Flexible Matching Imputation in the Manufactured Homes Survey[1]

Bonnie E. Kegan and Todd R. Williams
U.S. Bureau of the Census, Washington, D.C. 20233

1. Introduction

The Census Bureau conducts the Manufactured Homes Survey (MHS) each month to track a sample of shipments of Department of Housing and Urban Development (HUD) inspected manufactured homes from manufacturer to dealer to buyer, until they are placed or otherwise taken off the market. A 1 in 40 sample of shipped units is drawn and a questionnaire mailed to the dealers who then respond by telephone. Sampled homes are classified as being in inventory, placed for residential or nonresidential use, or in an "other" status (returned to manufacturer, destroyed, out-of-scope, etc.). If no response is received on the home, survey interviewers try to contact the dealer the next month for up to three consecutive months. If no information has been obtained on the home after three months, the Census Bureau classifies it as a permanent non-respondent.

If the home is placed for residential use the MHS collects detailed information, including length, width, number of bedrooms, presence of central air-conditioning, price, location, and method for securing it in place. As in any survey there is a certain amount of unit and item non-response that results in missing data for the MHS. Currently the Census Bureau compensates for unit and item non-response with a random backward-forward hot-deck procedure that selects donors within imputation cells formed mainly by sample selection month, the state where the home was placed, and number of sections in the home. Matches are found by defining imputation cells by selection month, placement state or number of sections and by placing the records into the appropriate cells. A record with an observed value for the variable being imputed is then matched to a record with missing information within the imputation cell. Note that for matches to be possible the month and state must be present in both the observed or donor record and the missing data or recipient record. The missing data are replaced using the observed data from the donor record. Sales price for sold homes is imputed using a regression on the square footage and number of sections, not by the hot-deck method.

Flexible matching imputation (FMI) (Williams, 2001) combines hot-deck imputation with model-based methodology. This method identifies and gives a ranking to the matching variables that should be used for any given missing variable or combination of missing variables. The set of matching variables for a particular combination of missing variables are determined by fitting regression models to the data in which the variables have observed values. Here the observed variables are the dependent variables in the models and the possible matching variables are the independent variables. For missing continuous variables, a multivariate linear regression model is fitted, whereas for each missing categorical variable, a polytomous logit model is fitted. The rankings of the matching variables are determined by fitting the models using a forward stepwise procedure in which the first independent variable kept in the model is the most important, the second variable kept is the next important, etc. Once the set of matching variables is found, the variables are used to find donor records. If no match can be found using all the matching variables, the lowest ranked variable is dropped and a new attempt is made at finding a match with the remaining variables. Dropping of variables continues in order of increasing rank until a match is found. The FMI procedure is an automated procedure meaning that, once the user selects a group of variables that are available as possible matching variables, the procedure will find the best matching variables and perform the hot-deck imputation without any need for user intervention.

The purpose of our research is to show that by using the FMI procedure for imputing missing data found in the MHS we obtain, on average, more accurately imputed data than by using the current method. We base our research assumption on the fact that the FMI procedure is designed to find the set of matching variables that have the strongest influence on the missing variables. As a result, the FMI method should preserve the inter-variable relationships found in the data that might not be preserved by the current method.

In the following section of the paper we will describe the methods used to conduct our research. In the third section we will present our results and in the fourth section state our conclusions and recommendations.

[1] Disclaimer: This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

2. Methods

In this discussion records that require no imputation will be referred to as observed data, and those that require imputation of at least one variable will be referred to as imputed data. The data for this study is taken from the June 2001 MHS data file that includes fifty months of data beginning May 1997 and ending June 2001. There are a total of 54,315 records in this file each corresponding to one manufactured home. Price of the home is the most frequently missing variable (46%). Length, width and number of bedrooms as a group are the second most frequently missing variable (34%). In the current imputation method length, width and bedrooms are the last home characteristics to be imputed before price is imputed using a regression model. Length, width, bedrooms and price are only collected for homes placed for residential use. Additionally price is only collected for homes that are sold. Residential placements and sold homes make up 99% and 96% of the observed records, respectively. Thus the main focus of this study will only be on sold homes placed for residential use since these homes make up the majority of the data records.

In order to perform the comparison between the current imputation method and the FMI method, we simulate missing data for length, width, number of bedrooms, and price of home by blanking out some of the observed values from the observed data. In this way, we can compare the results from the imputation methods to the observed or true values. Because we found that the missing variables are missing at different rates depending on the number of sections of the home and the dealer region, we created post-strata defined by the four possible sections and the four dealer regions. Within each post-stratum for all the data, we calculate the proportion of data containing missing information. After making the calculations, we separate the observed data into the correct post-strata and multiply the counts by the corresponding proportion to obtain the number of records that will have their values set to missing. We then randomly select observed data records within each post-stratum and blank out their variable values until the calculated total number of simulated missing data records has been reached. The total number of records that we are using is 17,949.

Our final stage of setting up our research is to create multiple imputations by creating twenty different simulated missing data files. We picked the number twenty arbitrarily, but it is more than enough for computing aggregate estimates (Schafer, 1997). We create the twenty different data files by randomly resorting the order of the remaining observed data records for each data file. The observed data records are possible donors for the current hot-deck and the FMI methods and resorting their order gives the opportunity for different donors to be used in each of the twenty files. By doing this, we can calculate averages of aggregate estimates and future variances from the imputed data.

3. Results

We show the matching variables that are identified by the FMI program in Tables 1 and 2. Table 1 shows the matching variables for length, width and number of bedrooms (LWBR) that are used when price is present and when it is absent. Note that when number of bedrooms is imputed it uses a different set of matching variables than length and width because it is a categorical variable and categorical variables are imputed after all of the continuous variables. This allows for price to be available for matching at all times when imputing values for the number of bedrooms. Table 2 shows the matching variables for price that are used when LWBR are present and when they are missing. The current method finds donor records for imputing LWBR within donor cells that are formed based on a state or group of states in which the home is placed and whether the home has one section or two or more sections (Table 3). Price is imputed using a linear regression based on square footage and number of sections (1, 2, or 3+). By comparing Tables 1 and 2 to Table 3 it can be seen that FMI uses matching variables that are not considered in the current method to impute price, length, width and bedrooms. Note that in Table 1 price is used as a matching variable for LWBR when it is present. This is not possible in the current method because price is imputed after LWBR. Type of placement site is used with or without price. Type of foundation, secured method, and presence of air-conditioning are also used to find donors to impute price or LWBR in the FMI procedure. Although these variables are all imputed prior to LWBR and price, none of them are used in forming the donor cells.

In order to ascertain how valuable these matching variables are in finding good donors, we fit regression models to the set of observed records that are available as possible donors in our simulation. Here the set of variables we are going to impute are the dependent variables and the matching variables are the independent variables. By doing this, we can see how well the matching variables can predict values for the missing variables. If they are strong predictors, then they should be able to match with a donor who can

donate an imputed value that maintains the correlations found between the missing variables and the matching variables. We are not concerned with matching variables that are weak predictors because either there is little relationship to maintain or they are highly correlated with another matching variable and should have their relationships maintained indirectly.

Table 1. Matching Variables for Length, Width and Bedrooms (FMI)

Rank	Price present (275 records)	
	Length and Width	**Bedrooms**
1	Number of sections	Air-conditioning
2	Price	Price
3	Dealer Region	
4	Type of site	
5	Type of setup	
Rank	**Price absent (6,242 records)**	
	Length and Width	**Bedrooms**
1	Number of sections	Air-conditioning
2	Dealer Region	Price
3	Type of site	
4	How home is secured	
5	Air-conditioning	

Table 2. Matching Variables for Price (FMI)

Rank	Length, width, and bedrooms present (2,721 records)	Length, width and bedrooms missing (6,242 records)
1	Width	Number of sections
2	Length	Dealer Region
3	Dealer Region	Type of site
4	Number of sections	How home is secured
5	Type of site	Air-conditioning

Table 3. Key Variables in Current Method

	Imputation of Length, Width and Bedrooms	Regression to Impute Price
Key Variables	Placement state	Square feet
	Number of sections	Number of sections

For missing length and width when price is also missing, we fit a multivariate linear regression model using length and width as the dependent variables and the matching variables as the independent variables. For missing price, we fit a multiple regression model with price as the dependent variable and the matching variables as the independent variables. We do this to see how important the matching variables are in predicting values for the missing variables. By applying t-tests for each parameter estimate associated with the levels of the independent or matching variables, we test the hypothesis that there is no relationship between the matching variables and the dependent or missing variables. We also look at the resulting R-Square value to get a feel for how well the entire model is fitting the data. The closer the R-Squared value is to 1.0, the more the variability in the data is explained by the model and the better the fit.

Referring to the variables for missing length and width in Table 3, the number of sections is an important predictor at a 0.0001 significance level for both length and width. However while placement state is an important predictor for missing width at the 0.01 significance level it is not an important predictor for length. In terms of model fitting the R-Squared value for length is 0.20 and for width is 0.74.

Referring to the variables for missing length and width when price is absent in Table 1, one level each for two of the variables is not a significant predictor, but overall the variables are important at least at a 0.05 significance level. The non-significant levels are the first dealer region and not secured home. The R-Squared value for length is 0.22 and for width is 0.84. The increase in R-Squared values is most likely due to the extra information added by the FMI procedure.

We next perform the same comparison using the fitted models for when price is missing. Referring to the variables in Table 3 for imputing price, all of the key variables are significant predictors of price at a 0.0001 significance level. The resulting R-Squared value is 0.46.

Referring to the matching variables listed in Table 2, all of the variables are important predictors of price at least at a 0.005 significance level. The R-Squared values are 0.52 and 0.39 for the models with and without LWBR respectively. We feel that the higher R-Squared value of 0.52 compared to 0.46 is due to extra information added by the FMI procedure. We also feel that by not having length and width in the model, we obtain a lower R-Squared value of 0.39 compared to 0.46.

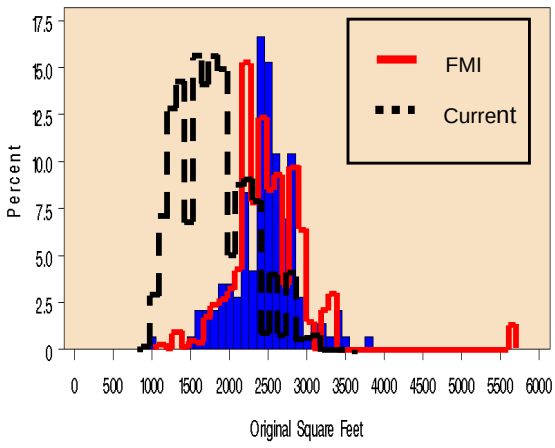
Based on the above analysis, we decided to look at imputed length and width in terms of

square footage by the number of sections since this is the most important matching variable in the FMI procedure. Table 4 gives the medians, means and standard deviations of square footage for the observed or true values and the values imputed by each of the methods. As expected, the two methods produce imputed values that are statistically close to the observed values for homes having only one section. The same is true with homes containing two sections. However for homes having three or four sections (about 2.5% of all homes), the FMI procedure performed considerably better. In Figure 1, we show the distributions based on the observed values shown in the bars and the imputed values for homes with three sections (about 2% of all homes). Here we can definitely see the lower bias associated with the imputed values from the current method.

Table 4: Estimates of Square Footage Grouped by Number of Sections

No. of Sections	Estimates	Observed Values	Method of Imputation	
			Flexible Matching	Current Method
1	Median	1,216	1,216	1,216
	Mean	1,137	1,125	1,129
	SD	187	189	189
2	Median	1,624	1,680	1,680
	Mean	1,665	1,672	1,698
	SD	368	375	392
3	Median	2,460	2,400	1,664
	Mean	2,459	2,478	1,696
	SD	414	520	428
4	Median	3,647	3,120	1,642
	Mean	3,446	2,847	1,694
	SD	1,113	284	431

Figure 1. Distribution of Square Footage for Three Section Homes



One item of interest to us is the relationship between the number of bedrooms and the price of the home since both are imputed and the number of bedrooms uses price as a matching variable in the FMI procedure. In Table 5, we see that imputed values for price from the FMI method come closer to the observed estimates than that of the current method. For imputed prices using the current method, there appears to be a lower bias when the number of bedrooms is small and an upper bias when the number of bedrooms is large. We also see that the standard errors are considerably smaller for the values imputed by the current method.

Table 5. Estimates of Price Grouped by Number of Bedrooms

# Bedrooms	Estimates	Observed Values	Method of Imputation	
			Flexible Matching	Current Method
1	Median	$30,406	$26,375	$19,085
	Mean	$31,888	$32,142	$22,729
	SD	$18,966	$20,912	$10,582
2	Median	$30,900	$29,400	$27,969
	Mean	$37,450	$36,050	$33,339
	SD	$22,184	$19,982	$11,195
3	Median	$45,900	$46,500	$49,688
	Mean	$47,915	$48,420	$48,685
	SD	$17,833	$18,142	$11,039
4	Median	$58,000	$58,300	$62,505
	Mean	$59,229	$60,097	$60,908
	SD	$15,717	$17,894	$8,541
5	Median	$55,830	$56,000	$62,538
	Mean	$58,598	$59,208	$63,409
	SD	$15,955	$16,906	$7,487

Next we compare the imputed values of price and length and width (as square footage) in Table 6. We see that when comparing the median, means and standard deviations both the FMI and the current methods come close to the observed or true values. However for imputed price, the median of the imputed values produced by the FMI procedure is closer to the observed values and once again the standard deviation of the current method's imputed values is a lot lower. Something else that is very striking in Table 6 is that the correlation coefficient between price and square footage produced by the current method is very high.

Table 6: Estimates of Price and Square Footage

	Estimates	Observed Values	Method of Imputation	
			Flexible Matching	Current Method
Price	Median Mean SD	$47,094 $48,800 $18,663	$47,500 $49,091 $19,097	$51,190 $49,358 $12,587
Square Feet	Median Mean SD	1,512 1,568 427	1,512 1,570 431	1,512 1,572 425
Price, Square Feet	Corr. Coeff.	0.6150	0.6237	0.9050

Figure 2

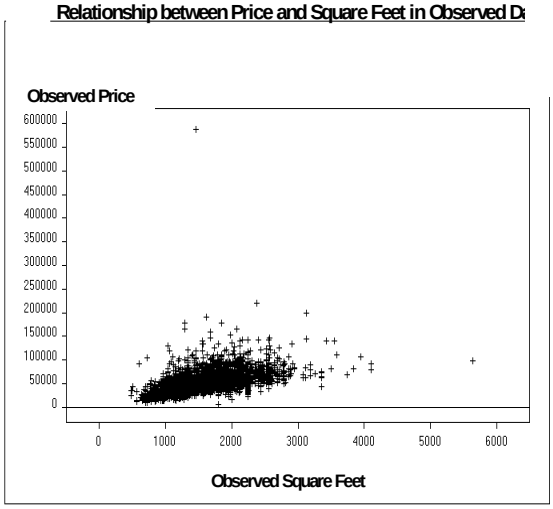


Figure 3

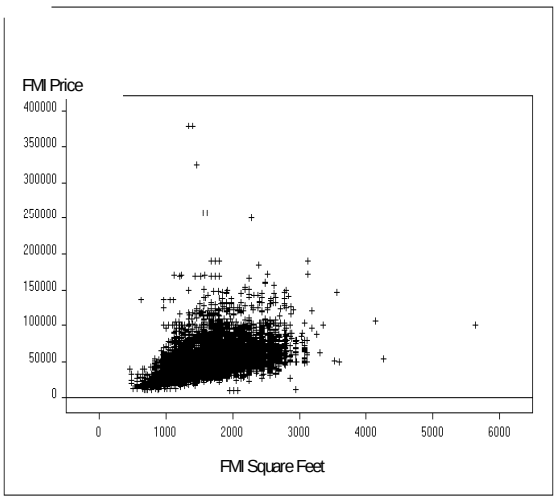
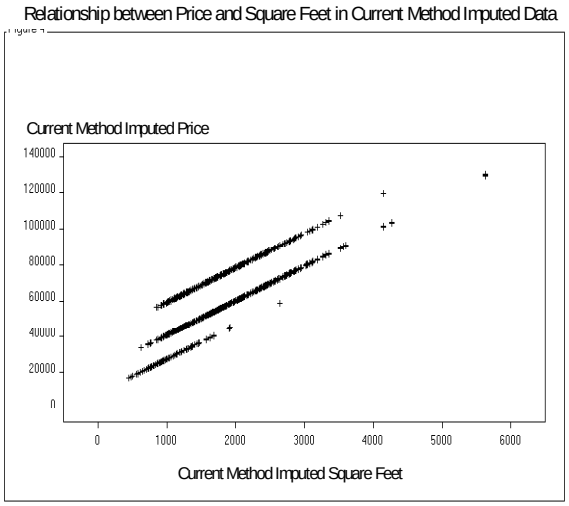


Figure 4



We provide a clearer understanding of what is happening between square footage and price in Figures 2, 3 and 4. We see that the imputed values produced by the FMI procedure are close to those of the observed values. Those produced by the current method are not close.

4. Conclusion

Through simulation of missing data in observed records from the MHS we are able to compare the performance of the current imputation method to the flexible matching imputation method. Our purpose in conducting this research is to show that the FMI method is able to automatically find matching variables for hot-deck imputation that strongly influence the imputed values. As a result, the inter-variable relationships that exist in the observed data between these variables are maintained and more accurate imputes on average are found than with the current method.

The current method uses primarily placement states and numbers of sections when forming its donor cells for imputing length, width and bedrooms (LWBR). When imputing price the current method uses a multiple regression model based on number of sections and square feet. Similar to the current method the FMI specifies number of sections and a location variable (dealer region) as important matching variables for imputing price or LWBR. However, the FMI method also specifies matching variables not considered by the current method such as type of site, secured method, type of setup, and air-conditioning. The FMI method also allows length and width to be matching variables for imputing price and price to be a matching variable for imputing length and width when length and width

or price is available, respectively. Another major difference between the two methods is the use of price as a matching variable to impute bedrooms in the FMI. The current method imputes bedrooms simultaneously with length and width, before price is imputed using regression.

We looked at imputed length and width as square feet by the number of sections of the home (Table 4). By maintaining the extra number of sections, we came closer to the observed or true estimates of square feet by sections with the FMI procedure than we did with the current method. This gives an example of the importance of using valuable extra information when matching to donors.

Length and width are used in both imputation methods to determine price, although, only when observed in the FMI method. So we compared the average square feet and average prices for the imputed values from both methods with the observed values (Table 6) and found that the FMI values were closer to the observed values than the current method's imputed values. Standard deviations for FMI values were also closer to the observed value standard deviations. The current method standard deviations for price were much smaller than the observed standard deviations. In the majority of records, length, width, number of bedrooms and price are likely to be missing together. In the current method length and width must be imputed first before they can be used in the regression to determine price. However in the FMI method, a specific set of matching variables is determined to impute price when all four variables are missing. As seen in the results, the approach of the FMI provides imputations for length, width, and price that have a distribution closer to that of the actual values. In all comparisons that were made, the FMI method yielded more accurate imputed values on average and preserved inter-variable relationships better than the current method by identifying and utilizing the most important relationships found in the data.

We also looked at imputed values for the number of bedrooms. In many cases we analyzed, there was not a significant improvement in the FMI imputed values compared to the current method imputed values. Due to lack of time, we did not pursue this as extensively as we would have liked. We plan on conducting further evaluations on the imputed number of bedrooms.

For the purposes of this research we only considered length, width, bedrooms and price to have possible missing values. In the MHS it is very possible to be missing all of the home characteristics on a record, including whether or not it is a residential placement or if it is purchased. Further research on the performance of the FMI method in imputing other possible missing variables is needed.

As we see with the direct modeling approach for imputing price by the current method (Figure 4), imputing a value directly from a fitted regression line removes the chance of displaying the variability found in the observed data. A direct modeling approach can be more successful if variability is added back to the imputed value taken from the fitted regression line. One way for us to accomplish this is to randomly select from the distribution of residuals found when fitting the model and add this residual value to the imputed value (Williams, 1998). We are planning further research to include imputed values using this approach and to compare the results to the FMI method.

We also plan to estimate variances due imputation. The twenty simulated missing data files allows for variances to be calculated for the current hot-deck and FMI methods. We can also calculate estimated variances directly from the procedures that find imputed values directly from fitted models.

Our research showed that the FMI method provides closer imputes on average and preserves the more important inter-variable relationships better than the current method when imputing price, length, width, and bedrooms for the MHS. While it appears that implementing the FMI method in the MHS would provide more accurate imputed data and maintain important variable relationships, more research must be done on the overall survey effect of changing the current imputation methodology.

References:

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data, First Edition*, Series: Monographs on Statistics and Applied Probability #72, London: Chapman & Hall.

Williams, Todd R. (1998). "Imputing Person Age for the 2000 Census Short Form: A Model-Based Approach," *in Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 680-685.

Williams, Todd R. (2001). "Flexible Matching Imputation: Combining Hot-Deck Imputation with Model-Based Methodology," *2001 Proceedings of the American Statistical Association, Section on Survey Research Methods [CD-ROM]*.