COMPARISON OF TWO IMPUTATION METHODS IN THE SURVEY OF DOCTORATE RECIPIENTS[1]

Jeffrey Stratton, John Finamore, Todd Williams
US Census Bureau, Washington DC, 20233-8700

Key Words: Hot Deck Imputation, Flexible Matching Imputation

1. INTRODUCTION

Nonresponse is a major concern in most sample surveys. Missing data are often imputed to provide users with a complete dataset. Many different methods of imputation have been proposed, and the methods used vary widely among different surveys. One of the most common techniques used to impute missing data is Hot Deck imputation. Even having chosen this technique, there are many different ways to implement it. This paper examines the results of two different Hot Deck imputation techniques applied to the 2001 Survey of Doctorate Recipients (SDR).

The SDR is one component of the National Science Foundation's Scientists and Engineers Statistical Data System (SESTAT), covering graduates in science and engineering. The other two components are the National Survey of College Graduates (NSCG) and the National Survey of Recent College Graduates (NSRCG). The SESTAT system is used to develop estimates of the total population of scientists and engineers in the United States, and to provide demographic and employment data on people who have been trained in or are working in natural sciences, social sciences, or engineering.

The SDR is a biennial survey and was last conducted in 2001 by the U.S. Census Bureau. The target population is those people under 76 years old, who are either citizens or non-citizens who planned to remain in the United States after receiving their degree, and who earned a doctoral degree from a U.S. college or university in a science and engineering (S&E) field. Doctoral level professional degrees, such as those in medicine, law, and education, are not included. The 2001 SDR included doctorate degrees earned between January 1, 1942, and June 30, 2000.

The sample design is a stratified sample design with probability proportional to size (PPS) sampling within each stratum. The overall sampling rate was about 5.8 percent, but sampling rates varied considerably across the strata. The total sample size was 40,000 people. The overall weighted response rate was 82.6 percent. For more information on the survey background and sampling and weighting procedures, please see references [1], [2], and [3].

Imputation is done for several reasons. One reason is to provide a complete dataset for users to analyze. Many users are unfamiliar with techniques used in analyzing missing data. Also, many multivariate software packages will simply ignore an entire observation with one missing response. This will eliminate a lot of useful data. Another reason we impute is to reduce nonresponse bias. The SDR's large sample size, small amount of missing data, and many questionnaire items make it a good candidate for imputation.

This paper compares the current SDR imputation methodology with a method called Flexible Matching Imputation (FMI). Both are Hot Deck methods that use modeling to determine appropriate sort and class variables. We wish to find if the less tedious and more automated FMI method yields comparable results, as this would greatly ease the post-collection processing burden.

The following two sections of this paper will describe the current 2001 SDR imputation procedure and the FMI procedure. The fourth section will describe how the research data file was constructed for testing the procedures. Sections five and six present our analysis and the results of our comparison. The final section will give our conclusions and describe further research.

2. CURRENT SDR IMPUTATION

The Survey of Doctorate Recipients uses a mixture of logical and statistical imputation. Some logical imputation is accomplished during the editing phase of data processing. The SDR also uses statistical imputation in the form of a Hot Deck class-donor method. In general, our Hot Deck procedure splits individuals into "similar" cells using class variables. Within a particular class, all individuals are sorted using sort variables. The 2001 SDR imputation method used a serpentine sorting procedure. An observation with a

[1] 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. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau or the National Science Foundation.

missing data value is assigned the same response as the observation just above it in the ordered sort. No donor may be used more than four times. This paper will focus on the Hot Deck imputation phase. The logical imputations were already performed during editing.

We perform a separate Hot Deck imputation for each questionnaire item. We begin by identifying important variables for a given item using log-linear modeling, namely the PROC CATMOD and PROC GLM procedures in the SAS statistical software. We use CATMOD to analyze categorical response variables, and GLM to analyze the near-continuous variables such as salary. Some categorical variables have many levels (such as type of employment) and PROC CATMOD has great difficulty with so many levels. In these cases, we use GLM to analyze the variable. To begin, we select a group of variables that we feel should be good predictors of the response. Since the SDR is longitudinal, we also have the previous year's responses for cases carried over from 1999. We then use PROC CATMOD or GLM to model the responses. We choose variables that are significant predictors in the model for use in our imputation.

Identifying Class Variables – Class variables subset the respondents into different groups. The sorting and matching step of the Hot Deck is performed within each class. The 2001 SDR questionnaire has certain "skip variables" directing which people answer certain questions. These "filter" variables must remain as class variables. Other important variables that are purely categorical (such as gender or race) could also be used as class variables. However, using too many class variables can result in having classes without enough donors to do the imputation. Thus, only the filter variables are kept as class variables.

Identifying Sort Variables – Any other variables found to be important in our model are put into the sort list, and are usually listed in order of importance. We sort the respondents in a certain class by the first variable in the sort list. Within that, they are sorted by the second variable, and so on. We generally determine the sort order by looking at the p-value from the SAS model output. We rank the variables by their chi-squared value statistics if they have similar p-values. However, this is not always the case. Some variables might have numerous levels. Sorting by these variables would effectively eliminate the effect of any variables placed behind them in the sort order. So, we usually move any large multi-level variables to the end of the sort order. Sometimes, the sort order must be manipulated to ensure consistency or some other editing feature. For example, age was sometimes moved to the front of the sort order in order to ensure that people were getting retirement or other information from a respondent of the same age, and we would not impute an age younger than their retirement age.

We divide the sampled cases into the classes specified by the class variables, and then sort by the sorting variables within each class. We use a serpentine sort for this step. For a case missing a response, we insert the response prior to it. No donor may be used more than four times. For more detailed information on the imputation methodology of the 2001 SDR, please see reference [4].

The current methodology has some limitations. Due to the number of questionnaire items needing imputation (98), a large amount of time is spent performing the repeated model fitting. In addition, there are subjective judgments that must be made when determining the order of the sorting variables. We would like to compare this method to the more automated and less subjective Flexible Matching Imputation procedure.

3. FLEXIBLE MATCHING IMPUTATION

We designed the Flexible Matching Imputation (FMI) procedure to perform two phases for missing data item imputation. The first phase uses the SAS statistical software to fit regression models to the observed data (data in which none of the data items are missing). We fit these regression models in order to find the best possible predictor variables for the missing item value in question. For phase two, we take the best predictor variables and use them as matching variables in a hot-deck imputation procedure. In the procedure, the data record with the missing item value will be matched to a record that can donate the value by matching on the matching variable values that are common on both data records.

In the first phase, we fit a separate regression model for each possible missing data item that will be imputed. We fit the models using forward selection methodology in which each possible predictor (independent) variable is fit individually for determining the missing (dependent) value. A test statistic is calculated for each independent variable where the variable that provides the lowest value is favored. A second term is added in the equation that penalizes independent variables that have a higher number of possible values or response levels. The independent variable that provides the lowest value for the test statistic is then kept in the model as a fixed predictor variable. We continue the forward selection procedure by fitting regression models in which the first independent variable is the fixed predictor variable and the second is one of the

remaining possible predictor variables being tested. Again, we calculate the test statistics associated with each new independent variable tried in the models. The tested independent variable in conjunction with the fixed first independent variable that provides the lowest test statistic value is kept in the model as the second fixed predictor variable. We repeat this cycle in the forward selection procedure to find a third, fourth, and fifth predictor variable. For finding the third, fourth, and fifth predictor variables, we also introduce two variable interactions between the independent variables.

For finding predictor variables for missing continuous data items, we use the SAS GLM procedure which uses the method of least squares to fit general linear models. In our case, we are fitting multiple linear regression models. Once the model is fitted, we use the number of data records *n*, the number of parameters *p*, and the sum of squared errors (SSE) to calculate Akaike's information criterion (AIC) [5] which is our test statistic. This calculation for multiple regression models is the following.

$$\mathrm{AIC} = n\,\mathbf{ln}\left(\frac{\mathrm{SSE}}{n}\right) + 2p$$

For finding predictor variables for missing categorical data items, we use the SAS CATMOD procedure which uses the maximum likelihood estimation of parameters for log-linear and logistic regression models. Here we are using the procedure to fit logistic regression models. After the model is fitted, we use the –2 log likelihood value from the model along with the total number of response levels minus one *k* and the number of predictors *s* to obtain our test statistic, Akaike's (1987) information criterion (AIC) [5] as follows.

$$\mathrm{AIC} = -2\,\mathbf{Log\ L} + 2(k+s)$$

After the model fitting, we perform the second phase. For each missing data item, the first predictor variable placed in the model is considered the most important matching variable. The second variable placed in the model is the second most important. The same reasoning follows for the third, fourth and fifth predictor variables with the fifth predictor variable added to the model being the least important matching variable. We find a donor for the missing item data record by matching the variable values on the missing item record with those on the next matching donor record. In the case of the matching variable being continuous, the variable is recoded into deciles in order to increase the chance of a match happening. If a donor cannot be found using all five matching variables, we drop the least important variable and try to find a donor by matching on the remaining four variables. This process of continuing to drop the next least important matching variable continues until a match is found. After the donor has been found, the missing data item is given the donor record's variable value.

4. THE RESEARCH DATASET

This section describes how we created our research dataset. We wish to evaluate how these two imputation methods perform by investigating how well they impute individual responses and how well they maintain the distribution of responses. To that end, we created a study dataset and simulated missing data for three variables: salary, faculty rank, and citizenship. These variables were selected because estimates of interest from SDR data are derived from them. These three study variables are also good examples of the types of variables we encounter on the SDR questionnaire. Salary is near-continuous. Citizenship is a categorical variable with several levels (five), and faculty rank is a categorical variable with many levels (nine).

The original 2001 SDR dataset had 31,366 completed interviews. Since item nonresponse is possible, we restricted this dataset to those cases not missing any of the three study variables. This was further restricted to cases not missing the filter variables for the faculty rank questionnaire item. This left us with 29,503 cases. We will refer to these filled citizenship, faculty rank, and salary values as the observed values. Then, we blanked out the three study variables at rates similar to the missing rates from the complete dataset. In addition to the three study variables, we kept 26 other variables on our datafile. The complete SDR dataset has many more than that, but we limited this to be less cumbersome for the FMI procedure. We chose these 26 additional variables because they had proven to be significant in imputing previous SDR surveys.

To determine the imputation rates in our survey, we stratified the complete dataset of 31,366 cases by a demographic group variable (eight levels), gender, and a collapsed field of degree variable (five levels). A similar stratification was used in the sample design of the SDR. Also, independent analysis showed the three item response rates varied across these variables. This resulted in 80 strata. For our three study variables, there are seven patterns of missing data:

1. Missing only salary
2. Missing only faculty rank
3. Missing only citizenship
4. Missing salary and citizenship
5. Missing salary and faculty rank
6. Missing faculty rank and citizenship
7. Missing salary, faculty rank, and citizenship

We tallied the number of each pattern of missing data in each stratum for the complete dataset.

We then stratified our study dataset of 29,503 cases into the same 80 strata, and determined the number of cases in each stratum. We then scaled the number of each pattern of missing data from the complete dataset by the ratio of our study dataset stratum size to the complete dataset stratum size. The result is the number of cases of each missing data pattern in each stratum to blank out for our study.

We blanked out the data for each pattern of missing data separately, beginning with missing data pattern 1 (missing only salary). We generated a uniform random number for all cases. Some cases are missing responses due to the skip pattern of the questionnaire. Responses for such cases are filled with a "logical skip" value. These should not be blanked out, so such cases received a random number of 1.0. Sorting by the stratum number and random number, we gave each case a within stratum number, and blanked out salary for those cases with a case number less than or equal to the number of pattern 1 cases to be blanked. We repeated these steps for the remaining six patterns of missing data within the same stratum. We made sure not to choose a case that had already been blanked out earlier by setting the random number for any cases altered previously to 1.0. This process was then completed over the remaining strata.

The salary variable was related to two of our additional variables. These also needed blanking. During telephone interviewing, people refusing to answer salary were asked if they would place their salary within a given range. The ICRANGE1 variable consisted of 12 intervals of $10,000. A total of 114 cases that did not respond to salary responded to ICRANGE1. Some respondents preferred to categorize their salary as greater than or less than $60,000 (ICRANGE2). Out of the 31,366 complete SDR 2001 cases, 1,489 (4.75%) did not respond to salary. For the procedure that follows, we restricted ourselves to those cases on the complete dataset and the study dataset where salary was blank. We found ICRANGE1 was present in the complete dataset at approximately the same rate for each level of the stratification variables mentioned earlier (demographic group, gender, collapsed field of degree). Thus we only stratified by the 12 levels of ICRANGE1. We counted the number of cases at each level of ICRANGE1 in the complete dataset, and scaled these by the ratio of the study blank salary (1,402) to the 1,489 complete cases missing salary. This resulted in the total number of each level of ICRANGE1 not to blank out in the study dataset. As before, we assigned each case a random number, and sorted by ICRANGE1 and the random number. We assigned a within stratum number, and blanked those cases with a within stratum number greater than to the number of cases to be kept. We kept ICRANGE1 for the remaining cases.

Once the narrower income range was blanked out, we blanked out ICRANGE2. Any cases with a nonblank ICRANGE1 must also have a nonblank ICRANGE2. Setting the random number for cases keeping ICRANGE1 to zero accomplished this. ICRANGE2 was present at approximately the same rate between genders, but differed between collapsed field of degree. Also, there was a difference between natives and foreign born, but not the other levels of demographic group. We stratified the cases by collapsed demographic group (natives/foreign born, 2 levels) and collapsed field of degree (5 levels). This resulted in 20 strata. Sorting by the stratum number and random number, we gave each case a within stratum number, and blanked out ICRANGE2 for those cases with a within stratum number greater than the number of stratum cases to be kept. We kept ICRANGE2 for the remaining cases.

Using our study dataset, we imputed the missing data using both the current imputation procedure and the flexible matching imputation procedure. The results are discussed in the next section.

5. ANALYSIS

We analyzed the performance of the current imputation procedure and the FMI procedure by comparing their results to the observed responses for each of the three study variables. We used several different measurements to assess performance. The categorical (citizenship, faculty rank) and semi-continuous (salary) variables required different techniques.

Categorical Variables We looked at several different statistics to analyze our results for categorical variables. Some of these are designed to measure the association between two categorical variables. Cohen's Kappa differs from these in that it is designed to test agreement between two categorical variables. The measures are:

1. Cross-classification matrix and percent agreement.
2. W Statistic – This is a statistic proposed by Stuart [6]. It is used to test the null hypothesis that the distribution of the variable is preserved. Suppose a variable has c+1 levels. W is calculated as:

$$W = (\mathbf{R}-\mathbf{S})^t \left[\mathrm{diag}(\mathbf{R}+\mathbf{S}) - \mathbf{T} - \mathbf{T}^t\right]^{-1} (\mathbf{R}-\mathbf{S}),$$

where **R** is a c-vector of imputed counts for the first c categories of the variable, **S** is the c-vector of observed counts for these categories, and **T** is the

cross-classification matrix for these categories. W is distributed approximately $\chi^2(c)$ for large n.

3. D Statistic – We can assess how well an imputation method preserves individual values by measuring the proportion of responses that change between the observed responses and the imputed responses [7]. We calculate: $D = 1 - 1/n \sum_{i=1}^{n} I\left(\hat{Y}_i = Y_i^*\right)$ and $\hat{V}(D) = n^{-1}(1-D)$. The indicator function in the summation above is 1 if the imputed value $(\hat{Y}_i)$ is the same as the observed value (Y_i^*). This measure is also called the Gross Difference Rate [8]. The percent agreement is 1 minus D, expressed as a percentage. A large GDR indicates serious response variance in the data. The value of D is zero if values are preserved perfectly. We form a confidence interval around D to test that D is significantly different from zero.
4. Kappa – We calculate Cohen's Kappa statistic to assess whether the agreement is larger than agreement due to chance [9]. This is calculated using the SAS PROC FREQ procedure [10]. A value closer to one indicates stronger agreement.
5. Index of Inconsistency (I) – This estimates the ratio of response variance to total variance for a question value. The L-fold index we use here averages I across all categories for the question. I ranges from 0 to 1, with values closer to 1 indicating higher variance between the two methods[2] [8].
6. Chi-Square based statistics – The adjusted contingency coefficient and Cramer's V statistic can be used to test the association in our cross-classification tables. $C_{adj} = \sqrt{\frac{\chi^2}{n+\chi^2}} \Big/ \sqrt{\frac{r-1}{r}}$, where r is the number of rows in the number of rows in our table. $\text{Cramer's V} = \sqrt{\chi^2/n(r-1)}$, where n is the total sample size. These test for association of the imputed responses with the observed responses, but do not indicate the magnitude of the association [11]. They are also only appropriate for tables with very few small cells. Thus they are inappropriate for our analysis of citizenship and faculty rank.
7. Fisher's test – We also assess the association between the observed and imputed values using this test. It is conducted using the SAS PROC FREQ procedure [10]. We use this test for citizenship and faculty rank because the many small cells in our cross-classification tables make the Chi-Squared statistic unreliable.
8. Lambda – This is a measure based on the proportional reduction in error using the imputed variable to predict the observed variable [11,12]. This statistic does give a degree of association. Higher values represent more association. It is calculated as:
   E1 = number of errors assuming no association
   E2 = number of errors assuming perfect association
   $\lambda_r = \frac{E1 - E2}{E1}$.

[2] The index of consistency is frequently used for response variance analysis. The index as used here is calculated the same way, but is based on different underlying assumptions.

Semi-Continuous Variables We analyzed the performance in imputing salary in four ways:

1. Categorize, calculate W and D – Semi-continuous variables may be categorized and analyzed as we did the categorical variables [7].
2. Univariate Statistics – Calculate the difference between the observed and the imputed data, and look at the mean, variance, etc.
3. Correlation – We may calculate the correlation between the imputed and observed values. We used both Pearson's and Spearman's correlation coefficients.
4. Regression – Finally, we can regress the observed salary values on the imputed salary values using the model $Y^* = \beta\hat{Y} + \varepsilon$. If the imputation procedure worked perfectly, we would have a model with a slope of one [7]. We can test that $\beta = 1$.

Summary Tables The National Science Foundation (NSF) publishes detailed statistical tables of characteristics of doctoral scientists and engineers [13]. We calculate some of these estimates of interest, and evaluate how they compare for the observed values, the FMI estimates, and the current imputation procedure estimates.

6. RESULTS

This section presents the results of our analysis of the two imputation procedures.

Citizenship – Current Method. Tables 1 and 2 show the cross-classification tables for the current imputation and FMI procedures, respectively. The evaluation statistics are presented in Table 5.

The responses imputed using the current procedure agree with the observed responses for 93.49% of the cases. The W statistic has a very high p-value,

indicating that we can conclude that the distribution of citizenship was preserved by our imputation procedure. The confidence interval of D indicates that we may conclude that individual values were also preserved. The Kappa statistic was quite high, indicating a high amount of agreement. The p-value from Fisher's test indicates that there is a statistically significant association between the counts imputed using the current method and the observed number of counts. In addition, the index of inconsistency is quite low, indicating good evidence of consistency between the imputed and observed values.

Table 1 – Current Imputation Procedure
Cross-Classification Table for Citizenship

		Imputed Responses					
	Values	1	2	3	4	5	Total
Observed Resp.	1	245	1	0	0	0	246
	2	1	38	4	0	0	43
	3	0	6	14	5	0	25
	4	0	0	4	5	0	9
	5	0	0	0	0	0	0
	Total	246	45	22	10	0	323

Citizenship – FMI Method. The conclusions using the FMI method are similar. There was 94.12% agreement. The W statistic leads us to conclude that the distribution is preserved, and the D statistic also indicates that individual values are preserved. The Kappa and lambda statistics are quite high, indicating a high amount of agreement and association, respectively. And finally, the index of inconsistency is quite low.

Table 2 – FMI Procedure
Cross-Classification Table for Citizenship

		Imputed Responses					
	Values	1	2	3	4	5	Total
Observed Resp.	1	246	0	0	0	0	246
	2	2	38	3	0	0	43
	3	0	5	14	6	0	25
	4	0	0	3	6	0	9
	5	0	0	0	0	0	0
	Total	248	43	20	12	0	323

Faculty Rank – Current Method. Tables 3 and 4 present the cross-classification tables for faculty rank, and the evaluation statistics are given in Table 5. We had generally lower performance with this variable. The percent agreement was 78.33%. The p-value on the W statistic was lower, but still leads us to conclude that the imputation method preserved the distribution. However, the D statistic confidence interval does not contain 0. We conclude that individual values were not preserved by the imputation procedure. The Kappa statistic still indicates a moderately high agreement, and the Kappa and lambda statistics are still favorable.

Table 3 – Current Imputation Procedure
Cross-Classification Table for Faculty Rank

		Imputed Responses									
	Values	1	2	3	4	5	6	7	8	9	Total
Observed Responses	1	1	0	0	0	0	0	0	0	0	1
	2	1	6	0	0	1	0	0	2	1	11
	3	2	1	32	1	1	0	0	0	0	37
	4	0	0	1	32	4	0	0	2	0	39
	5	1	0	0	2	18	0	2	0	0	23
	6	0	0	1	0	0	1	0	0	0	2
	7	0	1	0	0	0	0	1	1	0	3
	8	0	1	0	0	0	0	0	3	0	4
	9	0	0	0	0	0	0	0	0	0	0
	Total	5	9	34	35	24	1	3	8	1	120

Faculty Rank – FMI Method. While the FMI procedure had the same percent agreement as the current method, the conclusion of the W statistic is that the distribution of faculty rank is not preserved. However, the Kappa measure of agreement and lambda measure of association are quite similar to that of the current imputation method.

Table 4 – FMI Procedure
Cross-Classification Table for Faculty Rank

		Imputed Responses									
	Values	1	2	3	4	5	6	7	8	9	Total
Observed Responses	1	1	0	0	0	0	0	0	0	0	1
	2	1	5	1	0	3	1	0	0	0	11
	3	2	0	32	2	1	0	0	0	0	37
	4	0	1	4	31	3	0	0	0	0	39
	5	0	1	0	0	22	0	0	0	0	23
	6	0	0	0	0	2	0	0	0	0	2
	7	0	0	1	0	0	0	1	1	0	3
	8	0	1	0	0	0	1	0	2	0	4
	9	0	0	0	0	0	0	0	0	0	0
	Total	4	8	38	33	31	2	1	3	0	120

Salary – Current Method. We employed some different analysis methods for salary, as it is not a categorical variable. We can classify the variable into categories, and use the same techniques as those used for the citizenship and faculty rank variables. We divided salary into ten categories, based on the deciles of the observed responses to the salary question. The cross-classification tables are large, so we present the percentages of cases imputed to a different category than their observed category. The resulting statistics are found in Table 6.

The results of the current imputation procedure were less encouraging than those for the categorical

variables. There is 37.35 percent agreement, with 31.86% of the cases being imputed to a category two levels above or below their observed one. The W statistic still indicates that the distribution is preserved, but the p-value was much lower than for citizenship and faculty rank. The D statistic indicates that individual values are not preserved, and the Kappa statistic is rather low, indicating less agreement. The Chi-Squared p-value indicates a high level of association, however. The index of inconsistency is higher, indicating that there is less agreement between the imputed and observed responses. The Chi-Squared related measures of association (Adjusted contingency coefficient, Cramer's V) both indicate a more moderate association.

Table 5 – Summary of Evaluation Statistics

	Citizenship		Faculty Rank	
	Current	FMI	Current	FMI
% Agreement	93.50%	94.12%	78.33%	78.33%
W – Statistic	0.5111	3.5000	11.1912	14.5250
DF	3	3	8	7
P-Value	0.9164	0.3207	0.1911	0.0426
D – Statistic	0.0650	0.0588	0.2167	0.2167
Var (D)	0.0029	0.0029	0.0065	0.0065
lower CI	(-0.0425,	(-0.0491,	(0.0551,	(0.0551,
upper CI	0.1726)	0.1668)	0.3783)	0.3783)
Fisher p-val	1×10^{-81}	2×10^{-83}	NA*	NA*
Kappa	0.8355	0.8498	0.7184	0.7133
lower CI	(0.7749,	(0.7908,	(0.6269,	(0.6200,
upper CI	0.8961)	0.9089)	0.8099)	0.8066)
Lambda	0.7273	0.7532	0.7160	0.7037
index I	0.1645	0.1502	0.2816	0.2867

* We could not calculate the Fisher p-value for faculty rank due to lack of memory resources.

A plot of the observed values of salary versus those imputed using the current imputation method was created. The points do not lie on a straight line with a slope of 1 as we would like. The slope of the least-squares regression line is 0.7633, which is significantly less than one[3]. The Pearson correlation between the imputed values and the observed values is 0.4754. There are some outliers in the data, and Spearman's correlation is a bit higher (0.6908).

Salary – FMI Method. Our conclusions using the categorized salary variable are similar to those of the current imputation method. The percent agreement is 32.57%, and 33.64% of the missing cases were imputed to a category two or more levels away from the observed one. We have evidence that the distribution is preserved, but that individual values are not. The Kappa and lambda statistics are relatively low, and the adjusted contingency coefficient and Cramer's V statistic show a more moderate association.

Table 6 – Summary of Salary Statistics

	Salary	
	Current	FMI
% Agreement	37.35%	32.57%
W – Statistic	8.1972	6.7923
DF	9	9
P-Value	0.5144	0.6587
D – Statistic	0.6265	0.6743
Var (D)	0.0003	0.0002
lower CI	(0.5939,	(0.6438,
upper CI	0.6591)	0.7047)
X^2 p-val	<0.0001*	<0.0001*
Kappa	0.3038	0.2506
lower CI	(0.2757,	(0.2234,
upper CI	0.3320)	0.2779)
Lambda	0.3006	0.2472
index I	0.6962	0.7494
Percentage of missing responses imputed as a different category than observed		
Difference	Current	FMI
≥ 2	16.46%	15.54%
1	13.68%	14.47%
0	37.35%	32.57%
-1	17.11%	19.32%
≤ -2	15.40%	18.10%

* For Salary, we were able to use the Chi-Square statistic, rather than the Fisher test.

We also created a plot of the observed salary versus the FMI method salary. These do not lie on a slope of 1 either. The slope of a regression line through these data is 0.8755, significantly less than 1. The Pearson correlation for these data is 0.4263, and Spearman's correlation is 0.6942.

Table 7 – Univariate Salary Statistics

	Obs. Salary	Current Salary	FMI Salary	Current Diff.	FMI Diff.
Mean	80,856	81,531	77,757	-675	3,099
St. Dev.	51,372	63,399	48,459	59,683	53,523
Median	74,000	75,000	72,100	0	2,000
Mode	60,000	80,000	100,000	0	0
IQR	47,700	46,100	48,703	20,427	25,000

Salary – Univariate Statistics. Table 7 presents some univariate statistics for our different imputations. The current imputation seems to overestimate salary. It has a larger average, more variability, and the median and mode are larger. The average difference was slightly greater than zero ($675). The FMI method seems to underestimate salary. It has a lower average salary and

[3] Measured at the 90% confidence level.

median, but higher mode. It averaged $3,099 less than the observed values, but showed less variability.

Comparing Estimates from the Data. Using the imputed data, we calculated several estimates of interest. These are similar to some of the estimates in reference [13]. Specifically, we looked at totals of native born and naturalized citizens, permanent and temporary residents, as well as totals for different faculty ranks: professor, assistant professor, associate professor, instructor, lecturer, and adjunct. We found that both imputation methods give estimates that are quite similar. We found that the imputation methods had very little affect on the overall citizenship and faculty rank totals. This was true at even finer levels, such as gender by collapsed field of degree. However, these similarities can partially be attributed to the low item nonresponse rate for this survey.

7. CONCLUSION

In general, the two methods yield results that are quite similar. They both perform quite well for imputing citizenship, but less well for faculty rank, and even more poorly for salary. Even if the methods performed less well, the distribution was preserved for all but the FMI faculty rank imputation. The evaluation statistics for each method were all quite similar, also.

Given the similar performance of these two methods, the FMI procedure looks like a very feasible option. However, we found during processing that it is problematic when imputing categorical data. In addition, more manual decision work was required when dealing with the skip patterns of this survey. It was not quite as automated as we anticipated. Further work to determine and automate the decisions and steps of the FMI procedure is needed prior to its use for the Survey of Doctorate Recipients.

In the future, we plan to repeat these imputation processes multiple times, to get a better feeling for the variability introduced by each method. We also plan to investigate more detailed totals, such as citizenship $\times$ gender, to see how distributions are preserved at smaller levels. We would also like to investigate how to incorporate consistency checks for the imputed values in the FMI procedure.

The authors would like to thank Kelly Kang of the National Science Foundation for her support of this research, and David Pysh of the U.S. Census Bureau for his efforts processing the data for the current imputation method.

8. REFERENCES

[1] Finamore, J., and Stratton, J., "Sample Design for the 2001 Survey of Doctorate Recipients (SDR01-SAMP-4)," Internal Census Bureau Memorandum for Documentation for Chester E. Bowie from Alan R. Tupek, November 20, 2002.

[2] Stratton, J., "The 2001 Survey of Doctoral Recipients Weighting Plan (SDR01-WT-1)," Internal Census Bureau Memorandum for Documentation for Chester E. Bowie from Alan R. Tupek, October 15, 2002.

[3] U.S. Census Bureau, "2001 Survey of Doctorate Recipients Methodology Report," November 2002.

[4] Stratton, J., and Finamore, J., "Imputation Specification for the 2001 Survey of Doctorate Recipients (SDR01-IMPT-1)," Internal Census Bureau Memorandum for Chester E. Bowie from Alan R. Tupek, June 13, 2001.

[5] Akaike, H. (1987), "Factor Analysis and AIC," *Psychometrika* 52, 317 -332

[6] Stuart, A. A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika* 42 (1955), 412.

[7] Chambers, Ray. "Evaluation Criteria for Statistical Editing and Imputation," National Statistics Methodological Series No. 28, London, 2001.

[8] U.S. Bureau of the Census, Evaluating Censuses of Population and Housing, Statistical Training Document, ISP-TR-5, Washington, D.C., 1985.

[9] Bishop, Y.M.M., Fienberg, S.E. , Holland, P.W., Discrete Multivariate Analysis: Theory and Practice, MIT Press, Cambridge, 1975.

[10] SAS Institute. 1999. SAS/STAT User's Guide, Version 8. SAS Institute, Inc., Cary, North Carolina.

[11] Ott, L., Larson, R.F., Mendenhall, W., Statistics: A Tool for the Social Sciences: 3rd Edition. Prindle, Weber, & Schmidt, Boston, 1983.

[12] Goodman, L.A., Kruskal, W.H., "Measures of Association for Cross Classifications," JASA, Vol. 49, No. 268 (Dec, 1954).

[13] National Science Foundation, Division of Science Resources Statistics, *Characteristics of Doctoral Scientist and Engineers in the United States: 2001*, NSF 03-310, Project Officer, Kelly H. Kang (Arlington, Va 2003).