Estimation of Missing Income Data in Household Travel Surveys.

A. J. Richardson, Director, Transport Research Centre, RMIT University
&

M. Loeis, Research Scholar, Transport Research Centre, RMIT University

ABSTRACT

Personal or household income is often useful information for the development of transport demand models or for the evaluation of the equity impacts of transport policies. Unfortunately, the response rate to the income question in a household travel survey is often relatively low. As a result, the resulting data set has a significant number of records with missing income information. To fully utilise such a data set, it is therefore necessary beforehand to account for the income values in those incomplete records.

This paper describes a method for estimating personal income based on other attributes of the survey respondents. It uses data from the 1994 and 1995 Victorian Activity and Travel Survey (VATS) in which respondents were asked to provide socio-demographic information such as age, sex, occupation, employment status and personal income. First, for records containing complete information on all the above variables, a model was developed to describe average personal income in terms of combinations of the other four variables. A stochastic model was then used to estimate the personal income of respondents who did not disclose their income but who provided information on some or all of the other four variables.

It was found that, for any combination of sex, occupation and employment status, a second-degree polynomial equation, with age as the independent variable, could be constructed to estimate average personal income. It was also found that the distribution of personal income could also be described in terms of an analytical probability distribution function, thus enabling stochastic estimates of missing income to be made. The technique promises to be transferable to other travel surveys which collect the necessary socio-economic variables, either as a means of imputing missing income values or as a way of estimating income values for all respondents when the income was not asked of respondents.

Richardson, A.J. and Loeis, M. (1997). "Estimation of Missing Income in Household Travel Surveys". Forum Papers, 21st Australasian Transport Research Forum, Vol. 1, pp 249-266.

1. INTRODUCTION

Personal or household income is often useful information for the development of transport demand models or for the evaluation of the equity impacts of transport policies. Radbone (1994), for example, uses income to evaluate the equity implications of public transport usage in Adelaide, but warns about the necessity of having a good measure of income with which to undertake the analysis. Unfortunately, the response rate to the income question in a household travel survey is often relatively low. As a result, the resulting data set has a significant number of records with missing income information. To fully utilise such a data set in modelling or evaluation, it is therefore necessary beforehand to account for the income values in those incomplete records.

This paper describes a method for estimating personal income based on other attributes of the survey respondents. The data for this analysis is drawn from the Victorian Activity & Travel Survey (VATS) being conducted by the Transport Research Centre. VATS is an ongoing survey using a mail-out/mail-back self-completion questionnaire technique which has been developed and used over many years in Australia and overseas by members of the Transport Research Centre (Richardson and Ampt, 1995). The survey records all travel by all modes by all people in the responding households in the survey sample. In addition, respondents were asked to provide a range of socio-demographic information in addition to their personal income. Each household was asked to provide their travel and activity data information for a specified travel day. The survey is continuous, covering all 365 days of the year, thereby enabling temporal variations in activity patterns to be observed. It is intended to continue the VATS survey for at least five years, generating an expected total response of about 25,000 households in the first five years. The VATS survey began in December 1993 and has collected information from about 5000 responding households in each of the financial years from 1993-94 through 1996-97. The information being used in this paper is from the period January 1994 through December 1995.

The paper first examines the overall question of item non-response in household travel surveys, and then describes the methods of dealing with such item non-response. It then describes an imputation method developed for the VATS data, and shows how this method has been applied to the data. It concludes by showing the results of the imputation process, by comparison with the income data supplied directly by respondents.

2. ITEM NON-RESPONSE IN HOUSEHOLD TRAVEL SURVEYS

Non-response is a common problem in household travel surveys. This non-response can take one of three major forms; specific item non-response, non-reported trips, and unit non-response. Specific item non-response occurs where the respondent has provided answers to most questions, but has failed to answer a specific question (e.g. the time of arriving at a particular destination). Non-reported trips are more extreme in that the respondent fails to tell us anything about specific trips or activities. This often occurs with respect to short duration trips by non-motorised modes (e.g. walking to the sandwich shop at lunch time). Unit non-response occurs where, for example, an entire household fails to respond to the survey. While the latter two forms of non-response are particularly serious, they have been well covered in other publications (e.g. Richardson and Ampt, 1994; Polak, Ampt and Richardson, 1995; Richardson, Ampt and Meyburg, 1996). The current paper pays particular attention to the first form of non-response; item non-response.

While attention to good design and the use of follow-up quality control techniques can be used to minimise item non-response, there will always be some respondents who fail to provide complete information for all questions in the interview or questionnaire. As noted by Zmud and Arce (1997), item non-response is the result of five major factors, namely:

- ***Knowledge and recall.*** Sometimes, respondents are asked questions to which they simply do not know the answer. More frequently, respondents suffer memory lapses which produce missing or low quality data. Memory lapses can include the forgetting of minor events, such as certain trips or activities, or incorrect recall of these events. Events may sometimes be well recalled but the lapses affect the sequence or order of events, or their exact timing.

- ***Comprehension***. Often, questions are difficult to understand and to answer. As Sheatsley (1983) observes;

  "Because questionnaires are usually written by educated persons who have a special interest in and understanding of the topic of their inquiry, and because these people usually consult with other educated and concerned persons, it is much more common for questionnaires to be overwritten, overcomplicated, and too demanding of the respondent than they are to be simpleminded, superficial, and not demanding enough."

  Writing clear and simple questions requires attention being paid to four important principles: simple language, common concepts, manageable tasks, and widespread information. The need for intensive work on the part of question-writers is exacerbated because of the disproportionate distribution of comprehension gaps (e.g. among those from non-English speaking backgrounds).

- ***Perceived and/or real burden***. Stopher and Metcalf (1996) point out that the trend in household travel surveys in the 1990s has been an increasing level of detail being asked from respondents. They ask “how much more can people be asked to report or even whether the point of asking for too much information has already been passed.” In an age of increasing demands upon peoples’ time from all types of sources (e.g., family, job, community), perceptions about the time burden inflicted by travel surveys is an important factor.

- ***Desire for privacy or concerns about personal information***. Questions about a household's trip-making behaviour can be as threatening to some respondents as questions about gambling, drinking alcohol, or sexual activities. These questions are considered threatening for a variety of reasons, including fear of consequences of divulging data or distrust of the person asking the questions. This may be a special concern for persons living alone, who may feel threatened by divulging when they are at home, or when their house is empty.

- ***Deliberate mis-reporting***. For a variety of reasons, respondents may be tempted to give deliberately inaccurate answers. One motivation, as outlined above, can be the fear of consequences. In addition, the desire to present oneself in a favourable light or to give a good impression may be a strong motivation for some respondents. Most frequent is the desire to report socially acceptable behaviour or to not report socially sanctioned behaviour.

For all the above reasons, item non-response is to be expected in all household travel surveys, to a greater or lesser degree. However, the extent of item non-response will vary between surveys, and it will also vary between different variables in the one survey. For the VATS surveys, the levels of item non-response associated with a range of variables are shown in Table 1 for 1994 and 1995.

It can be seen that most variables have a level of item non-response in the range of 0% to 3%. Only two variables have levels of non-response that are significantly higher: the provision of a phone number and personal income. The provision of a phone number was asked to enable follow-up phone interviews to be done where questions existed concerning the data. In the circumstances, a response rate to this question of over 85% was considered very satisfactory. The relatively high item non-response rate for income is in line with expectations, but is less than has been observed in many other surveys where a non-response rate of 10-20% has often been observed. It can be seen that, overall, the level of item non-response was lower in 1995 than it was in 1994 (1.7% compared to 2.4%). This is primarily due to the learning effect from conducting the survey on a continuous basis. Greater attention was paid to telephone call-back interviews in 1995 to obtain missing information. In addition, some questions were re-designed in 1995 based on the experience from the 1994 survey. The question on bicycle ownership is a case in point here, with re-positioning of the question reducing the non-response from 6.0% to 1.2%. However, despite more careful attention to design and editing, some level of item non-response will always remain. Accounting for this item non-response is the subject of the rest of this paper.

Table 1 Item Non-Response for VATS94 and VATS95

	Item Non-Response %	
ITEM	**VATS94**	**VATS95**
Household Variables		
Dwelling Type	1.7	1.9
Ownership of Dwelling	2.5	2.6
Number of Passenger Vehicles	0.6	0.8
Number of Motorcycles	0.6	1.0
Number of Other Vehicles	0.6	1.0
Number of Bicycles	6.0	1.2
Phone Ownership	2.7	2.6
Provision of Phone Number	14.1	14.0
Person Variables		
Year of Birth	3.4	2.5
Sex	0.8	0.5
Relationship to Oldest Person	1.1	0.7
Country of Birth	1.7	1.5
Licence Holding	2.5	1.6
Employment Status	1.7	1.5
Educational Status	1.7	1.5
Other Activities	1.7	1.5
Occupation	3.1	2.4
Industry	3.7	2.9
Start of Day Location	3.0	1.9
Start Time for First Trip	3.4	1.7
Reason for No Travel	0.3	0.2
Personal Income	7.3	7.1
Vehicle Variables		
Make	1.8	1.4
Model	4.5	2.6
Year of Manufacture	3.2	2.3
Number of Cylinders	3.9	3.0
Company Car?	2.4	1.9
Stop Variables		
Origin Location	1.2	0.3
Origin Place Type	0.3	0.1
Origin Purpose	0.5	0.2
Destination Location	1.5	0.3
Destimnation Place Type	0.6	0.1
Destination Purpose	0.9	0.2
Type of Goods Purchased	0.2	0.1
Mode	0.7	0.3
Travel Time	2.8	2.8
Name of Bus Operator	0.6	0.1
Was Household Car Used	2.0	1.0
Household Vehicle Number	2.7	1.5
People in Car	2.7	1.4
Parking Place	3.4	1.9
Parking Fee Paid	3.6	2.0
Walk Time from Parking Place	4.0	2.1
Ticket Type	0.8	0.3
Ticket Zones	0.8	0.4
Ticket Fare Type	0.9	0.6
Start Time	2.0	2.1
End Time	2.0	2.3
Average	2.4	1.7

3. DEALING WITH ITEM NON-RESPONSE

Given that there is some level of item non-response in the data set, the question remains as to what can be done to account for this non-response. Essentially, there are four courses of action that can be taken to deal with item non-response:

- ***Ignore Missing Data***. The simplest option, and the one used most often, is to simply ignore the missing values on a case-by-case basis. That is, for each analysis (such as a frequency distribution, a cross-tabulation or a regression model) a record is ignored if it has a missing value for any of the required variables. This has the side-effect that totals of distributions and cross-tabulations will be different because different records will have been omitted from each calculation. Unless the level of item non-response is high, or unless a large number of variables are used in the analysis, this effect will not be significant.

- ***Remove Records with any Missing Data***. Because it is easier to deal with a "clean data matrix" (i.e. one that does not have any missing data), one way of achieving this is to remove all records with any missing data, thereby ensuring that the data matrix contains no missing data. This form of data editing is, however, rather extreme and wasteful of data. Romios (1996) illustrates this factor using the VATS data collected in the first half of 1994. Analysis of this data shows that there were 5,043 households which responded to the survey, providing 13,546 Person records, 7,753 Vehicle records and 50,515 Stop records. However, if all household records in which there were any missing data were removed, this would reduce the available sample significantly, as shown in Table 2. This "record censoring" results in over half the data being removed from the full data set in order to obtain a "clean data set".

Table 2 Loss of Survey Sample Size via Record Censoring

Total Number of Household, Person, Vehicle and Stop Records	Total Number of Perfect Household, Person, Vehicle and Stop Records.	% Loss of Sample
5,043 Households	2,487 Households	50.7%
13,546 Persons	6,100 Persons	54.9%
7,753 Vehicles	3,584 Vehicles	53.8%
50,515 Stops	21,782 Stops	56.9%

While Table 2 shows how wasteful the method of "record censoring" is in obtaining a clean data matrix, an even more important consideration is the biasing effect that this process has on the remaining data. Since any household with item non-response is omitted from the final data set (which may then be used for data analysis or modelling), it stands to reason that, if item non-response is distributed randomly through the data, those households with more people and those people making more trips are more likely to be omitted because they have a higher chance of item non-response (simply because they are providing more data from which something may be missing). Romios (1996) demonstrates this effect empirically using the VATS data, as shown in Table 3. The censored data set contains smaller households with fewer vehicles and making fewer trips than the complete data set.

Table 3 Number of Person, Vehicle and Stop Records per Household in Complete and Censored Data Sets

	Complete Data Set	Censored Data Set
	Number per Household	
Persons	2.68	2.45
Vehicles	1.54	1.44
Stops	10.02	8.76

It is also possible that non-English speaking people, people with disabilities, people with poor literacy skills, the elderly, and people without a telephone (and hence unable to be contacted for a follow-up interview to supply missing information) may also be under-represented if a "record censored" data set was adopted.

For the above reasons, the method of "record censoring" is strongly discouraged as a means of obtaining a "clean data set".

- ***Re-weighting of the Data***. As noted above in the first method of dealing with item non-response, ignoring the missing data on a case-by-case basis when performing the analysis will result in different totals being obtained in the marginals of any tables. This applies when performing calculations of the sample data, but also when performing analysis on data which has been weighted to allow for expansion to the total population (e.g. by comparison with Census data). This is because the expansion weights have usually been calculated by comparing a cross-tabulation of the sample data (e.g. persons by age and sex) with the population data to calculate the expansion weights. Any missing data in the sample cross-tabulation is treated in a specific manner, such as assigning the weight which corresponds to the average of the missing variable. These weights are then attached to the records in the data set. In later analyses, however, different records will be ignored depending upon which variables are being analysed, and hence the sum of the weights of those records included in the analysis will not always be equal.

One way around this problem is to recalculate the expansion weights for every specific analysis conducted. Thus, the analysis sample is first determined by removing records with missing values for the variables in question, and then the weights are calculated before the analysis is performed. In this way, the population estimates will always agree with the totals in the population data set. This method can become unwieldy, however, since every new analysis creates a new set of weights. Very soon, there are more weights in the data set than there is real data!

- ***Imputation of Missing Data***. The fourth method of dealing with item non-response is to impute (estimate) values for the missing data based on some other source of information. This method has the advantage that all data in the existing data set is used (i.e. no data is discarded), the imputation is done only once (compared to the multiple re-calculations of weights using the re-weighting method), and a clean data matrix is obtained for future analysis. For these reasons, imputation is the preferred method of dealing with item non-response.

4. METHODS OF IMPUTATION

As noted by Armoogum and Madre (1997), there are a number of different methods of imputation that can be used with household travel survey data, namely:

- ***Deductive Imputation***. This method allows a missing value to be replaced by a perfect prediction, based on a logical conclusion drawn from other data in the data set. This is often the case when redundant questions are asked in a survey, where missing responses to one question can be replaced by information derived from the other redundant questions.

- ***Overall Mean Imputation***. In this method, the missing value is replaced by the mean of that variable across all respondents in the sample. For example, a missing income would be replaced by the mean income of the respondents in the sample. This can be a dangerous method, unless the extent of item non-response is very small, because the method leads to reduced estimates of the variance (because all the imputed values are at the mean of the distribution) and hence invalid confidence intervals.

- ***Class Mean Imputation***. This method overcomes some of the problems of Overall Mean Imputation by first dividing the sample population into strata, based on other variables in the data set, and then calculating the mean of the variable to be imputed within each strata. The observation requiring imputation is then assigned to one of these strata, based on its values of the stratifying variables, and the mean of the variable within the stratum is assigned to the missing value. There will still be some reduction in variance using this method, but far less than would have occurred using Overall Mean Imputation.

- ***Hot-Deck Imputation***. In hot-deck imputation, missing responses are obtained by finding a record within the data set which is similar in all respects to the record with the missing value. The value of the variable (e.g. income) for this record is then substituted for the missing value. A variety of hot-decking procedures have been proposed including random overall hot-deck imputation (whereby a set of records with similar characteristics are formed, and the value to be imputed is obtained by random sampling from this set), random imputation within classes, sequential hot-deck imputation (where imputed values are obtained from the set of records by selecting each record in sequence) and hierarchical hot-deck imputation (where a set of records is developed with exact or non-exact matches to the target record, and then the better matches are used preferentially as the source of imputed data).

- ***Cold-Deck Imputation***. Whereas hot-deck imputation uses information from the data set of the current survey, cold-deck imputation uses data from sources other than the current survey. In most other respects, cold-deck imputation is very similar to hot-deck imputation.

- ***Regression Imputation***. In this 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. This method is useful when the use of Class Mean imputation stratification may result in a large number of empty cells within the stratification. Regression imputation allows these cells to be filled with information from neighbouring cells.

- ***Multiple Imputation***. In all the above methods, a single value of the imputed variable is obtained and substituted into the data matrix. With multiple imputation, a number of different values are imputed to create a number of "clean data matrices", which are then analysed as different representations of the complete data set.

The above techniques can be combined in various ways to create hybrid ways of imputing missing data. The main conclusion that appears to have emerged from the literature is that probabilistic imputation is better than deterministic imputation, since probabilistic imputation preserves the variance in the data to a much greater extent. Despite the relatively high item non-response for income and the importance of having an income measure for a variety of purposes, relatively little has appeared in the transport literature concerning the imputation of incomes (e.g. Bhat, 1994). The following sections seek to add to this literature.

5. IMPUTING MISSING INCOMES IN THE VATS DATA

The income imputation method developed for VATS combines several of the above methods to create a probabilistic, regression-based imputation method. Personal incomes are collected in the VATS survey using the same question as used by the Australian Bureau of Statistics in the 1991 Census of Population and Housing. This question has 15 income categories ranging from "less than $3001 per year" up to "more than $80,000 per year". Because the ABS income question (in 1991) does not have a specific "zero income" category, this category is created in the VATS data by assigning anyone in the lowest income category, and anyone not reporting an income, a value of zero if they state that they are unemployed and a student. Both conditions must be satisfied, since employed students can earn an income while unemployed non-students may be in receipt of welfare benefits (which are counted as income). After this step, there are still about 7% of respondents for which no value of personal income is available. It is for these respondents that income values need to be imputed.

The basis of the imputation method is to construct statistical relationships between personal income and a range of demographic variables provided by those respondents who did provide their income, and then use these relationships to impute the personal income of those respondents who have not provided this information. In this way, it is expected that maximum use will be made of the available information when performing subsequent analyses. It is also anticipated that such imputation methods will reduce any bias which might have arisen by excluding those who declined to disclose details about their income.

The demographic information used in the construction of the income models for employed persons was age, sex, work status (full-time or part-time) and occupation. The level of non-response on each of these variables was lower than for the income question (3.0%, 0.7%, 1.6% and 2.7% respectively). Hence it was expected that mean income would be able to be imputed from these questions which were more readily answered. The mean income was also estimated for those not in the paid workforce (such as those on pensions and welfare benefits). For these people, income models were constructed based on age, sex and activity status (e.g. keeping house, age pensioner etc). After the mean income had been estimated from the regression models, a probabilistic value was sampled from the distribution of income to obtain the final imputed value.

Regression Models of Personal Income

For each year of the VATS data, regression models were built which explained personal income in terms of the age, sex, work status and occupation of employed persons. Examples of these relationships are shown in Figures 1 and 2, for male and female full-time professionals in 1994.

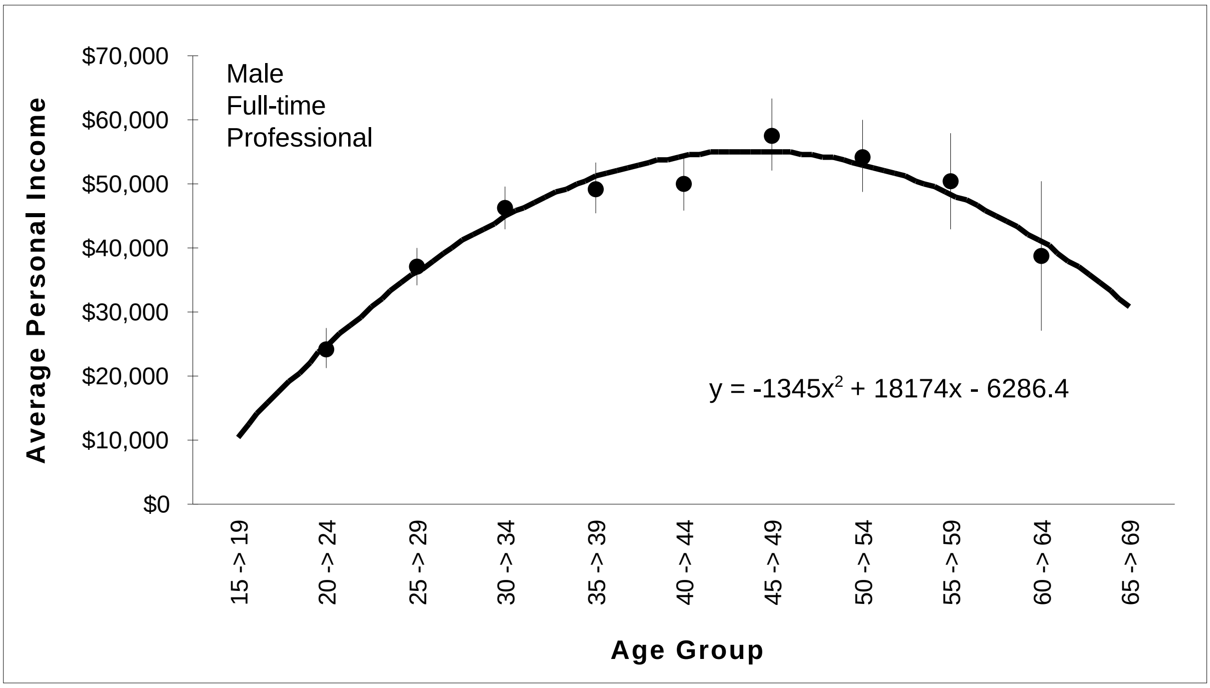

Figure 1 Average Income for Male, Full-time Professionals in 1994

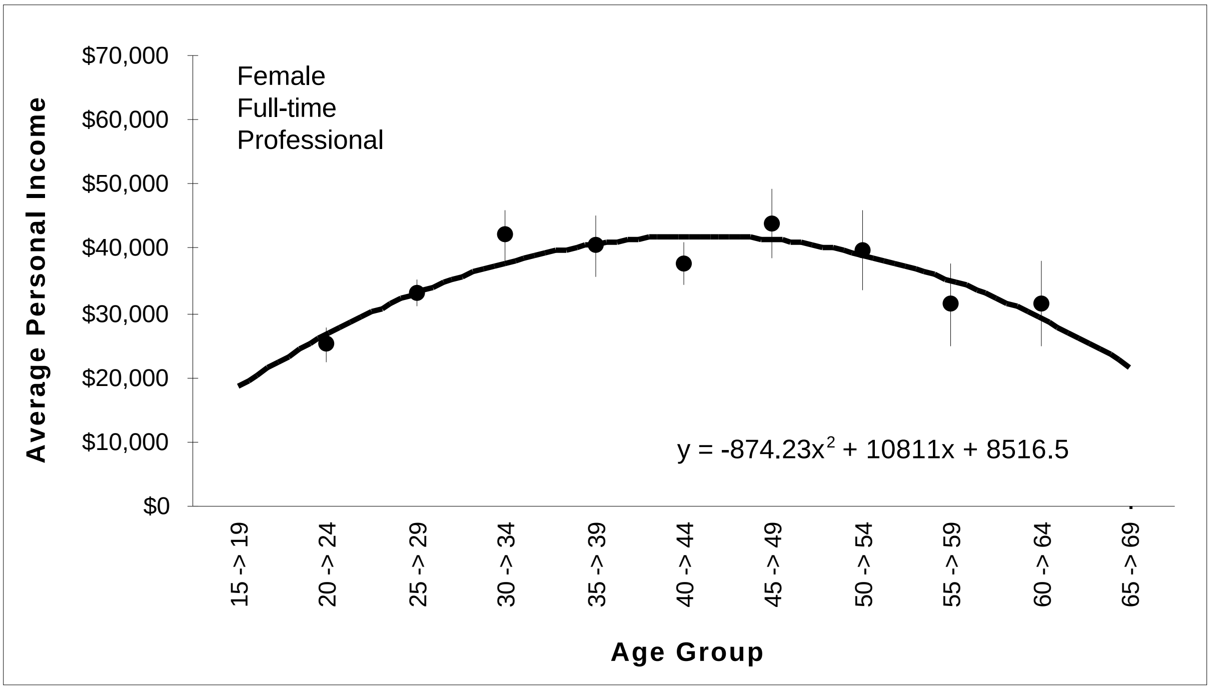

Figure 2 Average Income for Female, Full-time Professionals in 1994

For each of these employment groups a non-linear regression model was estimated to describe average personal income as a function of the age group (where those aged between 15 and 19 were in age group 1, i.e. x=1, and those aged between 65 and 69 were in age group 11). The estimated regression equations are shown in Figures 1 and 2. This process was repeated for each occupation (as given by ASCO codes), each sex and employment status for each year. Regression models were also estimated for the marginals data (e.g. both sexes, all employment statuses, all occupations) for use when the record for which an income value was to be imputed did not have complete data for all the independent variables. For example, if sex was not stated, then the average income value for both sexes was used.

An example of the regression models estimated for both years is shown in Figure 3, where the regression models for full-time professionals for males and females in 1994 and 1995 are shown graphically. It can be seen that all regressions have the same shape, with highest incomes for professionals in their forties. Both male and female incomes start at the same level, but by the mid to late twenties male full-time professionals are earning consistently more than females. The 1995 incomes are slightly higher than the 1994 incomes for both males and females.

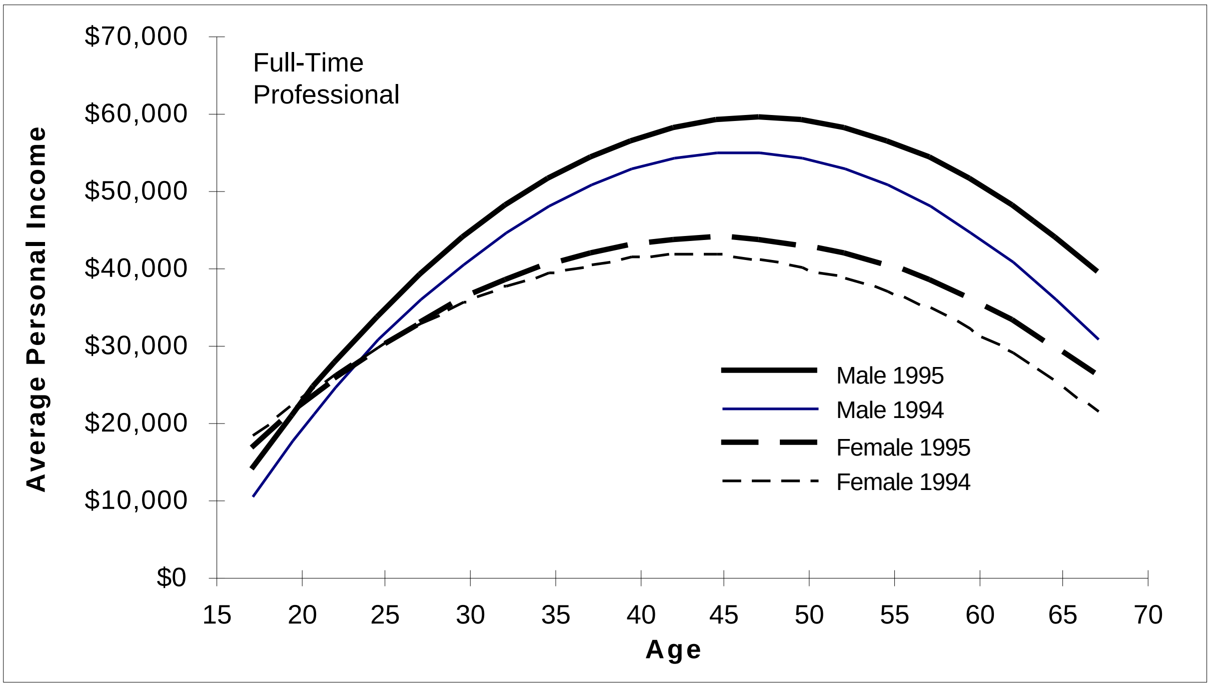

Figure 3 Income Regression Models for Full-time Professionals

The regression models for those not in paid employment were not as clear-cut as those shown above. Mainly because of limited sample sizes, it was not possible to show consistent relationships between age and income for any of the activity status categories for either year. The best that could be done was to calculate average incomes by activity status and sex for each year, as shown in Table 4. It can be seen that males consistently have higher incomes than females, and that generally the 1995 incomes are slightly higher than the 1994 incomes. Even though only a single income value is given for each cell in Table 4, this does not imply that the same value is used for imputation for all people belonging to each cell (which would be the case if Class Mean Imputation had been used). As will be shown below, explicit account is taken of the distribution of income around these mean values.

Table 4 Average Income by Activity Status and Sex

		Average Income	
Activity Status	Sex	1994	1995
Keeping House	Male	$30,000	$27,000
	Female	$15,600	$16,000
Currently Unemployed	Male	$8,000	$9,000
	Female	$6,500	$8,000
Retired	Male	$18,000	$21,000
	Female	$14,000	$16,000
Age Pensioner	Male	$8,500	$8,800
	Female	$8,000	$8,100
Other Pensioner	Male	$9,500	$9,300
	Female	$9,500	$9,300
Multiple Choice or Other	Male	$12,000	$11,100
	Female	$9,000	$10,500

The Distribution of Personal Incomes

As shown above, for both those in the workforce and those not in the workforce, an estimate of average personal income can be made (from other respondents in the VATS data set) based on age, sex, employment status, occupation and other activity status. It is also possible to examine the distributions of income within each of the categories defined by the above variables.

For those in employment, the distribution of personal incomes within each age/sex/occupation group was calculated. These distributions have then been standardised by dividing the incomes by the mean income for that group to obtain many distributions of relative incomes. The standardised income distributions for all age, sex and occupation groups have then been plotted on the same graph for 1994 and 1995 VATS data, as shown in Figure 4. It can be seen that, apart from a few outliers, the majority of the distributions lie along the same cumulative curve. This curve is sigmoid and skewed to the left with a long tail to the distribution.

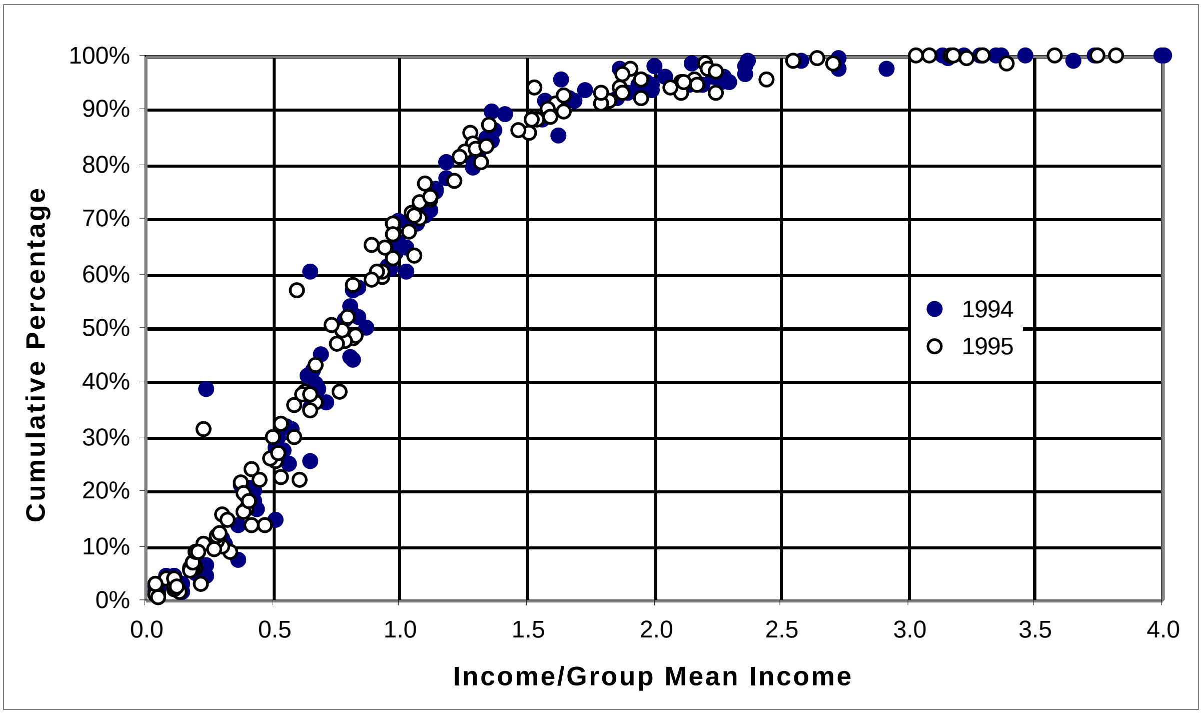

Figure 4 Distribution of Income Ratios for Employed Persons

A similar process was carried out for persons not in the paid workforce, and the cumulative distributions of relative income for these groups are shown in Figure 5. It can be seen that again the distributions tend to lie along the same cumulative distribution, although the agreement is not as high as with those in paid employment because of the smaller sample size for those not in paid employment. This curve is again sigmoid and skewed to the left, with a long tail to the distribution, but with a reduced variance compared to Figure 4.

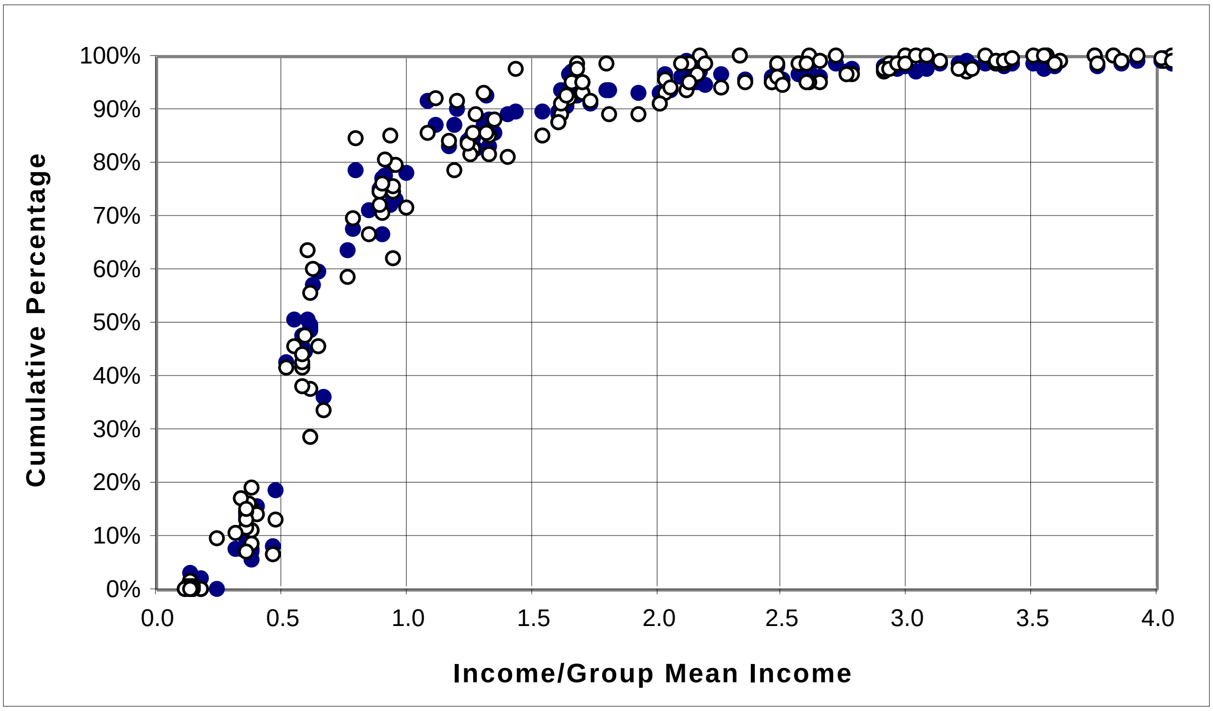

Figure 5 Distribution of Income Ratios for Non-employed Persons

To be able to use the information contained in Figures 4 and 5 in the stochastic imputation of missing incomes, it was necessary to find a theoretical distribution which captured the essentials of Figures 4 and 5. Given the sigmoid shapes, but the different variances, a gamma distribution was suggested of the form:

$$f(x) = \frac{1}{b^{a}\Gamma(a)} x^{a-1} e^{-x/b}, \qquad x>0$$

$$= 0 \qquad \text{elsewhere}$$

Different values of the parameters a and b were trialed until acceptable degrees of fit were obtained for both the employed and non-employed groups. The resulting distributions are shown in Figure 6, with the values of a and b for each group.

In applying the imputation procedure, the income regression models were used to first calculate the mean income for the respondents with a missing income (given their age, sex, work status and occupation). The gamma distributions were then used to sample a value of income ratio, which was then multiplied by the estimated mean income to derive a stochastic value of imputed income. This value was then substituted into the data record for that respondent, and treated thereafter as their real income.

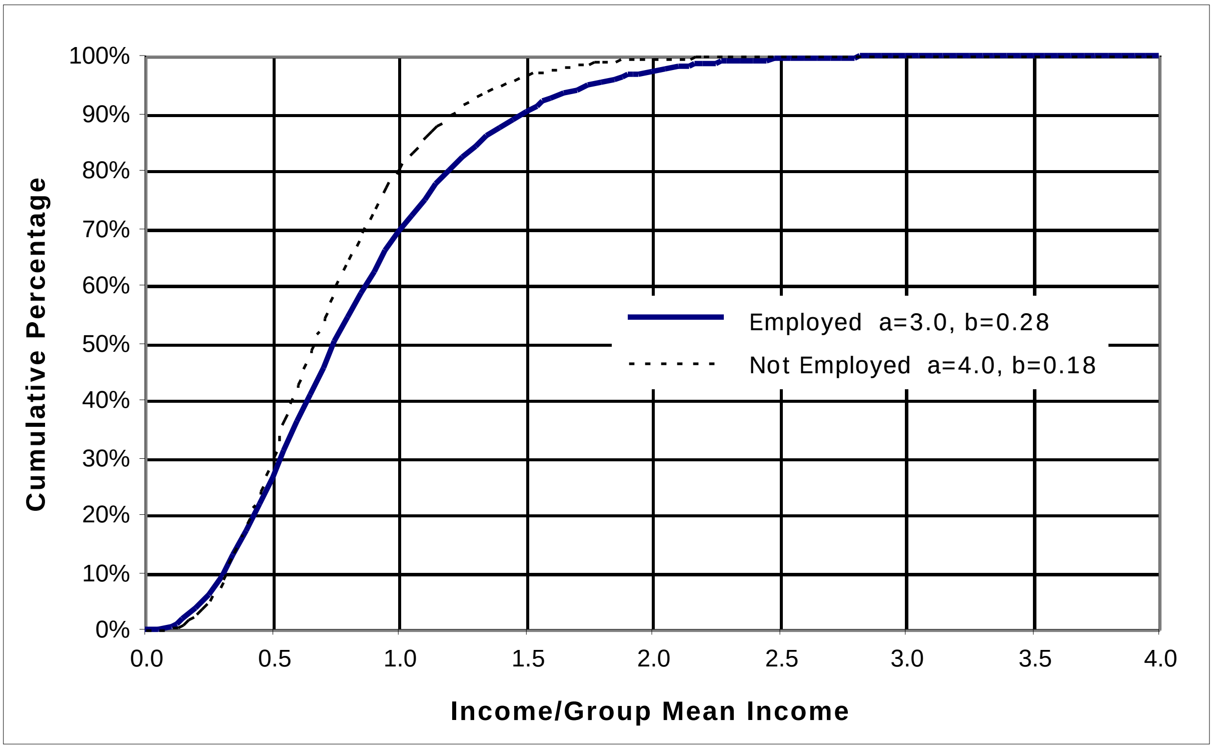

Figure 6 Gamma Distributions of Income Ratios

6. CHARACTERISTICS OF IMPUTED INCOMES

One of the advantages of stochastic regression-based imputation is that is tends to preserve the variance inherent in the original distribution of income, unlike the Overall Mean or Class Mean imputation methods, which tend to reduce the variance by selecting too many imputed values towards the middle of the range of incomes. This can be seen clearly by reference to Figure 7, which shows the distributions of imputed incomes in 1994 and 1995. It can be seen that imputed incomes occur across the full range of possible incomes, with imputed values in both the highest and lowest income categories. The spread of the imputed incomes is very similar to the spread of the reported incomes shown in Figure 8. The major difference is that there are relatively few imputed incomes of zero, whereas there are about 33% of reported incomes being zero. This is because most missing incomes which may have been zero would have been previously estimated using Deductive Imputation.

The outcome of the income imputation process is that every respondent in the sample ends up with a value of personal income. One question, however, is whether such imputation makes much difference. The first effect, of course, is that all the records can now be used when performing any analysis concerning income, rather than throwing away about 7% of the records (which originally had no income values). The second effect, as shown in Table 5, is that the imputed incomes are, on average, higher than the reported incomes. This is especially true when considering all incomes, but is also true when only considering non-zero incomes. In each year the (non-zero) imputed incomes are about $2500 - $3500 higher than the reported incomes. Bearing in mind the proportion of incomes which need to be imputed, the average income for the total data set (including imputed incomes) is about $250-$350 higher than the reported incomes.

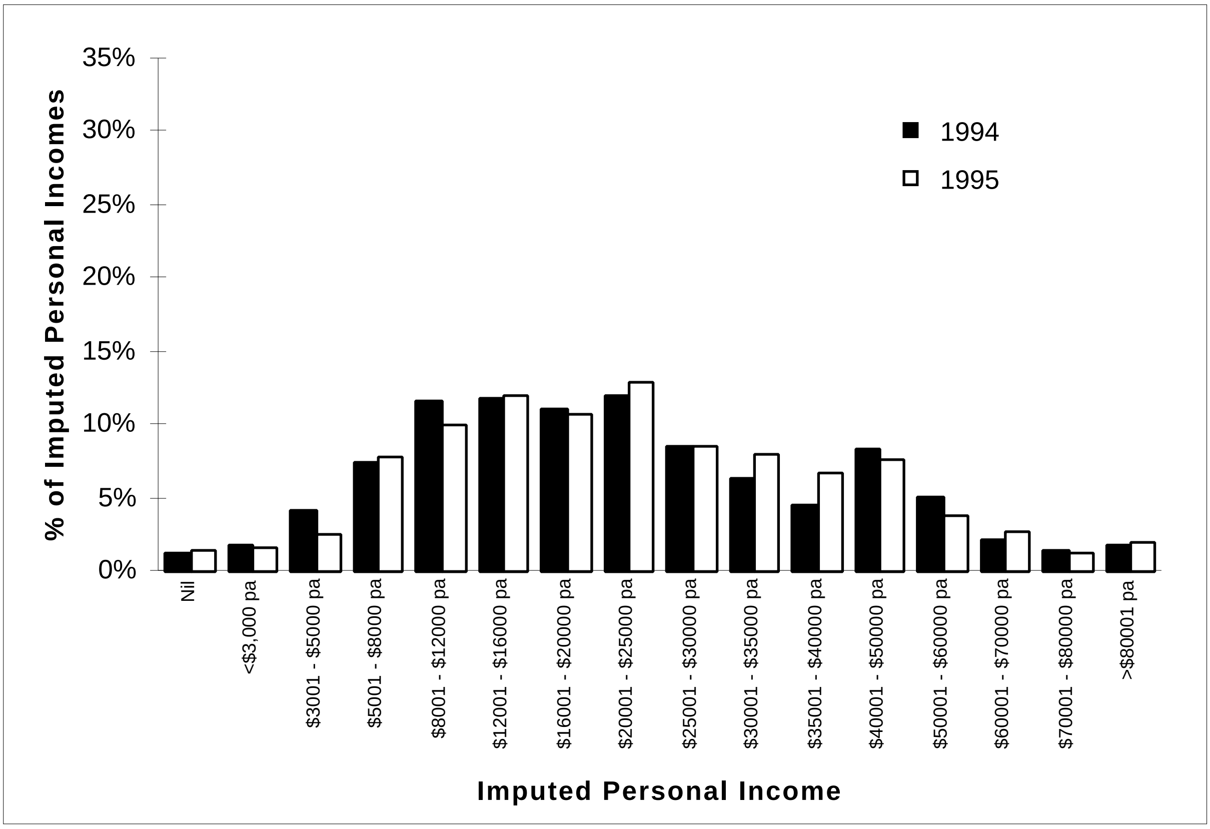

Figure 7 **Distribution of Imputed Personal Incomes**

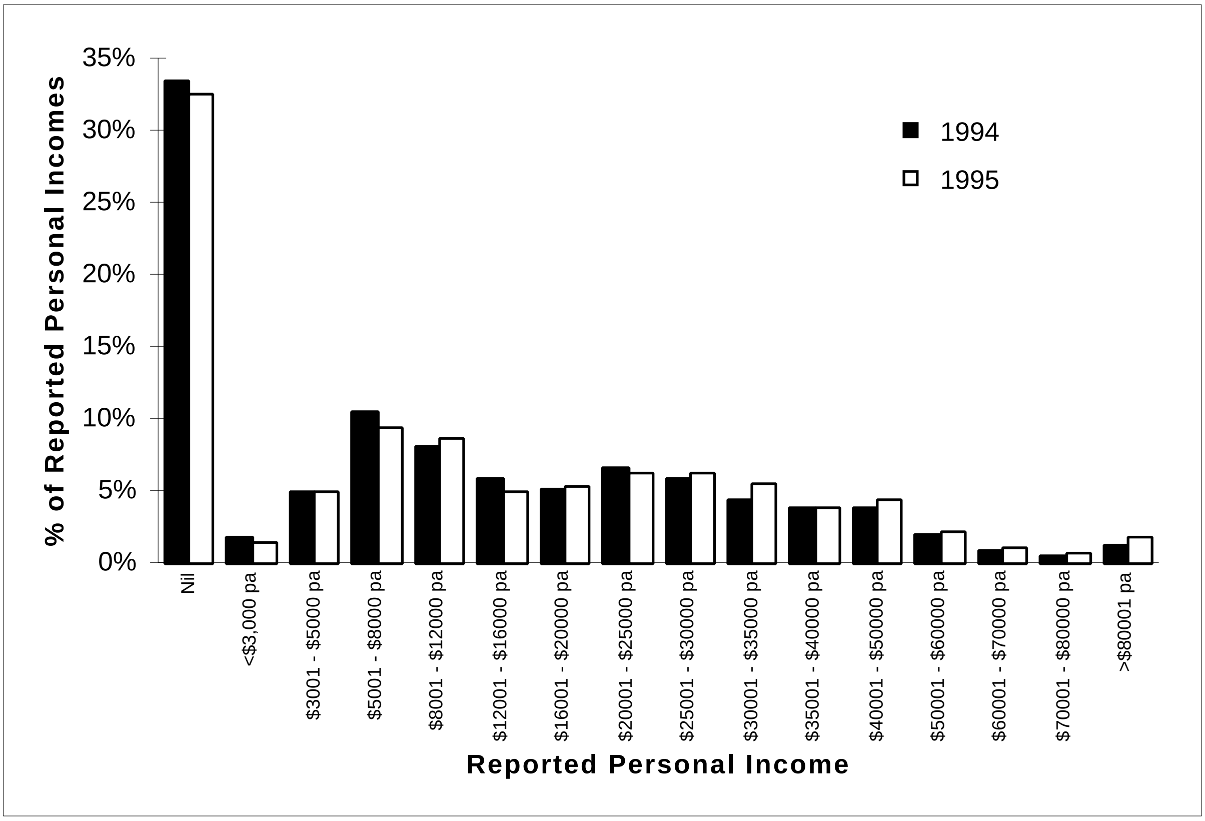

Figure 8 **Distribution of Reported Personal Incomes**

Thus imputation has not only made the data set more useable in its entirety, but it has also corrected a slight bias caused by having a higher income group not reporting their income in the first instance.

Table 5 Reported, Imputed and Total Income Distributions

	VATS94	VATS95	VATS94	VATS95	VATS94	VATS95
Personal Income	Reported	Reported	Imputed	Imputed	Total	Total
Nil	34%	33%	1%	1%	31%	30%
<$3,000 pa	2%	2%	2%	2%	2%	2%
$3001 - $5000 pa	5%	5%	4%	3%	5%	5%
$5001 - $8000 pa	11%	9%	7%	8%	10%	9%
$8001 - $12000 pa	8%	9%	12%	10%	8%	9%
$12001 - $16000 pa	6%	5%	12%	12%	6%	6%
$16001 - $20000 pa	5%	5%	11%	11%	6%	6%
$20001 - $25000 pa	7%	6%	12%	13%	7%	7%
$25001 - $30000 pa	6%	6%	9%	9%	6%	6%
$30001 - $35000 pa	4%	5%	6%	8%	5%	6%
$35001 - $40000 pa	4%	4%	5%	7%	4%	4%
$40001 - $50000 pa	4%	5%	8%	8%	4%	5%
$50001 - $60000 pa	2%	2%	5%	4%	2%	2%
$60001 - $70000 pa	1%	1%	2%	3%	1%	1%
$70001 - $80000 pa	1%	1%	1%	1%	1%	1%
>$80001 pa	1%	2%	2%	2%	1%	2%
TOTAL	100%	100%	100%	100%	100%	100%
Average->	$14,928	$16,249	$25,511	$26,103	$15,696	$16,953
Non-Zero Average ->	$22,468	$24,089	$25,855	$26,475	$22,821	$24,331

7. CONCLUSION

This paper has reviewed the causes of item non-response in household travel surveys, and has outlined methods of dealing with this item non-response by various methods of imputation. It has then described the development of a stochastic regression-based method of imputation and applied this method to estimate missing income values in the Victorian Activity & Travel Survey data for 1994 and 1995. This technique is now a standard part of the VATS editing procedures, in an attempt to provide the best data outputs from the survey. The use of stochastic imputation methods is seen to give reliable estimates of missing income which preserve the inherent variance in the data.

The use of imputation techniques is recommended as a standard procedure to obtain maximum value from the data, and to avoid subtle biases in the data. For example, where personal incomes are combined to obtain measures of household income (e.g. Loeis and Richardson, 1997), it is important to impute missing personal incomes before combining them into household incomes. Otherwise, any household in which one member has failed to provide a personal income would have to be treated as having a missing value for their household income. If the occurrence of missing personal income was assumed to be a random event, then larger households with more income earners would have more chance of having at least one missing personal income and hence a missing household income. This would result in the mean value of household income being lower than it actually was because more large households with high household income would have been ignored in the calculation. Therefore, even if missing personal incomes were distributed proportionately across the income distribution, the distribution of household incomes would be biased. However, we have seen above that missing personal incomes are likely to be higher than the reported personal incomes. Therefore, not imputing personal incomes would carry a double penalty in terms of the effect on household incomes.

This paper has sought to make a small contribution to raising the quality of data used in transport policy analysis. Only by the conduct, and reporting, of numerous research studies making these types of small contributions can we ensure that the quality of transport research, modelling and policy analysis is improved over time.

8. REFERENCES

Armoogum, J. and Madre, J-L. (1997). "Item Sampling, Weighting and Non-Response". *International Conference on Transport Survey Quality and Innovation*, Grainau, Germany.

Bhat, C. (1994). "Estimation of Travel Demand Models with Grouped and Missing Income Data", *International Association for Travel Behaviour Research Conferenc*e, Santiago, Chile, July.

Loeis, M., and Richardson, A.J. (1997). "A Welfare Index for Better Understanding of Travel Behaviour". *21st Australasian Transport Research Forum*, Adelaide.

Polak, J.W., Ampt, E.S. and Richardson, A.J. (1995). "An Analysis of Non-Response in Travel Diary Surveys", *23rd PTRC European Transport Forum*, The University of Warwick, England.

Radbone, I. (1994). "Taking Social Justice Seriously in the Provision of Public Transport". *19th Australasian Transport Research Forum*, Lorne, Victoria, pp.133-148.

Richardson, A.J. and Ampt, E.S. (1994). "Non-Response Effects in Mail-Back Travel Surveys", *International Association for Travel Behaviour Research Conference*, Santiago, Chile, July.

Richardson, A.J. and Ampt, E.S. (1995). "The Application of Total Design Principles in Mail-back Travel Surveys". *7th World Conference of Transport Research*, Sydney.

Richardson, A.J., Ampt, E.S. and Meyburg, A.H. (1996). "Non-Response Issues in Household Travel Surveys", In *Conference on Household Travel Surveys: New Concepts and Research Needs*, Transportation Research Board, Conference Proceedings 10, pp. 79-114.

Romios, G. (1996). "The Value of Missing Survey Data". *18th Conference of Australian Institutes of Transport Research (CAITR)*, Queensland University of Technology.

Sheatsley, P. (1983). "Questionnaire Construction and Item Writing". In P.H. Rossi et al. (Eds.) *Handbook of Survey Research*. New York: Academic Press.

Stopher, P.R. and Metcalf, H. (1996). *Methods for Household Travel Surveys*. Synthesis of Highway Practice 236, National Cooperative Highway Research Program. Washington, D.C.: National Academy Press.

Zmud, J.P. and Arce, C.H. (1997). "Item Nonresponse in Travel Surveys: Causes and Solutions". *International Conference on Transport Survey Quality and Innovation*, Grainau, Germany.