2008

The Impact of Data Imputation Methodologies on Knowledge Discovery

Marvin Lane Brown
Cleveland State University

THE IMPACT OF DATA IMPUTATION METHODOLOGIES

ON KNOWLEDGE DISCOVERY

MARVIN L. BROWN

Bachelor of Business Administration
Shepherd College
December, 1980

Master of Business Administration
Morehead State University
May, 1984

Submitted in partial fulfillment of requirements for the degree

DOCTOR OF BUSINESS ADMINISTRATION
at the
CLEVELAND STATE UNIVERSITY

July, 2008

This dissertation has been approved

for the COLLEGE of BUSINESS ADMINISTRATION

and the College of Graduate Studies by

_____

Dissertation Chairperson, Dr. Chien-Hua (Mike) Lin

_____

Department & Date

_____

Dr. Adam Fadlalla

_____

Department & Date

_____

Dr. Walter O. Rom

_____

Department & Date

_____

Dr. John F. Kros

_____

Department & Date

_____

Dr. Marc P. Lynn

_____

Department & Date

THE IMPACT OF DATA IMPUTATION METHODOLOGIES

ON KNOWLEDGE DISCOVERY

MARVIN L. BROWN

ABSTRACT

The purpose of this research is to investigate the impact of Data Imputation Methodologies that are employed when a specific Data Mining algorithm is utilized within a KDD (Knowledge Discovery in Databases) process. This study will employ certain Knowledge Discovery processes that are widely accepted in both the academic and commercial worlds. Several Knowledge Discovery models will be developed utilizing secondary data containing known correct values. Tests will be conducted on the secondary data both before and after storing data instances with known results and then identifying imprecise data values. One of the integral stages in the accomplishment of successful Knowledge Discovery is the Data Mining phase. The actual Data Mining process deals significantly with prediction, estimation, classification, pattern recognition and the development of association rules. Neural Networks are the most commonly selected tools for Data Mining classification and prediction. Neural Networks employ various types of Transfer Functions when outputting data. The most commonly employed Transfer Function is the s-Sigmoid Function. Various Knowledge Discovery Models from various research and business disciplines were tested using this framework.

However, missing and inconsistent data has been pervasive problems in the history of data analysis since the origin of data collection. Due to advancements in the capacities of data storage and the proliferation of computer software, more historical data is being

collected and analyzed today than ever before. The issue of missing data must be addressed, since ignoring this problem can introduce bias into the models being evaluated and lead to inaccurate data mining conclusions. The objective of this research is to address the impact of Missing Data and Data Imputation on the Data Mining phase of Knowledge Discovery when Neural Networks are utilized when employing an s-Sigmoid Transfer function, and are confronted with Missing Data and Data Imputation methodologies.

TABLE OF CONTENTS

Page:

LIST OF TABLES

LIST OF FIGURES:

Abbreviations

ANN	Artificial Neural Networks
CART	Classification and Regression Trees
CHAID	Chi-squared Auto. Interaction Detector
CLASSIT	Extended Conceptual Clustering Model
CM	Confusion Matrix
CMAR	Completely Missing At Random
COBWEB	Conceptual Clustering Model
CRISP-PM	Cross Industry Standard Process For DM
CRM	Customer Relations Management
DASL	Data and Story Library
DBMS	Database Management System
DM	Data Mining
DT	Decision Trees
EM	Expectation Maximization
GA	Genetic Algorithms
HSD	Honestly Significant Difference
IDA	Intelligent Data Analyzer
KDD	Knowledge Discovery In Databases
MAE	Mean Absolute Error
MAR	Missing At Random
MBA	Market Basket Analysis
NN	Neural Networks
NN	Nearest Neighbor
OLAP	On-Line Analytical Processing
RMS	Root Mean Square
SAS	Statistical Analysis System
SL	Supervised Learning
SPSS	Statistical Package for the Social Sciences
SQL	Structured Query Language
SSF	S-Sigmoid Function
TQM	Total quality Management
UL	Unsupervised Learning

CHAPTER I

INTRODUCTION

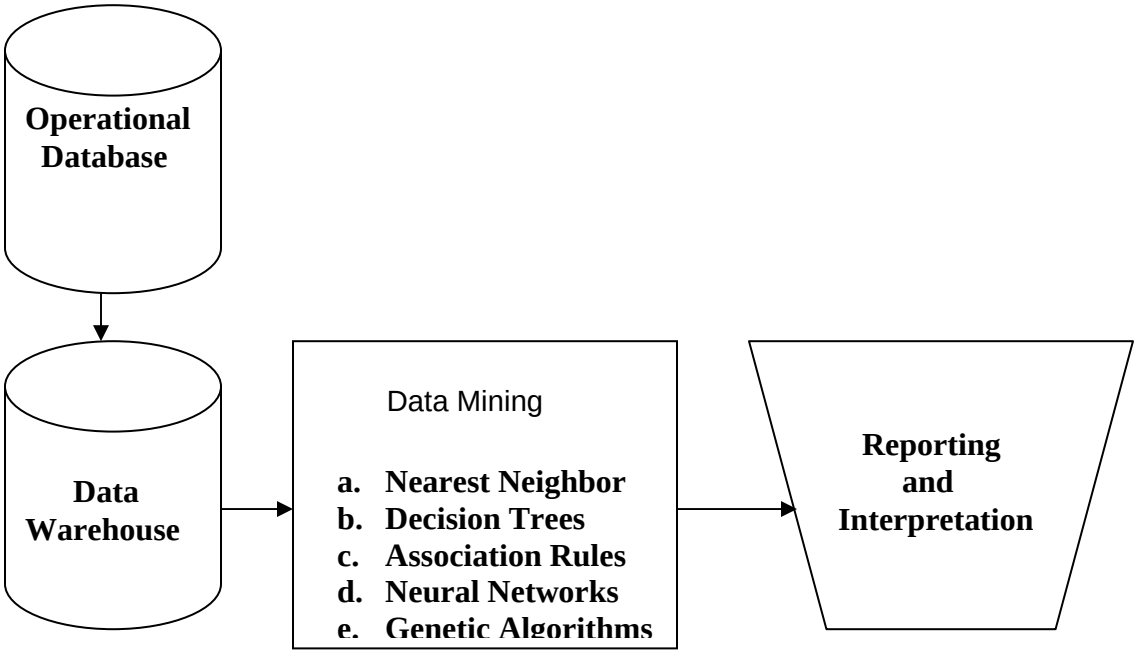
1.1 Background

As hardware capacity continues to grow at an increasingly affordable rate, the ability to house large volumes of case frequencies involving historical data in VLDB's (very large databases) is now possible. Software engaged to process these large volume data sets of increased case frequencies has recently been developed to aid the knowledge worker in an attempt to discover previously unknown patterns in that data (Fayyad 2001).

When Data Mining was first introduced, it was defined as "the application of Pattern Recognition Algorithms to large volumes of data to discover hidden patterns in the data not previously known" (Fayyad, 1996).Data mining is based upon searching the concatenation of multiple databases that usually contain some amount of missing data along with a variable percentage of inaccurate data, pollution, outliers and noise. The overall goal of the data mining process deals significantly with prediction, estimation, classification, pattern recognition and the development of association rules. Figure 1.1

illustrates the original data mining concept, while Figure 1.2 depicts the application of the scientific method to data mining.

Figure 1.1 Simple Data Mining



1.2 Knowledge Discovery In Databases

Since the advent of low-cost hardware and software to house and support high-volume historical data sets, a focus in the area of decision support has shifted to that of data warehousing and knowledge discovery. In addition, the technological advancements not only in data storage facilities and high-speed retrieval and processing mechanisms, but also advancements in pattern recognition software has given rise to new niches in areas of Knowledge Discovery. Furthermore, "old" data sets may now be utilized by modern applications (Loshin, 2004, p. 1).

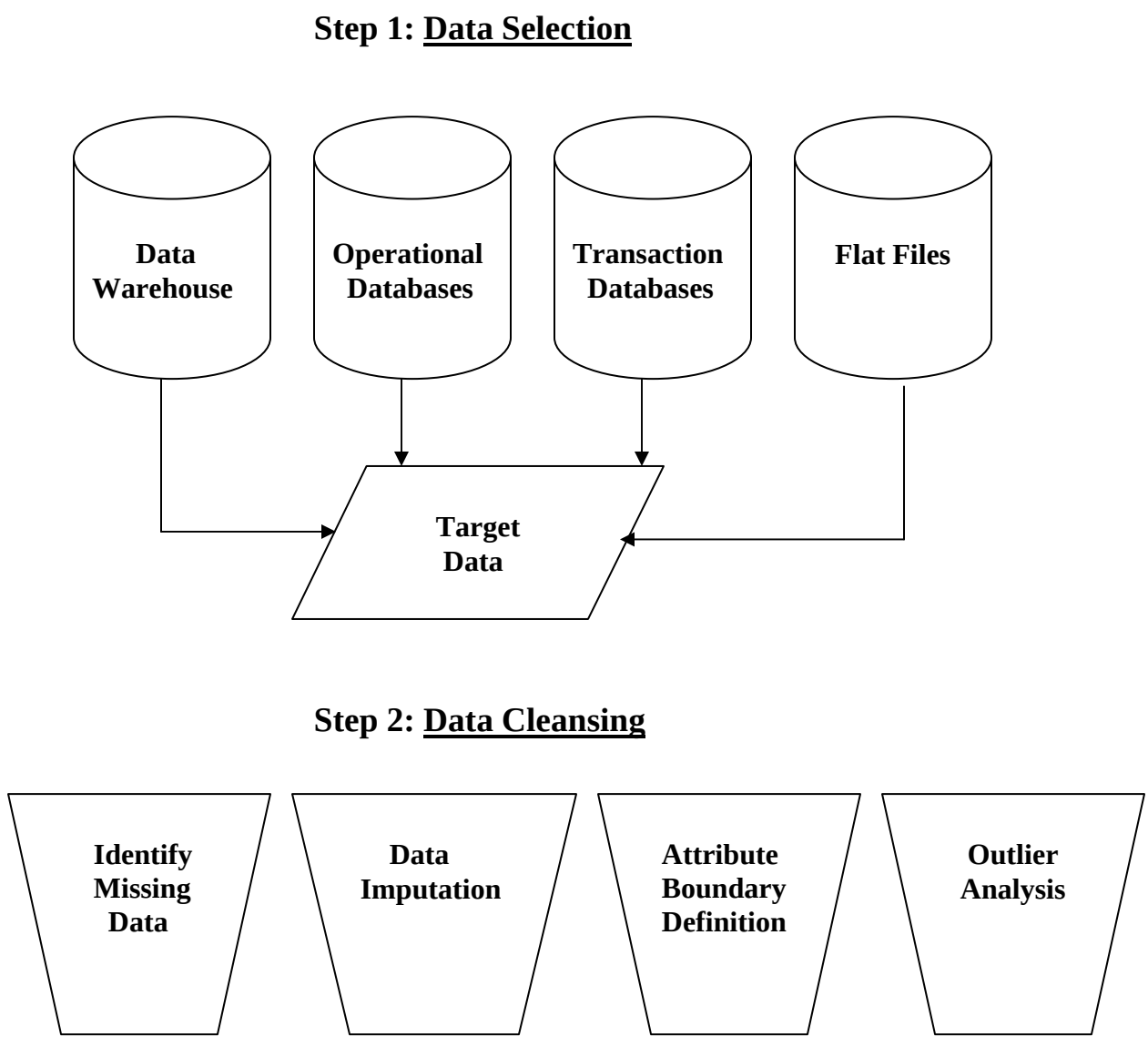
Initially described as simply “data mining”, Knowledge Discovery in Databases (KDD) is now described as applying the scientific method to data mining methodology (Roiger and Geatz, 2003, p. 150). As the application of Data Mining to large data sets became more widely practiced, the process was decomposed into a series of logical procedures. KDD is a process that can be utilized to identify and provide the use of previously undiscovered patterns and relationships within a large data warehouse containing historical data for an organization. Although various authors and researchers have decomposed the KDD process into a variable number of categories (some as many as fourteen steps or procedures), the following are a widely accepted series of procedures has been identified for use in the Knowledge Discovery process:

- Data Selection
- Data Cleansing
- Data Enrichment
- Data Coding
- Data Mining
- Reporting and Interpretation

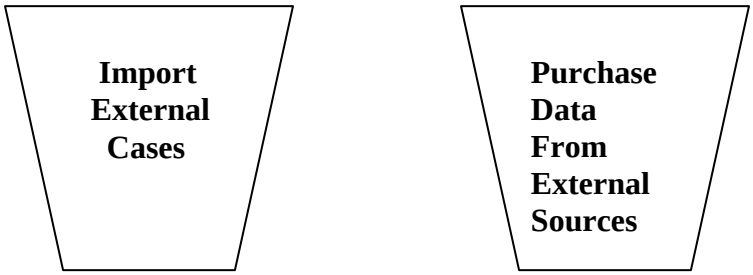
A growing number of knowledge engineers have added two additional steps to the aforementioned procedure to increase the efficiency of the entire process. A Goal Identification (Knowledge Requirements) Phase may be added to the front end of the process, and an Action Phase appended to the final step (Miller, 2000).

The KDD process has been successfully implemented in the fight against terrorism, financial and stock market investments, credit card fraud, general law enforcement and CRM (Customer Relations Management). However, flaws in the utilization of KDD have resulted in misclassification of cases, misrepresentation of data classes, invalid clustering and inconsistent forecasting.

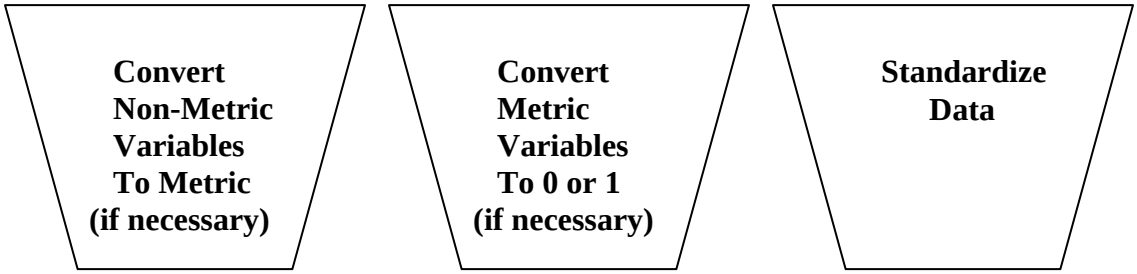
Figure 1.2 A 6-Step KDD Model



Step 3: Data Enrichment

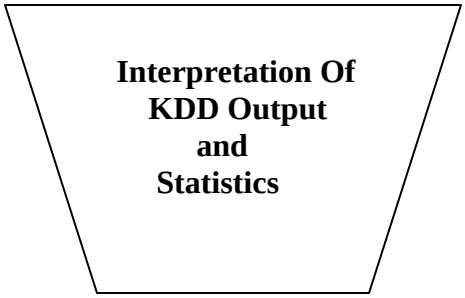


Step 4: Data Coding



Step 5: Data Mining

Step 6: Reporting



1.3 Data Mining

Data Mining is the term once used to describe the entire aforementioned KDD process. Otherwise stated, Data Mining was the KDD process.

While some researchers describe data mining as an “anything goes” process, various tools have been employed to give structure to the overall task. Query tools, statistical methods, data visualization techniques, and OLAP (On-Line Analytical Processing) software have all proven themselves beneficial to the task of Knowledge Discovery in Databases.

.

1.4 Data Mining Algorithms

Although many techniques have been developed and successfully implemented, five algorithms have been become widely accepted as standards in commercial data mining software packages:

- k-Nearest Neighbor
- Decision Trees
- Association Rules
- Neural Networks
- Genetic Algorithms

Each methodology has its own strengths and weaknesses that will be addressed in detail in Chapter 2. The algorithm should only be selected following an analysis of the type of data to be mined and the relationships within and the volume of the cases in question. The implementation of an algorithm to an inappropriate environment may result in improper data categorization, incorrect classification of cases and invalid test

conclusions (Miller and Han, 2001). A Neural Network may be chosen as the algorithm to be tested, due to its proven ability to adapt to changing environments and the inherent ability for the Artificial Neural Network Model to be refreshed and retrained with addition and more recent data instances. Various forms of Dependency Analysis, in which columns across rows are evaluated, may be employed to discover dependencies between data attributes. These dependencies may be altered and/or violated by data quality issues (Loshin, 2004).

1.5 Types of Missing Data

Data Quality issues and assessments have gained a great deal of notoriety in recent literature. A basic set of data quality dimensions include data accuracy, completeness, consistency, timeliness, interpretability and accessibility. Despite the recent trend in addressing these data quality issues, the development of new methodologies for data quality management has historically been given very low priority in IT development and operations.

Price WaterhouseCoopers reports that 88% of data integration projects fail due to damaged data, and that 33% of organizations either cancel or delay new IT start-ups due to poor data quality. According to the Gartner Group, poor data quality is the number one factor in the failure of CRM system failure and causes losses of $611 billion annually in the US alone. They also estimate that data typically degenerates at a rate of 2% per month, or approximately 25% annually.

Along with the globalization of business and information, mergers, acquisitions and universal dependence on business information, the exponential growth of data

storage into petabyte-sized (and larger) data warehouses creates even more data quality issues. The false assumption that data can be treated as a static, finite asset must be avoided. Missing, incomplete, incorrect, ambiguous, irrelevant, inconsistent, duplicate and aged data create erroneous data rules and business intelligence that lead to poor decision making and subsequent business failure. Although data quality issues have been recently addressed, it is still common business practice for businesses to tolerate poor data quality than to directly manage and/eliminate it in areas of decision Support (Cappiello, Francalanci and Pernici, 2004).

Traditionally, most stored data for Business Intelligence has been of the "hard" type. Hard data is inherently verifiable. However, increased use of "soft" data that has inherently unverifiable quality (such as projections of the future intentions of the competitors of a business entity) is yet another industry trend that must also be addressed (Ballou, Pazer, Tayi and Wang, 1998).

"The time has come to focus in the 'I' in IT" (Peter Drucker, 2004). In agreement, the Japanese data quality guru Kaoru Ishikawa affirmed that in order to "speak with the data", the data has to be accurate (Ishikawa, 2004). While there is no universally agreed upon metric for the measurement of information quality, the approaches addressed by Six Sigma and Cooperative Information Systems are helping to improve the overall efforts of user organizations and data quality issues.

Missing or inconsistent data has been a pervasive problem in data analysis since the origin of data collection. More historical data is being collected today due to the proliferation of computer software and the high capacity of storage media. The management of missing data in organizations has recently been addressed as more firms

implement large-scale enterprise resource planning systems (Vosburg and Kumar, 2001; Xu et al., 2002). The issue of missing data becomes an even more pervasive dilemma in the Knowledge Discovery process, in that as more data is collected, the higher the likelihood of missing data becomes. Data dependencies, data sparseness (especially within critical data clusters) and anomaly analysis are areas that are directly impacted by the issue of Missing Data (Loshin, 2004).

The following is a partial list of commonly encountered and accepted standard types of missing data:

- Data Missing At Random
- Data Missing Completely At Random
- Non-Ignorable Missing Data
- Outliers Treated As Missing Data

1.6 Methods of Addressing Missing Data

Methods for dealing with imprecise or missing data can be broken down into the following categories:

- Use Of Complete Data Only
- Deleting Selected Cases Or Variables
- Data Imputation
- Model-Based Approaches

These categories are based on the randomness of the missing data, and how the missing data is estimated and used for replacement.

According to Loshin (2004), some applications that may be profiled as types of Missing Data applications include:

- Anomaly Analysis
- Data Reverse Engineering
- Information Quality Discovery
- Data Model Integrity

1.7 Data Imputation Methods

Imputation methods are procedures resulting in the replacement of missing values by attributing them to other available data. A definition of imputation is as follows: "the process of estimating missing data of an observation based on valid values of other variables" (Hair et al., 1998). As Dempster and Rubin (Dempster and Rubin, 1983) commented, "imputation is a general and flexible method for handling missing-data problems, but is not without its pitfalls. Caution should be used when employing imputation methods as they can generate substantial biases between real and imputed data". Nonetheless, data imputation methods tend to be a popular method for addressing the issue of missing data.

Some of the most commonly used imputation methods include:

- Case Substitution
- Mean Substitution
- Hot Deck Imputation
- Cold Deck Imputation
- Regression Imputation
- Multiple Imputation

1.8 Scope of the Study

The scope of this study is to perform an experimental design method to explain the impact of imprecise and missing data on the process of Knowledge Discovery In Databases. More precisely, the Data Mining phase of the Knowledge Discovery process will be tested utilizing Neural Network software that employs the S-Sigmoid as its Transfer Function. Secondary data will be used for the experiment, utilizing data obtained from a collaborative project of the Center for Disease Control (CDC). The Behavioral Surveillance Branch (BSB) of the CDC has verified the data and made it available to researchers for statistical and analysis purposes. This data will be used in conjunction with software for Knowledge Discovery utilizing a Neural Network employing an S-Sigmoid Transfer Function. The network will be trained using the complete data set(s) with known values for the dependent variable(s). Using a verified SQL algorithm for randomization, various percentages of data elements within the data set(s) will be randomly identified and recorded, and then replaced with missing data. The model will

be retested using the same software and parameters. Missing values will then be imputed by utilizing several accepted imputation methods. The data set will be retested by the knowledge discovery process following each imputation method. The percentage of data missingness will be increased, retested and the results recorded. Percentages of missing data to be injected into the data sets under investigation will be set at 10%, 20%, 30%, 40%, 50%, 60% and 70% prior to imputation.

1.9 Experimental Model

Results from the proposed model will be evaluated using a series of Analysis of Variance (ANOVA) tests and two-tailed T-Tests to determine if significant differences exist between Knowledge Discovery outcomes whenever complete data is present, a certain percentage of data missingness exists and when missing data is replaced using various methods of data imputation. An overview of the experimental model is presented in Figure 3.

1.9.1 Contribution

The anticipated contribution of this experimental research is to better understand and employ various methods of data imputation in the environment of Knowledge Discovery in the presence of missing data when a specific type of software and associated algorithm are utilized. Expected benefits derived from the research include a better understanding of the impact that data missingness has on select types of Knowledge Discovery, the value of data imputation when missing data is confronted, and the identification of levels of data missingness that impact the success of data imputation.

Previous research in this area will be extended in this study to include a specific niche of popular commercial data mining software packages. Neural Networks that employ the S-Sigmoid Transfer Function. Contribution to the areas of missing data and the value of data imputation is also expected. The researcher also expects that the application of data imputation at various levels of data missingness can be identified.

Due to the lack of previous exploration in this specific area of Knowledge Discovery, this research provides an initial step for the examination of other data mining algorithms that are also confronted with various degrees of data pollution, noise, corruption and data missingness.

1.9.2 Organization of Dissertation

This dissertation proposal will be made up of five chapters. The first chapter will introduce the topic of Knowledge Discovery In Databases and the Data Mining Algorithms that may be utilized in the process. The topic of Missing Data is then addressed decomposed into various categories and discussed. The impact of Missing Data on the Data Mining phase of Knowledge Discovery is then addressed. “Methods of Data Imputation” is then introduced, as well as the reasons for their use. The implementation of various types of Data Imputation is then discussed and evaluated.

Chapter Two provides a review of both seminal and current literature relating to both the conceptual model as well as specific areas that impact the study.

Chapter Three presents the overall proposed research model and corresponding hypotheses to be tested.

Chapter Four discusses the overall experiment to be performed, the data gathered for testing, and the research methods to be employed in this study.

Chapter Five will present the analysis to be performed in the final evaluation of the experimentation performed.

Chapter Six will present the conclusions and contributions of this research and discuss future trends for research and industry.

CHAPTER II

LITERATURE REVIEW

Missing or inconsistent data has been a pervasive problem in data analysis since the origin of data collection. With the proliferation of information technology and the advent of increased storage space capacity, more and more data is being collected today than ever before. In turn, the impact of Missing Data becomes an even larger dilemma. An added complication is that as more data that is collected, the higher the likelihood of missing data becomes, and in turn the more likely one will need to address the problem of incomplete cases. During the last four decades, statisticians have attempted to address the impact of missing data on information technology. One objective of this research is to address the impact of missing data and its impact on data mining. A review of seminal and current literature will be conducted in the areas of Knowledge Discovery, Data Mining, Missing Data, Data Imputation and the Impact of Data Imputation on Data Mining Algorithms. The overall dissertation focuses on methods of addressing missing data and the impact that missing data has on the knowledge discovery process, depending on the data mining algorithm being utilized.

2.1 Knowledge Discovery in Databases (KDD)

Knowledge Discovery in Databases (KDD) is a new, multidisciplinary field that focuses on the overall process of information discovery in large volumes of warehoused data (Abramowicz and Zurada, 2001). KDD has been mostly utilized by artificial intelligence and machine learning researchers.

2.2 Data Mining

Data Mining (DM) involves searching through databases for correlations and/or other non-random patterns, while being utilized by statisticians, data analysts and the management information systems community. While the benefits derived from data mining when newly discovered patterns are interpreted correctly, disadvantages certainly exist.

2.3 Data Mining Disadvantages

Misinterpretation of discovered patterns can occur when correlations are skewed by poor data quality or missing data, or the use of too many variables in the study. Also, the system can be allowed to execute long enough to find evidence to support any or specific preconceptions that the researcher may have (Coy, 1997). Further, a high amount of knowledge concerning the application is required by the researcher, as well as the choice of the database/data warehouse under study (Sethi, 2003). Finally, it may be difficult to determine if a relationship or pattern has occurred at random or if it is indeed unique to a specific sample that has been taken (McQueen and Thorley, 1999).

While these disadvantages surely exist in the process of mining data, one distinct theme must be addressed: A model is only as good as the variables and data used to create it. The quality, freshness and accuracy of the data are imperative for successful data mining. The elimination of errors, removal of redundancies and filling in gaps that exist within the data (although tedious and time-consuming) are an integral phase of the overall process (Wang, 2003).

2.4 Data Mining Algorithms

2.4.1 k-Nearest Neighbor

The k-nearest neighbor concept is named so due to the fact that each data record is said to exist in a particular cluster or "neighborhood", and that the records that are closest to each other are referred to as neighbors. This method is used to predict the behavior of certain data elements (Adriaans and Zantinge, 1997).

The k-Nearest Neighbors to an observation are first identified. The *k* stands for a predetermined constant representing the number of neighbors that contain *no* missing data and will be considered in the analysis. According to Witten and Frank (2000), it is advised to keep the value for *k* small, say five, so that the impact of any noise present will be kept to a minimum. Since data sets with a large number of attributes or closely related cases may result in a high number of closely related neighborhoods, this algorithm is not recommended for large data sets (Adriaans and Zantinge, 1997).

2.4.2 Decision Trees

Decision Trees are analytical tools used to discover rules and relationships in data by systematically breaking down and subdividing information from a general view down to the detail level (Acharya and Mitra, 2003). Contained in this tree structure are branches that represent the outcomes of a particular test and leaf nodes that represent resulting classes or class distributions (Han and Kamber, 2001). The greatest benefit of decision trees is their ease of use and understandability (Groth, 2000). Decision trees also scale up very well for large data sets (Adriaans and Zantinge, 1997).

Breiman et al., (1984) developed methods known as CART®, or classification and regression trees. Classification trees are used to predict membership of cases or objects in the classes of categorical dependent variables from their measurements on one or more predictor variables. Loh and Shih (1997) expanded work on classification trees with their paper regarding split selection methods. Some popular classification tree programs include, FACT (Loh and Vanichestakul, 1988) and THAID (Morgan and Messenger, 1973), as well as the related programs AID, for Automatic Interaction Detection (Morgan and Sonquist, 1963, and CHAID, for Chi-Square Automatic Interaction Detection, (Kass 1980)).

2.4.3 Association Rules

Association Rules help to identify how various attribute values are related within a data set. They are developed to predict the value of an attribute (or sets of attributes) in the same data set (Darling, 1997), or to discover correlations or co-occurrences of transactional events (StatSoft, 2002). They are useful when performing exploratory

analysis, or when searching for interesting relationships that may exist within a data set (Westphal and Blaxton, 1998). Since Association Rules are many times developed specifically to help identify these various regularities (patterns) within a data set, algorithms that utilize association rules have been found to work best with large data sets.

Agrawal, Imielinski, and Swami (1993) introduced Association Rules for the first time in their paper *"Mining Association Rules Between Sets of Items in Large Databases"*. A second paper by Agrawal and Srikant (1994) introduced the Apriori algorithm, which is the reference algorithm for the problem of finding Association Rules in a database. Initial studies regarding the discovery of Association Rules was performed in "Fast Discovery of Association Rules" (Agrawal et al).

2.4.4 Neural Networks

An Artificial Neural Network (ANN) is a system loosely modeled after the human brain in an attempt to simulate the multiple layers of simple processing elements called neurons. Each neuron is linked to specific neighboring neurons with varying coefficients of connectivity that represent the strength of these connections. Learning is accomplished by adjusting these strengths (weights) by to cause the overall network to produce the best possible resulting output.

Neural networks can be used to build explanatory models by exploring datasets in search of relevant variables or groups of variables. Haykin (1994), Masters (1995), and Ripley (1996) provide information on neural networks. Warner and Misra (1996) provide a good overview of neural networks used as statistical tools. The neural net

literature of late also contains some good papers covering prediction with missing data (Ghahramani and Jordan (1997) and Tresp, Neuneier, and Ahmad (1995)).

Neural Networks have been found to perform very well on classification tasks, and it has also been discovered that they are both reliable and effective when applied to applications involving prediction, classification, and clustering (Adriaans and Zantinge, 1997).

2.4.5 Genetic Algorithms

Genetic Algorithms are methods of combinatorial optimization techniques based on processes that occur in natural biological evolution. The name of this method is derived from its similarity to the process of natural selection. Three natural processes mimicked by software packages using genetic algorithms, including selection, crossover and mutation. Applying this concept to a Knowledge Discovery application involves the optimization of a data model along with a genetic method to obtain the most fit model (Groth, 2000). Presently, genetic algorithms are considered among the most successful of the machine-learning techniques in use.

Genetic algorithms are also a learning based data mining technique. Holland (1975) introduced genetic algorithms as a learning based method for search and optimization problems. Michalewicz (1994) provides a good overview of genetic algorithms, data structures, and evolution programs. Further usage of genetic algorithms have been discussed by Flockhart and Radcliffe (1996), Szpiro (1997), and Sharpe and Glover (1999).

2.5 Base Theory of Missing Data

The analysis of missing data is comparatively recent. With the advent of the mainframe computer in the 1960's, businesses were capable of collecting large amounts of data on their customer bases. As large amounts of data were collected, the issue of missing data began to appear. Afifi and Elashoff (1966), Hartley and Hocking (1971), Orchard and Woodbury (1972) and Dempster, Laird, and Rubin (1977) provide early seminal works on the analysis of missing data.

Little (1982) discussed models for non-response, while Little and Rubin (1987) considered statistical analysis with missing data. These papers sparked numerous works in the area of missing data which include: Diggle and Kenward (1994), Graham et al., (1997), Little (1995), Little and Rubin (1989), and Howell (1998). The problem of missing data is also very complex. Little and Rubin (1987) and Schafer (1997) provide conventional statistical methods for analyzing missing data. However, the statistical literature on missing data (Little, 1992) deals almost exclusively with the training of models rather than model prediction. They describe training as follows: when dealing with a small proportion of cases with missing data, you can simply eliminate the missing cases for purposes of training. However, if one wishes to make predictions for cases with missing inputs, cases cannot be eliminated.

Before an analyst can began to address the issue of missing data, it is important to understand the types of missing data that may be encountered. According to Little and Rubin (1989), there are several categories of missing data:

- Data Missing At Random

- Data Missing Completely At Random
- Non-Ignorable Missing Data
- Outliers Treated As Missing Data

2.5.1 Data Missing at Random (MAR)

It is obvious that cases containing incomplete data must be treated differently than cases with complete data. Rubin (Rubin, 1978) defined missing data as MAR "when given the variables X and Y, the probability of response depends on X but not on Y". Simply stated, some correlation exists between an attribute containing missing values and some other attribute(s) within the data structure. The pattern of the missing data may be traceable or predictable from other variables in the database rather than being due to the specific variable on which the data are missing (Statistical Services of University of Texas, 2000).

2.5.2 Data Completely Missing at Random (MCAR)

MCAR data exhibits a higher level of randomness than does MAR. Rubin (1978) and Kim (2001) classified data as MCAR when "the probability of response [indicates that] independence exists between X and Y". In other words, the observed values of Y are truly a random sample for all values of Y, and no other factors included in the study may bias the observed values of Y. In contrast to the MAR situation where data missingness is explainable by other measured variables in a study; non-ignorable missing

data arise due to the data missingness pattern being explainable --- and *only* explainable --- by the very variable(s) on which the data are missing (Stat. Serv. Texas (2000)).

In practice, the MCAR assumption is seldom met. "Non-Ignorable missing data is the hardest condition to deal with, but unfortunately, the most likely to occur" (Kim, 2001). Most missing data methods are applied upon the assumption of MAR, although that is not always tenable.

2.5.3 Non-Ignorable Missing Data

Given two variables X and Y, data is deemed Non-Ignorable when the probability of response depends on variable X and possibly on Y. For example, if the likelihood of an individual providing his or her weight varied according to the weight values in each age category, the missing data is non-ignorable (Kim 2001). Thus, the pattern of missing data is non-random and is not predictable from other variables in the database. As again corresponded by Kim (2001), "Non-Ignorable missing data is the hardest condition to deal with, but unfortunately the most likely to occur as well".

2.5.4 Outliers Treated as Missing Data

Data whose values fall outside of predefined ranges may skew test results. Many times it is necessary to classify these outliers as Missing Data.

Pre-testing and calculating threshold boundaries are also necessary in the pre-processing of data in order to identify those values which are to be classified as missing.

For even greater precision, various levels of data missingness for specific attributes can be calculated for their volume, magnitude, percentage and overall impact on other attributes in order to determine their effect on overall data mining performance. A "trigger" may then be defined in the data mining procedure to identify which test samples may be polluted with an overabundance of missing data, thus skewing the sample taken.

2.6 Methods of Addressing Missing Data

Several methods have been developed for the treatment of missing data. While the use of complete data only is a common approach, the cost of lost data and information when cases containing missing value are simply deleted can be huge. Another alternative is Data Imputation, the replacement of missing values with other known or derived values.

Schafer (1999), Schafer and Olsen (1998), Rubin (1996), Schafer (1997), and Little (1992) have all published articles regarding imputation methodologies. In addition, a number of case studies have been published regarding the use of imputation in medicine ((Barnard and Meng, 1999), and (van Buren, Boshuizen, and Knook 1999)) and in survey research (Clogg et al., 1991).

A number of researchers have begun to discuss specific imputation methods. Hot deck imputation and nearest neighbor methods are very popular in practice, but have received little overall coverage in the literature (Ernst (1980), Kalton and Kish (1981), Ford (1981), and David et al. (1986).

2.6.1 Use of Complete Data Only

One of the most direct and simple methods of addressing missing data is to include only those values with complete data. Only when missing data is classified as MCAR can this method be used successfully. If missing data are not classified as MCAR, bias will be introduced and make the results non-generalizable to the overall population. This method is generally referred to as the "complete case approach" and is readily available in all statistical analysis packages. When the relationships within a data set are strong enough to not be significantly affected by missing data, large sample sizes may allow for the deletion of a predetermined percentage of cases. Overall, this method is best suited to situations where the amount of missing data is small.

2.6.2 Delete Selected Cases or Variables

The simple deletion of data instances or attributes that contain missing values may be utilized when a non-random pattern of missing data is present. However, it may be ill advised to eliminate ALL of the samples taken from a test. The researcher may determine other methods to gain new knowledge from the test without dropping all sample cases from the test.

If the deletion of an entire particular subset (cluster) significantly detracts from the usefulness of the data, case deletion may not be effective. Furthermore, it may also simply not be cost effective to simply delete a significant number of cases from a sample. Nie et al (1975) examined this strategy, however no firm guidelines exist for the deletion of offending cases.

Furthermore, if the deletion of an attribute (containing missing data) that is to be used as an independent variable in a statistical regression procedure has a significant impact on the dependent variable, various imputation methods may be applied to replace the missing data, rather than altering the significance of the independent variable on the dependent variable.

2.6.3 Data Imputation Methods

2.6.3.1 Case Substitution

This method is the most widely used to replace observations with completely missing data. Cases are simply replaced by non-sampled observations. Only a researcher with complete knowledge of the data (and its history) should have the authority to impute missing data with values from previous research.

2.6.3.2 Mean Substitution

This type of imputation is accomplished by estimating missing values by using the mean of the recorded or available values. This is a popular imputation method for replacing missing data. However, it is important to calculate the mean only from responses that been proven to be valid and are chosen from a population that has been verified to have a normal distribution. If the data is proven to be skewed, the median of the available data can also be used as a substitute.

Mean imputation is a widely used method for dealing with missing data. The main advantage is its ease of implementation and ability to provide all cases with complete information.

2.6.3.3 Cold Deck Imputation

Cold deck imputation methods select values or use relationships obtained from sources other than the current database (see Kalton and Kasprzyk, 1982 and 1986; and Sande, 1982 and 1986). With this method, the end user substitutes a constant value derived from external sources or from previous research for the missing values. It must be ascertained by the end user that the replacement value used is more valid than any internally derived value. Pennell (1993) contains a excellent example of using cold deck imputation to provide values for an ensuing hot deck imputation application.

Unfortunately, feasible values are not always provided using cold deck imputation methods. Many of the same disadvantages that apply to the mean substitution method apply to cold deck imputation. Cold deck imputation methods are rarely used as the sole method of imputation and instead are generally used to provide starting values for hot deck imputation methods.

2.6.3.4 Hot Deck Imputation

Generally speaking, hot deck imputation replaces missing values with values drawn from the next most similar case(s). Once the most similar case(s) has been identified, hot deck imputation substitutes the most similar complete case's value for the missing value. The implementation of this imputation method results in the replacement of a missing value with a value selected from an estimated distribution of similar responding units for each missing value. In most instances, the empirical distribution consists of values from responding units. This method is very common in practice, but has received little attention in missing data literature. One paper using SAS to perform hot deck imputation is Iannacchione (1982). Advantages of hot deck imputation include conceptual simplicity, maintenance and proper measurement level of variables, and the availability of a complete set of data at the end of the imputation process that can be analyzed like any complete set of data. One of hot deck's disadvantages is the difficulty in defining what is "similar". Hence, many different schemes for deciding on what is "similar" may evolve.

2.6.3.5 Regression Imputation

Regression Analysis is used to predict missing values based on the variable's relationship to other variables in the data set. Simple and/or multiple regression techniques may be utilized to impute missing values. The first step consists of identifying the independent variables and the dependent variables. In turn, the dependent

variable is regressed on the independent variable(s). The resulting regression equation is then used to predict the missing values

Although regression imputation is useful for simple estimates, it has several inherent disadvantages:

1. This method reinforces relationships that already exist within the data. As this method is utilized more often, the resulting data becomes more reflective of the sample and becomes less generalizable to the universe it represents.

2. The variance of the distribution is understated.

3. The assumption is implied that the variable being estimated has a substantial correlation to other attributes within the data set.

4. The estimated value is not constrained and therefore may fall outside predetermined boundaries for the given variable. An additional adjustment may necessary.

In addition to these points, there is also the problem of over-prediction. Regression imputation may lead to over-prediction of the model's explanatory power. For example, if the regression R^2 is too strong, multicollinearity most likely exists. Otherwise, if the R^2 value is modest, errors in the regression prediction equation will be substantial (Graham, Hofer and Piccinin, 1994).

Mean imputation can also be regarded as a special type of regression imputation. For data where the relationships between variables are sufficiently established, regression imputation is a very good method of imputing values for missing data.

Overall, regression imputation not only estimates the missing values, but also derives inferences for the population (see discussion of variance and covariance above). For discussions on regression imputation see, Royall and Herson (1973) or Hansen, Madow and Tepping (1983).

2.6.3.6 Multiple Imputation

Rubin (1978) was the first to propose multiple imputation as a method for dealing with missing data. Multiple imputation combines a number of imputation methods into a single procedure. In most cases, expectation maximization (Little and Rubin, 1987) is combined with maximum likelihood estimates and hot deck imputation to provide data for analysis. The method works by generating a maximum likelihood covariance matrix and a mean vector. Statistical uncertainty is introduced into the model and is used to emulate the natural variability of the complete database. Hot deck imputation is then used to fill in missing data points to complete the data set.

Multiple imputation differs from hot deck imputation in the number of imputed data sets generated. Whereas hot deck imputation generates one imputed data set to draw values from, multiple imputation creates multiple imputed data sets. Multiple imputation creates a summary data set for imputing missing values from these multiple imputed data sets.

Multiple imputation has a distinct advantage in that it is robust to the normalcy conditions of the variables used in the analysis and it outputs complete data matrices. The method is time intensive as the researcher must create the multiple data sets, test the models for each data set separately, and then combine the data sets into one summary set.

The process is simplified if the researcher is using basic regression analysis as the modeling technique. It is much more complex when models such as factor analysis, structural equation modeling, or high order regression analysis are used.

2.6.3.7 Model-Based Procedures

Model-based procedures incorporate missing data into the analysis. These procedures are characterized in one of two ways: maximum likelihood estimation or missing data inclusion.

Dempster (1977) and Little and Rubin (1987) give a general approach for computing maximum likelihood estimates from missing data. They call their technique the EM approach. The approach consists of two steps, "E" for conditional expectation step and "M" for the maximum likelihood step.

The EM approach is an interactive method. The first step makes the best possible estimates of the missing data and the second step then makes estimates of the parameters (e.g., means, variances, or correlations) assuming the missing data are replaced. Each of the stages is repeated until the change in the estimated values is negligible. The missing data is then replaced with these estimated values. This approach has become very popular and is included in commercial software packages such as SPSS. Starting with SPSS 7.5, a missing value module employing the EM procedure for treating missing data is included.

Cohen and Cohen (1983) prescribe inclusion of missing data into the analysis. In general, the missing data is grouped as a subset of the entire data set. This subset of missing data is then analyzed using any standard statistical test. If the missing data occur

on a non-metric variable, statistical methods such as ANOVA, MANOVA, or discriminate analysis can be used. If the missing data occur on a metric variable in a dependence relationship, regression can be used as the analysis method.

2.7 The Impact of Missing Data on Data Mining Algorithms

Missing data impacts the Knowledge Discovery process in various ways depending on which data-mining algorithm is being utilized. The impact of missing data on various types of data mining algorithms will now be addressed.

2.7.1 The Impact of Missing Data on *k*-Nearest Neighbor

The very nature of the k-Nearest Neighbor algorithm is based on the accuracy of the data. Missing and inaccurate data have a severe impact on the performance of this type of algorithm. If data is missing entirely, misrepresented clusters (data distributions) can occur depending upon the frequency and categorization of the cases containing the missing data. One method to help solve this problem is to use the k-Nearest Neighbor data mining algorithm itself to approach the missing data problem. The imputed values obtained can be used to enhance the performance of the Nearest Neighbor algorithm itself.

First, the k-Nearest Neighbors (those containing *no* missing data) to the observation that does contain missing data are identified. The *k* stands for a predetermined constant representing the number of neighbors containing *no* missing data to be considered in the analysis. According to Witten and Frank (2000), it is advised to

keep the value for *k* small, say five, so that the impact of any noise present will be kept to a minimum.

Hence, this algorithm is not recommended for large data sets (Adriaans and Zantinge, 1997). Once these "neighbors" have been identified, the majority class for the attribute in question can be assigned to the case containing the missing value. Berson, Smith, and Thearling (2000) maintained that a historical database containing attributes containing similar predictor values to those in the offending case can also be utilized to aid in the classification of unclassified records.

Of course, the three main disadvantages mentioned in the imputation section (variance understatement, distribution distortion and correlation depression) should be addressed whenever a constant value is used to replace missing data. The proportion of values replaced should be calculated and compared to all clusters and category identification that existed prior to the replacement of the missing data.

2.7.2 The Impact of Missing Data on Decision Trees

Decision trees are a good methodology for dealing with missing data occurs frequently (Berry and Linoff, 1997). Decision trees also scale up very well for large data sets (Adriaans and Zantinge, 1997). It is sometimes useful to prune the tree whenever there is an overabundance of missing data in certain branches (Berry and Linoff, 1997). Eliminating particular paths may be necessary to ensure that the overall success of the decision-making process is not inhibited by the inclusion of cases containing missing data. Witten and Frank (2000) advise the use of pre-pruning during the tree-building

process to determine when to stop developing sub-trees. Post-pruning can be utilized after a tree is completely built. If one chooses post-pruning, decisions for pruning rules can then be made after the tree has been built and analyzed.

2.7.3 The Impact of Missing Data on Association Rules

Association Rules help to identify how various attribute values are related within a data set. Since Association Rules are many times developed to help identify various regularities (patterns) within a data set, algorithms that utilize association rules have been found to work best with large data sets. They are developed to predict the value of an attribute (or sets of attributes) in the same data set (Darling, 1997). The main focus of association rule discovery is to identify rules that apply to large numbers of cases that the rules can directly relate to, missing data may overstate both the support and the confidence of any newly discovered rules sets (Witten and Frank, 2000).

Attributes containing missing or corrupted data values may easily result in the creation of invalid rule sets or in the failure of identifying valid patterns that normally exist within the data. However, if the data set used to train the algorithm contains only "pristine" data, over-fitting the model based on the patterns included in the training set typically results.

Therefore, rules need to be developed for the "exceptions-to-rule-sets" that have been constructed in violation of correct or "clean" data. It is then necessary to populate the training set for algorithms that utilize Association Rules with a sufficient percentage of "noisy data", representing all possible types of exceptions to existing rules.

In this way, exception rules can be developed to handle all patterns of noise that may be associated with a given data set rather than redesigning rule sets that deal with "clean" data or attempting to force cases that do not belong to existing rule sets into those sets. As exceptions are discovered for initial exceptions, a type of tree structure is created, forming a decision list for the treatment of missing and noisy data for the data set. It becomes necessary to utilize both propositional rules and relational rules in the rule set for the treatment of missing or noisy data.

Propositional rules test an attribute's value against a constant value thereby developing very concise limits to delineate between "clean" and "noisy" data. In extreme instances, the constants, breakpoints and values from associated attributes are used to grow a regression tree in order to estimate missing data values under various conditions.

Incorporating an additional rule or rule set to deal with exceptions (such as missing data) can easily be incorporated since some rules may be developed to predict multiple outcomes. Failure to allow for the missing data exception may easily misrepresent some of the associations between attributes.

Although a rule may have both high support and confidence, a subjective evaluation by the end-user may determine how interesting a newly discovered rule is (Groth, 2000). Some association rule software packages may be trained to automatically prune "uninteresting rules". Therefore, minimum values (breakpoints) must be established for both the confidence and support of newly discovered rules.

In some instances, a hierarchy of rules can be developed so that some rules may imply other rules. In some cases, only the strongest rule is presented as a newly

discovered rule and rules of “lesser strength” (support and confidence) are linked to the stronger rule for use at a later time (Han and Kamber, 2001).

2.7.4 The Impact of Missing Data on Neural Networks

Neural Networks have been found to be both reliable and effective when applied to applications involving prediction, classification, and clustering (Adriaans and Zantinge, 1997). Missing data has a similar impact on neural networks as it does on other types of classification algorithms, such as k-Nearest Neighbor. These similarities include variance understatement, distribution distortion, and correlation depression.

When using neural networks are used in the presence of missing data in the data mining process, it may be necessary to “train” the initial network with missing data if the data to be tested and evaluated later is itself going to contain missing data. By training the network with cases containing complete data only, the internal weights developed with this type of training set cannot be accurately applied to a test set containing missing values in later usage of the neural network model.

Missing data actually impacts the internal execution of the neural network in several ways. Since the internal weights used to calculate outputs are created and distributed within the network without providing the insight as to how a solution is created, missing or dirty data can distort the weights that are assigned as the associations between nodes in a manner unknown to the research analyst.

While the hidden layer is where the actual weights are developed for the network, the activation function combines the inputs to the network into a single output (Westphal

and Blaxton, 1998). The output remains low until the combined inputs reach a predetermined threshold, and small changes to the input can have a dramatic effect on the output (Groth, 2000). The activation function can also be very sensitive to missing data.

The activation function of the basic unit of a neural network has two sub-functions: the combination function and the transfer function. The combination function commonly uses the "standard weighted sum" (the summation of the input attribute values multiplied by the weights that have been assigned to those attributes) to calculate a value to be passed on to the transfer function. The transfer function applies either a linear or non-linear function to the value passed to it by the combination function. Even though a linear function used in a feed-forward neural network is simply performing a linear regression, missing values can distort the coefficients in the regression equation and therefore pass on invalid values as output (Berry and Linoff, 1997).

2.7.5 The Impact of Missing Data on Genetic Algorithms

Genetic Algorithms relate to evolutionary computing that solves problems through the application of natural selection and evolution (Sethi, 2003). In a GA application, a chromosome is a string of binary bits in which a possible solution is encoded. The quality of a solution is called a "fitness function". The search for the optimal solution creates a "gene pool" with associated evolving fitness values. Missing data can severely impact the evaluation of the most fitness functions, resulting in a potentially inappropriate solution being chosen as most optimal (Wang, 2003).

CHAPTER III

DEVELOPMENT OF HYPOTHESIS

Developed on the base theories of previous research, this chapter presents a research model that will address the use of several knowledge discovery models that utilizes neural networks as their data mining algorithm. An s-Sigmoid transfer function is employed in the neural network as the selected transfer function, and when the secondary data used in each model is injected with a specified increasing level of data missingness. The data imputation methods of case deletion and mean substitution are utilized in the presence of various levels of data missingness and then compared for effectiveness.

3.1 The Research Model

Grounded on the base theory of four various dimensions (knowledge discovery, data mining, missing data and data imputation), a research model is proposed to explore the performance effectiveness of the more popular methods of knowledge discovery and data mining techniques, when confronted with various increasing levels of data

missingness. Unique to this research model is the variation of secondary data in the volume of data instances employed in training and testing, as well as also varying the level of data missingness in each KDD model. Evaluation of the most popular data imputation methods for handling the problem of missing data is then tested and analyzed.

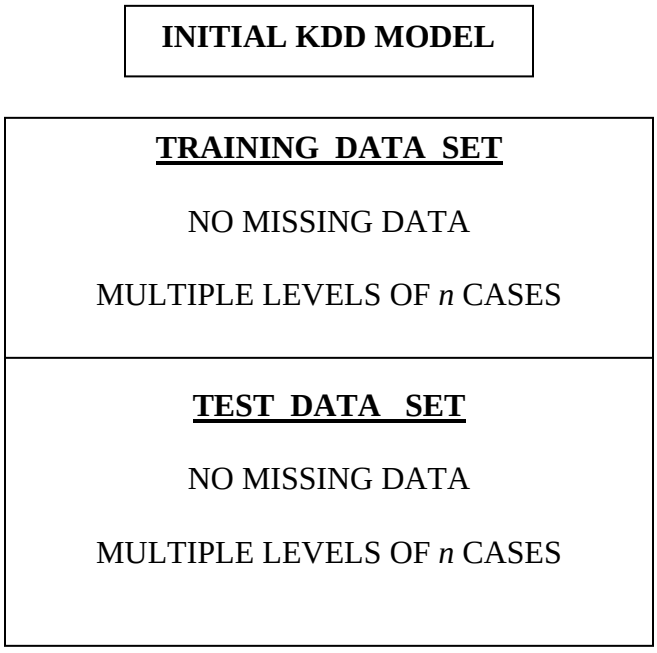
This research model extends the traditional methodology by extending the constructs of previous research (Howell, (1998) and Little and Rubin, (1989)).

The innovation proposed in this research model is to simultaneously alter the case volume of the secondary data and the level of data missingness within the data set, and to test the impact of Data Imputation and data Re-Sequencing on model performance.

Therefore, the Initial Research Model can be stated as follows:

An Unsupervised Knowledge Discovery Model Utilizing A Neural Network Algorithm with an S-Sigmoid Transfer Function Evaluating Root Mean Square Values In The Presence Of Various Increasing Levels of Data Frequency.

Figure 3.1: The Initial Research Model

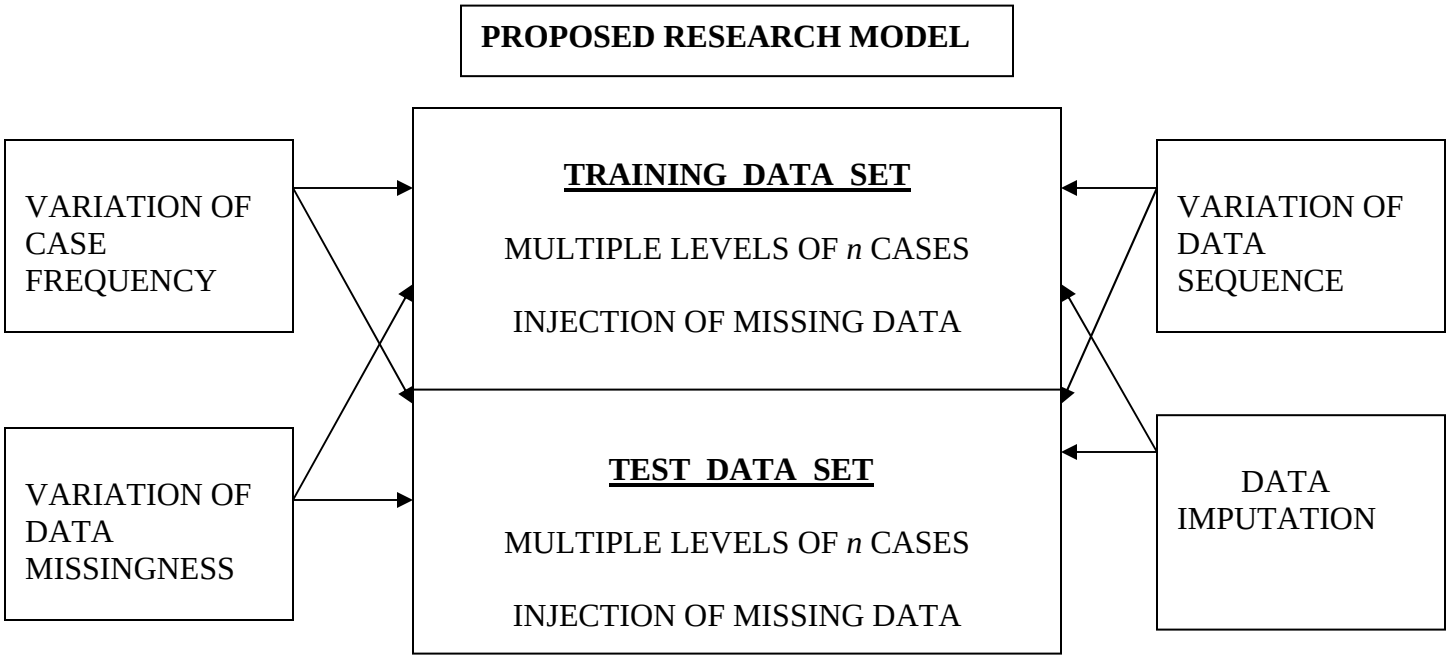


The overall innovation proposed in this research model is to simultaneously modify and/or add various constructs to Knowledge Discovery models, and to measure the impact of those additions/modifications. The Case Frequency Volume of the models, the Level of Data Missingness within the models, the Re-Sequencing of Data Instances within the models, as well as and the Impact of various methods of Data Imputation on model performance will be tested and measured.

Therefore, the proposed Research Model is stated as follows:

An Unsupervised Knowledge Discovery Model Utilizing A Neural Network Algorithm with an S-Sigmoid Transfer Function Evaluating Various Imputation Methods and Data Re-Sequencing In The Presence Of Various Increasing Levels of Data Missingness:

Figure 3.2: The Proposed Research Model



Root Mean Square (RMS) Values, ANOVA Testing, T-tests, and Tukey's Honestly Significant Difference Test will be used to evaluate the differences between performance levels of the various Knowledge Discovery and Neural Network Models, both in the presence and absence of Missing Data in the Training and Testing data sets under study.

3.2 Exploratory Hypothesis

An initial hypothesis is developed to test the impact of training, testing and evaluating a KDD Neural Network model employing an S-Sigmoid Transfer Function that contains no missing data, and training, testing and evaluating the same KDD model utilizing a low volume of case frequencies, and when missing data is injected into the into those KDD models at increasing levels.

Hypothesis 1:

Data set case frequency is a significant factor in the performance of KDD models that utilize a Neural Network as the Data Mining Algorithm and employ an S-Sigmoid Transfer Function, as measured by the Root Mean Square Value calculated for the model.

KDD models that contain at least 1000 cases are said to achieve results similar to data warehouses containing an extremely high volume of case frequency. Therefore, two KDD models from different disciplines will be selected for testing. Randomly selected cases will be selected from these models to form new KDD models containing case frequencies from N=500 to N=5000 at increments of 500 cases, for a total of twenty new MDD models.

Root Mean Square (RMS) Values will then be calculated for each newly constructed model. These RMS values will be analyzed to test the impact of varying the number of cases used for training and testing KDD models.

An ANOVA Test will be conducted to test the results of all Root Mean Square Values obtained from the newly constructed KDD models.

Hypothesis 2:

The Level of Data Missingness in the Training and Testing data sets of a KDD model is a significant factor in the calculation of the Root Mean Square Value for a KDD model in small models (Case Frequency N=500).

KDD models from five disciplines will be selected. Five hundred random cases will then be selected from each model to construct new models. Root Mean Square (RMS) Values will be calculated for each new model with N=500.

Missing Data will then be injected into each model at increasing levels, beginning at ten percent and continuing up to seventy percent, at ten percent increments.

Root mean Square (RMS) values will be calculated at each level of Data Missingness for the models.

An ANOVA Test will then be conducted to determine if, in KDD models containing 500 cases, the Level of Data Missingness is significant in the calculation of the Root Mean Square Value.

If the ANOVA test determines that a significant difference exists in the level of Data Missingness, the Tukey's Honestly Significant Difference Test will be performed as an ad hoc test.

Hypothesis 3:

The Level of Data Missingness in the Training and Testing data sets of a KDD model is a significant factor in the calculation of the Root Mean Square Value for KDD models containing various levels of Case Frequency.

When a KDD model utilizing a neural network uses complete data to train and test the model, its performance may begin to degrade and perform poorly when confronted with data that contains missing values. The level of Data Missingness in the data may indicate a point at which the model degradation may begin. KDD models from various disciplines will be selected for training and testing.

Missing Data will then be injected into each model at increasing levels, beginning at ten percent and continuing up to seventy percent, at ten percent increments.

Root mean Square (RMS) values will be calculated at each level of Data Missingness for the models.

An ANOVA Test will conducted to determine if the Level of Data Missingness is significant in the calculation of the Root Mean Square Value for KDD models with different volumes of Case Frequency.

Hypothesis 4:

Data Imputation (Mean Substitution and Case Deletion) is a significant factor in the performance of KDD models containing various volumes of Case Frequency and various levels of Missing Data.

KDD models from various disciplines will be selected for training and testing. A Root mean Square (RMS) Value will be calculated for each model using complete data.

Missing Data will then be injected into each model at increasing levels, beginning at ten percent and continuing up to seventy percent, at ten percent intervals.

Root mean Square (RMS) values will be calculated at each level of Data Missingness for the models.

The two most common types of Data Imputation will then be performed on the KDD models containing Missing Data, Mean Substitution and Case Deletion.

Following Data Imputation, each model will be re-trained and re-tested. Including the original KDD model, a total of twenty-two KDD models will be tested for each of the five original models, for a grand total of one-hundred and ten total models.

An ANOVA test will be performed to determine if performing Data Imputation is significantly different from not performing Data Imputation using KDD models from various disciplines. A marginal means plot will also be used to explore possible interactions between the factors of Imputation Type and the Level of Data Missingness..

Hypothesis 5:

There is no difference between the Data Imputation Methods of Mean Substitution and Case Deletion, when performed on Missing Values in KDD models containing various levels of Case Frequency and various levels of Data Missingness.

If it is discovered that Data Imputation is significantly different from not performing Data Imputation on the Missing Values in a KDD model, then an ad hoc a-priory test will be conducted on the Imputation Methods of Mean Substitution and Case Deletion to determine if there is a significant difference between the two methods.

Hypothesis 6:

The re-sequencing of data cases in the training and test data sets of KDD models containing various volumes of Case Frequency is a significant factor in the performance of those models.

This research will attempt to determine if a KDD model's performance is significantly different following re-sequencing the data instances used for the training and testing of the KDD model.

KDD models from five different disciplines will be trained and tested, and a Root Mean Square (RMS) Value calculated for each model.

The Data Instances in each model will be re-sequenced. The models will then be re-trained and re-tested, and new Root Mean Square (RMS) Values calculated.

A T-test will then be conducted to test the means of the Root Mean Square (RMS) Values discovered for the KDD models in their original sequence and the means of the KDD models after the Data Instances used for training and testing have been randomly re-sequenced.

Hypothesis 7:

The re-sequencing of data cases in the training and test data sets of KDD models containing various volumes of Case Frequency and various volumes of Data Missingness is significant in the performance of those models.

This research will attempt to determine if the sequence of the cases containing Missing Values in the data sets used for training and testing a KDD model has a significant impact on the performance of the KDD model.

KDD models from five different disciplines will be trained and tested, and a Root Mean Square (RMS) Value calculated for each model.

Data Missingness will be injected into each KDD model at increasing levels and model performance will be tested by calculating a Root Mean Square (RMS) Value for each new KDD model for each level of Data Missingness.

In each new KDD model, all Data Instances in used for training and testing in each model will be re-sequenced.

The KDD models will then be re-trained and re-tested, and new Root Mean Square (RMS) Values calculated.

A T-test will then be conducted to test the means of the Root Mean Square (RMS) Values discovered for the KDD models in their original sequence and the means discovered for the same KDD models after the Data Instances used for training and testing have been randomly re-sequenced.

Hypothesis 8:

There is no significant difference between the Imputation Methods of Mean Substitution and Case Deletion in the performance of KDD models when the level of Data Missingness in those models is increasingly varied and the Case Frequency is held constant, N=500.

According to Howell (1998), the results of tests performed on data sets containing 1000 cases can be used as valid estimates for data warehouses containing far greater volumes of Case Frequencies.

Tests will be conducted to ascertain if KDD models containing less than 1000 cases (N=500) perform differently when injected with varying increasing levels of Data Missingness (10%, 20%, 30%, 40%, 50%, 60% and 70%).

Therefore, tests will be performed on KDD models containing less than 1000 cases and with various increasing levels of Data Missingness. Although inaccurate performance statistics may be generated by "small" training and testing models, it will be interesting to observe if either greater performance degradation will be observed, or if seemingly positive, although inaccurate, results will be determined.

ANOVA tests will then be performed on the Root Mean Square (RMS) values calculated for all KDD models with N=500, at various levels of Data Missingness and following execution of the data imputation methods of Mean Substitution and Case Deletion.

Hypothesis 9:

There is no significant difference between the imputation methods of Regression Imputation, Mean Substitution and Case Deletion when performed on KDD models containing increasing levels of Data Missingness, and when the Case Frequency of the models are constant, N=1000.

Tests will be conducted to ascertain if KDD models containing exactly 1000 cases for model testing and training perform differently when injected with varying increasing levels of Data Missingness (10%, 20%, 30%, 40%, 50%, 60% and 70%) and after the imputation methods of Regression Imputation, Mean Substitution and Case Deletion on those Missing Values have been performed.

Root Mean Square (RMS) values will be calculated for each KDD model, both prior to and following Data Imputation using the three methods mentioned above.

Hypothesis 10:

There is no difference in the performance of KDD models when the level of Data Missingness in those models is increasingly varied and the Case Frequency is held constant, N=1000, and when the Imputation Methods of Multiple Imputation, Regression Imputation, Mean Substitution and Case Deletion are performed on those models.

Mean Substitution and Case Deletion are the most common Data Imputation Methods employed when dealing with the dimension of Data Missingness. Regression Imputation and Multiple Imputation are also frequently employed Data Imputation Methods.

A KDD model containing 1000 cases will be selected for training and testing.

New KDD models will be created by injecting Data Missingness into the original KDD model at the 10%, 20%, 30%, 40%, 50%, 60% and 70% levels.

New KDD models will be constructed by employing the Data Imputation Methods of Regression Imputation, Multiple Imputation, Mean Substitution and Case Deletion.

A total of thirty-six KDD models will be used for testing this hypothesis. Root Mean Square (RMS) values will then be calculated for each of the thirty-six KDD models.

Anova Tests and Tukey's Honestly Significantly Difference Test (if necessary) will be conducted to determine the effectiveness of Data Imputation on KDD models containing varying increasing levels of Data Missingness.

CHAPTER IV

RESEARCH METHODOLOGY

This research will explore the impact of data missingness at various increasing levels in KDD models that contain various volumes of case frequencies that employ Neural Networks as the Data Mining Algorithm. The s-Sigmoid is the Transfer Function employed by the Neural Network in this study. Four of the most commonly utilized Data Imputation methods, Case Deletion, Mean Substitution, Regression Imputation and Multiple Imputation will then be used to determine their effectiveness in dealing with the issue of data missingness.

While studies have been conducted independently in the areas Knowledge Discovery, Missing Data and Data Imputation, only a few have integrated all three dimensions. This research extends previous studies by isolating on KDD models that employ Neural Networks that utilize the s-Sigmoid transfer function. Further, tests are conducted utilizing secondary data sets containing multiple levels of case volume frequencies. Even more, the level of data missingness is also altered within each data set tested.

The methodology for this research is proposed at this point. Following are discussions detailing the software to be employed in the study, the population of the secondary data to be acted upon and the manipulation of that data during the study.

The overall research model is designed to bring together the three focal areas of study: Knowledge Discovery (in conjunction with Data Mining and Neural Networks utilizing an S-Sigmoid Transfer function), Data Missingness and Data Imputation.

First, a software product was chosen that contains a particular data mining algorithm, and also utilizes a specific type of internal architecture. Various algorithms may be selected for use in the data mining phase of the Knowledge Discovery process, and selected options and variations may then be selected.

Next, the population of secondary data is selected for study. This data has been deemed appropriate for use in Knowledge Discovery testing and has been used in previous studies. Five data sets of various case frequencies were selected for this research and analyzed. The data was then altered along several dimensions and retested.

Finally, the results of each research model is evaluated, compared, summarized and presented by the dissertation author.

One goal of this study is to present an imminent approach to practitioners for effective Knowledge Discovery in the presence of various levels of missing data. The results of this research hope to present interesting guidelines for users of commercial KDD software packages when confronted with the real-world problem of data pollution, primarily data missingness.

The dissertation author has concentrated on areas that are generalizable to the most common types of both available commercial software and types of data missingness

issues that may be confronted in industry. Therefore, some data imputation methods that require strong influence from human interaction with the data were excluded from this study, i.e. Cold Deck Imputation.

4.1 Software Utilized In This Research

A backpropogation Neural Network architecture employing an s-Sigmoid Transfer Function was chosen for this study. The network is trained using a data set(s) with known values for the dependent variable(s). The Root Mean Square (RMS) error (comparison between desired output and computed output) was selected as the metric to be evaluated in determining the performance of each backpropogation feed-forward Neural Network model. The Intelligent Data Analyzer (iDA) software product, developed by the Information Acumen Corporation, was selected to perform the data mining session.

The primary reason for the use of this particular software was its use of a backpropogation feed-forward Neural Network data mining algorithm. Further, iDA also utilizes an s-Sigmoid Transfer Function within the Hidden Layer of the Neural Network. The use of Neural Networks is one of the most commonly utilized data mining algorithms, and due its ability to handle both linear and non-linear data, the s-Sigmoid is the most commonly used Transfer Function employed in Neural Networks. Also, the iDA package attaches itself to Microsoft's EXCEL spreadsheet software for ease of use and data manipulation.

The Data Mining phase of the Knowledge Discovery process employs either a single or a hybrid of various algorithms. Commonly used algorithms for Data Mining

include Nearest Neighbor, Decision trees, Association Rules, Neural Networks and Genetic Algorithms. Neural Networks are commonly used in commercial data mining applications, due to their effectiveness in dealing with various types of input data and their ability to handle Missing Data.

Neural Networks may employ various types of Transfer Functions in the Hidden Layer of the network. The types of Transfer Functions that may be utilized include the Hyperbolic Tangent, Multiple Regression and the s-Sigmoid function. The s-Sigmoid is the Transfer Function selected for most commercial data mining software packages, due to its ability to handle both linear and non-linear data types.

4.1.1 Software Parameters Utilized in Testing

Five basic parameter values must be specified by the researcher prior to testing the secondary data with the iDA software package.

First, since Neural networks may contain a multiple hidden layers that are utilized in Feed Forward-Back Propagation networks, the number of hidden layers to be utilized must be specified. Hidden layers are layers of processing elements that are not directly connected to the external world and are used by the neural network in the calculation of output nodes. The iDA software used in this research can use one or two hidden layers, specified prior to the execution of the training phase of neural learning for the network. This research will be conducted utilizing a single Hidden Layer.

Second, a particular Learning Rate for the network must be set. The goal of robust and expedient network convergence is more easily obtained when the data to be used for training the neural network is linear in nature. However, when the instances used for the

training of a neural network is nonlinear in nature, the Learning Rate may be increased to permit robust convergence. The Learning Rate can range from 0.10 to 0.90. The lower the learning rate, the more training iterations will be required by the network in order for the network to converge. A higher learning rate will permit the network to converge more rapidly, sometimes allowing for a less than optimal solution to be derived by the neural network.

Third, a finite number of epochs to be used for training the neural network must be established by the researcher. The epoch parameter is the number of iterations (cycles) that the entire set of training data is passed through the neural network. At the end of an epoch, all weight parameters within the neural network are adjusted and updated with the results obtained from the calculations made during that particular iteration. The default parameter value for the number of epochs to be initially used during testing using the iDA software package is 20000.

Fourth, a convergence setting must be initialized to set a maximum level for the Root Mean Square Error value to be used as a threshold for terminating the training phase of the neural network. This convergence parameter should be set at a reasonably low value if the researcher determines that the termination of the training phase of the neural network should rather be based on the number of epochs rather than the Root Mean Square Error value.

A reasonable level for the maximum Root Mean Square value has been determined by past researchers of neural networks to be approximately 0.10.

Finally, the number of instances to be used by the neural network in the training phase must be established. If the secondary data to be tested contains *n* instances and *m*

instances are to be used for *training* the network, then the difference in the frequency of these instances (*n*-*m*) are then utilized for *testing* the neural network. However, if the entire set of *n* instances are selected for training the neural network, the entire data set (both the instances used for training as well as the instances initially selected for testing) will also be included in the testing of the neural network.

Therefore, the parameter value usually selected for initial research is commonly set at 50% of the number of total instances contained in the data set to be tested.

4.2 Secondary Data Selected for Research

The data for study in this research was selected by the dissertation author primarily due to the format of the data and the Case Frequencies contained within the data sets.

In order for data to be better utilized by a backpropogation feed-forward neural network, all input values must be converted to be within a range of zero to one. The data must also contain a dependent variable whose value can be derived from other variables in the data set, and can be set at either zero or one.

A frequency of approximately 1000 cases is deemed large enough for results to be applicable to a much larger data warehouse (Howell, 1998). Therefore, a data set with a case frequency of 1000 was first selected. A data set with a less optimal case frequency was then selected to observe the impact of the research parameters when the case volume is low. For this situation, a case frequency of 500 was chosen. The entire secondary data model used for this experiment included five data sets containing different frequencies of cases obtained from various sources, where the value of a dependent variable within each

data set had previously been determined and verified. Methods of data imputation are employed within a knowledge discovery model to attempt to identify how various levels of data missingness within data sets of varying frequencies of cases may impact a Data Mining study. The data sets contained case frequency levels of N=500, 1000, 3500, 5000 and 7000, respectfully.

4.3 Research Methodology

Each data set will initially be mined with no missing data, without altering the standard parameters necessary for data mining utilizing an untrained neural network (learning rate, number of input nodes, number of hidden layers, number of epochs) and obtaining a Root Mean Square (RMS) value. Each data set will then be injected with a particular level of data missingness (e.g., 10%, 20%, 30%, 40%, 50%, 60%, 70%) and mined again, using the same standard neural network parameters. A total of eight different tests will be performed on each data set. Data imputation will then be performed on the missing values using the Case Deletion and Mean Imputation methods, again with no missing data, and at the 10%, 20%, 30%, 40%, 50%, 60%, 70% levels of data missingness.

The RMS results obtained will then be analyzed using T-Tests, two and three factor ANOVA tests, and Tukey's Honestly Significant Difference (HSD) Test Statistic to determine if original data set size, level of data missingness, and/or data imputation method are significant factors in the performance of KDD models..

CHAPTER V

TEST RESULTS AND DISCUSSION

The overall mission of this research is to analyze the impact and significance of Missing Data on the process of Knowledge Discovery In Databases (KDD) on data sets that have varying volumes of case frequency and data missingness. In addition, multiple types of data imputation are tested and evaluated for overall performance when the Neural Network Algorithm employed in the Data Mining phase of Knowledge Discovery employs an s-Sigmoid Transfer Function .

The Data Mining step utilizing this methodology was performed on secondary data sets of complete data from various disciplines at various levels of case frequency (N=500, N=1000, N=3500, N=5000 and N=7000 cases). Various levels of Data Missingness were then injected into each of the selected test data sets prior to Neural Network training and testing (at the 10%, 20%, 30%, 40%, 50%, 60%, 70% levels of data missingness), and data mining performed on those data sets. Further testing of the KDD utilizing an ANN with an s-Sigmoid Transfer Function was performed when the missing data values were imputed utilizing the most common methods of data imputation, Case Deletion, Mean Substitution, Multiple Regression and Multiple Imputation. Comparisons of these test results were then performed using ANOVA, T-Tests, and Tukey's Honestly Significant Difference (HSD) test.

All secondary data sets containing various levels of Case Frequency Volumes (N=500, N=1000, N=3500, N=5000, N=7000) were selected and tested in similar fashion.

Each data set was mined using the Data Mining software package, Intelligent Data Analyzer. An unsupervised Data Mining methodology was employed, and a Root Mean Square value was determined by the software through the utilization of a Neural Network data mining algorithm, and with the s-Sigmoid algorithm as the Transfer Function employed when transferring data to the Neural Network Output Layer by the ANN.

After each KDD model was initially mined with No Missing Data and a Root Mean Square value calculated, each data set was injected with various levels of Data Missingness (10%, 20%, 30%, 40%, 50%, 60% and 70%) and the KDD model tested again at each level. A Root Mean Square value was determined for each level of Data Missingness (10%-70%).

Each of the new KDD models for each level of Case Frequency was then re-trained and re-tested using some of the most widely employed methods of Data Imputation: Case Deletion. Mean Substitution, Regression and Multiple Imputation

Root Mean Square values were then calculated for all data sets at various levels of Case Frequency. Root Mean Square (RMS) values were calculated for each data set with No Missing data, seven different levels of Data Missingness (10%-70%), and after the Data Imputation methods of Case Deletion and Mean Substitution were employed at the various levels of Data Missingness.

A total of one-hundred and twenty-six KDD models were initially developed for this study. Each Data Mining Session for every individual KDD model calculated a Root Mean Square (RMS) value for the KDD model being analyzed.

5.1 Validity

The validity of the secondary data sets selected for use in this study was first explored. The five KDD models initially selected were intended for the use of predicting the value of a dependent variable by using a set of independent variables. Each KDD model contained historical data from various scientific/medical/business disciplines, containing proven data values that had previously been used in the prediction of the dependent variables. Therefore, the content validity of the secondary data was inherent by the selection of data sets that had already been validated in the prediction of a selected dependent variable in each KDD model.

5.1.1 Content Validity

The first dimension considered in this study was that of KDD Case Volume Frequency. Adequate coverage for the investigation of the impact of Missing Data on the performance of a KDD application using a Neural Network as its Data Mining algorithm was addressed by selecting KDD models containing various levels of case frequency. According to Howell (1998), a data set containing one thousand data instances is sufficient in the testing of data warehouses. It was therefore determined to test in this research one data set containing less than one thousand instances (N=500), another data set containing exactly one thousand instances and several data sets with increasing levels

of case frequency (N=3500,N=5000 and N=7000). All data sets were selected from sources guaranteeing the validity of the data. A group of educators and practitioners reviewed the data sets prior to testing.

5.1.2 Criterion Validity

The predictive type of criterion validity is tested in this KDD study through the use of Data Mining software utilizing a Neural Network that employed an S-Sigmoid Transfer Function in its Activation Function as its Data Mining algorithm. In order to facilitate this type of validity, the data sets selected had to contain a dependent attribute whose value could be predicted by other independent attributes within the same KDD model. All secondary data sets selected for this study met all criteria of data dependence/independence and predictability.

5.1.3 Construct Validity

As Neural Network algorithms perform more effectively when both the dependent and independent attributes being tested contain values between zero and one (Berry and Linnoff, 2000), this feature was also a necessary factor in the selection of the KDD models to be tested. The secondary data selected for this study met all criteria necessary for maximum efficiency of Neural Network performance, as well as the aforementioned criteria for the prediction of a dependent attribute within each KDD model.

In order to test the second dimension of this study, the impact of data missingness, a methodology had to be developed to randomly inject specific levels of data missingness into the data sets. This issue was addressed and performed by the development and

execution of a Visual Basic module that allowed for the random selection of cases and a specific level of data missingness was be injected into the data set. Standard parameters for the Data Mining segment of the Knowledge Discovery process were also set by the module. The Neural Network Architecture's Backpropagation Learning Parameters were not altered from their initial default settings prior to training and testing each KKD model. These parameters included the Number of Hidden Layers to be used by the Neural Network, the Neural Network's Learning Rate, the number of Epochs to be used and the Convergence Level of the network.

A selected percentage of random attribute values were selected by the module and replaced those values with null (missing) values to complete the injection of data missingness at the specified level.

Root Mean Square values were then calculated by the iDA ANN software architecture for each data set containing complete data and at each of the seven levels of data missingness (10%-70%).

The third dimension tested in this study was the type of Data Imputation method employed on the selected KDD models in the presence of Missing Data. The most commonly used types of data imputation employed in KDD and data mining applications are case deletion and mean imputation. A Visual Basic module was developed to automatically prepare all data sets utilizing the aforementioned data imputation methods prior to performing the re-training and re-testing of the KDD models. Therefore, all criteria necessary to provide adequate data sets meeting all dimensions and constraints of this analysis were met.

5.2 Analysis

Root Mean Square (RMS) values were obtained from all data mining sessions performed on the KDD models under analysis by utilizing the Intelligent Data Analyzer Data Mining software (iDA), and are displayed in Table 5.1.

These calculations were performed on all data sets at various levels of case frequency (N=500, N=1000, N=3500, N=5000, N=7000), with different levels of injected data missingness (10%, 20%, 30%, 40%, 50%, 60%, 70%), and after employing the Data Imputation methods of Case Deletion and mean Substitution.

Table 5.1 Root Mean Square Statistics for all KDD Models

		% of Missing Data						
Original Data Set Size	**Imputation Method**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**
500	**No Method**	.449	.454	.454	.454	.454	.445	.453
	Mean Sub.	.255	.200	.250	.290	.199	.240	.270
	Case Del.	.249	.230	.210	.269	.270	.230	.200
1000	**No Method**	.144	.144	.144	.144	.144	.144	.144
	Mean Sub.	.087	.087	.087	.087	.085	.087	.087
	Case Del.	.083	.094	.098	.097	.096	.101	.107
3500	**No Method**	.105	.340	.020	.250	.096	.360	.420
	Mean Sub.	.254	.196	.248	.287	.195	.234	.261
	Case Del.	.243	.225	.208	.261	.267	.228	.198
5000	**No Method**	.125	.340	.020	.250	.430	.360	.420
	Mean Sub.	.251	.189	.241	.280	.190	.230	.259
	Case Del.	.240	.209	.200	.260	.259	.221	.193
7000	**No Method**	.105	.340	.200	.250	.430	.360	.420
	Mean Sub.	.250	.190	.240	.280	.190	.230	.260
	Case Del.	.240	.221	.200	.261	.260	.218	.191

5.2.1 Hypothesis Testing

Hypothesis 1 tested the significance of the volume of data set case frequency in KDD models that employ a Neural Network as the data mining algorithm, and utilizing a S-Sigmoid Transfer function within the Activation Function of the Artificial Neural Network (ANN).

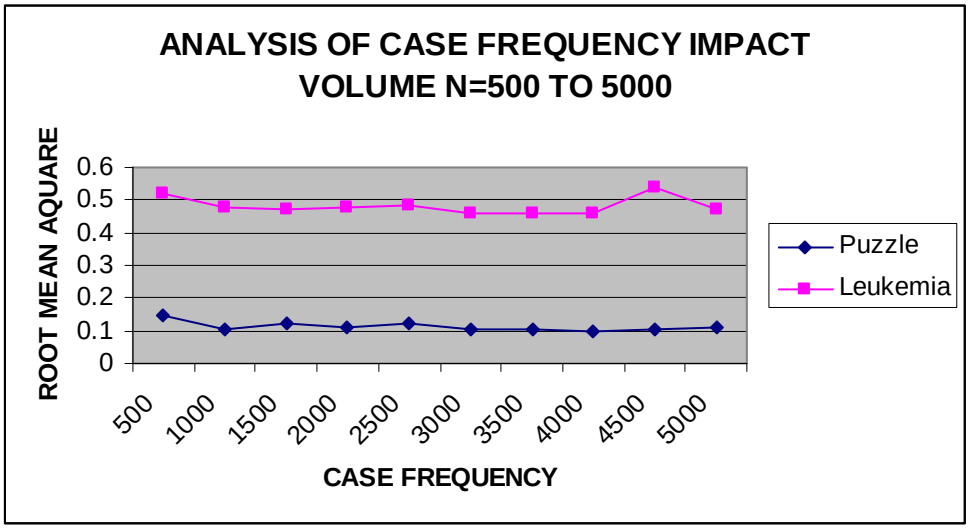
The two largest of the original five KDD models with complete data (N=5000, N=7000) were selected for testing.

This hypothesis tested the ten new models that were constructed for each of the two original KDD models. Random instances were selected to construct these ten new models with Case Frequencies of N=500, N=1000, N=1500, N=2000, N=2500, N=3000, N=3500, N=4000, N=4500 and N=5000 instances, respectively, for a total of twenty new KDD models.

These twenty new KDD models were trained and tested using the iDA data mining software. Root Mean Square (RMS) values were calculated employing iDA's Neural Network Algorithm utilizing an S-Sigmoid Transfer Function within the Activation Function. The standard default parameters of the ANN architecture were left intact prior to training and testing each KDD model.

Figure 5.1 illustrates the resulting Root Mean Square (RMS) values for the twenty KDD models that were analyzed when Complete Data was used for testing by the iDA software:

Figure 5.1 Root Mean Square Values For Two Models: Complete Data



It can be seen from the above graph that a slight positive variation (lower RMS value) was observed when the Case Frequency Volume was increased from 500 to 1000 instances in both of the original KDD models. As the Case Frequency Volume was increased in increments of five hundred, only slight positive or negative variation was observed.

An ANOVA test was then performed on the new Root Mean Square (RMS) values calculated by the iDA software. This test concluded that data set size (case frequency volume) of a KDD does not significantly impact the Root Mean Square values calculated by the proposed KDD's that utilize an S-Sigmoid transfer function employing a Neural Network as it's data mining algorithm.

Table 5.2 illustrates the results of the ANOVA test:

Table 5.2 ANOVA Results for Root Mean Square Values

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	**0.005756**	**9**	**0.00064**	**2.149023**	**0.134937**	**3.178893**
Columns	**0.680436**	**1**	**0.680436**	**2286.369**	**3.83E-12**	**5.117355**
Error	**0.002678**	**9**	**0.000298**			

Therefore, at the 0.05 level of significance, we accept the null hypothesis that data set case frequency is not a significant factor in the calculation of Root Mean Square Values when a Neural Network utilizing an S-Sigmoid Transfer Function within the Activation Function is employed as the Data Mining Algorithm by a KDD model.

.

Hypothesis 2 tested the significance of data missingness on KDD models from various scientific/medicine/business disciplines that contain a case frequencies of less than 1000 instances (N = 500). That is, no significant difference exists in the RMS values calculated in a KDD model containing less than 1000 instances using complete data, and when data missingness is injected into the model at the 10%, 20%, 30%, 40%, 50%, 60% and 70% levels.

Five new KDD models were constructed by randomly selecting 500 cases from the original KDD models (N=500, N=1000, N=3500, N=5000, N=7000) selected for this study. Using complete data for these new KDD models containing 500 cases, the models were trained and tested, and the resulting Root Mean Square (RMS) values recorded.

Data missingness was then injected into each of these five new KDD models (N=500) at the aforementioned levels (10%-70%). The KDD models were re-trained and re-tested and new Root Mean Square (RMS) values calculated.

The following graphs illustrate a comparison of all Root Mean Square (RMS) values that were calculated when the datasets used for training and testing the ANN in the KDD model contained N=500 instances only, and when noise was randomly injected into each model, 10%-70%.

Figure 5.2 500 Instances – Missing Data Plot

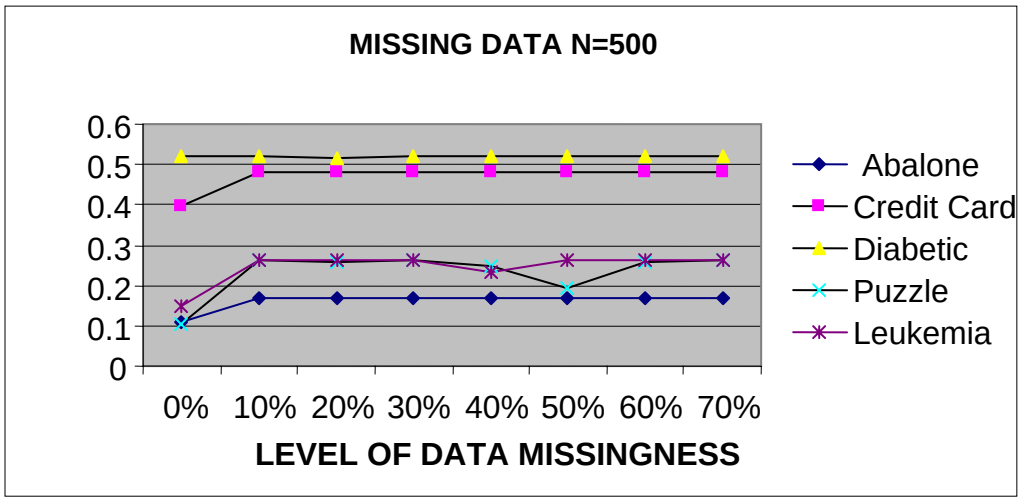
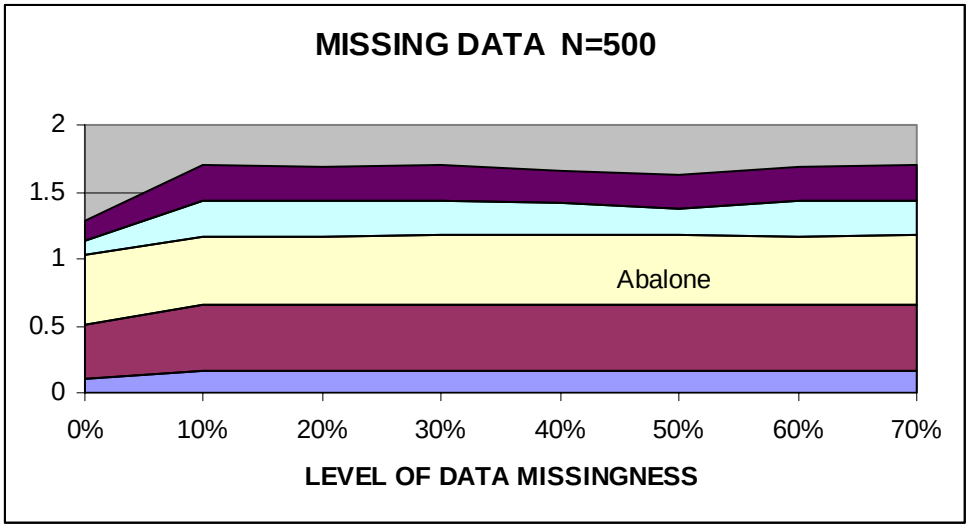


Figure 5.3 500 Instances – Missing Data Area Chart



Figures 5.2 and 5.3 illustrate that in four out of the five models tested (80%), significant degradation occurred immediately upon the original injection of data missingness into the model at the 10% level. It can also be seen that all five of the KDD models failed to significantly degrade (or improve) with increased levels of data missingness (20%-70%).

The only model that did not spike (show a significant increase in the calculated Root Mean Square value) at the 10% level of data missingness was the model that originally contained N=7000 cases. This may be attributed to inheritance factors contained in the original data set prior to case selection to create the smaller data subset of N=500 cases.

An ANOVA test was executed to test if the level of Data Missingness injected into the KDD models using only N=500 cases for ANN training and testing was a significant factor.

The resulting ANOVA testing the level of Data Missingness is shown in Table 5.3:

Table 5.3 ANOVA Test for Level of Data Missingness

ANOVA

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.813433	4	0.203358	162.7526	4.28E-22	2.641465
Within Groups	0.043732	35	0.001249			
Total	0.857165	39				

The ANOVA test indicates that we must accept the null hypothesis that the level of data missingness injected into a KDD model containing less than 1000 cases for training and testing is therefore not a significant factor.

However, it should be noted that the Root Mean Square value (RMS) value spiked significantly at the original introduction of Missing Data (at the 10% level) in four of the five models, with the exception being the model that had the largest original case frequency (N=7000).

Hypothesis 3, Hypothesis 4 and Hypothesis 5 were tested simultaneously.

Hypothesis 3 tested the significance of the level of data missingness in the training and testing of KDD models that utilize Artificial Neural Networks (ANN's) that employ an s-Sigmoid Transfer Function within the Activation Function when the KDD models are from multiple disciplines of study (Business, Science and Health) and contain various case frequencies (N=500, N=1000, N=3500, N=5000, N=7000).

Hypothesis 4 tested the impact of performing Data Imputation on KDD models when Missing Data is encountered, versus performing No Data Imputation (leaving Missing Data in the KDD model for training and testing) on KDD models that contain larger case frequencies (N=3500). Recall that for Hypothesis 2, at the 0.05 level of significance, we accepted the null hypothesis that the level of data missingness is not a significant factor in the calculation of the RMS value(s) in KDD models that contain less than 1000 cases.

Hypothesis 5 tested the performance of Mean Imputation versus Case Deletion performance as the method of Data Imputation in the replacement of Missing Data in KDD models that utilize ANN's that employ an s-Sigmoid Transfer Function within the Activation Function.

More precisely, the Data Mining phase of the Knowledge Discovery process is tested utilizing Neural Network software that employs the s-Sigmoid as its Transfer Function in the Activation Function. Previous studies have investigated missing data treatments applied within the context of data mining (Acuna and Rodriguez, (2004); Batista and Monard, (2003), Razi and Athappily, (2005) and Sehgal, Gondal, and Dooley, (2005)), but few have specifically studied the effects of missing data on the neural network s-Sigmoid function.

An experimental design to explain the impact of the level of data missingness and missing data, and the effect of Data Imputation, on the KDD process was developed.

The experiment in this study is comparable in breadth and depth (i.e., approximately equal in the number of missing data treatments, size of data sets, and number of data sets used) to previous missing data treatment experiments (Acuna and

Rodriguez (2004), Batista and Monard (2003), Razi and Athappily (2005), Sehgal, Gondal, and Dooley (2005)). Analogous to past experiments where missing data treatments were investigated, data sets were employed from the Machine Learning Database Repository at the University of California, Irvine and a standard algorithm was used to generate missing values within each data set (Matlab, 2007).

The Intelligent Data Analyzer (iDA) software product was selected to perform the data mining session (Roiger and Geatz, 2003). A backpropogation Neural Network architecture employing an s-Sigmoid Transfer Function was chosen for this study. The network is trained using the data set(s) with known values for the dependent variable(s). The Root Mean Square (RMS) error (comparison between desired output and computed output) was selected as the metric to be evaluated in determining the performance of each Neural Network model ad was normalized for scaling purposes. Methods of data imputation are employed within the data mining model to attempt to identify how various levels of data missingness within data sets and type of imputation method impact a Data Mining study.

Each data set was initially mined with no missing data, and without altering the standard parameters necessary for data mining utilizing a neural network (the default parameters for learning rate, number of hidden layers and number of epochs were accepted) and obtaining an RMS value.

Data imputation was then performed on the missing values in each KDD model using the Case Deletion and Mean Imputation methods. Each data set was then modified with respect to the amount of missing data within each model. Once again, data imputation was performed on the missing values using the two imputation methods, and

new RMS values calculated. The RMS results were then analyzed using a two factor ANOVA and Tukey's Honestly Significant Difference (HSD) statistic to determine if the percent level of data missing, and/or data imputation method employed, are significant factors.

A two-factor ANOVA test was conducted at the 0.05 significance level. The two factors include level of data missingness and imputation method.

The results of the two-factor ANOVA test are displayed in Table 5.4.

Table 5.4 ANOVA Results

Source	Sum of Squares	df	Mean Square	F	Significance	Partial Eta Squared
Corrected Model	0.0475	8	0.0059	3.354	.029	.691
Intercept	1.1870	1	1.1870	671.055	.000	.982
% Data Missingness	0.0055	6	0.0009	0.522	.782	.207
Imputation Method	0.0419	2	0.0210	11.852	.001	.664
Error	0.0212	12	0.0018			
Total	1.2550	21				
Corrected Total	0.0687	20				

Concerning Hypothesis 3, the ANOVA results at the .05 level of significance indicate that we must accept the null hypothesis, and conclude that the percentage level of data missingness is not a significant factor in the performance of KDD models from multiple disciplines containing a large frequency of cases (N=3500).

In support of the ANOVA and Multiple Comparisons tests, the following graphs illustrate the variation of the Root Mean Square values when various levels of data missingness are injected into each of the KDD models evaluated:

Figure 5.4 Comparison of RMS Values –500 Instances

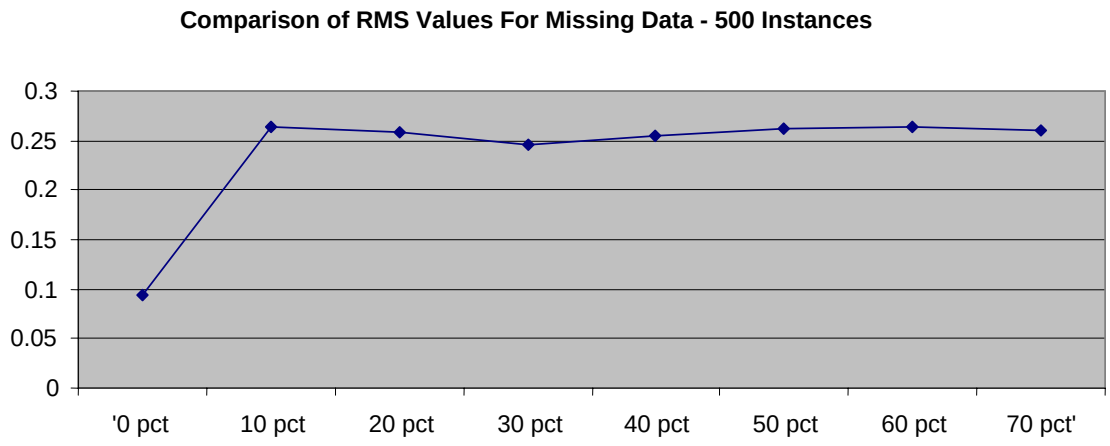


Figure 5.5 Comparison of RMS Values – 1000 Instances

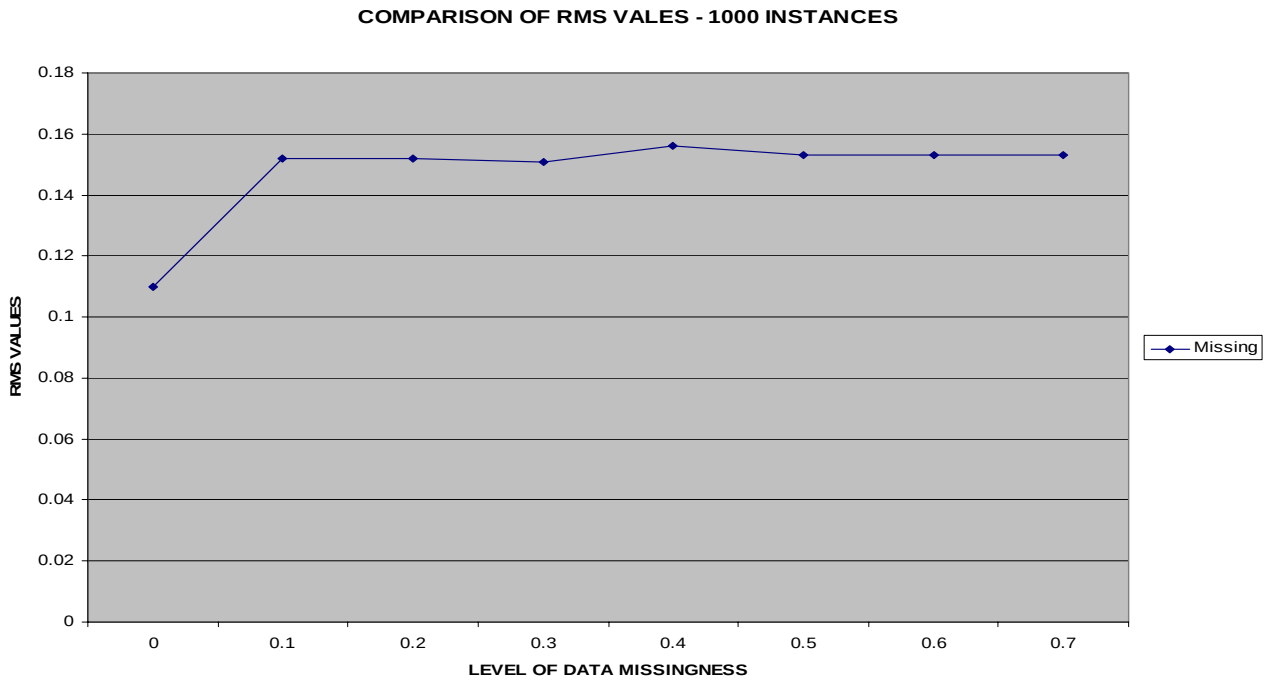


Figure 5.6 Comparison of RMS Values – 3500 Instances

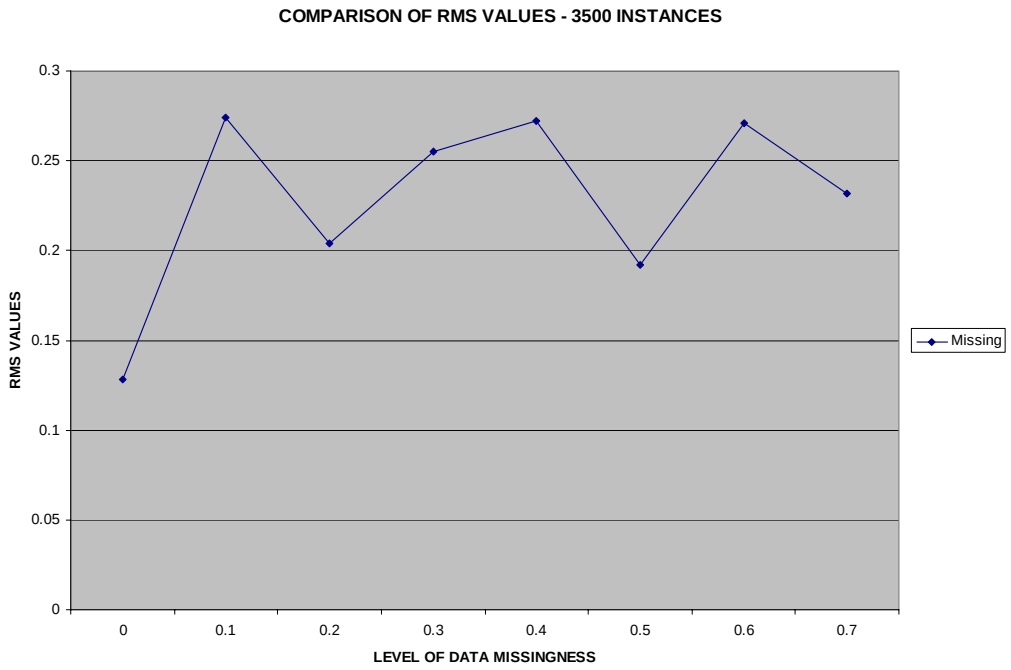


Figure 5.7 Comparison of RMS Values – 5000 Instances

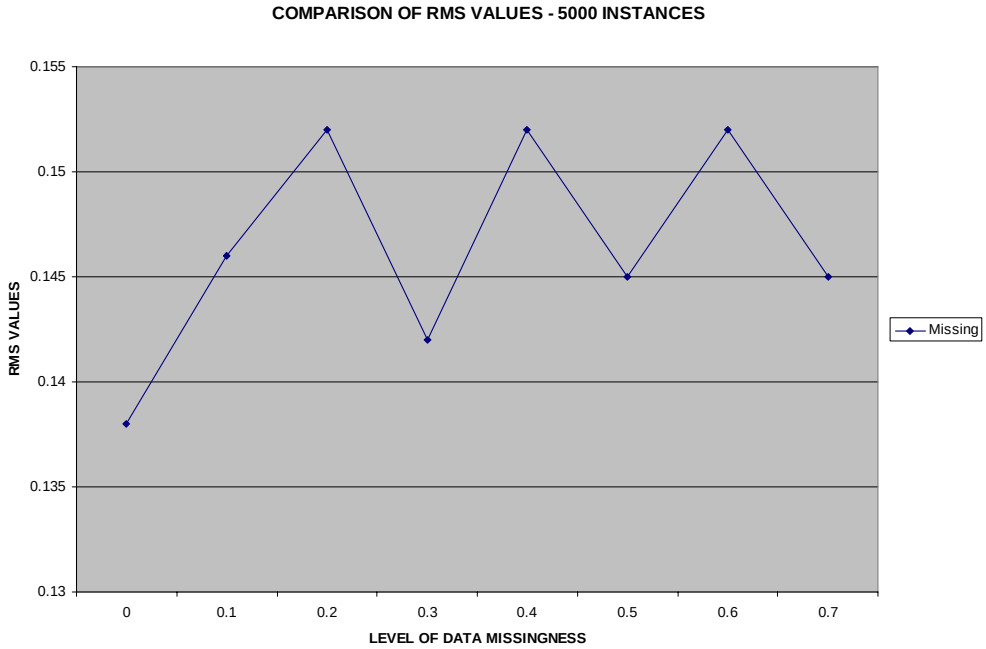
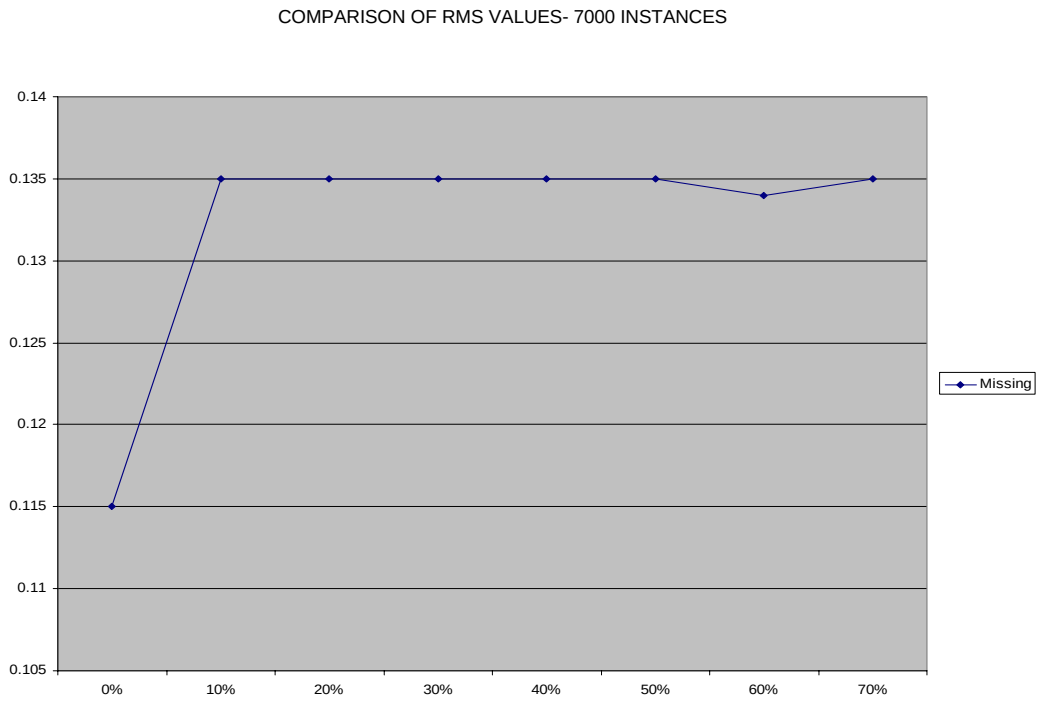


Figure 5.8 Comparison of RMS Values – 7000 Instances



Prior to this research, it was believed that the KDD models would begin to degrade at some point as the level of data missingness was increased in each data set, regardless of the original volume of case frequency. It can be seen that all models tested began showing a significant increase in degradation in the value of the Root Mean Square at the 10% level of data missingness, and in most cases did not significantly increase or decrease as higher levels of data missingness were injected into the model.

However, the two-factor ANOVA test and Tukey's Honestly Significant Test concluded that, at the 0.05 level of significance, the level of data missingness injected into the KDD models did not significantly impact the calculated Root Mean Square (RMS) values in those proposed models.

While the ANAOVA test and Tukey's Test indicated that while there was not a significant difference discovered in the levels of the Root Mean Square (RMS) values when various levels of data missingness were injected into the KDD models, all models illustrated a significant increase in the degradation of the Root Mean Square whenever data missingness is initially injected into the KDD models at the 10% level.

This further illustrates that once a model has been initially introduced to missing data, the model is not significantly altered by the injection of greater amounts of data missingness.

Hypothesis 4, Performing Imputation versus Not Performing Data Imputation, set out to prove if any difference exists between employing a data imputation methodology and "no data imputation being performed" on the missing values in the data sets of the proposed KDD model in the computation of RMS values.

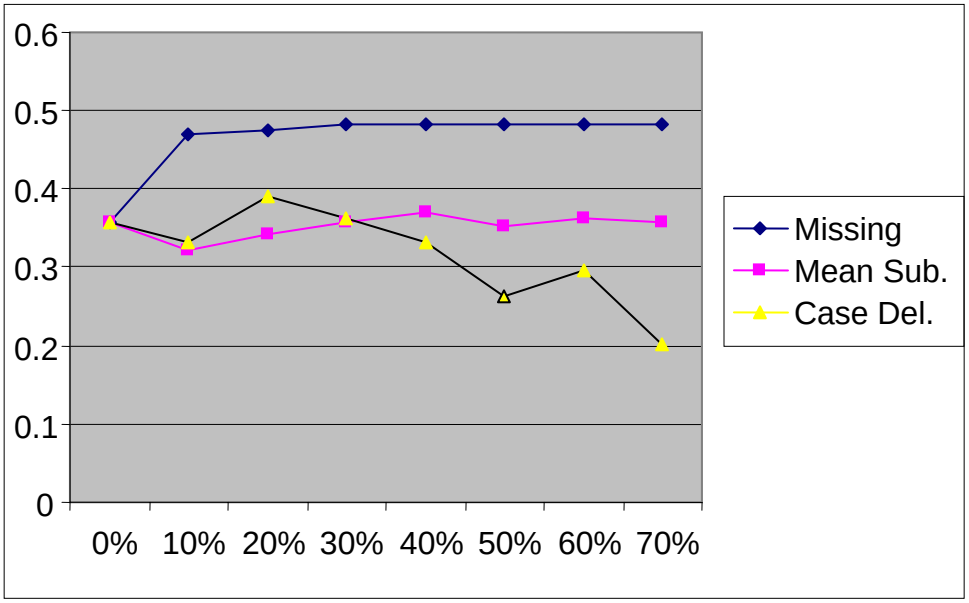
The two-factor ANOVA test displayed above indicates that one of the two factors tested (Data Imputation) is significant at the 0.05 significance level.

Therefore, we fail to accept the null hypothesis that Imputation Method does not have an effect on the calculated Root Mean Square values (RMS), and conclude that performing Data Imputation on KDD models that contain missing values does have an effect on the calculated RMS values in the KDD models tested when the ANN utilizes an s-Sigmoid Transfer Function within the Activation Function.

The following graph further illustrates the comparison of Root Mean Square (RMS) values obtained from the ANN in the various Knowledge Discovery Models when no Imputation Method was employed, and when Missing Data was permitted in the Training and Testing of the KDD models, and after the two most common methods of

data imputation were employed, Case Deletion and Mean Imputation. It can be seen that better (lower) Root Mean Square (RMS) values are obtained in all instances when Data Imputation is performed on the KDD models.

Figure 5.9 Comparison of Root Mean Square Values (RMS)

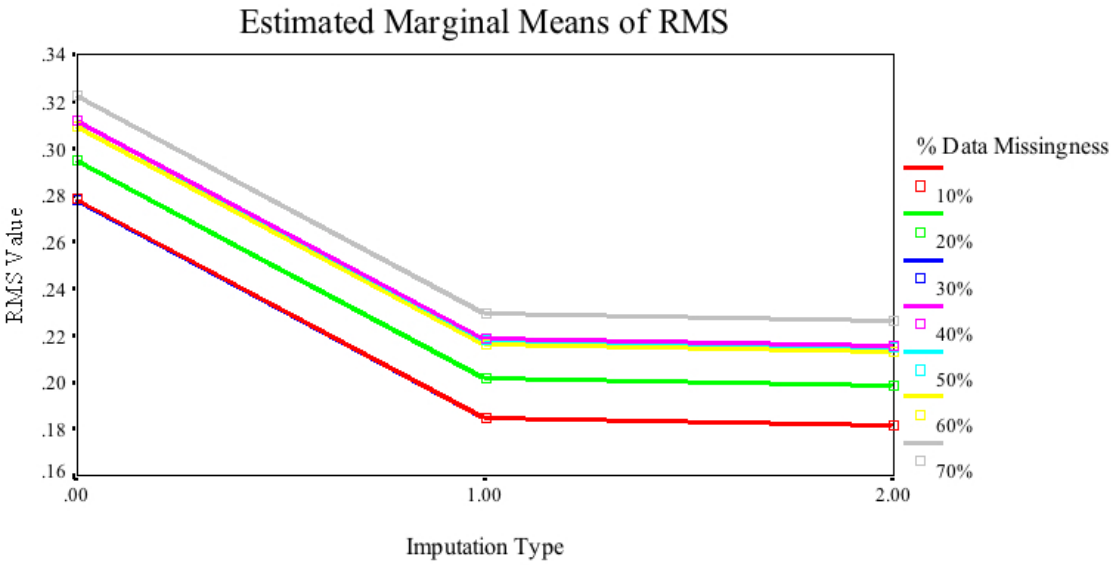


To further reinforce the results of Hypothesis 4, the interaction effect between the level of Data Missingness and Imputation Method was examined by using a marginal means plot.

Figure 5.10 displays the marginal means plot for the level of Data Missingness and Imputation Method. This plot illustrates interaction between the two factors and the level percent of Data Missingness and Imputation Method. Here, the lines of the effect of Imputation Type for all Levels of Data Missingness are parallel to each other. Parallel lines imply that if the level (type) of the Imputation Method is changed, the corresponding change in the Root Mean Square (RMS) value is the same regardless of the value of the percent level of Data Missingness. Similarly, a change in the level of

Data Missingness produce the same change in RMS value, regardless of the level (type) of Imputation Method employed.

Figure 5.10 Estimated Marginal Means of RMS



Marginal Means Plot for percent level of Data Missingness and Imputation Type

Hypothesis 5 tested the significance of Data Imputation Method (Case Deletion and Mean Substitution) performed prior to the training and testing of KDD models containing a large number of cases (N=3500).

Due to the fact that the two-factor ANOVA test displayed above resulted in a rejection of the null hypothesis (that performing Data Imputation was not significant), a test of multiple comparisons was then performed on the type of Data Imputation method (Case Deletion and Mean Substitution). Tukey's Honestly Significant Test was chosen for this test at the 0.05 significance level.

The results from Tukey's test are illustrated below in Table 5.5:

Table 5.5 Results of Tukey's HSD Multiple Comparisons for Imputation Method

Imputation Method	|Tukey's HSD Mean Substitution	Case Deletion
No Imputation	.0931 (*)	.0964 (*)
Mean Substitution	---	.0033

*The mean difference is significant at the .05 level.

The results of Turkey's Honestly Significant Difference Test indicate that there is no significant difference between the type of Data Imputation Method employed (Case Deletion vs. Mean Substitution) on Missing Data in KDD models of various case frequencies (N=500, N=1000, N=3500, N=5000, N=7000) and from multiple disciplines of study.

The following supporting graphs illustrate the variation of the RMS values when Data Imputation is performed on the large KDD models prior to training and testing:

Figure 5.11 Comparison of Imputation Methods - 3500 Instances

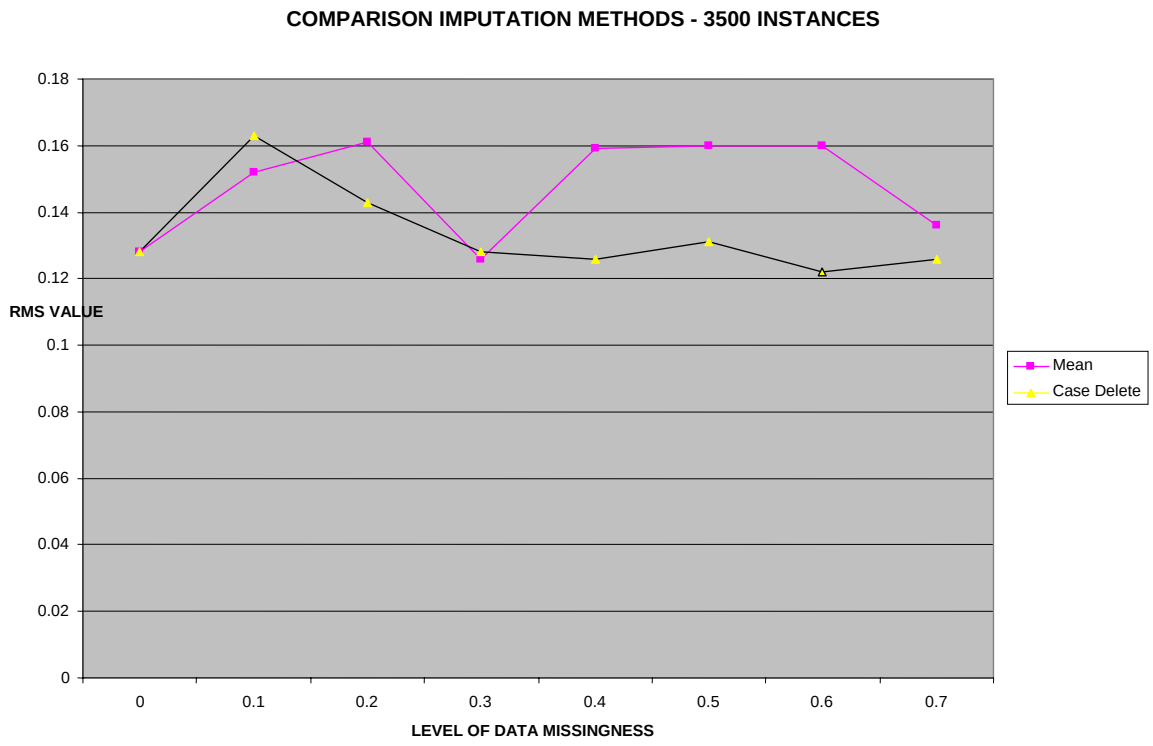


Figure 5.12 Comparison of Imputation Methods – 5000 Instances

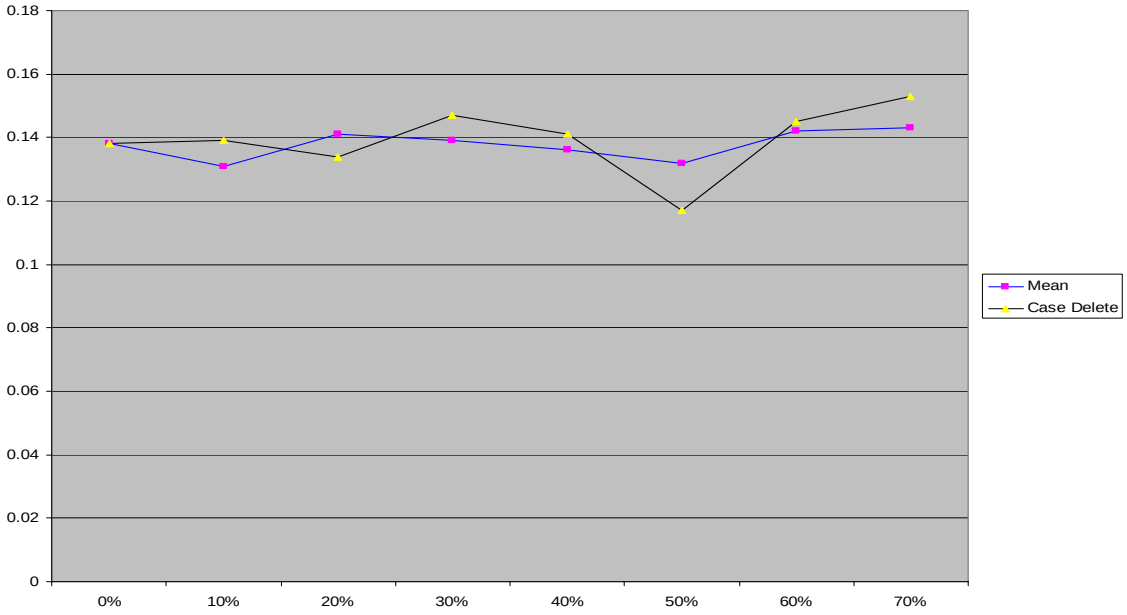
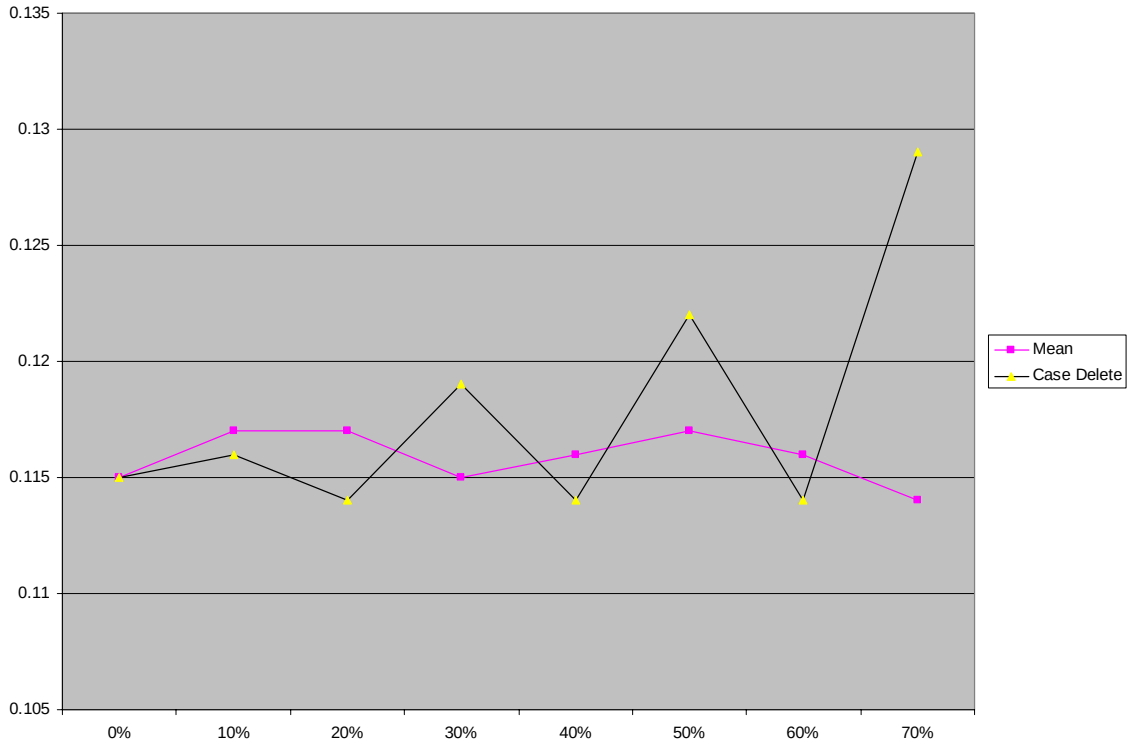


Figure 5.13 Comparison of Imputation Methods – 7000 Instances

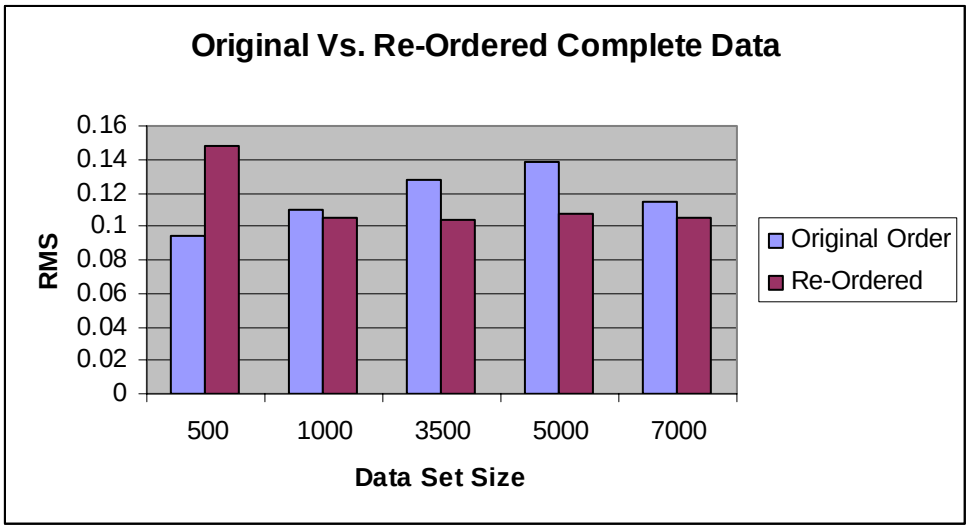


Hypothesis 6 was tested to determine if the Root Mean Square (RMS) value computed by an ANN in a KDD model using an S-Sigmoid Transfer Function is significantly different from a new RMS value obtained after the data instances used in the training and testing of the model has been introduced to an algorithm that re-sequences the data instances prior to re-training and re-testing the KDD model.

Due to factors such as the timeliness of data entry, implementation of indexes, data sorting, original data sequence, data deletion and data aging, data clusters may be formed within the training data for the ANN in KDD. Even if the data is presumed smooth and no clustering is known to exist, this test will determine if data re-sequencing has an effect on the calculated Root Mean Square values in KDD models of various case frequencies.

The following graph illustrates a comparison between the differences in RMS values calculated by the ANN in using an S-Sigmoid Transfer function in a KDD using data instances in their original order and after the instances have been randomly re-sequenced.

Figure 5.14 RMS Values for Original Data Sequence vs. Re-Ordered Data



It can be seen from Figure 5.12 that the data set containing 500 instances showed a considerable difference in the calculated RMS value, with a lower (better) RMS value than that obtained when the instances were reordered. However, in all other cases a slightly lower (better) RMS value was obtained when the instances were randomly re-sequenced prior to training and testing the ANN for the KDD.

A T-Test was then conducted to test the means of the Original Order vs. Re-sequenced data sets. The results shown in Table 5.6 indicate that the difference in means is not statistically significant at the .05 level.

Table 5.6 Test: Paired Two Sample for Means

	Variable 1	*Variable 2*
Mean	0.117	0.114
Variance	0.000286	0.000363
Observations	5	5
Pearson Correlation	-0.726514649	
Hypothesized Mean Difference	0	
Df	4	
t Stat	0.200625141	
P(T<=t) one-tail	0.425389857	
t Critical one-tail	2.131846782	
P(T<=t) two-tail	0.850779713	
t Critical two-tail	2.776445105	

In summation, this research indicates that the re-sequencing of cases in a KDD model has no statistically significant impact on the calculation of Root Mean Square (RMS) Values by an Artificial Neural Network (ANN) utilizing an S-Sigmoid Transfer Function as a component of the ANN's Activation Function, regardless of the volume of Case Frequency used for ANN training and testing for the KDD.

Hypothesis 7 tested the significance of re-sequencing ("Smoothing") the Data Missingness in a KDD model on the computed Root Mean Square (RMS) value for KDD models containing various volumes of case frequencies, both before and after the cases in each KDD model were re-sequenced prior to being tested with the S-sigmoid transfer function. Each test was conducted twice.

The first test computed a Root Mean Square (RMS) value for each KDD model at each level of Data Missingness (10%-70%), and then re-tested after the data instances in the data set had been re-sequenced.

Data sets with case frequencies of 500, 1000, 3500, 5000 and 7000 were all injected with missing data at the 10%, 20%, 30%, 40%, 50%, 60% and 70% level of data missingness, and the KDD models trained and tested. All KDD models were then re-sequenced, then re-trained and tested again.

The following graphs illustrate the variation in the calculation of RMS values when various levels of data missingness are injected into data sets containing a variable number of data instances (N=500, N=1000, N=3500, N=5000 and N= 7000), re-sequenced and re-tested.

Figure 5.15 Original vs. Re-Sequenced Data – 500 Instances

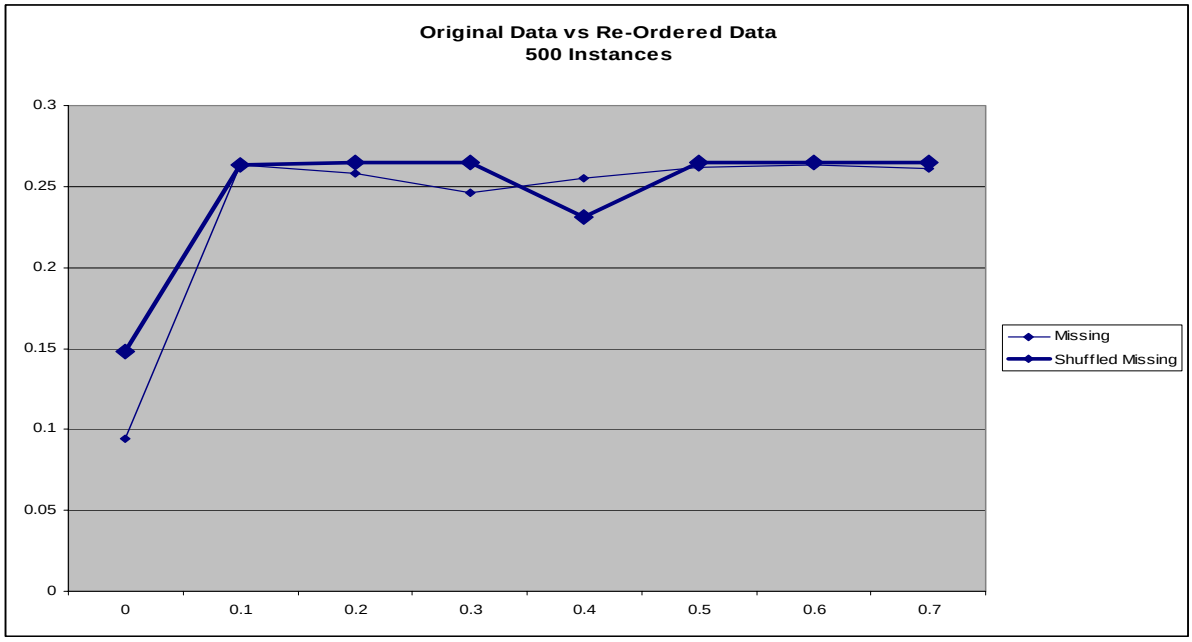


Figure 5.16 Original vs. Re-Sequenced Data – 1000 Instances

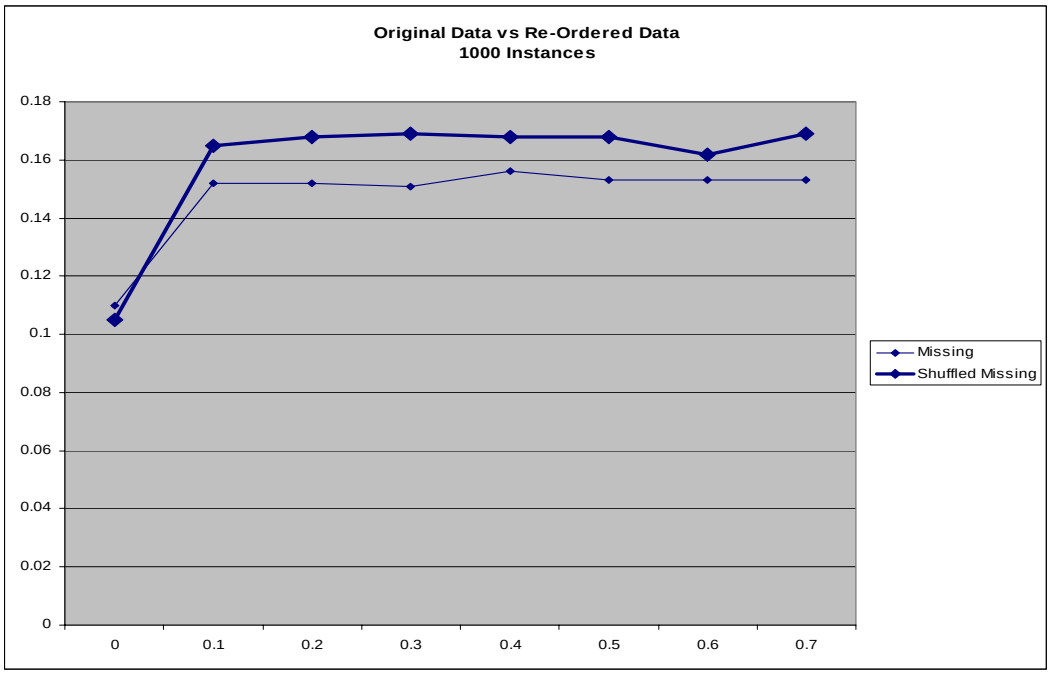


Figure 5.17 Original vs. Re-Ordered Data – 3500 Instances

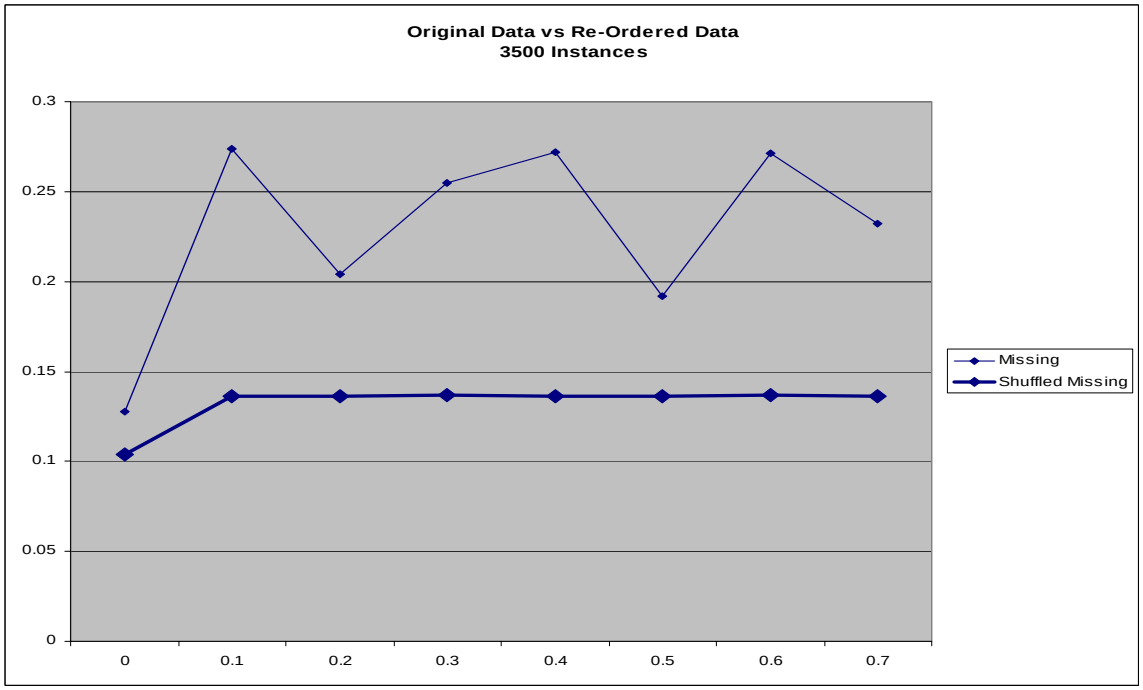


Figure 5.18 Original Data vs. Re-Ordered Data – 5000 Instances

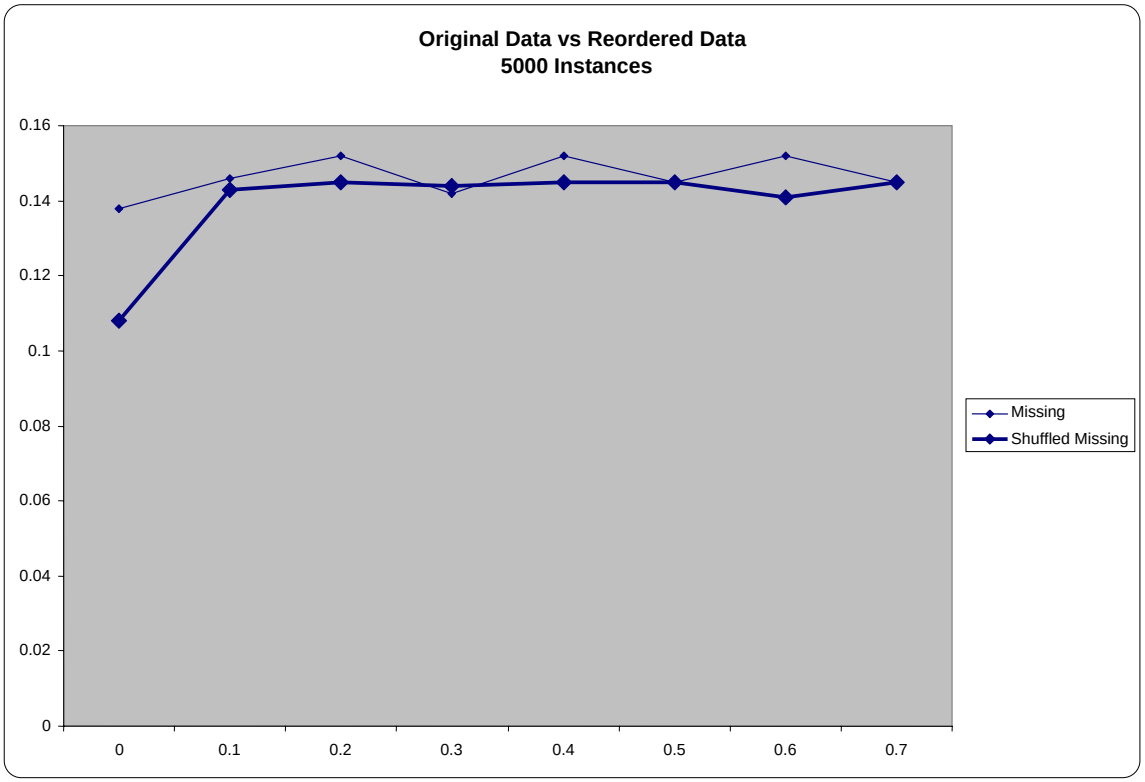
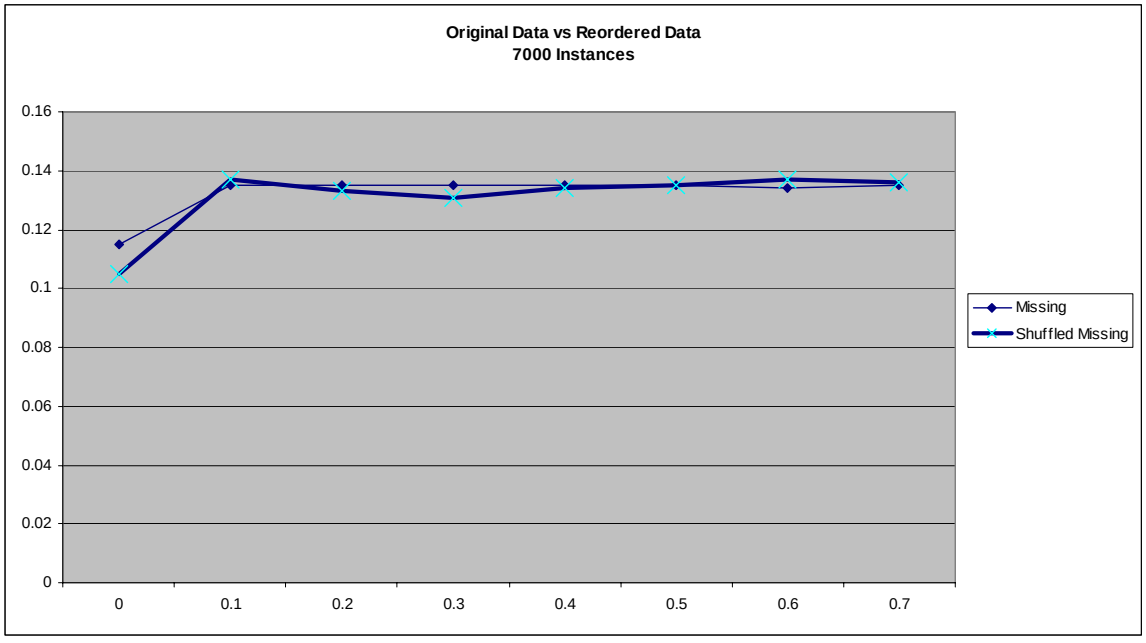


Figure 5.19 Original Data vs. Re-Ordered Data – 7000 Instances



It can be seen from the charts above that when the data set is small (N=500, N=1000), the RMS values calculated are consistently higher (worse) at all levels of data missingness. However, when the Case Frequency is higher (N=3500, N=5000, N=7000), the calculated RMS values are lower (better) at nearly all levels of data missingness.

At N=500 and at the 10% level of data missingness, the RMS values were equal for both the original data sequence and when the data was randomly re-sequenced prior to training and testing. Only at the 40% level did the re-sequencing of the training and testing data result in a Root Mean Square (RMS) value less than the RMS value obtained from the data being in its original sequence.

When N=1000, re-sequencing the data resulted in higher (worse) calculated RMS values at all levels of data missingness, from 0% through 70%.

When data missingness was injected into a data set containing 3500 instances, the impact was quite different. This data set, in it's original sequence, resulted in very erratic RMS calculations as greater levels of data missingness were injected into the model.

The best RMS values were obtained when the level of missingness was set at 20% and 50%. However, when the KDD model was re-trained and re-tested following the re-sequencing of the data instances, lower and more consistent RMS values were obtained.

A similar pattern was discovered when the KDD models with N=5000 and N=7000 were run. The RMS values calculated by the models that contained re-sequenced data resulted with better performance (lower RMS values) than when the models were run with their original data sequences.

An ANOVA test was used to test the significance of data order (original data sequencing vs. random re-sequencing) on the calculation of the Root mean Square (RMS) Value for a KDD at the 0.05 significance level.

The results of the ANOVA test are displayed in Table 5.7.

Table 5.7 ANOVA Test for RMS Values Original Data Sequence vs. Re-Sequenced

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	40	7.141	0.178525	0.003218
Column 2	40	6.463	0.161575	0.002393

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.005746	1	0.005746	2.048018	0.156399	3.963472
Within Groups	0.218842	78	0.002806			
Total	0.224588	79				

The results of the ANOVA test indicate that the re-sequencing of data instances containing Missing Data in KDD models using an ANN as the Data Mining Algorithm had no statistically significant impact on the calculation of the Root Mean Square (RMS) Value for those models.

Hypothesis 8 tested the impact of variable levels of Data Missingness on the calculation of RMS values in a KDD model using an ANN that utilized an S-sigmoid transfer function when the dataset used in the model for training and testing was small (N=500), and when the imputation methods of Mean Substitution and Case Deletion are employed on the Missing Data.

A Root Mean Square (RMS) value was first calculated using the KDD model originally containing N=500 instances only. Data Missingness was then injected into the model at the 10%, 20%, 30%, 40%, 50%, 60% and 70% levels and the KDD models re-trained and re-tested.

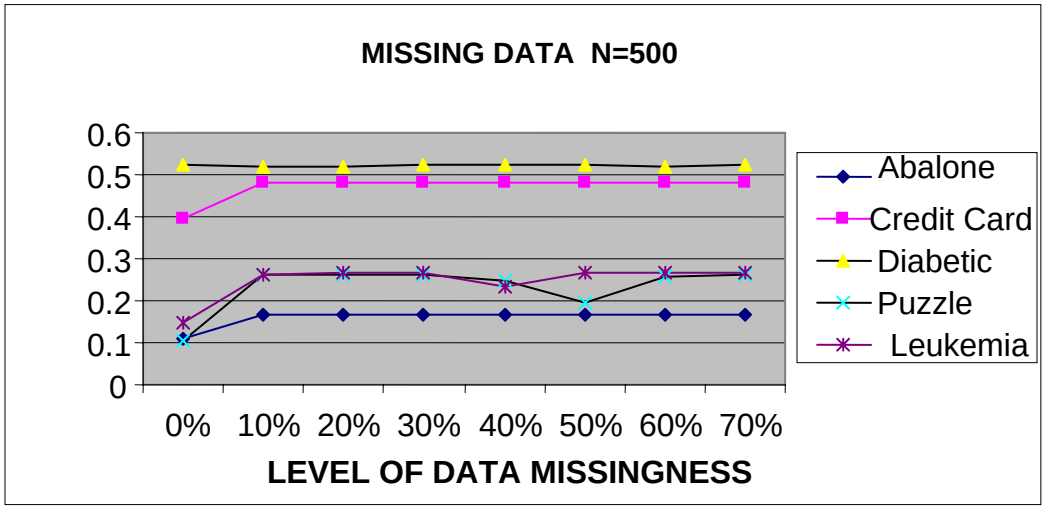
Similarly, 500 data instances from each of the other four data sets from other disciplines of study were randomly selected for the re-training and re-testing of those KDD models. Missing Data was then injected into each of those four new KDD models with N=500 data instances at the 10%, 20%, 30%, 40%, 50%, 60% and 70% levels. RMS values were then calculated for each new KDD model.

The data imputation method of Case Deletion was then performed on the datasets and the ANN's in the KDD models were re-trained and re-tested. RMS values were again calculated using the modified models.

In a like manner, the data imputation method of Mean Imputation was performed on the new KDD models with Missing data (10%-70% Missing Data). The KDD models were then re-trained and re-tested following Mean Imputation. New RMS values were again calculated from these modified models.

The following charts illustrate the results of the RMS calculations both before and after the data imputation methods of Case Deletion and Mean Imputation were employed in the models at N=500, and following the injection of Missing Data:

Figure 5.20 RMS Values N=500



Prior to running the KDD models at N=500, it was expected that an erratic pattern may emerge in the calculation of RMS values. Especially in regard to Case Deletion data imputation, the extremely small number of data instances used to train the ANN would expectedly result in inconsistent RMS calculations. However, even at N=500, all but one of the KDD models illustrated similar behavior as that of KDD models utilizing a greater number of data instances in the training and testing of the ANN.

In all of the KDD models tested, data imputation resulted in lower (better) RMS values being obtained in a majority of the models at various levels of noise injection than when the missing data (noise) was left in the model for training and testing.

Only one KDD model (Abalone data) resulted in higher (worse) RMS values at the 10%, 40% and 60% levels of noise injection when Case Deletion was employed as the data imputation method when compared to the KDD RMS calculations when missing data (noise) was left in the model for training and testing.

Although the RMS values calculated were no better (lower) in the KDD model utilizing Diabetic data when Mean Imputation was employed, they were not worse (higher) either.

In the other three KDD models, both methods of data imputation resulted in lower (better) RMS calculations than when missing data (noise) was left in the model for training and testing.

Surprisingly, this is not inconsistent with similar tests conducted with larger frequencies of data instances in other KDD models.

An ANOVA test was performed on the level of Data Missingness in the five KDD models with N=500 instances.

The results of the test are displayed in Table 5.8.

Table 5.8 ANOVA Results for Data Missingness

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
0	5	1.281	0.2562	0.036596
0.1	5	1.697	0.3394	0.023397
0.2	5	1.952	0.3904	0.026424
0.3	5	1.703	0.3406	0.023683
0.4	5	1.653	0.3306	0.025679
0.5	5	1.632	0.3264	0.027301
0.6	5	1.692	0.3384	0.023375
0.7	5	1.702	0.3404	0.023722

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.047123	7	0.006732	0.256238	0.966285	2.312741
Within Groups	0.840705	32	0.026272			
Total	0.887828	39				

The ANOVA results indicate that there is no significant difference in the Root Mean Square Values (RMS) calculated for the five KDD models from various disciplines when tested with N=500 randomly selected data instances containing Missing Data injected at the 10%, 20%, 30%, 40%, 50%, 60% and 70% levels.

Hypothesis 9 tested the impact of the Regression Imputation Method as compared to the data Imputation Methods of Mean Substitution and Case Deletion on KDD models that were injected with various increasing levels of Data Missingness, and the calculation of Root Mean Square (RMS) values in those KDD models (using an Artificial Neural network (ANN) and utilizing an S-sigmoid transfer function when the dataset used in the model for training and testing was considered to be large (N=1000).

A KDD model that originally contained 1000 data instances was selected for this test, and a Root Mean Square (RMS) value calculated for the model.

Data Missingness was then randomly injected into the independent variables in the model at increasing levels of Data Missingness, 10%, 20%, 30%, 40%, 50%, 60% and 70%. Root mean Square values were calculated for the new KDD models at each of the specified levels of Data Missingness.

Regression Imputation, Mean Substitution and Case Deletion were then performed on the Missing Data in each of these new KDD models. All KDD models were re-trained, re-tested and new Root Mean Square (RMS) values calculated.

The results of all tests were tabulated and are presented in Table 5.9.

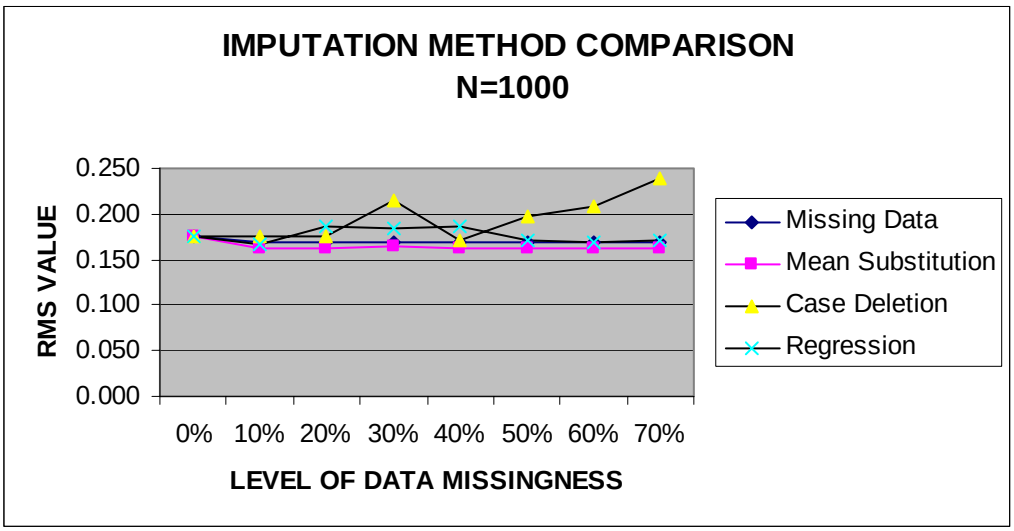
Table 5.9 Root Mean Square Statistics, N=1000

N=1000

	% of Missing Data							
Imputation Method	**0%**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**
No Imputation	**0.175**	**0.168**	**0.169**	**0.169**	**0.169**	**0.168**	**0.169**	**0.169**
Case Deletion	**0.175**	**0.176**	**0.176**	**0.216**	**0.172**	**0.197**	**0.209**	**0.238**
Mean Substitution	**0.175**	**0.163**	**0.162**	**0.164**	**0.162**	**0.162**	**0.163**	**0.163**
Regression Imputation	**0.175**	**0.167**	**0.186**	**0.185**	**0.186**	**0.172**	**0.169**	**0.171**

Figure 5.21 illustrates a comparison of Root Mean Square values after the methods of Regression Imputation, Mean Substitution and Case Deletion have been performed on a KDD model containing 1000 data instances, and after having been injected with increasing levels of Data Missingness (up to 70%):

Figure 5.21 Imputation Method Comparison



It can be seen from the figure above that the Mean Substitution imputation method was the only method of imputation that resulted in Root Mean Square (RMS) values that were lower (better) than performing no data imputation at all (with nearly identical RMS values when compared to no imputation being performed, only a very slight positive variance) at all levels of Data Missingness.

Also, performing Regression Imputation prior to the re-training and re-testing of the KDD models following the injection of Data Missingness actually resulted in lower (better) Root Mean Square (RMS) values than the Case Deletion method of Data Imputation. This is most likely due to the fact that the number of remaining data instances used for training and testing the KDD models following Case Deletion dipped blow the 1000 case level at the 10%-70% levels of Data Missingness.

Figure 5.21 indicates that as the level of Data Missingness increased, and as N became smaller following Case Deletion, the Root Mean Square (RMS) values grew increasingly large (worse).

A two-factor Analysis of Variance (ANOVA) test was then performed at the .05 level of significance, testing the level of Data Missingness and the type of Imputation Method employed (prior to training and testing the KDD models) when calculating Root mean Square (RMS) values for the KDD models.

Table 5.91 displays the ANOVA results of Root Mean Square values for Data Missingness and Data Imputation Method in KDD models with 1000 data instances:

Table 5.9.1 ANOVA Results for Level of Data Missingness and Imputation Method

N=1000

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Missing Data	8	1.356	0.1695	5.14E-06
Mean Substitution	8	1.314	0.16425	1.94E-05
Case Deletion	8	1.559	0.194875	0.000592
Regression	8	1.411	0.176375	6.46E-05
0	4	0.7	0.175	0
0.1	4	0.674	0.1685	2.97E-05
0.2	4	0.693	0.17325	0.000105
0.3	4	0.734	0.1835	0.00055
0.4	4	0.689	0.17225	0.000102
0.5	4	0.699	0.17475	0.000237
0.6	4	0.71	0.1775	0.000449
0.7	4	0.741	0.18525	0.001248

ANOVA

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Rows	0.004292	3	0.001431	7.766367	**0.001124**	3.072467
Columns	0.000896	7	0.000128	0.694888	**0.675797**	2.487578
Error	0.003868	21	0.000184			
Total	0.009056	31				

The two-factor ANOVA test shows that the null hypothesis stating that the level of Data Missingness is not significant at the .05 level of significance cannot be rejected.

However, the null hypothesis that the type of data imputation method performed on the KDD's tested is rejected at the .05 level of significance, and it is determined that at least one of the imputation methods is different from the others.

At this point it was necessary to perform an ad hoc test to determine which of the imputation methods performed differently from the others.

Tukey's Honestly Significant Different Test was selected as the ad hoc test to be used.

Table 5.92 displays the results of Tukey's Honestly Significant Different Test (HSD) on the type of Data Imputation Method employed:

Table 5.9.2 Tests of Between-Subjects Effects

Dependent Variable: RMS

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.008(a)	11	.001	3.608	.003
Intercept	1.195	1	1.195	6063.947	.000
IMPUTATION	.006	4	.002	7.790	.000
DATAMISS	.002	7	.000	1.217	.326
Error	.006	28	.000		
Total	1.208	40			
Corrected Total	.013	39			

a R Squared = .586 (Adjusted R Squared = .424)

Multiple Comparisons

Dependent Variable: RMS
Tukey HSD

(I) IMPUTATION	(J) IMPUTATION	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
.00	1.00	.0254(*)	.00702	.009	.0049	.0458
	2.00	.0306(*)	.00702	.001	.0102	.0511
	3.00	.0185	.00702	.091	-.0020	.0390
	4.00	.0356(*)	.00702	.000	.0152	.0561
1.00	.00	-.0254(*)	.00702	.009	-.0458	-.0049
	2.00	.0053	.00702	.943	-.0152	.0257
	3.00	-.0069	.00702	.862	-.0273	.0136
	4.00	.0103	.00702	.596	-.0102	.0307
2.00	.00	-.0306(*)	.00702	.001	-.0511	-.0102
	1.00	-.0053	.00702	.943	-.0257	.0152
	3.00	-.0121	.00702	.434	-.0326	.0083
	4.00	.0050	.00702	.952	-.0155	.0255
3.00	.00	-.0185	.00702	.091	-.0390	.0020
	1.00	.0069	.00702	.862	-.0136	.0273
	2.00	.0121	.00702	.434	-.0083	.0326
	4.00	.0171	.00702	.134	-.0033	.0376
4.00	.00	-.0356(*)	.00702	.000	-.0561	-.0152
	1.00	-.0103	.00702	.596	-.0307	.0102
	2.00	-.0050	.00702	.952	-.0255	.0155
	3.00	-.0171	.00702	.134	-.0376	.0033

Based on observed means.
* The mean difference is significant at the .05 level.

RMS

Tukey HSD

IMPUTATION	N	Subset 1	Subset 2
4.00	8	.1593	
2.00	8	.1643	
1.00	8	.1695	
3.00	8	.1764	.1764
.00	8		.1949
Sig.		.134	.091

Means for groups in homogeneous subsets are displayed.
Based on Type III Sum of Squares
The error term is Mean Square(Error) = .000.

a Uses Harmonic Mean Sample Size = 8.000.
b Alpha = .05.

Multiple Comparisons

Dependent Variable: RMS
Tukey HSD

(I) DATAMISS	(J) DATAMISS	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
.00	1.00	.0112	.00888	.905	-.0178	.0402
	2.00	.0008	.00888	1.000	-.0282	.0298
	3.00	-.0074	.00888	.989	-.0364	.0216
	4.00	.0112	.00888	.905	-.0178	.0402
	5.00	.0068	.00888	.994	-.0222	.0358
	6.00	-.0010	.00888	1.000	-.0300	.0280
	7.00	-.0044	.00888	1.000	-.0334	.0246
1.00	.00	-.0112	.00888	.905	-.0402	.0178
	2.00	-.0104	.00888	.934	-.0394	.0186
	3.00	-.0186	.00888	.442	-.0476	.0104
	4.00	.0000	.00888	1.000	-.0290	.0290
	5.00	-.0044	.00888	1.000	-.0334	.0246
	6.00	-.0122	.00888	.861	-.0412	.0168
	7.00	-.0156	.00888	.652	-.0446	.0134
2.00	.00	-.0008	.00888	1.000	-.0298	.0282
	1.00	.0104	.00888	.934	-.0186	.0394
	3.00	-.0082	.00888	.981	-.0372	.0208
	4.00	.0104	.00888	.934	-.0186	.0394
	5.00	.0060	.00888	.997	-.0230	.0350
	6.00	-.0018	.00888	1.000	-.0308	.0272
	7.00	-.0052	.00888	.999	-.0342	.0238
3.00	.00	.0074	.00888	.989	-.0216	.0364
	1.00	.0186	.00888	.442	-.0104	.0476
	2.00	.0082	.00888	.981	-.0208	.0372
	4.00	.0186	.00888	.442	-.0104	.0476
	5.00	.0142	.00888	.747	-.0148	.0432
	6.00	.0064	.00888	.996	-.0226	.0354
	7.00	.0030	.00888	1.000	-.0260	.0320
4.00	.00	-.0112	.00888	.905	-.0402	.0178
	1.00	.0000	.00888	1.000	-.0290	.0290
	2.00	-.0104	.00888	.934	-.0394	.0186
	3.00	-.0186	.00888	.442	-.0476	.0104
	5.00	-.0044	.00888	1.000	-.0334	.0246
	6.00	-.0122	.00888	.861	-.0412	.0168
	7.00	-.0156	.00888	.652	-.0446	.0134
5.00	.00	-.0068	.00888	.994	-.0358	.0222

	1.00	.0044	.00888	1.000	-.0246	.0334
	2.00	-.0060	.00888	.997	-.0350	.0230
	3.00	-.0142	.00888	.747	-.0432	.0148
	4.00	.0044	.00888	1.000	-.0246	.0334
	6.00	-.0078	.00888	.986	-.0368	.0212
	7.00	-.0112	.00888	.905	-.0402	.0178
6.00	.00	.0010	.00888	1.000	-.0280	.0300
	1.00	.0122	.00888	.861	-.0168	.0412
	2.00	.0018	.00888	1.000	-.0272	.0308
	3.00	-.0064	.00888	.996	-.0354	.0226
	4.00	.0122	.00888	.861	-.0168	.0412
	5.00	.0078	.00888	.986	-.0212	.0368
	7.00	-.0034	.00888	1.000	-.0324	.0256
7.00	.00	.0044	.00888	1.000	-.0246	.0334
	1.00	.0156	.00888	.652	-.0134	.0446
	2.00	.0052	.00888	.999	-.0238	.0342
	3.00	-.0030	.00888	1.000	-.0320	.0260
	4.00	.0156	.00888	.652	-.0134	.0446
	5.00	.0112	.00888	.905	-.0178	.0402
	6.00	.0034	.00888	1.000	-.0256	.0324

Based on observed means.

The results of Tukey's Honestly Significant Different Test (HSD) indicate that although the factor of the Level of Data Missingness is not significant in the KDD models tested (as previously enforced by the ANOVA test), the test does in fact indicate that performing some method of Data Imputation (as opposed to not performing any type of Data Imputation) is significant at the .05 level.

Further, Tukey's HSD Test illustrates that although only one of the Data Imputation methods tested in this research is not statistically significant at the .05 level, and that all Imputation Methods are statistically significant at the .10 level of significance.

Therefore, it can be seen from the ANOVA and Tukey's Honestly Significant Difference Test, that when the volume of Case Frequency is considered to be large (N=1000), the level of Data Missingness does not significantly impact the calculation of

the Root Mean Square Value by a KDD model that utilizes an Artificial Neural Network employing an S-Sigmoid Transfer Function as it's Data Mining Algorithm.

However, the employment of some type of Data Imputation Method is significant, and all but one of the Imputation Methods is significant at the .05 level of significance and all Imputation Methods are significant at the .10 level of significance. In summation, KDD models tested with a Case Frequency Volume of N=1000, the imputation method of Regression Imputation does not result in better (lower) Root Mean Square (RMS) values than the imputation methods of Mean Substitution and Case Deletion.

Hypothesis 10 tested the impact of performing Multiple Imputation on missing values in KDD models that were injected with various increasing levels of Data Missingness, and the calculation of Root Mean Square (RMS) values in those KDD models (using an Artificial Neural network (ANN) and utilizing an S-sigmoid transfer function) when the dataset used in the KDD model for training and testing was considered to be large (N=1000).

The Multiple Imputation Method employed consisted of randomly imputing 50% of the Missing Values using Regression Imputation and the remaining 50% of the Missing Values utilizing Mean Substitution.

The KDD model originally consisting of 1000 data instances was selected for this test, and a Root Mean Square (RMS) value calculated for the model.

Random Data Missingness was then injected into the independent variables to be used for the Regression Imputation in the model at increasing levels of Data Missingness, 10%, 20%, 30%, 40%, 50%, 60% and 70%.

Root mean Square values were calculated for each new KDD model at each of the specified levels of Data Missingness.

Multiple Regression Imputation (utilizing a hybrid method of data imputation combining Regression Imputation and Mean Substitution) was then randomly performed on the Missing Data Values in each of the new KDD models.

Fifty per cent of these Missing Values were imputed utilizing Regression Imputation, and the remaining 50% were imputed using Mean Substitution. All of the KDD models that had been injected with increasing levels of Data Missingness were processed in like manner.

All of the KDD models were re-trained and re-tested, and new Root Mean Square (RMS) values were calculated.

The results of all tests are displayed in Table 5.93:

Table 5.9.3 Root Mean Square Statistics

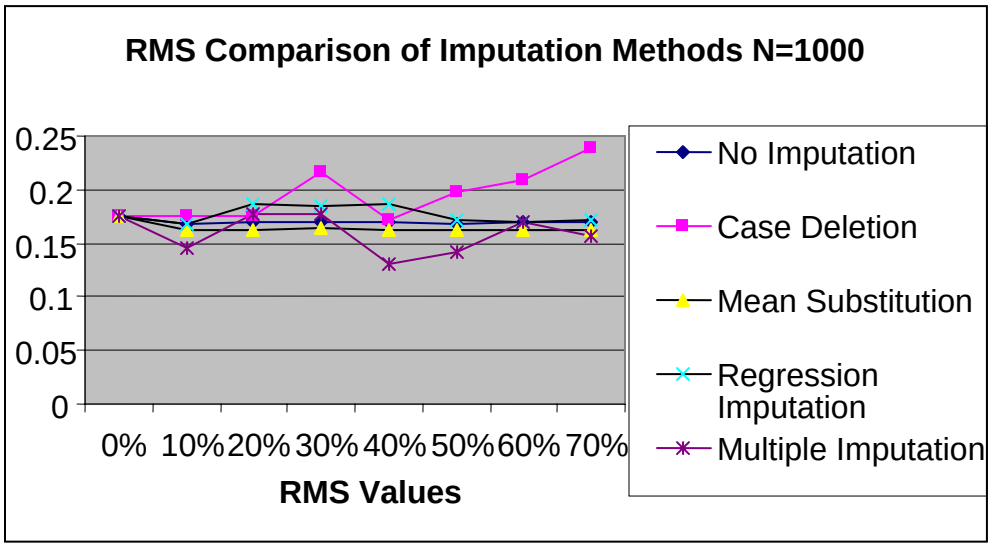
N=1000

	% of Missing Data							
Imputation Method	**0%**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**
No Imputation	**0.175**	**0.168**	**0.169**	**0.169**	**0.169**	**0.168**	**0.169**	**0.169**
Case Deletion	**0.175**	**0.176**	**0.176**	**0.216**	**0.172**	**0.197**	**0.209**	**0.238**
Mean Substitution	**0.175**	**0.163**	**0.162**	**0.164**	**0.162**	**0.162**	**0.163**	**0.163**
Regression Imputation	**0.175**	**0.167**	**0.186**	**0.185**	**0.186**	**0.172**	**0.169**	**0.171**
Multiple Imputation	**0.175**	**0.145**	**0.178**	**0.178**	**0.130**	**0.142**	**0.170**	**0.156**

Figure 5.22 illustrates a comparison of how Multiple Imputation compares in the performance in the calculation of Root Mean Square (RMS) values in KDD models against the same models when no Data Imputation method is performed and when three

other types of Data Imputation (Regression Imputation, Mean Substitution and Case Deletion) are utilized:

Figure 5.22 Multiple Imputation Method Comparison



Interestingly, the RMS values calculated for the KDD's by utilizing the Multiple Imputation Method (combining Regression Imputation and Mean Substitution) to impute Missing Data, were the lowest (best) in four of the KDD models tested when Data Missingness was injected into the original model.

When Data Missingness was first injected into the model at the 10% level, and also at the 40%, 50% and again at the 70% levels, Multiple Imputation resulted in the lowest (best) Root Mean Square (RMS) values.

However, when Mean Substitution alone was implemented as the Data Imputation method, the tests resulted in lower (better) Root Mean Square (RMS) values than when Multiple Imputation (combining Regression Imputation and Mean Substitution) at the

20%, 30% and 60% levels of Data Missingness, when the model contained 1000 data instances.

A two-factor ANOVA test was performed on the calculated Root Mean Square (RMS) values for KDD models containing 1000 data instances when Data Missingness was injected at the 10%, 20%, 30%, 40%, 50%, 60% and 70% levels, and when no Data Imputation was performed on the Missing Values and when the Imputation Methods of Multiple Imputation, Regression Imputation, Mean Substitution and Case Deletion were performed on the models prior to re-training, Re-testing and calculating new Root mean Square (RMS) values for the KDD models.

The results of the two-factor ANOVA test are displayed in Table 5.94:

Table 5.9.4 ANOVA Test for Level of Data Missingness and Imputation Method

N=1000

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
No Imputation	8	1.356	0.1695	5.14E-06
Case Deletion	8	1.559	0.194875	0.000592
Mean Substitution	8	1.314	0.16425	1.94E-05
Regression Imputation	8	1.411	0.176375	6.46E-05
Multiple Imputation	8	1.274	0.15925	0.000348
0	5	0.875	0.175	0
0.1	5	0.819	0.1638	0.000133
0.2	5	0.871	0.1742	8.32E-05
0.3	5	0.912	0.1824	0.000418
0.4	5	0.819	0.1638	0.000433
0.5	5	0.841	0.1682	0.000392
0.6	5	0.88	0.176	0.000348
0.7	5	0.897	0.1794	0.001107

ANOVA

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Rows	0.006141	4	0.001535	7.790414	**0.000237**	2.714076
Columns	0.00168	7	0.00024	1.217415	**0.326199**	2.35926
Error	0.005518	28	0.000197			
Total	0.013339	39				

The two-factor ANOVA test shows that the null hypothesis that the level of Data Missingness is not significant at the .05 level of significance cannot be rejected.

However, the null hypothesis that the type of data imputation method performed on the KDD's tested is rejected at the .05 level of significance. It is determined that at least one of the imputation methods is different from the others.

At this point it was necessary to perform an ad hoc test to determine which of the imputation methods performed differently from the others. Tukey's Honestly Significant Different Test was selected as the ad hoc test to be used.

Table 5.95 displays the results of Tukey's Honestly Significant Different Test on the type of Data Imputation Method:

Table 5.9.5 Tests of Between-Subjects Effects

Dependent Variable: RMS

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	.008(a)	11	.001	3.608	.003
Intercept	1.195	1	1.195	6063.947	.000
IMPUTATION	.006	4	.002	7.790	.000
DATAMISS	.002	7	.000	1.217	.326
Error	.006	28	.000		
Total	1.208	40			
Corrected Total	.013	39			

a R Squared = .586 (Adjusted R Squared = .424)

Multiple Comparisons

Dependent Variable: RMS
Tukey HSD

(I) IMPUTATION	(J) IMPUTATION	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval			
							Lower Bound	Upper Bound
.00	1.00	.0254(*)	.00702	.009	.0049	.0458		
	2.00	.0306(*)	.00702	.001	.0102	.0511		
	3.00	.0185	.00702	.091	-.0020	.0390		
	4.00	.0356(*)	.00702	.000	.0152	.0561		
1.00	.00	-.0254(*)	.00702	.009	-.0458	-.0049		
	2.00	.0053	.00702	.943	-.0152	.0257		
	3.00	-.0069	.00702	.862	-.0273	.0136		
	4.00	.0103	.00702	.596	-.0102	.0307		
2.00	.00	-.0306(*)	.00702	.001	-.0511	-.0102		
	1.00	-.0053	.00702	.943	-.0257	.0152		
	3.00	-.0121	.00702	.434	-.0326	.0083		
	4.00	.0050	.00702	.952	-.0155	.0255		
3.00	.00	-.0185	.00702	.091	-.0390	.0020		
	1.00	.0069	.00702	.862	-.0136	.0273		
	2.00	.0121	.00702	.434	-.0083	.0326		
	4.00	.0171	.00702	.134	-.0033	.0376		
4.00	.00	-.0356(*)	.00702	.000	-.0561	-.0152		
	1.00	-.0103	.00702	.596	-.0307	.0102		
	2.00	-.0050	.00702	.952	-.0255	.0155		
	3.00	-.0171	.00702	.134	-.0376	.0033		

Based on observed means.
* The mean difference is significant at the .05 level.

RMS

Tukey HSD

IMPUTATION	N	Subset 1	Subset 2
4.00	8	.1593	
2.00	8	.1643	
1.00	8	.1695	
3.00	8	.1764	.1764
.00	8		.1949
Sig.		.134	.091

Means for groups in homogeneous subsets are displayed.
Based on Type III Sum of Squares
The error term is Mean Square(Error) = .000.
a Uses Harmonic Mean Sample Size = 8.000.
b Alpha = .05.

Multiple Comparisons

Dependent Variable: RMS
Tukey HSD

(I) DATAMISS	(J) DATAMISS	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
.00	1.00	.0112	.00888	.905	-.0178	.0402
	2.00	.0008	.00888	1.000	-.0282	.0298
	3.00	-.0074	.00888	.989	-.0364	.0216
	4.00	.0112	.00888	.905	-.0178	.0402
	5.00	.0068	.00888	.994	-.0222	.0358
	6.00	-.0010	.00888	1.000	-.0300	.0280
	7.00	-.0044	.00888	1.000	-.0334	.0246
1.00	.00	-.0112	.00888	.905	-.0402	.0178
	2.00	-.0104	.00888	.934	-.0394	.0186
	3.00	-.0186	.00888	.442	-.0476	.0104
	4.00	.0000	.00888	1.000	-.0290	.0290
	5.00	-.0044	.00888	1.000	-.0334	.0246
	6.00	-.0122	.00888	.861	-.0412	.0168
	7.00	-.0156	.00888	.652	-.0446	.0134
2.00	.00	-.0008	.00888	1.000	-.0298	.0282
	1.00	.0104	.00888	.934	-.0186	.0394
	3.00	-.0082	.00888	.981	-.0372	.0208
	4.00	.0104	.00888	.934	-.0186	.0394
	5.00	.0060	.00888	.997	-.0230	.0350
	6.00	-.0018	.00888	1.000	-.0308	.0272
	7.00	-.0052	.00888	.999	-.0342	.0238
3.00	.00	.0074	.00888	.989	-.0216	.0364
	1.00	.0186	.00888	.442	-.0104	.0476
	2.00	.0082	.00888	.981	-.0208	.0372
	4.00	.0186	.00888	.442	-.0104	.0476
	5.00	.0142	.00888	.747	-.0148	.0432
	6.00	.0064	.00888	.996	-.0226	.0354
	7.00	.0030	.00888	1.000	-.0260	.0320
4.00	.00	-.0112	.00888	.905	-.0402	.0178
	1.00	.0000	.00888	1.000	-.0290	.0290
	2.00	-.0104	.00888	.934	-.0394	.0186
	3.00	-.0186	.00888	.442	-.0476	.0104
	5.00	-.0044	.00888	1.000	-.0334	.0246
	6.00	-.0122	.00888	.861	-.0412	.0168
	7.00	-.0156	.00888	.652	-.0446	.0134
5.00	.00	-.0068	.00888	.994	-.0358	.0222
	1.00	.0044	.00888	1.000	-.0246	.0334
	2.00	-.0060	.00888	.997	-.0350	.0230
	3.00	-.0142	.00888	.747	-.0432	.0148
	4.00	.0044	.00888	1.000	-.0246	.0334
	6.00	-.0078	.00888	.986	-.0368	.0212

	7.00	-.0112	.00888	.905	-.0402	.0178
6.00	.00	.0010	.00888	1.000	-.0280	.0300
	1.00	.0122	.00888	.861	-.0168	.0412
	2.00	.0018	.00888	1.000	-.0272	.0308
	3.00	-.0064	.00888	.996	-.0354	.0226
	4.00	.0122	.00888	.861	-.0168	.0412
	5.00	.0078	.00888	.986	-.0212	.0368
	7.00	-.0034	.00888	1.000	-.0324	.0256
7.00	.00	.0044	.00888	1.000	-.0246	.0334
	1.00	.0156	.00888	.652	-.0134	.0446
	2.00	.0052	.00888	.999	-.0238	.0342
	3.00	-.0030	.00888	1.000	-.0320	.0260
	4.00	.0156	.00888	.652	-.0134	.0446
	5.00	.0112	.00888	.905	-.0178	.0402
	6.00	.0034	.00888	1.000	-.0256	.0324

Based on observed means.

Similar to the results found in the testing of Hypothesis 9, the results of Tukey's Honestly Significant Different Test (HSD) indicate that the Level Of Data Missingness is not significant in the KDD models tested (also enforced by the ANOVA testing). The test also indicates that performing some method of Data Imputation (as opposed to not performing any type of Data Imputation) is significant at the .05 level.

Tukey's Honestly Significant Difference test (HSD) discovered that only one of the Data Imputation methods tested in this research is not statistically significant at the .05 level, but all Imputation Methods are statistically significant at the .10 level of significance.

Therefore, it can be seen from the ANOVA and Tukey's Honestly Significant Difference Test (HSD), that when the volume of Case Frequency is considered to be large (N=1000), the level of Data Missingness does not significantly impact the calculation of the Root Mean Square Value by a KDD model that utilizes an Artificial Neural Network employing an S-Sigmoid Transfer Function as it's Data Mining Algorithm.

Next, the employment of some type of Data Imputation Method was tested and found to be significant. All but one of the Imputation Methods are significant at the .05 level of significance, and, as also discovered in Hypothesis 9, all Imputation Methods are significant at the .10 level of significance.

In summation, utilizing large KDD models (tested at a Case Frequency Volume of N=1000), the imputation method of Multiple Imputation does not result in better (lower) Root Mean Square (RMS) values than the imputation methods of Regression Imputation, Mean Substitution and Case Deletion.

CHAPTER VI

IMPLICATIONS AND CONCLUSIONS

Knowledge Discovery in Databases (KDD) is defined as “the non-trivial multi-step process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data” (Fayyad, 2001). Data Mining is known as the “discovery” step of the Knowledge Discovery process.

Data mining is based on searching the concatenation of multiple databases that usually contain some amount of missing data along with a variable percentage of inaccurate data, pollution, outliers and noise. Various Data Mining Algorithms may be employed in the discovery step of Knowledge Discovery.

Nearest Neighbor, Decision Trees, Association Rules, Neural Networks, Genetic Algorithms and multiple hybrid algorithms are some of the methods employed when performing a Data Mining operation.

A Neural Network algorithm contains an Activation Function that is triggered when data has been processed and is ready to be sent to the Output Layer of the Neural Network. This Activation Function consists of a Combination Function and a Transfer Function.

Several types of Transfer Functions may be employed by the Neural Network Architecture, the most common being the S-Sigmoid Transfer Function. Due to its robust ability to handle both linear and non-linear data, this is the function chosen for study in this research.

The issue of Missing Data must be addressed, as ignoring this problem can introduce bias into the Knowledge Discovery models being evaluated and may lead to inaccurate data mining conclusions.

Therefore, the volume of original data under analysis, when confronted with various increasing levels of Missing Data, must be tested to challenge the sensitivity of Knowledge Discovery models when using a Neural Network employing an S-Sigmoid Transfer Function within its Activation Function.

Various Transfer Functions may be employed to transfer the value obtained from the Combination Function to the output nodes of the Neural network. Some of the commonly employed Transfer Functions include the Linear Transfer Function, The Hyperbolic Tangent and the S-Sigmoid Transfer Function.

Due to it's robust handling of both linear and non-linear data, the most commonly employed Transfer Function is the S-Sigmoid, by far (Berry and Linoff, 1997).

Although Missing Data may be categorized as Missing At Random (MAR), Missing Completely At Random (MCAR), Non-Ignorable Missing Data and Outliers Treated As Missing Data, they are all simply data that are not complete and may be treated as Missing Data.

Data Imputation Methods may be employed to replace missing values. Commonly employed Imputation Methods include Case Deletion, Mean Substitution, Hot Deck Imputation, Cold Deck Imputation, Regression Imputation and Multiple Imputation.

6.1 Implications of this Research

The significant goals of this research are to test the most commonly used architectures employed in a typical Knowledge Discovery environment along three dimensions. Those dimensions include Data Set Size (Case Frequency Volume), Level of Data Missingness and type of Data Imputation Method.

The most common algorithm used in “off the shelf” commercial Data Mining software packages is the Neural Network, employing an S-Sigmoid Transfer Function. Using the Default Parameters of a commercial KDD package, the three aforementioned dimensions were tested.

This research sought to determine if the size of a KDD model (in terms of the Number Of Data Instances) resulted in more effective results and if the amount of data that is missing in a KDD model also had an impact on those results, when utilizing the most commonly employed KDD methodologies.

6.2 Goals of this Research

This research took to task the goal of determining how the three aforementioned factors impact a data mining process. The factors of original KDD Case Frequency Volume, Level of Data Missingness and Data Imputation Method were selected, and a

KDD and Data Mining analysis completed for five independent KDD models.

Data missingness was injected into each KDD model at various levels and compared to Data Mining results obtained when complete data was used in each KDD model, and when no Data Imputation was performed on the injected Data Missingness.

ANOVA tests, T-Tests and Tukey's Honestly Significant Difference tests were performed to determine which factors were significant for a more effective KDD and Data Mining results. Root Mean Square (RMS) values were used as the metric for measuring the effectiveness of the KDD and Data Mining processes.

This research explores two specific steps in the Knowledge Discovery of Databases (KDD) process, Data Cleansing and Data Mining, as well as the impact of the volume of data being analyzed. The actual data mining process deals significantly with prediction, estimation, classification, pattern recognition and the development of association rules. Therefore, this analysis depends heavily on the accuracy of the database and on the chosen sample data to be used for model training and testing. One objective of this research is to address the Effects of the Neural Network S-Sigmoid Function on KDD models containing various levels of data instances in the Presence of Imprecise Data using three-factor ANOVA tests, two-factor ANOVA tests, T-Tests and Tukey's Honestly Significant Difference test.

This research further investigates the accuracy and impact of Data Imputation Methodologies that are employed when a specific Data Mining algorithm is utilized within a Knowledge Discovery In Databases (KDD) process. This study will employ certain Knowledge Discovery processes that are widely accepted in both the academic and commercial worlds. This work includes testing the impact of Missing Data on KDD

models that utilize the Neural Network S-Sigmoid Transfer Function type in the Data Mining process, by experimenting with three factors: Imputation Method, Level of Data Missingness, and the Volume of Case Frequency in the KDD model.

6.3 Contributions of this Research

The first contribution of this research was to test and analyze the performance of KDD models utilizing a Neural Network as it's Data Mining Algorithm, and when that algorithm employed an S-Sigmoid Transfer Function, and when the KDD models contained various Case Frequency volumes in the training and testing of those models.

It was discovered in this research that Data Model Size (Case Frequency Volume) was not a significant factor in the training and testing of KDD models, both in KDD environments where the volume of Case Frequency was considered to be low (N=500) and again when the KDD models contained their original volume of Case Frequencies.

The second contribution of this research was the testing and analysis of the sensitivity of KDD models that had different volumes of Case Frequency, and when those models were confronted with various increasing levels of Data Missingness.

This research first tested this dimension of Case Frequency Volume and Level of Data Missingness by testing and analyzing KDD models from five different disciplines at a low level of Case Frequency (N=500).

It was discovered in this research that the performance of all models used in this study (with N=500) degraded significantly immediately when exposed to Missing Data (at the 10% level), but did not suffer further degradation upon exposure to greater levels of Data Missingness (20%-70%).

The dimension of Level of Data Missingness was further tested on the KDD models containing their original Case Frequencies.

In total, these tests indicate that the Level of Data Missingness in the Data Instances used to train and test KDD models did not significantly impact the performance of those models past the 10% level of Data Missingness. This further supported the results of tests that had been performed on KDD models that contained a low volume of Case Frequency. That is, the volume of Case Frequency again did not significantly impact the performance of KDD models containing increasing levels of Missing Data.

The third contribution of this research was the testing and analysis of the performance of KDD models following the employment of Data Imputation methodologies on KDD models that contained various increasing levels of Data Missingness.

The tests performed in this research showed that performing Data Imputation did in fact have a significant impact on the performance of KDD models than when no imputation was performed. Further testing performed on the type of Data Imputation Method utilized did not distinguish a significant difference between the types of Imputation Method employed, Mean Substitution and Case Deletion.

The dimensions of Level of Case Frequency and Data Imputation were again tested in like manner. Five hundred randomly selected Data Instances from each KDD model were chosen for training and testing. These tests disagreed with the tests performed on the KDD models containing all Data Instances. When the Level of Case Frequency was below 1000 (N=500), Data Imputation did not significantly impact the performance of the KDD model.

Further testing in this research indicated that the Data Imputation methods of Regression Imputation and Multiple Imputation performed on KDD models that contained various levels of Data Missingness performed no better than the Imputation Methods of Mean Substitution and Case Deletion.

6.4 Limitations and Directions for Future Research

Research in the areas of Knowledge Discovery, Data Mining, Missing Data and Data Imputation may be limited by a lack of standardized information regarding new methodologies as the field evolves into the state of maturity. Currently, Data Mining is viewed as an evolving, but not yet mature, field (KDNuggets, 2007).

Also, as the focus on other technical dimensions (such as Data Warehousing and Data Shaping) continue to evolve concurrently, KDD and Data Mining software will also following adaptation to those areas and continue to evolve in order to adapt to them.

In future areas of research, other dimensions such as more complex methodology in the design of hybrid data mining algorithms employed (including merging concepts from Nearest Neighbor, Decision Trees, Association Rules, Genetic Algorithms and newly developed hybrids) in conjunction with varying the parameters within a specific algorithm which may be added/and tested for their impact of the Knowledge Discovery process.

Also, the parameters used within a particular data mining algorithm, such as the neural network, may be adjusted in an attempt to determine which combination of parameter settings perform most effectively when implemented upon various types (size and structure) of data sets.

Concerning neural networks specifically, a comparison of different combination and/or transfer functions within the Activation Function of an ANN can be performed to determine the most effective type of function or combination of functions is most desirable for data sets of various dimensions.

The type of imputation method utilized when confronted with missing data is yet another area of research that may be explored. The imputation techniques of Hot Deck, Cold Deck, Regression and Multiple Imputation are just a few methods that may be tested in conjunction with the aforementioned imputation methods utilized in this research for more effective Knowledge Discovery and Data Mining.

7.0 Conclusions

From the T-tests, ANOVA results and Tukey's Honestly Significant Difference (HSD) Tests, this research revealed that original the KDD Case Frequency and the type of Imputation Method employed are significant factors in the performance of KDD models that utilize a Neural network as its Data Mining Algorithm and employ an S-Sigmoid Transfer Function.

However, while the level of data missingness in a KDD model was found to promote a higher (worse) Root Mean Square (RMS) Value when Missing Data is first introduced to a KDD model, increased levels of Data Missingness was not proven to be significant in this study.

It was also discovered, via that Tukey's Honestly Significant Difference Test (HSD) analysis, that while there is a significant difference between employing and not employing a Data Imputation Method, there is no significant difference between the

Imputation Methods of Multiple Imputation (utilizing a hybrid of Regression Imputation and Mean Substitution), Regression Imputation, Mean Substitution and Case Deletion.

BIBLIOGRAPHY

Abramowicz, W., and Zurada, J, (2001), Knowledge Discovery For Business Information Systems. Boston, MA: Kluwer Academic Publishers.

Acharya, T. and Mitra, S. (2003), Data Mining, John Wiley & Sons, Inc., Hoboken, New Jersey, p. 181.

Acuña, J., Alonso, S., Hortelano, C., Martínez-Tello, F., Ortiz, P., Pajares, R., Pérez-Gómez, B., Piris, M., Pollán, M., Rodríguez-Peralto, J., Sánchez, L. (2004), "Progression in Cutaneous Malignant Melanoma Is Associated with Distinct Expression Profiles", American Journal of Pathology, 164:193-203.

Adriaans, P. and Zantinge, D. (1997), Data Mining, Addison-Wesley, New York.

Afifi, A., and Elashoff, R. (1966) "Missing Observations in Multivariate Statistics I: Review Of The Literature", Journal of the American Statistical Association, Vol. 61, pp. 595-604.

Agrawal, R., Imielinski, T., and Swami, A. (1993), "Mining Associations between Sets of Items in Large Databases". Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, DC, 207-216.

Agrawal, R. and Srikant, R. (1994), "Fast Algorithms for Mining Association Rules in Large Databases", Proceedings of the 20th International Conference on Very Large Databases, Santiago de Chile, Chile.

Agrawal, R.,Mannila, H.,Srikant, R.,Toivonen,H.,and Verkamo, A. (1995), "Fast Discovery of Association Rules", Advances in Knowledge Discovery and Data Mining, Chapter 12. Cambridge, MA: AAAI/MIT Press.

Ballou, D., Pazer, H., Tayi, G. and Wang, R. (1998) "Modeling Information Manufacturing Systems to Determine Information Product Quality", Management Science, Vol. 44, No. 4 pp. 462-484.

Batista, G. and Monard, M. (2003). "An Analysis of Four Missing Data Treatment Methods for Supervised Learning", Applied Artificial Intelligence, Volume 17, Numbers 5-6, Numbers pp. 519-533(15).

Barnard, J. and Meng, X. (1999), "Applications of Multiple Imputation In Medical Studies: From AIDS to NHANES", Statistical Methods in Medical Research, Vol. 8, pp. 17-36.

Berson, A. and Smith, S. (1997), Data Warehousing, Data Mining and OLAP, McGraw-Hill.

Berry, M., and Linoff, G. (1997), Data Mining Techniques, Wiley, New York.

Berry, M., and Linoff, G. (2000) "The Art And Science Of Customer Relationship", Industrial Management & Data Systems, Vol. 100, No 5, pp. 245-246.

Berson, A., Smith, S. and Thearling, K. (2000), Building Data Mining Applications for CRM, McGraw-Hill, New York.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984), Classification and Regression Trees, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA.

Brick, J. M., and Kalton, G. (1996) "Handling Missing Data In Survey Research", Statistical Methods in Medical Research, Vol. 5, pp. 215-238.

Cappiello, C., Francalanci, C. and Pernici, B. (2004), "Data Quality Assessment From The User's Perspective", IQIS, pp. 68-73.

Clogg, C., Rubin, D., Schenker, N., Schultz, B., and Weidman, L. (1991), "Multiple Imputation of Industry and Occupation Codes in Census Public-Use Samples Using Bayesian Logistic Regression", Journal of the American Statistical Association, Vol. 86, No 413, pp. 68-78.

Cohen, J., and Cohen, P. (1983), Applied Multiple Regression/Correlation Analysis For The Behavioral Sciences, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.

Coy, P. (1997), "He Who Minds Data May Strike Fool's Gold", Business Week, # 3531, 40.

Darling, Charles B. (1997), "Datamining for the Masses", Datamation, Vol. 52, pp. 5.

David, M., Little, R., Samuhel, M., and Triest, R. (1986) "Alternative Methods for CPS Income Imputation", Journal of the American Statistical Association, Vol. .81, pp. 29-41.

Dempster, A., and Rubin, D. (1983), "Incomplete Data in Sample Surveys", in Madow, W., G., Olkin, I., and Rubin, D. (Eds.), Sample Surveys Vol. II: Theory and Annotated Bibliography, Academic Press, New York, pp. 3-10.

Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum Likelihood From Incomplete Data Via The EM Algorithm (with discussion)", Journal of the Royal Statistical Society, Vol. B39, pp. 1-38.

Diggle, P., and Kenward, M. (1994), "Informative Dropout in Longitudinal Data Analysis (with discussion)", Applied Statistics, Vol. 43, pp. 49-94.

Dillman, D. A. (1999), Mail and internet surveys: "The Tailored Design Method", John Wiley Company, New York.

Drucker, P. (2004), "It Is Now Time To Focus on the "I" in "IT", http://www.opengroup.org/comm/newsletter/2004/09.htm.

Ernst, L. (1980) "Variance Of The Estimated Mean For Several Imputation Procedures", American Statistical Association 1980, Proceedings of the Survey Research Methods Section, pp. 716-720.

Fayyad, U., Haussler, D., and Stolorz, P., (1996), "Mining Scientific Data", Communications of the ACM, Vol. 39. No.11.

Fayyad, U. (2001), "Data Mining and Knowledge Discovery", International Journal, Vol. I Issue 2.

Flockhart, I and Radcliffe, N. (1996), "A Genetic Algorithm-Based Approach to Data Mining", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96).

Ford, B. (1981), "An Overview Of Hot Deck Procedures" in Madow, W. G., Olkin, I.., And Rubin, D. (Eds.), Sample Surveys Vol. II: Theory and Annotated Bibliography, Academic Press, New York, pp. 3-10.

Ghahramani, Z., and Jordan, M. (1997), "Mixture Models For Learning From Incomplete Data", In Cowan, J., Tesauro, G. and Alspector, J. (eds.), Advances In Neural Information Processing Systems 6. San Mateo, CA: Morgan Kaufman, 120-127.

Graham, J., Hofer, S., and Piccinin, A. (1994) "Analysis With Missing Data In Drug Prevention Research", in Collins, L. M. and Seitz, L. (Eds.), "Advances in Data Analysis for Prevention Intervention Research NIDA Research Monograph", Series (#142), National Institute on Drub Abuse, Washington, D.C.

Graham, J., Hofer, S., Donaldson, S., MacKinnon, D., and Schafer, J. (1997), "Analysis with missing data in prevention research", in Bryant, K., Windle, W., and West, S. (Eds.), New Methodological Approaches to Alcohol Prevention Research, American Psychological Association, Washington, D.C.

Groth, R. (2000), Data Mining: Building Competitive Advantage, Prentice-Hall, Upper Saddle River, NJ.

Hair, J., Anderson, R., Tatham, R., and Black, W. (1998), Multivariate Data Analysis, Prentice-Hall, Upper Saddle River, NJ.

Han, J., and Kamber, M. (2001), Data Mining: Concepts and Techniques, Academic Press, San Francisco.

Hansen, M., Madow, W., and Tepping, J. (1983), "An Evaluation Of Model-Dependent And Probability-Sampling Inferences In Sample Surveys", Journal of the American Statistical Association, Vol. 78, pp. 776-807.

Hartley, H., and Hocking, R. (1971), "The Analysis Of Incomplete Data", Biometrics, Vol. 27, pp. 783-808.

Heitjan, D.F. (1997), "Annotation: What Can Be Done About Missing Data? Approaches to Imputation", American Journal of Public Health, Vol. 87, No 4, pp. 548-550.

Herzog, T., and Rubin, D. (1983), "Using Multiple Imputations To Handle Nonresponse In Sample Surveys in Incomplete Data", Sample Surveys, Volume 2: Theory and Bibliographies, Madow, W. G., Olkin, I., and Rubin, D. (Eds.), Academic Press, New York, pp. 209-245.

Holland, J. (1975), "Adaptation in Natural and Artificial Systems", University of Michigan Press: Ann Arbor, MI.

Howell, D.C. (1998), "Treatment of missing data", D. C. Howell personal website, http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html/

Haykin, S. (1994), Neural Networks: A Comprehensive Foundation, New York: Macmillan Publishing.

Iannacchione, V. (1982), "Weighted Sequential Hot Deck Imputation Macros", Proceedings of the SAS Users Group International Conference, Vol. 7, pp. 759-763.

Ishikawa, K. (2004), "Information Quality: Your Decisions Are Only as Good as Your Information", http://www.opengroup.org/comm/newsletter/2004/09.htm

Jenkins, C. R., and Dillman, D. A. (1997), "Towards A Theory Of Self-Administered Questionnaire Design", in Lyberg et al. (Eds.), Survey measurement and process quality, John Wiley Company, New York.

Kalton, G., and Kasprzyk, D. (1982), "Imputing For Missing Survey Responses", American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 22-31.

Kalton, G., and Kasprzyk, D. (1986), "The Treatment Of Missing Survey Data", American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 22-31.

Kalton, G., and Kish, L. (1981), "Two efficient random imputation procedures", American Statistical Association 1981, Proceedings of the Survey Research Methods Section, pp. 146-151.

Kass, G. (1980), "An Exploratory Technique For Investigating Large Quantities Of Categorical Data", Applied Statistics, 29(2).

KDNuggets Poll (2007), "Is Data Mining and Knowledge Discovery A *Mature* Or *Emerging* Field?". www.kdnuggets.org/2007poll.

Kim, Y. (2001), "The Curse Of The Missing Data", Y. Kim personal website. Available http://209.68.240.11:8080/2ndMoment/978476655/addPostingForm/

Lee, S., and Siau, Keng (2001), "A Review Of Data Mining Techniques", Industrial Management & Data Systems, Vol. 101, No 1, pp. 41-46.

Li, K. (1985), "Hypothesis Testing In Multiple Imputation - With Emphasis On Mixed-Up Frequencies in Contingency Tables", Ph.D. Thesis, The University of Chicago, Chicago, IL.

Lim, T-S, Loh, W-Y and Shih, Y-S (1997), "An Empirical Comparison Of Decision Trees And Other Classification Methods", TR 979, Department of Statistics, UW Madison, June.

Little, R. (1982), "Models for Non-response in Sample Surveys", Journal of the American Statistical Association, Vol. 77, pp. 237-250.

Little, R. (1992), "Regression With Missing X's: A Review", Journal of the American Statistical Association, Vol. 87, pp. 1227-1237.

Little, R. (1995), "Modeling The Drop-Out Mechanism In Repeated-Measures Studies", Journal of the American Statistical Association, Vol. 90, pp. 1112-1121.

Little, R., and Rubin, D., (1987), Statistical Analysis With Missing Data, Wiley, New York.

Little, R., and Rubin, D. (1989), "The Analysis Of Social Science Data With Missing Values", Sociological Methods and Research, Vol. 18, pp. 292-326.

Loh, W. and Shih, Y. (1997), "Split Selection Methods for Classification Trees", Statistica Sinica, 7, 815-840.

Loh and Vanichestakul (1988), "Tree-Structured Classification Via Generalized Discriminate Analysis (with discussion)", Journal of the American Statistical Association, V83, 715-728.

Loshin, D. (2001), Enterprise Knowledge Management: The Data Quality Approach. Morgan Kuafmann.

Loshin, D. (2004), "Knowledge integrity: data standards and data models", DM Review, (January), p. 2-3.

Ma, C., Chou, D., and Yen, D. (2000), "Data Warehousing, Technology Assessment and Management", Industrial Management & Data Systems, Vol. 100, No 3, pp. 125-135.

Masters (1995), "Neural, Novel, and Hybrid Algorithms for Time-Series Predictions", New York: Wiley.

MatLab (2007), "Representing Missing Data Values", http://www.mathworks.com/access/helpdesk/data_analysis

McQueen, R. and Thorley, S (1999), "Mining Fool's Gold", JEL Classifications: G14, G11, C10, Working Paper Series.

Michalewicz, Z. (1994), Genetic Algorithms + Data Structures = Evolution Programs, Springer-Verlag (Berlin and New York).

Miller, H. J. (2000), "Geographic Representation in Spatial Analysis", *Journal of Geographical Systems, 2:55-60.*

Miller, H. J. and Han, J. (2000), "Discovering Geographic Knowledge In Data Rich Environments: A Report On A Specialist Meeting", SIGKDD Explorations: Newsletter of the, Association for Computing Machinery, Special Interest Group on Knowledge Discovery and Data Mining, 1(2): 105-108. http://www.acm.org/sigs/sigkdd/explorations.

Miller, H. and Han, J. (2001), Geographic Data Mining and Knowledge Discovery, London: Taylor & Francis.

Morgan, J, and Messenger, R. (1973), "THAID: A Sequential Analysis Program For The Analysis Of Nominal Scale Dependent Variables", Technical Report, Institute for Social Research, University of Michigan, Ann Arbor, MI.

Morgan, J, and Sonquist, J. (1963), "Problems In The Analysis Of Survey Data And A Proposal", Journal of the American Statistical Association, 58, 415-434.

Nie, N., Hull, C., Jenkins, J., Steinbrenner, K., and Bent, D. (1975), SPSS, 2nd ed., McGraw-Hill, New York.

Orchard, T., and Woodbury, M. (1972), "A Missing Information Principle: Theory and Applications", Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, pp. 697-715.

Pennell, S. (1993), "Cross-Sectional Imputation And Longitudinal Editing Procedures In The Survey Of Income And Program Participation", Technical report, Institute of Social Research, University of Michigan, Ann Arbor, MI.

Razi, M., and Athappilly, K. (2005), "A Comparative Predictive Analysis Of Neural Networks (Nns), Nonlinear Regression and Classification and Regression Tree (CART) Models", Elsevier Ltd.

Ripley, B. (1996), Pattern Recognition And Neural Networks, Cambridge, UK: Cambridge University Press.

Roth, P. (1994), "Missing Data: A Conceptual Review For Applied Psychologists", Personnel Psychology, Vol. 47, pp. 537-560.

Roiger R., and Geatz, M. (2003), Data Mining, Addison-Wesley.

Royall, R., and Herson, J. (1973), "Robust Estimation from Finite Populations", Journal of the American Statistical Association, Vol. 68, pp. 883-889.

Rubin, D. (1978), "Multiple Imputations In Sample Surveys - A Phenomenological Bayesian Approach To Nonresponse", Imputation and Editing of Faulty or Missing Survey Data, U.S. Department of Commerce, pp. 1-23.

Rubin, D. (1986), "Statistical Matching Using File Concatenation With Adjusted Weights And Multiple Imputations", Journal of Business and Economic Statistics, Vol. 4, pp. 87-94.

Rubin, D. (1996), "Multiple Imputation After 18+ Years (with discussion)", Journal of the American Statistical Association, Vol. 91, pp. 473-489.

Rubin, D., and Schenker, N. (1986), "Multiple Imputation For Interval Estimation From Simple Random Sample with Ignorable Nonresponse", Journal of the American Statistical Association, Vol. 81, pp. 366-374.

Sande, L. (1982), "Imputation In Surveys: Coping With Reality", The American Statistician, Vol. 36, pp. 145-152.

Sande, L. (1986), "Hot-Deck Imputation Procedures", in Madow, W. G. and Olkin, I. (Eds.), "Incomplete Data In Sample Surveys", Vol. 3, Proceedings of the Symposium, Academic Press, New York, pp. 339-349.

Schafer, J. (1997), Analysis Of Incomplete Multivariate Data, Chapman and Hall, London.

Schafer, J. (1999), "Multiple Imputation: A Primer", Statistical Methods in Medical Research, Vol. 8, pp. 3-15.

Schafer, J., and Olsen, M. (1998), "Multiple Imputation For Multivariate Missing-Data Problems: A Data Analyst's Perspective", Multivariate Behavioral Research, Vol. 33, pp. 545-571.

Sehgal, M., Gondal, I., and Dooley L. (2005), "Methods of DNA Data Analysis: Missing Value Imputation for Microarrays", Patent Number, *2005903507*. http://www.warezreleases.com/files/missing data california.rar.

Sethi I. (2003), "The Pitfalls Of Knowledge Discovery In Databases And Data Mining", Data Mining: Opportunities and Challenges, pp. 220-238.

Sharpe, P and Glover, R., (xxxx), Efficient GA Based Techniques for Classification, Kluwer Academic Publishers, Hingham, MA.

Statistical Services of University of Texas (2000), "General FAQ #25: Missing or Incomplete Data". http://www.utexas.edu/cc/faqs/stat/general/gen25.html/

StatSoft (2002), "Pre-Empting User Questions Through Anticipation: Data Mining FAQ Lists". http://www.statsoft.com/faq.

Szpiro, G.G. (1997), "A Search for Hidden Relationships: Data Mining with Genetic Algorithms", Computational Economics, Vol. 10, No. 3, Aug. 1997, pp. 267-277, Kluwer Academic Publishers.

Tresp, V., Neuneier, R., and Ahmad, S. (1995), "Efficient Methods For Dealing With Missing Data In Supervised Learning", in Tesauro, G., Touretzky, D. and Keen, T. (eds.), Advances In Neural Information Processing Systems 7. Cambridge, MA: The MIT Press, 689-696.

van Buren, S., Boshuizen, H., and Knook, D. (1999), "Multiple Imputation Of Missing Blood Pressure Covariates In Survival Analysis", Statistics in Medicine, Vol. 18, pp. 681-694.

Vosburg; J. and Kumar, A. (2001), "Managing Dirty Data In Organizations Using ERP: Lessons From A Case Study", Industrial Management & Data Systems, Vol. 101, No 1, pp. 21-31.

Wang, J., (2003), Data Mining: Opportunities and Challenges, Idea Group Publishing.

Warner, B., and Misra, M. (1996), "Understanding Neural Networks as Statistical Tools", The American Statistician, 50, 284-293.

Westphal, C., and Blaxton, T. (1998), Data Mining Solutions, Wiley, New York.

Witten, I., and Frank, E. (2000), Data Mining, Academic Press, San Francisco.

Wothke, W. (1998), "Longitudinal And Multi-Group Modeling With Missing Data", in Little, T. D., Schnabel, K. U., and Baumert, J. (Eds.), Modeling Longitudinal and Multiple Group Data: Practical Issues, Applied Approaches and Specific Examples, Lawrence Erlbaum Associates, Mahwah, NJ.

Xu, H., Horn Nord, J., Brown, N. and Nord, G. D. (2002), "Data Quality Issues In Implementing an ERP", Industrial Management & Data Systems, Vol. 102, No 1, pp. 47-58.

APPENDIX A

VISUAL BASIC MODULE EMPLOYED IN THIS STUDY

DATA IMPUTATION MODULE

```
Sub Imputation IDA()


Application.DisplayAlerts = False
Dim counter As Integer

Dim delrows(6500) As Integer


For I = 1 To 6500
delrows(I) = 0
Next

Sheets(1).Select
Sheets(1).Name = "Master"

strCol = InputBox("What is the letter(s) of the rightmost column?")
If strCol = "" Then
Exit Sub
End If


If Len(strCol) > 2 Then
MsgBox "You have too many columns for this program"
Exit Sub
End If

strOutputCol = InputBox("What is the letter(s) of your output column?")

intOutputCol = getColNum(strOutputCol)

strRow = InputBox("What is the row number of the last record of data?")
If strRow = "" Then
Exit Sub
End If
```

```
'
'
If strRow > 6500 Then
MsgBox "Cannot exceed 6500"
Exit Sub
End If


intCol = getColNum(strCol)



If intCol = 999 Then
MsgBox "Error"
Exit Sub
End If


intIgnore = InputBox("How many top rows do you want to ignore?")

If intIgnore = "" Then
intIgnore = 0
End If



intRecords = (strRow - intIgnore)

'pct = InputBox("There are " & intRecords & " records of data.  What percentage of
records do you wish to corrupt? (one attribute in each random record will be blanked)")

'If pct = "" Then
'Exit Sub
'End If

If MsgBox("Will create missing values, mean values and case deletion for 10% - 70% in
increments of 10%.  Proceed?", vbYesNo) = vbNo Then

Exit Sub
End If
```

```
For M = 1 To 7


Sheets("Master").Copy After:=Sheets("Master")
Sheets(2).Select
Sheets(2).Name = "Missing " & M & "0 pct"
Sheets("Master").Copy After:=Sheets("Master")
Sheets(2).Select
Sheets(2).Name = "Case Delete " & M & "0 pct"
'    Columns("A:A").Select
'    Selection.Insert Shift:=xlToRight
Sheets("Missing " & M & "0 pct").Select


pct = M & "0"

Dim intCorruptRecords As Integer

intCorruptRecords = intRecords * (pct / 100)


'If MsgBox("Will corrupt one attribute in each of " & intCorruptRecords & " random
records.  Proceed?", vbYesNo) = vbNo Then
'
'Exit Sub
'End If

For I = 1 To intCorruptRecords

   Do
      intLetter = Rand(1, intCol)

      If intLetter <> intOutputCol Then

         letter = getColLetter(intLetter)
         Number = Rand(intIgnore + 1, strRow)
         cell = letter & Number
         If InStr(1, strDeleted, Number & " ") = 0 Then
            Exit Do
         End If

      End If

   Loop
```

```
    strDeleted = strDeleted & Number & " "

'    Sheets("Case Delete " & M & "0 pct").Select
'
'    Range("A" & Number).Select
'    ActiveCell.Formula = 1

    delrows(I) = Number

    Sheets("Missing " & M & "0 pct").Select

    Range(cell).Select
    ActiveCell.FormulaR1C1 = ""
    Selection.Interior.ColorIndex = 3



Next


Sheets("Missing " & M & "0 pct").Copy After:=Sheets("Missing " & M & "0 pct")
Sheets(4).Select
Sheets(4).Name = "Mean " & M & "0 pct"




For I = 1 To intCol

letter = getColLetter(I)
cell = letter & strRow + 1
Formula = "=AVERAGE(" & letter & intIgnore + 1 & ":" & letter & strRow & ")"
Range(cell).Select
ActiveCell.Formula = Formula
colavg = ActiveCell.Value

    Columns(letter & ":" & letter).Select
    'Range("AD979").Activate
    Selection.Replace What:="", Replacement:=colavg, LookAt:=xlPart, _
        SearchOrder:=xlByRows, MatchCase:=False, SearchFormat:=False, _
        ReplaceFormat:=False
```

```
Next


    Rows(strRow + 1 & ":" & strRow + 1).Select
    Selection.Delete Shift:=xlUp

    Sheets("Case Delete " & M & "0 pct").Select

'    tempstr = ""
'    For R = 1 To 10
'    tempstr = tempstr & " " & delrows(R)
'    Next
'    MsgBox tempstr
'
'
'    For Q = 1 To 5999
'
'       temp = ""
'
'       'MsgBox delrows(Q) & " " & delrows(Q + 1)
'       If delrows(Q + 1) = 0 Then
'          Exit For
'       End If
'
'       If delrows(Q) > delrows(Q + 1) Then
'
'          temp = delrows(Q + 1)
'          delrows(Q + 1) = delrows(Q)
'          delrows(Q) = temp
'       'MsgBox delrows(Q) & " " & delrows(Q + 1)
'       End If
'
'    Next
'
'    tempstr = ""
'    For R = 1 To 10
'    tempstr = tempstr & " " & delrows(R)
'    Next
'    MsgBox tempstr


Sheets.Add
    For Q = 1 To 6500
       If delrows(Q) = 0 Then
          Exit For
       End If
```

```
    Range("A" & Q).Select
    ActiveCell.Formula = delrows(Q)
  Next

  Columns("A:A").Select
  Selection.Sort Key1:=Range("A1"), Order1:=xlAscending, Header:=xlGuess, _
    OrderCustom:=1, MatchCase:=False, Orientation:=xlTopToBottom, _
    DataOption1:=xlSortTextAsNumbers

  For Q = 1 To 6500
    Range("A" & Q).Select
    If ActiveCell.Formula = "" Then
    Exit For
    End If
    delrows(Q) = ActiveCell.Formula
  Next

  Application.DisplayAlerts = False
  ActiveWindow.SelectedSheets.Delete


'   tempstr = ""
'   For R = 1 To 10
'   tempstr = tempstr & " " & delrows(R)
'   Next
'   MsgBox tempstr
'



  For Q = 1 To 6500
    If delrows(Q) = 0 Then
      Exit For
    End If
    Rows(delrows(Q) - (Q - 1) & ":" & delrows(Q) - (Q - 1)).Select
    Selection.Delete Shift:=xlUp

  Next

  For Q = 1 To 6500
    delrows(Q) = 0
  Next



'   For Q = 1 To strRow
```

```
'
'       Range("A" & Q).Select
'       If ActiveCell.Formula = 1 Then
'          Rows(Q & ":" & Q).Select
'          Selection.Delete Shift:=xlUp
'          Q = Q - 1
'       End If
'
'    Next
'
'    Columns("A:A").Select
'    Selection.Delete Shift:=xlToLeft



    strDeleted = ""

Next 'M



For M = 1 To 7

    Sheets("Missing " & M & "0 pct").Select
    Application.Run "IDA.XLA!ESXMacro"
    Application.DisplayAlerts = False
    Sheets("Missing " & M & "0 pct").Delete

    Sheets("Mean " & M & "0 pct").Select
    Application.Run "IDA.XLA!ESXMacro"
    Application.DisplayAlerts = False
    Sheets("Mean " & M & "0 pct").Delete

    Sheets("Case Delete " & M & "0 pct").Select
    Application.Run "IDA.XLA!ESXMacro"
    Application.DisplayAlerts = False
    Sheets("Case Delete " & M & "0 pct").Delete


Next

    Sheets("Master").Select
    Application.Run "IDA.XLA!ESXMacro"
```

```
Dim RMS(0 To 7, 1 To 3) As String

Sheets("Master RES NN").Select
      Range("A1").Select
      Cells.Find(What:="Test Data RMS", After:=ActiveCell, LookIn:=xlFormulas, _
      LookAt:=xlPart, SearchOrder:=xlByRows, SearchDirection:=xlNext, _
      MatchCase:=False, SearchFormat:=False).Activate
      RMS(0, 1) = Right(ActiveCell.Formula, 5)
      RMS(0, 2) = Right(ActiveCell.Formula, 5)
      RMS(0, 3) = Right(ActiveCell.Formula, 5)


For M = 1 To 7
   Sheets("Missing " & M & "0 pct RES NN").Select
      Range("A1").Select
      Cells.Find(What:="Test Data RMS", After:=ActiveCell, LookIn:=xlFormulas, _
      LookAt:=xlPart, SearchOrder:=xlByRows, SearchDirection:=xlNext, _
      MatchCase:=False, SearchFormat:=False).Activate
      'MsgBox ActiveCell.Formula
      RMS(M, 1) = Right(ActiveCell.Formula, 5)
   Sheets("Mean " & M & "0 pct RES NN").Select
      Range("A1").Select
      Cells.Find(What:="Test Data RMS", After:=ActiveCell, LookIn:=xlFormulas, _
      LookAt:=xlPart, SearchOrder:=xlByRows, SearchDirection:=xlNext, _
      MatchCase:=False, SearchFormat:=False).Activate
      RMS(M, 2) = Right(ActiveCell.Formula, 5)
   Sheets("Case Delete " & M & "0 pct RES NN").Select
      Range("A1").Select
      Cells.Find(What:="Test Data RMS", After:=ActiveCell, LookIn:=xlFormulas, _
      LookAt:=xlPart, SearchOrder:=xlByRows, SearchDirection:=xlNext, _
      MatchCase:=False, SearchFormat:=False).Activate
      RMS(M, 3) = Right(ActiveCell.Formula, 5)
Next

Sheets.Add

Application.ActiveSheet.Name = "RMS"

Range("B1").Select
ActiveCell.Formula = "Missing"
Range("C1").Select
ActiveCell.Formula = "Mean"
Range("D1").Select
ActiveCell.Formula = "Case Delete"
```

```
For M = 0 To 7
Range("A" & M + 2).Select
ActiveCell.Formula = M & "0%"
Range("B" & M + 2).Select
ActiveCell.Formula = RMS(M, 1)
Range("C" & M + 2).Select
ActiveCell.Formula = RMS(M, 2)
Range("D" & M + 2).Select
ActiveCell.Formula = RMS(M, 3)
Next

Charts.Add
ActiveChart.ChartType = xlColumnClustered
ActiveChart.SetSourceData Source:=Sheets("RMS").Range("A1:D9"), PlotBy _
   :=xlColumns
ActiveChart.Location Where:=xlLocationAsNewSheet
With ActiveChart
   .HasTitle = False
   .Axes(xlCategory, xlPrimary).HasTitle = False
   .Axes(xlValue, xlPrimary).HasTitle = False
End With


End Sub




Public Function Rand(ByVal Low As Long, _
                ByVal High As Long) As Long
  Rand = Int((High - Low + 1) * Rnd) + Low
End Function


Public Function getColNum(ByVal strCol As String) As Integer

strCol = UCase(strCol)

If Len(strCol) = 2 Then

   letter1 = Left(strCol, 1)
   letter2 = Right(strCol, 1)
Else
   letter1 = strCol
```

```
End If



Select Case letter1
   Case "A"
      num1 = 1
   Case "B"
      num1 = 2
   Case "C"
      num1 = 3
   Case "D"
      num1 = 4
   Case "E"
      num1 = 5
   Case "F"
      num1 = 6
   Case "G"
      num1 = 7
   Case "H"
      num1 = 8
   Case "I"
      num1 = 9
   Case "J"
      num1 = 10
   Case "K"
      num1 = 11
   Case "L"
      num1 = 12
   Case "M"
      num1 = 13
   Case "N"
      num1 = 14
   Case "O"
      num1 = 15
   Case "P"
      num1 = 16
   Case "Q"
      num1 = 17
   Case "R"
      num1 = 18
   Case "S"
      num1 = 19
   Case "T"
      num1 = 20
```

```
    Case "U"
       num1 = 21
    Case "V"
       num1 = 22
    Case "W"
       num1 = 23
    Case "X"
       num1 = 24
    Case "Y"
       num1 = 25
    Case "Z"
       num1 = 26
End Select

Select Case letter2
    Case "A"
       num2 = 1
    Case "B"
       num2 = 2
    Case "C"
       num2 = 3
    Case "D"
       num2 = 4
    Case "E"
       num2 = 5
    Case "F"
       num2 = 6
    Case "G"
       num2 = 7
    Case "H"
       num2 = 8
    Case "I"
       num2 = 9
    Case "J"
       num2 = 10
    Case "K"
       num2 = 11
    Case "L"
       num2 = 12
    Case "M"
       num2 = 13
    Case "N"
       num2 = 14
    Case "O"
       num2 = 15
    Case "P"
```

```
      num2 = 16
   Case "Q"
      num2 = 17
   Case "R"
      num2 = 18
   Case "S"
      num2 = 19
   Case "T"
      num2 = 20
   Case "U"
      num2 = 21
   Case "V"
      num2 = 22
   Case "W"
      num2 = 23
   Case "X"
      num2 = 24
   Case "Y"
      num2 = 25
   Case "Z"
      num2 = 26
   Case Else
      num2 = ""
End Select



If Len(strCol) = 1 Then

   getColNum = num1
Else
   getColNum = num1 * 26 + num2

End If


End Function


Public Function getColLetter(ByVal intCol As Integer) As String
Dim num1 As Integer
Dim num2 As Integer



If intCol <= 26 Then
```

```
    getColLetter = numToLetter(intCol)
Else

    div = intCol / 26

    a = InStr(1, div, ".")
    If a > 0 Then
        num1 = Left(div, a - 1)
    Else
        num1 = div - 1
    End If

    num2 = intCol Mod 26

    If num2 = 0 Then
        num2 = 26
    End If

    getColLetter = numToLetter(num1) & numToLetter(num2)


End If

End Function



Public Function numToLetter(ByVal num As Integer) As String

Select Case num
    Case 1
        numToLetter = "A"
    Case 2
        numToLetter = "B"
    Case 3
        numToLetter = "C"
    Case 4
        numToLetter = "D"
    Case 5
        numToLetter = "E"
    Case 6
        numToLetter = "F"
    Case 7
        numToLetter = "G"
    Case 8
        numToLetter = "H"
```

```
    Case 9
        numToLetter = "I"
    Case 10
        numToLetter = "J"
    Case 11
        numToLetter = "K"
    Case 12
        numToLetter = "L"
    Case 13
        numToLetter = "M"
    Case 14
        numToLetter = "N"
    Case 15
        numToLetter = "O"
    Case 16
        numToLetter = "P"
    Case 17
        numToLetter = "Q"
    Case 18
        numToLetter = "R"
    Case 19
        numToLetter = "S"
    Case 20
        numToLetter = "T"
    Case 21
        numToLetter = "U"
    Case 22
        numToLetter = "V"
    Case 23
        numToLetter = "W"
    Case 24
        numToLetter = "X"
    Case 25
        numToLetter = "Y"
    Case 26
        numToLetter = "Z"
    Case Else
        numToLetter = "ERROR"
End Select


End Function
```