25th Australasian Transport Research Forum
Incorporating the BTRE Transport Policy Colloquium
Canberra 2-4 October 2002

A comparison of two methods for imputing missing income from household travel survey data

Min Xu , Michael Taylor and Anthea Page
Transport Systems Centre, University of South Australia

Abstract

Personal and household income are essential information in the development of activity-based travel demand forecasting models. Such models seek to explain travel behaviour and incorporate land use and socio-demographic information. However, it is common to find that
the income response rate is usually relatively low in a household travel survey, or similar, data set. One way of resolving this problem entails using a method of replacing the missing data in the data set before incorporating it into the model. Several techniques for imputing missing data have been outlined in previous literature. This paper introduces the process for repairing missing income data by two methods – hot-deck imputation and regression imputation – and provides comparisons of the two methods. The 1999 Metropolitan Adelaide Household Travel Survey (99MAHTS) data set is used for this process. The 99MAHTS data set includes the complete travel activities of participants, including the entire household, over a continuous two-day period, together with associated socio-demographic information. A simple object-oriented computer program developed for use in undertaking hot-deck imputation is outlined. The paper concludes by comparing some statistics obtained from the repaired (using hot-deck imputation and regression imputation) and unrepaired data sets.

Contact author

Min Xu
PhD candidate
Transport Systems Centre, University of South Australia
GPO Box 2471
Adelaide SA 5001

Tel: (08) 8302 2212
Fax: (08) 8302 1880
Email: min.xu@postgrads.unisa.edu.au

GO BACK TO START SCREEN GO BACK TO INDEX

Introduction

Social and economic characteristics are two of the main factors that influence travel behaviour. Personal and household income are commonly seen as essential elements in the development of activity-based travel demand forecasting models. Household travel survey data are often used to gain income information, however it is common to find that the income response rate is usually relatively low in household travel survey, or similar, data sets. Excluding households or persons whose income is missing from the travel behaviour analysis could lead to the loss of significant useful information and may also result in an inadequate sample size. Therefore it is necessary to find a solution to this problem by using a method to replace the missing data in the data set before incorporate it into the model.

This paper introduces the procedure for repairing missing income data by two methods – hot-deck imputation and regression imputation. Simple statistical analysis is employed to demonstrate some results of the comparisons between the repaired data set and unrepaired data set within each method, and between the two methods. The statistics also provide an insight into how the methods work. The 1999 Metropolitan Adelaide Household Travel Survey (99MAHTS) data set is used for this process. The 99MAHTS data set includes the complete travel activities of participants, including the entire household, over a continuous two-day period, together with associated socio-demographic information. An object-oriented computer program was developed to carry out the imputation processes.

The paper commences with a literature overview of the various methodologies used for data imputation. It then provides a brief introduction to the 99MAHTS data set. The process of data imputation using two methods, hot-deck and regression imputation are then outlined. The paper presents and compares some statistical results from the two imputation methods, and recommends a preferred imputation method.

Imputation methodologies

Data imputation means assigning a value to replace missing items in a data set. It is performed when a reasonable answer by inference can not be made from the observation (Lessler and Kalsbeek, 1992). Three primary methods are used to deal with missing data. Firstly, the data can be discarded. Secondly, a value may be inferred by presuming it from other related facts. For example, if a husband's age is missing, a value could be assigned by the inference based on the wife's age. Last, a value may be imputed.

The main aims of data imputation are both statistical and practical. The statistical aim of imputation is to minimise the mean square error of survey data set estimates. The mean square error has both a variance and a bias element. All imputation procedures increase the variance of estimates, but a better imputation method increases the variance less than a poorer one. A better

imputation method is able to correctly guess the missing value of individual items, which leads to a reduction in variance and bias of the estimation. The practical aim of imputation is the efficiency gained in using all of the information in a survey data set.

There are a number of ways to impute massing data, including mean imputation, hot-deck imputation, cold-deck imputation, regression imputation and multiple imputation (Armoogum and Madre, 1997). The problems associated with the mean imputation will not be discussed in this paper, but are detailed elsewhere. For example, see Ford, 1980, Rubin, 1987, or Dudala and Stopher, 2001. The multiple imputation method is somewhat different from the other methods listed above. Multiple imputation employs multi-stage methodology, whereby a number of different values are imputed to create a number of independent estimate sets. Analyses based on resulting sets of estimates are then averaged. The final result may be more reliable than one resulting from a single imputed value of missing data, as it averages over imputation error (Rubin, 1987).

Imputation methods applied to replace missing values in household travel survey data have been receiving more attention in recent years. Richardson and Loeis (1997) reported on imputation of income data using the regression method. The data used for the imputation was from the Victorian Activity and Travel Survey (VATS) in which socio-demographic information was provided by the respondents. A regression model was built to explain personal income based on four variables, age, sex, work status and occupation. A stochastic model was then used to estimate the personal income of the respondents who did not provide their income within each formulated group by the four variables. The results from the imputed data and data directly supplied by the respondents showed that imputed incomes cover the full range of income categories, with most imputed income values around the middle of the range of incomes. A notable difference was that only few zero incomes resulted from the imputed incomes, however approximately 33 per cent of reported incomes were zero. Richardson and Loeis concluded that the regression based imputation method preserved the variance inherent in the original income distribution.

Dudala and Stopher (2001) described their research work involving repairing missing items as well as entire person and trip records for travel survey data using the hot-deck imputation method. The Baton Rouge Personal Transportation Survey data collected in 1997 was used for this study. Firstly, the paper outlined the hot-deck imputation procedure for repairing missing items, in which missing income imputation was illustrated. The statistical results of reported incomes, imputed incomes by the mean imputation and by the hot-deck imputation were compared. A similar procedure to impute missing person and trip records was also detailed. In addition, a test was run in which complete data were changed by deliberately removing certain data items, and subsequently these data were replaced using the same procedure. The results suggested that hot-deck imputation provided estimates that were closer to the actual value than the other imputation methods (e.g. mean income imputation). The hot-deck imputation method has substantial potential for repairing transport

survey data, both for individual items and entire records. The study also emphasised the importance of applying inference for some missing data before imputations are carried out.

Imputing missing income for 99MAHTS

The 1999 Adelaide Household Travel Survey was conducted between 29 March and 31 July 1999 inclusive. The respondent households were randomly selected across the Adelaide metropolitan area. Interviews were then undertaken for the selected households in the survey. All the members of each household were asked to provide details of all their travel activities they made over two consecutive days, including where they went, at what time, and for what purpose, the forms of transport they used, as well as socio-demographic information.

The 99MAHTS data are arranged into five tables. The household table comprises 5886 households, with 14,004 persons included in the person table. The day table, stop table and disable table coving travel and activities, and related household information. After connecting the household table, the person table and the disable table, there were a total of 5732 households, in which the number of cars in the household is recorded. Within the 5732 households, 407 households had at least one person record (entire person record) missing. Based on the household type, if the missing person record was a parent or there are more missing persons than the number of persons recorded in a household, the household was excluded from the data set. Thus 317 households were excluded using this criterion. A total of 5415 households then remained for use in the income data analysis. Inference is employed prior to the commencement of the imputation to repair some missing or incorrect data in the database using related facts. For example, the result of data examination for the age difference of partners showed that 78.3 per cent of the age difference between partners is 0-5 years and 16.5 per cent of the age difference between partners is 6-10 years, which indicated that if a wife's age is missing or incorrect, then the husband's age could be used to correct her age, and vice-versa. Table 1 shows the percentages of the age difference between partners by the couple families within the data set.

Table 1. The percentages of the age difference between partners

Age difference group	< -5	-5 - -1	0 - 5	6 - 10	11 - 15	>15
Percentage	2.6%	12.1%	66.3%	14.4%	3.9%	0.7%

Two techniques are used for income imputation, the hot-deck method and the regression method. In hot-deck imputation, missing data items are obtained by finding values from similar respondents in the same survey. Two procedures are applied, one on household income, and the other on person income. A computer program is written in Delphi enabling household or person to be

picked randomly from the appropriate category of the donor file, which then is paired with the missing income household or person in the same category of the recipient file. In the regression method, a regression equation is estimated from the data set and then used to predict the variable to be imputed from other variables within the data set.

Hot-deck imputation performed on household income

According to the hot-deck imputation method of replacing missing items that was developed by Dudala and Stopher (2001), households may be separated into two groups. Households with complete person records or with complete main person records (main income supporter) are allocated to the donor file. Households with missing data are allocated to the recipient file. Based on analysis of variance (ANOVA) and correlation analyses, the households are categorized by 0, 1, 2 and 3+ vehicles and by 0, 1, 2 and 3+ workers in both the donor and recipient files. If, for example, a recipient household is one with 1 car and 1 worker, then the computer program randomly selects a donor household from that category, and the income of a household is assigned to replace the missing value of income in the recipient household. In the 99MAHTS database there are 4295 donor households and 1120 households with missing income. Table 2 shows a contingency table of the frequencies of households in terms of the number of workers per household and the household vehicle ownership for both the donor and recipient households. Results of the descriptive statistics from the hot-deck imputation and donor data are shown in Table 3.

Table 2: Frequencies of Donor and Recipient households in terms of workers and vehicle ownership (recipient household frequencies in parenthesis)

	Workers				
Cars	0	1	2	3+	Total
0	370 (89)	46 (5)	3 (0)	2 (0)	421 (94)
1	936 (258)	690 (162)	190 (26)	3 (2)	1819 (448)
2	189 (69)	553 (140)	781 (190)	39 (9)	1562 (408)
3+	18 (8)	99 (36)	250 (69)	126 (57)	493 (170)
Total	1513 (424)	1388 (343)	1224 (285)	170 (68)	4295 (1120)

Table 3: Descriptive statistics for household income before and after imputation (hot-deck imputation based on household incomes)

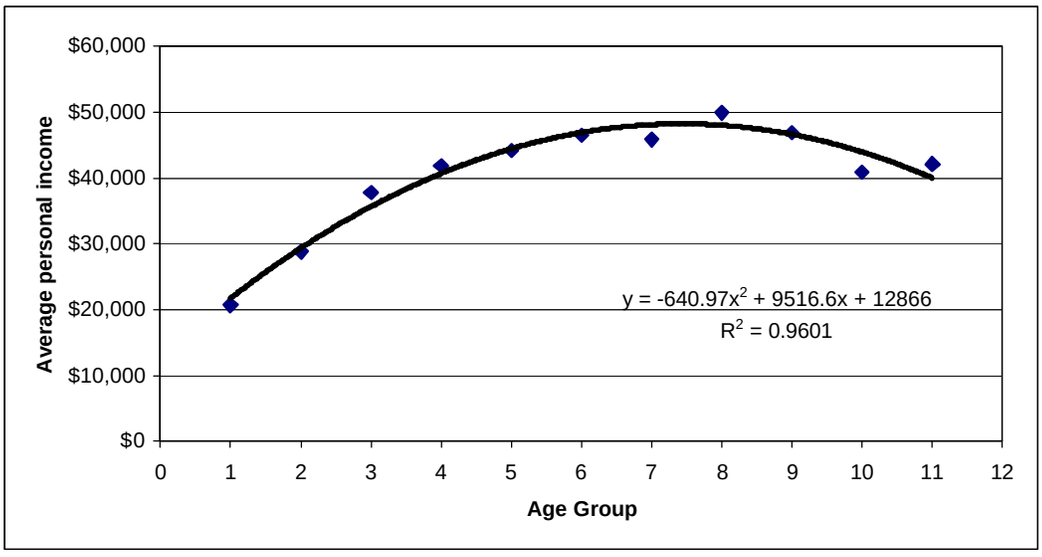
	Before Imputation	**After Imputation**
Mean	41717	42075
Standard Error	465	421
Median	36400	36400
Mode	9360	9360
Standard Deviation	30484	30982
Sample Variance	929303960	959908646
Kurtosis	2.241666861	2.036813697
Skewness	1.30265931	1.28017578
Range	196240	196240
Minimum	1040	1040
Maximum	197280	197280
Sum	179172378	227834798
Count	4295	5415
Largest(1)	197280	197280
Smallest(1)	1040	1040

Regression imputation performed on person income

The method of regression imputation used for this study is borrowed from Richardson and Loeis (1997). In the 99MAHTS data set 8706 persons have income recorded, 2245 person incomes are not available, and several income items are left blank, resulting in about 21 per cent of persons' income missing. Persons with complete income records and those with missing income records are allocated to the separate files. The demographic information - age, sex, work status (full-time or part-time) and occupation - are used in the construction of employed persons' categories. Based on the person's occupation, as provided by ASCO (Australian Standard Classification of Occupations), persons are then categorized into four groups in each occupation class, male full time, male part time, female full time and female part time. A regression model is used to identify average personal income as a function of the age group. Therefore, there are eleven age groups in each group. Figure 1 is an example of data for average income for full time male professionals.

Due to limited sample size, it is impossible to illustrate consistent relationships between age and income for any of the activity categories for those not in the paid workforce, so the average income for activity status categories, in terms of housekeeping, aged pensioner (including other pensioner) and retired pensioner, and gender are applied. After average incomes for each age/gender/occupation group and each gender/activity (not in the paid workforce) group are calculated in the recorded groups, then each income

Figure 1: Average income for full time professional males



record has been standardised by dividing the incomes by the mean income for that group. These standardised incomes for all age, gender, and work status are plotted on two graphs, Figure 2 shows the distribution of income ratios for employed persons and Figure 3 is the distribution of income ratios for unemployed persons. As the sample data for the housekeeping category is limited, male and female are shown as one category in the data plots.

Figures 2 and 3 illustrate that all the cumulative curves are sigmoid and skewed to the right with a long tail to the distribution. Full time male workers and full time female workers have similar curves. Part time male and part time female workers have almost the same distribution. Male and female aged pensioners (including other pensioners) and retirees also have similar curves. A theoretical model, the gamma distribution is proposed for describing and capturing the distribution. The form of the gamma distribution is as follows:

$$F(x) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{-x/b}$$

Where x: ratio of income to group mean income
a, b: model parameters
Γ: the gamma function

Figure 2: Distribution of income ratios for employed persons

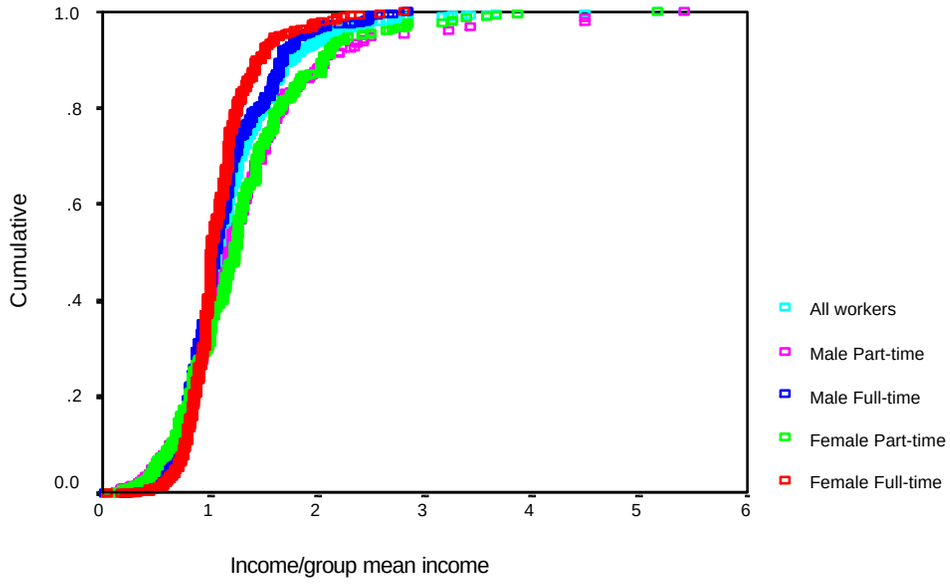
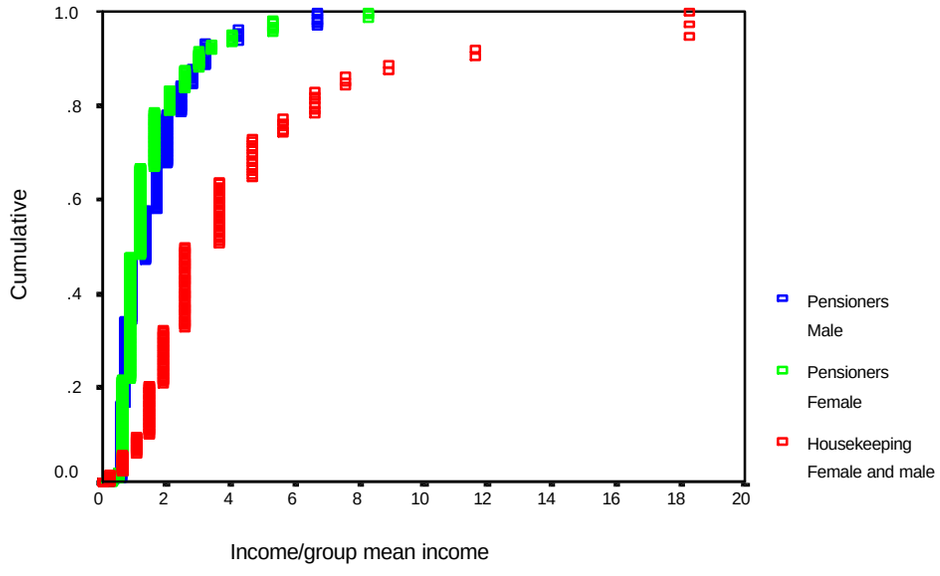


Figure 3: Distribution of income ratios for non-employed persons



Different values of model parameters *a* and *b* are tried until the best fit is found for the cumulative curve. Figure 4 is the gamma distribution of income ratios for both full time and part time workers. Figure 5 is the gamma distribution of income ratios for both housekeepers and pensioners.

As the mean incomes have been calculated for the respondents with a missing income in each group, the gamma distribution is used to randomly sample a value of income ratio, which is then multiplied by the estimated mean income to

obtain a value of imputed income. The calculated income values replace the missing incomes of the respondents. With 2245 missing person incomes in the 99MAHTS data set, 2073 persons receive a replacement income using this procedure, while 172 persons are excluded from the data set as they are also missing work status or occupation categories.

Figure 4: Gamma distribution of income ratios for workers

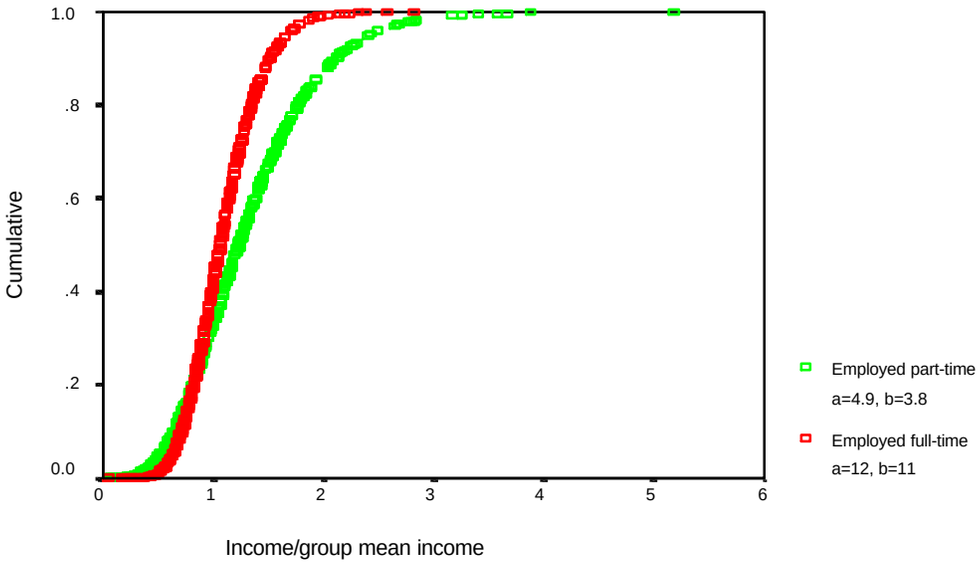
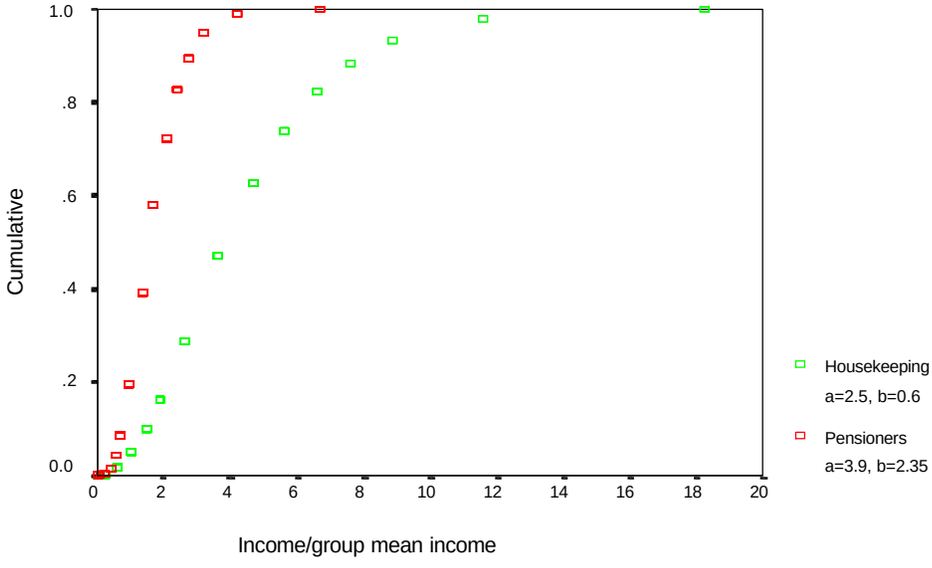


Figure 5: Gamma distributions of income ratios for non-employed persons



Following this procedure, households with complete person records and those with missing records are separated into two files. There are 4198 households with complete person income records and 1207 households with at least one person's income missing. After all the missing income persons receive an income replacement, they are then reallocated to their household in order to calculate the household income. Finally 1187 households have income replaced, with 20 households excluded due to other missing information (e.g. 127 persons were also missing work status or occupation). Table 4 shows the results of the descriptive statistics from regression imputation and recorded income households.

Table 4: Descriptive statistics for household income before and after imputation (regression)

	Before Imputation	**After Imputation**
Mean	41481	43968
Standard Error	472	425
Median	35620	38220
Mode	9360	9360
Standard Deviation	30586	31215
Sample Variance	935517250	974358180
Kurtosis	2.168424581	2.103102782
Skewness	1.290486952	1.206886222
Range	197280	274280
Minimum	0	0
Maximum	197280	274280
Sum	174138639	236768375
Count	4198	5385
Largest(1)	197280	274280
Smallest(1)	0	0

Comparison of the results from Hot-deck imputation and regression imputation

In Table 3 and Table 4, the results of donor data and hot-deck imputation show how the imputation methods modified household descriptive statistics. Both methods change the mean, but not the mode. Median and maximum values are altered in the regression imputation but not in the hot-deck imputation. The standard deviation and variance are only slightly increased and the skewness is slightly decreased for both methods. The regression imputation yields relatively larger increases in the standard deviation and variance, whilst the skewness is relatively largely decreased by regression imputation compared to the hot-deck imputation. This suggests that the hot-deck imputation preserves the standard deviation and variance better. This maybe because the values imputed from the regression method are higher than the reported income. Zero income is difficult to be imputed, while in the housekeeping group, there are considerable

zero incomes reported. This phenomenon was also found in Richardson and Loeis (1997). Kurtosis in both methods is slightly decreased, indicating that both methods make the data slightly flatter. Overall, the statistical results indicate that hot-deck imputation better preserves the basic characteristics of the donor data set.

Hot-deck imputation performed on person income

The previous analysis has used imputation (hot-deck) on household incomes directly. Further examination of the database suggests that in many instances missing household income results from the partial completion of income data by household members (i.e. some members of the household reported their personal income whilst others did not) rather than by complete omission of household income (which is, after all, merely the sum of the personal incomes of household members). This suggests that an alternative procedure is to impute missing personal income. This will use more of the reported data (i.e. the personal incomes reported by household members in households for which other members did not provide incomes). This could result in an overall improvement of the imputation process.

The first stage of the data process for this procedure is the same as the regression imputation method that is presented before, although the process used is hot-deck imputation at the person level, not regression. Persons with completed income records and those with missing income records first are allocated to the separate files. Four variables, age, sex, work status (full-time or part-time) and occupation are used in the construction of employed person's categories. Based on the person's occupation, persons are categorized into four groups (in both donor and recipient files) in each occupation class, for male full time, male part time, female full time and female part time. The computer program is then used to randomly choose an income of a person from a donor group, being the same group of the recipient person. For example, if a recipient person is a professional male, aged 23 and working full time, the computer program randomly selects a donor person from that group, and the income of that person is assigned to replace the missing value of income to the recipient person. Persons who are not in the paid workforce (such as those on pensions and welfare benefits) are categorized into housekeeping and aged pensioner and retired pensioner categories. In applying this procedure, 2073 persons receive an income replacement, after which they are re-allocated to their household in order to calculate the household income. Finally, 1187 households have their income replaced. Table 5 is the results of the descriptive statistics from hot-deck imputation and recorded income households.

Table 5: Descriptive statistics for household income before and after imputation (Hot-deck imputation based on person incomes)

	Before Imputation	**After Imputation**
Mean	41481	41889
Standard Error	472	420
Median	35620	36400
Mode	9360	9360
Standard Deviation	30586	30856
Sample Variance	935517251	952089823
Kurtosis	2.168424581	2.279030659
Skewness	1.290486952	1.277164511
Range	197280	242780
Minimum	0	0
Maximum	197280	242780
Sum	174138639	225574578
Count	4198	5385
Largest(1)	197280	242780
Smallest(1)	0	0

Comparison of hot-deck imputation on household income and on person income

The descriptive statistics from Table 3 and Table 5 indicate that hot-deck imputation performed on household income and on person income produces similar results. The mean is altered, but not the mode. The standard deviation, variance and skewness are quite close to the donor data in both hot-deck imputations, while they show only slight increases in the standard deviation and variance, and a slight decrease in the skewness. Kurtosis based on household income, shown in Table 3 is slightly decreased, while based on person income, shown in Table 5, is slightly increased. This suggests that hot-deck imputation performed on household income results in the data being slightly flatter, and hot-deck imputation performed on person income results in the data being slightly more peaked. The values of the changes in the hot-deck imputation performed on person income are relatively smaller when compared to the hot-deck imputation performed on household income. This indicates that hot-deck imputation performed on person income is even better at preserving the basic characteristics of the donor data set, which means the imputed values should be closer to the true values.

Conclusions

This paper demonstrates three procedures, the hot-deck imputation on household income, the hot-deck imputation on person income and the regression imputation, used for missing income data imputation. Results of the descriptive statistics from the imputations indicate that the three procedures are all suitable methods for missing data imputation, with hot-deck imputation having more merit than the regression method. Hot-deck imputation produces estimates that are closer to the true values. In addition, hot-deck imputation is easier to use as after the initial data process, the computer program carries out the imputation process. The regression method is especially time consuming in finding the values of the model parameters *a* and *b* in the Gamma distribution to make the best fit of the curve. Therefore, the hot-deck imputation method is the preferred method for missing data imputation.

The results demonstrate that having sufficient descriptive variables for categorisation is essential, they can help to produce very small variance within-category. This is the likely explanation as to why hot-deck imputation on person income produced better results than the hot-deck imputation on household income, as the former has four variables to describe person's income, while the second only has two variables to describe household's income. This suggests that hot-deck imputation would produce better estimates for replacing missing data if more person and household characteristics can be collected in the survey. The method of person income hot-deck imputation also makes better use of the available data.

The data from the 99MAHTS showed that households with more income earners have a greater chance of having a person's income missing. Therefore it is important to impute missing person income before combining them into the household data set. Otherwise it will result in the mean income being lower than it actually is as these households are placed into households with their income recorded, but actually at least one person has their income missing, although they may be not a main income earner of the household.

Further research is being undertaken to examine the use of imputation for other missing data items, such as car ownership. Tests of the methods using the Dudala-Stopher method of recreating observed data records by artificially removing some data items is also under consideration.

Acknowledgement

This study is part of a research project supported by a 'Linkage-Industry' collaborative research grant provided by the Australian Research Council (ARC), with the industry partner being the South Australian Department of Transport, Urban Planning and the Arts, which includes the agencies Transport SA, Planning SA and the Passenger Transport Board. The authors wish to thank the staff of these agencies for their kind cooperation and interest in the project.

References

Armoogum, J and Madre, J L (1997) Item sampling, weighting and non-response *International Conference on Transport Survey Quality and Innovation*, Grainau, Germany.

Dudala, T and Stopher, P (2002) Survey data repair using hot-deck imputation procedure *The Proceedings of the 8th Transportation Planning Methods Applications Conference of the TRB*

Ford, B (1980) An overview of hot-deck procedures, pp 185-206 of Madow, W L, Olkin I and Rubin, D B (eds) *Incomplete data in sample survey 2*, Academic Press, New York

Lessler, J and Kalsbeek, W (1992) *Nonsampling error in surveys,* New York, NY: Wiley and Sons, Inc

Richardson A and Loeis M (1997) Estimation of missing income in household travel surveys *Papers of the 21st Australasian Transport Research Forum*, Adelaide

Rubin, D (1987) *Multiple imputation for nonresponse in surveys,* New York, NY: Wiley and Sons, Inc.