Income Imputation in the Household, Income and Labour Dynamics in Australia (HILDA) Survey

Ms Rosslyn Starick
Australian Bureau of Statistics (on secondment to)
Melbourne Institute of Applied Economic and Social Research
161 Barry Street, University of Melbourne, Victoria 3010 AUSTRALIA
rosslyn.starick@abs.gov.au
r.starick@unimelb.edu.au

Abstract

This paper describes the income imputation method adopted in Release 3.0 of the HILDA data. The primary method for imputing income is based on a method developed by Little and Su (1989). It is an extension to the method used by the German Socio-Economic Panel. This longitudinal imputation method incorporates trend and individual level information into the imputed amounts and 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. The cross-sectional imputation method used was the nearest neighbour regression method adopted in Release 2.0 (similar to that used by the British Household Panel Study). This paper also presents a quantitative assessment of how well the Little and Su method performs against the nearest neighbour regression method.

Introduction

Non-response is a common source of non-sampling error in surveys. Non-respondents may have differing characteristics to respondents and this may result in non-response bias. 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. To reduce the effect of non-response bias, weighting adjustments and imputation are used.

In the Household, Income and Labour Dynamics in Australia (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.

The objective of this research was to find a suitable longitudinal imputation method to deal with item non-response with a focus on missing income data. Income data is considered a sensitive topic to collect in surveys and respondents may refuse to provide information on income or just may not know the answers to some of the income questions. Therefore, income data is subject to higher levels of item non-response than other topics.

This paper describes the income imputation method adopted in Release 3.0 of the HILDA data. The primary method for imputing income is based on a method developed by Little and Su (1989). It is an extension to the method used by the German Socio-Economic Panel. This longitudinal imputation method incorporates trend and individual level information into the imputed amounts and 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. The cross-sectional imputation method used was the nearest neighbour regression method adopted in Release 2.0 (similar to that used by the British Household Panel Study).

Also presented in this paper is a quantitative assessment of how well the Little and Su method performs against the nearest neighbour regression method. In finding an appropriate imputation methodology for use in longitudinal surveys such as HILDA, an evaluation study was undertaken to assess the performance of the Little and Su method and the nearest neighbour regression method against a set of criteria.

Imputation Methods

Nearest Neighbour Regression Method

The income imputation for Release 2.0 of the HILDA data was implemented using a nearest neighbour regression method, Watson (2004). The predicted values from a regression model for the variable of interest were used to identify the nearest case whose reported value could be inserted into the case with the missing value.

Little and Su Method

The income imputation provided in Release 2.0 was deficient in a number of ways as outlined in Watson (2004). Therefore, the income imputation methodology was revised for Release 3.0.

The primary method for imputing income is based on a method developed by Little and Su (1989). It is an extension to the method used by the German Socio-Economic Panel. This longitudinal imputation 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).

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 Little and Su methodology, therefore, was extended to take into account the characteristics of the donors and recipients. Donors and recipients are matched within imputation classes which have 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+.[1] The formulae for the Little and Su method are provided in Appendix 1, together with a worked example.

The Little and Su method requires at least one non-zero income estimate to exist before it can be applied. As a result, there is still a need to use a cross-sectional imputation method for some cases. For example new entrants interviewed in wave 3 who did not respond to some income questions, the imputation method used was the nearest neighbour regression method adopted in Release 2.0.

It is important to note that there is a fundamental flaw with the 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 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.

[1] 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.

Missing Income Data and the Extent of Income Imputation

The number and proportion of cases with missing income data in Release 3.0 are provided in Table 1. For most income variables, the proportion of missing income falls each wave. The reason for the decline in the proportion of missing income may be because respondents are becoming more comfortable with the survey. The variables with the highest proportion of missing cases are still business income, investments and private transfers.

Table 2 shows how much of the mean income was imputed, for each wave and income component. For respondents with item non-response, 6.1 percent of total financial year income was imputed in wave 3, compared to 7.4 percent in wave 2 and 8.0 percent in wave 1. Including the imputed income totals for non-respondents within responding households (but excluding children), the percentage of total financial year income imputed for enumerated persons is 14.6 percent in wave 3.

This shows that while approximately one in eight responding persons are missing some component of financial year income, only one sixteenth of the mean income comes from imputed values and the remainder is from reported values. At the household level, one in four households are missing some component of financial year income and one seventh of the mean income is from imputed values.

Table 1: Number and proportion of cases with missing income data, waves 1, 2 & 3

	Wave 1		*Wave 2*		*Wave 3*	
Variable	*Number of missing cases*	*Prop'n of cases, %*	*Number of missing cases*	*Prop'n of cases, %*	*Number of missing cases*	*Prop'n of cases, %*
RESPONDING PERSONS (non-zero cases only)						
Current income						
Wages and salaries	462	6.0	310	4.2	275	3.8
Benefits	136	3.2	80	2.1	74	1.9
Financial year income						
Wages and salaries	666	7.9	550	6.9	434	5.7
Aust govt pensions	67	1.5	52	1.2	54	1.3
Foreign govt pensions	1	0.5	3	1.4	0	0.0
Business income	404	29.1	366	28.6	354	25.4
Interest income	661	19.5	596	18.6	424	13.9
Dividends and royalties	584	14.6	521	14.5	402	11.8
Rent income	240	20.3	189	15.3	181	14.3
Private pensions	59	6.2	41	4.6	29	3.2
Private transfers	28	7.1	89	23.1	72	19.9
Total FY income	2054	15.6	1817	14.7	1462	12.1
Windfall income						
Windfall income	32	4.1	31	2.9	39	3.3
ENUMERATED PERSONS (zero and non-zero adult cases)						
Total FY income	3212	21.2	2795	19.9	2337	17.2
Windfall income	1190	7.9	1009	7.2	914	6.7
HOUSEHOLDS (zero and non-zero cases)						
Total FY income	2243	29.2	2009	27.7	1700	24.0
Windfall income	838	10.9	723	10.0	662	9.3

Table 2: Proportion of mean income imputed, waves 1, 2 & 3

	Wave 1		*Wave 2*		*Wave 3*	
Variable	*Mean*	*Prop'n of mean imputed, %*	*Mean*	*Prop'n of mean imputed, %*	*Mean*	*Prop'n of mean imputed, %*
RESPONDING PERSONS						
Current income (per week)						
Wages and salaries	400	5.2	411	3.9	426	3.5
Benefits	49	2.8	51	2.1	53	1.2
Financial year income						
Wages and salaries	21,049	5.8	21,748	4.8	22,603	3.9
Aust govt pensions	2,181	1.4	2,498	1.0	2,589	1.3
Foreign govt pensions	61	0.0	83	0.9	68	0.0
Business income	1,618	30.8	1,905	28.1	1,712	31.8
Investments	1,582	23.1	1,633	25.8	1,695	18.3
Private pensions	1,119	7.0	1,534	5.7	1,540	3.0
Private transfers	132	10.9	160	37.7	154	16.4
Total FY income	27,742	8.0	29,561	7.4	30,362	6.1
Windfall income						
Windfall income	310	3.1	1,415	1.2	1,670	3.1
ENUMERATED PERSONS (excluding children)						
Total FY income	27,807	16.1	29,714	16.0	30,220	14.6
Windfall income	314	12.4	1,549	18.5	1,625	9.1
HOUSEHOLDS						
Total FY income	57,775	16.1	60,932	16.0	62,629	14.9
Windfall income	457	12.4	2,196	18.2	2,537	9.1

Attributes of a Good Imputation Method

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 the Little and Su method against the nearest neighbour regression method.

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.

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

Predictive Accuracy

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

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

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

The imputation procedure should maximise the preservation of analysis. That is, it should reproduce the lower order moments[2] 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).

[2] Moments are mainly used to approximate the probability distribution of a random variable. All the moments for a random variable can be packaged into one expression, called a moment-generating function. A moment-generating function is a mathematical way to show if two random variables have the same probability distribution (Mendenhall, Wackerly and Scheaffer 1990).

Imputation Plausibility

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

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.

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.

Method of Evaluation

An evaluation was undertaken to assess how well the nearest neighbour regression method and the Little and Su method perform against a set of criteria. 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.

Rubin (1976) describes three assumptions that can be made about the types of non-response that can occur: missing completely at random (MCAR), missing at random (MAR) and not missing at random. MCAR means the probability of the variable being missing is independent of the values of that variable and of any other variable. MAR means the probability of the variable being missing is independent of the values of that variable, but is dependent on the variables that are not missing. Not missing at random means the probability of the variable being missing depends on the values of that variable. Most approaches to dealing with non-response assume the data are at least missing at random.

The missing data were simulated ten times 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.

Table 3 shows the number and proportion of cases set to missing in the evaluation data which was made to be as close as possible to the real HILDA data (the actual proportions of missing cases were previously presented in Table 1).

Table 3: Number and proportion of cases with missing income data, evaluation data

	Wave 1		*Wave 2*		*Wave 3*	
Variable	*Number of missing cases*	*Prop'n of cases, %*	*Number of missing cases*	*Prop'n of cases, %*	*Number of missing cases*	*Prop'n of cases, %*
RESPONDING PERSONS (non-zero cases only)						
Current income						
Wages and salaries	219	6.0	169	4.2	162	3.8
Benefits	75	3.2	53	2.1	49	1.9
Financial year income						
Wages and salaries	313	7.9	299	6.9	266	5.7
Aust govt pensions	38	1.5	34	1.2	38	1.3
Foreign govt pensions	na	na	na	na	na	na
Business income	120	29.1	112	28.4	119	25.4
Investments						
Interest income	291	19.5	298	18.6	229	13.9
Dividends and royalties	262	14.6	267	14.5	211	11.8
Rent income	86	20.2	76	15.3	77	14.3
Private pensions	32	6.2	24	4.6	18	3.2
Private transfers	14	7.3	42	22.8	49	19.1
Total FY income	920	14.2	926	13.2	804	10.9
Windfall income						
Windfall income	13	4.0	17	2.9	24	3.3
ENUMERATED PERSONS (zero and non-zero adult cases)						
Total FY income	1492	20.0	1484	18.6	1343	16.0
Windfall income	585	7.8	575	7.2	563	6.7
HOUSEHOLDS (zero and non-zero cases)						
Total FY income	1355	27.6	1353	26.1	1238	23.0
Windfall income	559	11.4	552	10.7	541	10.0

na not assessed

A sample of cases for foreign government pensions was not set to missing because of the small number of instances where item non-response occurred for this variable.

Limitations of the Evaluation

While the evaluation is as realistic as possible to the real HILDA environment, there are some limitations of the study that should be noted. Table 4 compares *unweighted* mean incomes estimated from the 8,720 evaluation cases of HILDA data (before a sample of cases were set to missing) with the *weighted* mean incomes from the real HILDA data.

Table 4: Comparison of mean income between real HILDA data and evaluation data, waves 1, 2 & 3

	Wave 1		*Wave 2*		*Wave 3*	
Variable	*Real data*	*Eval'n data*	*Real data*	*Eval'n data*	*Real data*	*Eval'n data*
RESPONDING PERSONS						
Current income (per week)						
Wages and salaries	400	381	411	399	426	405
Benefits	49	57	51	59	53	59
Financial year income						
Wages and salaries	21,049	19,864	21,748	20,669	22,603	21,285
Aust govt pensions	2,181	2,580	2,498	2,938	2,589	2,999
Foreign govt pensions	61	na	83	na	68	na
Business income	1,618	1,010	1,905	1,202	1,712	1,318
Investments	1,582	1,340	1,633	1,218	1,695	1,284
Private pensions	1,119	1,175	1,534	1,483	1,540	1,593
Private transfers	132	136	160	123	154	153
Total FY income	27,742	26,183	29,561	27,730	30,362	28,717
Windfall income						
Windfall income	310	286	1,415	1,280	1,670	1,569
ENUMERATED PERSONS (excluding children)						
Total FY income	27,807	26,600	29,714	27,867	30,220	28,850
Windfall income	314	294	1,549	1,262	1,625	1,546
HOUSEHOLDS						
Total FY income	57,775	40,434	60,932	43,018	62,629	44,884
Windfall income	457	446	2,196	1,947	2,537	2,405

na not assessed

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).

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.

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

Evaluation Criteria for Comparing Imputation Methods

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.

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

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.

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

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

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.

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.

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

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

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.

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.

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

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

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

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).

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

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

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

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 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.

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.

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.

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

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

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

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

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

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.

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

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

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

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.

Results

This section presents the results and discusses the comparison of the nearest neighbour regression method and the Little and Su method. Firstly, the performance of these imputation methods are looked at in a cross-sectional sense and then in a longitudinal sense.

Tables 5, 6 and 7 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.

Table 5: 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**
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
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**
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
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
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
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
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**
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**
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**
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**
Extended Little & Su	**0.71**	**-4.59**	**0.02715**	1,037	392,482,853	3.65	87.79

Table 6: 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**
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
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
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
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
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
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
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**
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**
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**
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**
Extended Little & Su	**0.67**	-4.53	**0.02470**	**1,235**	849,748,769	4.13	124.47

Table 7: 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**
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
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
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
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**
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**
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**
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
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**
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
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**
Extended Little & Su	**0.74**	-4.47	**0.02995**	**1,274**	**514,252,423**	**2.50**	64.58

Table 8 shows the comparison between the nearest neighbour regression method and the Little and Su method. These results show us which method performed better, the method adopted in Release 2.0 or the method adopted in Release 3.0.

The Little and Su method performed better than the nearest neighbour regression method for

- wages and salaries in all waves
- Aust govt pensions in wave 2 and wave 3, but not wave 1
- interest income in all waves
- dividends and royalties in wave 1 and wave 2, but not wave 3
- rent income in wave 1 and wave 2, but not wave 3
- private pensions in wave 1 and wave 2, but not wave 3

On the other hand, the nearest neighbour regression method performed better than the Little and Su method for

- business income in wave 1 and wave 3, but not wave 2
- private transfers in wave 1 and wave 2, but not wave 3

For total financial year income for responding persons, the Little and Su method performs better for wave 2 and poorer for waves 1 and 3; for non-responding persons, the Little and Su method performs better for waves 1 and 3 and poorer for wave 2; and for enumerated persons, the Little and Su method performs better for wave 3 and poorer for waves 1 and 2 (how well a method performs for enumerated persons seems to be primarily driven by how well it performs for non-responding persons).

Overall, the Little and Su method has performed better for each wave. However, the results show that when the Little and Su method performs poorly, it can perform quite poorly. For example, in wave 1 the Little and Su method has performed quite poorly for business income as evidenced by the evaluation measures presented in Table 5. In particular, the estimation accuracy measures differ greatly from those produced from the nearest neighbour regression method. A separate comparison of the Little and Su method used in Release 3.0 to the original Little and Su method which doesn't use imputation classes was undertaken but not reported in detail here. However, this comparison indicated 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 original Little and Su method. However, the age groupings did improve the imputation for other income components, such as wages and salaries and Aust govt pensions. Further work is required in exploring improvements in the formation of better imputation classes.

From these cross-sectional results, the Little and Su method has outperformed the nearest neighbour regression method.

Table 8: Cross-Sectional Comparison of NNRM and Extended Little & Su Method

	Predictive Accuracy		*Distributional Accuracy*	*Estimation Accuracy*			
Variable	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Wave 1							
Responding persons							
Wages and salaries	EL&S	EL&S	EL&S	EL&S	NNRM	NNRM	NNRM
Aust govt pensions	--	NNRM	NNRM	EL&S	NNRM	NNRM	EL&S
Business income	EL&S	NNRM	NNRM	NNRM	NNRM	NNRM	NNRM
Interest income	EL&S	NNRM	NNRM	EL&S	EL&S	EL&S	EL&S
Dividends & royalties	EL&S	EL&S	NNRM	EL&S	NNRM	EL&S	EL&S
Rent income	EL&S	EL&S	NNRM	EL&S	EL&S	EL&S	EL&S
Private pensions	EL&S	NNRM	NNRM	EL&S	EL&S	EL&S	EL&S
Private transfers	NNRM	EL&S	EL&S	EL&S	NNRM	NNRM	NNRM
Total FY income	NNRM	EL&S	EL&S	EL&S	NNRM	NNRM	NNRM
Non-responding persons							
Total FY income	EL&S	EL&S	EL&S	EL&S	NNRM	EL&S	NNRM
Enumerated persons							
Total FY income	EL&S	EL&S	EL&S	NNRM	NNRM	NNRM	NNRM
Wave 2							
Responding persons							
Wages and salaries	EL&S	EL&S	EL&S	EL&S	EL&S	EL&S	NNRM
Aust govt pensions	EL&S	EL&S	EL&S	--	EL&S	EL&S	EL&S
Business income	EL&S	NNRM	NNRM	EL&S	NNRM	EL&S	EL&S
Interest income	EL&S	EL&S	NNRM	EL&S	NNRM	EL&S	EL&S
Dividends & royalties	EL&S	EL&S	NNRM	NNRM	NNRM	EL&S	EL&S
Rent income	EL&S	NNRM	EL&S	EL&S	EL&S	EL&S	EL&S
Private pensions	EL&S	EL&S	NNRM	EL&S	EL&S	EL&S	EL&S
Private transfers	EL&S	EL&S	EL&S	NNRM	NNRM	NNRM	NNRM
Total FY income	EL&S	EL&S	EL&S	EL&S	NNRM	EL&S	NNRM
Non-responding persons							
Total FY income	EL&S	NNRM	EL&S	EL&S	NNRM	NNRM	NNRM
Enumerated persons							
Total FY income	EL&S	NNRM	EL&S	EL&S	NNRM	NNRM	NNRM
Wave 3							
Responding persons							
Wages and salaries	EL&S	EL&S	EL&S	EL&S	NNRM	NNRM	NNRM
Aust govt pensions	EL&S	NNRM	EL&S	EL&S	EL&S	EL&S	EL&S
Business income	EL&S	NNRM	NNRM	NNRM	NNRM	EL&S	EL&S
Interest income	EL&S	EL&S	NNRM	EL&S	EL&S	EL&S	EL&S
Dividends & royalties	EL&S	EL&S	NNRM	EL&S	NNRM	NNRM	NNRM
Rent income	EL&S	NNRM	EL&S	NNRM	NNRM	NNRM	NNRM
Private pensions	EL&S	NNRM	EL&S	NNRM	NNRM	NNRM	NNRM
Private transfers	EL&S	NNRM	EL&S	EL&S	EL&S	EL&S	EL&S
Total FY income	EL&S	EL&S	NNRM	NNRM	NNRM	EL&S	NNRM
Non-responding persons							
Total FY income	EL&S	NNRM	EL&S	EL&S	EL&S	EL&S	EL&S
Enumerated persons							
Total FY income	EL&S	NNRM	EL&S	EL&S	EL&S	EL&S	NNRM

-- indicates that the evaluation measures are the same for both imputation methods

Next, let us look at the performance of the imputation methods in a longitudinal sense. Table 9 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 9: 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**
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**
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**
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
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
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
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**
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**
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**
Extended Little & Su	0.50	**-3.58**	**0.02455**	**508**	**439,687,240**	4.66	170.45

Table 10: Longitudinal Comparison of NNRM and Little & Su Method

	Predictive Accuracy		*Distributional Accuracy*	*Estimation Accuracy*			
Variable	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Wave 1 to Wave 2							
Responding persons							
Total FY income	EL&S	EL&S	EL&S	EL&S	NNRM	NNRM	NNRM
Non-responding persons							
Total FY income	EL&S	EL&S	EL&S	EL&S	NNRM	NNRM	NNRM
Enumerated persons							
Total FY income	EL&S	EL&S	EL&S	EL&S	NNRM	NNRM	NNRM
Wave 2 to Wave 3							
Responding persons							
Total FY income	NNRM	EL&S	EL&S	EL&S	EL&S	EL&S	EL&S
Non-responding persons							
Total FY income	NNRM	EL&S	EL&S	EL&S	EL&S	EL&S	EL&S
Enumerated persons							
Total FY income	NNRM	EL&S	EL&S	EL&S	EL&S	EL&S	EL&S
Wave 1 to Wave 3							
Responding persons							
Total FY income	NNRM	EL&S	EL&S	EL&S	EL&S	NNRM	NNRM
Non-responding persons							
Total FY income	--	EL&S	EL&S	EL&S	EL&S	EL&S	NNRM
Enumerated persons							
Total FY income	--	EL&S	EL&S	EL&S	EL&S	NNRM	NNRM

-- indicates that the evaluation measures are the same for both imputation methods

Table 10 shows the comparison of the nearest neighbour regression method and the Little and Su method. In a longitudinal sense, the Little and Su method performs better than the nearest neighbour regression method, especially against the predictive accuracy, distributional accuracy and estimation accuracy (measure 4) criteria. Given that income data are highly skewed, the regression measure for predictive accuracy (2) may be more appropriate to use than the correlation coefficient measure (1).

Next, we look at the cross-wave correlations produced by the imputed data and compare these to the true data (Table 11). Bold table entries indicate which correlation coefficients derived from the imputed data are closest to the true correlation coefficients.

Based on cross-wave correlations, the Little and Su method performs 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 method are higher than the true correlations.

Table 11: Cross-Wave Correlations, Total FY Income (Evaluation Measure 8)

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

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 12) 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 method does 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 12: Chi-Square Test Statistics on Total FY Income Deciles (Evaluation Measure 9)

Imputation Method	*Wave 1 to Wave 2*	*Wave 2 to Wave 3*	*Wave 1 to Wave 3*
NNRM	354.85692	431.79889	156.88333
Little & Su	**52.364103**	**62.739884**	**54.314304**

Conclusions

An assessment of the performance of the 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.

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.

Overall, the Little and Su method outperforms the nearest neighbour regression method in a cross-sectional sense, and it is even more obvious that the Little and Su method performs better in a longitudinal sense. Evidence shows that the Little and Su method preserves the distribution of income between waves and is also better at predicting estimates of change between waves (although the calculated t-test statistics exceed the critical value, the t-values are lower for the Little and Su method). Furthermore, the Little and Su method performed better in maintaining cross-wave relationships and income mobility.

However, the Little and Su method did not consistently perform better against the estimation accuracy criteria in preserving unbiased parameter estimates of interest, such as mean income. This may be because under the Little and Su method, the imputed values are not necessarily restricted to the reported values of the donors and therefore 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.

In a cross-sectional sense, the Little and Su method performs better than the nearest neighbour regression method for most income components, such as wages and salaries but not for other income components, such as business income. The age groupings used in the creation of imputation classes in Release 3.0 did not improve the imputation for some income components but did improve the imputation for other components. Further work is required in exploring improvements in the formation of imputation classes.

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.

Further Research

Other Imputation Methods to Evaluate

Besides exploring what improvements can be made to the Little and Su method, other imputation methods worth evaluating include:

- last value carried forward (the method used by the Canadian Survey of Labour and Income Dynamics)
- random/population carryover method (the method used by the US Survey of Income and Program Participation), Williams and Bailey (1996)

Multivariate Imputation for Multiple Missingness

Both the Little and Su method and the nearest neighbour regression method impute one variable at a time. Therefore, if a recipient has more than one variable missing, it is likely that the imputed values may have come from different donors. A definite enhancement to the imputation procedure would be to modify the imputation method to impute a subset of variables simultaneously. The Little and Su method already imputes multiple missing waves simultaneously using the same donor, but could be modified to impute multiple missing items simultaneously. The evaluation criteria would need to be extended to include measures for assessing how well an imputation method preserves the relationships between income variables.

Imputing Income Components for Non-Respondents

The imputation process can be extended to impute the income components for non-respondents within responding households. The current approach is to impute only the income totals, which means that only income totals are available at the household level. The current set of evaluation criteria should provide a reasonable assessment of the quality of the imputation of non-respondents.

Varying the Response Mechanism

How much does the response mechanism matter? Further work may be undertaken to compare imputation methods when the response mechanism is not missing at random.

References

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

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

Mendenhall, W., Wackerly, D.D., and Scheaffer, R.L. (1990), *Mathematical Statistics with Applications* (4^{th} edition), PWS-KENT, Boston

Rubin, D.B. (1976), *Inference and Missing Data*, Biometrika, Vol.63, pp. 581-590

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

Acknowledgements

The author wishes to thank Frank Yu for his valued input and guidance in this research. In addition, the author gratefully acknowledges the valuable comments and suggestions provided by Nicole Watson, Paul Sutcliffe and members of the HILDA Technical Reference Group - Peter Boal, John Henstridge, Stephen Horn and Frank Yu.

Disclaimer

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). The views expressed are those of the author and do not necessarily reflect those of the MIAESR or the ABS.

Appendix 1 – Little and Su Method

Formulae

The 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).

The Little and Su method was implemented as follows:

(a) Column (wave) effects of the form

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

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

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

(b) Row (person) effects of the form

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

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 $\bar{Y}_{h}^{(i)}$, 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

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

$$= Y_{hlj}\frac{\bar{Y}_{h}^{(i)}}{\bar{Y}_{h}^{(l)}}$$

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.

Example

Suppose we have the following small sample of fictitious responses to current wages and salaries.

All cases

OBS	Wages & Salaries		
	Wave 1	Wave 2	Wave 3
1		400	420
2	675	235	700
3	345	690	800
4	200	480	210
5	200		
6	350	370	
7	400	450	470
8	0	790	790
9	360	450	600
10	135	130	200

From this example, we see that observation 1 did not respond to the current wages and salaries question in wave 1, but provided responses in subsequent waves. Observations 5 and 6 also partially responded and wages and salaries information are not provided in all 3 waves.

The first step in the Little and Su method is to calculate the column effects based on complete cases only. Complete cases were defined as individuals that were interviewed in all 3 waves and responded in all 3 waves for the variable of interest. In this example, the complete cases are:

Complete cases

OBS	Wages & Salaries		
	Wave 1	Wave 2	Wave 3
2	675	235	700
3	345	690	800
4	200	480	210
7	400	450	470
8	0	790	790
9	360	450	600
10	135	130	200

The column effects are calculated using formula (a) above and are computed to be:

Column effects

OBS	Wages & Salaries		
	Wave 1	Wave 2	Wave 3
1		400	420
2	675	235	700
3	345	690	800
4	200	480	210
5	200		
6	350	370	
7	400	450	470
8	0	790	790
9	360	450	600
10	135	130	200
	0.70	**1.06**	**1.24**

The Little and Su method incorporates trend information into the imputed amounts via the column effects. In this example, the wave 1 column effect of 0.70 indicates that the mean current wages and salaries in wave 1 is 30% lower than the overall mean current wages and salaries, and the means in waves 2 and 3 are 6% and 24% higher than the overall mean, respectively.

Next, the row effects are calculated using formula (b) above and are computed to be:

Row effects

OBS	Wages & Salaries			
	Wave 1	Wave 2	Wave 3	
1		400	420	**357**
2	675	235	700	**585**
3	345	690	800	**596**
4	200	480	210	**303**
5	200			**287**
6	350	370		**425**
7	400	450	470	**459**
8	0	790	790	**460**
9	360	450	600	**475**
10	135	130	200	**159**
	0.70	**1.06**	**1.24**	

The sample is then ordered by the row effects, and the closest donor is identified.

Sorted by row effects

OBS	Wages & Salaries			
	Wave 1	Wave 2	Wave 3	
10	135	130	200	**159**
5	200			**287**
4	200	480	210	**303**
1		400	420	**357**
6	350	370		**425**
7	400	450	470	**459**
8	0	790	790	**460**
9	360	450	600	**475**
2	675	235	700	**585**
3	345	690	800	**596**

Once the closest donor has been identified, the missing value is imputed by multiplying the actual value for the variable of interest of the donor with the row effect of the recipient divided by the row effect of the donor.

In this example, the imputed current wages and salary amounts using the Little and Su method are highlighted below.

Impute missing values

OBS	Wages & Salaries		
	Wave 1	Wave 2	Wave 3
10	135	130	200
5	200	455	199
4	200	480	210
1	236	400	420
6	350	370	436
7	400	450	470
8	0	790	790
9	360	450	600
2	675	235	700
3	345	690	800