

Use of alternative data sources in Canadian SPPIs

**30th Voorburg Group Meeting
Sydney Australia
September 21 – 25, 2015**

Mary Beth Garneau
Statistics Canada
Marybeth.Garneau@statcan.gc.ca

The author would like to thank the following people for their contribution to this paper:
Jean Le Moullec (Consumer Prices Division), Gaétan Garneau (Producer Prices Division)
and Lucy Opsitnik (Producer Prices Division)

Use of alternative data sources in Canadian SPPIs

At the 27th Voorburg Group meeting, Statistics Canada contributed a paper, “Use of Administrative Data in Statistics Canada’s Business Surveys – The Way Forward” which articulated how Statistics Canada uses administrative data from the Canada Revenue Agency (value-added tax and income tax data) in its annual and sub-annual turnover programs.

Over the last decade, Canada has been expanding its program of Service Producer Price Indexes (SPPIs). While the majority of SPPIs are based on traditional survey methods, the quest to minimize cost and response burden have led to greater use of alternative sources of data. Alternative data for SPPIs come in many forms such as list prices on the Internet, third party data files, and micro data from turnover or other surveys.

Focussing exclusively on SPPIs, this paper provides further information on the way Canada makes use of alternative data sources.

The first section of the paper summarizes the various sources of alternative data available to National Statistical Offices (NSOs), along with the challenges and potential solutions to make use of these data. Section 2 provides specific examples of how Statistics Canada has made use of alternative data to develop SPPIs in the Transportation, Finance and Professional Services sectors. The third section addresses some of the operational challenges of using a diverse set of alternative data. It describes how Statistics Canada is currently developing tools to integrate a diverse set of data. Finally, Section 4 concludes the paper providing other ideas for consideration.

1. Alternative data sources for SPPIs

Statistics Canada makes use of a wide range of alternative data sources in its price programs. Each source has its own unique set of challenges and risks. This section will look at some of those issues and describe ways to mitigate the specific risks. The four main sources of data used in Canada’s developing SPPI program are:

- capture of prices listed on the Internet
- data files from third parties in the private sector
- data collected by other government departments or regulators
- micro data files from 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 in any municipal office, corporation, business or organization for the purposes of the Act. The Act also requires

that the confidentiality of the data be strictly maintained. As a complement, Statistics Canada also uses information available to the public, under licence or not, with or without a fee.

1.1 List prices on the Internet

List prices captured from the web sites of firms in sample can be a valuable source of data. For example, Statistics Canada collects price information from the company websites of couriers. Individual company rates or pricing schedules are readily available. In some cases, 'rate-finders', or online invoice calculators, are used to obtain price estimates. 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.

At the moment, these data are manually collected from web sites. 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 are impacted by the number of views or the device used to access the web site. Repeated searches on travel sites may be counted as increased demand which could lead to price increases. In addition, some travellers have noted that prices quoted on mobile devices are sometimes higher than prices quoted on a desktop computer.

One potential challenge that Canada has identified in using public sources of data, relates to confidentiality. Data that are protected from disclosure under Canada's *Statistics Act* are either collected through surveys prescribed by the Chief Statistician or, as indicated earlier, from files obtained from other organizations under the *Statistics Act*. To ensure the data used in our statistical program are protected by the *Statistics Act*, the use of public sources of data is also prescribed by the Chief Statistician as a collection activity. Otherwise, this micro-data would be subject to Canada's *Access to Information Act* and the resulting release of this micro-data could lead to a number of undesirable outcomes. First and foremost is the public perception of our commitment to respondents to not disclose their data. Canadians are unlikely to distinguish between micro-data from survey and micro-data from public sources. In addition, once it is known that we are using a specific data source, those data could be subject to manipulation and unlike data collection through a survey, there is no legal recourse under the *Statistics Act* to the falsification of data.

Web data collection could potentially be handled by automated means such as "web-scraping. In fact, not only could web scraping reduce collection costs, it allows the collection of much more data in a timelier manner. Statistics Canada does not use web-scrapers but in exploring the option, has uncovered a few issues.

First, prices on some sites are provided in such a way that makes it difficult or impossible for a web scraper to pick up, at least not with any standard methods known to Statistics Canada. In the web price collection of the Freight Rail Services Price Index, Statistics Canada captures prices that are displayed on images of price sheets. In the web data

collection of Couriers, some sites only display a price when you hover over a specific area of the page with a mouse. In the collection of Couriers data, frequent changes to the web site have also been noted as a potential impediment to web scraping.

Another concern is that some web scrapers do not meet Government of Canada security protocols. For example, one software would communicate back to its developers information about the users (i.e., what data were being scraped).

In addition, the terms and conditions of some web sites prohibit the use of automated web-scrapers. Other sites have terms that limit the data to personal and non-commercial use, which would exclude use by the NSO, with some further sites explicitly prohibiting government use. One reason for firms to prohibit web-scraping is that web-scraping could generate too much traffic and bring down a web site. 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 use intrusive methods that could be perceived as security attacks against websites. The technology has at times been associated with illegal activity and malicious attacks.

To minimize risks associated with web scraping, it is recommended that NSOs first communicate with web site owners before scraping data. The web scrapers used should not be intrusive nor should they create a security risk to the NSO. Finally, scraping should be done in a minimal manner such that the website's server is not overloaded.

1.2 Data files from third parties in the private sector

An increasing amount of data can be accessed from firms in the private sector. A business case can be made for use of such data when collection costs of gathering a comparable data set are higher than the data purchase. 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. Section 2.2 illustrates how such data can be used to calculate an SPPI.

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. Where more than one company provides a similar service, the time it takes to procure the data could be lengthy. Procurement could result in breaks in the series when switching from one source to another. The NSO does not have the same control over costs when relying on market prices for third-party data. Finally, some firms may not want their aggregate data publicly disseminated. The latter is generally not a problem since many firms profit from small-area data while Canada's SPPIs are disseminated at a Canada-level of aggregation.

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 breaks in a series.

1.3 Data from other government 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 SPPIs. Section 2.3 provides one example of how Statistics Canada is using data collected by Canada's central bank to calculate an index for New Lending Services.

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 web site. Statistics Canada favours the use of formal data acquisition agreements as it permits both organizations to fully understand the terms of the agreement.¹

Some data suppliers may wish to include a specific reference to how the administrative data will be used. Agreements should include conditions around future availability of the data to ensure continuity of the data program should the collection of the data for administrative purposes no longer be required by the supplier. For example, the memorandum of understanding between Statistics Canada and the Bank of Canada 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, in particular, the New Lending Services Price Index, a monthly index published quarterly. It also notes that should the Bank of Canada discontinue collection of the specific return, it commits to work to transfer collection responsibility to Statistics Canada if requested. In addition, the Bank of Canada notified its respondents of the additional use of their data.

Also included in the directive are requirements 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, they check if the data are already available and in use elsewhere in Statistics Canada. Where a program has identified a new data set that could be beneficial to other programs, there is a consultative process to identify other potential users in the organization and their requirements.

1.4 Data from other surveys

¹ Directive on Obtaining Administrative Data under the *Statistics Act*, Revision Date: April 1, 2015

Finally, some SPPIs can be constructed using existing survey data. Influencing data collected on other surveys presents an efficient and viable alternative to developing new surveys aimed at directly collecting prices, and determining the possibility is a natural step to finding a solution to fill data gaps.

At the onset of SPPI development, industries are analyzed to determine the main activities of the industry, the homogeneity of activities or of price movements across activities, and the major sources of data used to estimate output. Having identified survey data as the source, and knowing how the data fit into the output 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. Turnover data are aggregate, therefore, coupling the information with other data, even if the data are disaggregated, results in unit value prices and indexes. Unit value indexes are often criticized for their inability to account for quality adjustment or changes in compositional product mix.

Although these characteristics are less than ideal, collecting information needed to calculate unit value prices has been supported given the potential for better response rates, ease of reporting, and given resource constraints. The requirement for less detail renders reporting less burdensome, and more favourable to certain respondents who might not otherwise divulge pricing information. In addition, using existing sources is less costly than producing a new survey using dedicated resources for one purpose. Furthermore, reported detailed turnover data, not often available through survey frames to SPPI programs that collect transaction prices, provide a good source of weight information at the product level. This disaggregation could help to mitigate issues of unit value bias described above.²

At the moment, Statistics Canada is experimenting with the development of price indexes using data from other sources. Section 2.4 summarizes how data from the Quarterly Telecommunication Survey are used to construct SPPIs for Telecommunications. Similarly, Section 2.5 describes how the addition of a few more variables on the Quarterly Survey of Trusteed Pension Funds will likely enable the development of an SPPI for that industry.

² Diewert, W. E. and 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.

2. Examples of alternative data use in Canadian SPPIs

2.1 Freight Rail SPPI – price index using list prices collected from the Internet

This index measures the monthly price movement for the Mainline Freight Rail industry, North American Industry Classification System (NAICS) 482113. This industry comprises establishments primarily engaged in operating railways for the transport of goods over a mainline rail network. A mainline rail network is a system that usually comprises one or more trunk lines, into which a network of branch lines feed.

No questionnaire is used for this price index. Information is collected monthly from public data available on carrier web sites. Prices are defined as the amount charged by a carrier for delivering a certain type of commodity and of a specific weight, on a specific distance (specific origin and destination) and under certain terms of shipment. The prices do not include taxes and other optional services fees.

For every monthly collection cycle, base rates and applicable fuel surcharges (mileage based or percentage based) are also collected for selected origin-destination pairs across Canada and US as well as for predetermined commodity groups. While requested for the calculation of the fuel surcharge based on mileage, the distance between every origin-destination pair is provided by third party software.

These data are used to compute a final price by origin-destination. If applicable, the final price by origin-destination in US dollars is converted to Canadian dollars using the Bank of Canada monthly average exchange rate (noon spot rate). Estimates are produced by calculating a weighted average of price relatives for a fixed sample, then aggregated similarly to other SPPI indexes using a Laspeyres formula.

2.2 Investment Banking SPPI (pilot study) – price index using third party private sector subscription data

Investment Banking Services Price Index (IBSPI) measures price changes for investment banking services, an activity in the Investment Banking and Securities Dealing industry 52311 of the NAICS.³ This industry is comprised of establishments primarily engaged in acting as principals in originating, underwriting and/or distributing securities of businesses, governments and institutions.

The financial industry is highly regulated in Canada, and financial institutions including investment banks are required to report extensively on their activities. In particular, new issues of debt and equity, as well as mergers and acquisitions are reported through the

³ The NAICS industry partially concurs with ISIC K6499, Other financial service activities, except insurance and pension funding activities, n.e.c.

System for Electronic Document Analysis and Retrieval (SEDAR). Administered by the Canadian Securities Administrator, SEDAR is a mandatory document filing and retrieval system for Canadian public companies. For new issues, some private data providers collect transaction information with commissions through administrative data sources, tabulate them into a database, and make them available to the public on a subscription basis. With data obtained from one such third party, Statistics Canada was able to produce an experimental index measuring price changes for investment banking services.

The experimental IBSPI is based on a census of transactions featuring all new issues of financial securities from 2004-2013. The data allows product definition by type (corporate debt, ownership, preferred shares, etc.), term, and deal size, where the service price is calculated as a unit value of commissions (i.e., total commission income divided by total proceeds).

2.3 New Lending SPPI- price index using administrative data from Canada's central bank

The New Lending Services Price Index (NLSPI) is Statistics Canada's first attempt to price Financial Intermediation Services Indirectly Measured (FISIM). While NLSPI does not serve as an overall deflator of total FISIM, it represents the price movements of new loans (and refinanced loans) including new residential mortgages. Over time, the index could serve as a component of a broader index.

The NLSPI measures monthly price changes over time for new lending services in Canada; the estimates are produced on a quarterly basis. Prices are derived as the difference between annual percentage rates for new loan products and weighted averages of yields on financial market instruments. The variables used to derive the prices are weighted annual percentage rates for new lending services, funds advanced by product, and market rates. Since the value of money erodes over time, the calculated spread is deflated using the implicit price index for Final Domestic Expenditure.

The principal data for the NLSPI are obtained from the Report on New Lending, collected monthly by the Bank of Canada through a survey of all Canadian chartered banks. Each bank is required to provide data on interest rates and funds advanced for 10 lending products by 6 interest rate term maturities as well as for the aggregate of all products and all maturities.

In order to calculate prices for each lending product, a reference rate is deducted from each product's lending rate (by bank and maturity). The reference rate is derived from data on Financial Market Statistics compiled by the Bank of Canada and published in Statistics Canada's dissemination database, CANSIM. Certain market instruments were chosen and their yields were aggregated in order to produce the reference rate. The deflator is obtained from the Canadian System of National Accounts' Gross Domestic Product by Income and Expenditure. As all data come from already existing sources,

Statistics Canada has been able to produce this product with no additional burden.

2.4 Telecommunications SPPI (pilot study) – price index from existing turnover survey

Statistics Canada has been piloting a quarterly Telecommunications Services Producer Price Index (TSPI) that covers telecommunications services under NAICS 517-Telecommunications. Statistics Canada collects data for this industry under a partnership with The Canadian Radio-Television and Telecommunications Commission (CRTC). The quarterly telecommunications survey collects revenues and corresponding quantities, such as the number of lines, or the number of minutes, by the following service categories: Local & Access (LAC), Wireline Long Distance, Mobile Wireless, Internet, Data & Private Line, and Broadcast Distribution Undertaking. Retail versus wholesale and residential versus business components are also identified, where applicable.

The TSPI is produced by aggregating unit value prices from a fixed basket of telecommunications services using the Laspeyres formula, and will update weights every two years.

2.5 Trusteed Pension Funds SPPI (pilot study) – price index by adding additional questions to an existing survey

Statistics Canada currently collects data on the revenues, expenses, liabilities, and assets of trusteed pension funds in its Quarterly Survey of Trusteed Pension Funds. The survey had been identified as a potential source of data for an index measuring the price of the services of NAICS 526111, Trusteed pension funds, given its close conceptual relationship to output. However, the variable of interest to the price index program suffers from heterogeneity given the underlying inclusions and potentially variable weights of the components across the different funds. As a result, a program was launched to assess the viability of calculating an index using data at a more detailed level.

Under a pilot survey, the questionnaire has been modified to break out data reported on administration costs into three components: administrative operating costs, investment management costs, and investment management costs related to a fund's performance. Using these data Statistics Canada has the extra level of detail needed to experiment with calculating a price for each of the following components:

- i. Administrative operating costs are taken as reported into the calculation of the price of the service
- ii. Investment management costs are included per dollar of assets under management (less the assets on which performance related payments are calculated)

- iii. Investment management costs relative to fund performance are included per dollar of assets on which performance related payments are calculated

Price relatives are calculated for each component and price updated weights are applied to the different components using revenue share in the base year. Total administration costs reported (the sum of the three components) for each fund in the base year is taken as the weight applied to each fund in order to generate the overall industry level index.

3. Operational challenges in using diverse sources of alternative data

With growing use of data from a variety of alternative sources, it is a challenge to integrate the diverse sources of data into production systems. Producer Prices Division noted over 100 unique sources of data supplying all of its goods, services and construction producer price indexes. Statistics Canada is currently developing a set of solutions to this challenge for producer price programs (see Figure 1).

The solution is comprised of four key elements:

- Alternative Data Register
- Alternative Data Repository
- Data capture from non-standard sources
- Custom adaptors to load data files

Alternative Data Register

The data register will contain metadata regarding the data source, cost and licences if applicable, renewal dates, descriptions and any other relevant data for the management of the supply of data. Many of the metadata requirements are corporate requirements to facilitate the documentation of administrative data required in the “Directive on Obtaining Administrative Data under the *Statistics Act*” noted in Section 1.3 above.

Alternative Data Repository

The Alternative Data Repository allows for central storage of all the alternative data enabling additional use by other programs in the division or the department.

Data capture from non-standard source

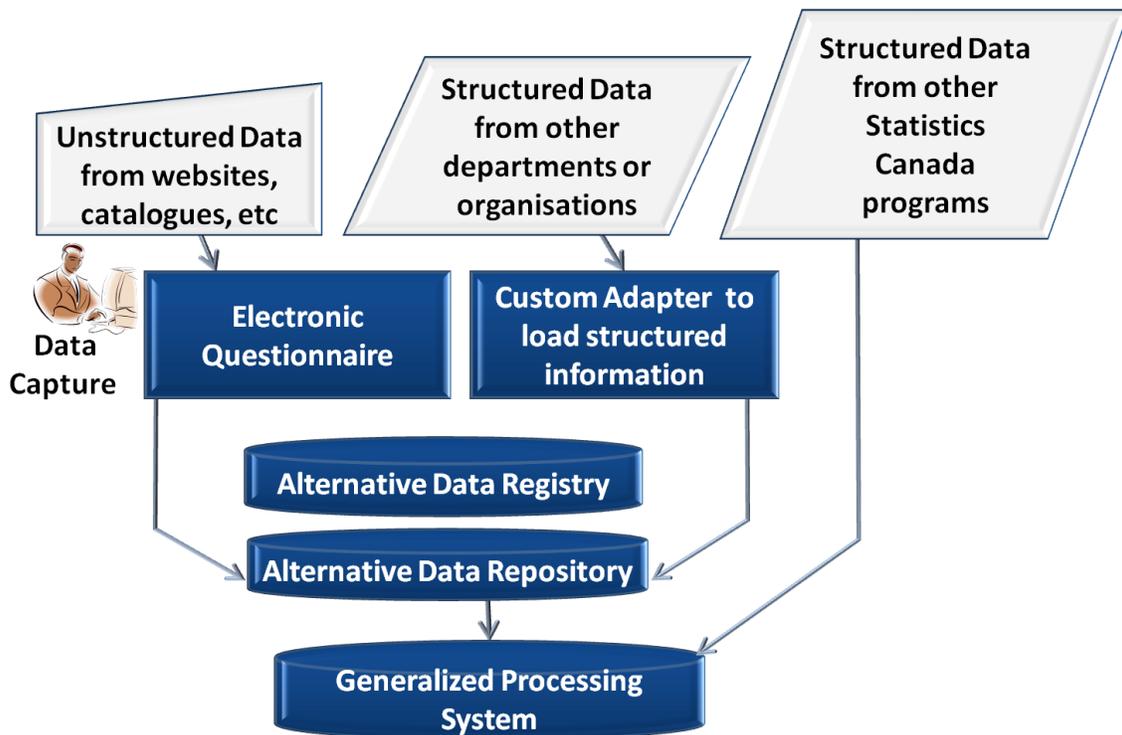
Producer Prices Division is developing standard templates with which to collect data using corporate collection tools. Statistics Canada has invested in an electronic questionnaire system that can be used for both respondent completed on-line reporting or for Computer-Assisted Telephone Interviews. The same platform will be leveraged to capture data from any unstructured source such as the Internet or catalogues.

Custom adaptors to load data files

Where structured data files are available on a recurring basis, custom adaptors are being developed to load data directly into the data repository. These adaptors allow for the application of complex calculations and the creation of derived variables that are required

by the processing system. The use of loaders is an exception, and account for less than 10% of the alternative data sources used in prices.

Figure 1: Operational Model



4. Conclusion

Despite the challenges and complexities of using alternative data sources, the potential for filling data gaps while controlling costs and minimizing response burden is undeniable. To help manage the various challenges associated with alternative sources of information, Statistics Canada has divided roles and responsibilities of various aspects across specialized divisions in the department. 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.

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

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. They are also responsible to develop and implement corporate strategies that facilitate the acquisition, use, management and disposal of administrative data in the department.

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 departments, or any other body, to consult with Statistics Canada when they create or redevelop administrative systems.

An important challenge to NSOs is the ability to influence what data are gathered, especially by government departments, regulators and the private sector. Establishing positive relationships with custodians of these data and a track record of producing relevant quality indexes from various alternative 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 development of Services Producer Price Indexes, the department will continue to seek new sources of alternative data to produce high quality statistics as efficiently as possible.

References

Diewert, W. E. and 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.

Directive on Obtaining Administrative Data under the *Statistics Act*, Revision Date: April 1, 2015

Nael Hajjar, Lucy Opsitnik and Min Xie (2014), Investment Banking Services Price Index: An experimental approach, Producer Prices Division, Statistics Canada (Internal Document)

New Lending Services Price Index (NLSPI), Statistics Canada:
<http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5207>

Quarterly Survey of Telecommunications, Statistics Canada:
<http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=2721>

Yung, Wesley and Lys, Peter: “Use of Administrative Data in Statistics Canada’s Business Surveys – The Way Forward”, Statistics Canada, October 2012
http://voorburg/Documents/2012%20Warsaw/Papers/4016%20-%20use_of_admin_data_VG_2012_Canada.pdf

Appendix

Current use of alternative data sources in Statistics Canada's SPPIs

Index Title (Status)	Key Variables	Price Method	Data Source(s) ¹
Couriers and Messengers (Production)	Prices (Base rate excluding taxes and Fuel charge)	List price	Internet prices for couriers sample Couriers and Messengers Services Price Index for messengers sample
Freight Rail (Production)	Prices (Base rate excluding taxes and Fuel charge) Distance Canada-US exchange rate Weights	List price	Internet data from carrier websites Third party data on distances Bank of Canada Railway - Annual Report, a Statistic Canada survey conducted by Transportation Division used for weights
Informatics Professionals (Production)	Number of employees and salary and wage expenses Operating revenue, other expenses, total revenue	Price index is the result of applying a revenue to wage expense net multiplier to contract fees and wage rates	Informatics Professional Services Price Index Payroll Deduction Remittances Corporation Income Tax Returns
Investment Banking (Pilot)	Total commissions per deal Total proceeds per deal Implicit price index (for deflation)	Unit value of total commission income divided by total proceeds (which are deflated to account for money erosion)	Third party data on new issues Gross fixed capital formation; existing Statistic Canada program used for deflation
New Lending (Production)	Annual percentage rates for loans by type and maturity Market rates for select securities Implicit price index (for escalation)	Price defined as spread between rates for new loan and the reference rate, escalated by GDI	Bank of Canada (Report on New Lending <u>and</u> Financial Market Statistics) Gross Domestic Product by Income and Expenditure; existing Statistic Canada program used for deflation

Index Title (Status)	Key Variables	Price Method	Data Source(s) ¹
Passenger Air (Production)	Fares excluding taxes	Unit value	Survey conducted by Statistics Canada's Transportation Division furnishes all the data to this program
Passenger Rail Services Price Index (Pilot)	Prices are base rate excluding taxes, by class, passenger type and ticket group Weights	List price	Internet data from carrier website Weights are obtained from an administrative file obtained by Statistics Canada's Transportation Division, from Transport Canada
Pension Funds (Development)	Administrative cost, investment cost, performance related cost, and assets	Cost per assets for administrative and investment purpose, escalated by an appropriate factor (TBD)	Quarterly Survey of Trusteed Pension Funds; existing Statistic Canada survey modified to provide more detail in one previously reported variable Existing Statistic Canada program to be used for deflation
Property & Casualty Insurance (Development)	Premiums, Investment income, net premiums earned, and assets (investment)	Model price	Third party private source of data (two sources for different variables), and Canadian regulatory data
Securities Dealing (Development)	Price quotes for certain Canadian government securities and associated trading volume	Spread between bid and ask quotes taken as price	Third party private source of data Canada's international transactions in securities, an existing Statistic Canada survey
Telecommunications (Pilot)	Revenues and quantities for service categories; Retail customers vs. Wholesale; and Residential vs. Business	Unit value from a fixed basket of telecommunication services	Quarterly Telecommunication Survey; an existing Statistic Canada survey conducted in partnership with the Canadian regulator (Canadian Radio-Television and Telecommunications Commission)
Traveller Accommodation (Production)	Prices (Base rate excluding taxes)	List price	Consumer Price Division (internet collection)

Note: ¹ Services Producer Price Index surveys are shown in bold in the Data Source(s) column.