BANK OF CANADA
BANQUE DU CANADA
CELEBRATING 75 YEARS
CÉLÉBRONS 75 ANS

Working Paper/Document de travail
2010-6

Assembling a Real-Financial Micro-Dataset for Canadian Households

by Umar Faruqui

www.bank-banque-canada.ca

Bank of Canada Working Paper 2010-6

February 2010

Assembling a Real-Financial Micro-Dataset for Canadian Households

by

Umar Faruqui

Financial Stability Department
Bank of Canada
Ottawa, Ontario, Canada K1A 0G9
ufaruqui@bankofcanada.ca

ISSN 1701-9397

Acknowledgements

I thank Brian Murphy (Statistics Canada), Ron Morrow, Allan Crawford, Christopher Reid, Cesaire Meh, Bob Fay, Jason Allen, Evren Damar and Jonathan Witmer for their comments and suggestions. Wendy Kei provided excellent research assistance for this paper. All remaining errors and omissions are my own.

Abstract

The lack of consolidated Canadian micro data on household balance sheets and expenditures has been an important impediment to empirical research into real-financial linkages in the Canadian household sector. Our paper attempts to fill this data gap by merging household balance sheet data from the Canadian Financial Monitor survey with household expenditure data from the Survey of Household Spending. The merge process uses a categorical matching framework aimed at preserving the heterogeneity in the underlying datasets. The resulting combined dataset is a novel source of Canadian micro data on household finances and spending patterns. The dataset covers the period 1999 till 2005 and contains roughly 11,000 observations (households) for each year. We plan to use these combined data to test key real-financial linkages (like those between house prices, debt and household expenditures) for the Canadian household sector.

JEL classification: D10, C81
Bank classification: Sectoral balance sheet

Résumé

L'absence de microdonnées regroupant les informations sur les bilans des ménages canadiens et leurs dépenses a sensiblement entravé la recherche empirique sur les liens entre variables réelles et variables financières dans le secteur des ménages. Les auteurs cherchent à combler cette lacune en fusionnant les données sur les bilans des ménages tirées de l'enquête Canadian Financial Monitor et celles provenant de l'Enquête sur les dépenses des ménages. Ils procèdent à cette fusion en appliquant un cadre d'appariement par catégories qui préserve l'hétérogénéité des données sous-jacentes. L'ensemble de données réuni représente une nouvelle source de microdonnées canadiennes sur les finances des ménages et leurs profils de dépenses. Il s'étend de 1999 à 2005 et renferme quelque 11 000 observations (ménages) par année. Les auteurs prévoient utiliser les données regroupées pour étudier certains des principaux liens entre variables réelles et variables financières dans le secteur canadien des ménages, par exemple ceux qui unissent les prix des maisons, la dette et les dépenses des ménages.

Classification JEL : D10, C81
Classification de la Banque : Bilan sectoriel

1.0 Introduction and summary

In recent years, mushrooming interest in financial accelerator effects (see Bernanke et. al, 1999) has spurred empirical research on real-financial linkages within both the household and business sectors. On the household front, a number of international studies have used micro data to empirically examine the links between house prices, debt and consumption.[3] For example, one key question that the micro data provide insight into is the estimated magnitude of the wealth effect and the collateral channel[4] in the house price - consumption relationship. These analyses based on micro-data are important for two reasons. First, the results can be used to identify key behavioural relationships that are not easily disentangled with aggregate data. This information can be used to calibrate macro models with financial accelerators, amongst other things. Second, micro data studies allow us to better understand how household expenditure behaviour is related to asset prices and how (and if) this relationship changes over time.

There are no studies to date (that we are aware of) that examine the household real-financial linkages for Canada at the micro level. The main reason is likely the lack of a consolidated micro data set for Canada containing both financial and real variables. Our paper attempts to fill this data gap by putting forward a second-best solution: merging household balance sheet data from the Canadian Financial Monitor survey with household expenditure data from the Survey of Household Spending. The merge process is structured to preserve the heterogeneity in the underlying datasets, which is a key attraction of micro data. The resulting combined dataset is a novel source of Canadian micro data on household finances and spending patterns.

The remainder of this paper is organized as follows. The next section provides some background on the subject of data combination from multiple sources. Section 3 describes the data used in our analysis. Section 4 outlines methodology used for merging. The following section outlines the robustness check of the merged data. Section 6 concludes with a summary of main findings and a brief discussion on how we plan to use these data going forward.

[3] These studies include Attanasio, Blow, Hamilton, and Leicester (2005), Campbell and Cocco (2005), and Benito and Mumtaz (2006) for the U.K. household sector.

[4] The collateral channel (also referred to as the financial accelerator channel) works as follows: an increase in house prices raises the value of housing collateral, improving access to credit and supporting consumption. This effect is particularly important for households that might otherwise have been constrained by the availability of credit.

2.0 Background

This section is divided into three parts: section 2.1 presents a short discourse on the main techniques that have been used in the literature to combine information from multiple data sources, the following sub-section reviews selected papers on the subject, and section 2.3 outlines reasons why this paper uses the statistical matching approach to combine real and financial data for Canadian households.

2.1 Techniques for combining data from multiple datasets

The literature on combining information from multiple datasets largely revolves around three strategies: exact matching, categorical matching and regression-based techniques. The first two methods aim to create a single merged dataset containing all relevant variables from multiple datasets, while the third method relies on regression-based tools to leverage information from multiple datasets. Some advantages and disadvantages of each approach are outlined in Table 1 below.

Table 1

Method	Advantage	Disadvantage
Exact matching	- Full consistency of information in the combined dataset. - Retains richness in terms of heterogeneity of the micro data.	- Not possible for most datasets. - Expensive in terms of time and effort. - Privacy concerns.
Statistical matching, in particular, categorical matching method	- Ease of use. Categorizing, sorting and matching are relatively straightforward. - Ensures consistency of information in the combined dataset. - Heterogeneity of micro data is largely retained.	- Requires assumption of conditional independence between merged data. - Time consuming, though less so than exact matching. - Matching is not exact and this can lead to additional noise in the combined dataset.
Regression-based techniques	- Requires minimal manipulation of the data. - Does not require conditional independence assumption.	- Hard to use with pseudo panel data estimation. - Affords less flexibility than if datasets are merged.

Exact matching is relevant if for two datasets, A and B, the records are drawn from the same population base. If the samples are large and the draws are random, there is a possibility that a number of records from both datasets are overlapping. In this case, we can do an exact match of the overlapping records as long as the records are consistently and uniquely identified. An example of a scenario where exact matching might be practical is if tax records were being linked with pension records. Both the tax dataset and the pension dataset would be large (if not comprehensive) and the records can be uniquely and consistently identified using an individual's social insurance number. In practice the possibilities for applying exact matching are rare.

Categorical or "sort-and-merge" technique is a specific example of statistical matching method for combining multiple datasets. Categorical matching requires a number of steps. Typically the first step is to identify a number of key common variables (Z) between the two datasets along with the unique variables (Y) in the donor dataset that need to be added to the host dataset. Common bins or intervals of each of the Z variables are then created and records in each dataset are grouped into these bins. This process is repeated until the bins are as fine as possible and then each dataset is sorted according to the X variables. The partitioned and sorted records from the two datasets are then matched. Finally, the unique variables (Y) are copied from the donor records to the matched host records.[5]

Regression based techniques to combine information from more than one dataset have largely revolved around the two-sample instrumental variable (2SIV) approach. Originally proposed by Klevmarken (1982), the framework has been further developed by a number of other authors including Ridder and Moffitt (2007). The basic steps in the 2SIV approach for a single-equation linear estimation are outlined in equations 1-3 below.

$$Eq.\ 1:\ Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$$

$$Eq.\ 2:\ X = \gamma Z + \Phi$$

$$Eq.\ 3:\ Y = \beta_0 + \beta_1 \gamma Z + \beta_2 Z_o + \varepsilon + \beta_1 \Phi$$

Given two datasets A and B, Z is the vector of common variables present in both datasets and Z_o is a subset of Z. Meanwhile, X and Y are variables unique to datasets A and B, respectively. In order to estimate Eq. 1, which combines information from the two datasets, two steps are required. First, Eq. 2 is estimated using dataset A. Second, the information from Eq. 2 is used to estimate Eq. 3. Details of this approach, including a discourse on statistical properties, are discussed in Ridder and Moffitt (2007).

[5] The sort-and-merge methodology is described in more detail in Sections 4.1 - 4.3.

2.2 Review of selected papers[6]

Arguably, the seminal piece on using statistical merging techniques to combine information from multiple micro-data sets was written by Okner (1972). In his paper, Okner used a sort-and-merge technique to merge data from the 1970 Survey of Economic Opportunity with the 1970 IRS tax file, in order to study the distribution of total household income in the U.S. population. Subsequently, Ruggles and Ruggles (1974 and 1977) have written extensively on the motivation and techniques for merging micro datasets. In particular, Ruggles and Ruggles assert that amongst the various techniques for combining data from different sources, the sort-and-merge technique is arguably the most useful, especially when the datasets are large. Adler (1974) used the sort-and-merge methodology to link together information from the 1970 Canadian Survey of Consumer Finances with the 1970 Family Expenditure Survey for Canadian households. Meanwhile, Bordt et al. (1990) undertook the daunting task of linking information from four separate datasets (Survey of Labour Income and Dynamics, Canadian Revenue Agency tax returns data, employment insurance claims and the Survey of Household Spending). The merged dataset developed by Bordt et al. (1990) has been used in Canada for taxation modeling. More recently, Sutherland et al. (2001) have built on Okner's and Ruggles' work and used statistical matching to link the UK Family Resources Survey with the Family Expenditure Survey.

Statistical matching has faced considerable criticisms since Okner's seminal piece. Sims (1972) and others have pointed out that the use of categorical matching imposes the implicit assumption that the unique variables in the two datasets are independent conditional on the set of variables common to both datasets. Ridder and Moffitt (2007) provide an excellent synopsis on data combination techniques and argue that the conditional independence assumption negates the usefulness of the statistical matching of independent samples. Instead, the authors assert that data merging is sub-optimal to data combination via regression-based methods, either using a two-sample instrumental variable, or two-sample maximum likelihood approach.

A number of empirical studies have used regression methods to deal with information that is contained in multiple datasets. Angrist and Kreuger (1992) apply two-sample instrumental variable approach to study the links between school-entry age and completed years of schooling using data from the 1960 and 1980 census. Caroll and Weil (1994) study the relation between wealth income ratios and income growth using data from the U.S. Survey of Consumer of Finance and the Panel Study of income dynamics. Similar regression based approaches are used by Currie and Yellowitz (2000) and Dee and Evans (2003) to combine data from multiple datasets.

The debate on the appropriateness and efficacy of using statistical matching to combine information from multiple datasets is well summarized in Rassler (2002).The author notes that statistical matching is rarely the first-best option for combining information

[6] Given that exact matching is not a viable approach for our datasets we do not cover this topic in our review of literature. Instead we focus on selected papers on statistical matching and regression-based data combination techniques.

from multiple data sources. However, from a practical standpoint, statistical matching is one technique of leveraging information from multiple data sources.

2.3 Data combination approach adopted in this note

Statistical matching is used in this paper to combine consumer spending information from the Canadian Survey of Household Spending with the financial information contained in the Canadian Financial Monitor survey. Our choice of data combination technique is driven by practical factors. The primary driver is that we plan to use the unbalanced panels from the Canadian Financial Monitor for our analytical work and without an actual merge of the two datasets the panel aspect of the Canadian Financial Monitor data cannot be effectively utilized. In addition, there is a strong precedent in the use of statistical matching to combine micro data in Canada. In particular, Statistics Canada has used statistical matching to create the Social Policy Simulation Database[7] (see Bordt et al. (1990)), which links the Survey of Labour Income Dynamics, Survey of Household Spending, personal income tax records and employment insurance records.

3.0 The Data

This paper uses data from the Canadian Financial Monitor (CFM) survey and the Survey of Household Spending (SHS).

3.1 The CFM Data

The CFM survey is conducted by Ipsos Reid Canada and collects detailed household balance sheet information.[8] The survey has a sample size of approximately 12,000 households per year responding through a mail-in questionnaire. An important concern of a household financial survey is to capture the distribution of income and wealth across households. Since income and wealth are highly concentrated within a few "rich" households, CFM survey over samples high-income households.

Households in CFM are sampled from a pool of 60,000+ units that have already indicated their openness to being part of a survey. This pool is frequently refreshed, with some new households joining the pool and others dropping out of the pool. One implication of drawing CFM's sample from this household pool is that some households appear in the in multiple CFM surveys over time, though not necessarily in contiguous years.[9] This aspect of the CFM data is of particular interest to us as it allows a way to create an (unbalanced) panel for future analytical work.

---

[7] See http://www.statcan.gc.ca/microsimulation/pdf/spsdm-bdmsps-overview-vuedensemble-eng.pdf for further details on the dataset.

[8] See Ipsos Reid Canada's website for more information: *http://www.ipsos.ca/pdf/ipsos_canFinMon.pdf*

[9] Each household is assigned a unique identifier, much like a Social Insurance Number, when it enters the pool. This allows the household to be identifier across different years.

The survey content and collection methodology has remained mostly unchanged since its inception in 1999. The 2008 survey consisted of ten sections of which three sections were on assets, two on debt, two on banking behaviour and one section each on household characteristics, attitudes, financial advice and retirement. The household characteristics section collects information on the age group of the household head, family income, family size and marital status of the household head, amongst other things. Up until 2006, CFM data have been primarily used by Canadian financial institutions for market research.

More recently, CFM data have been used extensively at the Bank for assessing and monitoring the financial situation of Canadian households. In particular, Faruqui (2008) uses these data to evaluate the changes in the distribution of the household debt service ratio (DSR) over the 1999-2006 period. Dey et. al (2008) uses CFM data to simulate changes in the distribution of the DSR under various stress scenarios. The authors show how this framework can be used by analyzing the effects of two different scenarios on the distribution of the debt-service ratio and the impact on vulnerable households. The CFM data have also featured prominently in external Bank publications (e.g. FSR) and in internal conjunctural analysis of household financial health. Finally, the means of payments information in the survey is also currently being explored in several research projects at the Bank.

3.2. The SHS Data

The SHS survey is cross-sectional survey conducted by Statistics Canada, and provides information on household spending and dwelling characteristics.[10] The effective sample size (i.e. the number of respondents to the Survey) of the SHS varies each year, ranging from 14,000 to 17,000 households, annually. The survey data are collected via personal interviews. The SHS is a purely cross-sectional survey with no panel component to it.

The SHS data are used widely both in the private and public domain. For example, these data are used to benchmark the consumption basket for CPI calculations by Statistics Canada, and are employed by labour and contract negotiators to discuss wage and cost-of-living clauses.

3.3. Comparison of the CFM and SHS data

Table 2 compares the main characteristics of the CFM and SHS datasets. Similarities across the two surveys include the target population (i.e. a representative sample of the Canadian household population), demographic information about the respondent,[11] and broadly similar sample sizes. The main differences include the focus of the survey questions (balance sheet information for CFM and household expenditure for SHS), and data collection methodologies.

[10] More information about the SHS survey including a discussion of imputation, estimation, and survey design can be found on the Statistics Canada website: *http://www.statcan.gc.ca/*

[11] A comparison of the main demographic characteristics of the two datasets is outlined in Appendix 1.

Table 2: General characteristics of the CFM and SHS survey

	CFM	SHS
Starting point	1999	1997*
Last data point	2008	2005**
Frequency	Monthly	Annual
Sample size in database	~12,000	~14,000 – 17,000
Collection method	Mail-in questionnaire	Personal interview
Panel or cross-section data	Partial panel. Some households appear in multiple CFM surveys. However, the panel is not balanced.	Cross-Section

Notes:
* Prior to 1997, similar information was gathered by a combination of the Statistics Canada Family Expenditure Survey and the Household Facilities Equipment Survey.
** When this project was started SHS data was only available up to (and including) 2005. Latest currently available year of SHS public-use micro-data is 2007.

Both the similarities and some of the differences make the CFM and SHS surveys good candidates for merging. The similarities across surveys are crucial for the merging process to be successful. For example, common structure of the two surveys and common variables are important to create a link between the two datasets. Meanwhile, the complimentarity of data collected by each survey is important for the combined dataset to have value-added over-and-above the individual surveys.

4.0 Framework for merging the CFM and SHS datasets

This section details the procedure we follow to link the CFM balance sheet information with the SHS expenditure data using statistical matching methods. The particular matching method we use for our work is commonly referred to as categorical matching (also known as "sort-and-merge" method).

4.1 Preparing the datasets for the merge:

For the merge process, we assign CFM as the host, while SHS is the donor dataset. This means that information from the SHS will be added to the CFM data to create the final merged dataset.

Our matching process relies on using demographic characteristics of households to match records across the two datasets.[12] While both datasets collect information on a number of

[12] There are several other non-demographic variables that are common between the CFM and SHS surveys, like value of vehicles owned/leased, number and value of recreational vehicle owned/leased, and the

key overlapping household demographic variables, the data are not directly comparable in their raw form. For example, household income is collected as ranges (i.e. between $15,000-$19,999, etc.) in CFM, while SHS asks households their exact income. In this case, for categorical matching to work, datasets should store information on household income in the same manner.

Categorical matching requires both the CFM and SHS datasets to have a common set of household demographic variables, defined in a similar fashion. To facilitate categorical matching we redefine key demographic variables in the two datasets such that they have the same groupings. This ensures concordance between the relevant demographic variables from the two datasets. Bin or group sizes are determined largely by the existing grouping used in the CFM data. Appendix 2 lists the key demographic variables used for the merge and defines a common set of groups for each variable. These variables include housing tenure, household size, household income, age of household head, province of residence and marital status.

4.2 Creating new expenditure variables in the SHS:

The SHS survey collects household expenditures at a very detailed level. While this level of detail is important for studying particular issues, they are less important for examining the broad links between the financial position of households and their spending behaviour. Therefore, we aggregate household expenditures data by key groups and use these groups in the merge process. Our aggregated expenditure categories are constructed to follow the National Income and Expenditure Accounts (NIEA) groupings, and include expenditures on durables, semi-durables, non-durables, shelter, and services excluding shelter. The construction of these aggregates based on the SHS data is detailed in Appendix 3.

Concordance between the NIEA expenditure definitions and our newly created expenditure variables are not exact due to the following reasons:

a. Some SHS expenditure variables within the public database can belong to more than one consumption categories. For example, home entertainment equipment and services (M148) can belong to durable, semi-durable, and services categories. In this analysis, most unpublished small components that make up M148 are service-related; thus, M148 is considered as “service excluding shelter”.
b. Certain SHS expenditure variables are not considered “value-added” and as such, would not be listed under the NIEA. An example would be health insurance premiums.

In addition to creating these aggregated expenditure variables in SHS, we also define the ratio of these expenditure variables to total household income before tax. The ratios play a key part in our merge process as the household spending information is transferred from SHS to CFM records via these ratios. The main reason for using ratios rather than levels

---

current value of stocks and bonds owned by the household. However, for the matching process we judge it sufficient to use demographic characteristics to group and match households.

to transfer data is to minimize adding noise in the data during the merge process, given differences in the way that income is captured by SHS and CFM (as point-estimate in SHS and via ranges in CFM).

4.3 Merge procedure:

The broad steps in our merge process are shown in Figure 1. The first step -- preparation of the databases for the merge procedure -- has been already discussed in the previous sub-sections.

The second step involves sorting the records in each database by key demographic variables (Appendix 2). The sorting is done sequentially starting with most important variable (income in our case).[13] To illustrate this sort process, consider the following example. Suppose we wanted to sequentially sort the records in database A by home tenure status, age, and income groups. The process for the sorting would be as follows: we would first sort according to whether the household owned or rented their primary residence, then repeat this process for age and income groupings (Figure 2). At the end of the sort process, the records of database A would be grouped into narrow bins (e.g. homeowners, who are old and rich) with each bin containing a small number of records.

[13] Correlations between the demographic variable and household expenditure are used as a guide for ordering the selected demographic variables in Appendix 2.

Figure 1: Flowchart of merge procedure

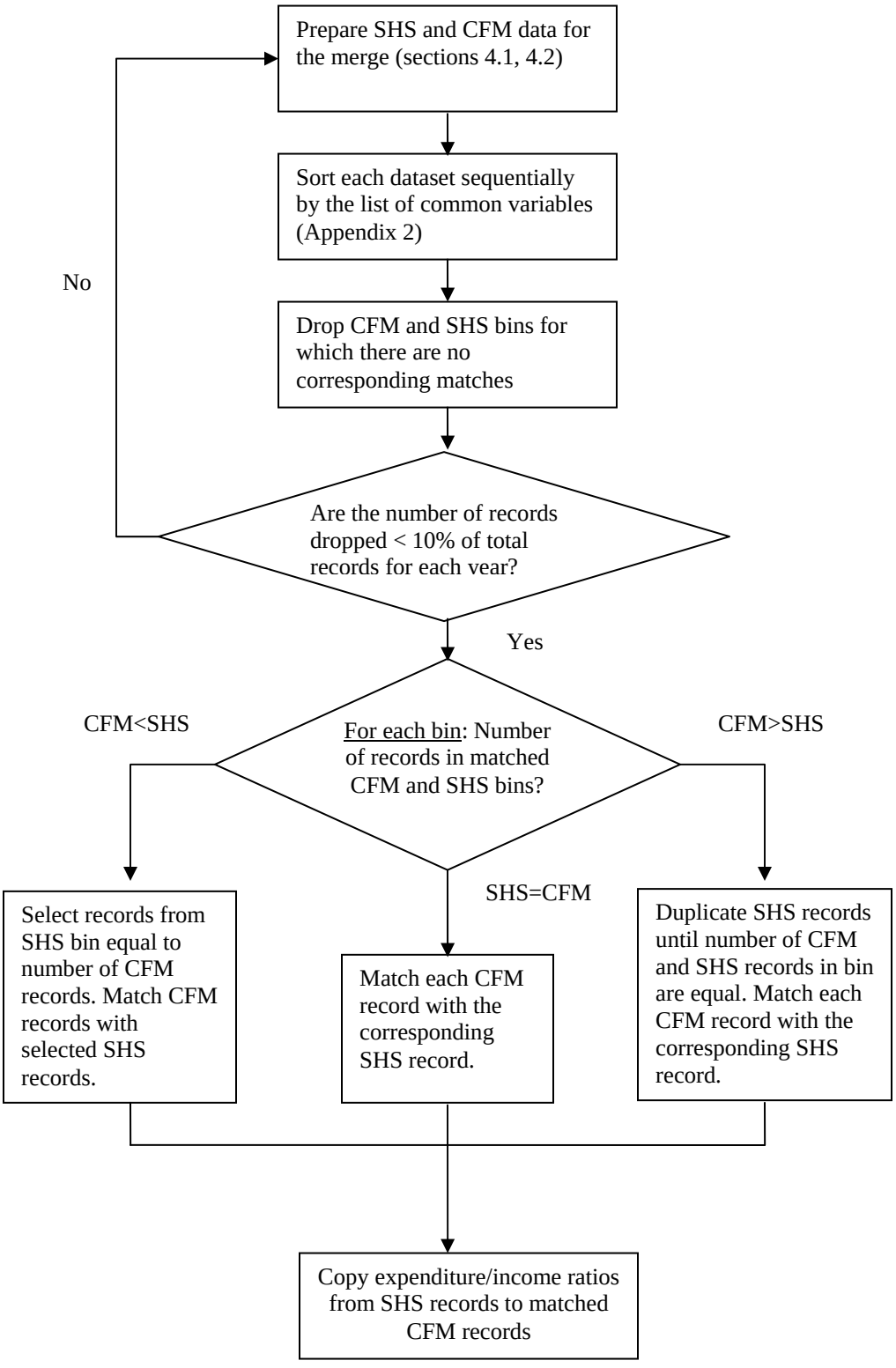
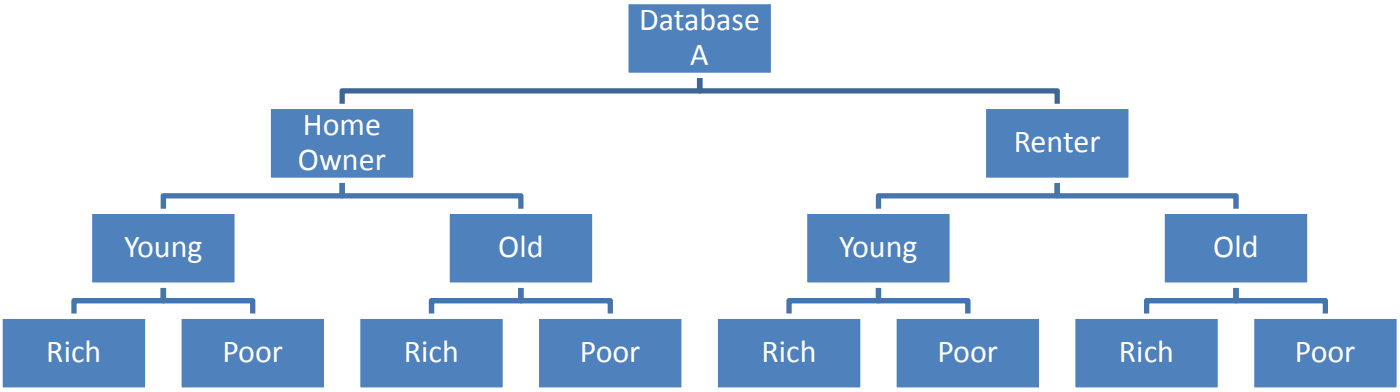


Figure 2: Sequential sort process -- An illustration



In similar fashion to the illustrative example above, CFM and SHS datasets are sorted separately and the records are grouped into bins. Given the large number of bins that are created during this sort process, there is a possibility that some bins in either CFM or SHS or both may contain zero records. We discard these bins (in both datasets) as it would add multiple layers of complications to the merge process. We then take stock of the number of records from CFM and SHS that are dropped because of this step. If a significant number of records are eliminated,[14] then this may signal a need to reassess the sorting procedure (e.g. perhaps fewer variables/bins should be used for the sort) and the sorting step has to be repeated.

Once the datasets are properly sorted, they are ready for merging. The merge algorithm relies on the matching records in common bins across the two datasets. The merge is conditional on the number of records in a given bin across the two databases. If the number of records in a bin is the same for SHS and CFM data, the matching is straightforward, as each CFM record is paired with a corresponding SHS record. For cases where the number of matched records is not the same, we impose the following rules:

a. For a given bin, if the number of CFM records is greater than the number of SHS records, then SHS records are duplicated until the number of SHS records is the same as for CFM. For example, if there are five CFM records and three SHS records in a given bin then the first two records in SHS bin are duplicated so that there are a total of five entries in the SHS bin (same as for CFM).
b. For a given bin, if the number of CFM records is less than the number of SHS records, then some SHS records are dropped until the number of records is aligned. Choosing which SHS records to keep (or drop) is subjective. Our approach chooses records randomly from the set of SHS records such that total number of selected SHS records is the same as the number of CFM records in the bin.

[14] We use 10% of total number of records (per year) as our benchmark, however this criteria is subjective.

The final step in the merge process is to copy the relevant household expenditure information from SHS records to the matched CFM records. As noted above, we have elected to transfer household expenditure information from SHS to CFM using expenditure-to- income ratios rather than expenditure levels. However, we are interested in both ratios and levels. Therefore, once the ratios are copied over from SHS records to matched CFM records, they are multiplied by the household income for each CFM record to get the level of expenditure as well.

5.0 Robustness check of the merged data

There are two main questions that we address in this section: (a) does the merged dataset compare favourably with the original data sources? and, (b) is our merge procedure efficient?

5.1 Comparing the merged CFM/SHS dataset with the original CFM data

The first check of the merged dataset is to compare it with the original CFM dataset. During the merge procedure, a number of records had to be dropped (see Section 4.3) and it is important to verify that these adjustments did not change the overall demographic properties of the CFM data.

Table 3 shows the total number of records by year in the original CFM and the final merged datasets. While some records had to be dropped during the merge process, this did not exceed 7% of total original CFM records in any given year.

Table 3: Total records in dataset by year

	Original CFM	Merged	% dropped
1999	12410	11644	6.2%
2000	11765	10976	6.7%
2001	11856	11226	5.3%
2002	11852	11213	5.4%
2003	12047	11437	5.1%
2004	12655	11959	5.5%
2005	11648	10905	6.4%

In addition, as Figures 3 and 4 show, the overall age and income distributions of the data are maintained in the final merged dataset.

Figure 3: Age distribution

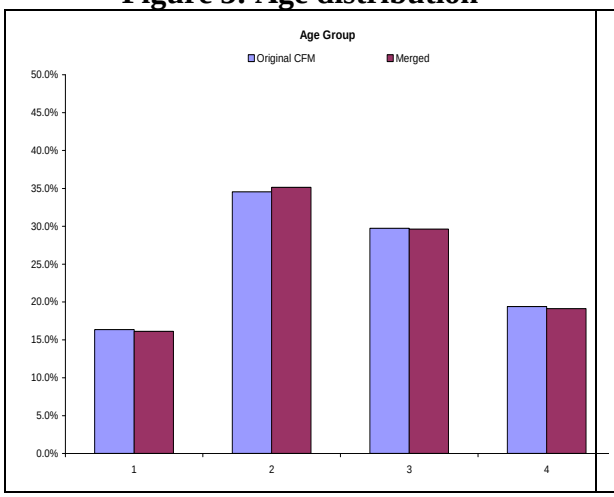
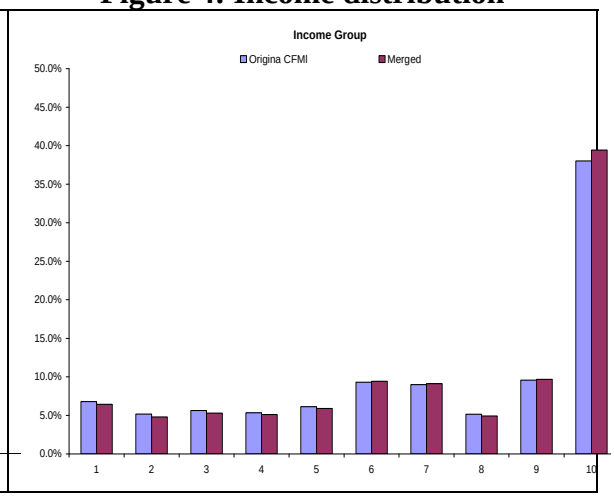


Figure 4: Income distribution



5.2 Comparing the merged CFM/SHS dataset with the original SHS data

One important aim of the merge procedure was to maintain the underlying heterogeneity of the SHS data in the consolidated dataset. To assess whether this criteria is met, we can compare the distributions of selected expenditure-to-income ratios from the merged dataset with those from the (donor) SHS dataset.

Figure 5a depicts the kernel density distribution[15] of the durable expenditures-to-income ratios from the merged dataset (solid line) and the SHS data (dashed line). The figures show that there is no discernable difference in the two density distributions. Similar observations can be made when we compare the distributions for semi-durables-to-income, non-durables-to-income and expenditure on services relative to income across the two datasets (Figures 5b-5d). These graphics suggest that the heterogeneity in the source SHS data has been largely preserved in the merged data.

[15] Kernel density is a fitted density curve. Smoothing parameter for kernel distributions shown in Figures 5a-d is set to 0.03. Density distributions in Figures 5a-d are for the 1999-2005 period (combined) and exclude records with expenditure-to-income ratios (in per cent) of less than zero or greater than 100.

Figure 5a: Durable/Income

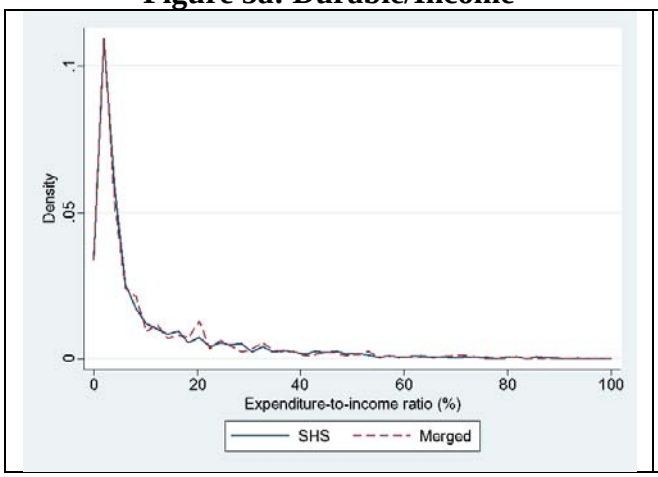


Figure 5b: Semi-durables/Income

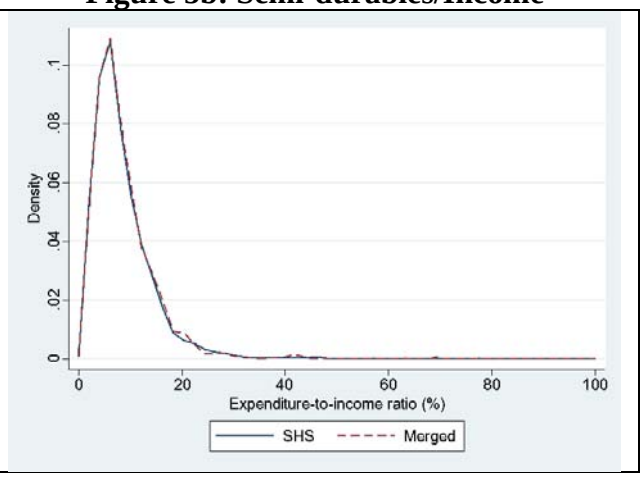


Figure 5c: Non-Durables/Income

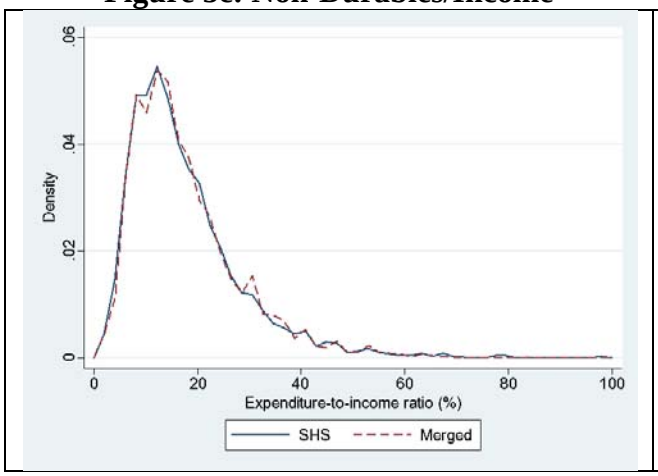
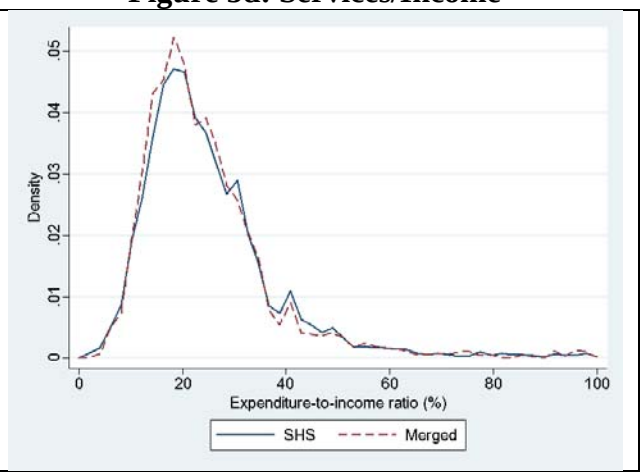


Figure 5d: Services/Income



To formally assess the closeness of distributions of consumption-to-income ratios from the merged dataset and the SHS data (Figures 5a-5d), we perform a Wald test on the difference in various moments of the distributions using a bootstrap framework. The null hypothesis in our tests is that the difference in moments is zero. The difference is bootstrapped 20,000 times. The results from these tests (Table 4) show that for the most part we cannot reject the null hypothesis that the moments of the distributions are equal.

Table 4: Test results

	p-Value for H_o = Equality across the two distributions			
	Mean	**Variance**	**Skewness**	**Kurtosis**
Durables/Income	0.32	0.29	0.99	0.99
Non-Durables/Income	0.95	0.91	0.60	0.38
Semi-Durables/Income	0.79	0.74	0.47	0.56
Service/Income	0.59	0.03*	0.27	0.56

* Statistically significant at the 95% confidence level

5.3 Is categorical matching efficient?

One test of how efficiently the categorical matching methodology replicates SHS's (actual) consumption-to-income ratios in the final merged dataset, is to try the categorical matching method on a sub-set of the SHS data.

The experiment is set-up as follows:

a. For a given year we randomly split the SHS data into two sub-samples. We use 2005 for our simulation exercise.
b. Treat the first sub-sample as the "host" data, and second sub-sample as the "donor data". Records from the host and donor data are matched as per the categorical matching method described in section 4.
c. Estimated consumption-to-income ratios are generated for matched host records.
d. Estimated and actual consumption-to-income ratios are then compared for the host records. If the categorical matching is efficient, the actual and estimated consumption-to-income ratios would be a tight match.

The results from this experiment are depicted in Figures 6a-d and Table 5.[16]

Figure 6a: Durable/Income

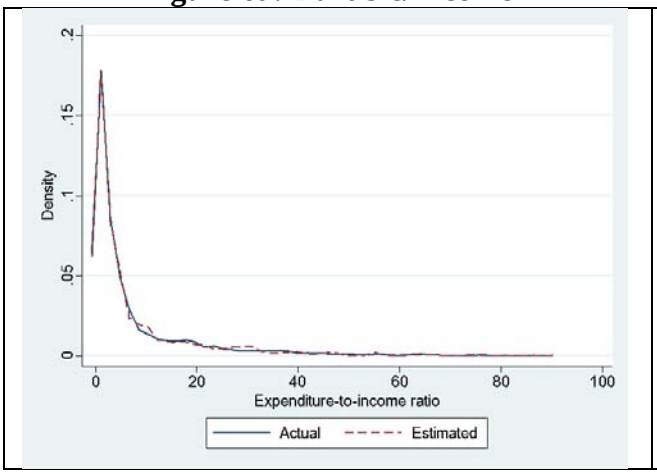


Figure 6b: Semi-Durable/Income

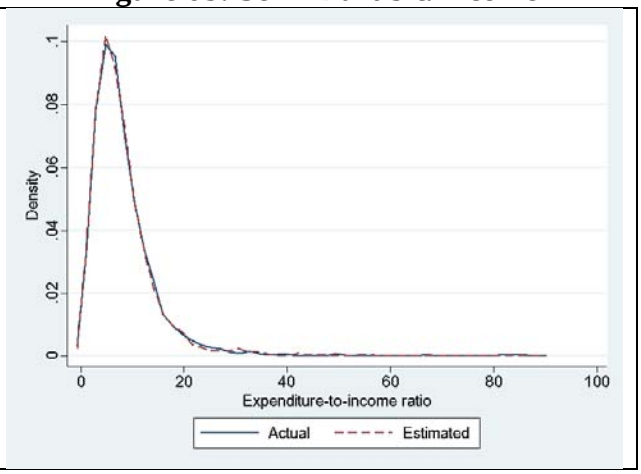


Figure 6c: Non-Durable/Income

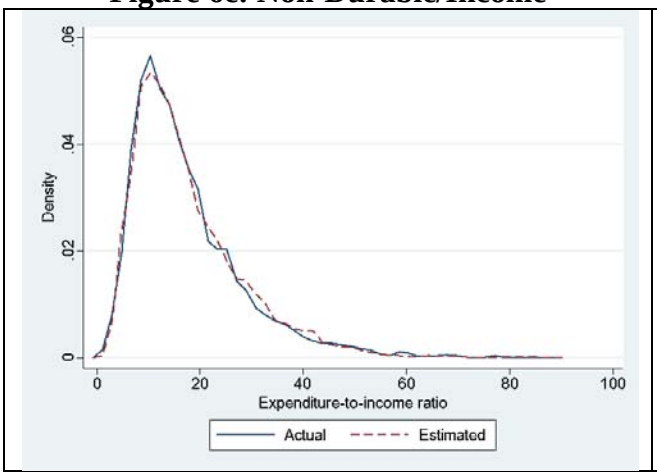


Figure 6d: Services/Income

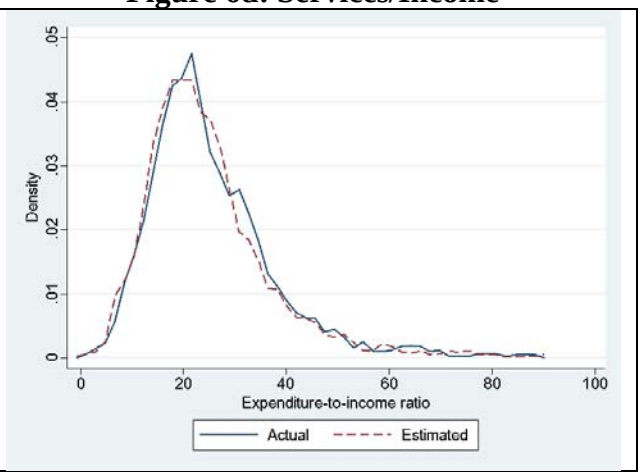


Table 5: Test results

	p-Value for H_o = Equality across the two distributions			
	Mean	**Variance**	**Skewness**	**Kurtosis**
Durables/Income	0.90	0.96	0.83	0.96
Non-Durables/Income	0.99	0.28	0.22	0.54
Semi-Durables/Income	0.88	0.99	0.64	0.94
Service/Income	0.82	0.95	0.65	0.66

[16] The kernel distributions in figures 6a-d are based on data for 2005 only and use a smoothing factor of 0.75. Expenditure-to-income ratios are expressed in percent in the figures.

The results depicted above suggest that the categorical matching method adopted in our work does a good job in estimating actual consumption-to-income ratios using demographic data as a guide.

5.4 Caveats about the merging process and limitations of the data

A number of important caveats should be noted about our categorical matching procedure and the analysis presented in sections 5.1-5.3 (above).

First, as noted in section 2.3, categorical matching is a second-best approach to data combination given constraints. Second, our data combination method involves a number of subjective decisions including the criterion used to assess whether the bin sizes are appropriate, decision rules that are applied if the number of CFM records are not the same as the number of SHS records in the matched bin, etc. These subjective choices may have an impact on the results of the merge.

Third, there are differences in the way that the CFM and SHS surveys capture information, which may introduce noise in the final merged database. In particular, CFM collects information on household income before taxes as income ranges. The estimated income for a given CFM household is the mid-point of the income range. SHS collects income information as a point estimate, not a range. Furthermore, there is some ambiguity about whether a household (i.e. a record in a dataset) is defined similarly across the CFM and SHS datasets.

Finally, the kernel distributions shown in this paper involve a filtering of the raw data to exclude records that may be contaminated with erroneous data. For example, records with an expenditure-to-income ratio of less than zero or greater than one are excluded from our graphs. Authors using the merged data would have to make their own judgements about how to filter the data.

6.0 Conclusions and future work

This paper details a framework for creating a micro-dataset for Canadian households containing both financial and real information. These data are important inputs into examining real-financial linkages in the Canadian economy. Amalgamated real and financial household micro data did not exist in the past, and thus this paper contributes to the literature by filling this data gap. There are a number of caveats with respect to the merge procedure used in the paper; however, they do not detract from the fact that we present a unique cross-sectional/time-series Canadian household dataset containing both balance-sheet data as well as expenditure data.

Future work will focus on using this combined dataset to examine the links between house prices, borrowing and household spending.

Bibliography

Adler, H. 1974. “Creation of a synthetic data set by linking records of the Canadian survey of consumer finance with the family expenditure survey.,” *Annal of Economic and Social Measurement*, 3(2).

Angrist, J.D., and A.B. Krueger. 1992. “The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples”. *Journal of the American Statistical Association* 87, 328–336.

Attanasio, O., L. Blow, R. Hamilton, and A. Leicester. 2005. "Consumption, House Prices and Expectations." Bank of England Working Paper No. 271.

Benito, A. and H. Mumtaz. 2005. "Consumption Excess Sensitivity, Liquidity Constraints and the Role of Housing." Bank of England.

Bernanke, B., M. Gertler, and S. Gilchrist. 1999. "The Financial Accelerator in a Quantitative Business Cycle Framework." NBER Working Paper No. 6455.

Bordt, M., G. Cameron, S. Gribble, B. Murphy, G. Rowe and M. Wolfson. 1990. "The Social Policy Simulation Database and Model: An Integrated Tool For Tax/Transfer Policy Analysis", *The Canadian Tax Journal*, January/February 1990.

Campbell, J.Y. and J.F. Cocco. 2005. "How do House Prices Affect Consumption? Evidence from Micro Data." NBER Working Paper No. 11534.

Carroll, C.D., and D.N. Weil. 1994. “Saving and growth: A reinterpretation”. *Carnegie–Rochester Conference Series on Public Policy* 40, 133–191.

Currie, J., and A. Yelowitz. 2000. “Are public housing projects good for kids?” *Journal of Public Economics* 75, 99–124.

Dee, T.S., and W.N. Evans. 2003. “Teen drinking and educational attainment: Evidence from Two-Sample Instrumental Variables (TSIV) estimates”. *Journal of Labor Economics* 21, 178–209.

Dey, S., R. Djoudad, and Y. Terajima. 2008. “A Tool for Assessing Financial Vulnerabilities in the Household Sector.” *Bank of Canada Review*, Summer 2008

Faruqui, U. 2008. “Indebtedness and the Household Financial Health: An Examination of the Canadian Debt Service Ratio Distribution.”. Bank of Canada Working Paper 2008-46

Klevmarken, W.A. 1982. “Missing variables and two-stage least-squares estimation from more than one data set.” *Proceedings of the American Statistical Association.*

Okner, B. 1972. “Construction a new data base from existing micro data set: the 1966 merge file,” *Annal of Economic and Social Measurement*, 1(3).

Raessler, S. 2002. *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches.* Springer, New York.

Ridder, G, and R. Moffitt. 2007. “The econometrics of data combination.” In *Handbook of Econometrics*. Eds. J.J. Heckman and E.E. Leamer, Ch 75.

Ruggles, N., Ruggles, R. 1974. “A strategy for merging and matching microdata sets”. *Annals of Economic and Social Measurement* 3, 353–371.

Sims, C. A. 1972. “Comments on Okner (1972)”. *Annal of Economic and Social Measurement*, 1(3).

Sutherland, H., R. Taylor and J. Gomulka. 2001. “Combining household income and expenditure data in policy simulations.” University of Cambridge, Department of Applied Economics. Working paper No. 0110.

Appendix 1: Demographic characteristics of the CFM and SHS surveys

Chart A1:

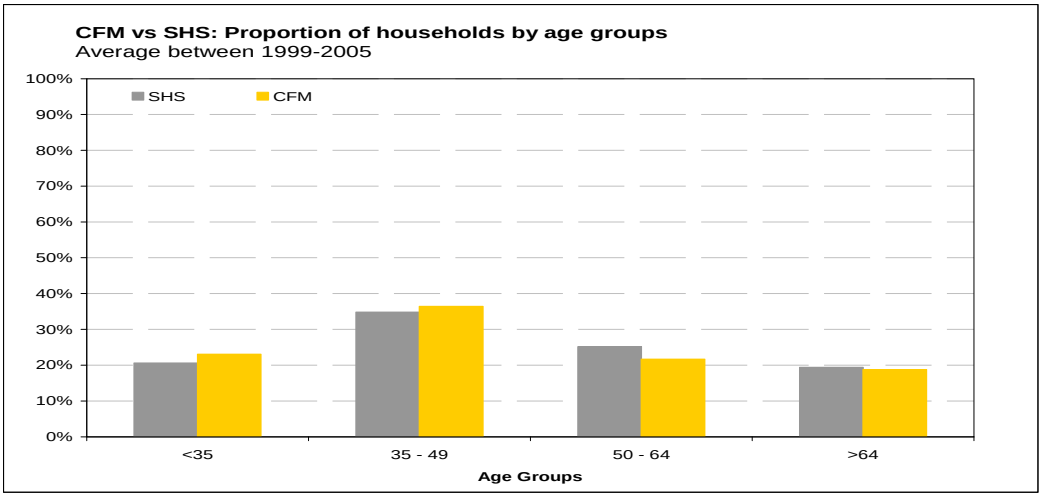


Chart A2:

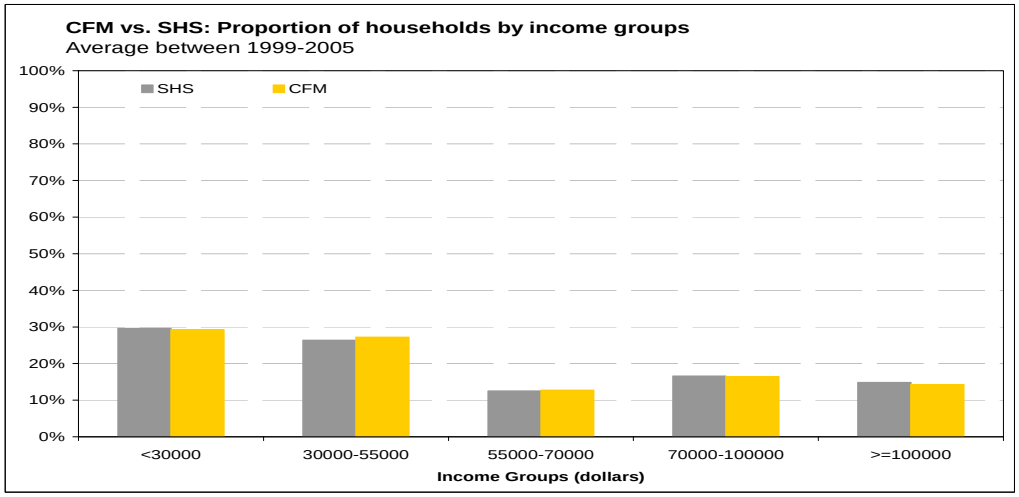
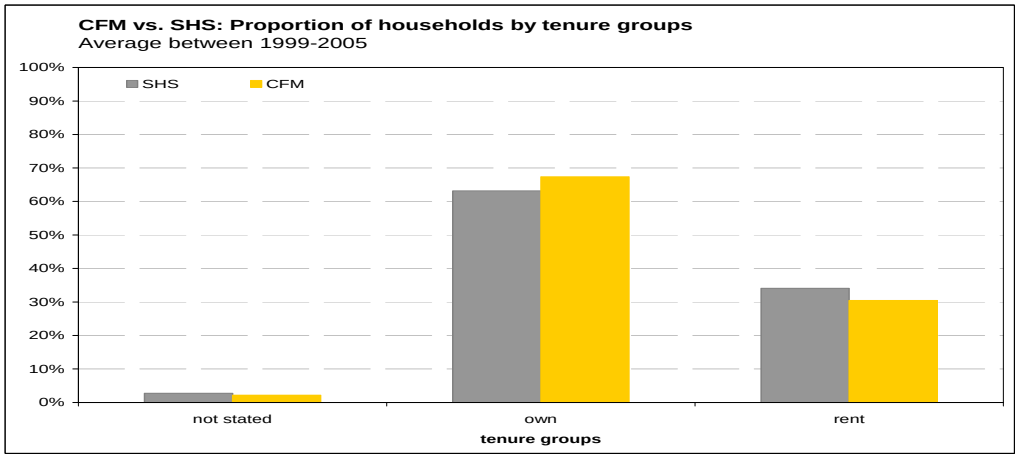


Chart A3:



Appendix 2: Common groupings for key demographic variables in CFM and SHS

Demographic variable	Original variable name from CFM	Original variable name from SHS	New Common Grouping
Total household income	L12Q10	HHINCTOT	1==under $15000 2==15000-19999 3==20000-24999 4==25000-29999 5==30000-34999 6==35000-44999 7==45000-54999 8==55000-59999 9==60000-69999 0==70000+
Household size	HLDSIZE	HHSZTOT	1==1 person 2==2 people 3==3 people 4==4 people 5==5 people 6== 6+ people
Marital status	MARRY	RPMARP	1==married or common law 2==never married 3==other (widowed/ divorced/ separated)
Age (of household head)	HLDCOMP	RPAGEGRP	1== less than 35 2==35-50 3==50-64 4==65+
Housing tenure	L12Q3	TENURYRP	1==own 2==rent 0==not specified / mixed tenure
Area of residence	REGION	PROVINCP	1==BC, AB, SK, MB 2==ONT, QC 3==NB, PEI, NS, NFLD

Appendix 3: SHS variables used to build new consumption variables

VARIABLES	DEFINITIONS
NON DURABLES	
F002	Food purchased from stores
H022	household cleaning supplies
H023	paper, plastic and foil household supplies
H026	garden supplies and services
L103	Health care supplies
L104	Medicinal and pharmaceutical products
L108	Eye-care goods and services
L202	personal care supplies and equipment
N101	tobacco
SEMI DURABLES	
H017	pet expenses
H001 – (H002 + H011 + H016 + H017 + H022 + H023 + H026)	other household supplies
I005	Window coverings and households textiles
I006	art, antiques, and decorative
I010	household equipment
J001	clothing
M102-(M103+M110)	Other recreation equipment and associated services
M201	reading materials
M302TOT	Education supplies and texts
DURABLES	
H004	Purchase of telephone and equipment
I003	Furniture
I004	Rugs, mats, and under padding
K003	purchase of automobiles and trucks
K007	purchase of automotive accessories
M103	Sports equipment
M110	Computer equipment and supplies
M127	Purchase of recreational vehicles
SERVICES excl shelter	
F001-F002-F008	Food, board paid to private household
F008	Food purchased from restaurants
H003-H004	telephone and installation service
H008	cellular service
H009	internet service
H070	online services *(effective in surveys after 2003)*
H010	postal and other communication services
H011	child care expenses
H016	domestic and other custodial services

Appendix 3: SHS variables used to build new consumption variables (continued…)

VARIABLES	*DEFINITIONS*
SERVICES excl shelter (continued)	
I042	maintenance and repairs of furniture and equipment
I046	services related to furnishing and equipment
K008	Rented and leased vehicles
K019	Operation of owned and leased vehicles
K031	public transportation
L107	physicians' care
L112	dental services
L114	hospital care
L116	health care practitioners
L117	other medical services
L118	health insurance premiums
L207	Personal care services
M139	Operation of recreational vehicles
M148	home entertainment equipment and services
M159	recreation services
M301-M302TOT	education services
N201	games of chance
O101	miscellaneous expenditures
SHELTER	
G001	shelter