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

Challenges Associated with Using Scanner Data for the Consumer Price Index

Catherine Deshaies-Moreault and Nelson Emond[1]

Abstract

Practically all major retailers use scanners to record the information on their transactions with clients (consumers). These data normally include the product code, a brief description, the price and the quantity sold. This is an extremely relevant data source for statistical programs such as Statistics Canada's Consumer Price Index (CPI), one of Canada's most important economic indicators. Using scanner data could improve the quality of the CPI by increasing the number of prices used in calculations, expanding geographic coverage and including the quantities sold, among other things, while lowering data collection costs. However, using these data presents many challenges. An examination of scanner data from a first retailer revealed a high rate of change in product identification codes over a one-year period. The effects of these changes pose challenges from a product classification and estimate quality perspective. This article focuses on the issues associated with acquiring, classifying and examining these data to assess their quality for use in the CPI.

Key words: Consumer Price Index, scanner data, classification.

1. Introduction

Thanks to technological evolution, a variety of data sources containing a wealth of information have become accessible in the past few decades. Statistical agencies, interested in taking advantage of these alternative data sources, are innovating in order to continue producing high-quality statistics without unduly increasing costs. One of these data sources, already being used in a few countries, is scanners. Scanner data are electronic records of transactions collected by reading a product bar code. They contain information such as the total revenue from sales and the quantities sold on the products in the file, identified by the Universal Product Code (UPC). This type of data source could be beneficial to a number of statistical programs, especially the Consumer Price Index (CPI). The CPI consists of classes of mutually exclusive elementary aggregates that comprise different products with homogenous price changes. The challenge will be to link each product from the scanner data to a class at the elementary aggregate level. Sections 2 and 3 contain an overview of the CPI and of scanner data, respectively. A few key concepts to be considered when introducing scanner data to the CPI are reviewed in Section 4. Section 5 describes a five-phase project developed by Statistics Canada to analyze and include scanner data in the monthly CPI. Lastly, Section 6 contains a brief conclusion.

2. Consumer Price Index

The CPI is an indicator of the changes in consumer prices experienced by households. Canada's CPI measures price changes by comparing the cost of a fixed basket of goods and services over time. The Canada All-items CPI is based on an annual sample of more than 950,000 price quotes. Some prices are collected by telephone interview, from alternative sources (such as the Internet), by questionnaire or from other Statistics Canada programs. However, since 55% of the basket is based on prices collected in person, significant resources are allocated to data collection.

Sampling for the CPI is multi-level. First, certain geographic areas with significant economic activity are selected. Points of sale or outlets are then sampled within these areas. To obtain a price sample by product category, the outlets

[1] Catherine Deshaies-Moreault, Statistics Canada, 100 Tunney's Pasture Driveway, Canada, K1A 0T6 (catherine.deshaies-moreault@canada.ca); Nelson Emond, Statistics Canada, 100 Tunney's Pasture Driveway, Canada, K1A 0T6 (nelson.emond@canada.ca).

are chosen based on the type of products to be collected. This ensures a representative sample of not only the products and services bought by households, but also of the location where they are purchased. Since not all products can be selected, a product offer is chosen from a product definition: the representative product. This representative product is developed to represent the price change of several similar products. The price quote for a product selected is the posted price. For this reason, there is no guarantee that there were, in fact, transactions for this product in the collection month.

The CPI's aggregation structure consists of eight main components divided into various intermediate aggregates and then elementary aggregates. The elementary aggregates level represents the lowest-level class for which a fixed estimate of quantities is available. For more information, the reader may consult the Canadian Consumer Price Index Reference Paper (Statistics Canada 2015).

3. Scanner data

3.1 Information available

As mentioned earlier, scanner data are electronic records of transactions collected by reading product bar codes. A considerable amount of information is maintained and stored by retailers in scanner files. Some of the variables typically available are store identifier (including address information), week end-date, store classification variables, UPC, UPC description, sales revenue and quantities sold.

It should be noted that the week end-date indicates that the file sent does not contain all individual transactions, but rather a summary of all sales for each UPC, by week. If a UPC was not sold in a given week, then there would be no record for that UPC in that week.

3.2 Interest in scanner data

It might be possible to enhance various aspects of the quality of the CPI by using scanner data, notably via sample size, representativeness, geographic coverage and use of unit values built from transactions and not posted prices. As mentioned in Section 2, the current CPI collection method involves selecting products and outlets in certain specific geographic areas. With scanner data, the product universe of a single store will potentially be available along with the approximate set of stores. Stores using a scanner system and that provide their data to Statistics Canada will represent the approximate store universe. Thus, the size of the sample will definitely be greater than that able to be collected in person, while reducing collection costs considerably through the discontinuation of in-person collection.

There would also be a gain in terms of representativeness since data for potentially all products sold in a particular store could be obtained. Rather than selecting one or a few product offers per elementary aggregate, several or all products could be used. Furthermore, with data on quantities sold, products could be weighted by quantity sold so that those purchased most by consumers contribute appropriately to the index.

There is also no physical constraint on collection with scanner data, which makes it possible to increase geographic coverage. With scanner data partially replacing in-person collection, non-sampling errors, such as capture errors, will be greatly reduced. However, it would be unrealistic to assume there would be no errors in the scanner files.

Finally, other Statistics Canada programs might benefit from using scanner data, such as the Survey of Household Spending, the Survey of Retail Trade and the System of National Accounts.

4. Review of concepts

Certain key elements need to be compared to determine if the format of scanner data meets the key concepts of the Canadian CPI, notably in terms of the prices and products collected, the fixed basket and the time to obtain the data.

The current CPI collection involves collecting the prices of certain representative products. Interviewers visit outlets to note the prices on specific products on a given day. In short, it is a point-in-time price captured at the same time each month, for example, on Tuesday of week 2. With scanner data, the prices received are calculated from the week's sales revenue and the quantities sold–they are weekly unit prices. This is an important difference in the price universe and numerous formulas have been developed to make use of unit prices. This concept is not the main focus of this article, but the reader may consult Richardson (2000), among others, on this topic. Furthermore, prices for all UPCs with non-zero sales are received, thus eliminating the constraint to rely on a list of representative products. However, the receipt of weekly data creates an issue with respect to UPC aggregation. Section 5.2.4 examines a number of scenarios for product identification using UPCs. There is also an issue with respect to zero sales–are they temporary, thus justifying a missing value adjustment, or are they permanent, warranting product replacement and subsequent adjustments (so-called "quality adjustments"; see Statistics Canada 2015)?

The implicit volumes of spending in the elementary aggregates are set in the current definition of the Canadian CPI. Given that the intent is to classify UPCs under the elementary aggregates, the fixed-basket concept will be respected. Long term, with detailed information on spending available from scanner data, the weights (based on quantity) used to combine different types of products may be reviewed or the basket itself restructured.

The data will be provided weekly. This means that the timelines for the monthly publication can be met. No decision has been made on whether the data from an entire month will be used to compile the CPI or only the first two or three weeks due to constraints related to the time for processing and analysis.

5. Five-phase project

A multidisciplinary team at Statistics Canada developed a five-phase project to include scanner data in the CPI. The following subsections cover these phases and the various issues encountered.

5.1 Phase 1 – Exploratory research

Other statistical agencies are already using scanner data in their national CPI. The Netherlands in particular (Schut 2001; van der Grient and de Haan 2010) have already published numerous articles on this topic. The strategy for using scanner data, the source of these data or the collection method varies from agency to agency. Some agencies purchase scanner data processed by a third party, or receive only a data sample specified to retailers. A few agencies use the store classification variables rather than aligning the UPCs with their traditional CPI classification. The Canadian context is different in many respects: the products in the sample are confidential (so Statistics Canada cannot provide a list of desired products to retailers) and Canada is a very large country geographically, naturally leading to considerable variety in retailers–using the store classification rather than the current CPI classification might get out of hand.

5.2 Phase 2 – Acquisition, analysis, feasibility study and pilot studies

An initial agreement was reached with a major retailer in the Canadian food industry to receive scanner data from a sample of six stores over two years. The analyses and feasibility studies described below are based on these data.

5.2.1 Data acquisition issue

One issue related to scanner data is the acquisition of data. Statistics Canada has the authority to collect data (Statistics Canada 2016) under the *Statistics Act* but also has a responsibility to consider the burden on respondents. The organization strives to maintain positive relations with its respondents and wanted to establish a framework conducive to productive collaboration right from the start of the scanner data integration project. The next step was to identify what data the agency wanted to acquire. While seemingly simple at first glance, this question becomes complex given the potential diversity of storage standards for such data in the industry (where each respondent might use its own data dictionary). The experience of other countries was a good starting point to identify which variables are normally present in a scanner data file and relevant to the needs of a program such as the CPI. Furthermore, the large volume of such files created technical issues related to data transmission, storage and processing.

5.2.2 Data examination issues

The files received are quite substantial, some 9 million records in a single year, including 190,000 unique UPCs. On average, there are 85,000 unique UPCs each month. The distribution of the UPCs is very asymmetrical. In effect, only 3.5% of UPCs were observed each month in the six stores and alone they account for 41% of total sales. This means that a great many UPCs could be removed from the files with minimal impact on total sales. However, care must be taken in reducing the size of the file. Certain UPCs might have only a slight impact on total sales, but be important within their product category. This issue led to a "chicken or egg" type discussion in the sense that the size of the file needs to be reduced to facilitate the classification work but the UPCs must be classified to determine which ones can be eliminated without negatively impacting CPI quality. UPCs also have a high turnover rate with only 33% of UPCs being present in two consecutive years in a given store. This suggests that the classification process will have to be ongoing.

5.2.3 Data classification issue

To be usable in building a CPI, scanner data must be classified at the finest level of the aggregation structure, which is the elementary aggregate in this case. Two types of variables found in the scanner data can be useful for classification purposes, specifically, the store classification variables and the UPC description. This issue can be addressed using techniques to process character strings. It would be interesting to use a generalized system to perform the classification task. Statistics Canada has such systems for data matching and coding. That being said, classification is a different issue, requiring an alternative tool. Some classification systems exist, but are more suited to processing stable classifications and are not revised on a continuous basis. The idea of a dictionary seems quite promising but would require considerable maintenance efforts. Thus, the process needs to remain generic enough to easily adapt to changes. It is not that simple to develop a generic solution suited to all types of products. For example, using product brands to help with classification seems like a good idea at first glance. While this would work with certain brands associated with very specific products, the idea is a non-starter for brands associated with a wide range of different products classified in different locations in the CPI structure.

The classification process currently being developed will use the retailer's classification variables and UPC descriptions to direct products to the appropriate CPI elementary aggregate. In addition, given that the links between the scanner data and CPI aggregates are word-based, several words could be linked to an item. A weighting based on the frequency of word appearance will be applied to prioritize linked words–the less frequent a word, the greater its weight if it is linked. This approach will reduce the impact of words frequently found in the vocabulary such as adjectives, flavours ("raspberry", "orange", etc.) or condition ("fresh", "frozen", etc.). More complex methods, such as classification by machine learning, were considered briefly and will be examined more thoroughly at a later date.

5.2.4 Index simulations

As mentioned earlier, scanner data are received weekly. To obtain a monthly price, the weekly prices have to be aggregated in some way. In addition, since the concept of product is different for scanner data, it raises the question of what is the new definition of product: is it a specific UPC or a UPC grouping? A number of scenarios for how to define the product and how to aggregate the weekly data are presented below. However, we start with the Jevons formula, (formula 5.1) which is used to calculate the price index for the majority of elementary aggregates.

Formula 5.1
Jevons index

$$I_j^{0:t} = \prod_{i=1}^{n}\left(\frac{p_i^t}{p_i^0}\right)^{1/n}$$

Where:
$I_j^{0:t}$ is the index of elementary aggregate *j* between the periods *0* and *t*;
p_i^t is the price of product *i* at time *t*;
p_i^0 is the price of product *i* at time *0*;
n is the number of products in aggregate *j*.

Table 5.1 presents the different scenarios proposed for the definition of product and how to aggregate weekly prices. The definitions of p_i and *n* refer to the Jevons formula defined earlier. The objective is to obtain a monthly p_i; thus, aggregation is always by week as well as by UPC, if indicated by the product definition.

Table 5.1
Proposed scenarios

	Formula	Definitions	Product
1	$p_i^t = \sum_{y=1}^{r}\sum_{k=1}^{m_y} unit\ price_y^{k,t} \,/ \sum_y m_y$	*r*: number of UPCs in month *t* m_y: number of weeks in month *t* with sales for UPC *y* *n* = 1	All similar UPCs represent the same product (*n*=1)
2	$p_i^t = \frac{\sum_{k=1}^{m} unit\ price_i^{k,t} * \pi_i^k}{\sum_k \pi_i^k}$	*m*: number of weeks in month *t* with sales for UPC *i* *n*: number of UPCs in month *t* π_i^k: quantities sold for UPC in week *k*	Each UPC is a different product
3	$p_i^t = \prod_{k=1}^{m}\left(unit\ price_i^{k,t}\right)^{1/m}$	*m*: number of weeks in month *t* with sales for UPC *i* *n*: number of UPCs in month *t*	Each UPC is a different product
4	$p_i^t = unit\ price_i^{k,t}$	*n* = 1	The unit price of a single given week for a single selected UPC is the representative product

The different scenarios will be explained using the fictitious example of an extract from a scanner data file contained in Table 5.2. Certain variables typically found in scanner files are shown in the table, including unit price, which is calculated by dividing sales revenue by the quantities sold.

Table 5.2
Example of scanner data

Week end-date	Store classification	UPC	UPC description	Sales revenue	Quantities sold	Unit price
6FEV15	CANNED SOUP	123	CRM OF BROCCOLI	$12	12	$1.00
13FEV15	CANNED SOUP	123	CRM OF BROCCOLI	$10	10	$1.00
20FEV15	CANNED SOUP	123	CRM OF BROCCOLI	$11.25	15	$0.75
27FEV15	CANNED SOUP	123	CRM OF BROCCOLI	$10	8	$1.25
6FEV15	CANNED SOUP	456	CHICKEN BROTH	$9	6	$1.50
13FEV15	CANNED SOUP	456	CHICKEN BROTH	$9.80	20	$0.49
27FEV15	CANNED SOUP	456	CHICKEN BROTH	$9.95	5	$1.99

The scenarios for the product concept and for weekly price aggregation proposed in Table 5.1 are described in greater details below.

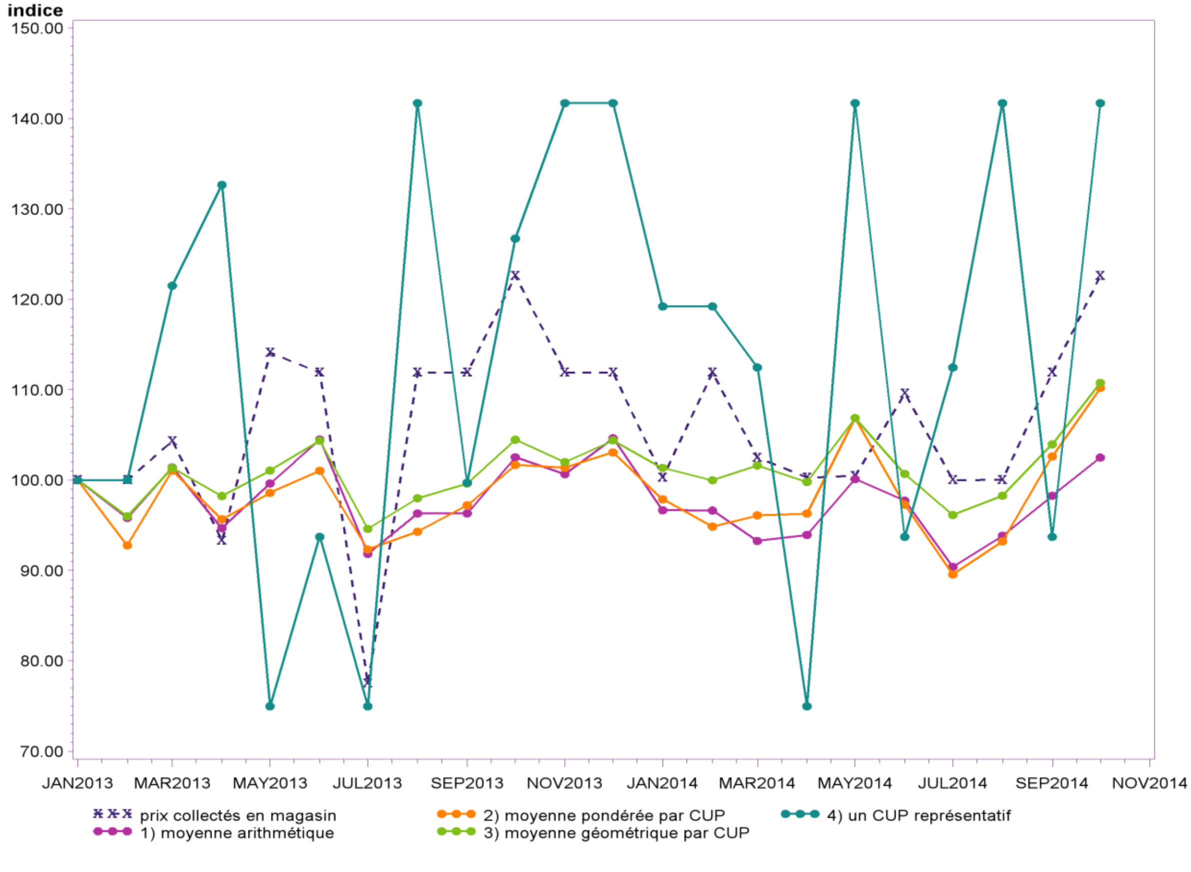
Scenario 1 considers the product to be a grouping of similar UPCs. Referring to the example in Table 5.2, *canned soups* would be the product in this case. The monthly price of the product is obtained by calculating the average of all unit prices.

In scenarios 2 and 3, each UPC is a different product. In Table 5.2, this means that there would be two distinct products, UPC 123 and UPC 456. The monthly price for each UPC in scenario 2 is obtained by calculating a weighted average on the quantities sold to obtain a price representative of buyer behaviour. The monthly price for each UPC in scenario 3 is calculated using a geometric average of the unit prices in keeping with the spirit of the Jevons formula.

Finally, in scenario 4, the unit price of a single UPC in a given week is selected to represent the price for the month of a group of similar products. In Table 5.2, this might be, for example, the unit price of UPC 456 from February 13, 2015, which would represent the price for the month of February for *canned soups*. This scenario is close to the current field collection approach where the price of a given representative product is collected.

Chart 5.1 contains the comparison of the different scanner data scenarios described above to the current index of CPI data collected in person. The data correspond to a given product in a specific store. For scenario 4, the UPC with the most sales in the first week is tracked month to month (always retaining the price from the first week of the month to be consistent with the frequency of CPI collection for food products).

Chart 5.1
Jevon indexes for different scenarios



As noted in the above figure, indexes using a single product (representative product in store and scenario 4) are much more volatile than those calculated under the other scenarios. It should be pointed out that the current CPI is not as volatile as the example illustrated above because it combines a number of different products collected from multiple points of sale.

5.3 Implementation and systems development

Scanner data will be introduced gradually to the calculation of the CPI at different stages of implementation. This section describes phases 3, 4 and 5 covering implementation and systems development.

5.3.1 Phase 3 – Acquisition of production data and simple, systematic implementation

Phase 3 of the project involves acquiring production data, developing the necessary systems and integrating the scanner data. In the short term, scanner data from an initial retailer could be integrated in the CPI calculation once they are available on a regular basis. The scenario envisioned is to replace each price collected in the stores of this retailer as part of the current sample with scanner data. This initial introduction will have the immediate benefit of reducing collection costs and will not require any changes to existing system (or very few). However, this approach will not increase the size of the sample and will not use the quantities sold to weight the different products. Once Phase 3 has been completed, the goal will be to use scanner data to increase the size of the CPI sample and then to introduce an implementation method that takes advantage of the information available on quantities sold while remaining compatible with the concept of a “fixed basket”.

5.3.2 Phase 4 – Expanded simple, systematic implementation

Phase 4 will expand the method developed in Phase 3 to more retailers across various industries (not only the food industry). However, the data will still not be used to their full potential–that will be the focus of Phase 5.

5.3.3 Phase 5 – Advanced implementation

The index formulas for the Canadian CPI and concepts will have to re-evaluated in order to use scanner data to their full potential. Since information on quantities sold is available, scanner data open the door to adopting formulas using more information, such as “superlative” index formulas that are considered better than those normally used with the concept of a “fixed basket”. At present, a number of studies of scanner data and the development of superior formulas are under way. The objective of Phase 5 will be to analyze these new formulas and implement them.

6. Conclusion

Scanner data offer a number of elements of interest to the CPI program. In particular, they will improve the quality and representativeness of the Index and help reduce collection costs. Using scanner data will also reduce non-sampling errors. In contrast, they come with their own challenges. Once data acquisition and examination issues have been addressed, the next challenge to overcome will be to classify the data at a satisfactory success rate to be usable in the CPI. Machine learning techniques have been considered and are being examined in order to complete this aspect of the work. Furthermore, it will be necessary to address appropriately the new data format, which presents weekly unit prices and offers a number of possible scenarios for product definition and data aggregation. Production data must be received on a regular basis before launching the simple, systematic implementation. A contingency plan will have to be developed in parallel to react quickly should there be a break in data transmission. Finally, the acquisition of data from other retailers is planned to expand market coverage and maximize use of scanner data. In the long term, the idea is to move the CPI to superior formulas that make full use of the potential of this rich data source.

References

Richardson, David H. (2000), « Scanner Indexes for the CPI », paper presented at *Sixth Meeting of the International Working Group on Price Indices*, Australie.

Schut, Cecile. (2001), « Using scanner data to compile price indices: practical problems », article présenté au *Joint Statistical Commission and Economic Commission for Europe / International Labour Organisation (ECE/ILO) Meeting on Consummer Prices*, Genève.

Statistics Canada (2015). *The Canadian Consumer Price Index Reference Paper, 2015, 62-553-X, 91 p.* http://www.statcan.gc.ca/pub/62-553-x/62-553-x2015001-eng.htm

Statistics Canada (2016). *The Companion guide to the* Statistics Act, http://icn-rci.statcan.ca/10/10b/10b_001-eng.html

van der Grient, Heymerick et de Haan, Jan. (2010), « The use of supermarket scanner data in the Dutch CPI », paper presented at *Joint ECE/ILO Workshop on Scanner Data*, Genève.