Imputation Bias Based on Different Formulations of Imputation Classes for the IPEDS Finance Survey

Marcus Berzofsky[1], June Z. X. Cong[1], Shiying Wu[1] and Roy Whitmore[1]
RTI International[1]

Introduction

IPEDS, the Integrated Post-Secondary Education Data System, is a set of annual surveys that all U.S. post-secondary institutions that participate in the Title IV student aid programs are required to complete. It is sponsored by the National Center for Education Statistics (NCES). IPEDS has nine required components. One of these components is the Finance survey, which collects information on an institution's assets as well as their revenue and expenditures. Institutions provide data for this survey using one of three forms determined by whether they are a public, a private not-for-profit or a private for-profit institution.

IPEDS is a census of participating Title IV institutions. To this end, NCES would like to provide data for all Title IV institutions. Therefore, when a participating Title IV institution is a nonrespondent for one of the components, NCES requires that their information be imputed. In these cases, we do not have a response value for any of the items in the component and, therefore, we must impute for all items in that component. So while this type of nonresponse is technically at the unit level we treat it as if it were item nonresponse. This is done at NCES's request because they want to avoid weighting the data, which is the standard solution for unit nonresponse. Currently, since IPEDS is a census, all institutions have a weight of one which makes data manipulation very easy. IPEDS data is made publicly available on the internet and a majority of the users of the data do not have a statistical background. Therefore, since IPEDS only has a 2% to 3% nonresponse rate, NCES decided that it was better to provide the user a complete data set that user could download and use with ease. Alternatively, weight adjustments could be made to account for the nonrespondents, however, NCES felt that using the data with weights would be more cumbersome for the public and the data could be used improperly if the weights were not used applied correctly.

A nonresponding institution's data are imputed via one of three imputation methods. First, if an institution has completed the survey in either of the past 2 years, a carry forward method is used. If an institution has not completed the survey in the past 2 years, a nearest neighbor procedure is used by identifying a donor from an imputation class. If a donor cannot be determined, a group median procedure which takes the data from the median institution in the imputation class is used.

At this time, it will be helpful to review the imputation methods used by IPEDS in relation to their bias and appropriateness as found in the literature. Sande (1982) suggests the most relevant concerns to consider are the bias and variance of the estimates.

Most studies have found that imputations based on a unit's previous values performed the best (Engels and Diehr, 2003). In the carry forward procedure used by IPEDS, an institution's last known value is ratio adjusted based on the change found in the nonrespondent's imputation class. This is an improvement over the carry forward method used by Engels and Diehr because they simply replaced the missing value with the most recent reported value without making any time adjustments.

The nearest neighbor procedure is preferred as a second alternative since it allows for the use of auxiliary information that may be highly correlated with the characteristic of interest in choosing a donor (Montaqula and Ponikowski, 1995). In addition, the literature suggests that the nearest neighbor procedure yields point estimates with small or negligible bias, assuming that a linear relationship exists between the variable of interest and the variable used for nearest neighbor identification (Rancourt, Sarndal and Lee, 1994; Chen and Shao, 1997; Hu, Salvucci and Cohen, 1998).

Mean or median imputation has been found to be, while the simplest, the least accurate and found to have the largest bias and underestimate variances when compared to the carry forward method, the nearest neighbor method and other methods that use additional information when determining a donor (Hu, Salvucci and Cohen, 1998; Engels and Diehr, 2003).

When previous data are not available, how the imputation classes are constructed will have a major impact on the bias of a nonrespondent's data. The goal is to construct pools that explain as much of the variance in the variables to be imputed as possible. This allows the use of the assumption that the mechanism that leads to missing values is 'ignorable' (Little & Rubin, 2002). In other words, the missing values are as though they were missing at random (Administration for Children & Families, 2005). Moreover, if imputation classes are constructed poorly then bias can be introduced (Durrant, 2005).

The Finance survey is a large survey with over 200 variables. Due to the fact that several of these variables throughout the survey form are related, NCES requires that, during imputation, a single donor be used for all 200 variables to maintain consistency across the survey. Moreover, due to the interrelated items, the choice of techniques for imputing involve considerations somewhat different from those when only a single variable is being imputed (Sande, 1982). In this case, it is important to identify donor institutions that are as similar to the nonrespondent for as many components of the survey as possible. This makes it extremely important to construct the 'best' donor pools, or imputation classes, where 'best' is defined as being the pool which results in generating estimates with the smallest amount of error (Robertson, Tou and Huff, 1995).

Currently, the Finance survey uses a combination of Principle Component Analysis (PCA) and Chi-squared Automatic Interaction Detector (CHAID) to determine the

imputation classes. While the imputation classes defined by this method are based on statistical algorithms, they are data driven and may change from year to year. This study attempts to determine if the non-statistical method could be used for defining imputation classes for the Finance component. The motivation to conduct this study is three fold. Since the non-statistical method has been the existing method for all other components of IPEDS it is important to determine if the newer method, PCA/CHAID, is performs better. Furthermore, the non-statistical method is easier to implement and would save resources. Third, estimates created based on the PCA/CHAID imputation class method may not be comparable from year to year. In other words, if would not necessarily be clear if changes in the estimates based on imputed data were attributable to changes in the imputation classes or an actual change. We hypothesized that the PCA/CHAID method for creating imputation classes would produce better imputations than the non-statistical method.

Methods

In order to achieve our goal, there were three major issues that needed to be resolved. First, we defined and compared the two types of imputation classes. Second, we conducted an analysis to determine if nonresponse was missing completely at random. Prior to testing the imputation procedures, we determined if the set of nonrespondents was random across all institutions or if institutions with particular characteristics had a higher propensity for being a nonrespondent. Third, the methodology for analyzing the bias of the imputation classes was designed.

The non-statistical imputation classes were defined prior to development of the PCA/CHAID imputation classes. These imputation classes were based solely on subject matter expertise without any aid from statistical analysis. Table 1 displays the variables that were selected for defining the imputation classes.

Census Division was included for the public institutions because public institutions within the same state, and to a lesser extent the same Census Division, share common characteristics in terms of the level of funding they receive from their state government and the level of non-public revenue that is generated. However, since private institutions receive no formal public funding, their finances were not believed to have regional ties.

Imputation classes under the non-statistical method had a minimum of 9 institutions. If an imputation class based on its initial definition was smaller than 9 institutions it was collapsed with an imputation class with similar characteristics to form a class of at least 9 institutions. Under this method 31 imputation classes were created.

We formed PCA/CHAID imputation classes by first performing PCA, using PROC PRINCOMP in SAS version 8.2, to create a summary index for each institution and then we used CHAID, which partitions the data into mutually exclusive, exhaustive subsets that best describe the dependent variables to cluster the institutions into homogeneous imputation classes, using Answer Tree 2.0 software (Kass, 1980). The Finance survey forms contain 20 summary variables. These variables contain the most information about an institution's most recent finances and are the best source to use for grouping institutions. The values of these variables were used in the PCA to generate a weight for each variable. An index for each responding institution was generated using the formula $I_j = \sum_{i=1}^{20} w_i x_{ij}$ where w_i is the PCA weight for variable i and x_{ij} is the value of variable i for institution j. We selected predictor variables to be used in the CHAID analysis based on prior knowledge of the data set. Variables used had to be known for both respondents and nonrespondents (Department of Veteran Affairs, 2004) Through regression analysis, using the index variable as the dependent outcome, the initial set of predictor variables was subsetted based on each variable's significance to predict the index. We inserted the variables that had the greatest significance into the CHAID algorithm, with the institution's index as the outcome variable, to determine the best imputation classes. We performed separate analyses using the 2002, 2003 and 2004 surveys. The variables that were most commonly found in the resulting CHAID trees were incorporated into the final set of variables. This imputation method yielded 45 imputation classes.

The universe of eligible institutions for this analysis was defined as all institutions that responded

Table 1: Variables Used in Definition of Non-Statistical and PCA/CHAID Imputation Classes

Non-Statistical	PCA/CHAID
Form used (Public, Private not-for-profit or Private for-profit)	Form used (Public, Private not-for-profit or Private for-profit)
Institutional level and control	Degree granting status (yes/no)
Medical school (yes/no)	Offer graduate or first-professional courses (yes/no)
Levels of offering (undergraduate, graduate, first-professional	Medical school (yes/no)
Census division (Public schools only)	FTE students (categorical)
	Student services offered (yes/no)
	Athletics offered (yes/no)

to the 2005 survey (N=5,084). A second universe, a nonrespondent universe, was defined as those that responded to the 2005 survey, but were a nonrespondent in at least one of the 2002, 2003 or 2004 surveys (N=388). Table 2 illustrates the distribution of institutions from both universes across institutional level and institutional control.

Five hundred independent samples of size n=49 where taken from each of these universes. The sample size was based on the number of complete nonrespondents in the 2005 survey and would, therefore, best simulate how our imputation procedures would be implemented under normal conditions. Within the universe of eligible institutions, the sample was drawn proportional to the number of institutions in each institutional level and control. This was done because under the assumption that complete nonresponse is random one would expect that the complete nonrespondents follow a distribution similar to the population distribution. For the nonrespondent universe, the sample was drawn via a simple random sample among all previous nonrespondents. Since each of these institutions has shown a previous propensity to be a nonrespondent, under the assumption that complete nonresponse is not random, each of them was assumed equally likely to be a nonrespondent in the future regardless of their institutional level or control. In order to increase the power of our analyses, 500 replicate samples were drawn from each distribution. Also, using multiple simulations in this manner provides the opportunity to observe what happens to the estimates as different sets of units are deleted to simulate nonrespondents and are represented by the remaining respondents (Robertson, Tou and Huff, 1995).

Within each replicate, the data for the sampled institutions were removed and the nearest neighbor procedure was implemented to identify a donor institution using both the non-statistical imputation classes and the PCA/CHAID imputation classes. Imputed values for each variable were calculated by multiplying the ratio of full time equivalent (FTE) students in the imputed institution over the FTE students in the donor institution by the donor institutions value. Once the imputed values were determined, the no intercept regression model $imputed = \beta_i(response) + e_i$ was fitted to determine the beta value for the *i*-th variable which represents the slope of the regression line in this simple linear model and, thus, represents the bias of the imputation. If the imputation was perfect, then β_i would equal 1 for the *i*-th variable.

In order to determine which imputation class method created the best imputed values, the distribution of the beta estimates was compared for each variable across the 500 replicates and across all of the finance variables. Similar to the methods used by Engels and Diehr (2003), in order to compare the distributions across the 200 finance variables, the mean square error (MSE) was calculated for each β_i using the formula $MSE(\beta_i) = \left(\overline{\beta}_i - 1\right)^2 + V(\beta_i)$ where $\overline{\beta}_i$ is the mean beta coefficient among the 500 replications for the i^{th} variable, 1 is the expected value of the beta coefficient and $V(\beta_i)$ is the variance of the distribution of β_i estimated by the 500 replicates.

Table 2: Distribution of Eligible Samples and Institutional Level and Control

Institutional Level and Control	% Dist. of Eligible Population (N=5,084)	% Dist. of Nonrespondent Universe (N=388)	Avg. % Random Sample Dist. (n=49)	Avg. % Nonrandom Sample Dist. (n=49)
Administrative unit only	0.00	0.00	0.00	0.00
Public, 4 year and above	11.15	1.29	11.14	1.31
Private not-for-profit, 4 year and above	28.40	15.98	28.41	16.04
Private for-profit, 4 year and above	4.45	2.58	4.44	2.57
Public, 2 year	18.17	10.05	18.17	10.10
Private not-for-profit, 2 year	2.68	4.38	2.69	4.08
Private for-profit, 2-year	11.53	16.24	11.50	16.05
Public, less than 2-year	3.11	7.73	3.20	7.76
Private not-for-profit, less than 2-year	1.24	1.55	1.19	1.54
Private for-profit, less than 2-year	19.28	40.21	19.27	40.55

Analysis

After defining the two universes in which institutions were eligible, a chi-square test of homogeneity was performed to determine if institutions that had previously been nonrespondents were missing completely at random based on their distribution by institutional level and control. This test yielded a test statistic of 215.36 and a p-value less than 0.0001. Therefore, it was concluded that complete nonresponse was not missing completely at random, and all further analyses were based on the samples from the nonrespondent distribution.

Figures 1 – 3 compare the distribution of the beta values based on imputations using the non-statistical classes and the PCA/CHAID classes for the net asset variables for public, private not-for-profit and private for- profit institutions, respectively. The 'X' in each box plot represents the mean of the beta coefficients across the 500 replications.

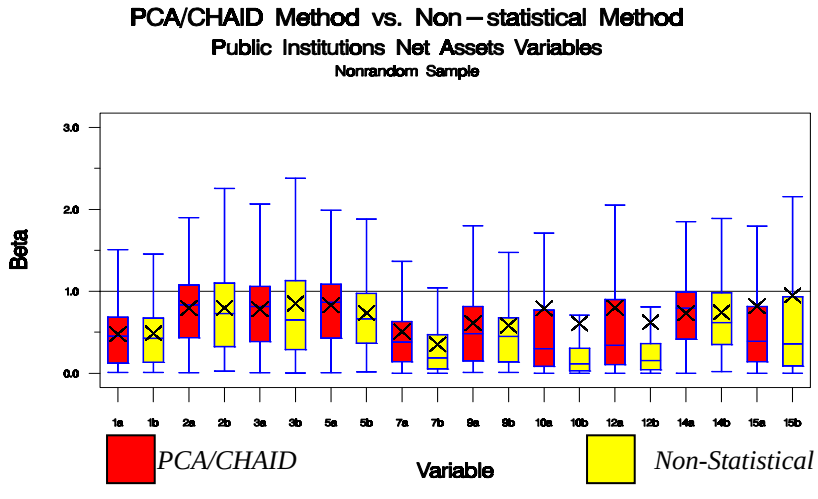
In all three figures, the distribution of the net asset variables is very similar between the imputation class methods. The one exception was with the private not-for-profit institutions. For this control type the median beta estimate, for all of the net asset variables, is closer to one under the non-statistical method than under the PCA/CHAID method. In addition, for all institutional control types and under both imputation class methods, the median beta estimate is less than 1. This implies that, regardless of imputation class method, the imputed values were biased downward, meaning the imputed values for each variable were less than the actual reported value.

Figure 4 shows the distribution of the MSE for all variables by institutional control and Figure 5 shows the distribution of MSE across all institutions. Similar to the distribution of the beta estimates of each individual variable, the distribution of the MSES is very similar for each institutional control type. Moreover, under both imputation methods, the median MSE is never greater than 1.5, which indicates that overall the imputed values were a good proxy for their true respondent value. Table 3 displays the distribution of the mean square errors. It can be seen from this table that the interquartile range is smaller for the non-statistical method for all institutional control types (22.9 for PCA/CHAID vs. 13.9 for non-statistical in public institutions, 9.6 vs. 4.6 in private not-for-profit institutions and 1.0 vs. 0.5 in private for-profit institutions).

Discussion

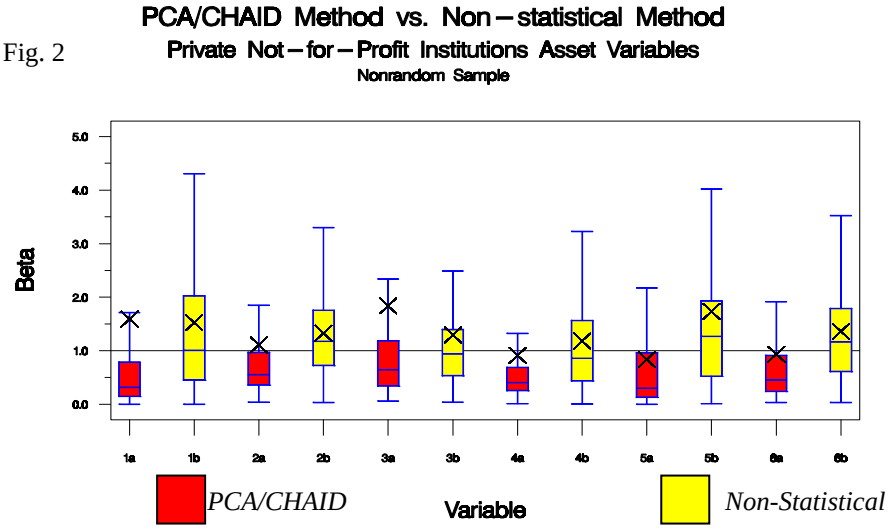
The results of this analysis were contrary to our expectations that the PCA/CHAID method for creating imputation classes would produce more accurate imputations. Instead, the results, as seen in Table 3, only indicate an overall difference in MSE of 0.10, which in this situation can be considered negligible. This caused us to review the process in which the PCA/CHAID imputation classes were created and determine places that may have hampered the effectiveness of this method. In doing this, we identified three places where the effectiveness of the PCA/CHAID method was diminished.

Fig. 1



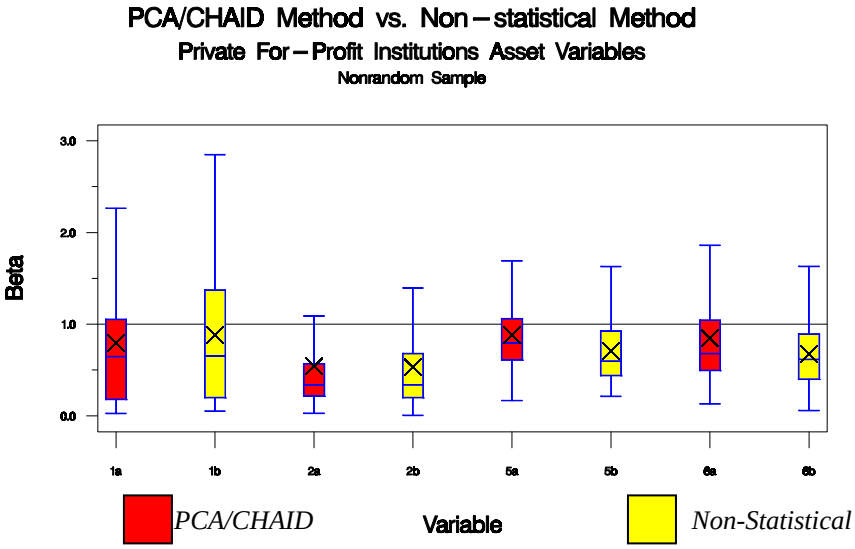
1=Current Assests 2=Capital Assests 3=Accumulated Depreciation
5=Total Noncurrent Assets 7=Long Term Debt, current portion 9=Total Current Liabilities
10=Long Term 12=Total Noncurrent Liabilities 14=Invested in Capital Assets
15=Restricted-expendable 16=Restricted-nonexpendable

Fig. 2



1=Long-term investment *2=Total assets* *3=Total liabilities*
4=Total unrestricted net assets *5=Total restricted net assets* *6=Total net assets*

Fig. 3



1=Total assets *2=Total liabilities* *5=Total revenues* *6=Total expenses*

Fig. 4

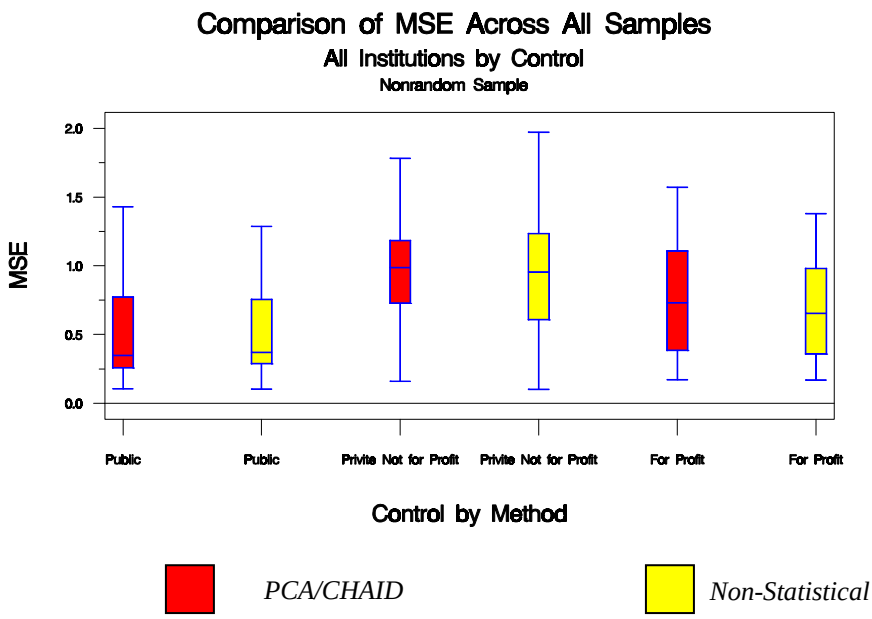
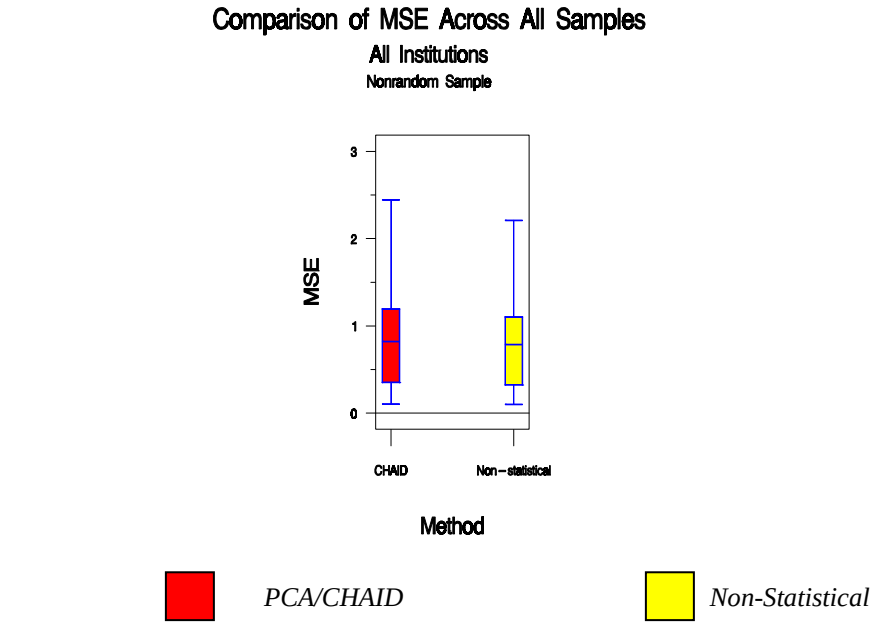


Fig. 5



First, the initial constraint during the imputation process was that only one donor would be used for all 200 finance variables. One advantage of using CHAID is that it allows donors to be found for specific variables or groups of variables. Since no two institutions are exactly the same, it stands to reason that different components of a nonresponding institution's finances align themselves more closely to different institutions. We were not able to take advantage of this benefit; however, we are not proposing that IPEDS imputations should be done this way because the need to have consistency across all variables outweighs the desire to be more accurate on any particular variable. Second, the index created for each institution that was used as the outcome measure in the CHAID algorithm was based on 20 summary variables. While these variables comprised 53% to 70% of the variation among all the finance variables, they did not explain all of it. In review of the variables included in the PCA, we determined that a large number of additional variables would be needed to make an appreciable difference in explaining the variation. This would have been logistically cumbersome in the PCA and, therefore, not pursued.

Table 3. Distribution of MSEs Across All Variables by Control and Imputation Class Method

	Public		Private Not-for-Profit	
	CHAID	Non-Stat	CHAID	Non-Stat
Mean	46,201,054	10,108.86	37.12	1,571.54
Min	0.10408	0.10312	0.15912	0.09906
25th Percentile	0.30496	0.33615	0.96128	0.92724
Median	0.99953	0.92544	1.44954	1.32948
75th Percentile	23.2158	14.3219	10.5242	5.50132
Max	2.5 Billion	296,119.62	1,884.01	110,965

	Private for Profit		Overall	
	CHAID	Non-Stat	CHAID	Non-Stat
Mean	5.67	7.49	22537115	5,570.8
Min	0.16982	0.16721	0.10408	0.46153
25th Percentile	0.54107	0.55595	0.56407	0.46153
Median	0.93824	0.75244	1.19236	1.03093
75th Percentile	1.57107	1.04083	10.9853	7.27908
Max	95.1646	152.71	2.5 Billion	296,119

Finally, the CHAID algorithm is dependent on the source data. Variables that were significant using one year's data were not necessarily found to be significant in either of the other two years analyzed. This was especially true when looking at the third or fourth branch of the CHAID tree. Due to the fact that we wanted to have relatively consistent imputation classes across years, only the variables that consistently were in the higher branches were used when determining the imputation classes. Therefore, variables that may have significantly improved the imputation classes for a specific year's data were not included. This probably diminished the effectiveness of the imputation classes for a given year. When reviewing the differences in how the imputation classes were defined for each method, it was interesting to note that the PCA/CHAID method did not find any region variable significant for public institutions. This was contrary to the assumptions used when creating the non-statistical imputation classes where we believed that public institutions in the same state or region would have correlated finance data and, therefore, needed to be grouped in the same imputation class. The median MSE across all variables for public institutions under both methods was very similar suggesting that region is not significant as found under the PCA/CHAID method. However, the mean of the PCA/CHAID method was much more skewed with a mean value of 46,201 compared to 10,108 under the non-statistical method. This suggests that incorporating region into the imputation class may help minimize extreme imputations.

Furthermore, unlike what some of the literature has found our analysis consistently produced a regression coefficient less than one for all variables under both methods. This implies a slight bias toward the null which means that our imputations underreported an institution's finance values. This is in line with the findings of Hu, Salvucci and Cohen (1998) in their evaluation of the nearest neighbor procedure which found a small, but consistently negative bias. While we do think that there is a linear relationship between the finance variables and FTE, which was used to identify a donor, the fact that so many variables were being imputed simultaneously may have diminished the effectiveness of the donor identification.

Moreover, after reviewing the literature, one thing that was not found was a method for comparing imputation bias across several variables. While the analysis methods used in this research are similar to those used by Engels and Diehr (2003), they only looked at the MSE for individual variables. Our analysis looked at ways in which an aggregate statistic could be created to summarize the bias of the imputation across several variables. In doing so, we found that the MSE of the distribution of all variables could be used as a summery statistic.

Conclusions

These findings suggest that using PCA/CHAID does not appreciably improve the imputations for the Finance survey. Therefore, due to the ease in which the non-statistical imputation classes can be implemented and the assurance of year to year consistency in class definition, if the same substantive expertise is available to form the classes, we recommend using the non-statistical method for creating imputation classes in future IPEDS Finance surveys.

While these results indicate that a non-statistical method in this situation performs as well a statistical method it must be pointed out that this result may not be as generalizable as we would like. Even though subject matter expertise was used when developing the non-statistical method, it cannot be assumed that the same results would occur for another set of data. Before using the results of this paper on another data set a researcher should confirm the findings for the specific data being used.

Furthermore, the literature confirms the hierarchy in which IPEDS determines which imputation procedure should be implemented using the carry forward method as the first imputation option followed by the nearest neighbor and group median methods.

While this paper focused on the bias created through the imputation process through the regression coefficient, we did not attempt to quantify the variance of the regression error. Future work should attempt to quantify this amount to give a complete picture of the imputation quality.

Additional future research that arises from these findings may be the need to derive a correction factor that can be used to correct the bias due to the imputation procedures used in IPEDS. This correction factor may need to differ depending on the imputation method used and the institutional level and control of the nonresponding institution.

References

Sande, I.G. (1982). Imputation in Surveys: Coping with Reality. *The American Statistician*, Vol. 36, No. 3, Part 1, 145 – 152.

Robertson, K.W., Tou, A. & Huff, L. (1995). A Study of Donor Pools and Imputation Methods for Missing Employment Data. *ASA Proceedings of the Section in Survey Research Methods*.

Kass, G.V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, 29, 119 – 127.

Engels, J.M. & Diehr, P. (2003). Imputation of Missing Longitudinal Data: A Comparison of Methods. *Journal of Clinical Epidemiology*, 56, 968 – 976.

Rancourt, E., Sarndal, C., & Lee, H. (1994). Estimation of the Variance in the Presence of Nearest Neighbor Imputation. *ASA Proceedings of the Section in Survey Research Methods.*

Chen, J., & Shao, J. (1997). Biases and Variances of Survey Estimators Based on Nearest Neighbor Imputation. *ASA Proceedings of the Section in Survey Research Methods*.

Hu, M., Salvucci, S.M., & Cohen, M.P. (1998). Evaluation of Some Popular Imputation Algorithms. *ASA Proceedings of the Section in Survey Research Methods*.

Montaquila, J.M., & Ponikowski, C.H. (1995). An Evaluation of Alternative Imputation Methods. *ASA Proceedings of the Section of in Survey Research Methods*.

Little, R.J. & Rubin, D.B (2002). *Statistical Analysis with Missing Data* (2nd ed.). Hoboken, NJ: Wiley & Sons, Inc.

Administration for Children and Families. Office of Planning, Research and Evaluation. Head Start Impact Study: First Year Findings. Appendix 4.1. Retrieved November 5, 2005 from http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/first_yr_finds/firstyr_finds_app4_1.html

Durrant, G. (June 2005). Imputation Methods for Handling Item-Nonresponse in the Social Sciences: A Methodological Review. NCRM Methods Review Papers. Retrieved November 5, 2005 from http://www.ncrm.ac.uk/publications/documents/MethodsReviewPaperNCRM-002_000.pdf