1352.0.55.075

Australian Bureau of Statistics

Research Paper

Imputation in Longitudinal Surveys: The Case of HILDA

www.abs.gov.au

New Issue

Research Paper

Imputation in Longitudinal Surveys: The Case of HILDA

Rosslyn Starick

ABS Outposted Officer
Melbourne Institute of Applied Economic and Social Research
University of Melbourne VIC 3010

Methodology Advisory Committee

18 November 2005, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) THURS 23 MAR 2006

ABS Catalogue no. 1352.0.55.075
ISBN 0 642 48179 2

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Mr Paul Sutcliffe, Statistical Services Branch on Canberra (02) 6252 6759 or email <paul.sutcliffe@abs.gov.au>.

IMPUTATION IN LONGITUDINAL SURVEYS: THE CASE OF HILDA

Rosslyn Starick
ABS Outposted Officer
Melbourne Institute of Applied Economic and Social Research
The University of Melbourne VIC 3010

email: r.starick@unimelb.edu.au
telephone: (03) 8344 1660
fax: (03) 8344 2111

EXECUTIVE SUMMARY

This paper outlines the methodological issues relating to imputation for longitudinal household surveys. A brief review of the imputation methods adopted by major longitudinal surveys is reported. The main objective of this paper is to develop an appropriate imputation methodology for use in longitudinal surveys, by evaluating alternative imputation methods and to adopt the best method in the Household, Income and Labour Dynamics in Australia (HILDA) Survey.

This paper describes a methodological evaluation framework for assessing a good imputation method and presents a quantitative comparison of the performance of alternative imputation methods, using HILDA data.

ACKNOWLEDGMENTS

This paper reports results from research undertaken by the author whilst on secondment to the Melbourne Institute of Applied Economic and Social Research (MIAESR), from the Australian Bureau of Statistics (ABS). This research uses data from the Household, Income and Labour Dynamics in Australia (HILDA) Survey, which is funded by the Commonwealth Department of Family and Community Services (FaCS).

Views expressed in this paper do not necessarily represent those of the MIAESR or FaCS.

DISCUSSION POINTS FOR MAC

- Are there any comments on the evaluation strategy?
- Are the evaluation criteria for comparing imputation methods appropriate?
- Are there any suggestions on alternative imputation methods not considered in the evaluation study?
- Are there any suggestions on approaches to multivariate imputation for multiple missingness in longitudinal surveys?

CONTENTS

Imputation in Longitudinal Surveys: The Case of HILDA

Rosslyn Starick
ABS Outposted Officer
Melbourne Institute of Applied Economic and Social Research
The University of Melbourne VIC 3010

email: r.starick@unimelb.edu.au
telephone: (03) 8344 1660
fax: (03) 8344 2111

1. Introduction

The Household, Income and Labour Dynamics in Australia (HILDA) Survey is a broad social and economic survey, with particular attention paid to family and household formation, income and work. It is a large-scale longitudinal survey that commenced in 2001 and the fieldwork is conducted annually.

2 The HILDA Survey began with a large national probability sample of Australian households occupying private dwellings. All members of the households providing at least one interview in wave 1 form the basis of the panel to be pursued in each subsequent wave. The sample has been gradually extended to include any new household members resulting from changes in the composition of the original households.

3 Continuing Sample Members (CSMs) are defined to include all members of wave 1 households. Any children subsequently born to or adopted by CSMs are also classified as CSMs. Further, all new entrants to a household who have a child with a CSM are converted to CSM status. CSMs remain in the sample indefinitely. All other people who share a household with a CSM in wave 2 or later are considered Temporary Sample Members (TSMs). TSMs are followed for as long as they share a household with a CSM.

4 Figure 1 shows the evolution of the sample across the three waves. The wave 1 sample consisted of 19,914 people. A further 442 births and 54 parents of newborns who were not originally CSMs have been added to the sample in waves 2 and 3. A total of 177 deaths have been identified across the two follow-up waves and 256 people have moved overseas, though 24 returned after being away for one wave. Of the TSMs joining the sample in wave 2, a third had moved out by wave 3.

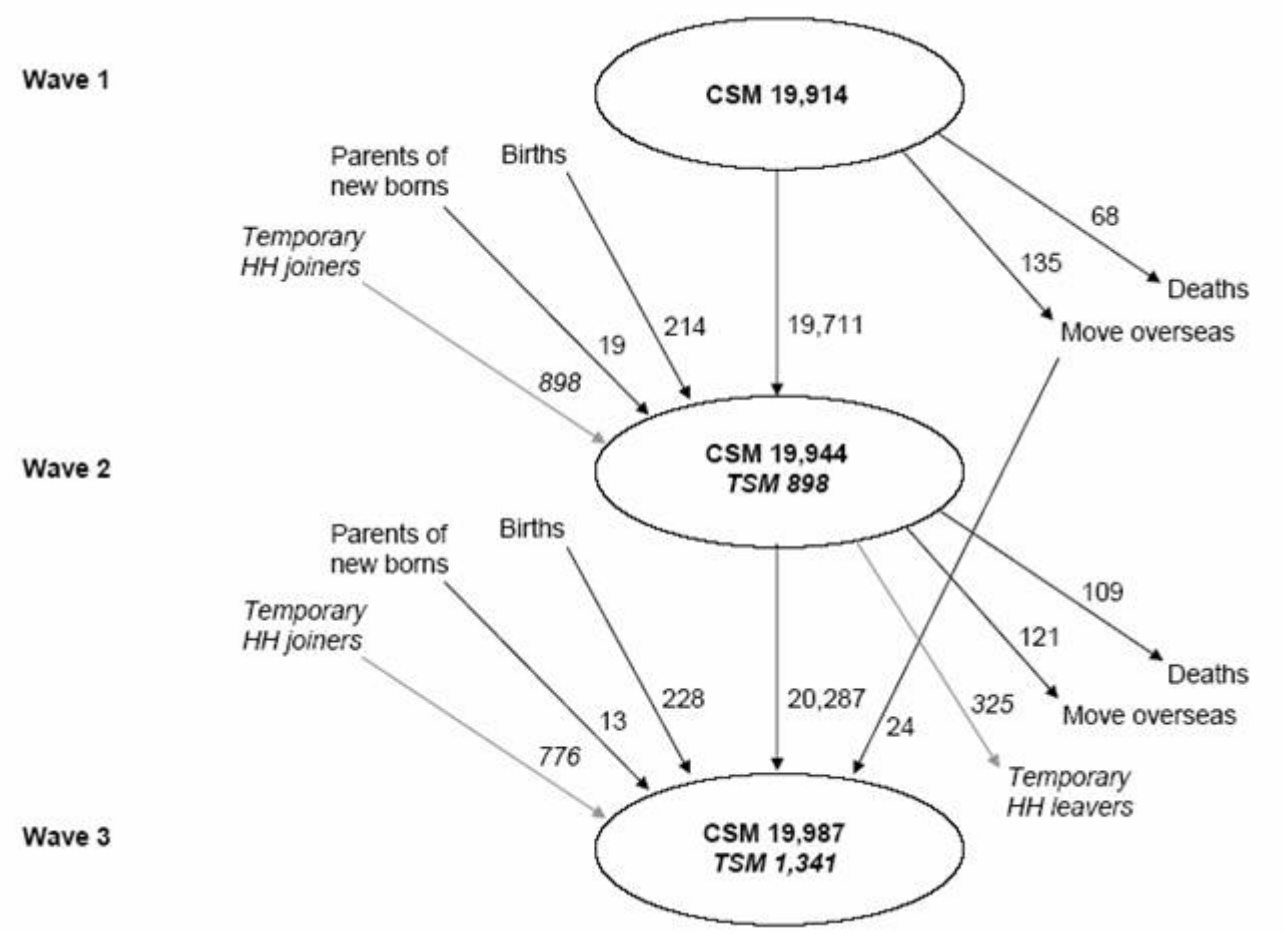
5 The author is currently outposted to the Commonwealth Department of Family and Community Services (FaCS) to provide statistical support in the further development of the HILDA Survey and is located at the Melbourne Institute of Applied Economic and Social Research at the University of Melbourne. The Melbourne Institute conducts the HILDA Survey for FaCS. The author's key role is to research and develop an appropriate imputation methodology for use in longitudinal surveys such as HILDA.

6 The purposes of this paper are to:

- highlight the methodological issues relating to imputation for longitudinal surveys;
- articulate the evaluation criteria for assessing a good imputation method; and
- report the results from an evaluation study using HILDA data.

7 The paper is organised as follows. Section 2 briefly reviews the imputation methods adopted by major longitudinal surveys. Section 3 discusses the methodological issues relating to imputation in longitudinal surveys. Section 4 describes the attributes of a good imputation method. The next five sections describe the imputation methods considered in an evaluation study comparing the performance of these methods, the evaluation methodology, the evaluation criteria, and the results from a comparison of imputation methods using HILDA data.

Figure 1: The Evolution of the HILDA Survey Sample



2. Imputation Methods Adopted by Major Longitudinal Surveys

8 There is an abundance of literature on imputation methods. Many of the methods have been developed for cross-sectional surveys. For longitudinal surveys, cross-sectional imputation restricts information for predicting missing values to the current wave whereas longitudinal imputation allows information from other waves to be used. The disadvantage of using cross-sectional imputation in a longitudinal survey is that the imputation may introduce distortion in estimates of change between waves.

9 A review of the imputation methods adopted by longitudinal surveys conducted overseas is briefly highlighted here and the current approach to imputation in HILDA is also briefly mentioned in this section.

Imputation in Panel Studies Overseas

10 In the British Household Panel Study, two main methods of imputation are used. For continuous variables, a nearest neighbour regression method is used, whilst for categorical variables, a hot deck method is used (Buck, 1997).

11 The German Socio-Economic Panel uses an imputation method developed by Little and Su (1989). It is a simple stochastic longitudinal imputation method for repeated measures data.

12 The Canadian Survey of Labour and Income Dynamics uses a 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.[1]

13 The US Panel Study of Income Dynamics, in general uses hot deck procedures to impute missing data (Hofferth et al. 1998).

14 The US Survey of Income and Program Participation currently uses two methods of imputation. Item non-response is imputed using a sequential hot deck imputation procedure (Pennell, 1993) and wave non-response is imputed using a longitudinal imputation procedure referred to as the random carryover method (Williams and Bailey, 1996).

Imputation in HILDA

15 In HILDA, the primary method for imputing income is based on the Little and Su method, which has been modified to match donors and recipients within imputation classes. However, for some cases there is still a need to use a cross-sectional imputation method and the method used is the nearest neighbour regression method (similar to that used by the British Household Panel Study).

[1] 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.

3. Imputation in Longitudinal Surveys – Methodological Issues

Dealing with Wave Non-Response

16 Non-response is a common source of non-sampling error in surveys. In general, there are two types of non-response. Unit non-response occurs when no data are collected for a sampled unit. Item non-response occurs when the sampled unit provides data for some but not all of the survey data items. In a longitudinal survey, there is another type of non-response – wave non-response. Wave non-response occurs when responses are provided for some but not all waves of the survey.

17 To reduce the effect of non-response bias, weighting adjustments and imputation are used. Literature on this topic suggests that it is common practice that weighting adjustments are used to deal with unit non-response and imputation is used to deal with item non-response. Wave non-response may be viewed as a set of item non-responses in the longitudinal record, suggesting that imputation may be used to deal with wave non-response from a longitudinal perspective but from a cross-sectional perspective, it may be viewed as unit non-response, therefore a weighting adjustment may be appropriate (Kalton (1986)).

18 In the HILDA Survey, the following approach was undertaken to deal with non-response. Non-responding households were accounted for by adjusting the weights of the responding households. Non-responding persons (from responding and non-responding households) were also accounted for by adjusting the weights of responding persons. For respondents with item non-response, the income components have been imputed and the totals are the sum of the relevant components. However, for non-respondents within responding households just the income totals have been imputed. Therefore, for income, only imputed totals are available at the household level.

19 There has been some debate about whether to impute the income components, so that the components are available at the household level. Researchers are interested in studying these components at the household level, but this issue requires further investigation to assess the quality of these imputations.

Preserving Estimates of Change between Waves

20 For a longitudinal survey, such as HILDA it is important that the imputation method performs well over time since there are repeated observations made on the same set of cases. The imputation method should preserve the estimates of change between waves.

21 The income imputation for Release 2.0 of the HILDA data was implemented using a nearest neighbour regression method or predictive mean matching, Little (1988). An analysis was undertaken to assess the quality of the imputation of income and a number of issues arose regarding the imputation method. The most serious of these quality issues was the biases introduced into the estimates of change by the imputation method itself.

22 As the income imputation provided in Release 2.0 was deficient in a number of ways as outlined in Watson (2004), the income imputation methodology was revised for Release 3.0. Hence, the extended version of the Little and Su method is now employed.

Multivariate Imputation for Multiple Missingness

23 There are two types of multiple missingness in a longitudinal survey. A sampled unit with item non-response may have multiple missing items. A sampled unit may also have multiple missing waves.

24 In HILDA, both the nearest neighbour regression method and the Little and Su method impute one item at a time. This means that if a recipient has multiple missing items, the imputed values for these missing items can come from different donors. To preserve the correlations between the items imputed, it is desirable to use an imputation method that can impute multiple missing items simultaneously.

25 This problem of multiple missingness can be extended to recipients that need imputing in multiple waves. The Little and Su method already imputes multiple missing waves simultaneously using the same donor, but the nearest neighbour regression method doesn't. The advantage of imputing multiple missing waves at the same time is to try to preserve the correlations between waves.

Revision of Imputation

26 In a cross-sectional survey, in general imputed values are not revised as new data becomes available. Usually, imputation methods employed in cross-sectional surveys use historical and/or cross-sectional information that is available at the time.

27 In a longitudinal survey, longitudinal information should be used in the imputation where possible. This means that information from future waves should be used to revise the imputation for previous waves. The question then becomes how far into the future and how far back into the past should information be included in the imputation.

4. Attributes of a Good Imputation Method

28 The objective of this research was to develop an appropriate imputation methodology for use in longitudinal surveys such as HILDA. The remainder of this paper describes the methodological evaluation framework for assessing a good imputation method and presents a quantitative comparison of the performance of alternative imputation methods based on simulation which uses HILDA data.

29 A good imputation method must have good statistical properties and be operationally efficient. Ideally, an imputation procedure should be capable of effectively reproducing the key outputs from a "complete data" statistical analysis of the dataset of interest. However, this is usually impossible, because the "true" values are unknown.

30 Chambers (2000) proposed the following desirable properties for an imputation procedure. These properties are not mutually exclusive.

Predictive Accuracy

31 The imputation procedure should maximise the preservation of true values. That is, it should result in imputed values that are as "close" as possible to the true values.

Ranking Accuracy

32 The imputation procedure should maximise the preservation of order in the imputed values. That is, it should result in ordering relationships between imputed values that are the same (or very similar) to those that hold in the true values.

Distributional Accuracy

33 The imputation procedure should preserve the distribution of the true data values. That is, marginal and higher order distributions of the imputed data values should be essentially the same as the corresponding distributions of the true values.

Estimation Accuracy

34 The imputation procedure should maximise the preservation of analysis. That is, it should reproduce the lower order moments of the distributions of the true values. In particular, it should lead to unbiased and efficient inferences for parameters of the distribution of the true values (given that these true values are unavailable).

Imputation Plausibility

35 The imputation procedure should lead to imputed values that are plausible. In particular, they should be acceptable values as far as the editing procedure is concerned.

Other Attributes

36 As well as these statistical properties that are desirable in an imputation method, another important aspect is the operational efficiency of an imputation method. That is, the ease with which it can be implemented, maintained and applied.

37 Furthermore, Chambers suggested that an imputation system should produce measures of the quality of its imputations. He suggested one important quality measure being the imputation variance (assuming that the imputation method preserves distributions). This is the additional variability, over and above the "complete data variability", associated with inference based on the imputed data. It is caused by the extra uncertainty associated with randomness in the imputation method. This imputation variance can be measured by repeating the imputation process and applying multiple imputation theory.

5. Univariate Imputation Methods

38 The following section describes the imputation methods considered in the evaluation study. They are grouped here under the heading of univariate imputation methods because the imputation is applied to a variable at a time. Multivariate imputation methods are considered in section 6.

Nearest Neighbour Regression Method

39 The income imputation for Release 2.0 of the HILDA data (see Watson, 2004) was implemented using a nearest neighbour regression method (similar to that used by the British Household Panel Study). For the evaluation, a 3 wave version of the method adopted in Release 2.0 was implemented as follows.

40 For each wave and for each variable imputed, regression models[2] using information from the same wave as well as information from other waves (if available) were constructed. Therefore, four different types of models were developed for each wave and for each variable imputed.

Wave 1 income imputation models

- where wave 2 and wave 3 income known
- where wave 2 income known and wave 3 income unknown
- where wave 2 income unknown and wave 3 income known
- where wave 2 and wave 3 income unknown

[2] For Release 2.0, Appendix 1 in Watson, 2004 provides a full list of the variables considered in the income models, together with tables showing the variables kept in the final models. This illustrates the size and complexity of the models constructed.

Wave 2 income imputation models

- where wave 1 income known or imputed and wave 3 income known
- where wave 1 income known or imputed and wave 3 income unknown
- where wave 1 income unknown and wave 3 income known
- where wave 1 and wave 3 income unknown

Wave 3 income imputation models

- where wave 1 income known or imputed and wave 2 income known or imputed
- where wave 1 income known or imputed and wave 2 income unknown
- where wave 1 income unknown and wave 2 income known or imputed
- where wave 1 and wave 2 income unknown

41 As in Release 2.0, a statistical package called MARS (Multivariate Adaptive Regression Splines) was used to construct the income models.[3] MARS is an automatic regression package which finds the best model for the specified variable from the host of variables it is instructed to consider. Main effects and two-way interactions were considered. MARS provided a practical solution to the resource intensive problem of constructing good predictive models.

42 The predicted values from a regression model for the variable being imputed were used to identify the nearest case whose reported value could be inserted into the case with the missing value.

$$\hat{Y}_i = Y_k$$

where $|\hat{\mu}_i - \hat{\mu}_k| \leq |\hat{\mu}_i - \hat{\mu}_l|$ for all respondents l, $\hat{\mu}_i$ is the predicted mean of Y for individual i, and Y_k is the observed value of Y for respondent k.

Little and Su Method

Basic Little and Su Method

43 The longitudinal imputation method proposed by Little and Su (1989), and used by the German Socio-Economic Panel will be referred to as the basic Little and Su method, to distinguish it from the modified version that was implemented in Release 3.0 of the HILDA data. The modified version will be referred to as the extended Little and Su method.

[3] See the Salford Systems website for an overview of the MARS package: www.salford-systems.com.

44 The basic Little and Su method incorporates trend and individual level information into the imputed amounts by using a multiplicative model based on row (person) and column (wave) effects. The model is of the form

imputation = (row effect) x (column effect) x (residual).

(a) Column (wave) effects of the form

$$c_j = \frac{\bar{Y}_j}{\bar{\bar{Y}}}$$

where $\bar{\bar{Y}} = \frac{1}{m}\sum_j \bar{Y}_j$

were computed for each wave j = 1, …, m, where $\bar{Y}_j$ is the sample mean of variable Y for wave j, based on complete cases and $\bar{\bar{Y}}$ is the global mean of variable Y based on complete cases.

(b) Row (person) effects of the form

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

were computed 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 (a).

(c) Cases were ordered by $\bar{Y}^{(i)}$, and incomplete case i is matched to the closest complete case, say *l*.

(d) Missing value Y_{ij} was imputed by

$$\hat{Y}_{ij} = \left[\bar{Y}^{(i)}\right]\left[c_j\right]\left[\frac{Y_{lj}}{\bar{Y}^{(l)} c_j}\right]$$

$$= Y_{lj}\frac{\bar{Y}^{(i)}}{\bar{Y}^{(l)}}$$

where the three terms in square parentheses 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.

45 It is important to note that there is a fundamental flaw with the basic Little and Su method. There is an underlying assumption that the individual effects must be non-zero. However, it is quite valid to have an individual reporting zero income in previous waves and then report that they have income but refuse to provide it and therefore this missing value needs to be imputed. Unfortunately, this individual's effect would be zero which means that any imputed amount under the basic Little and Su method would always be zero, which we know cannot be the case. There is also a problem with individual effects of donors that are zero as the imputed amount requires division by zero. Therefore, recipients with zero individual effects must be imputed using another method and in this case, they were imputed using the nearest neighbour regression method.

Extended Little and Su Method

46 Ideally, the record with missing information (called the recipient) should be imputed using information from a record with complete information (called the donor) that has similar characteristics for the variable of interest. The basic Little and Su method, therefore, was extended to take into account the characteristics of the donors and recipients. Donors and recipients were matched within imputation classes which had similar characteristics. The imputation classes used were age groups defined by the following ranges: 15-19, 20-24, 25-34, 35-44, 45-54, 55-64, 65+.[4]

47 The extended Little and Su method was implemented in Release 3.0 of the HILDA data as follows:

(a) Column (wave) effects of the form

$$c_{hj} = \frac{\overline{Y}_{hj}}{\overline{Y}_{h}}$$

where $\overline{Y}_{h} = \frac{1}{m}\sum_{j}\overline{Y}_{hj}$

were computed for each wave j = 1, …, m, and for each age group h = 1, …, c, where $\overline{Y}_{hj}$ is the sample mean of variable Y for wave j, age group h based on complete cases and $\overline{Y}_{h}$ is the global mean of variable Y for age group h based on complete cases.

(b) Row (person) effects of the form

$$\overline{Y}_{h}^{(i)} = \frac{1}{m_i}\sum_{j}\frac{Y_{hij}}{c_{hj}}$$

[4] Age groups were used to create the imputation classes because it is a simple characteristic and it is known for almost all donors and recipients. For a few cases, age was missing and was therefore imputed from a family of similar relationship structure to the missing case.

were computed 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_{hij} is the variable of interest for case i, wave j, age group h; and c_{hj} is the simple wave correction from (a).

(c) Cases were ordered by $\overline{Y}_h^{(i)}$, and incomplete case i is matched to the closest complete case, say *l* within age group h.

(d) Missing value Y_{hij} was imputed by

$$\hat{Y}_{hij} = \left[\overline{Y}_h^{(i)}\right]\left[c_{hj}\right]\left[\frac{Y_{hlj}}{\overline{Y}_h^{(l)} c_{hj}}\right]$$

$$= Y_{hlj}\frac{\overline{Y}_h^{(i)}}{\overline{Y}_h^{(l)}}$$

where the three terms in square parentheses 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.

Carryover Method

Last Value Carried Forward

48 Similar to the Canadian Survey of Labour and Income Dynamics, a last value carried forward method will be assessed. Where reported information from the previous wave is available, the missing value Y_{ij} for case i, wave j is imputed by

$$\hat{Y}_{i,j} = Y_{i,j-1}$$

49 Where reported information from the previous wave is absent, the nearest neighbour regression method is used.

Random Carryover Method

50 The random carryover method imputes single missing wave data that is bounded on both sides by an interviewed wave. This means that this method does not impute data where there are two or more consecutive missing waves nor the first or last wave. This method will be adapted to the HILDA environment as follows.

51 The random carryover method will be used to impute item non-response in HILDA. In general, there are three types of situations where imputation is required, where X denotes a reported value and O denotes a missing value.

XOO – last value carried forward applied
OOX – next value carried backward applied
XOX – random carryover method applied

52 The random carryover method is applied by randomly determining which wave (preceding or subsequent) is used to donate the imputed amount. A value r is randomly assigned to each case for each missing item, where r = 0 or 1. If r = 0 then the imputed value comes from the preceding wave. If r = 1 then the imputed value comes from the subsequent wave.

Population Carryover Method

53 A variation of the random carryover method is referred to as the population carryover method. Rather than choosing a donor by assigning a random value r, a donor is determined by reflecting the population changes in the reported income amounts between waves.

54 An indicator variable c is created which equals 1 when the reported change between waves 2 and 3 is smaller than the reported change between waves 1 and 2 for the interviewed sample; and 0 otherwise. The proportion p, of the interviewed sample where the change between waves 2 and 3 is smaller than the change between waves 1 and 2 is then determined. Whether the preceding wave or the subsequent wave donates the imputed amount is determined by reflecting the probabilities associated with the occurrences of change between waves found in the interviewed sample.

6. Multivariate Imputation Methods

55 As mentioned above, it is desirable to impute multiple missing items and multiple missing waves simultaneously. The Little and Su method imputes multiple missing waves simultaneously but not multiple missing items.

56 Researching and developing multivariate imputation methods is an area of further work that is planned. The following section explores preliminary multivariate imputation options being considered using the Little and Su method as the base method.

Option 1

57 One option[5] is to determine which income component is the most important variable and using the Little and Su method, impute the missing value for this item. Then, impute the remaining missing items (if there are multiple missing items), using the same donor.

58 For example, suppose we have a case which has missing values for current wages and salaries and for current benefits. We decide that wages and salaries is more important than benefits. So, we impute wages and salaries using the Little and Su method and, using the same donor we impute benefits.

[5] Thanks to Robert Clark for suggesting this option.

Option 2

59 Another option is to calculate a combined row effect, then ordering cases by this combined row effect to identify the nearest neighbour. The combined row effect would be a function of the row effects for each income variable.

60 Let Y_k denote the income variables being imputed, k = 1, …, K.

61 For each Y_k compute the column effects c_{khj} and the row effects $\overline{Y}_{kh}^{(i)}$. Then calculate a combined row effect using a distance function such as Euclidean distance or Mahalanobis distance.

$$\overline{Y}_{.h}^{(i)} = D\left(\overline{Y}_{1h}^{(i)}, \overline{Y}_{2h}^{(i)}, ..., \overline{Y}_{Kh}^{(i)}\right)$$

62 Cases are ordered by this combined row effect, and the nearest suitable donor is found.

63 Missing value Y_{khij} is imputed by

$$\hat{Y}_{khij} = Y_{khlj} \frac{\overline{Y}_{kh}^{(i)}}{\overline{Y}_{kh}^{(l)}}$$

- Are there any suggestions on alternative imputation methods not considered in the evaluation study?
- Are there any suggestions on approaches to multivariate imputation for multiple missingness in longitudinal surveys?

7. Evaluation Criteria for Comparing Imputation Methods

64 This section defines the evaluation criteria that were used in the evaluation study and form the framework for comparing imputation methods. The following criteria were based mainly on those proposed by Chambers (2000) and the criteria considered appropriate in the HILDA context were applied.

65 Unless otherwise stated, all measures are defined on the set of n imputed values within a dataset, rather than the set of all values. Let $\hat{Y}$ denote the imputed version of variable Y and Y^* denote the true version of variable Y.

Predictive Accuracy

66 The imputation procedure should maximise the preservation of true values. That is, it should result in imputed values that are as "close" as possible to the true values.

67 It is desirable to compute the predictive accuracy of each of the imputation methods. If this property holds, then $\hat{Y}$ should be close to Y^* for all cases where imputation has been carried out. For data that are reasonably "normal" looking, the sample Pearson correlation between $\hat{Y}$ and Y^* for those n cases where an imputation has actually been carried out should give a good measure of imputation performance. The formula for the sample Pearson correlation is

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

where $\bar{Y}$ denotes the sample mean of Y-values for the same n cases. A good imputation method will have r close to ±1.

68 For data that are highly skewed, Chambers (2000) recommended a regression approach to evaluate the performance of the imputation method. The regression approach evaluates the performance of the imputation method by fitting a linear model of the form

$$Y^* = \beta\hat{Y} + \varepsilon$$

to the imputed data values using a robust estimation method. Let b denote the fitted value of β that results. A measure of the regression mean square error

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i^* - b\hat{Y}_i\right)^2$$

can be computed as well. A good imputation method will have b close to 1 and a low value of $\hat{\sigma}^2$. For comparing imputation methods, the t-test statistic was calculated and the better imputation method will have the lower t-test statistic.

$$H_0 : \beta = 1$$

$$T = \frac{b-1}{s\sqrt{h_{ii}}} \qquad (2)$$

Note: The income variables were transformed by taking the natural logarithm of the variables.

$$transform(Y) = \log(Y+1)$$

69 Only cases with non-negative incomes were included in the regression models for this criterion. Negative incomes occurred for business income, rental income and total income.

Distributional Accuracy

70 The imputation procedure should preserve the distribution of the true data values. That is, marginal and higher order distributions of the imputed data values should be essentially the same as the corresponding distributions of the true values.

71 One measure that can be used to assess the preservation of the distribution of the true values is to compute the empirical distribution functions for both the imputed and true values and then measure the distance between these functions.

$$F_{Y^*n}(x) = \frac{1}{n}\sum_{i=1}^{n} I\left(Y_i^* \le x\right)$$

$$F_{\hat{Y}n}(x) = \frac{1}{n}\sum_{i=1}^{n} I\left(\hat{Y}_i \le x\right)$$

72 The "distance" between these functions can be measured using the Kolmogorov-Smirnov distance

$$d_{KS}\left(F_{Y^*n}, F_{\hat{Y}n}\right) = \max_x\left(\left|F_{Y^*n}(x) - F_{\hat{Y}n}(x)\right|\right) = \max_j\left(\left|F_{Y^*n}(x_j) - F_{\hat{Y}n}(x_j)\right|\right) \qquad (3)$$

where the $\{x_j\}$ values are the jointly ordered true and imputed values of Y. A good imputation method will have a small distance value.

Estimation Accuracy

73 The imputation procedure should maximise the preservation of analysis. That is, it should reproduce the lower order moments of the distributions of the true values. In particular, it should lead to unbiased and efficient inferences for parameters of the distribution of the true values (given that these true values are unavailable).

74 In considering the preservation of aggregates when imputing values, the most important case is the preservation of the raw moments of the empirical distribution of the true values. For k = 1, 2, …, a measure of how well these are preserved is given by

$$m_k = \left| \frac{1}{n} \sum_{i=1}^{n} \left(Y_i^{*k} - \hat{Y}_i^k \right) \right| = \left| m\left(Y^{*k} \right) - m\left(\hat{Y}^k \right) \right|$$

75 In the evaluation study, the parameter estimates of the mean, variance, skewness and kurtosis were computed for both the distribution of true values and the distribution of imputed values. A good imputation method will have a low absolute difference in moments.

Absolute difference in mean (1st order moment):

$$m_1 = \left| m\left(Y^{*1} \right) - m\left(\hat{Y}^1 \right) \right| \qquad (4)$$

Absolute difference in variance (2nd order moment):

$$m_2 = \left| m\left(Y^{*2} \right) - m\left(\hat{Y}^2 \right) \right| \qquad (5)$$

Absolute difference in skewness (3rd order moment):

$$m_3 = \left| m\left(Y^{*3} \right) - m\left(\hat{Y}^3 \right) \right| \qquad (6)$$

Absolute difference in kurtosis (4th order moment):

$$m_4 = \left| m\left(Y^{*4} \right) - m\left(\hat{Y}^4 \right) \right| \qquad (7)$$

Other Measures

76 For a longitudinal survey, such as HILDA it is important that the imputation method performs well over time since there are repeated observations made on the same set of cases.

77 The imputation procedure should preserve the longitudinal nature of the true data values. That is, change in estimates between waves should be essentially the same for both the imputed and true values.

78 One measure that can be used to assess the preservation of the change between waves is to compute the cross-wave correlations for both the imputed and true values. For example, the formulae for the correlations between wave 1 and wave 2 for both the imputed and true values are

$$r_{\hat{Y}_1\hat{Y}_2} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_{i1}-\hat{\bar{Y}}_1\right)\left(\hat{Y}_{i2}-\hat{\bar{Y}}_2\right)}{\sqrt{\sum_{i=1}^{n}\left(\hat{Y}_{i1}-\hat{\bar{Y}}_1\right)^2\sum_{i=1}^{n}\left(\hat{Y}_{i2}-\hat{\bar{Y}}_2\right)^2}} \quad (8)$$

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

where Y_1 denotes the Y-values in wave 1 and Y_2 denotes the Y-values in wave 2. A good imputation method will have cross-wave correlations close to the true cross-wave correlations.

79 In a longitudinal survey context, it is also important to assess the consistency of the income distribution between waves. One measure that can be used to assess the distribution consistency between waves is to compute income mobility by measuring the change in income decile group membership from one wave to another for both the imputed and true data values and then test if the consistency of the distribution between waves is the same for imputed and true values. Note that this measure uses all data values in the dataset rather than just the imputed values.

80 As in distributional accuracy, the empirical distribution functions for both the imputed and true values are computed.

$$F_{Y^*n}(x) = \frac{1}{n}\sum_{i=1}^{n} I\left(Y_i^* \le x\right)$$

$$F_{\hat{Y}n}(x) = \frac{1}{n}\sum_{i=1}^{n} I\left(\hat{Y}_i \le x\right)$$

81 Let $\hat{x}_p$ denote the decile corresponding to p. Find x_j and x_{j+1} by $F\left(x_j\right) \le p < F\left(x_{j+1}\right)$. Let np = j + g where j is the integer part of np, and g is the fractional part of np. Then

$$\hat{x}_p = \begin{cases} \frac{1}{2}\left(x_j + x_{j+1}\right) & if \quad g = 0 \\ x_{j+1} & if \quad g > 0 \end{cases}$$

82 To test if the consistency of the distribution between waves is the same for imputed and true values, a Chi-Square test can be used where the observed cell frequencies are the imputed cell frequencies and the expected cell frequencies are the true cell frequencies.

$$H_0 : \hat{n}_{ij} = n^*_{ij}$$

$$\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{\left(\hat{n}_{ij} - n^*_{ij}\right)^2}{n^*_{ij}} \qquad (9)$$

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

84 Another approach to assess how well an imputation method performs in a longitudinal sense is to analyse the impact of the imputation method on the movement of income between waves. To do this, movement estimates of income can be computed

$$\Delta Y^*_i = Y^*_{i2} - Y^*_{i1}$$

$$\Delta \hat{Y}_i = \hat{Y}_{i2} - \hat{Y}_{i1}$$

and then ΔY_i can be used as the variable of analysis in the previously mentioned measures under predictive accuracy, distributional accuracy and estimation accuracy.

- Are the evaluation criteria for comparing imputation methods appropriate?

8. Evaluation Methodology

85 The main objective of this project is to evaluate alternative imputation methods and to adopt the best method in the HILDA Survey. The previous section outlined the evaluation criteria for comparing imputation methods. This section outlines how the evaluation study was set up, in particular the process for modelling the response mechanism, and also includes the limitations of the study.

86 As it is impossible to compare the imputed values with the true values using the entire HILDA dataset because some true values are not reported, the evaluation was based on a subset of the HILDA data. More specifically, the evaluation was based on persons who responded and provided all income items in the waves they were eligible for and a sample of these cases were set to missing. That way, the actual responses were treated as the true values. There were 8,720 cases from the HILDA data for the evaluation. The sample of cases set to missing was based on modelling the response mechanism. The response mechanism was modelled in the case where the missing values are missing at random.

Modelling the Response Mechanism

87 Logistic regression analysis was used to investigate the relationship between the response probability and the explanatory variables.

88 The response indicator, denoted by R_i is defined as

$$R_i = \begin{matrix} 1 \\ 0 \end{matrix} \Big\} \begin{matrix} \textit{if individual i is a respondent} \\ \textit{if individual i is a non-respondent} \end{matrix}$$

89 In the case where the missing income values are missing at random, the response probability model is

$$P\left(R_i = 1 \middle| \mathbf{Z}_i\right) = \frac{1}{1+\exp\left(-\alpha - \mathbf{Z}_i\delta\right)}$$

90 In the case where the missing income values are not missing at random, that is the probability of response to the survey question on Y depends on Y (the variable being imputed), the response probability model is

$$P\left(R_i = 1 \middle| Y_i, \mathbf{Z}_i\right) = \frac{1}{1+\exp\left(-\alpha - \gamma Y_i - \mathbf{Z}_i\delta\right)}$$

where

$\mathbf{Z}_i$ is a 1 x m vector of characteristics of individual i
α and γ are scalar parameters
$\boldsymbol{\delta}$ is a m x 1 parameter vector

91 So far, only the missing at random response probability models have been constructed and the results that follow are based on this response mechanism.

Non-Responding Persons in Responding Households

92 For each wave, regression models were developed to predict the presence of response. The predicted values for all cases were calculated from the models. A random sample of cases were selected, proportional to the probabilities of response, and all income values were set to missing in line with the proportion of non-responding persons observed in the real HILDA data.

Responding Persons

93 The sample of non-responding persons was determined first, and the remaining sample became the responding persons.

94 For each wave and for each income variable being imputed, regression models were developed to predict the presence of response to that income item. Only cases that reported that they had that particular income were included in the models. The predicted values for all non-zero cases were calculated from the models.

95 For practical reasons, the response probability models were constructed by undertaking sequential modelling of multivariate missingness, to take into account the dependence between income variables. That is, a response probability model was constructed to predict the presence of response to wages and salaries. Then, a response probability model was constructed to predict the presence of response to benefits, which included modelling a response indicator to wages and salaries, and so on.

96 The response probability models were constructed using the entire HILDA dataset and the missing data were simulated on the subset of the HILDA data, also in a sequential manner. That is, a random sample of cases was set to missing for wages and salaries. Then, a random sample of cases was set to missing for benefits, which took into account previously simulated response indicators (in this case, previously simulated missing wages and salaries), and so on.

97 The missing data were simulated ten times, by generating different random starts, in order to produce ten different datasets for the evaluation. Once the simulated evaluation datasets were created, the missing data were imputed using each imputation method. The results from the ten imputed datasets were averaged to form a single set of results for presentation in this paper.

Limitations of the Evaluation

98 While the evaluation is as realistic as possible to the real HILDA environment, there are some limitations of the study that should be noted.

99 As the evaluation was conducted on a subset of HILDA data (that is, only the individuals who reported income in the waves they were eligible for), the HILDA survey weights were not able to be used in the evaluation. In addition, the evaluation

data consists of a larger proportion of the older Australian population and less of the younger population. This may be because the older population has a greater tendency to stay in the survey. For comparison, the average age of enumerated persons was 35 years in the real HILDA sample and the average age in the evaluation was 43 years (in wave 3).

100 Also, it is difficult to construct a realistic notion of households from the evaluation dataset. The average household size from the evaluation is about 1.5 compared to the average household size of 2.0 (excluding children) in the real HILDA data.

101 Another limitation of the evaluation is that, while it measures the overall accuracy of the imputation, it does not distinguish between imputation bias and imputation variance.

102 Nevertheless, even with these limitations of the study the evaluation data is still a useful basis for comparing imputation methods.

- Are there any comments on the evaluation strategy?

9. Comparison of Imputation Methods

103 This section presents the results and discusses the comparison of the imputation methods evaluated so far – the nearest neighbour regression method, the basic Little and Su method, and the extended Little and Su method.

104 Firstly, the performance of these imputation methods are looked at in a cross-sectional sense and then in a longitudinal sense.

105 Tables 1, 2 and 3 summarise the evaluation measures (1) to (7) for waves 1, 2 and 3 respectively. Bold table entries indicate which imputation method performed better for each income item against each of the evaluation criteria.

106 It is not easy to draw conclusions from these results, but we can begin to analyse the results by looking at the main income components – wages and salaries, Aust govt pensions and business income. For wages and salaries, the extended Little and Su method performs only slighter better in waves 1 and 2, but not in wave 3. The basic Little and Su method performed better in wave 3.

107 One reason why the extended Little and Su method does not perform better than the basic version may be because the restriction of the choice of donors to the matched age variable has the impact of increasing the imputation variance, especially for some imputation classes which have a large number of recipients and a small number of donors.

108 However, for Aust govt pensions the extended Little and Su method performed better, especially in wave 2. This means that the age groupings used in the imputation classes in Release 3.0 improved the imputation for Aust govt pensions because age is highly correlated with pensions and hence the donors and recipients match well. The reduction in bias could compensate for any variance increase resulting from stratifying by age. The reverse could be said for the investment income items.

109 The third major contributor to total FY income is business income, and the nearest neighbour regression method performed better for this income component. The results show that when the extended Little and Su method performs poorly, it can perform quite poorly. For example, in wave 1 the extended Little and Su method has performed quite poorly for business income as evidenced by the evaluation measures presented in Table 1. In particular, the estimation accuracy measure for variance differs greatly from those produced from the nearest neighbour regression method and the basic Little and Su method.

110 These comparisons indicate that the age groupings used in the imputation classes in Release 3.0 did not improve the imputation for some income components, such as dividends and royalties income and actually made the imputation worse compared to the basic Little and Su method. However, the age groupings did improve the imputation for other income components, such as Aust govt pensions. Further work is required in exploring improvements in the formation of better imputation classes.

111 Another way to analyse the results is to look at how the imputation methods performed against the groupings of evaluation measures – predictive accuracy, distributional accuracy and estimation accuracy.

112 Both the Little and Su methods performed better than the nearest neighbour regression method under the predictive accuracy criteria. Under the distributional accuracy criteria, the basic Little and Su method performed the worst. Yet, under the estimation accuracy criteria, the basic Little and Su method performed the best.

113 In comparing the benefits of the extended Little and Su method over the basic Little and Su method, both methods performed equally well against the predictive accuracy criteria, the extended Little and Su method performed better against distributional accuracy and not so well against estimation accuracy. In fact, the extended Little and Su method performed just as well as the nearest neighbour regression method under the estimation accuracy criteria. The reason for this would be to do with the imputation classes used, which would have led to restrictive pools of suitable donors. The problem could also be exacerbated by the imputed values not necessarily being restricted to the reported values of the donors, under the Little and Su method. Therefore, imputed values may fall outside the range of possible donor values, leading to some unusual imputed values. Under the Little and Su method, the imputed values take into account a residual effect (that is, any specific deviations from the general average) whereas under the nearest neighbour regression method, the imputed values are that of the donor.

114 Further improvements planned to assist in interpreting the results include producing standard errors of the evaluation measures to assess whether these results are statistically significant.

Table 1: Summary of Evaluation Measures for FY Income, Wave 1

	Predictive Accuracy		*Distributional Accuracy*	*Estimation Accuracy*			
Variable	*1*	*2*	*3*	*4*	*5*	*6*	*7*
RESPONDING PERSONS							
Wages and salaries							
NNRM	0.70	-3.38	0.08946	3,021	**123,358,877**	**1.16**	**14.53**
Basic Little & Su	**0.75**	-3.28	0.07412	1,986	254,238,013	1.88	23.00
Extended Little & Su	**0.75**	**-2.42**	**0.05719**	**1,388**	161,991,137	1.30	16.37
Aust govt pensions							
NNRM	**0.38**	**-1.11**	**0.16053**	620	**6,106,120**	**0.75**	2.24
Basic Little & Su	0.36	-1.17	0.17632	727	7,750,870	0.81	2.46
Extended Little & Su	**0.38**	-1.12	0.17368	**554**	6,262,113	0.77	**1.98**
Business income							
NNRM	0.17	**-1.38**	**0.12833**	**3,664**	**759,385,959**	3.78	24.03
Basic Little & Su	**0.39**	-2.10	0.14417	7,041	800,285,332	**3.03**	**16.07**
Extended Little & Su	0.25	-2.02	0.13417	6,549	3,006,626,655	4.16	25.56
Interest income							
NNRM	0.33	-3.07	**0.07285**	597	95,662,007	3.78	77.53
Basic Little & Su	0.51	**-2.91**	0.08076	504	68,901,694	2.98	58.40
Extended Little & Su	**0.56**	-3.28	0.08625	**370**	**48,652,287**	**1.92**	**40.66**
Dividends and royalties							
NNRM	0.33	-3.30	**0.05382**	417	26,837,477	1.94	28.49
Basic Little & Su	**0.43**	**-1.62**	0.06451	**198**	**18,980,315**	2.28	39.62
Extended Little & Su	**0.43**	-2.20	0.06947	247	28,295,026	**1.77**	**27.86**
Rent income							
NNRM	0.00	0.34	**0.12352**	2,683	746,198,659	3.82	33.23
Basic Little & Su	**0.12**	**-0.17**	0.20408	**1,574**	**573,999,214**	**2.77**	**25.59**
Extended Little & Su	0.05	-0.26	0.14807	2,021	593,801,876	3.19	29.35
Private pensions							
NNRM	0.35	-1.09	**0.16250**	3,907	453,657,975	1.86	10.87
Basic Little & Su	0.43	**-0.99**	0.17188	**2,574**	265,902,036	1.20	6.55
Extended Little & Su	**0.50**	-1.14	0.16875	3,356	**264,733,798**	**1.08**	**6.47**
Private transfers							
NNRM	**0.62**	-0.78	0.25714	984	**9,552,455**	**0.55**	**1.97**
Basic Little & Su	0.58	-0.70	0.25714	1,081	16,681,828	0.70	3.27
Extended Little & Su	0.60	**-0.51**	**0.25000**	**972**	22,881,050	0.76	3.35
Total FY income							
NNRM	0.74	-3.01	0.04471	1,378	**129,209,373**	**1.09**	**13.68**
Basic Little & Su	**0.79**	-3.67	0.04273	1,716	187,674,866	1.11	19.40
Extended Little & Su	0.72	**-2.94**	**0.03698**	**1,339**	424,551,854	3.26	50.40
NON-RESPONDING PERSONS							
Total FY income							
NNRM	0.59	-5.76	0.08846	1,124	309,009,920	1.48	**19.10**
Basic Little & Su	0.69	-4.57	0.05315	**1,054**	**221,970,486**	1.68	25.24
Extended Little & Su	**0.73**	**-3.98**	**0.04406**	1,111	403,262,406	**1.45**	28.48
ENUMERATED PERSONS							
Total FY income							
NNRM	0.68	-6.27	0.04199	**1,033**	183,456,105	**1.13**	**15.02**
Basic Little & Su	**0.75**	-5.37	0.02858	1,151	**146,355,578**	1.20	23.07
Extended Little & Su	0.71	**-4.59**	**0.02715**	1,037	392,482,853	3.65	87.79

Table 2: Summary of Evaluation Measures for FY Income, Wave 2

	Predictive Accuracy		*Distributional Accuracy*	*Estimation Accuracy*			
Variable	*1*	*2*	*3*	*4*	*5*	*6*	*7*
RESPONDING PERSONS							
Wages and salaries							
NNRM	0.56	-5.99	0.14582	5,950	165,819,992	0.94	**9.08**
Basic Little & Su	**0.78**	**-1.64**	**0.05719**	1,310	241,699,187	1.43	15.88
Extended Little & Su	**0.78**	-1.75	0.05953	**1,079**	**134,727,881**	**0.84**	9.68
Aust govt pensions							
NNRM	0.51	-0.83	0.17059	**585**	5,458,848	0.58	1.67
Basic Little & Su	0.50	-0.48	0.18529	817	9,640,589	1.13	2.96
Extended Little & Su	**0.52**	**-0.40**	**0.16471**	**585**	**3,572,887**	**0.36**	**1.19**
Business income							
NNRM	0.45	**-1.23**	**0.10490**	5,713	**3,061,241,258**	3.52	31.24
Basic Little & Su	**0.46**	-1.45	0.12888	6,088	3,120,248,488	3.68	25.59
Extended Little & Su	**0.46**	-1.59	0.13604	**5,195**	4,288,904,731	**2.99**	**30.51**
Interest income							
NNRM	0.43	-2.88	**0.06141**	425	18,041,864	2.11	39.61
Basic Little & Su	**0.56**	**-0.52**	0.07886	**270**	**12,972,961**	**1.17**	**21.54**
Extended Little & Su	0.50	-0.86	0.07886	325	30,999,657	1.56	27.35
Dividends and royalties							
NNRM	0.43	-3.17	**0.06598**	592	65,905,743	2.58	42.69
Basic Little & Su	0.54	**-1.00**	0.07911	**579**	**41,345,501**	**1.50**	**24.11**
Extended Little & Su	**0.55**	-1.32	0.07088	628	71,710,825	2.28	42.40
Rent income							
NNRM	0.20	**-0.35**	0.12895	1,228	129,395,554	2.88	11.06
Basic Little & Su	**0.42**	-0.98	0.14869	1,518	167,827,119	**2.27**	12.19
Extended Little & Su	0.31	-0.71	**0.11184**	**752**	**92,076,395**	2.45	**9.70**
Private pensions							
NNRM	0.10	-1.43	0.24583	8,295	1,882,551,126	1.33	6.81
Basic Little & Su	**0.23**	-1.60	**0.24167**	9,454	2,184,095,440	1.23	6.44
Extended Little & Su	0.19	**-1.39**	0.25417	**7,927**	**1,101,219,053**	**0.77**	**3.53**
Private transfers							
NNRM	0.38	-2.08	0.17752	1,058	**11,740,564**	**0.95**	**6.21**
Basic Little & Su	**0.45**	-0.84	0.15201	**734**	23,957,274	1.20	8.80
Extended Little & Su	0.41	**-0.83**	**0.14724**	1,113	41,752,362	1.28	9.60
Total FY income							
NNRM	0.79	-4.60	0.05631	2,093	462,628,023	3.48	**58.85**
Basic Little & Su	**0.81**	**-1.73**	0.03030	**970**	**410,512,455**	3.41	61.33
Extended Little & Su	**0.81**	-1.79	**0.02948**	990	620,215,313	**2.97**	63.89
NON-RESPONDING PERSONS							
Total FY income							
NNRM	0.19	**-1.71**	0.20896	9,397	1,262,526,163	**3.91**	99.56
Basic Little & Su	0.49	-4.13	0.04875	**1,503**	**1,205,682,907**	4.56	**99.23**
Extended Little & Su	**0.51**	-4.32	**0.04660**	1,968	1,622,250,597	4.99	112.52
ENUMERATED PERSONS							
Total FY income							
NNRM	0.58	**-3.00**	0.05624	2,312	640,209,992	3.75	**82.20**
Basic Little & Su	0.66	-4.33	**0.02389**	**1,017**	**615,225,734**	**3.39**	85.86
Extended Little & Su	**0.67**	-4.53	0.02470	1,235	849,748,769	4.13	124.47

Table 3: Summary of Evaluation Measures for FY Income, Wave 3

	Predictive Accuracy		*Distributional Accuracy*	*Estimation Accuracy*			
Variable	*1*	*2*	*3*	*4*	*5*	*6*	*7*
RESPONDING PERSONS							
Wages and salaries							
NNRM	0.66	-3.53	0.09474	2,745	**160,627,865**	**0.67**	**8.93**
Basic Little & Su	**0.73**	**-1.27**	**0.05301**	**1,437**	310,018,136	1.92	23.72
Extended Little & Su	0.70	-1.84	0.05940	1,720	330,976,197	2.07	25.77
Aust govt pensions							
NNRM	0.53	**0.15**	0.17105	748	6,516,823	0.61	1.66
Basic Little & Su	**0.62**	-0.70	**0.14737**	528	**3,734,095**	0.51	**1.34**
Extended Little & Su	**0.62**	-0.81	0.16052	**483**	4,295,123	**0.49**	1.44
Business income							
NNRM	0.26	**-0.35**	**0.08235**	**5,688**	2,387,965,534	5.21	34.18
Basic Little & Su	0.32	-0.89	0.12437	6,963	**1,705,057,094**	5.32	32.66
Extended Little & Su	**0.41**	-1.07	0.11597	7,537	3,464,728,075	**3.74**	**26.93**
Interest income							
NNRM	0.40	-1.70	**0.07380**	555	84,200,395	3.71	65.78
Basic Little & Su	0.64	**-1.04**	0.07904	475	64,439,226	**2.01**	**42.30**
Extended Little & Su	**0.67**	-1.14	0.07511	**385**	**58,544,778**	2.50	47.01
Dividends and royalties							
NNRM	0.35	-2.07	0.07359	937	99,141,465	2.17	30.68
Basic Little & Su	0.47	**-1.21**	**0.06794**	**486**	**58,160,501**	**1.38**	**19.44**
Extended Little & Su	**0.53**	-1.22	0.08027	871	140,050,978	3.05	51.36
Rent income							
NNRM	0.05	**-0.50**	0.12987	**1,213**	**57,381,476**	**2.80**	**16.17**
Basic Little & Su	**0.26**	-1.33	0.14935	3,442	3,084,942,918	3.12	21.39
Extended Little & Su	0.23	-0.70	**0.12857**	2,449	1,409,016,628	3.18	21.93
Private pensions							
NNRM	0.52	**-0.96**	0.24444	**11,327**	**2,588,977,979**	**1.52**	**7.10**
Basic Little & Su	0.54	-1.06	0.22222	15,905	4,313,933,576	1.61	7.54
Extended Little & Su	**0.56**	-1.12	**0.21667**	19,408	7,688,136,805	1.81	8.31
Private transfers							
NNRM	0.31	**-0.33**	0.16434	966	62,513,279	2.98	23.33
Basic Little & Su	0.39	-0.74	**0.15443**	566	**28,584,641**	**1.81**	**14.00**
Extended Little & Su	**0.42**	-0.86	0.16073	**313**	30,701,428	1.91	14.81
Total FY income							
NNRM	0.78	-1.90	0.03388	**1,237**	**584,238,671**	3.36	**55.77**
Basic Little & Su	0.78	**-1.23**	**0.03328**	1,931	780,213,397	3.94	75.97
Extended Little & Su	**0.79**	-1.75	0.03408	1,943	751,621,846	**2.58**	62.04
NON-RESPONDING PERSONS							
Total FY income							
NNRM	0.31	**-3.85**	0.12134	5,050	1,631,439,050	8.62	140.41
Basic Little & Su	**0.63**	-4.03	0.05473	**955**	**693,886,932**	**3.39**	**53.76**
Extended Little & Su	**0.63**	-4.25	**0.05455**	1,015	751,792,747	3.61	61.26
ENUMERATED PERSONS							
Total FY income							
NNRM	0.62	-4.13	0.04390	1,763	685,007,798	3.89	**57.55**
Basic Little & Su	0.73	**-4.09**	**0.02986**	**1,161**	639,735,488	3.33	75.18
Extended Little & Su	**0.74**	-4.47	0.02995	1,274	**514,252,423**	**2.50**	64.58

115 Next, let us look at the performance of the imputation methods in a longitudinal sense. Table 4 summarises the evaluation measures for total financial year income. For responding persons, the measures were computed on cases where they were respondents in at least one of the two waves; and for non-responding persons, the measures were computed on cases where they were non-respondents in at least one of the two waves. This means that a person could contribute to the results for both responding persons and non-responding persons. Bold table entries indicate which imputation method performed better against each of the evaluation criteria.

Table 4: Summary of Longitudinal Evaluation Measures, Total FY Income

	Predictive Accuracy		*Distributional Accuracy*	*Estimation Accuracy*			
Variable	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Wave 1 to Wave 2							
Responding persons							
NNRM	0.51	-8.95	0.11217	2,159	359,642,221	8.21	231.79
Basic Little & Su	**0.53**	**-5.92**	0.04291	694	**276,349,232**	**6.94**	**201.62**
Extended Little & Su	0.52	-6.26	**0.02911**	**464**	452,591,418	8.70	264.18
Non-responding persons							
NNRM	0.22	-5.90	0.21675	6,193	**392,439,408**	8.60	**174.63**
Basic Little & Su	0.19	**-4.49**	0.05424	962	581,877,351	**8.48**	178.47
Extended Little & Su	**0.28**	-5.00	**0.03737**	**786**	646,534,426	9.83	212.56
Enumerated persons							
NNRM	0.50	-9.09	0.11650	2,319	363,856,004	8.11	227.53
Basic Little & Su	**0.52**	**-5.96**	0.04248	677	**272,363,604**	**6.85**	**200.34**
Extended Little & Su	**0.52**	-6.40	**0.02843**	**454**	451,688,710	8.62	260.01
Wave 2 to Wave 3							
Responding persons							
NNRM	**0.50**	-12.48	0.10215	1,352	763,612,374	7.65	175.82
Basic Little & Su	0.49	-6.83	**0.03780**	902	**203,908,821**	8.41	132.39
Extended Little & Su	0.44	**-6.65**	0.03888	**662**	254,518,413	**7.33**	**130.76**
Non-responding persons							
NNRM	**0.31**	-11.80	0.16617	3,131	1,383,397,620	11.37	203.59
Basic Little & Su	0.25	**-5.13**	**0.04771**	1,003	**467,777,672**	**7.72**	128.56
Extended Little & Su	0.18	-5.30	0.05129	**1,000**	593,032,801	8.49	**111.45**
Enumerated persons							
NNRM	**0.49**	-12.62	0.10188	1,380	778,943,720	7.56	174.09
Basic Little & Su	0.47	**-6.83**	**0.03754**	872	**199,654,950**	8.17	**123.97**
Extended Little & Su	0.43	**-6.83**	0.03889	**655**	260,366,862	**7.45**	128.40
Wave 1 to Wave 3							
Responding persons							
NNRM	0.51	-7.16	0.10960	1,669	672,429,574	**4.40**	**85.15**
Basic Little & Su	**0.53**	**-2.49**	0.03225	679	**282,635,389**	4.55	103.70
Extended Little & Su	0.50	-3.44	**0.02453**	**545**	453,093,104	4.47	168.35
Non-responding persons							
NNRM	**0.28**	-5.23	0.19033	3,359	973,851,569	5.81	79.07
Basic Little & Su	0.27	**-1.62**	0.06072	**440**	**225,111,960**	**4.74**	**73.01**
Extended Little & Su	**0.28**	-2.65	**0.05396**	444	266,035,413	5.72	92.28
Enumerated persons							
NNRM	0.50	-7.23	0.11258	1,734	657,839,721	**4.55**	**84.94**
Basic Little & Su	**0.52**	**-2.54**	0.03250	636	**274,213,255**	4.71	104.64
Extended Little & Su	0.50	-3.58	**0.02455**	**508**	439,687,240	4.66	170.45

116 The results show that the nearest neighbour regression method does not perform very well in a longitudinal sense. The extended Little and Su method performs better against the distributional accuracy and estimation accuracy (measure 4) criteria. But, overall the basic Little and Su method has performed better, and in particular against the predictive accuracy (measure 2) and the estimation accuracy measure for variance (measure 5).

117 Next, we look at the cross-wave correlations produced by the imputed data and compare these to the true data (Table 5). Bold table entries indicate which correlation coefficients derived from the imputed data are closest to the true correlation coefficients.

118 Based on cross-wave correlations, both the basic and extended Little and Su methods perform better than the nearest neighbour regression method for movement estimates. It is interesting to note that the cross-wave correlations for non-responding persons based on the Little and Su methods are higher than the true correlations.

Table 5: Cross-Wave Correlations, Total FY Income (Evaluation Measure 8)

	Cross-Wave Correlations			
Variable	*true*	*NNRM*	*Basic L&S*	*Extended L&S*
Wave 1 to Wave 2				
Responding persons				
Total FY income	0.72	0.54	**0.71**	0.67
Non-responding persons				
Total FY income	0.73	0.51	**0.82**	**0.82**
Enumerated persons				
Total FY income	0.72	0.54	**0.71**	0.67
Wave 2 to Wave 3				
Responding persons				
Total FY income	0.72	0.51	0.69	**0.71**
Non-responding persons				
Total FY income	0.66	0.26	**0.66**	0.72
Enumerated persons				
Total FY income	0.72	0.50	0.69	**0.71**
Wave 1 to Wave 3				
Responding persons				
Total FY income	0.74	0.51	**0.70**	0.65
Non-responding persons				
Total FY income	0.77	0.44	0.80	**0.78**
Enumerated persons				
Total FY income	0.74	0.51	**0.70**	0.65

119 The final evaluation criterion addresses the distributional consistency between waves by considering the change in income decile group membership from one wave to another. Based on this evaluation measure, the results (in Table 6) clearly show that the nearest neighbour regression method does not preserve the consistency of the income distribution between waves (the calculated χ^2 exceeds the critical value). On the other hand, the Little and Su methods do preserve the consistency of the income distribution between waves. The critical value of χ^2 for $\alpha = 0.05$ and 81 degrees of freedom is 101.879.

Table 6: Chi-Square Test Statistics on Total FY Income Deciles (Evaluation Measure 9)

Imputation Method	*W1 to W2*	*W2 to W3*	*W1 to W3*
NNRM	354.85692	431.79889	156.88333
Basic Little & Su	57.667667	**61.517467**	65.857736
Extended Little & Su	**52.364103**	62.739884	**54.314304**

10. Conclusions

120 An assessment of the performance of the basic Little and Su method, the extended Little and Su method (adopted in Release 3.0) and the nearest neighbour regression method (adopted in Release 2.0) was conducted using data from the first three waves of the HILDA Survey. A set of evaluation criteria, based on the statistical properties of a good imputation method, were used to compare these imputation methods.

121 The results of this evaluation study did not identify an imputation method that consistently performed better against each of the evaluation measures and for each income item in each wave.

122 Overall, in a cross-sectional sense both the Little and Su methods perform better than the nearest neighbour regression method. In a longitudinal sense, the Little and Su methods perform much better when compared to the nearest neighbour regression method. Evidence shows that the Little and Su methods preserve the distribution of income between waves. Furthermore, the Little and Su methods perform better in maintaining cross-wave relationships and income mobility.

123 However, while the extended Little and Su method generally outperformed the nearest neighbour regression method, the imputation classes used were not always beneficial. The age groupings used in the creation of imputation classes in Release 3.0 did not improve the imputation for some income components, such as dividends and royalties income but did improve the imputation for other components, such as Aust govt pensions. Further work is required in exploring improvements in the formation of imputation classes.

124 Given that the objective was to find a suitable longitudinal imputation method and based on findings from the evaluation study, the recommended strategy for the income imputation for Release 4.0 is to continue to implement the Little and Su method with ongoing enhancements, such as improvements in the formation of imputation classes and an extension to allow for the simultaneous imputation of multiple missing items.

125 This paper reports on work in progress. Further work is planned or is already underway. The extended Little and Su method will be enhanced to allow for improvements in the formation of imputation classes. Other imputation methods will be evaluated, namely the last value carried forward method, the random carryover method and the population carryover method. In addition, some options for multivariate imputation for multiple missingness will be assessed. The evaluation criteria will need to be extended to include measures for assessing how well an imputation method preserves the relationships between income variables.

126 Further work may also be undertaken to evaluate and vary the missing at random assumption. Future investigations may include simulations evaluating a non-ignorable response mechanism. In addition, there may also be investigations into the use of multiple imputation to measure imputation variance.

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, 1997 (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, University of Southampton, Southampton, UK.

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

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

Little, R.J.A. (1988), *Missing Data Adjustments in Large Surveys*, Journal of Business & Economic Statistics, Vol. 6, No. 3, pp. 287-296.

Little, R.J.A., and Su, H.L. (1989) 'Item Non-Response in Panel Surveys' in Panel Surveys, edited by Kasprzyk, D., Duncan, G., and Singh, M.P., Wiley, New York.

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

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

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

FOR MORE INFORMATION...

INTERNET **www.abs.gov.au** the ABS web site is the best place for data from our publications and information about the ABS.

LIBRARY A range of ABS publications is available from public and tertiary libraries Australia wide. Contact your nearest library to determine whether it has the ABS statistics you require, or visit our web site for a list of libraries..

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our web site, or purchase a hard copy publication. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice..

PHONE 1300 135 070

EMAIL client.services@abs.gov.au

FAX 1300 135 211

POST Client Services, ABS, GPO Box 796, Sydney 2001

FREE ACCESS TO PUBLICATIONS

All ABS statistics can be downloaded free of charge from the ABS web site.

WEB ADDRESS www.abs.gov.au

2000001524244
ISBN 0 642 48179 2

RRP $11.00

Produced by the Australian Bureau of Statistics