Missing Data Imputation Method Comparison in Ohio University Student Retention Database

Dyah Hening and David A. Koonce
Ohio University, USA

Abstract

Ohio University has been conducting research on first-year-student retention to prevent dropouts (OU Office of Institutional Research, First-Year Students Retention, 2008). Yet, the data set has more than 20% missing values. Missing data affects the ability in result generalization of the target population. This study categorizes the missing data into one of three types of missing data: missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). After the missing data is identified, the proper method of handling the data is discussed. Five methods were utilized in the research: mean, median, zero, hot-deck and multiple imputations. Despite the poor performance on the accuracy comparison test, multiple and hot-deck imputation have proven to improve the retention prediction rate. Mean and median imputation perform better in accuracy and are sufficient for the prediction model.

Keywords
Data imputation, missing at random (MAR), missing completely at random (MCAR), missing not at random (MNAR), multiple imputation, hot-deck imputation

1. Introduction
Quality is very important for many organizations as one of the parameters of success. Total Quality Management (TQM) is one of the quality control tools that initiate continuous improvement, and it has been implemented by profit and higher education organizations. The most important phase of the TQM implementation is customer identification. In higher education institutions, students are considered customers (Sirvanci 2004). Customer identification leads to tailoring systems for customer satisfaction, which has been the focus of TQM implementation. Aware of the importance of the students as their customers, many universities are striving to provide quality education by putting serious efforts into preventing student dropouts. One such institution is Ohio University (OU), which has been conducting research on first-year student retention in an effort to improve its quality of service to the student body. Still, the rate of students' retention at OU was generally declining over the years from 2003 to 2009, which can be seen in Figure 1 (OU Office 2008).

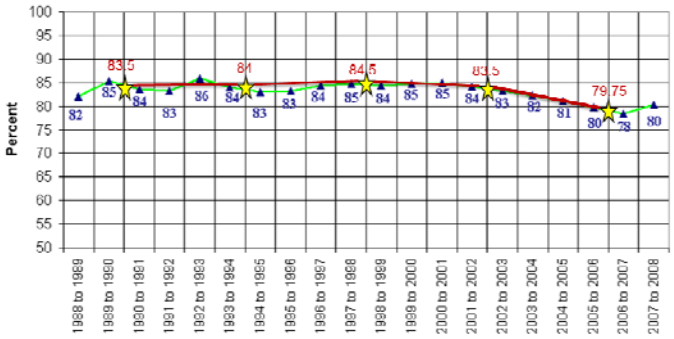

Figure 1: Ohio University overall first-year student retention.

Several research studies related to predicting student retention have been conducted. The research has led to some theoretical models on student retention development and highlighted the significant factors affecting student retention. Tinto (1975) synthesized his theory from social psychology and education economics in describing the interaction between individuals and college institutions, and his model led to the development of an attrition model.

This model involves a cost/benefit analysis and includes factors about students' family backgrounds, various individual characteristics, past educational experiences, and goal commitment levels.
In previous research, Roth (2008) developed a model to predict the retention of students based on the datasets of OU student behavior. However, large numbers of values in these sets are missing, especially numbers linked to attributes relevant to the prediction model. Roth used simple mean value imputation to fill in missing values. In general, mean value imputation affects the distribution of a given variable. However, most predictive analysis techniques do not include a specific method of handling missing data; hence it is necessary to address this problem. Current literature contains many proposed methods in dealing with missing data; yet, these techniques are not applicable to every situation and condition of missing values. The purpose of this paper is to develop a better understanding of how to handle missing data, especially for developing models to predict student attrition in student retention datasets. A comparison of the missing data imputation methods provide evidence of the most appropriate method to be used to fill in data for predicting student attrition.

2. Background

Missing data are a nuisance for statistical analysis. A threat for institutional research is that the missing data can affect a study's internal validity. In addition, missing data may have an effect on a study's external validity and limit its generalizability across a target population. Therefore, explore and identify ways to deal with missing data are important. According to Cohen et al. (2003), even when investigators employ conventionally appropriate strategies for coping with missing data, different approaches may lead to significantly different conclusions. To address missing data appropriately, it is helpful to understand the types and characteristics of the missing data. The most commonly occurring reason for missing data is non-response to items which, according to Umbach (2005), can stem from a variety of reasons. For instance, errors that might emerge during coding or data entry, respondents' inability to answer the survey questions, and the limitation of the study design can elicit responses (Umbach 2005).

2.1 Missing Values Categories

Gelman and Hill (2007) posit several reasons data may be missing. They group missing data into four types: missing completely at random (MCAR), missing at random (MAR), missing that depends on unobserved predictors, and missing that depends on the missing value itself. Missing values that depend on unobserved predictors and missing values that depend on the missing value itself can also be considered missing not at random (MNAR). These categories are meant to identify the characteristics of the data that will be missing, not the missing value itself. The three main categories of missing data—MCAR or missing completely at random, MNAR or missing not at random and MAR or missing at random— are summarized in table 1 below.

Table 1: Classification of missing data

Category	Characteristic	Example
MCAR	Missing values are independent of any factor	Unrecognized scanned responses
MAR	Attribute depends only on observed data	The pattern of missing data is traceable or predictable from other variables in the database.
MNAR	Missing values are based on the value of attribute	Value is sensitive and respondents choose not to respond.

MCAR occurs when any values of a variable have the same probability of being missing. The occurrences of non-response are quite common in sampling surveys, and in Gelman and Hill's study the mechanism of non-response is assumed as MCAR, (for more detailed explanation, see Rubin 1987). When missing data are MCAR, no specific clue could be derived from the other responses as to what the missing value should be. MAR, or missing at random, can be considered to be semi-MCAR. It occurs when the probability of any variable instance to be missing is the same for all units. However, what distinguishes MAR from MCAR is that with MAR the variable can be predicted from other available data. When data are MAR, omitting cases with missing data is accepted because doing so will reduce the bias of the inferences. MNAR can be subcategorized into: missing values which depend on unobserved predictors and missing values that depend on the missing value itself. In these cases the likelihood of a value being missing is dependent on some value. A good example of this comes from medical studies--when any particular

treatment causes discomfort to a patient, the likelihood of that patient walking out or dropping out will increase (Rubin 1987).

Another consideration that Rubin (1987) has put into his classification method is whether the missing data are ignorable or not. By "ignorable" he means that the whole variable can be omitted or disregarded in the model building. In cases of MAR, the ignorable missing data mechanism occurs when variables are less important or less related to the model than are other variables. This assumption has the same underlying philosophy as the causal framework, in which ignoring something can be done if sufficient evidence and information have already been gathered. So, in these cases, few correlated variables can be omitted. For example, suppose we want to predict someone's athletic capability or performance. The variable of favorite color would most likely not be related to the prediction model; therefore, excluding this variable will probably have few negative effects on the prediction model's accuracy.

2.2 Common Imputation Method

A simple method for supplying missing values is single imputation. There are three types of single imputation, based on the types of values: constant, randomly selected, and non-randomly derived values (Mcknight et al. 2007). Constant substitution refers to replacing the missing values with constant values such as mean substitution (either the arithmetic mean or the estimated mean of the population), median substitution, or zero imputation. Random imputation, which uses random values, consists of two major divergences: hot-deck and cold-deck imputation. The non-random imputations are derived values from regression, conditional imputation, or data that have been previously recorded from a subject. Despite different types of single variable imputations, these methods all have something in common--they assume that the standard error for the estimate is low.

2.2.1 Mean imputation

Due to the ease and simplicity of the following single imputation method, the most used type for supplying missing values is constant replacement (Mcknight et al. 2007). One constant replacement method is mean imputation, which consists of predicting the missing observation by simply filling in the missing values with the mean of the observed values. However, this method is less desirable because it tends to underrepresent extreme values, which biases the analysis by yielding a variable with greater central tendency than should be expected. This invalidates the estimates of variance and covariance, affecting the internal validity of the work. Another mean imputation method is based on the maximum likelihood (ML) algorithm. The arithmetic of this method slightly enhances the traditional mean imputation method regarding its sensitivity to the outliers' values. The methods draw on the assumption of normal distribution of the data. Although the ML substitution provides an estimate mean of the population (μ) instead of the sample mean, this method is still considered a less desirable method because the substantial deviations from the assumed normal distribution provide poor estimation.

2.2.2 Median imputation

Median substitution, is used when the data are not normally distributed, in which case the curve can be skewed, flat or peaked and cannot be represented by the mean replacement. Median imputation tends to produce larger standard errors, which is not optimal to avoid type I error. However, compared to the two previous constant imputation methods, it is better at reducing type II error (Mcknight et al. 2007).

2.2.3 Zero value imputation

The last common type of constant replacement method is replacing the missing values with a value of 0 based on logical rules. If the missing data happen to be in the outcome variable and the probability of the predictors fully depends on recorded variables, then the missing values can be modeled by adding another parameter having value of 0 or 1. The added parameter will have value 1 for recorded data and 0 for missing data. For example, in student retention datasets, one data element is the accumulated GPA. If the value of the current GPA is missing, the rule allows a substitution from the previous quarter's GPA. If there were no recorded GPAs in the previous quarters, the accumulative GPA value would be set at zero.

2.2.4 Hot-deck imputation

Random imputation of a single variable is needed when more than a small fraction of data has missing values. Random imputation involves replacing the missing values with randomly generated values. Randomly generated values can come from the available values in the current dataset, also known as hot-deck imputation, or from similar

datasets containing matching variables, also known as cold-deck imputation. In random imputation, the estimation of suitable values for replacing the missing values is generated based on the available data. Due to the unavailability of similar datasets, this paper uses hot-deck imputation for the random imputation method.

According to McKnight et al. (2007), there are different strategies for hot-deck imputation. The first strategy is simple random imputation, by imputing the missing value of any missing variable with randomized values based on the available data. If the missing data is MCAR, then there is no method for defining the missing value. Thus, if the observable values occur in the same proportion as the sampled population, supplying missing values from this predicted population will not introduce bias to the variable. This approach is considered to be a good starting point for preliminary data analysis. The strategy is hot-deck within adjustment cells—that is, blocking the relevant covariates and imputing the missing data based on the randomly generated values of the available data. Yet another approach uses the nearest neighbor's value in order to replace the data. This method imputes the missing value with the closest criteria from the available data. For example, if the ethnicity of a participant is missing from a group with a similar ethnicity, the missing values will be imputed with the particular ethnicity in that group.

Matching and hot-deck imputation determine each missing unit (y) with a value from similar value of predictors (x) in the observed data. Matching can become challenging when the matching vectors need to be built with a small amount of available data. To solve this problem, random imputation of the five closest resolved cases or other available information can be used. One can also predict the missing values based on several other variables that are fully observed; thus, the predicted data can be matched and imputed to the datasets. The most common problem that arises from this method is that it underestimates the standard errors due to the decreased variability. This is caused by the missing data being imputed by values that already exist in the dataset.

According to Seastorm et al. (2002), hot-deck imputation preserves the distribution of the original data and increases the variance compared to mean imputation. Consequently, according to Mundform and Whitcomb (1998), the estimate of the prediction accuracy would be too dependent to the randomly selected value, due to its variation from one selection value to another. In their research, Mundform and Whitcomb were running 1000 repetitions for hot-deck imputation and took the average value of the 1000 results of each 99 entries to obtain the value used in his research.

2.2.5 Multiple imputation

Multiple imputation is a method of supplying multiple values for a missing value. By utilizing Markov Chain Monte Carlo (MCMC) simulation, multiple values can be generated (Mcknight et al. 2007). MCMC is using computer simulation of Markov chains where the posterior distribution of the statistical inference problem is the asymptotic (Muller 2003). The imputed values can be analyzed for mean and variation. These statistics can then be used to derive expected values and associated confidence intervals.

Two common methods of multiple imputation using MCMC-method-derived Bayesian estimated values are routine multivariate imputation and iterative regression imputation. In routine multivariate imputation, a fitted multivariate model is built using all the variables containing missing values. The predictors (x) and the outcome (y) are considered vectors. This method has some difficulties, one of them being that much effort is required to set up a reasonable multivariate regression model. The *t*-distribution or multivariate normal distribution is commonly used for continuous outcomes, while the multinomial distribution is used for discrete outcomes. According to Rubin (1978), the efficiency of an estimate (relative efficiency in %) based on *m* imputations is shown by:

$$RE = \left(1 + \frac{\gamma}{m}\right)^{-1} \quad (1)$$

γ = rate of missing information for estimated quantity.

The multiple imputation efficiencies for various values of *m* and γ are shown in Table 2.

Table 2: Rubin's Multiple Imputation Efficiency

	γ				
m	0.1	0.3	0.5	0.7	0.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

3. Methodology

3.1 Data set

The original data were retrieved in from four resources: student applications, the Student Information System (SIS), the students' financial aid records, and students' involvement survey carried out by the Office of Institutional Research in 2006. Student applications contain students' demographics, high school information, and standardized test scores. SIS, the second source, is a software program that manages student information. It provides students' registration information from the past until present academic information. The third source is the information of students' financial aid records. These variables were entered into university databases through student's Free Application for Federal Student Aid, or FAFSA. The last source was a student involvement survey conducted by the Office of Institutional Research. This survey was conducted at the end of Winter quarter and provides information on students' attitudes and behaviors related to social and academic involvement in their first year.

3.2 Data Cleaning

The dataset from institutional research described above was received as one dataset that combined the information from all four sources. Yet, the dataset indicated a large amount of missing data. Roth (2008) had taken several steps to clean the data by keeping the valid data in preparation for the data modeling in predicting student attrition. Her second step after data cleaning was data imputation. Several simple imputation methods, such as mean and zero imputation, were utilized. Due to the biases from the decreased number of entries in sample sizes that can be created from applying complete-case analysis, or the elimination of any entry with a missing data point, complete-case analysis was not utilized (Gelman and Hill 2007). In this research, data imputation techniques comparison will be the main focus in order to find out the best imputation technique to be utilized in creating the predicting model. Table 3 shows a summary of the variables containing missing data points and the number of students missing information in each category.

Table 3: Variables and Number of Students with Null Values (4061 Observations)

Variable	**# of Students**	**% of missing**
HS Size	829	20.41%
HS Percentile Rank	830	20.43%
HS GPA	128	3.15%
State	14	0.35%
County Code	375	9.24%
ACT Composite	488	12.02%
ACT Math	488	12.02%
SAT Total	1847	45.49%
Expected Family Contribution	1219	30.02%
Involvement Survey Variables (28 Variables)	815	20.07%

After the variables with missing values are identified, the number of student entries is analyzed. The analysis indicated a stopout behavior after comparing student enrollment statuses in winter, spring and sophomore fall quarters, which resulted in student entries reduction. "A stopout student is one that demonstrates non-permanent attrition behavior, or who drops out one quarter, only to return in a following quarter" (Roth 2008). Since the purpose of the model is permanent attrition, 22 stopout students were removed from the dataset, resulting in a dataset of 4,039 entries for the model.

3.3 Summary of Imputation Method Procedure

For the imputation methods comparison, five variables with the largest number of missing values from the original data were chosen. As can be seen in Table 2, Expected Family Contribution has the second highest rate of missing information with a total of 1219 missing values. However, this variable is considered as missing not at random because the information was missing from due to students not filling out a FAFSA. So this variable was excluded from this research. Another variable with high missing values that was excluded from the imputation method comparison is the Involvement Survey with 815 missing values. The involvement survey data was conducted at the end of Winter Quarter and this can be considered too late for any typical retention intervention, which usually takes place at the time of fall quarter pre-registration. Then, the rest of the variables with missing data were chosen based on data available from the beginning of the fall quarter, 2007. The five variables with missing data chosen for the research are: SAT Total (1847 missing values), High School Percentile Rank (830 missing values), High School Size (829 missing values), ACT Composite (488 missing values) and ACT Math (488 missing values).

For each of these variables, a new dataset of 10 replications with similar distributional characteristics was randomly generated. For each of these sets, values were removed according to MCAR. For each variable, the missing values were imputed with one of the five different methods. These five methods are: mean imputation, median imputation, zero value imputation, hot-deck imputation and multiple imputation. After imputing the missing values, each set was analyzed for accuracy in the imputed values. First, each imputed value was compared to the removed value. The mean and variance of each imputed variable set was compared to the original to determine if the mean or variance has been affected. Based on the comparison results, for each variable in the student retention dataset, domain knowledge is used to classify the reasons for the missing values and the best imputation method will be implemented. Finally, a prediction model of student retention was built. This model then is compared with the Roth's model (2008) for prediction accuracy.

4. Imputation Method Comparison

A new dataset with similar distributional characteristics was generated randomly with 10 replications. Before a dataset can be generated, the distributional characteristics of the five variables need to be determined. MINITAB was used to fit in the distribution characteristics of the variables; this can be seen in Table 4.

Table 4 : Original Dataset Distribution Analysis

Variable	MINITAB distribution
SAT	Lognormal
ACTC	Lognormal
ACTM	Beta
HsSIze	3 parameter Lognormal
HSRank	3 parameter Weibull

4.1 SAT

Table 5 shows a summary of variable SAT imputation for the 10 sets. In mean comparison, as expected, mean imputation has the lowest total percentage difference between the initial values. The second lowest total percentage difference is multiple imputation, and then hot-deck imputation with 0.205% and 0.265% of difference. Although mean imputation has the lowest value of total mean average difference, it has a large standard deviation percentage difference. Multiple imputation and hot-deck imputation are superior to mean and median imputation for preserving the standard deviation values. Multiple imputation has the lowest percentage difference for standard deviation values with a 1.61% average of difference.

In Figure 2, the mean for each imputation is different due the randomness of the random number generated dataset. With mean imputation, small differences between the imputed mean and the complete dataset are expected. This is indicated by the low values that mean imputation has throughout the generated dataset, except in Random Number generated dataset 3 or RN3 that has a high value of mean compared to other imputed random number generated datasets. Yet, the overall performance of mean imputation, compared to the dataset or initial dataset mean value, is superior to other imputation methods. Mean imputation and median imputation seem to have similar behavior in standard deviation comparison.

Table 5 : Mean and Standard Deviation Result for SAT

Mean												Standard Deviation											
Dataset	(initial value)	mean imputation		median		zero		hot deck imputation		MI		Dataset	(initial value)	mean imputation		median		zero		hot deck imputation		MI	
		mean	delta %	mean	delta %	mean	delta %	mean	delta %	mean	delta %			StDev	delta %	StDev	delta %	StDev	delta %	StDev	delta %	StDev	delta %
RN1	1093.3	1092.3	0.091%	1087.6	0.521%	595.53	45.529%	1096.7	0.311%	1094.6	0.119%	RN1	144.6	108.5	24.97%	108.6	24.90%	554.71	283.62%	148.1	2.42%	146.0941	1.03%
RN2	1100.8	1102.9	0.191%	1099.7	0.100%	601.27	45.379%	1101.1	0.027%	1100.04	0.069%	RN2	149.3	109.4	26.72%	109.5	26.66%	560.04	275.11%	146	2.21%	147.2277	1.39%
RN3	1095.4	1100.5	0.466%	1095.3	0.009%	599.99	45.226%	1098.5	0.283%	1099	0.329%	RN3	142.7	107.2	24.88%	107.4	24.74%	558.47	291.36%	144.2	1.05%	146.1817	2.44%
RN4	1096.2	1096.7	0.046%	1092.7	0.319%	597.88	45.459%	1096.5	0.027%	1096.88	0.062%	RN4	144	104.7	27.29%	104.8	27.22%	556.1	286.18%	144.7	0.49%	144.2399	0.17%
RN5	1095.5	1097.2	0.155%	1091.6	0.356%	598.17	45.398%	1084	1.050%	1085.88	0.878%	RN5	146.1	109.9	24.78%	110.1	24.64%	557.35	281.49%	130.8	10.47%	135.0646	7.55%
RN6	1091.6	1094.6	0.275%	1085.5	0.559%	596.78	45.330%	1094.6	0.275%	1094.62	0.277%	RN6	155	117.6	24.13%	118.1	23.81%	557.69	259.80%	157.7	1.74%	157.4017	1.55%
RN7	1099	1099.3	0.027%	1096	0.273%	599.35	45.464%	1096	0.273%	1098.56	0.040%	RN7	145.1	108.6	25.16%	108.7	25.09%	558.17	284.68%	147.3	1.52%	146.2845	0.82%
RN8	1095.2	1096.7	0.137%	1093.2	0.183%	597.93	45.404%	1099.1	0.356%	1096.98	0.163%	RN8	147.8	110.4	25.30%	110.5	25.24%	557.24	277.02%	149.5	1.15%	147.6167	0.12%
RN9	1098	1098	0.000%	1094.8	0.291%	598.6	45.483%	1098.5	0.046%	1097.36	0.058%	RN9	147.9	109.8	25.76%	109.8	25.76%	557.71	277.09%	146.2	1.15%	147.4026	0.34%
RN10	1095.8	1095.9	0.009%	1090.9	0.447%	597.45	45.478%	1095.8	0.000%	1096.42	0.057%	RN10	146.4	109.3	25.34%	109.4	25.27%	556.59	280.18%	145.5	0.61%	147.4545	0.72%
average			**0.140%**		0.306%		45.415%		0.265%		0.205%	average			25.43%		25.33%		279.65%		2.28%		**1.61%**

4.2 ACTC

Tables 6 show a summary of variable ACTC imputation results for 10 random number generated datasets. The multiple imputation method has the lowest average difference in standard deviation compared to other imputation methods. In the mean imputation result, as expected, the mean imputation method has the lowest percentage difference. Yet, the difference between the total mean average for multiple imputation and hot-deck imputation is only 0.011%. Multiple imputation has the second lowest total average mean difference, 0.084%. And the third lowest total average mean difference is hot-deck imputation with a total difference of 0.093%. Multiple imputation still outperformed the other imputation methods in standard deviation difference by having the lowest total standard deviation difference (0.41%). The mean for each imputation is different due to the randomness of the random number generated dataset. It can be seen that the median imputation has the highest mean difference compared to the other methods. As can be seen in Figure 3, the standard deviations for all imputations tend to show similar behavior. Multiple imputation and hot-deck imputation seems to have very low standard deviations compared to mean and median imputation. Seven out of ten results of the imputation, multiple imputation has the smallest value of standard deviation compared to hot-deck, mean, and median imputations.

Table 6 : Mean and Standard Deviation Result for ACTC

Mean										Standard Deviation									
Dataset	(initial value)	mean imputation		median		hot deck imputation		MI		Dataset	(initial value)	mean imputation		median		hot deck imputation		MI	
		mean	delta %	mean	delta %	mean	delta %	mean	delta %			StDev	delta %	StDev	delta %	StDev	delta %	StDev	delta %
RN1	23.47	23.431	0.166%	23.379	0.388%	23.443	0.115%	23.4256	0.189%	RN1	3.486	3.268	6.25%	3.271	6.17%	3.49	0.11%	3.477889	0.23%
RN2	23.437	23.452	0.064%	23.397	0.171%	23.422	0.064%	23.4548	0.076%	RN2	3.451	3.234	6.29%	3.238	6.17%	3.44	0.32%	3.449418	0.05%
RN3	23.396	23.353	0.184%	23.311	0.363%	23.369	0.115%	23.368	0.120%	RN3	3.565	3.333	6.51%	3.335	6.45%	3.528	1.04%	3.54766	0.49%
RN4	23.549	23.529	0.085%	23.465	0.357%	23.502	0.200%	23.5254	0.100%	RN4	3.579	3.358	6.17%	3.362	6.06%	3.555	0.67%	3.575497	0.10%
RN5	23.354	23.355	0.004%	23.312	0.180%	23.371	0.073%	23.3604	0.027%	RN5	3.443	3.207	6.85%	3.209	6.80%	3.415	0.81%	3.422241	0.60%
RN6	23.442	23.456	0.060%	23.401	0.175%	23.424	0.077%	23.453	0.047%	RN6	3.48	3.259	6.35%	3.262	6.26%	3.46	0.57%	3.460935	0.55%
RN7	23.474	23.469	0.021%	23.413	0.260%	23.471	0.013%	23.4592	0.063%	RN7	3.498	3.292	5.89%	3.296	5.77%	3.491	0.20%	3.502247	0.12%
RN8	23.467	23.448	0.081%	23.394	0.311%	23.459	0.034%	23.4404	0.113%	RN8	3.577	3.344	6.51%	3.347	6.43%	3.56	0.48%	3.54803	0.81%
RN9	23.527	23.528	0.004%	23.464	0.268%	23.488	0.166%	23.5216	0.023%	RN9	3.554	3.317	6.67%	3.321	6.56%	3.517	1.04%	3.526938	0.76%
RN10	23.529	23.515	0.060%	23.453	0.323%	23.512	0.072%	23.511	0.077%	RN10	3.528	3.326	5.73%	3.33	5.61%	3.521	0.20%	3.540803	0.36%
average			**0.073%**		0.279%		0.093%		0.084%	average			6.32%		6.23%		0.54%		**0.41%**

Figure 2: ACTC Imputation Mean and Standard Deviation Comparison Results

4.3 ACTM

Tables 7 show summaries of variable ACTM imputation results for ten random number generated datasets. Compared to ACTC, multiple imputation method has the lowest average difference in standard deviation compared to other imputation methods. Again, mean imputation, as expected, has the lowest percentage difference in mean imputation comparison results. Yet, the hot-deck imputation method has the same low percentage difference as the mean imputation method, which is 0.0568%. Multiple imputation has the third lowest total average mean difference, 0.082%. Multiple imputation still outperformed the other imputation methods in standard deviation difference by having the lowest total standard deviation difference (0.367%).

Table 7 : Mean and Standard Deviation Result for ACTM

Mean										Standard Deviation									
Dataset	(initial value)	mean imputation		median		hot deck imputation		MI		Dataset	(initial value)	mean imputation		median		hot deck imputation		MI	
		mean	delta %	mean	delta %	mean	delta %	mean	delta %			StDev	delta %	StDev	delta %	StDev	delta %	StDev	delta %
RN1	22.924	22.922	0.009%	22.931	0.031%	22.923	0.004%	22.9046	0.085%	RN1	4.035	3.772	6.518%	3.772	6.518%	4.028	0.173%	4.02225	0.316%
RN2	22.898	22.923	0.109%	22.932	0.148%	22.911	0.057%	22.911	0.057%	RN2	4.059	3.796	6.479%	3.796	6.479%	4.035	0.591%	4.048428	0.260%
RN3	22.81	22.802	0.035%	22.826	0.070%	22.81	0.000%	22.8052	0.021%	RN3	4.012	3.759	6.306%	3.76	6.281%	4.004	0.199%	4.014369	0.059%
RN4	23.568	23.569	0.004%	23.501	0.284%	23.532	0.153%	23.567	0.004%	RN4	3.826	3.547	7.292%	3.552	7.162%	3.742	2.196%	3.774814	1.338%
RN5	22.86	22.885	0.109%	22.899	0.171%	22.869	0.039%	22.909	0.214%	RN5	4.039	3.779	6.437%	3.779	6.437%	4.014	0.619%	4.039639	0.016%
RN6	22.867	22.866	0.004%	22.882	0.066%	22.874	0.031%	22.8638	0.014%	RN6	4.036	3.802	5.798%	3.803	5.773%	4.054	0.446%	4.04164	0.140%
RN7	22.909	22.899	0.044%	22.911	0.009%	22.927	0.079%	22.879	0.131%	RN7	4.052	3.808	6.022%	3.808	6.022%	4.054	0.049%	4.054993	0.074%
RN8	22.942	22.95	0.035%	22.956	0.061%	22.948	0.026%	22.9306	0.050%	RN8	3.992	3.765	5.686%	3.765	5.686%	4.019	0.676%	4.014534	0.564%
RN9	22.881	22.839	0.184%	22.859	0.096%	22.846	0.153%	22.8344	0.204%	RN9	4.041	3.786	6.310%	3.787	6.286%	4.038	0.074%	4.033583	0.184%
RN10	22.82	22.828	0.035%	22.848	0.123%	22.827	0.031%	22.8296	0.042%	RN10	4.071	3.789	6.927%	3.79	6.902%	4.026	1.105%	4.041712	0.719%
average			**0.0568%**		0.106%		0.0572%		0.082%	average			6.378%		6.355%		0.613%		**0.367%**

4.4 High School Size

In table 8, High School Size imputation for ten random number generated datasets are summarized. In the table below, it can be seen that the multiple imputation method has the lowest average difference in mean and standard deviation compared to other imputation methods. Multiple imputation outperformed the other imputation methods in standard deviation differences by having the lowest total standard deviation difference (0.41%). The total average mean of the multiple imputation method is 0.24%, which is also the lowest among the other imputation methods. Unlike the previous variables results, the variable High School Size has more than 10% difference in total average difference in mean compared to the hot-deck imputation and mean imputation methods. Median imputation shows the highest mean among the other imputation methods except in one dataset, RN3. It can be seen that median imputation has a higher mean difference than the other imputation methods. The performance of multiple imputation has done better than the other imputation methods due the lower value of mean difference. Again, in seven out of ten sets, the mean of the multiple imputation method has lower values than do hot-deck imputation, mean imputation and median imputation. As can be seen in Figure 5, the standard deviation for each imputation tends to show similar behavior. Multiple imputation and hot-deck imputation seem to have very small variances and standard deviations compared to mean and median imputation. It can also be seen that multiple imputation and hot-deck imputation have slight differences, except in RN6 and RN7, where hot-deck imputation has higher variance difference than multiple imputation.

Table 8 : Mean and Standard Deviation Result for High School Size

Mean										Standard Deviation									
Dataset	(initial value)	mean imputation		median		hot deck imputation		MI		Dataset	(initial value)	mean imputation		median		hot deck imputation		MI	
		mean	delta %	mean	delta %	mean	delta %	mean	delta %			StDev	delta %	StDev	delta %	StDev	delta %	StDev	delta %
RN1	275.99	275.13	0.312%	272.15	1.391%	276.77	0.283%	276.398	0.148%	RN1	140.97	126.25	10.44%	126.4	10.34%	142.86	1.34%	141.990	0.72%
RN2	281.79	282.93	0.405%	279.51	0.809%	282.1	0.110%	281.974	0.065%	RN2	145.59	131.05	9.99%	131.2	9.88%	146.48	0.61%	146.108	0.36%
RN3	280.72	282.57	0.659%	280.1	0.221%	281.52	0.285%	281.956	0.440%	RN3	143.15	128.9	9.95%	129	9.88%	143.65	0.35%	144.212	0.74%
RN4	280.15	279.53	0.221%	276.32	1.367%	280.82	0.239%	279.244	0.323%	RN4	143.76	127.55	11.28%	127.7	11.17%	144.23	0.33%	142.849	0.63%
RN5	278.78	279.23	0.161%	271.63	2.565%	280.48	0.610%	280.23	0.520%	RN5	171.32	153.16	10.60%	153.9	10.17%	170.5	0.48%	171.7198	0.23%
RN6	276.9	275.09	0.654%	271.98	1.777%	274.73	0.784%	276.326	0.207%	RN6	141.65	124.79	11.90%	124.94	11.80%	138.79	2.02%	140.915	0.52%
RN7	277.01	276.03	0.354%	273.42	1.296%	275.35	0.599%	276.922	0.032%	RN7	141.22	125.37	11.22%	125.46	11.16%	139.32	1.35%	141.492	0.19%
RN8	277.7	276.94	0.274%	274.09	1.300%	278.37	0.241%	277.924	0.081%	RN8	140.46	124.77	11.17%	124.92	11.06%	141.67	0.86%	140.635	0.12%
RN9	276.14	275.44	0.253%	271.98	1.506%	276.4	0.094%	276.270	0.047%	RN9	142.05	127.48	10.26%	127.67	10.12%	143.29	0.87%	142.873	0.58%
RN10	281.27	279.57	0.604%	276.38	1.739%	279.43	0.654%	279.858	0.502%	RN10	142.49	127.12	10.79%	127.33	10.64%	142.02	0.33%	142.556	0.05%
average			0.39%		1.40%		0.39%		**0.24%**	average			10.76%		10.62%		0.85%		**0.41%**

4.5 High School Rank

In the table 9, it can be seen that the multiple imputation method has the lowest average difference in mean compared to other imputation methods. Surprisingly, unlike the previous variable comparison results, multiple imputation outperformed the other imputation methods only in mean difference, yet, the lowest total standard deviation difference (0.57%) was performed by hot-deck imputation. The mean for each imputation is different due to the randomness from the random number generated dataset. For variable High School Rank, the trend seems inconsistent. Yet, median imputation still holds the highest mean average among the other imputation methods. Six out of ten imputations results with the high values of mean difference for median imputation method. Hot-deck imputation has the three highest values of mean difference in RN6, RN7 and RN9. This leads to high total average of mean difference for hot-deck imputation. However, the total average percentage difference between multiple imputation and hot-deck imputation is only 0.55%. As can be seen in Figure 6, the standard deviation for each imputation tends to show similar behavior. Multiple imputation and hot-deck imputation seem to have a small variance and standard deviation compared to mean and median imputation. For High School Rank, hot-deck imputation has a lower difference value compared to the multiple imputation method.

Table 9 : Mean and Standard Deviation Result for High School Rank

Mean										Standard Deviation									
Dataset	(initial value)	mean imputation		median		hot deck imputation		MI		Dataset	(initial value)	mean imputation		median		hot deck imputation		MI	
		mean	delta %	mean	delta %	mean	delta %	mean	delta %			StDev	delta %	StDev	delta %	StDev	delta %	StDev	delta %
RN1	68.385	68.138	0.361%	68.208	0.259%	68.391	0.009%	68.3290	0.082%	RN1	20.011	17.746	11.32%	17.744	11.33%	19.757	1.27%	19.81229	0.99%
RN2	68.453	68.594	0.206%	68.771	0.465%	68.5	0.069%	68.5378	0.124%	RN2	19.667	17.572	10.65%	17.578	10.62%	19.623	0.22%	19.68115	0.07%
RN3	68.776	68.713	0.092%	68.937	0.234%	68.782	0.009%	68.7264	0.072%	RN3	19.621	17.566	10.47%	17.566	10.47%	19.617	0.02%	19.60684	0.07%
RN4	68.543	68.553	0.015%	68.804	0.381%	68.589	0.067%	68.3862	0.229%	RN4	19.611	17.444	11.05%	17.454	11.00%	19.651	0.20%	19.40444	1.05%
RN5	68.55	68.777	0.331%	68.922	0.543%	68.437	0.165%	68.5942	0.064%	RN5	20.243	18.142	10.38%	18.148	10.35%	20.256	0.06%	20.1091	0.66%
RN6	68.174	67.861	0.459%	67.974	0.293%	67.831	0.503%	68.0838	0.132%	RN6	19.856	17.706	10.83%	17.706	10.83%	19.863	0.04%	19.71822	0.69%
RN7	67.946	67.964	0.026%	68.026	0.118%	67.638	0.453%	67.8606	0.126%	RN7	19.894	17.738	10.84%	17.741	10.82%	19.864	0.15%	19.77367	0.60%
RN8	68.468	68.576	0.158%	68.894	0.622%	68.368	0.146%	68.2856	0.266%	RN8	19.778	17.55	11.27%	17.561	11.21%	19.526	1.27%	19.57847	1.01%
RN9	67.978	67.708	0.397%	68.077	0.146%	67.693	0.419%	67.7736	0.301%	RN9	19.775	17.526	11.37%	17.536	11.32%	19.559	1.09%	19.60874	0.84%
RN10	68.339	68.379	0.059%	68.7	0.528%	68.498	0.233%	68.4236	0.124%	RN10	19.728	17.486	11.36%	17.495	11.32%	19.456	1.38%	19.48139	1.25%
average			0.210%		0.359%		0.207%		**0.152%**	average			10.95%		10.93%		**0.57%**		0.73%

The accuracy results show that zero imputation is the poorest method. According to the RMSE results, mean imputation and median imputation performed with better accuracy than hot-deck and multiple imputation. Although mean and median imputation tend to center the distribution and decrease the variance and standard deviation, they

still have a better performance for accuracy. As for hot-deck and multiple imputation, they have lower variance, yet the RMSE results show that they are less accurate than mean and median imputation.

5. Prediction Model

From Table 10, it can be seen that an alternating decision tree (ADTree) model using the winter 2007 dataset was able to predict a student's retention status in fall of 2007 with 82.84 % overall accuracy. The overall accuracy had decreased from the previous model by 0.08%. There were 19,075 retention predictions made, and 82.86% of them were accurate. Attrition was predicted for just 15 student entries and the predictions were accurate 66.6% of the time. Yet, this prediction result cannot be considered useful, since it only predicts 15 out of a total 19,090 entries for attrition.

Table 10: Predicted Fall Enrollment from Winter Alternating Decision Tree vs. Actual Fall Enrollment

		Prediction			Accuracy	
		retention	attrition	Total	Retention	82.86%
Actual	retention	15805	5	15810	Attrition	66.67%
	attrition	3270	10	3280	Overall	82.84%
Total		19075	15	19090		

The ADTree result shows no significant difference with Roth's result for winter alternating decision tree in predicting fall enrollment. WEKA and MINITAB were utilized to derive logistic regression and linear regression models from the imputed dataset. WEKA was utilized for the logistic regression with the total imputed dataset with 19,090 student entries, resulting 83.47% overall accuracy (see table 11). The result shows that the prediction with the logistic regression has increased the overall accuracy to 83.47% from Roth's model, which was only 82.74% overall accuracy. A total of 15,810 retention prediction was made with 98.12% accuracy. Attrition predictions were 3,280, and those predictions were accurate 12.87% of the time. Table 12 shows that a 3,209 retention prediction was made, and 97.23% of the time it was accurate. For the attrition prediction, the accuracy rate was 3.94%, with a total of 609 of prediction.

Table 11: Predicted Fall Enrollment from Winter Logistic Regression vs. Actual Fall Enrollment

		Prediction			Accuracy	
		retention	attrition	Total	Retention	98.12%
Actual	retention	15513	2858	18371	Attrition	12.87%
	attrition	297	422	719	Overall	83.47%
Total		15810	3280	19090		

Table 12: Predicted Fall Enrollment from Winter Linear Regression vs. Actual Fall Enrollment for Individual Cases

		Prediction			Accuracy	
		retention	attrition	Total	Retention	97.23%
Actual	retention	3120	585	3705	Attrition	3.94%
	attrition	89	24	113	Overall	82.35%
Total		3209	609	3818		

6. Conclusion and future research

From this research, it can be concluded that multiple and hot-deck imputations perform poorly in the accuracy comparison test, yet can still slightly increase the prediction accuracy rate. The characteristic of missing data is an important factor in data analysis. Overall, it can be concluded that multiple imputation and hot-deck imputation are not useful for improving the attrition power prediction. But, they can be considered useful to improve the retention prediction rate, which is not useful for the attrition prediction model due to the high expectation of students who are least likely to drop out. It can also be concluded that for the purpose of building a prediction model, mean imputation can be utilized for its simplicity and easy implementation. This research has addressed some of the problems that Roth (2008) found while creating the student prediction model. The model was imputed with what was considered an appropriate method for each of the variables, and the result shows slight improvement. The

sample size was also increased due to the multiple imputation method, leading to a better prediction rate. Recommendation for future endeavor would be using alternative imputation technique such as maximum likelihood multiple imputation and hot-deck imputation combined.

References

Cohen, J., Cohen, P., West, S., and Aiken, L. *Applied multiple regression/correlation analysis for the behavioral sciences*, 3rd Edition, Lawrence Erlbaum Associates, New Jersey, 2003.

Gelman, A., and Hill, J. *Data analysis using regression and mulitlevel/hierarchical models*. Cambridge University Press, New York, 2007.

Mcknight, P. E., Mcknight, K. M., Sidana, S., and Fiqueredo, A. J. *Missing data: A gentle introduction*. New York: The Guilford Pres, 2007.

Muller, P. Monte carlo methods and bayesian computation: MCMC, Available: http://odin.mdacc.tmc.edu/~pm/class/422/mcmc-tutorial.pdf, August 10, 2009.

Mundform, D. J., and Whitcomb, A. Imputing missing values: The effect on the accuracy of classification. American Educational Research Association Annual Meeting , pp. 2-12, 1998.

Ohio University Office of Institutional Research. Factors associated with first-year student attrition and retention at Ohio University Athens campus. Available: http://www.ohiou.edu/intres/retention/RetenAthens.pdf, February 20, 2009.

Roth, S. A model to predict Ohio University student attrition from admission and involvement data. *Unpublished master's thesis*, Ohio University, 2008.

Rubin, D.B. Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 20-34, 1978.

Rubin, D. B. *Multiple imputation for nonresponse in surveys*, J. Wiley & Sons , New York , 1987.

Seastorm, M., Kaufman, S., and Lee, R. National Center for Education Statistics. Available: http://nces.ed.gov/statprog/2002/appendixb.asp. October 20, 2001.

Sirvanci, M. B. Critical issues for TQM implementation in higher education. *TQM Magazine*, vol. 16, pp. 382–386, 2004.

Tinto, V. Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, vol. 45,pp. 89–125, 1975.

Umbach, P. D. (Ed.) *Survey research: Emerging issues. New Directions for Institutional Research (Vol. 127)*, Jossey-Bass San Francisco, 2005.