Proceedings of Statistics Canada Symposium 2018
Combining to Conquer: Innovation in the Use of Multiple Sources of Data

Modernizing the Household Expenditure Program

Christiane Laperrière, Denis Malo and Johanne Tremblay[1]

Abstract

The Survey of Household Spending gathers information that is important for updating the Consumer Price Index basket and for Canada's System of National Accounts. These data are also used by a large user community that is generally interested in analyzing expenditures based on the socioeconomic characteristics of households. The detailed content and the traditional data-collection practices based on personal interviews and an expenditure diary place a heavy burden on respondents and result in high collection costs. As a result, the Household Expenditure Program has been exploring the potential for new data sources, in line with Statistics Canada's modernization objectives. In this article, a number of alternative data sources that are under study will be presented, and their potential and limitations will be discussed within the context of the program. The article will also describe the challenges encountered in exploring these new data sources, the innovative ideas envisaged for classifying and integrating them, and the results of the ongoing evaluations.

Key words: Modernization; Household Expenditure Program; Alterative Sources; Data Integration.

1. Introduction

In light of the challenges they face, household expenditure survey programs throughout the world are seeking ways to modernize (Eurostat, 2017). The Canadian context is no different. Even though the Canadian program has adopted the international collection model, it wants to position itself to better react to current and future data collection challenges. Therefore, searching for and integrating alternative sources of data while protecting personal information is natural for this program. This article describes the studies on the search for and use of new sources that are currently underway. A description of the existing Household Expenditure Program (HEP) is provided in Section 2, and the objectives of the modernization project are outlined in Section 3. Sections 4 and 5 describe data source issues related to housing expenditures and transactional data. Section 6 concludes with a summary of the issues encountered and a description of future projects.

2. Household Expenditure Program

The current HEP is based primarily on the Survey of Household Spending (SHS), which is a voluntary annual survey with a sample of approximately 17,500 households in the 10 provinces (Statistics Canada, 2018). The SHS combines two collection methods: an interview and an expenditure diary. The computer-assisted personal interview is used primarily to collect information on larger, less frequent expenditures. The reference periods are based on the type of expense (one month, three months, last payment, four weeks). The paper diary is used to collect information on smaller, more frequent expenses over a two-week period. These expenses would be more difficult to recall in a retrospective interview. The data on household income are taken from administrative sources and are added to the data obtained from the respondents during the collection process. There is a heavy response burden because, in

[1] Christiane Laperrière, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (christiane.laperriere@canada.ca); Denis Malo, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (denis.malo@canada.ca); Johanne Tremblay Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (johanne.tremblay@canada.ca)

addition to having to complete the diary, the average interview length is approximately 60 minutes. Interview response rates are around 65%, and diary response rates are between 40% and 45%.

Data from the SHS are used to update the weights of the Consumer Price Index basket of goods and services. They are also used by the System of National Accounts, in particular as inputs to derive the Gross Domestic Product, and by various federal and provincial departments in developing social and economic policies and programs. Lastly, various groups wanting to better understand the issues related to the spending habits of Canadians use data obtained from the SHS. So there is a vast array of users with varying needs, and this is an important element that needs to be considered when a modernization plan is being formulated. There is also a high demand for expenditure data with a higher analytical value. This translates into requests for the addition of content to the survey and an increased capacity to conduct analyses for specific populations. Under the current model, responding to these requests would translate into an increase in the response burden and a substantial increase in the size of the sample and therefore, an increase in the collection costs. For these reasons, and because maintaining the response rates over the past several years was achieved at the cost of an increased effort, the program is currently exploring the use of alternative data as a replacement for or in addition to survey data.

3. Objectives of modernizing the HEP

The major objectives of the program's modernization plan fall within the scope of Statistics Canada's overall modernization objectives. The program is aiming for increased utilization of administrative and alternative data as well as an improvement in the analytical value of the data for current and future users. Where the collection of data is required, the program also aims to reduce the response burden and collect the data using more modern methods, such as through the use of electronic tools.

The users described in the previous section have highly varied needs; some use the SHS data only at an aggregated level (with or without socio-demographic information), while others require to conduct their own analyses based on microdata to, for example, assess the factors related to food insecurity or expenses needed to meet the needs of children. Some of the available alternative sources may not be at the level required and therefore fail to meet these needs. It seems obvious at this point that various models or strategies will be required. So it is necessary to identify groupings of needs and determine which combinations of alternative data or survey plus alternative data could meet those needs. Obviously, issues relating to the coherency of the estimates derived from these types of models will need to be taken into consideration.

The process of acquiring alternative data at the Agency level is one that generally involves three phases:

- First, the need and the utilization potential are identified for a given source;
- Second, enough preliminary data are made available to allow for an evaluation and to determine whether the acquisition process should continue;
- Lastly, certain sources are officially acquired and are available for evaluation and production purposes.

The HEP is currently exploring which sources could be useful to it and is dependent on the Agency's acquisition process. With strict confidentiality and data-protection controls, the potential and limitations of the available sources are assessed and feedback is provided according to the acquisition phase of the source. Later in this article, the results of these assessments will be presented for two categories of sources: housing-related expenditures and transaction data.

4. Alternative housing-related data

Housing expenses are important because they represent approximately 30% of current household consumption expenditure. This expense category includes such things as mortgage payments, rent, utilities and municipal taxes. Together, these four components represent 75% of housing-related expenditures. Even though these expenses are relatively regular, interview respondents may need to refer to invoices or statements in order to report the exact amounts. The appeal of alternative sources for this category resides in the fact that they are related by the concept of address.

One initial source under study relates to consumer debt. This source contains quarterly data on payments and balances for a wide variety of loans, such as mortgages, lines of credit and automobile and student loans. The data are for each individual and aggregated by type of loan, which means, for example, that for an individual with several loans of a given type, only the payment and balance totals are available. This is a very interesting source as far as the program is concerned because it would allow the collection of data on mortgage payments and balances. While mortgage balances do not represent expenses as such, some users are interested in them. Preliminary evaluations of this source have highlighted a number of challenges. First, joint loans (which is often the case with mortgages) are shown with the same information for each individual involved. Identifying and resolving this duplication should be part of any data-processing strategy. For some users, it is important to be able to distinguish the information associated to mortgages on principal residences from those on secondary residences. Since the data are aggregated at the individual level, this type of differentiation will represent a major challenge, absent the assistance of auxiliary sources. Lastly, some mortgage payments include municipal taxes or homeowner's insurance. A strategy would need to be developed such that it would be possible to identify and deduct these amounts from the regular payments.

A second source involving mortgage loans has been identified and could prove useful in assessing and processing the principal source described in the previous paragraph. There is a federal agency that provides insurance to mortgage borrowers whose down payment is below a given threshold. The data from this source therefore involve the mortgages covered by this insurance, representing approximately one-third of Canadian mortgages. Therefore, this product does not provide complete coverage of households having a mortgage, but it has very good coverage of those mortgages that are insured and therefore of mortgage-issuing financial institutions, so this product could be used to assess the coverage of our primary source in terms of financial institutions.

Keeping with the family of alternative data sources tied to housing, one more source under study involves electricity suppliers. On the SHS, respondents are asked to report the latest payment for electricity applicable to their living accommodations. The Agency assessed the data from certain suppliers, and preliminary results show that the quality of the information needed for integration into the programs is generally good. For example, the quality of the addresses is good in urban settings, whereas it could pose certain challenges in a rural setting because of the use of postal addresses or non-civic addresses, such as coordinates. It is important to note that the electricity production and distribution market is relatively segmented in Canada, because there are several dozens of distributors. This is a key factor in the acquisition and standardization of data. In the case of electricity costs, there were few missing values or outliers in the data that were evaluated. The outliers found are often explained by the nature of a dwelling which would at any rate be excluded from the survey universe, such as a collective dwelling or a business.

Preliminary assessments have also identified certain potential challenges in the use of these data. First, it is critical to fully understand the conceptual differences in the data from the various suppliers. For example, the service address, the billing address and the address where the smart meter is located may or may not be available, depending on the source. The variables that are transmitted and their definitions can therefore differ, and they may need to be processed individually. This is especially important when it comes to obtaining a precise geolocation of the dwelling to which the electricity charges apply. The second challenge observed concerns the level of detail of the metadata associated with certain sources, which does not allow all HEP requirements to be met. In fact, electricity costs may or may not include one-time charges (connection, arrears, or interest charges) and it has been observed that the metadata do not necessarily include this level of precision. This is a good example of the challenges posed by the use of data collected or generated by external sources, and for which the statistical agency is not in control of the concept. Lastly, the lack of standardization of the reference periods to which the payments relate would of necessity translate into having to develop standardization strategies specific to each source. These adjustments would make it possible to reflect expenditure from a specific reference year, which is often necessary for expenditure programs. Overall, despite the challenges identified here, the replacement potential of this type of source is attractive when it comes to electricity costs.

The data on consumer debt and on electricity costs are only two examples of a broader collection of housing-related alternative data that were evaluated. These two sources were selected for more in-depth analysis because of the potential that they had for replacing survey data and reducing the response burden that they would involve, especially when it comes to debt-related data. Despite the apparent a priori ease of integration, the evaluations quickly revealed that a number of challenges relating to coverage, the conceptual differences, level of metadata detail, the multitude of sources and pre-processing would mean that integration would not be as easy as anticipated. These obstacles are clearly surmountable but would require a number of adjustment methods to be developed.

5. Transaction data

In this section, two more types of alternative data of interest to the HEP will be described: financial data and data from optical readers (referred to as “scanner data” in the remainder of this document).

Financial data are defined as data on expenses paid through various modes of payment, such as credit card, debit card, electronic transfer and preauthorized cheque. This type of data source presents an interesting potential for the HEP because it relates directly to the spending of individuals. Coverage of these data can vary from source to source. For example, financial data from banking institutions offer a very interesting coverage of modes of payment (credit cards, debit cards, electronic transfers, lines of credit and preauthorized cheques). The financial data provided by credit card companies cover just a single mode of payment. It should be noted that these two sources of financial data (i.e. data from banking institutions and credit card companies) do not cover payments in cash; adjustments would therefore be necessary to correct this undercoverage.

Scanner data refer to sales transactions recorded at cash registers and made in-store. In Section 5.1, the context and the results from a feasibility study on financial data will be presented, and then in Section 5.2 the potential and the limitations of scanner data within the context of the HEP will be discussed. Lastly, Section 5.3 will present a method of combining these two sources, financial data and scanner data, in order to take advantage of the benefits that they offer.

5.1 Financial data: Feasibility Study

To confirm the potential of data supplied by banking institutions and identify their limitations, a feasibility study was carried out. To collect data for exploratory purposes and before proceeding with a pilot project in collaboration with the financial institutions, certain employees of Statistics Canada were invited to supply, on a voluntary and confidential basis, their bank and credit card statements for the reference year 2017. A total of 52 employees agreed to participate, and 42 of these provided complete information for the targeted reference period. Since the financial data obtained through this process came from volunteers, the objective was not to draw inferences on the population, but rather to explore the potential of these data within the HEP context. The data were collected confidentially and did not contain any personal identifying information. In total, the file combining all the transactions collected contained 31,000 transactions for the reference year 2017. Table 5.1-1, below, contains fictitious examples of transactions. The information available for these data includes the transaction date, descriptive variables relating to each transaction, the cost of the transaction and the type of bank account used. The level of detail for the descriptive variables varies from transaction to transaction. For example, as seen in Table 5.1-1, certain transactions are easily identifiable (e.g. “Hydro North” can easily be classified as an electricity cost). Other transactions are difficult, and sometimes impossible, to classify (e.g. “Cheque no.” and “Email trfs” provide no details regarding the type of expense).

Table 5.1-1
Fictitious transactions from the feasibility study on financial data

Date	Description 1	Description 2	Debit	Credit	Account Type
23/08/2017	BT FOOD #1254	PURCHASE 55695	58.87		Cheque
12/12/2017	CHEQUE No.	54	680.00		Cheque
30/10/2017	RED RIBBON PUB	RR PUB OTT ON	94.02		Credit card
06/09/2017	HYDRO NORTH		128.56		Cheque
26/10/2017	CANADA PAY/PAY			1,925.33	Cheque
01/02/2017	EMAIL TRFS	INTERAC E-TRF-58964	200.00		Savings
26/10/2017	MORTGAGE BANK	WEST RED BANK	789.63		Line of credit

Within the HEP context, classifying financial data into expense categories according to a predefined classification system (such as the one used for the SHS) is necessary. This step can represent a major challenge, depending on the type of expense, especially since the descriptive variables often contain just the name of the store, without identifying the goods purchased. For certain "services" or "retail sales" (purchased in a store or on-line) expenditures, it is sometimes possible to find a direct link between the description and a category of expenses. For example, restaurants, taxis and pet shops are easy to identify since the name of the restaurant, taxi company or store, respectively, is shown in the description; also, these transactions are associated directly with a single category of expense. Respondents to the SHS often underreport restaurant and taxi expenditure through mere oversight or because they do not keep their receipts. Financial data therefore offer an interesting potential for reducing the effect of the underreporting of certain small, more frequent expenditures on the estimates. Other "retail sales" expenses are harder to classify. For example, purchases made in grocery stores or department stores are identifiable by the name of the store shown in the transaction description, but it is impossible to find a list of the goods purchased. For these cases, other sources, such as scanner data, could be used to obtain the desired level of detail. This point will be discussed in greater detail in sections 5.2 and 5.3.

In looking at housing-related transactions, some can occasionally be easily associated with a unique category of expenditure; this is especially so for mortgage payments, which can be identified with the help of key words, such as "mortgage" appearing in the description, or public services or in communications which can be identified by the name of the supplier. However, the level of detail is lower than that which is supplied by the SHS. Indeed, mortgage payments can include insurance expenses or municipal taxes. As for communications expenditures, it is impossible to distinguish between Internet, telephone and television expenses. Lastly, some expenses are normally paid by cheque or electronic transfer and in those cases, the description does not supply enough detail to allow for the expense to be classified; this is the case, for example, for rent and for childcare expenses.

The data from the feasibility study have shown the magnitude of the challenge of classifying expenditures into product categories. The descriptive variables for a transaction are essential in order to be able to arrive at a classification. For this type of classification exercise, the recommendation is to consider the use of machine learning algorithms. In fact, supervised algorithms would allow for an automatic classification of textual variables (description of the transaction, in this case) into predetermined product categories. These methods, however, would not be able to associate an expenditure category in cases where the description was too vague. In some cases, the description provides no details (e.g., "Cheque no."), but the fact that the payment recurs could yield important information. For example, a cheque for a substantial amount that recurs every month could possibly be associated with a rent payment. A number of hypotheses would be required to allow such conclusions to be drawn, but these could be validated through a broader study of the entirety of the transactions for a household.

The findings from the feasibility study show the enormous potential of this type of data for the HEP. Most modes of payment are included (credit cards, debit cards, cheques and electronic transfers) and therefore a large portion of expenses is covered. However, transactions made in cash are not covered. Adjustments would have to be made to the data for this undercoverage, all while taking into account the fact that cash transactions are not uniform within the expenditure categories. In fact, cash is used more often for lower-value purchases and more often for certain types of expenditures than for others (e.g. restaurant meals, entertainment and parking) (Henry et al., 2015). A global adjustment therefore would not be sufficient; consideration would have to be given to an adjustment specific to each expenditure category. In classifying financial data, other data sources may be necessary in order to obtain details regarding the goods purchased. Scanner data are a good example of this.

5.2 Scanner data

Scanner data cover sales transactions recorded at cash registers and made in stores. These data come from retailers and the records currently being received by the Agency are at an aggregated product level. Total sales and the total number of units sold for a given product (in a given store and for a given week) are available; information is not provided for each individual transaction. Table 5.2-1 includes a fictitious example showing how scanner data are currently formatted.

Table 5.2-1
Scanner data as currently formatted

UPC	Product Description	Location within Store	Week	Store Name	Sales ($)	Quantity Sold
2174060000	EXTRA LEAN GROUND BEEF	Meat	4	Store #2	1,154.95	100
1122334455	ALLERGY TABLETS	Non-prescr. Medication	23	Store #6	83.88	12
1112223334	NAIL POLISH, PINK	Cosmetics	7	Store #3	1.99	1
6568400537	GREEK YOGURT 0%, VANILLA	Middle of store	15	Store #1	13.98	4
1020304050	BAR SOAP, SENSITIVE SKIN	Health and Beauty Products	15	Store #3	19.90	10

File variables provide information on the Universal Product Code (UPC) and a highly detailed description that allows products to be classified into categories. The attractive advantage of this type of data when compared with the financial data described in the previous section, is the high potential for classifying the data into predetermined product categories, but one drawback is the fact that the information is aggregated at the product level, and does not include any socio-demographic information. Also, scanner data do not cover only expenditures made by Canadian households; the impact of including expenditures of businesses and international travellers could be significant for certain product types.

Scanner data from retailers currently available to the Agency represent approximately 50% of the market share of the sale of food products in Canada. The total amount spent on food cannot be derived because coverage from retailers is incomplete but the distribution according to expense category could be calculated, that is, the total could be allocated into major categories, such as meat, fruits and vegetables, dairy products, etc. This type of distribution would be aggregated, and not available according to socio-demographic domain, but this information could still be useful for certain users. In order to confirm the potential of such a distribution, the quality would have to be evaluated. The quality of the distribution will be dependent on the performance of the classification algorithm used and coverage of the scanner data currently available.

At the present time, the scanner data from a major Canadian grocery chain have been classified into food product categories using a machine learning algorithm. The distribution of total food sales into principal categories from scanner data for this grocery store chain was compared to that from the SHS. The results are shown in Table 5.2-2 for reference year 2015.

Table 5.2-2
Distribution of scanner data from a grocery store chain and the Survey of Household Spending (SHS) for 2015

Category	Scanner Data (grocery store chain)	SHS
MEAT	18.1%	19.5%
FISH AND SEAFOOD	2.8%	3.6%
DAIRY PRODUCTS AND EGGS	15.2%	14.8%
BAKERY AND CEREALS	14.5%	14.7%
FRUITS AND NUTS	12.7%	12.3%
VEGETABLES	11.8%	11.1%
OTHER FOOD PRODUCTS	24.2%	23.9%

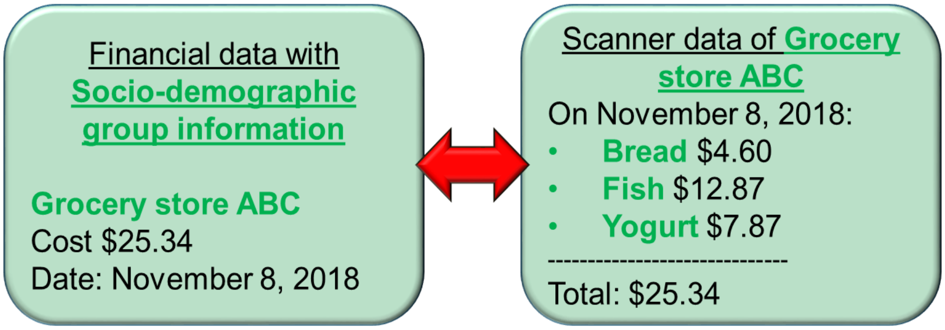
It is interesting to note that both distributions are very similar, even though just one grocery store chain was considered (compared to the SHS, which collects expenditures by households in all stores). Some differences between the two distributions are expected, especially at the more detailed product levels, because the data come from two distinct sources, each with sources of error. For example, scanner data can suffer from classification and coverage errors, and survey data can suffer from sampling errors and errors not caused by sampling (e.g. recall errors and underreporting). However, it is encouraging to note that the two distributions are similar, at least in the case of the major food-product categories, and it is expected that the two distributions will approach one another once scanner data from other retailers become available.

5.3 Combining financial data with scanner data

The two previous sections presented two data sources of interest for the HEP: financial data and scanner data. These alternative data sources have both advantages and disadvantages. The financial data cover several modes of payment and have the potential of supplying socio-demographic information but do not provide a sufficient level of detail regarding the products purchased. For their part, the scanner data provide highly detailed information on products, allowing for them to be classified into predetermined categories, but do not offer any detailed socio-demographic information. This section will look at the idea of combining these two data sources, using a record linkage process, in order to leverage the respective advantages of the two sources. In order for such a linkage to take place, it is important to note that scanner data should be available at the level of the individual transaction, and not at the aggregated product level. As a result, the acquisition of this level of scanner data is of interest for the HEP.

The example of a purchase made at ABC grocery store can be used to illustrate this. The transaction description from the financial data makes it possible to identify ABC grocery, the total cost and the date of the transaction, as well as the socio-demographic domain. The transaction description from the scanner data from ABC grocery provides details regarding the products purchased, the unit cost for each product (totalling the same cost as supplied in the financial data) and the purchase date. A record linkage based on the name of the store, the cost and the date of the purchase makes it possible to obtain a list of the products purchased, along with a socio-demographic domain. The example produced by the combination of the two sources is shown in Figure 5.3-1 below.

Figure 5.3-1
Combining financial data with scanner data



Scanner data, by virtue of the fact that they are now available at an aggregated level, can be used to produce aggregated distributions by category of product, as discussed in Section 5.2. These data also have additional potential when they are available at the transaction level. In fact, scanner data could then be integrated with the financial data in order to benefit from the advantages of the two sources, and thereby meet the needs of certain users. It would then be extremely important that the two data sources contain good-quality common linkage variables.

6. Conclusion

This article presents various sources of alternative data of interest to the HEP. The data relating to consumer debt and the data relating to electricity costs provide an attractive potential for replacing survey data but would require a major investment in order to provide a better understanding of, and mitigate the issues surrounding, coverage, conceptual differences, the level of metadata detail, the multitude of sources and pre-processing. A feasibility study on financial data has shown the enormous potential of this type of data for the HEP, especially when it comes to broad coverage of the various modes of payment. However, cash expenditures are not covered, and the challenges seen with respect to classifying transactions show that supplementary data sources, such as scanner data, would be needed in order to obtain the required level of detail. Currently available scanner data can provide an aggregated distribution of total food sales in primary product categories. If these data were to become available at the transaction level, an interesting potential for integrating data could be considered through a linkage to the financial data. The data explored up until now show interesting potential for the HEP, but there is a lot of evaluation work that needs to be done in order to determine whether they could replace or be integrated with data obtained through surveys.

References

Eurostat (2017), “Household Budget and Time Use Surveys – Launch of the Work of the Task Forces”, Meeting of the European directors of social statistics, Luxembourg, March 2017.

Henry, C., K. P. Huynh, and Q. R. Shen (2015), “2013 Methods-of-Payment Survey Results”, Bank of Canada Discussion Paper No. 2015-4.

Statistics Canada (2018), *User Guide for the Survey of Household Spending, 2017*.