Journal of Official Statistics, Vol. 27, No. 4, 2011, pp. 693–715

Evaluation of Alternative Income Imputation Methods for a Longitudinal Survey

Nicole Watson[1] *and Rosslyn Starick*[2]

This article evaluates various methods for imputing income in a household-based longitudinal survey. Eight longitudinal methods are evaluated in a simulation study using five waves of data from the Household, Income and Labour Dynamics in Australia (HILDA) Survey. The quality of the imputed data is evaluated by considering eleven criteria that measure the predictive, distributional and estimation accuracy of the cross-sectional estimates and the estimates of change between waves. Many of the imputation methods perform well cross-sectionally, but when the methods are placed in a longitudinal context their strengths and weaknesses are more apparent. The method that combines the Little and Su method with the population carryover method performs the best overall.

Key words: Carryover method; HILDA Survey; imputation evaluation framework; Little and Su method; longitudinal hot deck method; longitudinal nearest neighbour regression method.

1. Introduction

All large-scale surveys face the unavoidable problem of missing data. For longitudinal surveys there are three types of nonresponse: i) some individuals do not provide complete answers to all questions, either because they do not know or because they refuse, resulting in *item* nonresponse; ii) some do not provide an interview in every wave, resulting in *wave* nonresponse; and iii) some do not provide an interview in any wave, resulting in *unit* nonresponse. As nonrespondents may have different characteristics to respondents, restricting analyses to complete cases may result in biased population or regression estimates. To counter such problems, a combination of weighting and imputation is typically used (Lepkowski 1989; Nordholt 1998; Kalton 1986; Kalton and Brick 2000; Dillman et al. 2002).

[1] Melbourne Institute of Applied Economic and Social Research, University of Melbourne VIC 3010, Australia. Email: n.watson@unimelb.edu.au
[2] Australian Bureau of Statistics, GPO Box 2796Y, Melbourne VIC 3001, Australia. Email: rosslyn.starick@abs.gov.au
Acknowledgments: This research has been supported, in part, by an Australian Research Council Discovery project grant (#DP1095497). It makes use of unit record data from the Household, Income and Labour Dynamics in Australia (HILDA) Survey. The HILDA Project was initiated and is funded by the Australian Government Department of Families, Housing, Community Services and Indigenous Affairs (FaHCSIA) and is managed by the Melbourne Institute of Applied Economic and Social Research. The findings and views reported in this article, however, are those of the authors and should not be attributed to either FaHCSIA, the Melbourne Institute or the Australian Bureau of Statistics. The authors also thank the members of the HILDA Technical Reference Group (John Henstridge, Frank Yu, Paul Sutcliffe, Stephen Horn, Peter Boal, Robert Breunig and Clinton Hayes) for their comments and suggestions on the design of this evaluation study and earlier reports on the results.

The imputation methods adopted in cross-sectional settings have been used for many years and are reasonably well understood, but less is known about the performance of imputation methods for longitudinal surveys where additional demands are placed on the methods. Which method preserves both the cross-sectional estimates and the estimates of change across waves? Should multiple wave nonresponse be imputed with the same donor or different donors at each wave? How do we best use subsequent and previous wave data in the imputation method? How far into the future or into the past should we go?

The early experience of the Household, Income and Labour Dynamics in Australia (HILDA) Survey, a large nationally-representative longitudinal survey of Australian households, highlights one of the difficulties with imputation in the longitudinal context. (An overview of the HILDA Survey is given by Wooden and Watson (2007).) Watson (2004), following the British Household Panel Survey (BHPS), adopted a nearest neighbour regression method (described later in this article) to impute missing income data in the HILDA Survey. This method led to an overstatement of the change in income between waves, even though the regression models incorporated many respondent characteristics including income reported in other waves.

The aim of this evaluation study is to better understand the strengths and weaknesses of various imputation methods in a longitudinal survey context, and as a result identify a better method for the HILDA Survey. The first five waves of HILDA data are used to construct simulated datasets on which various imputation methods are tested. Eleven evaluation criteria are used to assess the predictive, distribution and estimation accuracy of each method. Eight longitudinal imputation methods are evaluated: one nearest neighbour regression method, two variants of the Little and Su method, three carryover methods, one hotdeck method, and a combined carryover-Little and Su method (all described below). Two cross-sectional methods are tested as the fallback option when the longitudinal methods cannot be used (e.g., when only one wave of data is available and the income component is missing).

There are four design features of the HILDA Survey relevant to the development of the imputation strategy. First, interviews are sought with every adult aged 15 and over within a household and a household roster provides some basic information on all household members. Second, the income module is repeated every wave to gather information about seven income components: wages and salaries; government pensions and benefits; business income; interest income; dividends and royalties; rental income; and private transfers. Total individual and household income are calculated from these components. Third, reimputation occurs at each release to take advantage of information collected in subsequent waves (the weights, data editing and derived variables may also be improved so users have become accustomed to using all waves from the latest release). Fourth, imputation is used to complete the missing income data for person-level item nonresponse and (within a responding household) person-level unit and wave nonresponse. Weights are used to adjust for household-level unit and wave nonresponse.

Our contribution to the literature on imputation methods for longitudinal surveys is threefold; we are testing a wider range of methods via a greater number of criteria across a larger variety of variables than other studies. Previous studies have tended to focus on a single income variable, usually total income, and as a result identified a single method suitable for that variable alone (for example, Tremblay 1994; Quintano et al. 2002;

Laaksonen 2003; Spiess and Goebel 2004; Frick and Grabka 2005). Williams and Bailey (1996) considered four income components but the evaluation criteria were limited. The comparison of results across other studies has been hampered by the varying evaluation criteria used. In contrast, we evaluate the performance of the imputation methods on all income components to understand which methods work "best" for which components and consider why this might be the case.

The remainder of this article is organized as follows: Section 2 describes a methodological evaluation framework – in terms of simulated data and evaluation critera – for assessing the performance of an imputation method; Section 3 details alternative imputation methods; Section 4 compares the performance of these methods in the HILDA Survey; and Section 5 concludes.

2. Evaluation Methodology

2.1. Simulated Data

Ideally we want to compare the imputed value with the value the respondent would have reported if they had not refused or did not know the value. As this is impossible, we simulate a series of datasets with missing values using the first five waves of HILDA data. This is done by taking 8,193 complete cases (i.e., those who provide all income items in each wave for which they are eligible to be interviewed) and setting a portion to missing.

The sample of cases set to missing are selected on the basis of logistic regression models of the response mechanism from the full HILDA dataset that assume the missing values are missing at random (Rubin 1976). That is, the probability that the income component is missing depends on a range of characteristics of the respondent but not on the value of the income component itself. To take account of the dependence of wave nonresponse across waves, the model for each wave includes a response indicator for each prior wave together with some basic characteristics of the respondent (age, sex, labour force status, relationship in the household, place of residence, value of the house, usual rent/mortgage repayments, and whether person has a long term health condition). The dependence of missingness between income components within a wave and across waves is taken into account by sequentially modelling the income components and including a response indicator for each income component in prior waves and the previous income components of that wave together with a range of other characteristics of the respondent. For example, the model to predict missingness of pensions and benefits in Wave 3 includes an indicator for missingness in each of the seven income components (total income is not included) in Waves 1 and 2 along with an indicator for missingness for wages and salaries in Wave 3. The other respondent characteristics are marital status, highest level of education, occupation status, multiple job holder, usual hours worked, whether the person speaks a language other than English, time since school spent working, time since school spent unemployed, several variables relating to the last financial year such as time spent in education, time spent employed, number of jobs held, and the basic respondent characteristics in the wave-level response models described earlier. Only cases reporting a particular source of income are included in the models (to mimic the fact that we almost always know from the filter question that the missing income for an interviewed person is

nonzero). The max-rescaled R^2 for these models are provided in Table 1 and range from 0.07 to 0.50.

Thirty datasets are simulated from the set of all complete cases by using the above predicted probabilities to assign missingness to the various income components. A sample of complete cases are set to contain item, wave and unit nonresponse in line with the proportions observed in the entire HILDA dataset. The missingness is set sequentially through the waves and income components. The predicted probabilities for missingness are adjusted if earlier income components are set to missing so that the dependent nature of missingness is replicated.

The missing data in each simulated dataset is imputed via each imputation method. The true and imputed values are compared via a series of evaluation measures described in the next section. We use the term "true" value to mean what a respondent actually reported before his/her data were set to missing.

Table 2 provides summary measures of the simulated datasets, including the number of cases to be imputed for each income component and various characteristics of interviewed individuals (respondents) reporting an amount for that component. The characteristics provided are the mean, the coefficient of variation, the coefficient of skewness, and the correlation with age. For respondents with item nonresponse, nonzero amounts need to be imputed and thus the characteristics provided in the top half of the table are restricted to respondents reporting nonzero amounts. In contrast, zero or nonzero amounts are acceptable for individuals with unit or wave nonresponse within responding households (referred to here as "nonrespondents"). As a result, the characteristics reported in the second half of the table include respondents who report zero or nonzero amounts. Some income components (such as business income, interest income, dividends and royalties, rental income and private transfers) are highly variable or highly skewed, and thus present a challenge for the imputing process. For business income, there are a large number of respondents that need to be imputed compared to the number reporting an amount for business income. When respondents with zero income for a component are included, the variability and the skewness of the data are much greater.

While the simulated datasets are as realistic as possible, a difference in household size occurred – the average number of adults per household in the simulation datasets is 1.5 compared to 1.9 in the entire HILDA data. This is because nonrespondents in partly responding households (where some but not all adults provided an interview) could not be

Table 1. Max-rescaled R^2 for the models of the response mechanism, Waves 1 to 5

	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5
Respondents					
Wages and salaries	0.205	0.202	0.246	0.265	0.262
Govt pensions and benefits	0.340	0.498	0.428	0.412	0.446
Business income	0.321	0.068	0.366	0.359	0.402
Interest income	0.131	0.182	0.279	0.279	0.314
Dividends and royalties	0.224	0.339	0.381	0.419	0.439
Rent income	0.196	0.245	0.335	0.321	0.359
Private transfers	0.287	0.252	0.292	0.390	0.292
Nonrespondents	0.101	0.398	0.442	0.470	0.467

Table 2. Characteristics of simulated datasets, averaged across Waves 1 to 5

		Respondents reporting an amount				
	Number to be imputed	Number	Mean	Coeff of variation	Coeff of skewness	Corr with age
	Respondents	*Nonzero amounts*				
Wages and salaries	207	3,468	37,329	0.83	3.9	0.24
Govt pensions and benefits	38	2,200	8,477	0.51	0.6	0.36
Business income	75	250	15,734	2.91	−1.3	0.00
Interest income	186	1,159	2,008	2.91	12.5	0.08
Dividends and royalties	154	1,306	3,056	4.19	11.4	0.11
Rent income	44	356	2,144	4.14	1.6	0.19
Private transfers	24	152	4,889	1.13	2.4	0.12
	Nonrespondents	*All amounts (zero and nonzero)*				
Wages and salaries	450	5,988	21,588	1.38	3.4	−0.18
Govt pensions and benefits	450	6,158	3,032	1.59	1.5	0.47
Business income	450	6,120	636	15.38	−1.5	0.00
Interest income	450	6,008	387	6.95	27.0	0.13
Dividends and royalties	450	6,041	648	9.28	24.2	0.07
Rent income	450	6,151	124	17.70	8.8	0.04
Private transfers	450	6,171	120	9.54	15.5	−0.03
Total financial year income	450	5,601	27,956	1.07	4.3	−0.03

included in the simulation datasets as they are not "complete". There were only minor differences between the simulated datasets and the full HILDA datasets for age, sex, labour force status, marital status and education level. It is not expected that this difference will substantially affect the results of this study.

2.2. *Evaluation Criteria*

This section defines the evaluation criteria that provide the framework for comparing the imputation methods. A good imputation method must reproduce key statistical properties of the complete data.

Seven of the eleven criteria used here are based on those proposed by Chambers (2000) for the Euredit Project which evaluated various imputation methods using mainly cross-sectional economic and social survey data from European countries. These seven are Criteria 1, 2, 6, and 8 through 11 (described below). We include four additional criteria to help assess how well the changes between waves (Criteria 3 and 7 below) and the relationships between variables are maintained (Criteria 4 and 5 below). The criteria measure three aspects of imputation accuracy – predictive, distribution and estimation. When undertaking regression analysis, all eleven criteria are important; however, when producing aggregate estimates, only distributional accuracy and estimation accuracy are necessary.

For a longitudinal survey it is important that the imputation method performs well both cross-sectionally and longitudinally. Most of the criteria are applied to both the *level* of income at each wave and the *change* in income between waves (the change between each two successive waves is examined as well as the change between the first and last wave). The exceptions are that Criteria 3 and 7 apply only to the *change* in income between waves, and Criteria 4 and 5 apply only to the *level* of income at each wave.

Apart from Criterion 7, the criteria are defined on the set of n imputed values within a dataset, rather than the set of all values. $\hat{Y}$ denotes the imputed version of variable Y and Y^* denote the true version of the same variable.

2.2.1. Criteria 1 to 5: Predictive Accuracy

The first five criteria assess the predictive accuracy of the imputation by considering how close the imputed value ($\hat{Y}$) is to the true value (Y^*). The imputation method should preserve the true values as far as possible.

The first criterion is the Pearson correlation between $\hat{Y}$ and Y^*:

$$r_{\hat{Y}Y^*} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \hat{\bar{Y}}\right)\left(Y_i^* - \bar{Y}^*\right)}{\sqrt{\sum_{i=1}^{n}\left(\hat{Y}_i - \hat{\bar{Y}}\right)^2 \sum_{i=1}^{n}\left(Y_i^* - \bar{Y}^*\right)^2}} \tag{1}$$

where $\bar{Y}$ denotes the mean of Y-values. This criterion works well for data that are reasonably normal, and the closer r is to 1, the better the imputation method.

The second criterion uses a regression approach to evaluate the performance of the imputation method which is useful for highly skewed data. The imputed and true values are first transformed by taking the natural logarithm ($log\ (Y+1)$). Only cases with

nonnegative incomes are included in the regression models for this criterion (negative incomes occur for business income, rental income and total income). The transformed imputed values $\hat{Y}_t$ are then regressed against the transformed true values Y_t^* using a linear model $Y_t^* = \beta \hat{Y}_t + \varepsilon$. For comparing imputation methods, the t-test statistic for $\beta = 1$ is calculated and the better imputation method has the t-test statistic closest to zero.

$$T = \frac{b-1}{\widehat{se}(b)} \tag{2}$$

where b denotes the estimated value of β and $\widehat{se}(b)$ is the estimated standard error of b.

The third criterion assesses the preservation of the change between waves by comparing the cross-wave correlations for the imputed and true values. The formula for the absolute change in correlations of the imputed and true values between Wave 1 and Wave 2 is

$$d_{corr1,2} = \left| \frac{\sum_{i=1}^{n} \left(\hat{Y}_{i1} - \hat{\bar{Y}}_1\right)\left(\hat{Y}_{i2} - \hat{\bar{Y}}_2\right)}{\sqrt{\sum_{i=1}^{n} \left(\hat{Y}_{i1} - \hat{\bar{Y}}_1\right)^2 \sum_{i=1}^{n} \left(\hat{Y}_{i2} - \hat{\bar{Y}}_2\right)^2}} - \frac{\sum_{i=1}^{n} \left(Y_{i1}^* - \bar{Y}_1^*\right)\left(Y_{i2}^* - \bar{Y}_2^*\right)}{\sqrt{\sum_{i=1}^{n} \left(Y_{i1}^* - \bar{Y}_1^*\right)^2 \sum_{i=1}^{n} \left(Y_{i2}^* - \bar{Y}_2^*\right)^2}} \right| \tag{3}$$

where Y_1 denotes the Y-values in Wave 1 and Y_2 denotes the Y-values in Wave 2. The closer the cross-wave correlations from the imputed data are to the true cross-wave correlations (that is, d_{corr} close to zero), the better the imputation method.

The fourth and fifth criteria assess the preservation of the relationships between income variables. The fourth criterion is the Euclidean distance between the imputed and true data values in multi-dimensional space. Let k denote the number of income variables being imputed simultaneously. Let y_{ij}^* denote the true data value for observation i and the jth variable, where $j = 1$ to k and let $\hat{y}_{ij}$ denote the imputed data for the same observation i and variable j. The mean of the Euclidean distances of the n imputed cases is then calculated.

$$mean\,(d_i) = \frac{1}{n} \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{k} \left(y_{ij}^* - \hat{y}_{ij}^2\right)} \tag{4}$$

A low mean indicates better performance of the method. While the Mahalanobis distance could be used here, the Euclidean distance is preferred as it gives equal weight to each unit of difference between the true and imputed values regardless of the size of the income components being compared.

The fifth evaluation criterion measuring the predictive accuracy compares the true correlations between the income variables being imputed with the correlations from the

imputed data for each two-way combination of $j = 1$ to k (the formula below is for Variables 1 and 2).

$$\left| r_{Y_1^* Y_2^*} - r_{\hat{Y}_1 \hat{Y}_2} \right| = \left| \frac{\sum_{i=1}^{n} \left(Y_{i1}^* - \bar{Y}_1^*\right)\left(Y_{i2}^* - \bar{Y}_2^*\right)}{\sqrt{\sum_{i=1}^{n} \left(Y_{i1}^* - \bar{Y}_1^*\right)^2 \sum_{i=1}^{n} \left(Y_{i2}^* - \bar{Y}_2^*\right)^2}} - \frac{\sum_{i=1}^{n} \left(\hat{Y}_{i1} - \hat{\bar{Y}}_1\right)\left(\hat{Y}_{i2} - \hat{\bar{Y}}_2\right)}{\sqrt{\sum_{i=1}^{n} \left(\hat{Y}_{i1} - \hat{\bar{Y}}_1\right)^2 \sum_{i=1}^{n} \left(\hat{Y}_{i2} - \hat{\bar{Y}}_2\right)^2}} \right| \tag{5}$$

The better the method, the closer this difference will be to zero (indicating the between-variable correlations are close to the true between-variable correlations).

2.2.2. Criteria 6 and 7: Distributional Accuracy

The next two criteria measure the distribution accuracy by considering whether the imputation method preserves the distribution of the true values.

The sixth criterion measures the distance between the empirical distribution functions for both the imputed and true values. The distance between these functions is measured using the Kolmogorov-Smirnov distance:

$$d_{KS} = \max_j \left(\left| \frac{1}{n} \sum_{i=1}^{n} I\left(Y_i^* \leq x_j\right) - \frac{1}{n} \sum_{i=1}^{n} I\left(\hat{Y}_i \leq x_j\right) \right| \right) \tag{6}$$

where the x_j values are the jointly ordered true and imputed values of Y. The smaller the distance, the better the imputation method.

In a longitudinal survey context, it is also important to assess the consistency of the income distribution between waves. The seventh criterion compares the distribution of income mobility in the dataset that includes the imputed values with the one that includes only true values (this measure includes all cases rather than just those imputed). Income mobility is measured by the change in income decile group membership from one wave to another. A Chi-Square test is used where the observed cell frequencies are those from the imputed dataset and the expected cell frequencies are the true cell frequencies. The null hypothesis is $H_0 : \hat{n}_{ij} = n_{ij}^*$ for all row i and column j. The test statistic is

$$\chi^2 = \sum_{j=1}^{10} \sum_{i=1}^{10} \frac{\left(\hat{n}_{ij} - n_{ij}^*\right)^2}{n_{ij}^*} \tag{7}$$

The better imputation method will have the lower χ^2 statistic.

2.2.3. Criteria 8 to 11: Estimation Accuracy

The final four criteria measure the estimation accuracy of the imputation methods by assessing whether the lower order moments of the distributions of the true values are preserved. Whilst Chambers (2000) suggested the absolute difference in the mean, variance, skew and kurtosis between the true and imputed cases be used, we have sought

scale invariant versions of these measures. Criterion 8 measures the absolute relative difference in means, Criterion 9 measures the absolute difference in the coefficient of variation, and Criteria 10 and 11 assess the absolute difference in the third and fourth standardised moment (i.e., the coefficient of skewness and kurtosis):

$$m_1 = \left| \left(\bar{Y}^* - \bar{\hat{Y}} \right) / \bar{\hat{Y}} \right| \tag{8}$$

$$m_2 = \left| \frac{\sigma^*}{\bar{Y}^*} - \frac{\hat{\sigma}}{\bar{\hat{Y}}} \right| \tag{9}$$

$$m_3 = \left| \frac{1}{n} \sum_{i=1}^{n} \left[\left(Y_i^* - \bar{Y}^* \right)^3 / \sigma^{*3} - \left(\hat{Y}_i - \bar{\hat{Y}} \right)^3 / \hat{\sigma}^3 \right] \right| \tag{10}$$

$$m_4 = \left| \frac{1}{n} \sum_{i=1}^{n} \left[\left(Y_i^* - \bar{Y}^* \right)^4 / \sigma^{*4} - \left(\hat{Y}_i - \bar{\hat{Y}} \right)^4 / \hat{\sigma}^4 \right] \right| \tag{11}$$

A good imputation method will have low absolute values for these measures.

3. Imputation Methods Tested

The following section describes the imputation methods considered in this evaluation study. The imputation methods adopted by large national household-based longitudinal surveys similar to the HILDA Survey provide guidance on which methods to include:

- In the British Household Panel Survey, two main methods of imputation are used. For continuous variables, a nearest neighbour regression method is used, whilst for categorical variables, a hotdeck method is used (Buck 1997).
- The German Socio-Economic Panel primarily uses the Little and Su method for the imputation of income (Frick and Grabka 2005).
- The Canadian Survey of Labour and Income Dynamics uses the last value carried forward method as the primary method. In the absence of data from the previous year, imputation using a nearest neighbour technique is employed. [4]
- The U.S. Panel Study of Income Dynamics, in general, uses hotdeck procedures to impute missing data (Hofferth et al. 1998).
- The U.S. Survey of Income and Program Participation uses two methods of imputation. Item nonresponse is imputed using a sequential hotdeck imputation procedure (Pennell 1993) and wave nonresponse is imputed using a longitudinal imputation procedure referred to as the random carryover method (Williams and Bailey 1996).

The methods tested in this evaluation study (all described in detail below) include:

- A nearest neighbour regression method
- A hotdeck method
- Two methods based on the Little and Su method, being with and without imputation classes

[4] Information on the imputation method used in the Canadian Survey of Labour and Income Dynamics was obtained from the documentation about the SLID methodology from www.statcan.ca.

- Three carryover methods, being the last value carried forward, the random carryover method, and the population carryover method
- A method which combines the population carryover method and the Little and Su method

As a single imputation solution is required for the HILDA data release at this time, we did not test any multiple imputation methods. All of the imputation methods considered, except for the carryover methods, could be extended to provide multiple imputes. It is not expected that the findings of this study would be substantially different if this were done.

In certain situations where the Little and Su methods and the carryover methods do not work on their own (e.g., the respondent is only interviewed in one wave and does not provide information about an income component), a cross-sectional method is needed. Two fallback options are assessed: a nearest neighbour regression method and a hotdeck method. Only information about the respondent from within the wave is used for covariates in the regression equation for the nearest neighbour regression method or to form hotdeck classes.

All of these methods are univariate imputation methods, where the imputation is applied one variable at a time. Indeed, only the nearest neighbour regression method and the hotdeck method use information about the receipt of a particular income when imputing another. We also assessed two variants of the Little and Su method that imputed two or more income variables simultaneously, however they did not demonstrate substantial improvement over the univariate Little and Su methods and have not been reported here.

3.1. Longitudinal Nearest Neighbour Regression Method

Like many imputation methods, the nearest neighbour regression method (also known as predictive mean matching (Little 1988)) uses the concept of donors and recipients. The record with missing information is called the "recipient" (i.e., it needs to be imputed). The "donor" has complete information that is used to impute the recipient's missing value. The nearest neighbour regression method seeks to identify the "closest" donor to each record that needs to be imputed via the predicted values from a regression model for the variable to be imputed. The donor's reported value for the variable being imputed replaces the missing value of the recipient.

For each wave and for each variable to be imputed, log-linear regression models using information from the same wave as well as information from other waves (if available) are constructed. Over 30 variables are considered for inclusion in the income models, covering demographic characteristics, employment characteristics, the partner's characteristics and income (if they had a partner), and income in other waves (reported or imputed in prior waves and reported in subsequent waves). A backwards elimination process in SAS is used to identify the key variables for each variable, wave and simulation.

The predicted values from these regression models for the variable being imputed are used to identify the nearest case (donor d) whose reported value (Y_d) can be inserted into the case with the missing value ($\hat{Y}_i = Y_d$). Donor d has the closest predicted value to the respondent i, that is $|\hat{\mu}_i - \hat{\mu}_d| \leq |\hat{\mu}_i - \hat{\mu}_p|$ for all respondents p (potential donors) where $\hat{\mu}_i$ is the predicted mean of Y for individual i that needs to be imputed, and Y_d is the observed value of Y for respondent d.

For respondents, the missing income is imputed for each variable. For nonrespondents (within responding households), only donors for total income are identified and the income components are taken from this donor.

For wages and salaries, government pensions, and rental income, we restrict the donor to the same age group as the recipient: 15–19, 20–24, 25–34, 35–44, 45–54, 55–64, and 65+. Imputation classes are formed using age because it is correlated with most income components and is known for almost all donors and recipients. As such, it helps impute realistic income amounts, especially for the young. For interest income, dividends and royalties, and private transfers, the age classes are: 15–24, 25–54, and 55+. No age class restrictions are applied for business income. The more detailed imputation class is used for total income for nonrespondents.

This method provides one of two fallback solutions when the Little and Su and carryover methods cannot be used. For example, the Little and Su method (described later) cannot be used if a respondent has not reported any income for the component being imputed and the nearest neighbour method is used to provide initial imputed values. When this fallback solution is adopted, the nearest neighbour regression method relies only on cross-sectional information (as longitudinal data is not available). We have termed this method a "nearest neighbour regression fallback method".

3.2. Longitudinal Hotdeck Method

The hotdeck method randomly matches suitable donors to recipients within imputation classes. The donor's reported value for the variable being imputed replaces the missing value of the recipient.

Up to 15 categorical variables are used to define the imputation classes for each income component. Suitable classes are derived from subject matter knowledge and investigative regression analysis using the data in Simulation 1. The sample in Simulation 1 is randomly divided into ten parts and log-linear regression models are fitted (via a stepwise process) ten times each wave, dropping 1/10th of the sample each time. Those variables most frequently included in the regression models (which signals their relative importance) are considered as imputation classes. These classes are then used in all waves and all simulations. The variables considered in the formation of the imputation classes are the categorical equivalent of those in the nearest neighbour modelling process, including income from other waves (where available).

Where there are not sufficient donors within a class, the list of variables defining the class is reduced, removing the least important variable first, until a suitable donor is found. A donor d is randomly matched to recipient i within an imputation class c (i.e., $c_i = c_d$). The donor's reported value is inserted into the recipient's missing value $\hat{Y}_i = Y_d$. A hotdeck macro (hesimput) written by the Statistical Services Branch of the Australian Bureau of Statistics is used.

This method provides an alternative fallback solution when the Little and Su and carryover methods cannot be used. When this fallback solution is adopted, it only uses cross-sectional information to form the imputation classes (the income bands from other waves are not available). We have termed this method a "hotdeck fallback method".

3.3. Basic Little and Su Method

The imputation method proposed by Little and Su (1989) is referred to here as the "basic Little and Su method" to distinguish it from the modified version using imputation classes which is referred to as the "Little and Su method with imputation classes".

The basic Little and Su method incorporates (via a multiplicative model) the trend across waves (column effect), the recipient's departure from the trend in the waves where the income component has been reported (row effect), and a residual effect donated from another respondent with complete income information for that component (residual effect). The model is of the form $imputation = \left(roweffect\right)\left(columneffect\right)\left(residualeffect\right)$. The column (wave) effects are calculated by

$$c_j = \frac{\bar{Y}_j}{\bar{Y}} \text{ where } \bar{Y} = \frac{1}{m}\sum_j \bar{Y}_j \tag{12}$$

for each wave $j = 1, \ldots, m$. $\bar{Y}_j$ is the sample mean of variable Y for wave j, based on complete cases and $\bar{Y}$ is the global mean of variable Y based on complete cases. The row (person) effects are calculated by

$$\bar{Y}^{(i)} = \frac{1}{m_i}\sum_j \frac{Y_{ij}}{c_j} \tag{13}$$

for both complete and incomplete cases. Here, the summation is over recorded waves for case i; m_i is the number of recorded waves; Y_{ij} is the variable of interest for case i, wave j; and c_j is the simple wave correction from the column effect. The cases are ordered by $\bar{Y}^{(i)}$, and incomplete case i is matched to the closest complete case d. The missing value Y_{ij} is imputed by

$$\hat{Y}_{ij} = \left(\bar{Y}^{(i)}\right)\left(c_j\right)\left(\frac{Y_{dj}}{\bar{Y}^{(d)}c_j}\right) = Y_{dj}\frac{\bar{Y}^{(i)}}{\bar{Y}^{(d)}} \tag{14}$$

where the three terms in brackets represent the row, column, and residual effects. The first two terms estimate the predicted mean, and the last term is the stochastic component of the imputation from the matched case.

It is important to note that due to the multiplicative nature of the Little and Su method, a zero individual effect will result in a zero imputed value. Nevertheless, it is quite valid to have an individual reporting zero income in earlier waves and then reporting that they have income but either do not know its value or refuse to provide it. The individual's effect would be zero and any imputed amount via the Little and Su method would also be zero, which we know is not true. Therefore, recipients with zero individual effects are imputed using a fallback method. In addition, the donors must have nonzero row effects to avoid dividing by zero.

3.4. Little and Su Method With Imputation Classes

Ideally, the donor and the recipient should have similar characteristics that are associated with the variable being imputed. The basic Little and Su method, therefore, is extended to

take into account basic characteristics of the donors and recipients. Donors and recipients are matched within longitudinal imputation classes defined by the following age ranges in the latest wave: 15–19, 20–24, 25–34, 35–44, 45–54, 55–64, 65+. The column and row effects are calculated within each imputation class and donors are matched to recipients which share the same imputation class.

3.5. Last Value Carried Forward

In the last value carried forward method, reported information from the previous wave where available is used to fill in the missing variable. That is, the missing value $Y_{i,j}$ for case i, wave j, is imputed by $\hat{Y}_{i,j} = Y_{i,j-1}$. Where this information is absent, a fallback method is used.

3.6. Random Carryover Method

The random carryover method imputes single missing wave data that is bounded on both sides by an interviewed wave (Williams and Bailey 1996). The value from either the preceding or subsequent wave is donated to the recipient with probability 0.5. That is, the missing value $Y_{i,j}$ for case i, wave j, is imputed randomly by $\hat{Y}_{i,j} = Y_{i,j-1}$ or $\hat{Y}_{i,j} = Y_{i,j+1}$. For this evaluation study, we relax the requirement that the missing wave data be bounded on both sides and permit information to be carried forward or backward if only one value is available. Where no information is available in surrounding waves, a fallback method is used.

3.7. Population Carryover Method

A variation of the random carryover method is also implemented and is referred to as the "population carryover method" (Williams and Bailey 1996). Rather than carrying information forward or backward with equal probability, the probability is chosen to reflect the changes in the reported income amounts between waves observed in the population.

The probability that the value is carried backward is calculated in the following way. An indicator variable is created which equals 1 when the reported change between waves j and $j + 1$ is smaller than the reported change between waves j and $j - 1$ for the complete cases; and 0 otherwise. The proportion p of the interviewed sample where the change between waves j and $j + 1$ is smaller than the change between waves j and $j - 1$ is then determined. The next value is carried backwards with probability p and the last value is carried forwards with probability $1 - p$, reflecting the probabilities associated with the occurrence of change between waves found in the complete cases.

3.8. Combination Method

During the course of this evaluation, we noticed the strengths of several methods and sought to combine two of the most promising methods to improve the imputation accuracy, particularly for nonrespondents. For this combined method, the population carryover method is used for nonrespondents to identify whether an income component is

a zero or nonzero amount. The Little and Su method is then used to impute the nonzero amounts. Age imputation classes (as described in Section 3.4) are adopted for wages and salaries and for pensions and benefits.

4. Comparison of Imputation Methods

Of the eight imputation methods described above, the Little and Su methods, the carryover methods and the combined method are examined both when the nearest neighbour regression fallback method is used and when the hotdeck fallback method is used, thus resulting in 14 imputation methods being implemented in this study. These imputation methods are compared via the eleven evaluation criteria. While we consider the performance in both the cross-sectional and longitudinal context, we place more emphasis on the longitudinal performance as this is the primary focus of a longitudinal survey.

4.1. Summarising Performance

To help draw conclusions from the many criteria and methods considered, we standardise the values from the various measures following the approach adopted by Chambers and Zhao (2003) in the Euredit project. The median value $\tilde{c}$ over all methods, variables and waves is subtracted from each evaluation measure c, giving residual $r = c - \tilde{c}$. This is then divided by the median of the nonzero absolute values of the residuals $|\widetilde{r_{nz}}|$. While this removes much of the variability in the measures, they are still somewhat skewed, so we add a small constant k and take the log of the standardised scores (to ensure the resultant value is valid). The log standardised score is therefore

$$\log c_s = \log\left(\frac{r}{|\widetilde{r_{nz}}|} + k\right) \tag{15}$$

As Criterion 1 does not have an absolute minimum at zero, it is modified to be the distance from 1 prior to the standardisation.

The log standardised scores are then averaged within the three classes of evaluation measures – predictive, distribution and estimation – for the cross-sectional and longitudinal estimates. Table 3 shows which criteria contributed to each accuracy class. By averaging the scores within these six dimensions, we treat the measures within each class equally. This process is undertaken separately for respondents and nonrespondents; the imputation for respondents only involves nonzero values whereas the imputation for nonrespondents could be zero or nonzero, resulting in quite different distributions of imputed values. The results for wages and salaries, pensions and benefits, business income and total income are reported in Table 4 (the other income components are not provided due to space limitations). Methods with low averaged log standardised scores are better than methods with high scores (the lowest scores are indicated with a bold entry in the table). Statistically significant differences in the performance of methods for each variable and respondent group are identified using Tukey's test which reduces the chance of a false positive result when multiple comparisons are made to 5 per cent.

Table 3. Grouping of criteria to assess imputation accuracy

	Cross-sectional accuracy			Longitudinal accuracy		
	Predictive	Distribution	Estimation	Predictive	Distribution	Estimation
Wages and salaries; Govt pensions and benefits; Interest income; Dividends and royalties; Rent income	1, 2, 4 and 5	6	8, 9, 10 and 11	1, 2 and 3	6	8, 9, 10 and 11
Business income; Private transfers	1 and 2	6	8, 9, 10 and 11	1, 2 and 3	6	8, 9, 10 and 11
Total financial year income	1 and 2	6	8, 9, 10 and 11	1, 2 and 3	6 and 7	8, 9, 10 and 11

Table 4. Average log standardised cross-sectional (X) and longitudinal (L) evaluation measures of predictive (P), distribution (D) and estimation (E) accuracy

Longitudinal method (base method)	Wages and salaries						Government pensions and benefits					
	XP	XD	XE	LP	LD	LE	XP	XD	XE	LP	LD	LE
Respondents												
NNRM longitudinal (base = NNRM)	1.49	1.30	1.53	1.73*	1.61*	1.81	1.59*	1.72	1.45*	1.63*	1.93*	1.86
Basic Little & Su (NNRM)	1.50	1.35*	1.50	1.66*	1.51	1.81	1.56*	1.71	1.44*	1.56*	1.77	1.82
Little & Su w imp classes (NNRM)	1.47	1.33	1.50	1.66*	**1.43**	1.79	1.55*	1.70	1.44	1.56*	1.75	1.80
LVCF (NNRM)	1.46	1.31	1.51	1.55	2.14*	1.90*	1.52	1.74	1.43	1.47	2.31*	1.84
Random carryover (NNRM)	1.44	1.29	1.49	1.55	1.95*	1.79	**1.48**	1.70	1.42	1.47	2.10*	1.79
Population carryover (NNRM)	**1.44**	**1.29**	1.49	**1.55**	1.95*	**1.78**	1.48	1.70	**1.42**	**1.47**	2.10*	1.80
Combined method (NNRM)	1.47	1.32	1.50	1.66*	1.44	1.80	1.54*	1.69	1.43	1.55*	**1.74**	1.80
Hotdeck longitudinal (HD)	1.54*	1.33	1.55*	1.73*	1.54*	1.80	1.56*	1.71	1.44*	1.55*	1.77	1.81
Basic Little & Su (HD)	1.50	1.35*	1.50	1.66*	1.50	1.81	1.57*	1.71	1.44*	1.57*	1.78	1.82
Little & Su w imp classes (HD)	1.48	1.33*	1.50	1.67*	1.43	1.79	1.55*	1.70	1.43	1.56*	1.74	1.79
LVCF (HD)	1.50	1.32	1.50	1.58	2.16*	1.87	1.55*	1.74	1.43	1.50	2.31*	1.81
Random carryover (HD)	1.47	1.31	1.48	1.56	1.96*	1.79	1.50	1.71	1.42	1.48	2.10*	**1.79**
Population carryover (HD)	1.47	1.31	**1.48**	1.56	1.96*	1.79	1.50	1.71	1.42	1.48	2.10*	1.79
Combined method (HD)	1.55*	1.32	1.56*	1.73*	1.55*	1.85	1.57*	**1.68**	1.44*	1.55*	1.77	1.79
Nonrespondents												
NNRM longitudinal (NNRM)	1.66*	1.66	1.44*	1.53*	2.29*	1.84	1.81*	1.68*	1.39*	1.64*	2.16*	1.91
Basic Little & Su (NNRM)	1.51	1.76*	1.42	1.48	1.87	**1.73**	1.65	1.92*	1.40*	1.56*	2.00*	1.85
Little & Su w imp classes (NNRM)	1.49	1.74*	1.41	1.46	1.85	1.75	1.64	1.90*	1.40*	1.55*	1.99*	1.84
LVCF (NNRM)	1.63*	1.67	1.43	1.51*	2.36*	1.84	1.73*	1.62	1.39	1.57*	1.94*	1.91
Random carryover (NNRM)	1.52	1.64	**1.41**	1.45	2.22*	1.80	1.61	**1.60**	1.38	1.49	1.87*	1.89
Population carryover (NNRM)	1.52	1.64	1.41	1.45	2.22*	1.81	1.61	1.60	**1.38**	1.49	1.87*	1.89
Combined method (NNRM)	**1.45**	1.68*	1.41	**1.43**	**1.81**	1.74	**1.57**	1.68*	1.38	**1.47**	**1.75**	**1.79**
Hotdeck longitudinal (HD)	1.69*	1.66	1.44*	1.51*	2.14*	1.82	1.83*	1.72*	1.40*	1.61*	2.05*	1.90
Basic Little & Su (HD)	1.65*	1.82*	1.43*	1.55*	2.07*	1.76	1.81*	2.03*	1.43*	1.65*	2.13*	1.98*
Little & Su w imp classes (HD)	1.65*	1.78*	1.42	1.55*	2.08*	1.77	1.81*	2.02*	1.43*	1.64*	2.12*	1.95
LVCF (HD)	1.70*	1.66	1.43*	1.52*	2.36*	1.84	1.80*	1.67*	1.39	1.57*	1.96*	1.90
Random carryover (HD)	1.56	1.63	1.41	1.46	2.21*	1.78	1.67	1.64	1.39	1.50	1.86*	1.91

Table 4. Continued

Longitudinal method (base method)	Wages and salaries						Government pensions and benefits					
	XP	XD	XE	LP	LD	LE	XP	XD	XE	LP	LD	LE
Population carryover (HD)	1.56	**1.63**	1.41	1.46	2.21*	1.78	1.67	1.63	1.39	1.50	1.86*	1.91
Combined method (HD)	1.62*	1.65	1.44*	1.47	2.10*	1.82	1.71*	1.64	1.39	1.50	1.86*	1.89
Longitudinal method (base method)	Business income						Total financial year income					
	XP	XD	XE	LP	LD	LE	XP	XD	XE	LP	LD	LE
Respondents												
NNRM longitudinal (base = NNRM)	1.65*	1.69*	1.75	1.86*	1.79*	2.33	1.32	1.19	1.63	1.95*	1.73*	2.24
Basic Little & Su (NNRM)	1.55	1.67*	1.69	1.66	1.65	2.18	1.28	1.18	1.57	1.64	1.35	2.19
Little & Su w imp classes (NNRM)	1.59	1.64*	1.80*	1.70	**1.62**	2.33	1.31	1.19	1.72*	1.70*	1.35	2.35*
LVCF (NNRM)	1.63*	1.63	1.77*	1.72*	1.79*	2.32	1.34	1.21	1.61	1.55	1.66*	2.28
Random carryover (NNRM)	1.60	1.59	1.74	1.71	1.73*	2.30	1.35*	**1.18**	1.60	1.57	1.50*	2.25
Population carryover (NNRM)	1.60	1.59	1.74	1.71	1.73*	2.30	1.35	1.18	1.60	1.56	1.50*	2.25
Combined method (NNRM)	1.56	1.66*	**1.69**	1.66	1.64	2.19	**1.26**	1.18	1.57	1.66*	**1.33**	2.18
Hotdeck longitudinal (HD)	1.62*	1.62	1.74	1.80*	1.70*	2.22	1.28	1.19	1.60	1.84*	1.57*	2.18
Basic Little & Su (HD)	**1.55**	1.69*	1.72	**1.65**	1.65	**2.18**	1.29	1.19	**1.55**	1.63	1.36	**2.16**
Little & Su w imp classes (HD)	1.59	1.65*	1.78*	1.67	1.62	2.28	1.32	1.20	1.70*	1.68*	1.34	2.30
LVCF (HD)	1.60	1.60	1.74	1.69	1.76*	2.28	1.33	1.22*	1.58	**1.54**	1.68*	2.23
Random carryover (HD)	1.57	1.58	1.73	1.68	1.71*	2.28	1.31	1.19	1.57	1.55	1.48*	2.20
Population carryover (HD)	1.57	**1.58**	1.73	1.68	1.71*	2.28	1.31	1.19	1.57	1.55	1.48*	2.20
Combined method (HD)	1.62*	1.62	1.73	1.79*	1.69*	2.22	1.28	1.19	1.60	1.83*	1.57*	2.21
Nonrespondents												
NNRM longitudinal (NNRM)	1.79*	1.44*	1.76	1.66	1.47*	**2.15**	1.08*	1.71	1.47	1.40	2.18*	1.80
Basic Little & Su (NNRM)	1.66*	1.40	1.76	1.70*	1.49*	2.18	1.00	1.77*	1.46	1.37	1.73*	1.81
Little & Su w imp classes (NNRM)	1.68*	1.41	1.76	1.69*	1.48*	2.20	0.98	1.71	1.47*	1.38	1.67	1.81
LVCF (NNRM)	1.73*	1.42*	1.81	1.65	1.45	2.20	1.11*	1.73*	1.46	1.37	2.29*	1.84
Random carryover (NNRM)	1.55	1.40	1.76	1.59	1.43	2.19	1.00	1.68	1.44	1.32	1.91*	1.77

Table 4. Continued

Longitudinal method (base method)	Business income						Total financial year income					
	XP	XD	XE	LP	LD	LE	XP	XD	XE	LP	LD	LE
Population carryover (NNRM)	1.55	1.40	1.76	1.59	1.43	2.19	1.00	1.68	1.44	1.32	1.91*	**1.77**
Combined method (NNRM)	**1.54**	**1.38**	**1.74**	1.58	**1.42**	2.18	**0.94**	1.68	1.45	**1.32**	**1.60**	1.79
Hotdeck longitudinal (HD)	1.80*	1.43*	1.83	1.67	1.46	2.22	1.12*	1.72*	1.47	1.38	1.97*	1.83
Basic Little & Su (HD)	1.70*	1.39	1.82	1.74*	1.52*	2.24	1.11*	1.89*	1.47*	1.44*	1.95*	1.80
Little & Su w imp classes (HD)	1.72*	1.40	1.82	1.75*	1.52*	2.27	1.10*	1.82*	1.47*	1.45*	1.91*	1.82
LVCF (HD)	1.74*	1.41*	1.81	1.65	1.45	2.19	1.15*	1.72*	1.45	1.37	2.26*	1.80
Random carryover (HD)	1.56	1.39	1.76	1.59	1.42	2.21	1.03*	1.67	**1.44**	1.33	1.91*	1.78
Population carryover (HD)	1.56	1.39	1.76	1.59	1.42	2.21	1.03*	**1.67**	1.44	1.33	1.91*	1.78
Combined method (HD)	1.62	1.42*	1.80	**1.56**	1.44	2.16	1.11*	1.72*	1.47	1.36	1.95*	1.82

Notes: i) The scores were standardised across the 30 replicates, eight variables, five waves, 14 methods and two respondent groups. The scores reported in this table are the average log standardised score for the evaluation measures in the same class (being predictive, distributional and estimation accuracy). This treats the measures within each class equally.
ii) Bold entries indicate the lowest score within each class (XP, XD, XE, LP, LD, LE) for respondents and for nonrespondents. Methods with low scores are better than those with high scores.
iii) *indicates that the score is significantly different (i.e., worse) from the method with the lowest score at the 5 per cent level.

4.2. Performance of the Imputation Methods

In order to focus our attention on a smaller number of methods, we first consider which cross-sectional fallback method works the best. With one exception, the methods using the nearest neighbour regression fallback method perform the same as, or significantly better than, their counterparts using the hotdeck fallback method. The exception is private transfers for nonrespondents (though this variable is not included in Table 4). The following discussion therefore concentrates on the first eight of the 14 imputation methods reported in Table 4.

The longitudinal nearest neighbour regression method and the longitudinal hotdeck method do not perform particularly well on either cross-sectional or longitudinal accuracy. This perhaps reflects the challenges faced in identifying the most relevant covariates for the regression models and the best imputation classes for the hotdeck method. While there is mixed performance on most accuracy measures, the longitudinal hotdeck method tends to perform better on longitudinal distribution accuracy than the longitudinal nearest neighbour method.

The Little and Su method with imputation classes seems to perform better than the basic Little and Su method where there is reasonably good correlation between the imputation class variable (age) and the variable being imputed (as occurred for wages and salaries and to a lesser extent pensions and benefits). When the imputation class variable is weakly associated with the variable being imputed, the basic Little and Su method performs slightly better. This indicates that adding unhelpful imputation classes can result in a poorer performance than if no imputation classes are used (especially when the donor pool is small). While few of these differences in accuracy are significant, the results are reasonably consistent across the different variables and respondent groups considered.

Of the carryover methods, there is little to distinguish the performance of the random carryover method and the population carryover method. However, the last value carried forward method performs poorly against these two, particularly on longitudinal distribution and estimation accuracy. As the population carryover method attempts to take into account any shift in income in the population between waves, it is preferred to the random carryover method.

For respondents, we find that the random or population carryover methods often perform the best in providing cross-sectional accuracy, but are very poor in maintaining the longitudinal distribution accuracy and, for some variables, also the longitudinal predictive accuracy. By comparison, the Little and Su methods tend to provide better longitudinal distributional accuracy whilst maintaining reasonably good accuracy on the other five accuracy dimensions.

For nonrespondents, it is clear that the Little and Su method does not perform as well as the random or population carryover method for many variables. The carryover methods are more likely to understate change and overstate the correlation between waves – however, this may be preferable to overstating change as occurs with some of the other methods (Heeringa and Lepkowski 1986).

Suspecting that carryover methods are better for nonrespondents because they more accurately impute zero amounts than the Little and Su methods, we introduced the combined method into our evaluation study. The combined method employs the

population carryover method to determine whether the nonrespondent should have a zero or nonzero value and a suitable Little and Su method is used to determine a suitable nonzero value. This evaluation shows that the combined method significantly improves the longitudinal predictive and distributional accuracy of pensions and benefits, business, interest and rental income for nonrespondents (though details of the last two variables are not shown in Table 4). Small (but admittedly insignificant) gains are also made on many of the other accuracy measures for both nonrespondents and respondents. Accuracy improvements for respondents are unexpected, but this suggests that better donors are being selected when the zeros in nonresponding waves are imputed more accurately.

If we had considered the imputation accuracy for total financial year income only we would have come to different conclusions. The basic Little and Su method seems to be best for respondents and the Little and Su method with imputation classes seems to work best for nonrespondents. The advantages of the carryover methods for nonrespondents are not as obvious for total financial year income, so we probably would not have thought a combined method would be beneficial. This shows that it is important to look at the imputation performance for at least three or four of the largest income components.

5. Conclusions

The results of this evaluation study do not demonstrate that one imputation method performs consistently better against each criterion for each income item.

The evidence shows that different imputation methods perform better for different income items. Using a variety of imputation methods best suited to each variable should produce superior results to using one imputation method for all variables. For items that have a large pool of donors and are well correlated with age (such as wages and salaries, and pensions and benefits), the Little and Su method with imputation classes performs the best for respondents. For all other income components that are not well correlated with age or have a smaller donor pool, the basic Little and Su method works well. For nonrespondents, the combination of the carryover method (to determine zeros and nonzeros) together with the Little and Su method (to determine the nonzero amount) provides good results. These methods are best supplemented by the nearest neighbour regression fallback method when the longitudinal method cannot be used.

We found that the evaluation framework is useful in comparing the different methods. It is instructive to apply this framework across multiple income components and to consider the results for both respondents and nonrespondents. Summarising the results into the three accuracy components – predictive, distribution and estimation – is valuable as some methods perform extremely well on some aspects yet poorly on others. The number of measures in each of these dimensions could possibly be reduced. The most useful measure seems to be longitudinal distribution accuracy as the methods tend to show their strengths and weaknesses via this measure. The least useful measures are cross-sectional and longitudinal estimation accuracy – many of the methods are indistinguishable on these measures. Identifying significant differences in how the methods perform helped focus our attention on the important differences. We did nevertheless find it difficult to compare so many methods in one evaluation, and suggest that any future comparisons for the HILDA Survey be made against the best method identified in this study.

It is worth noting that the evaluation framework is somewhat constrained by the response mechanism built into the simulations. In our study, we assumed that the response propensities were based on a range of characteristics but not on the value of income component itself. In reality, some respondents may decide not to report a particular income component because they feel the value is sensitive (e.g., they had a large windfall gain, or made a large business loss). Alternatively, they might not know the value because it was relatively small or irregularly paid or there may be events that occur after the last interview that result in subsequent wave nonresponse (e.g., a death of a family member, significant worsening of the individual's health). These reasons for missingness could not be modelled using the available data and as a result our evaluation study may overstate the effectiveness of the methods to reproduce the original data. We believe it is nevertheless unlikely to change the broad conclusions. Assessing the robustness of the results to the response mechanism employed is an area of potential future research. It would be particularly helpful if this could be done in the context of register information which provides "true" values for all sample members (such as the Finnish data used by Spiess and Goebel 2004).

Where comparisons can be made, our evaluation produced similar results to other studies of income imputation methods for longitudinal data. In assessing imputation methods for the U.S. Survey of Income and Program Participation, Tremblay (1994) and Williams and Bailey (1996) included various measures of predictive and estimation accuracy but did not include measures of distribution accuracy. Both studies concluded that the carryover methods were better than the basic Little and Su method. We may have formed the same conclusion had we not evaluated the distribution accuracy, thus highlighting the importance of such measures. None of the other studies that we have found include two or more of the methods we considered: Quintano et al. (2002) compares the last value carried forward method using a hotdeck fallback method with a cross-sectional hotdeck method and a longitudinal variant of the hotdeck method; Frick and Grabka (2005) compare the basic Little and Su method with a cross-sectional regression-based method; Spiess and Goebel (2004) compare a carryover method with several multiple imputation methods; and Laaksonen (2003) uses longitudinal data but only tests the performance of purely cross-sectional imputation methods.

This project has highlighted a number of possible areas for future work to improve our understanding of income imputation in longitudinal surveys. Firstly, we should investigate modifications to the Little and Su method, such as the choice of imputation classes, how the row, column and residual effects are combined, or how the trend and variability around the trend could be incorporated. Secondly, other imputation methods should be investigated using this evaluation framework, including multivariate imputation methods (such as the hierarchical imputation method used in the Euredit Project (Pannekoek 2002)) or multiple imputation methods. Thirdly, the response mechanism could be modified to one that is not missing at random (as mentioned above) to determine how much this matters in the evaluation of the imputation methods (an example is given by Champney and Bell (1982)). Finally, the imputation methods could be tested on other large national household-based longitudinal surveys with a view to harmonising the methods used (Frick and Grabka (2010) showed, for example, that some of the cross-national variation in an analysis of wages resulted from the different imputation methods adopted by different countries).

6. References

Buck, N. (1997). Imputation for Missing Income Data in a Panel Study. Paper presented at the IASS/IAOS Satellite Meeting on Longitudinal Studies, Jerusalem, 27–31 August. Draft Paper, ESRC Research Centre on Micro-Social Change, University of Essex.

Chambers, R. (2000). Evaluation Criteria for Statistical Editing and Imputation. Working Paper for the Euredit Project on the Development and Evaluation of New Methods for Editing and Imputation. Southampton: University of Southampton.

Chambers, R. and Zhao, X. (2003). Evaluation of Edit and Imputation Methods Applied to the UK Annual Business Inquiry. In Towards Effective Statistical Editing and Imputation Strategies – Findings of the Euredit Project (Volumes 1 and 2), J. Charlton (ed). Available at www.cs.york.ac.uk/euredit/

Champney, T.F. and Bell, R. (1982). Imputation of Income: A Procedural Comparison. Proceedings of the American Statistical Association, Survey Research Methods Section, 431–436.

Dillman, D.A., Eltinge, J.L., Groves, R.M., and Little, R.J.A. (2002). Survey Nonresponse in Design, Data Collection and Analysis. Survey Nonresponse, R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds). New York: Wiley, 3–26.

Frick, J. and Grabka, M. (2005). Item Nonresponse on Income Questions in Panel Surveys: Incidence, Imputation and the Impact on Inequality and Mobility. Allgemeines Statistisches Archiv, 89, 49–61.

Frick, J. and Grabka, M. (2010). Item Nonresponse and Imputation of Annual Labour Income in Panel Surveys from a Cross-National Perspective. Survey Methods in Multinational, Multiregional, and Multicultural Contexts, J.A. Harkness, M. Braun, B. Edwards, T.P. Johnson, L.E. Lyberg, P.P. Mohler, B. Pennell, and T.W. Smith (eds). New Jersey: Wiley, 355–372.

Heeringa, S.G. and Lepkowski, J.M. (1986). Longitudinal Imputation for the SIPP. Proceedings of the American Statistical Association, Survey Research Methods Section, 206–210.

Hofferth, S., Stafford, F.P., Yeung, W.J., Duncan, G.J., Hill, M.S., Lepkowski, J.M., and Morgan, J.N. (1998). A Panel Study of Income Dynamics: Procedures and Codebooks – Guide to the 1993 Interviewing Year. Michigan: Institute for Social Research, The University of Michigan.

Kalton, G. (1986). Handling Wave Nonresponse in Panel Surveys. Journal of Official Statistics, 2, 303–314.

Kalton, G. and Brick, J.M. (2000). Weighting in Household Panel Surveys. Researching Social and Economic Change: the Uses of Household Panel Studies, D. Rose (ed.). London: Routledge, 96–112.

Laaksonen, S. (2003). German Socio-Economic Household Panel. In Towards Effective Statistical Editing and Imputation Strategies – Findings of the Euredit Project (Volumes 1 and 2), J. Charlton (ed). Available at www.cs.york.ac.uk/euredit/

Lepkowski, J.M. (1989). Treatment of Wave Nonresponse in Panel Surveys. Panel Surveys, D. Kasprzyk, G.J. Duncan, G. Kalton, and M.P. Singh (eds). New York: Wiley, 348–374.

Little, R.J.A. (1988). Missing Data Adjustments in Large Surveys. Journal of Business and Economic Statistics, 6, 287–296.

Little, R.J.A. and Su, H.L. (1989). Item Nonresponse in Panel Surveys. Panel Surveys, D. Kasprzyk, G.J. Duncan, G. Kalton, and M.P. Singh (eds). New York: Wiley, 400–425.

Nordholt, E.S. (1998). Imputation: Methods, Simulation Experiments and Practical Examples. International Statistical Review, 66, 157–180.

Pannekoek, J. (2002). (Multivariate) Regression and Hot Deck Imputation Methods. Euredit Deliverable 5.1.1. Statistics Netherlands.

Pennell, S.G. (1993). Cross-Sectional Imputation and Longitudinal Editing Procedures in the Survey of Income and Program Participation. Michigan: Institute for Social Research, The University of Michigan.

Quintano, C., Castellano, R., and Regoli, A. (2002). A Mixed Imputation Procedure in a Split Panel Survey. Paper presented at International Conference on Improving Surveys, Copenhagen, 25–28 August.

Rubin, D.B. (1976). Inference and Missing Data. Biometrika, 63, 581–590.

Spiess, M. and Goebel, J. (2004). A Comparison of Different Imputation Rules. Harmonisation of Panel Surveys and Data Quality – CHINTEX, M. Ehling and U. Rendtel (eds). Wiesbaden: Federal Statistical Office, 293–316.

Tremblay, A. (1994). Longitudinal Imputation of SIPP Food Stamp Benefits. Proceedings of the American Statistical Association, Survey Research Methods Section, 809–814.

Watson, N. (2004). Income and Wealth Imputation for Waves 1 and 2. HILDA Project Technical Paper Series No. 3/04. Melbourne: Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Wooden, M. and Watson, N. (2007). The HILDA Survey and its Contribution to Economic and Social Research (So Far). Economic Record, 83, 208–231.

Williams, T.R. and Bailey, L. (1996). Compensating for Missing Wave Data in the Survey of Income and Program Participation (SIPP). Proceedings of the American Statistical Association, Survey Research Methods Section, 305–310.

Received March 2009
Revised April 2011