspapers
- data files from third parties in the private sector
- administrative data or data from other departments or regulators
- data files of other surveys.

The *Statistics Act* provides the authority to Statistics Canada to obtain access to any data maintained by any federal, provincial or territorial department or by any municipal office, corporation, business or organization for the purposes of the Act. The Act also requires that the confidentiality of the data acquired under the *Statistics Act* be strictly maintained. As a complement, Statistics Canada also uses information available to the public, under licence or not, with or without a fee.

2.2 Prices on websites

Prices posted on the websites of companies in the samples of producer price index programs can be a source of valuable data, especially if these prices (list prices) are similar to or only slightly different from actual transaction prices. Statistics Canada and the Producer Prices Division collect information on prices or rates on many websites such as the prices of courier services. In this specific example, individual company rates or pricing schedules are readily available. Rate finders or online invoice calculators can also be used to get price estimates. Finally, prices are collected monthly for a detailed set of price specifications covering geography (i.e., origin and destination of service), type of parcel and type of service (express or non-express).

At this time, these data are collected from websites manually. Besides the known limitations of using list prices instead of actual transaction prices, there are a number of things to consider when using the Web as a source of data. In some industries, prices may be impacted by the number of views or the device used to access the website. For example, repeated searches on travel sites may be counted as increased demand which might lead to price increases. The use of a device or medium to access the website may also impact the price. Some travellers have noted that prices quoted on mobile devices are sometimes higher than prices quoted on a desktop computer. It is therefore important to ensure that the website is reliable and will be available for ongoing price collection. Its disappearance could create headaches for price index programs that rely on the availability of prices on an ongoing basis.

Data collection on the Web could potentially be handled at Statistics Canada by automated means, such as “web scraping”. In fact, not only could web scraping reduce collection costs, but it could also collect much more data quicker. To date, Statistics Canada has not used web scraping for regular data collection for its programs. A pilot project to learn more about the option is under way. Since it was launched, the project has identified a number of potential problems associated with the use of this collection method, in addition to the benefits mentioned earlier.

First, prices on some Web pages or websites are provided in such a way that it is difficult or impossible for a web scraper to find them, at least not with any standard methods known to Statistics Canada. In collecting prices on the Web for the Freight Rail Services Price Index, for example, Statistics Canada captures prices that are displayed on images of price lists. In collecting data on the Web for courier services, some sites only display a price when you hover over a very specific area of the page with a mouse, and frequent changes to the website are also a potential impediment to web scraping.

In addition, the terms and conditions of some websites prohibit the use of automated web scrapers. One reason for firms to prohibit web scraping is that it could generate too much traffic and bring down a website. If that were to happen, it could be detrimental to the reputation of the NSO responsible. It could also leave the NSO liable for damages.

Finally, there are concerns about public perception of an NSO's use of web scraping. Some web scraping software (not used by Statistics Canada) uses intrusive methods that might be perceived as security attacks against websites. This technology has at times been associated with illegal activity and malicious attacks.

To minimize the risks associated with web scraping, it would be a good idea for NSOs to first contact website owners before scraping data. The web scrapers used should not be intrusive nor should they create a security risk for the NSO. Finally, scraping should be minimal such that the website's server is not overloaded. Statistics Canada is adhering to these approaches in its feasibility study on the use of web scraping as a data/price collection method. One new method is to get direct access to the database supporting the website. This database is the source of the prices generated when users enter search parameters for product prices. This application programming interface (API) is a solution that sets out how programs seeking to access the database may do so. With the consent of the website, this solution is increasingly used to provide access to the databases of the target sites and thus avoid the problems associated with web scraping.

2.3 Prices in catalogues or newspapers

This source consists of information collected from product catalogues or product lists or from newspapers. These prices may represent, for example, the list prices for tool rental or the averages of transactions for products or services on local or international markets. Access to these catalogues and newspapers may or may not require a paid subscription. For the companies selected for one or more products in a sample, the use of these alternative prices, which are representative of transaction prices, will be a way to reduce their response burden. At the moment, we collect many prices from newspapers such as the price of gold and silver and of several metals. For a defined set of products, if the average transaction price obtained from a market accurately represents all respondents and if the product is homogeneous, it can be used as an approximation of the price of all respondents for this product. Thus, these alternative data save time and reduce respondent burden. In all cases, collection is often done manually and a collection and capture system needs to be put in place to integrate these prices in NSO processing systems (see the solution identified for the Producer Prices Division in Section 3).

2.4 Data files from third parties in the private sector

Private sector companies access a growing quantity of data that they purchase themselves or that they collect. A business case can be made for use of such data when collection costs of gathering a comparable data set for the NSO are higher than the cost of purchasing or acquiring the data. As with the use of information from any data source, the NSO must evaluate whether the data align with the program's classifications and concepts and if they meet the data quality criteria. If the data meet the criteria and it is economically viable to obtain them, it is important to also evaluate the risks.

Relying on a third party for data can add risks of possible supply interruptions if the firm stops the service or goes out of business. When more than one company is offering a similar service, the time to acquire the alternative data could be lengthy. The acquisition could lead to breaks in series when changing from one source to another. The NSO may not have control of costs when relying on market value to acquire data from third parties through a request for services. The Producer Prices Division, for example, had to deal with this type of risk as part of a request for bids for the Computer and Peripherals Price Index and the Commercial Software Price Index.

Ideally, many of the risks identified above can be mitigated within the procurement process. Contracts can be negotiated with options of annual renewal for a number of years. This provides stable costs and data supply for multiple years and minimizes the potential number of alternative solutions to address a break in a series. Finally, an onsite data verification process could also be written into the contract to ensure the quality of the product and obtaining a complete set of information on these data (metadata).

2.5 Administrative data or data from other departments or regulators

Government departments and regulators are an important source of data for NSOs. They can provide detailed information for large populations at minimal cost, allowing statistical offices to develop statistical outputs without additional response burden. Many countries make use of administrative data from corporate taxes for turnover estimates. There are also opportunities to use administrative data for producer price indexes.

When acquiring data, program managers at Statistics Canada follow an agency-wide Directive on Obtaining Administrative Data under the *Statistics Act*. The directive includes requirements to document arrangements and conditions of access to ensure that Statistics Canada respects all legal requirements and maintains public trust by communicating on administrative data use via its website. Statistics Canada favours the use of formal data acquisition agreements as it permits both organizations to fully understand the terms of the agreement.[2]

Some data suppliers may wish to include a specific reference to how their administrative data will be used. Agreements should include conditions around future availability of the data to ensure continuity of the data program should collection of the data for administrative purposes no longer be required by the supplier. The memorandum of understanding between Statistics Canada and another organization specifies explicitly that the data are to be used solely for purposes of the *Statistics Act*. It states that Statistics Canada will use the data to produce aggregate statistics and specifies the name of the index. It also notes that should the organization discontinue collection of the specific return, it commits to work to transfer collection responsibility to Statistics Canada if requested. In addition, the organization in question may notify its respondents of the additional use of their data.

Also included in the directive is the requirement to document the data in an effort to maximize the utility of the administrative data source. When a program area first identifies a potential data source, it checks if the data are already available and in use elsewhere in Statistics Canada. Where a program has identified a new data set that may be beneficial to other programs, there is a consultative process to identify other potential users in the organization and their requirements.

2.6 Data files of other surveys

Finally, some producer price indices can be constructed using existing survey data. Adding questions to or modifying existing surveys presents an efficient and viable alternative to the development of new price surveys. Assessing the feasibility of leveraging other surveys is a natural step in finding a solution to fill data gaps.

At the onset of the development of a producer price index, industries are analyzed to determine their main activities, the homogeneity of activities or of price movements across activities, and the major sources of data used to estimate outputs. Having identified data from another survey as the source, and knowing how the data fit into the output price calculation, allows for the analysis of methods to supplement the data in order to calculate prices. For example, existing data collection could be disaggregated to produce finer levels of detail to reflect the homogeneity of prices and price movements of major activities within an industry. Another possibility, where a survey collects detailed turnover data, is to supplement the information with the number or value of transactions in order to calculate a unit price.

While using data from other surveys is occasionally possible and efficient, various limitations exist. Output data are aggregate rather than transactional. Coupling the information with other data, even if the data are at a detailed level of aggregation, results in unit prices and indexes. Unit value indexes are often criticized for their inability to account for quality adjustments or changes in compositional product mix.

Although these characteristics are less than ideal, collecting information needed to calculate unit prices has been supported given the potential for better response rates, ease of reporting and resource constraints. The requirement for less detail renders reporting less burdensome and more favourable to certain respondents who might not otherwise disclose pricing information. In addition, using existing sources is less costly than producing a new survey using dedicated resources for one purpose. Furthermore, detailed turnover data that capture transaction prices provide a good

[2] Directive on Obtaining Administrative Data under the *Statistics Act,* April 1, 2015 version.

source of weight information at the product level. This disaggregation may help mitigate the problems of unit value bias described above (Diewert, W. E. and Peter Von der Lippe (2010)) .

At the moment, Statistics Canada is experimenting with the development of price indexes using alternative data. For example, data from the Quarterly Telecommunication Survey are used to construct SPPIs for the telecommunications industry. Similarly, the addition of a few more variables to an existing survey, the Quarterly Survey of Trusteed Pension Funds, will likely enable the development of an SPPI for that industry.

3. A solution for the use of diverse data sources

Growth in the use of alternative data compared with data collected regularly or traditionally and the fact that they come from diverse sources make their integration in production systems difficult. Other than data collected regularly, the Producer Prices Division noted over 250 unique sources of data supplying all of its goods, services and construction producer price indexes. At this time, Statistics Canada is working to develop a generic solution to this challenge for all its programs because the Producer Prices Division is not the only division relying on alternative data collection. Until this agency-wide product is available, the Producer Prices Division has developed a solution that is expected to serve as a possible model for a future corporate system. In developing its solution, the Division created a conceptual hierarchy to construct the system's components and ensure a strong connection between the data modules.

The ***data provider*** is the highest level and represents the organization from whom the data are obtained–it is the starting point. For a given data provider, the data can be obtained in a number of ways, for example, from several web pages or in a number of publications. We will call this second level an ***item***. Lastly, for a given item, we can collect a number of prices for different elements of interest that we will call ***products***. Here is an example to better understand this concept:

Data provider: Agriculture and Agri-Food Canada

Item 1: Main web page
Product 1: Price of a dozen eggs, large, Quebec
Product 2: Price of a dozen eggs, large, Ontario
Item 2: Agriculture magazine
Product 1: Price of a cattle, Quebec

Keeping this hierarchy in mind, we developed two data modules and two different data capture methods based on how the data are organized:

Data modules:
- Data Source Register
- Data repository

Capture methods:
- Data capture from non-standard sources
- Custom adaptors to load structured data files

3.1 Data Source Register

The Data Source Register contains metadata on the data source, costs, licences and renewal dates, if applicable. It also contains descriptions and any other relevant data for management of the supply of data. Many of these metadata meet Statistics Canada's requirements to facilitate the documentation of administrative data required in the Directive on Obtaining Administrative Data under the Statistics Act mentioned in Section 2.5.

3.2 Data repository

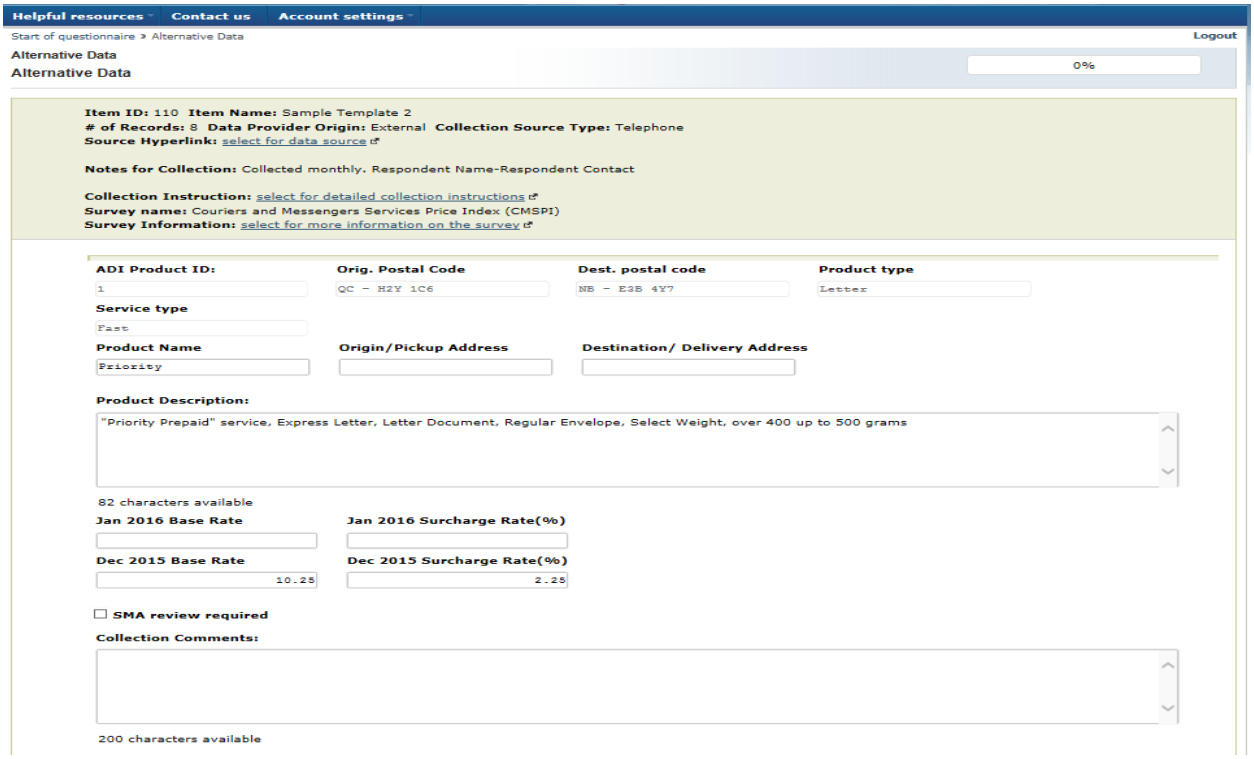
The data repository allows for central storage of all data from alternative data sources to facilitate additional utilization by other programs in the Division.

3.3 Data capture from a non-standard source

The Producer Prices Division has developed two standard models or templates for data collection that use existing corporate collection tools. Statistics Canada has invested in an electronic questionnaire system that can be used for both respondent-completed on-line reporting and computer-assisted telephone interviews. This platform is now used to capture data from unstructured sources such as the Internet or catalogues. This makes it possible to use a set of defined variables in the register and data repository to customize one of the two templates to capture product data including characteristics and prices.

The section at the top of the questionnaire (see Figure 1 below) contains information on the data provider, the survey for which the data will be used and a hyperlink to the web page where the person capturing the data can find more information on the data provider or selected item. The middle section contains a set of variables defining the uniqueness of the product collected and which are used as key variables to link the product to the item that is the target of the collection. Variables are also available to identify characteristics for a given product, along with a text box providing additional information on the product collected. Immediately below this section are the variables identifying the month and the product. This is where the prices are entered and where the prices from previous months can be revised. Furthermore, prices for previous periods are recorded to ensure that the product is properly identified across months. This avoids many capture or identification errors that might have gone undetected in the past when only the price for the current period was requested and recorded (i.e., by seeing that the price for the previous period was $1.20, it is unlikely that the person inputting the data for the current period will enter $12.50 rather than $1.25). Lastly, there is a place to enter comments to give subject matter experts and price experts a further means of verification. It is also possible to check off a box to attract the attention of these same experts. This capture screen can be used by collection specialists or by price experts depending on the complexity of the data capture. The screen is very flexible and can be adapted to capture many products for a single item up to a maximum of 260 price cells. This approach was used to avoid having to design a new capture tool each time.

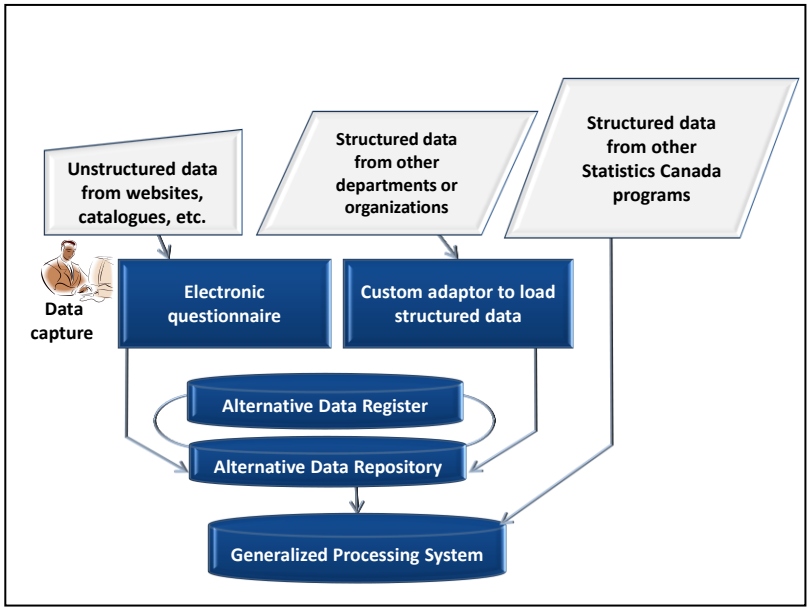
Figure 1: Capture screen template



3.4 Custom adaptors to load structured data files

Where structured data files are available on a recurring basis, custom adaptors are developed to load data directly into the data repository. These adaptors will enable complex calculations and the creation of derived variables required by the processing system. The use of adaptors is a solution for less than 10% of the alternative data sources used for prices.

Figure 2: Operational Model



4. Conclusion

Despite the challenges and complexities of using alternative data sources, the potential for filling statistical gaps while controlling costs and minimizing response burden is undeniable. To help manage the various challenges associated with these sources, Statistics Canada has divided the roles and responsibilities for various aspects across its specialized divisions. Program divisions are accountable for the quality of their programs. It is the role of program areas to identify, evaluate and implement use of alternative data to replace or complement direct collection when such data would reduce response burden and collection costs, fill data gaps, or result in higher quality data than that directly obtained from respondents through surveys and censuses.

The Information Management Division is responsible for supporting information management and access at Statistics Canada and ensuring compliance with applicable legislation, policies and directives. It sees that access, use, storage and management of data by Statistics Canada respect all legal requirements, such as the confidentiality provisions of legislation and the *Statistics Act*, in particular, as well as policies and directives of Statistics Canada and the federal government.

The Administrative Data Division is responsible for the acquisition, common processing, storage and access of administrative data sources that have a broad scope, such as tax data. It is also responsible to develop and implement corporate strategies that facilitate the acquisition, use, management and disposal of administrative data in the agency.

While Statistics Canada has provisions in its *Statistics Act* to obtain access to data from other organizations for statistical purposes, there is no formal requirement for federal government depart

A major challenge for NSOs is the ability to influence the type of data collected, particularly by departments, regulators and the private sector. Building positive relations with the custodians of these data and a history of producing quality relevant indexes from diverse sources is the first step in that direction. With a collaborative working arrangement, the NSO would be in a position to make recommendations to improve the usability of alternative data for statistical purposes.

As Statistics Canada continues the development of goods, services and construction producer price indexes, it will also continue to seek alternative data sources to produce high quality statistics as efficiently as possible.

References

Journal articles

Diewert, W. E. et Peter von der Lippe (2010), "Notes on Unit Value Index Bias", Discussion Paper No. 10-08, Department of Economics, University of British Columbia, Vancouver (Canada). http://econ.sites.olt.ubc.ca/files/2013/06/pdf_paper_erwin-diewert-10-8-notes-unit-value.pdf.

Garneau, Mary Beth (2015), « Use of alternative data sources in Canadian SPPIs », 30th Voorburg Group Meeting, Sydney, Australia, September 21-25, 2015. http://www.voorburggroup.org/Documents/2015%20Sydney/4004.pdf

Statistics Act

Unpublished document

Directive on Obtaining Administrative Data under the Statistics Act, date: April 1, 2015