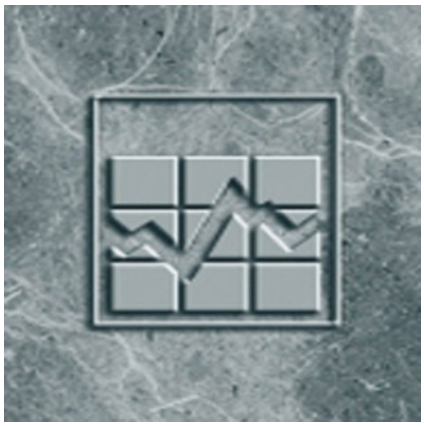
Catalogue no. 75F0002MIE — No. 010
ISSN: 1707-2840
ISBN: 0-662-41064-5

Research Paper

Income Research Paper Series

# General Housing Imputation (excluding utilities) in the Survey of Labour and Income Dynamics (SLID)



by Georgina House

Income Statistics Division and Social Survey Methods Division
Jean Talon Building, Ottawa, K1A 0T6

Telephone: 613 951-7355

Statistics Canada Statistique Canada

Canada

## How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Income Statistics Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: (613) 951-7355; (888) 297-7355; income@statcan.ca).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

| | |
|---|---|
| National inquiries line | 1 800 263-1136 |
| National telecommunications device for the hearing impaired | 1 800 363-7629 |
| Depository Services Program inquiries | 1 800 700-1033 |
| Fax line for Depository Services Program | 1 800 889-9734 |
| E-mail inquiries | infostats@statcan.ca |
| Website | www.statcan.ca |

## Information to access the product

This product, catalogue no. 75F0002MIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Income Statistics Division

Income research paper series

# General Housing Imputation (excluding utilities) in the Survey of Labour and Income Dynamics (SLID)

Published by authority of the Minister responsible for Statistics Canada



July 2005

Catalogue no. 75F0002MIE, Vol. 10

Frequency: Occasional

ISSN: 1707-2840

ISBN: 0-662-41064-5

Ottawa

La version française de cette publication est disponible sur demande (nº 75F0002MIF au catalogue).

---

**Note of appreciation**

*Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.*

# Table of Contents

# Introduction

For some time, Canada Mortgage and Housing Corporation (CMHC[1]) has used data on housing characteristics and housing-related expenditures from the Census of Population. Although the Census data source serves CMHC's purposes to a large extent, the federal government agency turned to the annual household surveys of Statistics Canada to provide information on a more frequent basis. This would allow them to have a better picture of annual trends, and perhaps have a greater choice of other characteristics with which to cross housing data on Canadian households. In 2001, CMHC began to sponsor additional content in both the Survey of Labour and Income Dynamics (SLID) and the Survey of Household Spending (SHS), starting with reference year 2002[2].

The Survey of Labour and Income Dynamics (SLID) is a longitudinal survey initiated in 1993. The survey was designed to measure changes in the economic well-being of Canadians as well as the factors affecting these changes. The target population consists of all persons living in Canada with the following exclusions: persons living in Yukon, the Northwest Territories, and Nunavut, persons living on Reserves, persons living in institutions, and military personnel living in barracks.

The SLID sample is comprised of 2 panels. Each panel remains in the survey for 6 consecutive years and a new panel is rotated in every 3 years. In January following the reference year, SLID sample households are contacted by telephone interviewers. Demographic information is collected for every person in the household. Complete survey data are collected for every eligible person in the household over the age of 15. Questions are asked on labour (labour market activity, work experience, jobless spells, and job information), educational attainment and income sources. At the end of the January interview, respondents are informed that they will be contacted again in May when they will be asked to supply data on income as well as certain expense items. However, the respondent may elect to grant permission to Statistics Canada to retrieve all the data required from the T1 tax file, thereby avoiding the necessity of a second interview. Collection of income data is deferred until May so that the respondent will be more familiar with the required data (having just filed an income tax return).

Although originally designed as a longitudinal survey, SLID has always maintained the capability of producing cross-sectional estimates. All persons who are members of selected SLID households in the beginning of the first year of a panel's existence are longitudinal sample persons for SLID. As such, it is these individuals that are followed longitudinally. Any (non-longitudinal) person living in a household with a longitudinal person is referred to as a cohabitant. Cohabitants living with cross-sectionally eligible longitudinal persons will also be cross-sectional sample persons.

1. Canada Mortgage and Housing Corporation (CMHC) is a federal government agency with the mandate to promote: housing construction, repair and modernization; housing affordability and choice; improvements to overall living conditions; the availability of low-cost financing; and, the national well-being of the housing sector.
2. At the time of this report, this sponsorship is continuing every year.

For more information about survey concepts, definitions and design please refer to Statistics Canada publication: "*Survey of Labour and Income Dynamics - A survey overview*", http://www.statcan.ca:8096/bsolc/english/bsolc?catno=75F0011X

To fill CMHC's need for more data, over 20 housing related questions were added to the SLID labour interview. Some questions apply to homeowners and renters, some only to homeowners, and some only to renters. SLID started to collect information on mortgage payments, property taxes, condominium fees, rent payments and what is included in their rent.

Because of non-response to specific questions, imputation of housing related content was introduced in SLID[3]. The purpose of imputation is to replace erroneous or missing data with values that will provide reasonable estimates. Two methods of imputation were used, longitudinal imputation and cross-sectional donor imputation. They are discussed in the next two sections.

## Longitudinal Imputation

Longitudinal imputation takes into account, as much as possible, information from the previous wave. The effectiveness of the imputation is greatly increased when data from consecutive waves are highly correlated.

Longitudinal imputation for the housing variables was performed wherever possible. Longitudinal imputation is preferred, especially when the household has not moved. We assume that if the household did not move residence from the previous year, then the information for this year would likely be the same as that collected in the previous year.

Since reference year 2002 was a new panel year, longitudinal imputation was performed differently depending on whether the household was part of the first wave of the new panel or not. For households not part of the first wave of the new panel, information for three housing variables (dwelling tenure, dwelling type and number of bedrooms) could be obtained from previous years. These were the only three housing variables that were not new to SLID in reference year 2002. (Mortgage on the dwelling was also available from past years, but it was decided not to do longitudinal imputation on this variable, because the mortgage status is more likely to change even if it appears that the address has not changed.) Missing information was imputed for households using information from that same household from previous years if the postal code remained the same. This was considered a good indication that the household had not moved.

The Labour Force Survey (LFS) and the Rent Survey, an LFS supplement, were used as sources of information for longitudinal imputation for those households part of the first

3. SLID has always done imputation of missing values for its income content.

wave of the new panel[4]. Missing information was imputed for households using information from that same household from the LFS data if the postal code remained the same. To use the Rent Survey, it was necessary to ensure that the dwelling tenure variable on SLID indicated that the household was still a renter.

Longitudinal imputation was successful in lowering the number of households with missing values for variables which were available from previous SLID years. For others, however, it had little to no impact. The new housing variables for which the only source of information was the rent survey had very few successful longitudinal imputations. The main reason for this was that the Rent Survey was only available for households in panel four who were considered renters. Of this small number, only those with valid data on the Rent Survey and whose postal codes matched were then imputed.

For reference year 2003 longitudinal imputation was done on all housing variables for both panels. Missing information was imputed for households using information from that same household from reference year 2002 if the postal code remained the same. If dwelling tenure was available for reference year 2003 then missing information was only imputed if the tenure also matched.

Table 1 contains the weighted percentage of missing values for each variable. Where longitudinal imputation was possible the weighted percentage of missing values after longitudinal imputation is also included. Since longitudinal imputation was only available for certain variables in 2002 some values are missing. Some variables apply only to certain households (i.e. homeowners and renters, only homeowners, and only renters); as a result many have different denominators. Those who require a value other than 'not applicable' to be imputed make up these different denominators. For variable definitions see appendix 1.

4. SLID uses households rotating out of the LFS in December SLID reference year minus one and January SLID reference year.

**Table 1- Impact of longitudinal imputation for 2002 and 2003**

| Variable | Group of households that require imputation for this variable | Missing Values (%) | | | | Number of households longitudinally imputed | |
|---|---|---|---|---|---|---|---|
| | | Before | | After | | | |
| | | 2002 | 2003 | 2002 | 2003 | 2002 | 2003 |
| dwldet25 | Owners and Renters | 6.9 | 7.4 | 0.6 | 0.5 | 1646 | 1871 |
| dwtenr25 | Owners and Renters | 6.6 | 7.4 | 0.4 | 0.5 | 1637 | 1908 |
| dwltyp25 | Owners and Renters | 6.9 | 7.4 | 0.6 | 0.5 | 1646 | 1871 |
| rooms25 | Owners and Renters | 6.9 | 8.0 | 2.7 | 0.6 | 1215 | 2031 |
| opbu25 | Owners and Renters | 8.1 | 9.4 | 8.0 | 0.8 | 9 | 2649 |
| rnre25 | Renters | 8.8 | 10.8 | 8.7 | 0.9 | 7 | 798 |
| rnpk25 | Rent payers | 8.0 | 9.5 | 7.9 | 0.6 | 6 | 689 |
| rnht25 | Rent payers | 8.0 | 9.5 | 7.9 | 0.6 | 4 | 689 |
| rnwa25 | Rent payers | 8.0 | 9.5 | 7.9 | 0.6 | 5 | 689 |
| rnec25 | Rent payers | 8.0 | 9.5 | 7.9 | 0.6 | 2 | 689 |
| rntv25 | Rent payers | 8.0 | 9.5 | 7.9 | 0.6 | 1 | 689 |
| rnfg25 | Rent payers | 8.0 | 9.5 | 7.9 | 0.6 | 5 | 689 |
| rnst25 | Rent payers | 8.0 | 9.5 | 7.9 | 0.6 | 6 | 689 |
| rnwd25 | Rent payers | 8.0 | 9.5 | 7.9 | 0.6 | 3 | 689 |
| rnfu25 | Rent payers | 8.0 | 9.5 | 8.0 | 0.6 | 0 | 689 |
| rnno25 | Rent payers | 8.0 | 9.5 | 8.0 | 0.6 | 0 | 689 |
| rentm25 | Renters | 12.1 | 14.4 | 10.2 | 2.6 | 109 | 946 |
| repa25 | Owners and Renters | 7.0 | 7.8 | n/a[2] | 0.6 | n/a | 1987 |
| heat25 | Owners and Renters | 11.6 | 12.8 | n/a | 2.3 | n/a | 2703 |
| heatg25 | Owners and Renters | 11.6 | 12.8 | n/a | 2.3 | n/a | 2703 |
| opfm25 | Owners and Renters | 3.4[1] | 7.0 | n/a | 0.7 | n/a | 1716 |
| mortg25 | Owners | 7.7 | 8.1 | n/a | 0.8 | n/a | 1426 |
| cond25 | Owners | 6.5 | 7.1 | n/a | 0.6 | n/a | 1267 |
| mortgn25 | Mortgage payers | 9.5 | 10.3 | n/a | 1.9 | n/a | 861 |
| mortgm25 | Mortgage payers | 29.4 | 29.5 | n/a | 8.6 | n/a | 2297 |
| prtxm25 | Owners | 25.7 | 26.7 | n/a | 5.4 | n/a | 4546 |
| condm25 | Condominium members | 13.8 | 15.7 | n/a | 5.4 | n/a | 95 |
| rnbs25 | Renters | 9.6 | 11.3 | n/a | 1.2 | n/a | 834 |

1. In reference year 2002 only households (outside of New Brunswick) who have a postal code indicating rural (second digit of postal code= 0) were asked if the dwelling operated a farm on the property (OPFM25), all others are set to 'not applicable'. The true non-response rate is 16.6% with a denominator of 11,018.

2. Longitudinal imputation not available in 2002.

## Cross-sectional Donor Imputation

Donor imputation consists of identifying a group of households sharing several characteristics with the household to be imputed, then selecting one of them to be the donor. The value reported by the donor replaces the missing field of the imputed household. For example, SLID imputed the income of an individual by picking the income of a person living in the same province and with the same age range, sex, level of education, type of job (employee, self-employed) and occupation.

Donor imputation was performed for the remaining variables and for all those households where longitudinal imputation was not available. Donor imputation was broken down into seven modules. This was done in order to facilitate the use of imputed variables from one module as auxiliary or matching variables in another. Imputed records may also act as donors in successive rounds of imputation.

Each module has different lists of auxiliary and matching variables. Auxiliary variables are used to create imputation groups[5], whereas matching variables are used in a score function to find the best suited donor within an imputation group. In some cases, auxiliary variables created receiver groups[6] with no matching donor groups[7]; when this was the case, collapsing of auxiliary variables was performed. For example, for all but one module, the province variable was collapsed into five regions. When collapsing was not sufficient enough to create possible donors for all receivers, auxiliary variables were then turned into matching variables. Within each module, variables on the "variables to be imputed" list were also used as matching variables when a value for that variable was available. For example, in the Dwelling type and tenure module, if dwelling type was missing and dwelling tenure was not, then it was attempted to find a donor with a matching value of dwelling tenure. Within each imputation group, donors were evaluated based on a score function. This score function was set up based on the score function used in SLID income imputation:

$$s(X,Y) = \sum_{k=1}^{K} p_k \, \mathrm{I}(X_k, Y_k), \qquad \text{where } \mathrm{I}(X_k, Y_k) = \begin{cases} 1 & \text{if } X_k = Y_k \\ 0 & \text{if not.} \end{cases}$$

Note that $p_k$ is a weight allowing us to assign more or less importance to the matching variable k. It was decided that all matching variables are of equal importance, hence $p_k=1$. $X_k$ is the value of the receiver's variable k and $Y_k$ is the value of the donor's variable k.

5. Both donors and receivers (records requiring imputation) are separated into imputation groups. Every receiver within an imputation group is compared to every donor within the same imputation group.
6. Receiver groups are the groups, based on auxiliary variables, in each module of records requiring imputation.
7. Donor groups are the groups, based on auxiliary variables, in each module of possible donor records.

A donor is then selected based on this score. The donor with the highest score is selected. If there are several donors with the high score value then a donor is randomly chosen out of this list. For those modules where there are no matching variables a donor is selected randomly from the entire imputation group.

The following is a breakdown of each module. Each table contains a list of variables to be imputed along with lists of auxiliary and matching variables. It was decided that for modules in which imputation of numeric variables was required, those that fell in the top one percent would be excluded from the donor pool. This occurs in four different modules; the Owners and renters module, the Mortgage module, the Property tax and condominium charges module and finally the Rent module. One reason for this was to lessen the chance of imputing a value higher than household income for dollar amounts such as rent or mortgage.

### Dwelling type and tenure module

The Dwelling type and tenure module is the first module to run. This is done because every proceeding module uses either dwelling type or dwelling tenure as an auxiliary or matching variable. Note that the number of codes is in brackets.

| | |
|---|---|
| Records entering module: | All households |
| Variables to be imputed: | Detailed dwelling type, Type of dwelling, Dwelling tenure |
| Auxiliary variables: | Province (10), Urban size group (3), Number of persons in the household as at December 31 (3) |

### Owners and renters module

The Owners and renters module is the next to run. Variables to be imputed in this module are variables for which all records require a value other than ‘not applicable’. When choosing possible donors in this module, all those who had a value of 9 or higher for the number of bedrooms were excluded.

| | |
|---|---|
| Records entering module: | All households |
| Variables to be imputed: | Number of bedrooms, Dwelling repairs needed, Principal heating fuel, Principal heating fuel group, Household operates farm, Household operates business |
| Auxiliary variables: | Province (5), Urban size group (3), Type of dwelling group (3), Number of persons in the household as at December 31 (3) |
| Matching variables: | Ownership of dwelling (2) |

### Owners module

The next module to be run is the Owners module. This module uses imputed variables from both previous modules as auxiliary variables. All households who have a value of dwelling tenure indicating ownership of the dwelling enter into this module. All other records with missing values are set to 'not applicable'.

| | |
|---|---|
| Records entering module: | Owners |
| Variables to be imputed: | Mortgage on the dwelling, More than one mortgage on the dwelling, Dwelling is part of condominium development |
| Auxiliary variables: | Province (5), Urban size group (3), Type of dwelling group (2), Number of bedrooms (4) |
| Matching variables: | Number of persons at December 31 (3) |

### Mortgage module

The Mortgage module follows the Owners module. This is done to ensure that only households who have a mortgage on the dwelling (MORTG25) enter into this module. All others with missing values are set to 'not applicable'. When choosing possible donors in this module, all those who had a value of monthly amount of regular mortgage payment higher than $3,000 were excluded.

| | |
|---|---|
| Records entering module: | Owners with a mortgage on the dwelling |
| Variables to be imputed: | Monthly amount of regular mortgage payments |
| Auxiliary variables: | Province (5), Type of dwelling (2), Number of bedrooms (4), More than one mortgage on the dwelling (2) |

### Property taxes and condominium charges module

The Property taxes and condominium charges module can be run in succession with the Mortgage module. All homeowners enter into this module. All others with missing values are set to 'not applicable'. Since the variable, dwelling is part of a condominium complex (COND25), is one of the matching variables, homeowners who are not a part of a condominium complex will have their monthly condominium charges imputed to 'not applicable'. When choosing possible donors in this module, all those who had a monthly amount of property taxes higher than $500 or a value of monthly condominium charges higher than $800 were excluded.

| | |
|---|---|
| Records entering module: | Owner |
| Variables to be imputed: | Monthly amount of property taxes, Monthly condominium charges |

| | |
|---|---|
| Auxiliary variables: | Province (5), Type of dwelling group-owners (2), Dwelling is part of a condominium development (2) |
| Matching variables: | Urban size group (3), Number of bedrooms (4) |

### Renters module

The Renters module must follow the Owners and renters module. All households who have a value of dwelling tenure indicating that they are a renter enter into this module. All others with missing values are set to 'not applicable'. Those who are renters, but do not pay a monthly rent (RENTM25=0) have variables such as rent includes electricity, parking, etc. are set to 'not applicable'.

| | |
|---|---|
| Records entering module: | Renters |
| Variables to be imputed: | … included in rent (several items), Rent includes fully or partially furnished, Reduced rent reason, Rent calculated on the basis of income |
| Auxiliary variables: | Province (5), Urban size group (3), Type of dwelling group-renters (3) |
| Matching variables: | Number of bedrooms (4) |

### Rent module

Finally, the last module is the Rent module. All households who pay a monthly rent enter into this module. All other households with missing values are set to 'not applicable'. When choosing possible donors in this module, all those who had a monthly amount of rent higher than $1,500 were excluded.

| | |
|---|---|
| Records entering module: | Renters |
| Variables to be imputed: | Monthly amount of rent |
| Auxiliary variables: | Province (5), Type of dwelling group–renters (3), Number of bedrooms (4), Reduced rent reason(2) |
| Matching variables: | …included in rent (several items), Number of bedrooms (4) |

There was a wide range of percentages of missing values prior to donor imputation for reference year 2002. Certain variables have a relatively high percentage of missing values resulting in small donor pools. Table 1 contains the weighted percentage of missing values before cross-sectional imputation for both reference years 2002 and 2003.

# Evaluation of the Imputation Methods for reference year 2002

To evaluate the success of the imputation methods, before and after frequencies for categorical variables and before and after means for continuous variables were examined. Changes of less than 1% in before and after percentages and a less than 6% change in before and after means were considered successful. A cross validation analysis was also performed by using 1,000 records with complete data. These records were imputed. Then, the original data was compared with the imputed data.

After imputation was completed using the methods described above, several analyses were done. These analyses were performed in order to ascertain whether or not the resulting data was similar to the original data. Before and after frequencies were taken on all imputed categorical variables. As an example, table 2 contains the before and after percentages for dwelling type. Prior to donor imputation, 4% of the households were missing this variable. Table 2 shows that the before and after percentages are quite similar. Any difference that does exist is less than 1% which is very reasonable.

**Table 2: Percentage before and after donor imputation for dwelling type**

| DWLDET25 Code set | Before donor imputation (excluding missing values in the denominator) | After donor imputation |
|---|---|---|
| 01 Single detached | 66.5 | 65.9 |
| 02 Double | 3.53 | 3.50 |
| 03 Row or terrace | 4.13 | 4.14 |
| 04 Duplex | 4.02 | 4.02 |
| 05 Low-rise apartment | 14.0 | 14.4 |
| 06 High-rise apartment | 5.01 | 5.17 |
| 07 Hotel, rooming house, camp | 0.26 | 0.27 |
| 08 Mobile Home | 2.42 | 2.42 |
| 09 Other | 0.15 | 0.17 |

It is important to also take a close look at variables for which the percentage of missing values is much higher. As an example, table 3 contains the before and after percentages for principle heating source. Prior to donor imputation, 24% of the households were missing this variable. Once again, however, table 3 shows that the difference in percentages is always less than one percent. These results are typical of what was seen with all categorical variables

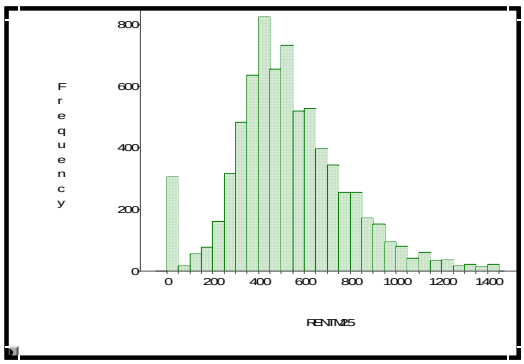
**Table 3: Percentage before and after donor imputation for principle heating source**

| HEAT25 Code set | Percentage before donor imputation (excluding missing values in the denominator) | Percentage after donor imputation |
|---|---|---|
| 01 Oil or other liquid fuel | 17.1 | 16.7 |
| 02 Piped gas (natural gas) | 41.0 | 41.8 |
| 03 Bottled gas (propane) | 0.90 | 0.88 |
| 04 Electricity | 33.0 | 33.5 |
| 05 Wood | 7.39 | 7.07 |
| 06 Other | 0.06 | 0.06 |

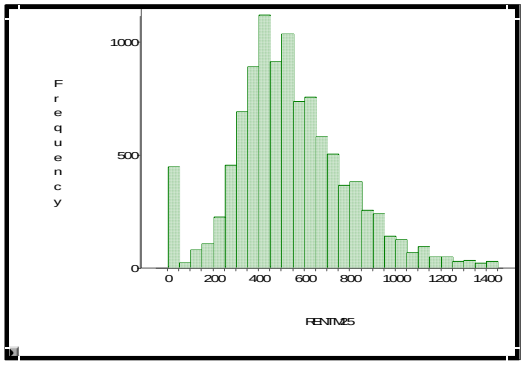
When looking at continuous variables, the before and after means were studied. This gave a good indication as to whether or not the data has been changed significantly. As an example of this, look at the before and after means of amount of monthly rent (RENTM25) and property taxes (PRTXM25). The mean prior to imputation of RENTM25 was $604.12 with 29% of households missing this variable; after imputation, the mean has dropped slightly to $589.53. This is a 2% reduction. The mean prior to imputation of PRTXM25 was $160.68 with 35% of households missing this variable; after imputation, the mean has again dropped slightly to $154.53. This is a 4% reduction. These results are characteristic of all continuous variables. The reductions range from 2 to 6 percent. A drop in the mean is to be expected, largely due to the fact that the top one percent of the donor pool was excluded.

A visual check was also performed. Continuous variables were plotted in a histogram to compare the donor pool to the newly imputed file. In order to make the comparison simpler, the top 1% of the data from the histogram of the newly imputed file was excluded. The histograms in graphs 1 and 2 show that the distribution before and after imputation is very similar. Histograms for all continuous variables were compared in this way and the example chosen is typical of what was seen. This is a good indication that the imputation methods chosen were successful in that they did not alter the data to any great extent.

**Graph 1: Donor pool histogram for RENTM25**




**Graph 2: File after imputation (with RENTM25<1500) histogram for RENTM25**



As an added analysis, a cross validation was done. A cross validation is a comparison on households for which we have complete data between values imputed using the methods described above and reported values. A cross validation is useful because it simulates the imputation process in a way that makes it easy to test its accuracy. A set of 1,000 records for which complete data was available was imputed using the methods described above. The original data for these records was compared with the newly imputed data. This cross validation comparison was only done on categorical variables. The data was imputed with an incorrect value between 10 and 40 percent of the time. This cross validation was run several times and the results below are characteristic of what was seen.

When looking at the results of the cross validation, one must keep in mind the fact that the imputation is done in sequential modules. The Dwelling type and tenure module is run first, thus the 11% of households with incorrect dwelling tenure will then be incorrect for most of the other modules. For example, those who are considered owners are the only ones that would be imputed a value for mortgage on the dwelling (MORTG25) all others would be given a value of not applicable.

For comparison purposes, random imputations were performed on the same 1,000 households. Values were randomly assigned to each variable based on the initial frequencies.

Since imputation is done sequentially, the imputation of one module is done before the imputation of another module. This generates two ways to proceed with the analysis. The imputation could have used the values created during the cross validation from the previous file, or could have used the new values that were randomly assigned from the previous file. In order to maintain a logical analysis of the variables being imputed, we use the cross validation data. In doing so respondents who were originally renters stayed renters, which meant that variables applicable to renters, like rent includes heat (RNHT25) would have a valid imputed value rather than a ‘not applicable’ which would happen if their randomly imputed value in a previous module changed them to an owner. Therefore those who are considered incorrectly imputed by the random method are based on the same subset of the population as those in the cross validation.

From these results, it can be seen that there is a difference between variables for which longitudinal imputation was successful and variables for which it was not. For example, random imputation was performed on DWTENR25. This was done several times and typically the data was imputed incorrectly 40% of the time, which is much higher than when the imputation methods discussed in this paper are used. As an example of a random imputation on a variable where longitudinal imputation was not available, look at the variable MORTG25. The results show that the imputation method produces only a slight improvement over the random imputation method. Typically, the incorrect value was imputed 43% of the time. It would appear from these results that the longitudinal imputation method is much better than random imputation, but that the cross-sectional donor imputation method is only slightly better. Table 4 is a table of the percentage

difference between the original data and the imputed data for the cross validation along with typical results of the random imputations.

**Table 4: Percentage of incorrectly imputed categories for two imputation strategies**

| Variable | Cross sectional imputation (%) | Random imputation (%) |
|---|---|---|
| Dwtenr25 | 11 | 40 |
| Dwldet25 | 20 | 54 |
| Mortg25 | 40 | 43 |
| Repa25 | 35 | 36 |
| Heat25 | 39 | 67 |
| Heatg25 | 39 | 67 |
| Obpu25 | 16 | 16 |
| Mortgn25 | 38 | 38 |
| Cond25 | 13 | 17 |
| Rnpk25 | 17 | 18 |
| Rnwa25 | 16 | 19 |
| Rnec25 | 19 | 21 |
| Rnht25 | 20 | 22 |
| Rntv25 | 15 | 15 |
| Rnfg25 | 16 | 20 |
| Rnst25 | 16 | 21 |
| Rnwd25 | 16 | 17 |
| Rnfu25 | 13 | 13 |
| Rnno25 | 12 | 12 |
| Rnre25 | 17 | 18 |
| Rnbs25 | 17 | 19 |

# Conclusion

Based on the initially high item non-response for housing variables, it was useful to put in place an imputation methodology. One of the strengths of the described imputation procedure is that the end result is a dataset with no missing values and internally consistent imputed records.

From these analyses, one can see that the imputation methods chosen were quite successful. Before and after frequencies for categorical variables and before and after means for continuous variables show only relatively small changes less than 1% and 6% respectively. As an added confirmation, before and after histograms are very similar in appearance. This was a confirmation that the imputation procedure did a good job in preserving univariate distributions.

A more challenging question is to what degree multivariate distributions were altered. To address this question, we looked at estimates for specific groups. For example, since age is not explicitly taken into account in the imputation, an existing correlation between age

and dwelling tenure may be lost. Analyses performed on the data in reference to the age of the oldest household member indicate that the imputation process does not affect the estimates of home owners based on age. The distribution of the age of the oldest member of the household of home owners is very similar before and after imputation. The average age of the person with the highest net income in households with a monthly mortgage for imputed versus non-imputed households is also quite similar. This is yet more evidence to suggest that the imputation method is quite good. Since it is impossible to test all variables for which there may be a correlation, one can't be sure that all estimates for specific groups are not affected by what appears to be, according to the cross validation analysis, a high error rate.

The bottom line is that, although the measurement error rate for some variables is high due to weakness in the underlying imputation models, the impact on univariate and some multivariate distributions seems to be quite limited. In other words, although initial non-response rates were high, the missing values were replaced without causing great changes to the overall estimates.

# Appendix 1

## Variable list

| Variable | Variable Description | Module | Group of households that require imputation for this variable |
|---|---|---|---|
| dwldet25 | Detailed type of dwelling | Dwelling type/Tenure | Owners and renters |
| dwtenr25 | Ownership of dwelling | Dwelling type/Tenure | Owners and renters |
| dwltyp25 | Dwelling type | Dwelling type/Tenure | Owners and renters |
| rooms25 | Number of bedrooms | Owners/renters | Owners and renters |
| opbu25 | Business on this property | Owners/renters | Owners and renters |
| repa25 | Repairs needed | Owners/renters | Owners and renters |
| heat25 | Principal heating fuel | Owners/renters | Owners and renters |
| heatg25 | Principal heating fuel group | Owners/renters | Owners and renters |
| opfm25 | Farm on this property | Owners/renters | Owners and renters |
| mortg25 | Mortgage on the dwelling | Owners | Owners |
| cond25 | Dwelling is a condominium | Owners | Owners |
| mortgn25 | More than one mortgage | Owners | Mortgage payers |
| mortgm25 | Monthly mortgage payment | Mortgage | Mortgage payers |
| prtxm25 | Monthly property taxes | Property-tax and condo | Owners |
| condm25 | Monthly condominium fee | Property-tax and condo | Condominium members |
| rnre25 | Reduced rent and reason | Renters | Renters |
| rnpk25 | Parking included in rent | Renters | Rent payers |
| rnht25 | Heat included in rent | Renters | Rent payers |
| rnwa25 | Water included in rent | Renters | Rent payers |
| rnec25 | Electricity included in rent | Renters | Rent payers |
| rntv25 | Cable TV included in rent | Renters | Rent payers |
| rnfg25 | Refrigerator included in rent | Renters | Rent payers |
| rnst25 | Stove included in rent | Renters | Rent payers |
| rnwd25 | Washer and dryer included in rent | Renters | Rent payers |
| rnfu25 | Furniture included in rent | Renters | Rent payers |
| rnno25 | No amenities included in rent | Renters | Rent payers |
| rnbs25 | Rent calculated on basis of income | Renters | Renters |
| rentm25 | Monthly rent | Rent | Renters |